Ultra Low Power CMOS Design

by

Kyungseok Kim

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 9, 2011

Keywords: Ultra-Low Power Design, Subthreshold Circuits, Dual Voltage Design, Mixed Linear Integer Program, Gate Slack Analysis

Approved by

Vishwani D. Agrawal, Chair, James J. Danaher Professor of
Electrical and Computer Engineering
Victor P. Nelson, Professor of Electrical and Computer Engineering
Fa Foster Dai, Professor of Electrical and Computer Engineering

Abstract

The ubiquitous era of emerging portable devices demands long battery lifetime as a primary design goal. Subthreshold circuit design can reduce energy per cycle in an order of magnitude of nominal operating circuits by scaling power supply voltage (V_{dd}) below the device threshold voltage. But, it lowers significantly circuit performance as a penalty. Stringent energy budget and moderate speed requirements of ultra low power systems in the market may not be best satisfied just by scaling a single supply voltage. Optimized circuits with dual supply voltages provide an opportunity to resolve these demands.

Utilizing the time slack for dual-V_{dd} is a well-known technique for a circuit operating with nominal V_{dd} for reducing the power consumption with small extra cost in physical design. Most previous works in subthreshold circuit design only used a single supply voltage scaled down to reduce the energy consumption without considering the time slack.

We propose a method for minimum energy digital CMOS (Complementary Metal Oxide Semiconductor) circuit design using dual subthreshold supply. The delay penalty of a traditional level converter is unacceptably high when the voltages are in the subthreshold range. In this work, level converters are either not used at all or special multiple logic-level gates are used only when, after accounting for their cost, they offer advantage. Starting from a lowest energy per cycle design whose single supply voltage is in the subthreshold range, a new mixed integer linear program (MILP) finds a second lower supply voltage optimally assigned to gates with time slack. The MILP accounts for the energy and delay characteristics of logic gates interfacing two different signal levels. New types of linearized AND and OR constraints are used in this MILP. We show energy saving up to 24.5% over the best available designs of ISCAS'85 benchmark circuits.

For modern large VLSI systems, the MILP may suffer from unacceptable run-time as the MILP algorithm for dual voltage design has exponential-time complexity. Gate slack analysis gives an opportunity to reduce the time complexity as linear for assigning the optimal lower supply voltage (V_{DDL}) to initially all higher supply voltage (V_{DDH}) gates in a single-V_{dd} circuit. The slack of a gate in a digital circuit is the difference between the critical path delay and the delay of the longest path through that gate. Using the previous work on static timing analysis, we have developed a linear-time algorithm for computing the slack for all gates in a circuit.

We propose a new slack-time based algorithm for dual-V_{dd} design to achieve maximum energy saving. For a given lower supply voltage, we first compute slacks for all gates of the circuit and then partition them into three groups. In one group, all gates can be unconditionally assigned the low voltage. In the second group, no gate can be assigned low voltage. In the third group, low voltage assignment to any single gate will not violate the critical path timing and, therefore, the low voltage must be sequentially assigned to gates one at a time. Because all steps of the voltage assignment algorithm rely on linear-time analysis, the overall complexity of this energy optimization method is close to linear in the number of gates. We apply our algorithm to optimize ISCAS'85 benchmark circuits and compare the results with those from MILP. Energy savings from the new slack-time based algorithm is very closed to the global optimum MILP solutions. The optimization time using gate slack can be as low as 1/43 when compared to that of the MILP method for dual-V_{dd} design. The new slack-time based algorithm is especially beneficial for large circuits, which may contain few critical or near-critical paths and many paths with large slack.

Acknowledgments

Without seamless encouragement, guidance, and support from my advisor, Professor Vishwani D. Agrawal, the dissertation would not have been written. First, I am deeply thankful to him as a very generous mentor throughout my doctoral studies. The work has been delightful and successful under his valuable advice.

I would like to thank Professor Victor P. Nelson and Professor Fa Foster Dai for their great suggestions as my advisory committee members and through their distinguished lectures. I am grateful to Professor Allen Landers for serving as the outside reader for my dissertation and his valuable suggestions. I am also grateful to Professor Prathima Agrawal, the Director of Wireless Engineering Research and Education Center (WEREC), for providing financial support for my research.

I sincerely appreciate our former and current colleagues for invaluable discussion and encouragement. Thanks to Nitin, Jins, Lu, Hillary, Khushboo, Ashfaq, Fan, Wei, Yu, Manish, Mridula, Priya, Rakshith, Jia, Murali, Lixing and Suraj. I would like to thank my friends for unforgettably joyful memories at Auburn.

Finally, I would like to thank my parents for their endless love and support during my whole life. I am grateful to my brother and his family for their encouragement. I am greatly thankful to my wife and lovely daughter for their patience and support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ultra-low power applications such as micro-sensor networks, pacemakers, and many portable devices require extreme energy constraint for long battery lifetime. Subthreshold operation presents an opportunity for such energy-constrained applications with its very low energy consumption [32, 62, 69, 76, 77, 84]. Subthreshold circuits offer a promising solution for implementing highly energy-constrained systems in clock ranges of low to medium frequencies for remote or mobile applications.

As the power supply voltage (V_{dd}) is scaled below the device threshold voltage (V_{th}), the subthreshold current ever so slowly charges and discharges nodes for the circuit's logic function [76]. This weak driving current inherently limits the performance but minimum energy operation of the circuit is achieved with reduced dynamic and leakage power, resulting in long battery life [36, 37, 38].

In the past decades, subthreshold circuit design was not well recognized in the area of digital circuits as high performance demand was a major concern. Lately, however, portability has become a trend in the electronics marketplace. Low energy per operation is a primary design parameter in such applications. Without the performance requirement, a subthreshold circuit can operate at its minimum energy operating point that is only slightly above the absolute minimum voltage [81] that would guarantee the correct logic function. Even for applications requiring high peak performance, ultra-dynamic voltage scaling (UDVS) [8] can provide an opportunity for subthreshold circuit design that would switch between a nominal voltage high performance mode and an energy efficient subthreshold mode according to the system workload.

To support more features or long uninterrupted operation in energy constrained systems, subthreshold circuit designers strive to further increase the performance or reduce the energy consumption, as much as possible. These enhancements can be achieved by utilizing the time slack in subthreshold circuits using the new design methodologies proposed in this dissertation.

1.1 Motivation

Subthreshold circuit design is suitably applicable for emerging portable applications that need tremendously low energy operation. The limitation of this technique is very slow speed of operation due to the extremely scaled down supply voltage. Despite a very high energy efficiency, the subthreshold design has been applied only in niche markets due to its low performance. Depending upon the application, size, weight and cost can be equally important as performance. Especially for remote, portable and mobile applications, low-power has significance. Reduced power consumption makes the circuits lighter, reduces or eliminates cooling subsystems, and reduces the weight and extends the life of the energy source.

According to the available literature, most low-power techniques exploit time slack on non-critical paths of a circuit to reduce power consumption without performance loss. These techniques have been applied to circuits operating with the nominal supply voltage by sizing device widths, using multi-V_{th} devices, or using multiple V_{dd} [64, 50, 79]. For subthreshold circuits, the technique of sizing device width affects the correct logic function of CMOS (Complementary Metal Oxide Semiconductor) circuits at low supply voltage [76]. The multi-V_{th} technique does not adequately utilize the time slack in the subthreshold regime [4], because semiconductor foundries normally provide standard cell libraries with two to three fixed V_{th} values, namely, high V_{th}, standard V_{th}, and low V_{th}, for low-power design. Gate delay exponentially depends on V_{th} in a subthreshold circuit. Therefore, we cannot utilize all possible

time slack on non-critical paths in a subthreshold circuit without further manipulation of these device threshold voltages.

The multi-V_{dd} technique has been widely implemented for two supply voltages [41]. The dual-V_{dd} design is best suited for exploiting the time slack in a subthreshold circuit as well. Although the gate delay exponentially depends on V_{dd} in the subthreshold region it may be possible to find an optimal lower supply voltage for the available time slack in the circuit. A DC to DC voltage converter [57] will then allow the voltage management.

There are two scenarios for applying dual-V_{dd} design to subthreshold circuits in energy constrained low-performance applications. Consider a digital circuit working in an absolutely minimum energy consumption mode. The supply voltage for such an operation is known to be in the subthreshold range [76]. First, we can further reduce the energy consumption without changing the performance by assigning an extra lower supply voltage. The lower voltage is supplied to gates on non-critical paths. Alternatively, the subthreshold circuit can be sped up by several times by selecting two supply voltages, one of which is higher than the optimal single V_{dd}. In this scenario, the dual-V_{dd} design retains the energy consumption close to that of the minimum energy point but operates at a higher speed obtained by using the higher supply for gates on critical paths.

1.2 Problem Statement

The aim of this dissertation is:

- Investigate the validation of dual-V_{dd} design for bulk CMOS subthreshold circuits.
- Develop new mixed integer linear programs (MILP) that automatically and optimally assign gate voltages and maintain a wide range of speed requirements for a given circuit, while minimizing the total energy per cycle.
- Develop new methods for dual-V_{dd} design using linear-time gate slack analysis to reduce computation time for optimization.

1.3 Contribution of the Dissertation

In this dissertation, we propose a framework for finding the optimal dual-V_{dd} assignment in subthreshold circuits to achieve minimum energy design. The minimum energy per cycle operation with a very low single voltage in the subthreshold region is known [76]. We further lower the energy per cycle below that point by using dual subthreshold supply. Without a proper level converter for this mode, special considerations are used in the design for eliminating or substituting the level converters that otherwise would have unacceptable delay overhead. For a wide range of speed requirements, new mixed integer linear programs (MILP) globally determine an energy-efficient circuit configuration by assigning an extra supply voltage V_{DDL} to gates on non-critical paths. This work could provide solutions for the demands of either lower energy or higher performance in subthreshold design applications. A subthreshold circuit is susceptible to process variation [20, 72], which affects the delay of gates. We investigate the benefit of dual-V_{dd} design for reducing the delay variability of a subthreshold circuit with process variation. To the best of our knowledge this work is the first to present a dual-V_{dd} scheme for subthreshold logic circuits to achieve lower minimum energy, which is an improvement over the known minimum energy operating point.

The new design procedure formulates mixed integer lineal programs (MILP) that, given today's computing capabilities, can deal with moderately large circuit complexity [19]. But, the exponential time complexity of the MILP method for energy optimized circuits may not be acceptable for modern VLSI (Very Large Scale Integration) systems. We propose a new slack-time based algorithm to save computation time and obtain a nearly global solution similar to that obtained by an MILP. The new technique is highly efficient and gives a quality of solution very close to the MILP. The time complexity of the basic slack analysis algorithm is linear in total number of gates, while the heuristic algorithms of dual-V_{dd} design in the literature still have polynomial time complexity $O(n^2)$ [13]. The proposed method of gate slack analysis can be applicable for other low-power design techniques to quickly

classify positive slack gates available for possible power-optimization in a large circuit. This approach reduces the optimization effort and saves run-time of the algorithms.

1.4 Organization of the Dissertation

The dissertation is organized as follows. Chapter 2 briefly provides an overview of subthreshold circuit design with a perspective of minimum voltage and minimum energy operation.

Chapter 3 demonstrates a new MILP algorithm for minimum energy design using dual-V_{dd} in the subthreshold regime. Unacceptable delay overhead of a level converter is avoided in the optimized circuit by using topological constraints in the MILP.

In Chapter 4, we propose another new MILP algorithm for minimum energy design with dual subthreshold supply and multiple logic-level gates. Multiple logic-level gates that suppress DC leakage currents are inserted to remove topological constraints and further improve the energy saving for the optimized circuit.

Chapter 5 investigates process variation effects on minimum energy design using dual subthreshold supply. An optimized circuit shows more immunity to process variation with technology scaling.

In Chapter 6, we propose a new slack-time based algorithm for dual-V_{dd} design. Gate slack analysis is used to reduce the time complexity of the optimization process in the minimum energy design.

Finally, the conclusion and ideas for the future advancement of this work are given in Chapter 7.

Chapter 2

Overview of Subthreshold Circuit Design

In this chapter, we provide the fundamental aspects of subthreshold design for ultra-low power circuits [76]. A description of subthreshold circuit properties as given here will be helpful to illustrate our proposed methods in this dissertation.

2.1 Origin of Subthreshold Circuit Design

The MOS (Metal Oxide Semiconductor) transistor conducts current, majority carriers, through an inverted channel between the source and drain caused by a nominal voltage applied to the gate. When a low voltage is applied to the gate, majority carriers in the substrate are repelled from the surface directly below the gate. Then, a depletion charge of immobile atoms forms a depletion region beneath the gate. The minority carriers in the depletion layer are made to move by diffusion and induce a drain current by applying a voltage between the drain and source in the MOS device. This weak inversion current was considered to be insignificantly small and ignored in digital circuit design until the recent decade.

As is relevant to the electronic wrist watch design [74, 75], the properties of MOS transistors have been investigated at a very low current level. The study uncovered an unusual exponential relationship of the drain current with the gate voltage. Figure 2.1 shows the first measurement of drain current of an MOS transistor below the device threshold voltage. This weak inversion current has been named the subthreshold current.

The early exploration of subthreshold design was focused on analog circuits such as amplitude detector, quartz ring oscillator, bandpass amplifier, and transconductance amplifier [29, 44, 73]. In the past years, subthreshold digital CMOS designs have been implemented

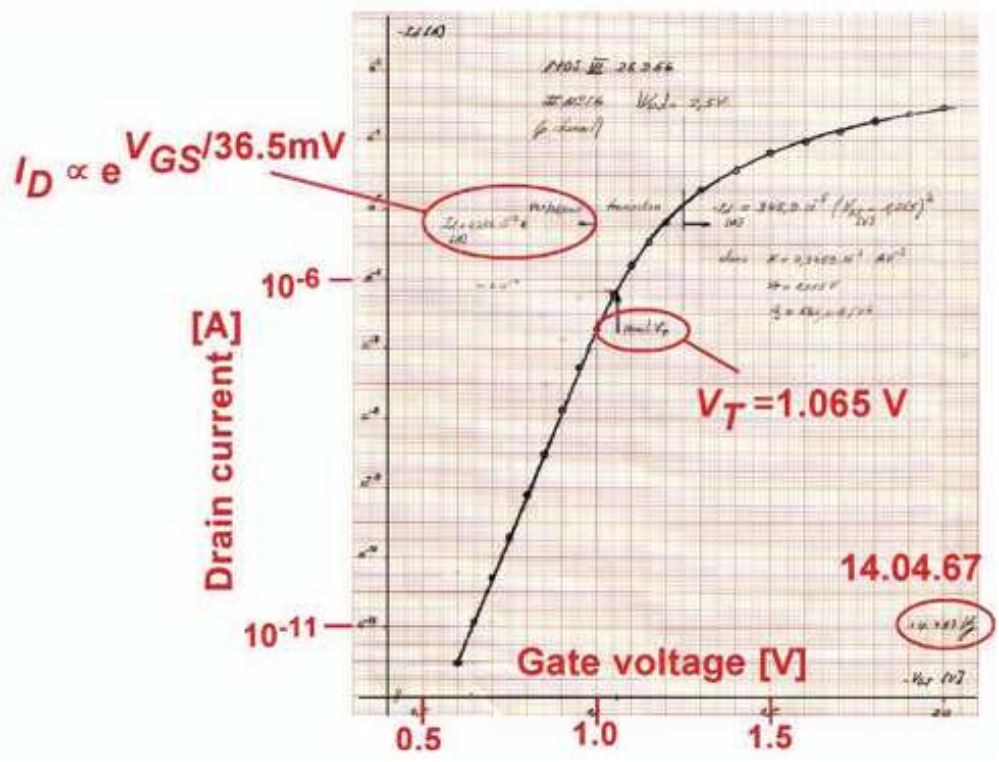


Figure 2.1: First measurement of an MOS transistor at very low current (annotated copy of Vittoz's notebook [75]).

for biomedical devices, FFT processors, and SRAMs [24, 32, 62, 77, 83, 43]. This unintended discovery provides an opportunity for meeting the demands of extreme energy efficient systems.

2.2 Minimum Voltage Operation

In 1972, Swanson and Meindl built a revised charge based model for an inverter, considering the weak and strong mixed inversion region [66]. Previously, their model [49] only considered both weak and strong inversion currents, but there was discontinuity in the model at the point where two regions meet. The revised model was used to analyze the voltage transfer characteristic (VTC) of the inverter that demonstrated operation down to 100mV, as shown in Figure 2.2. The off-currents for PMOS and NMOS transistors were equated and the gain of the inverter was calculated in the subthreshold region for finding the minimum

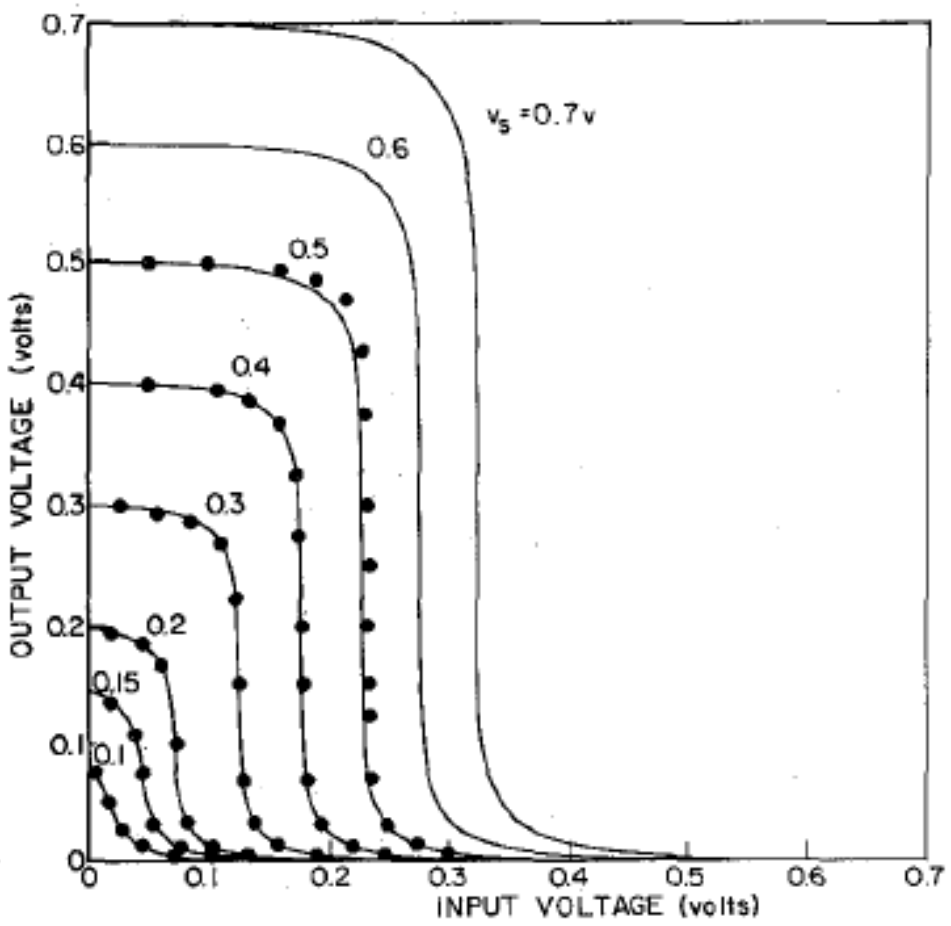


Figure 2.2: CMOS inverter voltage transfer characteristics (VTC) [66].

voltage. For sufficient gain at $V_{dd}/2$, the minimum voltage was considered as $8kT/q$, or 200mV at room temperature, based on device parameters at that time. The term kT/q is the thermal voltage (V_T).

The ideal limit for lowest operable voltage was expected to be $2kT/q$, or 57mV at room temperature, in 2001 [6]. To achieve this ideal limit, the PMOS and NMOS device threshold voltages in the inverter must be adjusted to ensure comparable off-currents for the two MOS devices. Otherwise, minimum voltage larger than $2kT/q$ is needed to guarantee the correct logic function. The circuits with very low supply voltages were successfully fabricated in standard 1.5V 180nm CMOS technology.

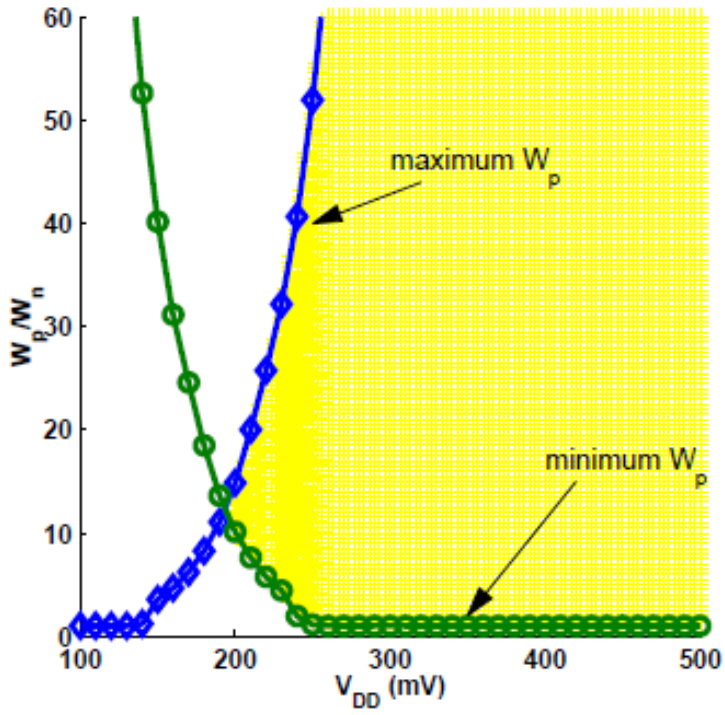


Figure 2.3: Minimum voltage operation for 10%-90% output swing for a 0.18μm ring oscillator [10].

Another approach for the minimum voltage limit was derived by balancing the threshold voltages of PMOS and NMOS transistors [52]. The use of the proposed V_{th} matching scheme reduces the lowest required supply voltage to 0.15V$\sim$0.30V for SRAM and enables CMOS LSI minimum supply voltage at 0.1V.

At very low supply voltage, sizing of a transistor affects the functionality of CMOS logic circuits. The minimum voltage operation (V_{min}) occurs when the currents of PMOS and NMOS devices are the same [61]. In Figure 2.3, the shaded region is the operational region of a ring oscillator. The line of maximum W_p guarantees the output voltage of an inverter for logic zero below 10% of V_{dd}. Large width of a PMOS device increases logic 0 level at the output from the subthreshold leakage through the PMOS device for a smaller NMOS device. Conversely, the minimum W_p line shows the output voltage of the inverter for logic 1 always maintains above 90% of V_{dd}. The output voltage of the inverter is reduced by

the subthreshold leakage through the larger NMOS device. The minimum voltage operation occurs at the point where maximum W_p is equal to minimum W_p and maintains the 10% to 90% output voltage swing. The ratio of the PMOS size to NMOS size is 12 for V_{min} in 0.18μm technology [10]. This ratio means that the subthreshold current of a unit width NMOS transistor is 12 times larger than that of a unit width PMOS transistor by technology imbalance.

Process variations affect the strength of the current for both devices [9]. To find minimum voltage operation considering process variations, maximum W_p should be defined at the worst case process corner, i.e., the strong PMOS and weak NMOS corner. For minimum W_p, the worst case corner of the weak PMOS and strong NMOS should be considered. Minimum energy operation of a circuit always occurs above V_{min} for the correct logic function.

2.3 Minimum Energy Operation

The minimum energy operation point (E_{min}) for a digital circuit means that the circuit consumes less *Energy per cycle* than any other point in the parameter space. Among the different parameters, power supply voltage (V_{dd}) and device threshold voltage (V_{th}) are mainly considered for the minimum energy point. The energy and delay contours for a ring oscillator circuit with varying V_{dd} and V_{th} show that E_{min} occurs in the subthreshold region [78].

For given V_{dd} and V_{th}, the minimum energy point for a circuit is determined by the relationship between energy and latency. As V_{dd} scales down, dynamic energy is quadratically reduced, while the delay of a circuit exponentially increases at supply voltages below V_{th}. The increased delay induces an exponential increase of leakage energy. The minimum energy point occurs where the magnitudes of dynamic energy and leakage energy are equal, as shown in Figure 2.4.

The switching activity of a circuit affects its minimum energy point. When the dynamic energy is decreased by reducing switching events, the leakage energy remains constant with switching activity. Thus, the leakage energy contributes substantially more to the total

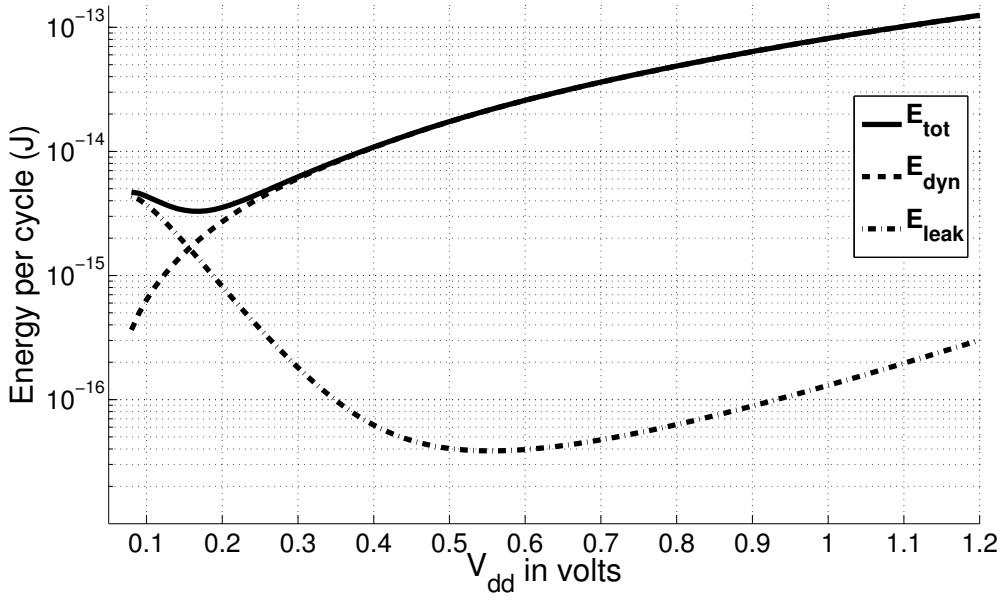


Figure 2.4: Energy per cycle for an 8-bit ripple carry adder through HSPICE [27] simulation in PTM 90nm CMOS, E_{min} = 3.29fJ at V_{dd} = 0.17V ($V_{th,pmos}$ = -0.21V and $V_{th,nmos}$ = 0.29V).

energy of a circuit. In that case, the minimum energy point occurs at higher supply voltages compared to higher activity circuits. Adversely, higher switching circuits move the minimum energy point to lower supply voltages to suppress the dynamic energy.

There are two representative minimum energy models in the literature. First, when the operating frequency and technology of a subthreshold circuit are given, the minimum energy model is derived to obtain the closed forms for optimal V_{dd} and V_{th}, respectively [7, 76]. This model uses fitting parameters normalized to a characteristic inverter for the given technology, where the minimum sized inverter, for simplicity, is a good choice. All other gates are normalized with respect to the inverter.

The delay of a characteristic inverter with output capacitance C_g is derived in subthreshold region as [51],

$$t_d = \frac{K \cdot C_g \cdot V_{dd}}{I_{o,g} \exp\left(\frac{V_{dd}-V_{th,g}}{mV_T}\right)} \tag{2.1}$$

where K is a delay fitting parameter, m is the subthreshold slope coefficient, and $I_{o,g}$ and $V_{th,g}$ are fitted parameters for the on-currents of a NMOS and PMOS transistor that are not symmetrical.

The longest (critical) path delay of a circuit is obtained as,

$$T_D = t_d L_{DP} \tag{2.2}$$

where L_{PD} is the logic depth of the longest path normalized to the characteristic inverter delay.

Subthreshold leakage current is not the only component for the leakage of nanometer CMOS transistors. But, the leakage energy mainly comes from subthreshold leakage in a circuit operating in the subthreshold region. From this assumption, total energy per cycle (E_{tot}) and its components, dynamic energy (E_{dyn}) and leakage enrgy (E_{leak}), are expressed as,

$$\begin{aligned} E_{dyn} &= C_{eff} V_{dd}^2 \\ E_{leak} &= I_{leak} V_{dd} T_D \\ &= W_{eff} I_{o,g} \exp\left(\frac{-V_{th,g}}{m V_T}\right) V_{dd} t_d L_{DP} \\ &= W_{eff} K C_g L_{DP} V_{dd}^2 \exp\left(\frac{-V_{dd}}{m V_T}\right) \\ E_{tot} &= E_{dyn} + E_{leak} \\ &= V_{dd}^2 \left(C_{eff} + W_{eff} K C_g L_{DP} \exp\left(\frac{-V_{dd}}{m V_T}\right)\right) \end{aligned} \tag{2.3}$$

where C_{eff} is the average total switched capacitance for the circuit and W_{eff} is the average total width that contributes to the leakage current. The derivative of total energy with respect to V_{dd} is given by

$$\frac{\partial E_{tot}}{\partial V_{dd}} = 2 C_{eff} V_{dd} + \left(2 - \frac{V_{dd}}{m V_T}\right) W_{eff} K C_g L_{DP} V_{dd} \exp\left(\frac{-V_{dd}}{m V_T}\right) \tag{2.4}$$

To solve for the optimal voltage (V_{opt}) for minimum energy, Equation (2.4) is set to zero and an analytical solution for V_{opt} is obtained:

$$V_{opt} = mV_T \left(2 - \text{lambert}W\left(\frac{-2C_{eff}}{W_{eff}KC_gL_{DP}}\exp(2)\right)\right) \tag{2.5}$$

The Lambert W function is subject to the constraint [16]:

$$\frac{-2C_{eff}}{W_{eff}KC_gL_{DP}}\exp(2) > -\exp(-1) \tag{2.6}$$

For obtaining $V_{th,opt}$, the operating frequency for the circuit is given by

$$f = \frac{1}{t_dL_{DP}} \tag{2.7}$$

and Equation (2.1) substitutes t_d for a given f:

$$V_{th,opt} = V_{opt} - mV_T \ln\left(\frac{fKC_gL_{DP}V_{opt}}{I_{o,g}}\right) \tag{2.8}$$

When the natural log argument exceeds 1, the circuit no longer operates in subthreshold region, $V_{th,opt} < V_{opt}$. This limits the maximum operating frequency for a subthreshold circuit. From Equations (2.5) and (2.8), the energy optimal voltage and device threshold voltage are determined for a given performance. For a given V_{th} with respect to the technology, the energy optimal voltage is still determined by Equation (2.5) and the corresponding operating frequency is given by Equation (2.7).

When V_{dd} reduces, the delay and leakage current of a circuit change simultaneously. The leakage current reduces due to *drain-induced barrier lowering* (DIBL) effect, while the delay increases exponentially in subthreshold regime. The leakage energy is the product of delay and leakage current, but the delay induces the overall leakage energy increase. Figure 2.5 shows the trends of normalized t_d and I_{leak} for an inverter in the Predictive Technology

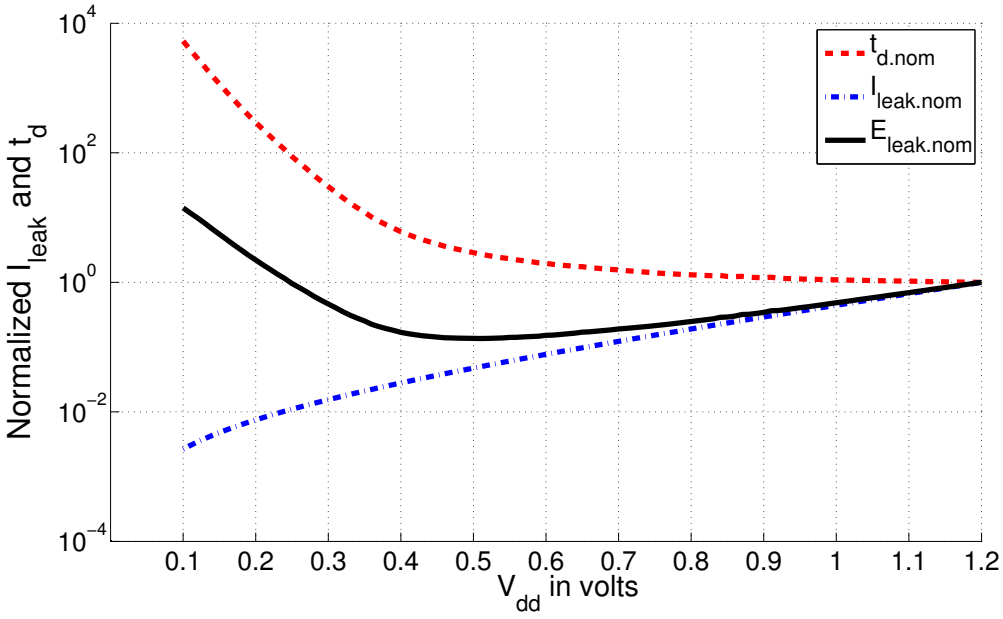


Figure 2.5: The delay and leakage current normalized to an inverter at $V_{dd} = 1.2$V through HSPICE simulation in PTM 90nm CMOS.

Model (PTM) 90nm CMOS technology [85]. The normalized leakage energy, $E_{leak,nom}$, starts to increase at the beginning of the subthreshold region.

Another minimum energy model is derived from an analytical expression for the energy consumption of an n-stage inverter chain as a function of V_{dd} [81]. The total energy per cycle of an n-stage inverter chain with switching activity α is given by:

$$\begin{aligned} E_{tot} &= E_{dyn} + E_{leak} \\ &= \alpha \cdot n \cdot E_{switch,inv} + P_{leak} \cdot T_d \\ &= \alpha \cdot n \cdot \left(\frac{1}{2} \cdot C_s \cdot V_{dd}^2\right) + (n \cdot V_{dd} \cdot I_{leak}) \cdot (n \cdot t_d) \\ &= \frac{1}{2} \cdot \alpha \cdot n \cdot C_s \cdot V_{dd}^2 + n \cdot V_{dd} \cdot I_{leak} \cdot n \cdot \frac{\eta C_s V_{dd}}{2 I_{on}} \\ &= \frac{1}{2} n C_s V_{dd}^2 \cdot \left(\alpha + \eta \cdot n \cdot \frac{I_{leak}}{I_{on}}\right) \\ &= \frac{1}{2} n C_s V_{dd}^2 \cdot \left(\alpha + \eta \cdot n \cdot e^{-\frac{V_{dd}}{m V_T}}\right) \end{aligned} \tag{2.9}$$

Where, the symbols used in these expressions are listed below:

- n: number of inverter stages.
- $E_{switch,inv}$: switching energy of an inverter.
- P_{leak}: total leakage power of the inverter chain.
- T_d: delay of the inverter chain.
- C_s: total switched capacitance of an inverter.
- t_d: delay of an inverter.
- I_{on}: average on-current of an inverter in subthreshold region.
- η: technology-dependent linear coefficient for the gap of inverter delay between actual and step delay.

The energy optimal voltage is obtained by equating $\partial E_{tot}/\partial V_{dd} = 0$. From setting $u = \eta \cdot n/\alpha$ and $t = V_{dd}/mV_T$, the minimum energy is achieved by the supply voltage V_{dd} that satisfies the following equation:

$$e^t = \frac{u}{2} \cdot t - u \tag{2.10}$$

Equation (2.10) is solved using curve-fitting to get the closed-form expression due to its non-linear characteristic:

$$t = 1.587 \ln u - 2.355 \tag{2.11}$$

By replacing u and t with the original variables, the energy optimal voltage is finally obtained as:

$$V_{opt} = \left(1.587 \ln \left(\eta \cdot \frac{n}{\alpha}\right) - 2.355\right) \cdot mV_T \tag{2.12}$$

The energy optimal voltage only depends on η and m for technology trends. Also, V_{th} does not affect the minimum energy and energy optimal voltage as seen in Equations (2.9)

and (2.12). The dependency of the leakage current and delay on V_{th} is the same, but opposite. Therefore, the leakage energy is constant with different V_{th} values, not as V_{dd} as shown in Figure 2.5, in subthreshold regime. The minimum energy and optimal voltage are strongly determined by α and n, which account for the relative amounts of dynamic and leakage energies in the total energy, respectively.

For large complex circuits, Equation (2.9) is extended as follows:

$$\begin{aligned} E_{dyn} &= \alpha \cdot S_{HD} \cdot C_{w0} \cdot W_{tot} \cdot V_{dd}^2 \\ E_{leak} &= I_{leak} \cdot V_{dd} \cdot T_c \\ &= (\gamma \cdot W_{tot} \cdot I_{leak0}) \cdot V_{dd}^2 \cdot (n_d \cdot t_{d,FO4}) \end{aligned} \tag{2.13}$$

$$\begin{aligned} E_{tot} &= E_{dyn} + E_{leak} \\ &= C_{w0} W_{tot} V_{dd}^2 \left(\alpha S_{HD} + 2\gamma \cdot n_d \cdot e^{-\frac{V_{dd}}{mV_T}} \right) \end{aligned}$$

where the delay of an inverter with fanout of four (FO4) is given with I_{on0}, on-current of a unit width inverter:

$$t_{d,FO4} = \frac{\frac{1}{2} \cdot (4W_{inv} \cdot C_{w0}) \cdot V_{dd}}{W_{inv} \cdot I_{on0}} \tag{2.14}$$

where,

- S_{HD}: switching factor to model the hamming distance of inputs [21].
- C_{w0}: capacitance of a unit width transistor.
- W_{tot}: total width of transistors in a circuit.
- T_c: critical path delay of a circuit.
- γ: leaking factor to model leakage stack effect and input pattern dependency.
- I_{leak0}: leakage current of a unit width transistor.

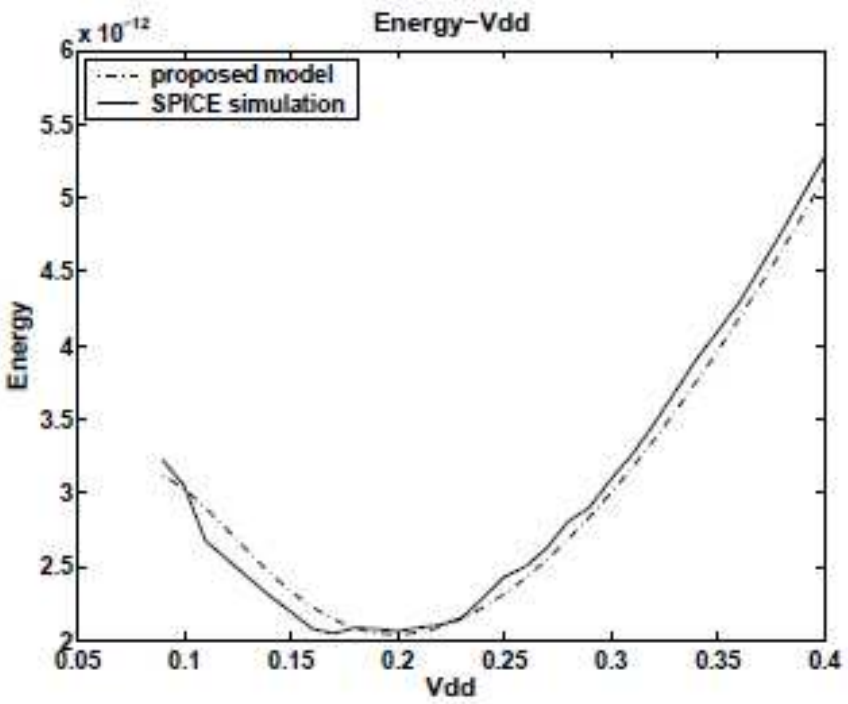


Figure 2.6: Total Energy vs. V_{dd} for a 16×16 multiplier [81].

- n_d: logic depth in terms of inverter delay with fanout of four.

As shown in Figure 2.6, the proposed total energy model is compared to SPICE simulation results for a 16×16 multiplier circuit, where the parameters used in the SPICE simulation are $S_{HD} \approx 0.55$, $\gamma \approx 0.5$, and $n_d \approx 65$. The switching activity for each block has a different value. Thus, we should consider the switching activity difference across the entire chip for minimum energy point. Low switching activity in a circuit corresponds to greater logic depth with normal switching activity when V_{dd} is scaled down to achieve E_{min}.

Chapter 3

True Minimum Energy Design Using Dual Below-Threshold Supply Voltages

This chapter investigates subthreshold voltage operation of digital circuits. Operation in the subthreshold voltage region has been long predicted and since verified [76]. To exploit the time slack on non-critical paths, some designs use dual voltages within a circuit. Although dual voltage operation for above threshold V_{dd} has been studied [11, 39, 65, 67, 68], below-threshold dual voltages have not been examined until the work presented here. Utilizing the time slack for dual-V_{dd} assignment can give valuable energy saving with small extra cost in physical design. This results in circuit operation below the minimum energy point for a single-V_{dd} circuit. Therefore, we call this the true minimum energy point.

We provide a framework for optimizing subthreshold circuits using dual-V_{dd} assignments with given speed requirements, where the design procedure formulates mixed integer lineal programs (MILP). In a dual-V_{dd} circuit, signal level converters are considered essential. Level converters insert delays and consume power [54, 80]. In the absence of level converters, certain interfaces become unsatisfactory. Especially, driving a high V_{dd} gate with a low voltage signal presents problems of high leakage and long delay. We characterize the multi-level interfaces and our MILP contains constraints to avoid the use of level converters.

3.1 Subthreshold Circuits

Before optimizing the minimum energy of subthreshold circuits by dual-V_{dd} assignments, we briefly summarize the properties of subthreshold circuits in terms of functional operation and failure, performance, and energy in this section.

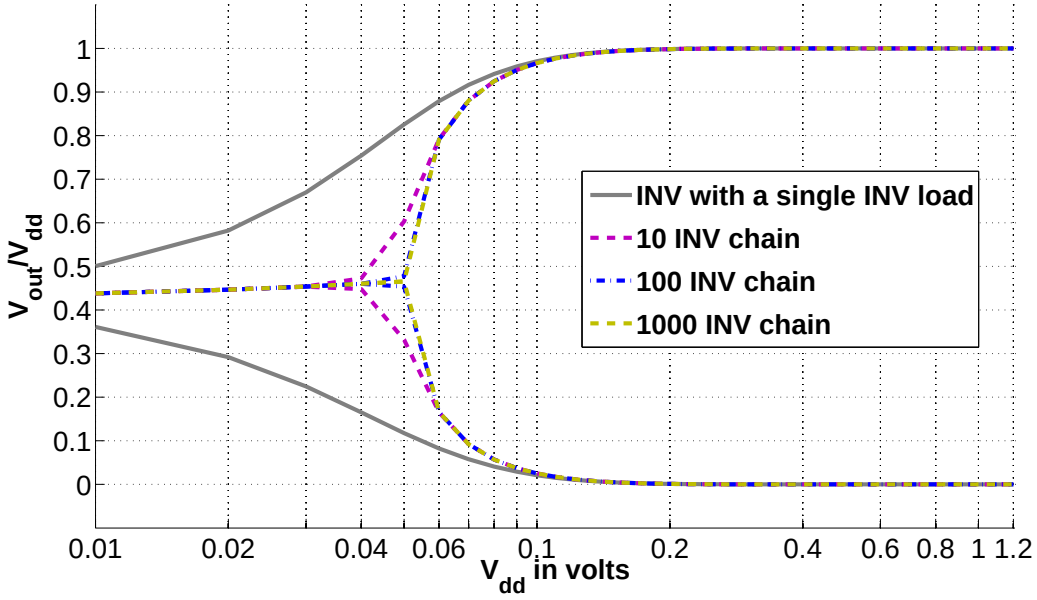


Figure 3.1: HSPICE [27] simulations for the output logic levels of inverter chains normalized to nominal supply voltage, 1.2V, with scaling V_{dd} in PTM 90nm CMOS (INV: $W_p = 5.5 \cdot L_g$, $W_n = 2.4 \cdot L_g$).

3.1.1 Minimum Operating Voltage

For the correct functional operation of a subthreshold logic circuit, the supply voltage V_{dd} should be higher than a certain minimum voltage (V_{min}). For bulk CMOS technology, the theoretical V_{min} is given as [48, 81],

$$V_{min} = 2 \cdot V_T \cdot ln\left(1 + \frac{S}{ln10 \cdot V_T}\right) \tag{3.1}$$

where $V_T = kT/q$ is the thermal voltage, $k = 1.381 \times 10^{-23}$ J/K is Boltzmann's constant, T is absolute temperature in Kelvin, $q = 1.602 \times 10^{-19}$ C is electronic charge and S is the subthreshold swing. From [23], S is degraded with the downscaling trend of CMOS technology, which means that the reduced ratio of on-current I_{on} at $V_{gs} = V_{ds} = V_{dd}$ to off-current I_{off} at $V_{gs} = 0$ and $V_{ds} = V_{dd}$ in subthreshold region ($V_{dd} < V_{th}$) causes smaller noise margins and possible functional logic failures at or below V_{min}. Figure 3.1 shows the

inverter chains work properly at lower supply voltages. The minimum operating voltage of the inverter chains, 80mV, guarantees 10% to 90% output voltage swing. The increased number of inverters in a chain slightly degrades V_{min}, but the degradation is saturated. Basically, this means that the logic 0 and 1 levels stabilize close to ground and supply voltages, respectively, and do not continue to degrade with the depth of the circuit.

3.1.2 Delay

The delay of a gate in a subthreshold circuit can be simply formulated from the CMOS gate delay equation [23],

$$t_d = \frac{K \cdot C_L \cdot V_{dd}}{I_{on}} \tag{3.2}$$

where K is a fitting parameter and C_L is the load capacitance of the gate. If it is assumed that total subthreshold current is equal to subthreshold drain current (I_{sub}), we replace I_{on} with I_{sub} [76]

$$I_{sub} = I_o \cdot 10^{\left(\frac{V_{gs}-V_{th}+\eta V_{ds}}{S}\right)} \cdot \left(1 - e^{\frac{-V_{ds}}{V_T}}\right) \tag{3.3}$$

where η is the *drain-induced barrier lowering* (DIBL) coefficient and I_o is the drain current at $V_{gs} = V_{th}$ in the weak inversion [58].

$$I_o = \mu_o \cdot C_{ox} \cdot \frac{W}{L} \cdot (m-1) \cdot V_T^2 \tag{3.4}$$

μ_o is the zero bias electron mobility, C_{ox} is the gate oxide capacitance, and m is the subthreshold slope coefficient.

When $V_{gs} = V_{ds} = V_{dd} \gg V_T$ ($\approx$ 26mV at 300K), we get gate delay as,

$$t_d = \frac{K \cdot C_L \cdot V_{dd}}{I_o \cdot 10^{\left(\frac{(\eta+1)V_{dd}-V_{th}}{S}\right)}}. \tag{3.5}$$

Thus, t_d is exponentially dependent on V_{dd}, V_{th}, η, and S.

3.1.3 Energy

Energy per cycle of a circuit is a key parameter for energy efficiency in ultra-low power applications. Because computing workload is characterized in terms of clock cycles, this measure directly relates energy consumption to the workload. Before considering the energy consumed by a circuit, we start by examining the total energy per cycle (E_{tot}) of a single gate, which is composed of dynamic energy (E_{dyn}) and leakage energy (E_{leak}):

$$\begin{aligned} E_{dyn} &= \alpha_{0\to1} \cdot C_L \cdot V_{dd}^2 \\ E_{leak} &= P_{leak} \cdot t_d \\ &= I_{off} \cdot V_{dd} \cdot t_d \\ &= K \cdot C_L \cdot V_{dd}^2 \cdot 10^{\frac{-V_{dd}}{S}} \\ E_{tot} &= E_{dyn} + E_{leak} \\ &= \left(\alpha_{0\to1} + K \cdot 10^{\frac{-V_{dd}}{S}}\right) \cdot C_L \cdot V_{dd}^2 \end{aligned} \tag{3.6}$$

where $\alpha_{0\to1}$ is the low to high transition activity for the gate output node and P_{leak} is static leakage power. I_{off} is static leakage current and presented by (3.3) :

$$I_{off} = I_o \cdot 10^{\left(\frac{-V_{th}+\eta V_{ds}}{S}\right)} \qquad V_{ds} \gg V_T \tag{3.7}$$

3.2 Dual-V_{dd} Scheme for Subthreshold Operation

Scaling V_{dd} down in circuits reduces both dynamic power and static leakage power besides reducing the performance. To reduce power consumption without degrading performance, a multi-V_{dd} technique exploits time slacks and lowers voltage V_{DDL} for gates on non-critical paths.

As shown in Figure 3.2(a), a *clustered voltage scaling* (CVS) algorithm [67] does not allow the V_{DDL} cells to feed directly into V_{DDH} cells and so level converting is implemented inside the filp-flop (LCFF) [28]. This topological limitation reduces full use of time slacks that

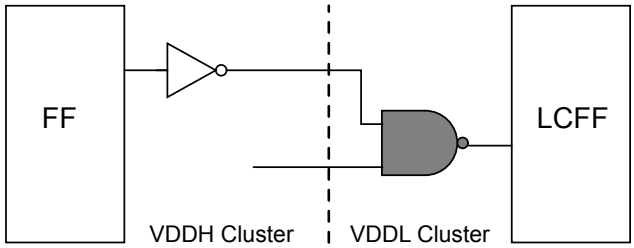
exist in a circuit. The *extended clustered voltage scaling* (ECVS) in Figure 3.2(b) eliminates this constraint by inserting a level converter (LC) with each V_{DDL} cell feeding into a V_{DDH} cell. ECVS gives better power saving than CVS but LC adds to power and delay overheads.

Without a level converter the low to high output transition delay of the second stage inverter in Figure 3.3 is not affected by the input voltage swing V_{DDL} from the previous stage, because the delay of the pull-up PMOS is only dependent on its own power supply V_{DDH} [59]. During the high to low output transition of the second inverter, the pull-down NMOS delay is affected by both the input swing V_{DDL} and the power supply V_{DDH}. Therefore, lower input swing reduces discharge current through the NMOS, which increases the pull-down delay. Because the pull-up PMOS in the inverter could not be shut off completely by the lower input swing level, severe DC current from the power supply V_{DDH} induces higher static leakage power consumption.

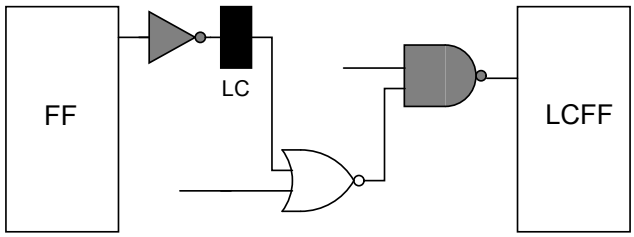
In subthreshold operation, the lower input swing exponentially increases the delay (3.5) of the driven gate. We investigate the delay and leakage power penalty from lower input swing voltage. For simplicity, we use only four types of cells, namely, INV, NAND2, NAND3 and NOR2, to synthesize example circuits. For cell characterization, all simulation results are from HSPICE using the Predictive Technology Model (PTM) for 90 nm CMOS [85]. CMOS device threshold voltages are $V_{th,PMOS} = 0.21V$ and $V_{th,NMOS} = 0.29V$ at nominal $V_{dd} = 1.2V$ and room temperature ($300K$).

Various input and output configurations interfacing gates in dual V_{dd} assignments are shown in Figure 3.4. Table 3.1 summarizes the delay and static leakage power for each case where $V_{DDH} = 250mV$ and $V_{DDL} = 200mV$ such that the entire operation is in subthreshold region. The difference between LL and HH delays shows that gate delay (3.5) is exponentially sensitive to the power supply voltage, while P_{leak} has a smaller change.

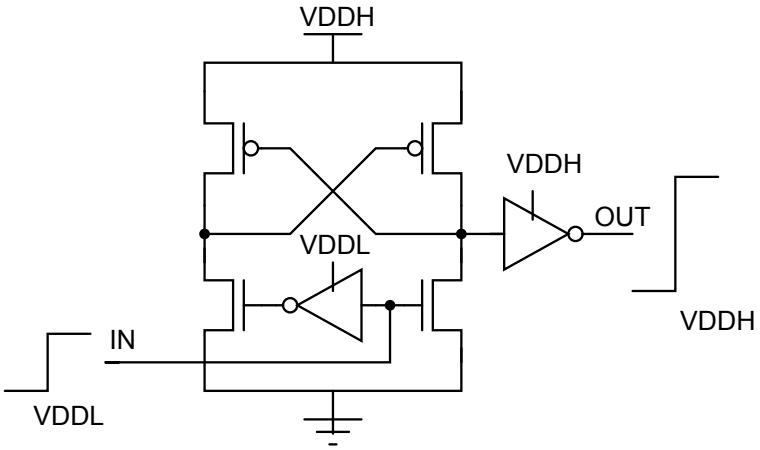
In Table 3.1, as expected, due to smaller discharging time constants, HL delays for NAND2 and NAND3 gates are lower than those for the LL configuration. However, that is not the case for INV and NOR2 gates, which are faster in the LL configuration. This



(a) Clustered voltage scaling (CVS).



(b) Extended clustered voltage scaling (ECVS).



(c) Level converter (LC).

Figure 3.2: Dual-V_{dd} schemes and level converter schematic [67, 68].

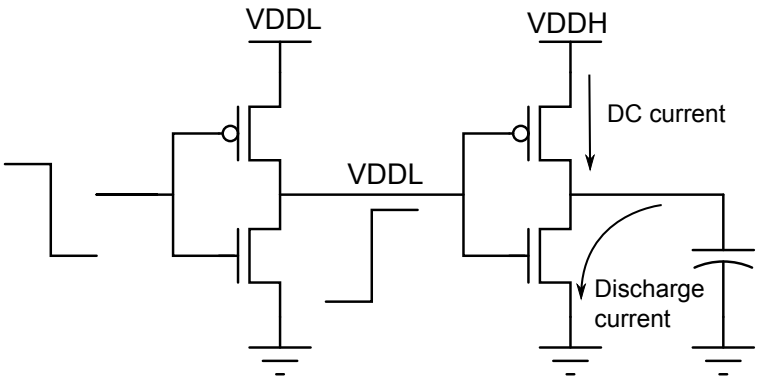


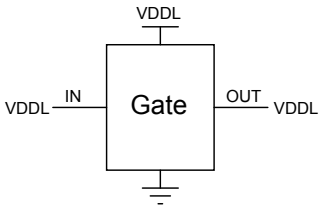
Figure 3.3: A two-inverter chain without level converter.

Table 3.1: Measurement of a gate delay with a single INV load and static leakage power in Figure 3.4 configurations at $V_{DDH} = 250mV$ and $V_{DDL} = 200mV$ through HSPICE simulation for PTM 90 nm CMOS.

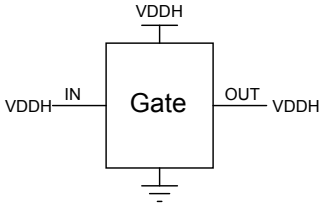
	Gate delay, t_d (ns)					Leakage power, P_{leak} (pW)				
Gate	(a) LL	(b) HH	(c) HL	(d) LH	(e) L-LC-H	(a) LL	(b) HH	(c) HL	(d) LH	(e) L-LC-H
INV	2.81	0.83	2.98	2.70	255.04	30.9	46.2	22.8	126.2	260.8
NAND2	6.82	2.10	5.31	7.92	260.32	31.1	45.3	26.2	101.5	259.9
NAND3	9.72	3.04	7.31	11.17	264.16	53.1	75.6	49.0	135.5	290.2
NOR2	8.33	2.54	8.91	5.73	262.27	32.6	48.4	20.8	156.6	263.0

speed increase is due to a higher logic 0 level for the LL configuration in charging time. In the case of leakage power for HL, all gates suppress the leakage current through the pull-up PMOS ($V_{gs} > 0$) from the power supply. Severe increases of the delay and power in dual-V_{dd} schemes are from LH, which is prohibited in CVS methodology and is allowed in ECVS with LC. But, a common LC used for above-threshold in Figure 3.2(c) cannot be used due to its unacceptable delay overhead, besides the power overhead.

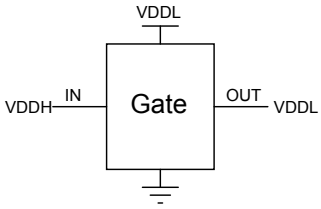
From Table 3.2, the LC delay penalty in subthreshold operation is around 80 fanout-of-four (FO4) inverter delays, which exceeds a clock cycle time of a pipelined microprocessor (13-15 FO4 delays) or an ASIC processor (44 FO4 delays) [14]. A new LC design suitable for subthreshold circuits may be needed but is out of the scope of the present work. In the next section, we include additional constraints in the MILP that will not allow the LH configuration (similar to CVS) for energy optimization.



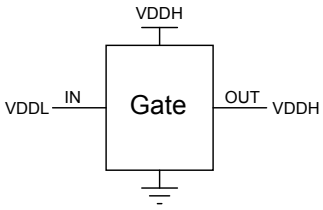
(a) LL: Low input swing driving a low V_{dd} gate.



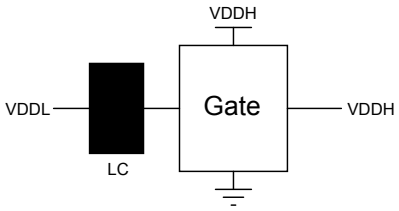
(b) HH: High input swing driving a high V_{dd} gate.



(c) HL: High input swing driving a low V_{dd} gate.



(d) LH: Low input swing driving a high V_{dd} gate.



(e) L-LC-H: Low input swing driving a high V_{dd} gate through a level converter.

Figure 3.4: Driven gates and input swing levels.

Table 3.2: Comparison of conventional LC (Figure 3.2(c)) delays normalized to INV(FO=4) delay ($V_{DD} = V_{DDH}$) for normal and subthreshold operations through HSPICE simulation in PTM 90 nm CMOS.

Gate delay	Normal $V_{DDH} = 1.2V$ $V_{DDL} = 0.8V$	Subthreshold $V_{DDH} = 300mV$ $V_{DDL} = 250mV$
INV(FO=4)	23.64 ps	1.52 ns
LC	112.33 ps	121.86 ns
LC norm. to INV(FO4)	4.8	80.2

3.3 MILP for V_{DDL} Assignment

In this section, we design minimum energy circuits with dual-V_{dd} assignments using mixed integer linear programming (MILP) [19]. First, the optimal (i.e., minimum energy per cycle) supply voltage (V_{opt}) for a single V_{dd} operation is determined. The critical path delay (or clock cycle time) of this design is used as the timing requirement for the dual voltage design. Thus, the MILP automatically applies higher supply voltage $V_{DDH} = V_{opt}$ to gates on critical paths to maintain the performance and finds an optimal lower supply voltage V_{DDL} assigned to gates on non-critical paths to reduce the total energy consumption by a global optimization considering all possible V_{DDL}. This differs from the backward traversal CVS heuristic algorithms that tend to be non-optimal. Note that more paths now may have delays that are either equal or close to the critical path delay.

Let X_i be an integer variable that is 0 for V_{DDH} or 1 for V_{DDL} for the power supply assignment of gate i. Let T_c be a predetermined critical path delay for the circuit. The optimal minimum energy voltage assignment problem is formulated as an MILP model:

$$\text{Minimize} \sum_{i \in \text{ all gates}} \Big[E_{tot,V_{DDL},i} \cdot X_i + E_{tot,V_{DDH},i} \cdot (1 - X_i) \Big] \tag{3.8}$$

$E_{tot,i}$ for V_{DDL} and V_{DDH} are given by (3.6)

$$E_{tot,i} = \alpha_i \cdot C_{L,i} \cdot V_{dd,i}^2 + P_{leak,V_{dd},i} \cdot T_c \tag{3.9}$$

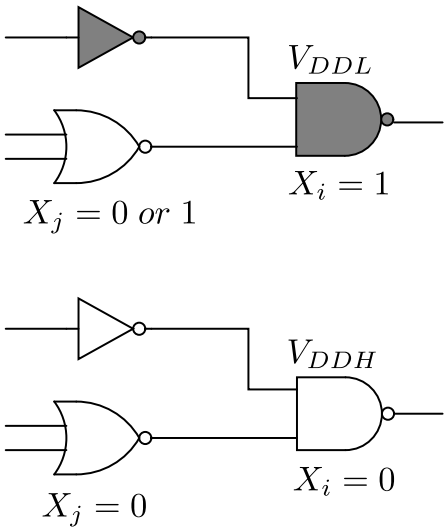


Figure 3.5: Topological constraints.

Subject to timing constraints:

$$t_{d,i} = t_{d,V_{DDL},i} \cdot X_i + t_{d,V_{DDH},i} \cdot (1 - X_i) \quad \forall i \in \text{all gates} \tag{3.10}$$

$$T_i \geq T_j + t_{d,i} \qquad \forall j \in \text{all fanin gates of gate } i \tag{3.11}$$

$$T_i \leq T_c \qquad \forall i \in \text{all primary output gates} \tag{3.12}$$

Subject to topological constraints:

$$X_i - X_j \geq 0 \qquad \forall j \in \text{all fanin gates of gate } i \tag{3.13}$$

In above constraints, T_i is the latest arrival time at the output of gate i corresponding to a primary input event [55, 56]. As mentioned in Section 3.2, the unacceptable delay penalty of asynchronous LC prohibits its use in a dual-V_{dd} scheme in the subthreshold region. The MILP model does not allow a V_{DDL} cell to drive a V_{DDH} cell as its fanout gate on account of topological constraint (3.13) as shown in Figure 3.5. Thus, the LH configuration of

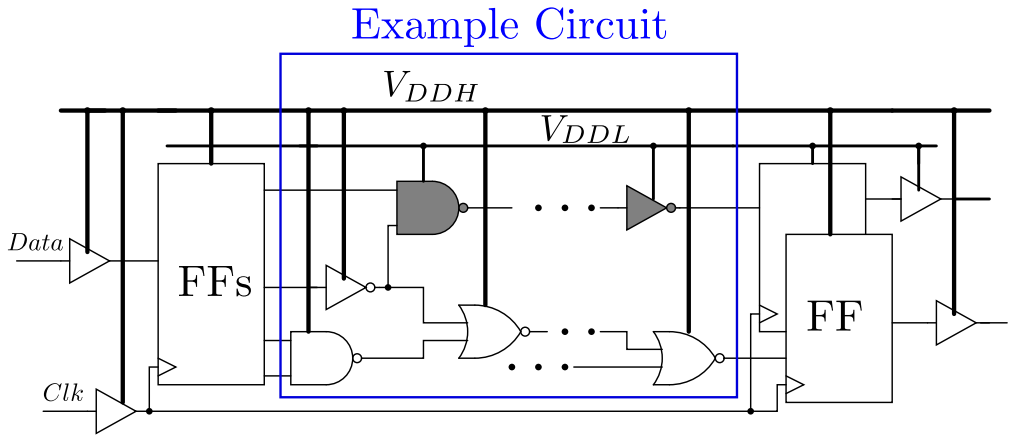


Figure 3.6: Simulation setup.

Figure 3.4(d) never occurs in the optimized circuit. Within the given timing constraint T_c, originally obtained for the best energy per cycle for single subthreshold V_{DDH} operation, the MILP searches for the best V_{DDL} such that the energy per cycle is further reduced to a true minimum.

3.4 Simulation Results

As mentioned before, we use only four basic cells (INV, NAND2, NAND3 and NOR2) for synthesizing two example circuits, a 16-bit ripple carry adder and a 4×4 multiplier, and ISCAS'85 benchmark circuits in PTM $90nm$ CMOS technology. The delay, capacitance and average leakage power of these four basic cells are characterized for the MILP model by scaling V_{dd} with a 10mV resolution in HSPICE simulations.

Switching activity α is the average number of low to high transitions at circuit nodes, which is calculated using a logic simulator with randomly generated input vectors. These randomly generated input vectors are the same as input signal vectors to the circuit for HSPICE simulation to measure energy consumption.

As shown in Figure 3.6, our example circuit, embedded in a test bench, is driven by randomly generated high input swing flip-flops. Two subthreshold voltages may be provided

by a DC to DC voltage converter [57, 77, 41]. The energy per cycle measurement is for the combinational circuit, excluding flip-flops.

From Figure 3.7(a), the minimum energy point for a 16-bit ripple carry adder with an activity factor $\alpha = 0.21$ is $9.65fJ$ at $V_{dd} = 0.21V$. The clock frequency was found to be 2.15MHz. With dual V_{dd} assignments the optimized circuit with $V_{DDH} = 0.21V$ and $V_{DDL} = 0.14V$ reduces the energy per cycle by up to 23.6% retaining the same performance. This energy reduction is shown by the downward arrow in Figure 3.7(b).

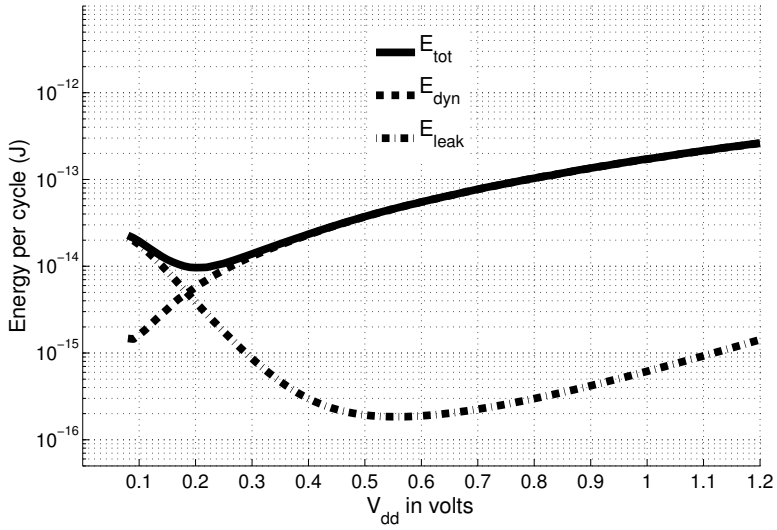
Consider again the minimum energy per cycle ($9.65fJ$) operation of the 16-bit ripple-carry adder circuit with a single subthreshold voltage 0.21V and a clock frequency of 2.15MHz. In an alternative design, we may hold the minimum energy constant and improve the performance.

From the MILP results in Table 3.3, we find that operation with two supply voltages 0.27V (V_{DDH}) and 0.19V (V_{DDL}) consumes $9.42fJ$, which is just under the minimum energy but has a clock frequency 8.41MHz. This, as shown by the right arrow in Figure 3.7(b), has about 4X speed improvement.

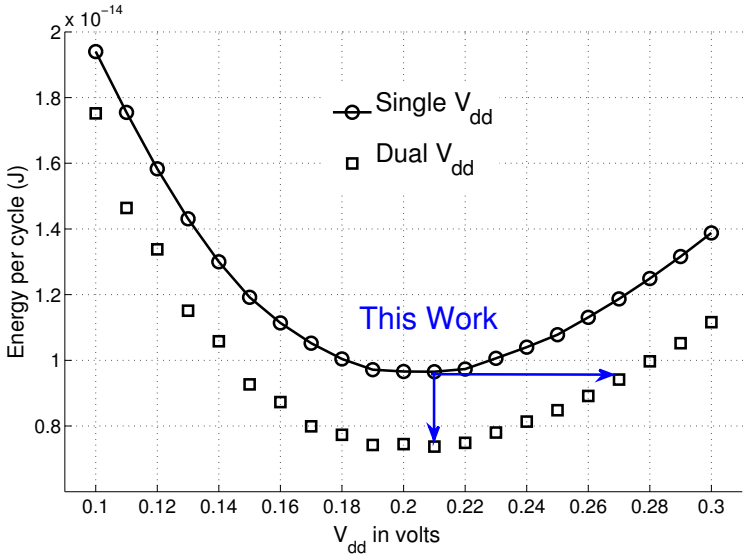
As a worst case example, a path balanced 4×4 multiplier reduces the energy per cycle to 5% below the minimum energy point with V_{DDH}=0.17V and V_{DDL}=0.12V, where the performance is not degraded. For better performance, the 4×4 multiplier can operate at 1.67MHz from a clock frequency 1MHz on minimum energy with single-V_{dd}, where two supply voltages 0.19V (V_{DDH}) and 0.13V (V_{VDDL}) are provided and minimum energy increases slightly.

Two example circuits using dual-V_{dd} show that performance improves largely for a circuit with large positive slack. Figure 3.8 (a) and (b) illustrate gate slack distribution of a 16-bit ripple carry adder and a 4×4 multiplier, respectively, for single and dual V_{dd} (Optimized) design at the minimum energy point.

Table 3.3 summarizes HSPICE simulations giving the total energy per cycle for the single voltage $V_{dd} = V_{DDH}$ reference and the optimized dual voltage $V_{dd} = \{V_{DDH},\ V_{DDL}\}$

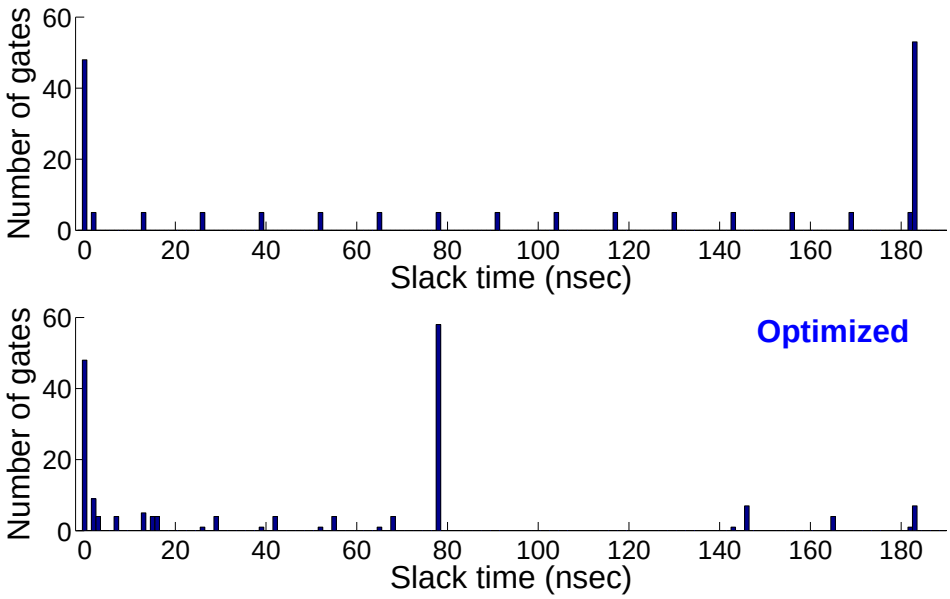


(a) Energy per cycle for single V_{dd}.

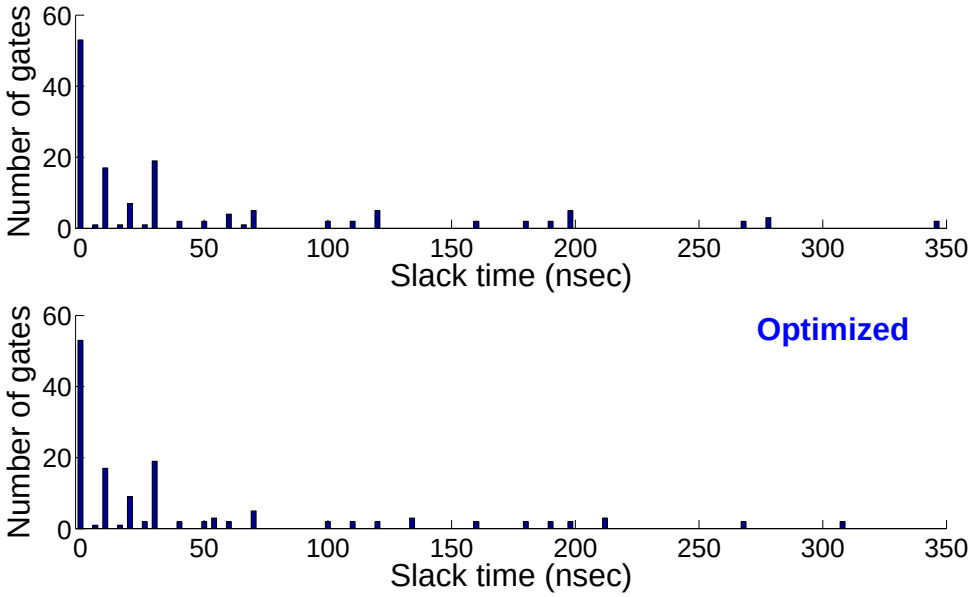


(b) Energy per cycle for single and dual subthreshold supply voltages.

Figure 3.7: Energy per cycle for a 16-bit ripple carry adder for single-V_{dd} and dual-V_{dd} in subthreshold region, activity factor $\alpha = 0.21$, PTM 90nm CMOS.



(a) 16-bit ripple carry adder at $V_{DDH} = 0.21$V and $V_{DDL} = 0.14$V.



(b) 4×4 multiplier at $V_{DDH} = 0.17$V and $V_{DDL} = 0.12$V.

Figure 3.8: Gate slack distribution (number of gates vs. slack) of a 16-bit ripple carry adder and a 4×4 multiplier for single-V_{dd} ($= V_{DDH}$) and dual-V_{dd} ($= V_{DDH}$, V_{DDL}) at the minimum energy point; slacks obtained by static timing analysis using gate delays for PTM 90nm CMOS.

Table 3.3: Total energy per cycle with optimal V_{DDL} for given V_{DDH} and maximum corresponding speed.

	16-bit ripple carry adder ($\alpha = 0.21$, total gates = 176)						4×4 multiplier ($\alpha = 0.32$, total gates = 140)					
V_{DDH} (V)	V_{DDL} (V)	V_{DDL} gate #	$E_{tot,single}$ (fJ)	$E_{tot,dual}$ (fJ)	reduction (%)	Freq. (MHz)	V_{DDL} (V)	V_{DDL} gate #	$E_{tot,single}$ (fJ)	$E_{tot,dual}$ (fJ)	reduction (%)	Freq. (MHz)
0.10	0.09	108	19.40	17.52	9.7	0.13	0.09	18	13.78	13.35	3.1	0.16
0.11	0.09	106	17.55	14.64	16.6	0.17	0.09	18	12.44	11.80	5.1	0.21
0.12	0.10	106	15.83	13.38	15.5	0.22	0.10	18	11.41	10.85	4.9	0.27
0.13	0.10	101	14.31	11.51	19.6	0.28	0.10	15	10.61	10.08	5.0	0.35
0.14	0.11	101	13.00	10.58	18.6	0.37	0.11	15	10.04	9.56	4.8	0.46
0.15	0.11	99	11.92	9.27	22.3	0.48	0.11	15	9.69	9.13	5.8	0.60
0.16	0.12	99	11.14	8.73	21.6	0.62	0.12	15	9.51	8.98	5.6	0.78
0.17	0.12	95	10.52	7.99	24.0	0.80	0.12	13	9.48	8.99	5.2	1.00
0.18	0.13	95	10.04	7.73	23.0	1.02	0.13	13	9.59	9.11	5.0	1.30
0.19	0.13	88	9.72	7.42	23.6	1.32	0.13	13	9.74	9.19	5.6	1.67
0.20	0.14	88	9.66	7.45	22.9	1.68	0.14	13	10.21	9.65	5.5	2.14
0.21	0.14	84	9.65	7.37	23.6	2.15	0.15	13	10.66	10.08	5.4	2.73
0.22	0.15	84	9.73	7.49	23.1	2.72	0.15	12	11.06	10.60	4.2	3.46
0.23	0.16	84	10.06	7.80	22.5	3.44	0.16	12	11.83	11.24	5.0	4.37
0.24	0.17	84	10.40	8.14	21.8	4.33	0.17	12	12.53	11.93	4.8	5.50
0.25	0.18	84	10.78	8.48	21.3	5.43	0.18	13	13.28	12.61	5.0	6.87
0.26	0.18	78	11.31	8.91	21.2	6.77	0.19	13	14.14	13.43	5.0	8.55
0.27	0.19	78	11.87	9.42	20.7	8.41	0.19	12	15.03	14.30	4.9	10.60
0.28	0.20	78	12.49	9.97	20.2	10.39	0.20	12	15.98	15.22	4.8	13.06
0.29	0.22	88	13.16	10.52	20.1	12.79	0.21	12	16.98	16.19	4.7	16.02
0.30	0.23	88	13.88	11.16	19.6	15.65	0.22	12	18.03	17.21	4.5	19.54
Average					20.5						4.9	

circuits. Voltages vary from $0.1V$ to $0.3V$. Both single and dual V_{dd} circuits have the same speed because all gates on critical paths have the same V_{DDH} for either circuit.

The energy savings at minimum energy operating points using dual-V_{dd} are obtained from HSPICE simulations for ISCAS'85 benchmark circuits, as shown in Table 3.4. The optimized c880 (an 8-bit ALU) shows 22.2% energy saving as the best case. The energy saving for c6288 (a 16×16 multiplier) is only about 2.1%. Gate slack distribution is shown for c880 and c6288, respectively, in Figure 3.9.

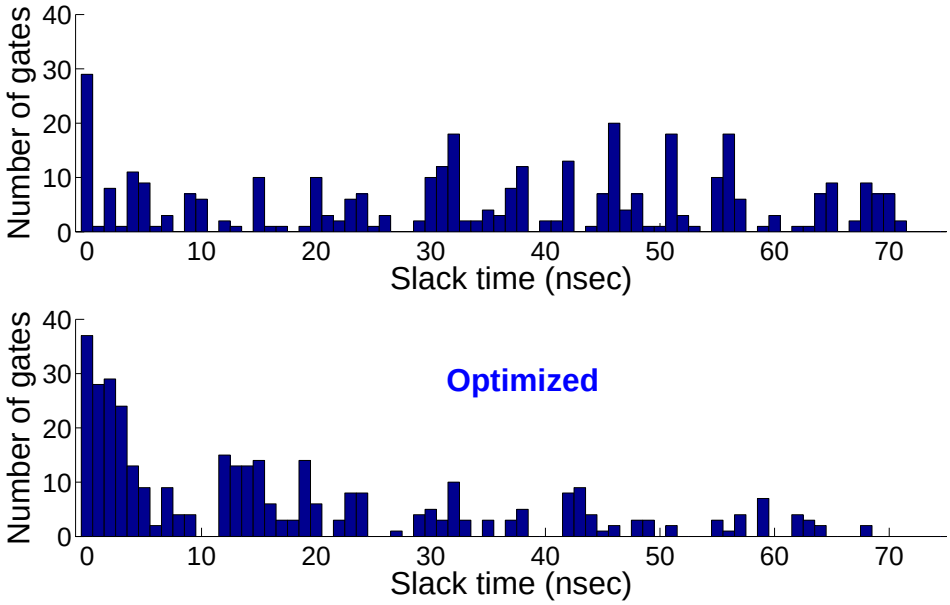
Logic function failure occurs at $0.08V$ in NAND3, so the possible lowest V_{DDL} assignment in MILP optimization is $0.09V$. This minimum operating voltage guarantees 10% to 90% output voltage swing for all four cells in the full range of operational voltages used. Figure 3.10 shows sample signal waveforms from an optimized 16-bit ripple carry adder circuit for $V_{DDH} = 0.11V$ and $V_{DDL} = 0.09V$. This has V_{DDL} assigned to cells on a non-critical path that leads to the least significant sum bit (s1). The output flip-flop (s1q) holds correct signal values at the minimum operating voltage on positive clock edges.

Table 3.4: Energy saving with optimal V_{DDL} for given V_{DDH} (minimum energy operating point) in ISCAS'85 benchmark circuits for PTM 90nm CMOS.

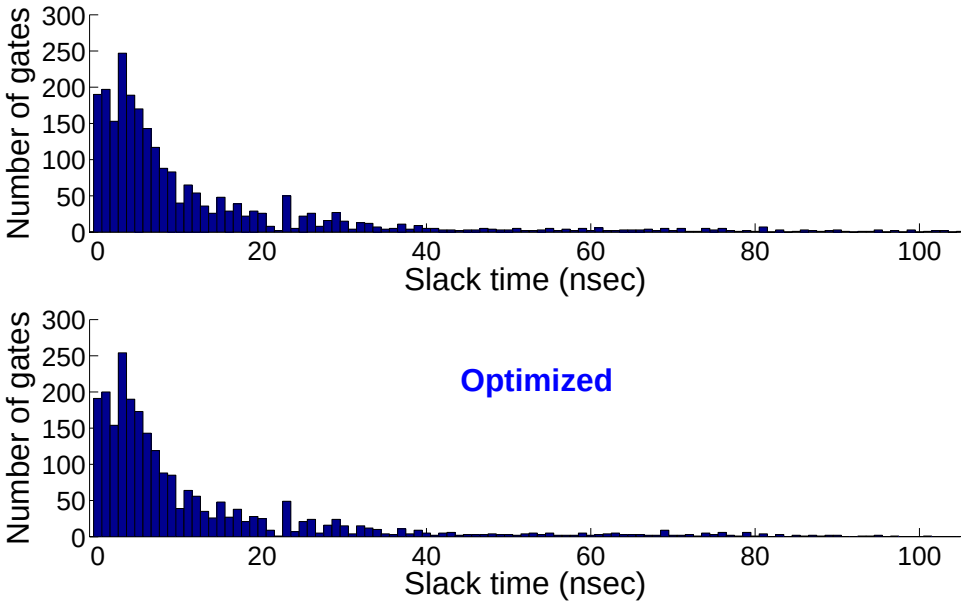
Benchmark circuit	Total gates	Activity α	V_{DDH} (V)	V_{DDL} (V)	V_{DDL} gates (%)	E_{single} (fJ)	E_{dual} (fJ)	$E_{reduc.}$ (%)	Freq. (MHz)
c432	154	0.19	0.25	0.23	5.2	7.9	7.8	1.1	14.4
c499	493	0.21	0.22	0.18	9.7	20.2	19.8	2.0	11.9
c880	360	0.18	0.24	0.18	46.4	14.4	11.2	22.2	13.6
c1355	469	0.21	0.21	0.18	10.2	19.5	19.0	2.5	9.8
c1908	584	0.20	0.24	0.21	24.3	26.5	25.0	5.8	11.8
c2670	901	0.16	0.25	0.21	46.4	32.8	28.0	14.8	17.4
c3540	1270	0.33	0.23	0.14	7.0	88.0	84.6	3.8	7.2
c5315	2077	0.26	0.24	0.19	47.1	116.8	98.0	16.1	9.8
c6288	2407	0.28	0.29	0.18	2.7	165.4	162.0	2.1	9.4
c7552	2823	0.20	0.25	0.21	42.3	131.7	117.1	11.1	13.6
Average					24.1			8.2	

When V_{DDH} is 100mV, it is approaching the lower end of its range beyond which the circuit would fail to operate. The MILP now has limited choices for a solution and gives a V_{DDL} that provides smaller energy saving. The 16-bit ripple carry adder has better energy reduction because it can utilize more time slack from non-critical paths compared to the 4×4 multiplier with more balanced paths. The gate delay in subthreshold operation increases exponentially with reducing supply voltage, which forces the optimal V_{DDL} close to V_{DDH}.

Even though the MILP model only allows HL configuration and eliminates the use of LC for a dual V_{dd} circuit block, level conversion may be needed at outputs to match signal levels across block to block connections of a system. The differential cascode voltage switch (DCVS) based level converter of a normal standard cell library in Figure 3.2(c) is not suitable for dual subthreshold design due to its huge delay penalty. Realizing that the design of LC for ultra low voltage is an open problem, our design refrains from using level converters while taking the penalty of energy saving into account. For level converting, we always assign V_{DDH} to primary output (PO) gates before the output flip-flops at multiple voltage boundaries between circuit blocks. The PO gates driven by V_{DDL} cells are found to correctly execute their logic functions if, for a given V_{DDH}, V_{DDL} is bounded as shown in Figure 3.11.



(a) c880 at $V_{DDH} = 0.24$V and $V_{DDL} = 0.18$V.



(b) c6288 at $V_{DDH} = 0.29$V and $V_{DDL} = 0.18$V.

Figure 3.9: Gate slack distribution of c880 and c6288 for single-V_{dd} and dual-V_{dd} at the minimum energy point in PTM 90nm CMOS.

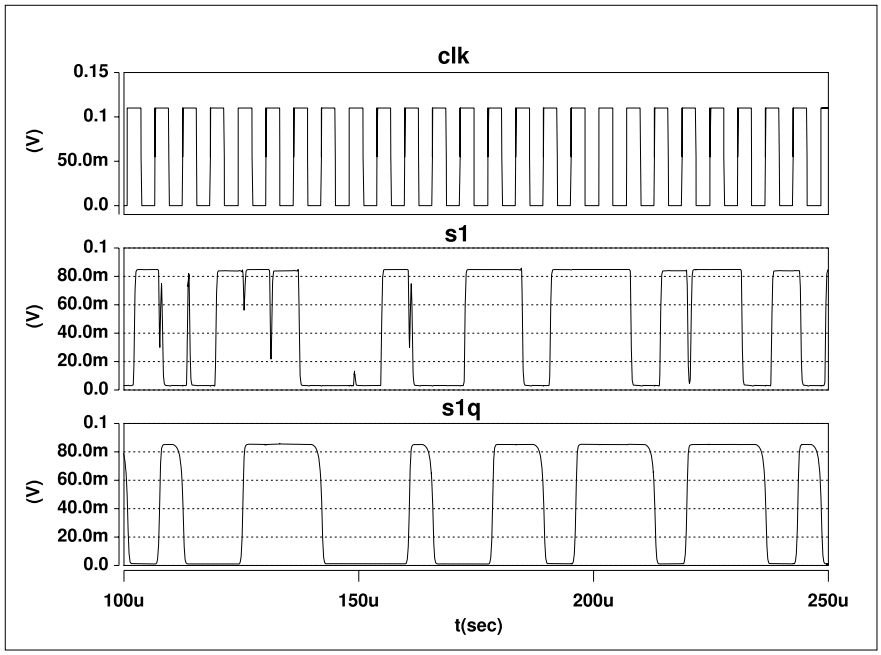


Figure 3.10: Output signal waveforms of s1 and s1q in a 16-bit ripple carry adder at minimum operating voltage, $V_{DDL} = 0.09V$, in HSPICE simulation, PTM 90nm CMOS.

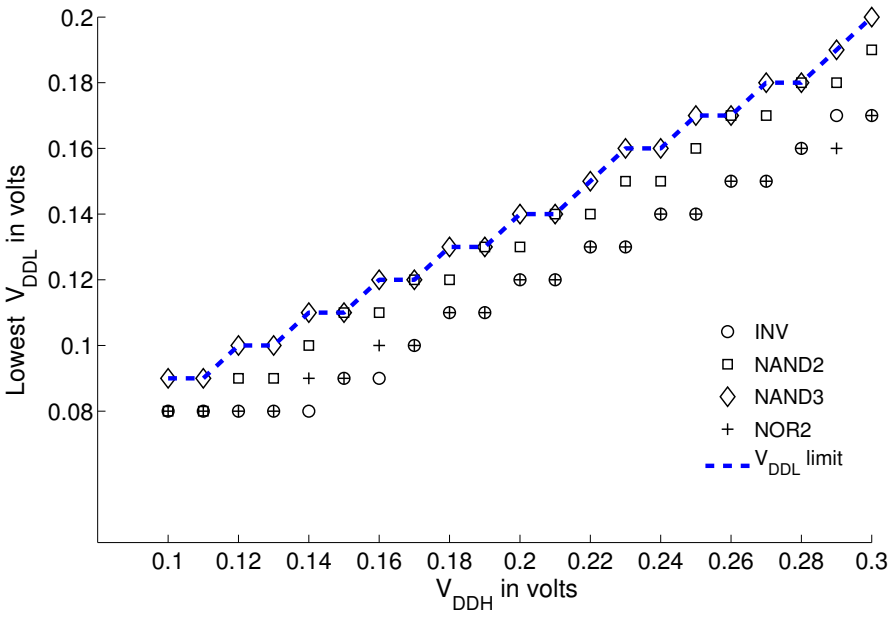


Figure 3.11: V_{DDL} bound for given V_{DDH} with LH configured cells.

This lowest possible V_{DDL} raises the minimum operating voltage for the dual voltage optimized circuit block. The optimal V_{DDL} in the MILP model can be higher than its true optimal value to suppress DC leakage power of the LH configured PO gates. Using two small example circuits, a 16-bit ripple-carry adder and a 4×4 multiplier show average reduced energy savings of 11.9% and 2.6%, respectively. The penalty of energy saving from level converting may be negligible for a large system in which most blocks would operate at V_{DDL} and only a few need V_{DDH}.

3.5 Summary

In this chapter, we first introduced dual-V_{dd} design for a bulk CMOS subthreshold circuit [35]. Some applications in the market may need minimum energy consumption without a performance concern. This work could solve those design problems. For a wide range of speed requirements, the MILP determines globally the energy optimized circuit by assigning the optimal V_{DDL} to gates on non-critical paths. A 16-bit ripple carry adder shows on average 20.5% reduced energy consumption, while maintaining same performance as the original single V_{dd} circuit. The worst case example of a 4×4 multiplier still gives on average 4.9% reduction. Further, allowing a small amount of increase in the energy consumption can significantly speed-up the subthreshold operation of a logic circuit. The methodology of dual V_{dd} assignment is valid for substantial speed-up without energy increase, as well as for energy reduction below the minimum achievable in a single voltage circuit.

The proposed MILP algorithm is not restricted to subthreshold operation alone. When a higher performance, impossible to achieve in the subthreshold region, is required we would then obtain two above-threshold voltages that will satisfy the performance criteria and minimize the energy per cycle. There may be potential for greater energy saving as circuit size increases due to larger critical path delay leading to greater slack for many gates. The process variation of the device threshold voltage (V_{th}) can seriously affect a subthreshold voltage design and this will be studied for nanometer technologies later. Higher leakage technologies

display higher speed in the subthreshold region because the logic operation relies on leakage currents. These aspects of dual-V_{dd} design in subthreshold region are worth exploring.

Chapter 4

Minimum Energy CMOS Design with Dual Subthreshold Supply and Multiple Logic-Level Gates

Some energy constrained applications that require moderate speed may not aggressively scale the supply voltage down to the minimum energy point to maintain the performance. Small energy increase from the absolute minimum energy point of a subthreshold circuit can notably improve performance. Near-threshold operating circuit design is another choice to cover a wider range of system performances for applications with tolerable energy increase ($\sim$2X) from E_{min} by scaling V_{dd} to near V_{th} [18, 47, 30]. Technology down-scaling improves the speed of a subthreshold circuit, but greater variability may adversely affect E_{min} for extremely small feature size [5].

In Chapter 3, the presented MILP limits full use of the time slack by topological constraints considering multiple voltage boundaries without level converters. Thus, the energy saving of dual V_{dd} design is not as much as expected. We are motivated to exploit full time slack on non-critical paths in a subthreshold circuit using multiple logic-level gates to further reduce E_{min} at its original speed or alternatively have the circuit operate at a higher speed holding the energy consumption close to E_{min}.

Figure 4.1 shows the benefit of dual voltage design for a 32-bit ripple carry adder in 90nm CMOS technology operating in the subthreshold regime. Energy per cycle for the optimized dual voltage design (E_{dual}) is reduced $\sim$0.67X from E_{min} that is obtained by scaling down a single supply voltage to its minimum energy operating point at V_{dd}=0.31V. This 32-bit ripple carry adder can also operate $\sim$7X faster with same energy as E_{min} in another dual voltage design using V_{dd}=0.45V. Finding an optimal lower supply voltage (V_{DDL}) for a given higher supply voltage (V_{DDH}) and its assignments is the main problem in dual voltage design.

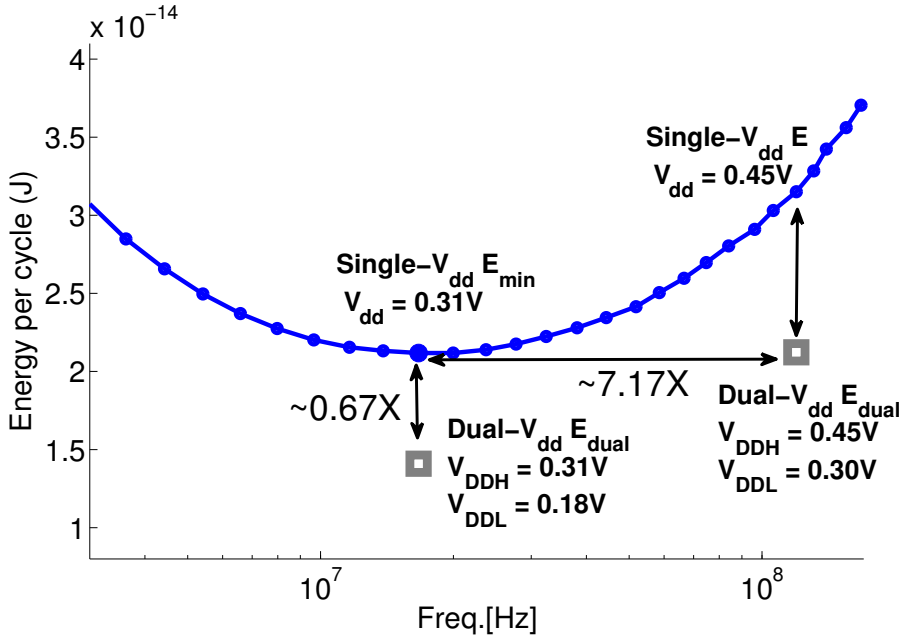


Figure 4.1: Energy and speed benefits of dual V_{dd} design in subthreshold voltage operation for a 32-bit ripple carry adder through HSPICE simulation in PTM 90nm CMOS (activity factor $\alpha = 0.17$, number of gates = 352).

We formulate a mixed integer linear program (MILP) to solve this problem with multiple logic-level gates considering multiple voltage boundaries.

4.1 Operation of Conventional Level Converters in Subthreshold Regime

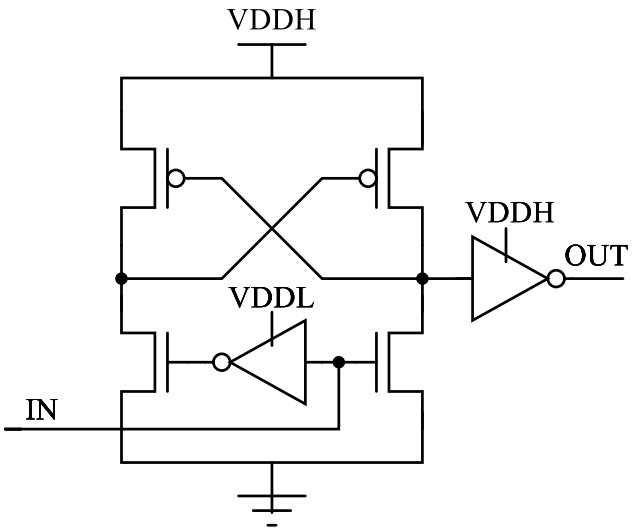
In a dual-V_{dd} design, assigning lower supply voltage (V_{DDL}) only to gates on non-critical paths reduces both dynamic and static leakage power of the circuit. Higher supply voltage ($V_{DDH} = V_{dd}$) is assigned to gates on critical paths to maintain the overall circuit performance. By utilizing the time slack, we ensure that there is no performance loss. But, an asynchronous level converter (ALC) is considered essential to suppress DC leakage current and guarantee the correct switching of a V_{DDH} gate driven by a low voltage input signal. Level converting cost, however, reduces the power saving of the dual-V_{dd} scheme.

Clustered voltage scaling (CVS) [67] assigns V_{DDL} to gates with positive time slack starting from primary outputs to primary inputs and so does not allow the V_{DDL} gates to

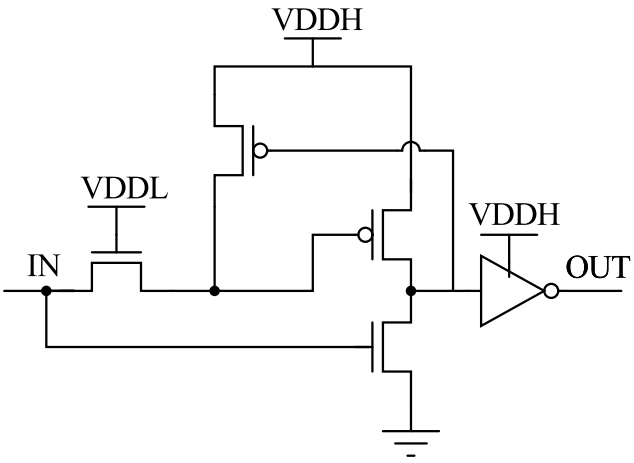
feed directly into V_{DDH} gates by grouping gates into V_{DDH} and V_{DDL} clusters. V_{DDH} cluster is always located upstream as signals flow. This topological constraint reduces the potential power saving from full use of the time slack that exists inside a circuit. Asynchronous level converters are not needed inside a combinational circuit block, but the level converting flip-flops (LCFF) are needed in sequential elements [28]. No overheads of power and delay from ALCs exist in CVS. For removing the topological constraint in CVS, *extended clustered voltage scaling* (ECVS) [68] inserts an ALC at a point, where a V_{DDL} gate drives a V_{DDH} gate, to assign V_{DDL} to more gates with time slack. This gives more power saving than CVS.

We apply the dual voltage technique to subthreshold supply combinational circuits. To maximize energy saving from the time slack, a level converter is still considered essential. In Figure 4.2, two traditional ALCs, a differential cascode voltage switched (DCVS) level converter and a pass gate (PG) level converter, are shown. The PG level converter consumes less energy than the DCVS level converter due to fewer devices in it and reduced contention [40]. Compared to the delay of a circuit operating with nominal V_{dd}, the delay of a subthreshold circuit increases exponentially as supply voltage V_{dd} reduces [76]. This means that the time slack is consumed quickly by assigning V_{DDL}, quite close to V_{DDH}, to gates on non-critical paths. With such delay characteristic, the delay overhead of the ALC is more critical for implementing a dual-V_{dd} design in the subthreshold regime.

We use the HSPICE simulator [27] to size properly for reducing the delay of two ALCs in subthreshold region. Predictive Technology Model (PTM) for 90 nm CMOS [85] was used in the simulations. Table 4.1 shows the delay penalty of the two optimized ALCs in a range of 28$\sim$ 60$\times$ INV(FO4) delays, where INV(FO4) is the delay of a standard inverter with fanout of four. The normal ALC delay is considered as 2$\times$ INV(FO4) delays [17] for a nominal supply voltage. A low voltage microprocessor has $\sim$ 400$\times$ INV(FO4) delays for a single pipeline stage. A microprocessor operating in subthreshold region would prefer a shallow pipeline to mitigate variability and a 40$\times$ INV(FO4) delay is considered as a typical design



(a) Differential cascode voltage switched (DCVS) level converter.



(b) Pass gate (PG) level converter.

Figure 4.2: Two traditional level converter schematics [40].

Table 4.1: Delays of two optimal sized ALCs with a single INV load at $V_{DDL} = 230mV$ and $V_{DDH} = 300mV$ in PTM $90nm$ CMOS.

ALCs	Delay	Norm. to INV(FO4)
DCVS	79.1 ns	60.4
PG	37.6 ns	28.7

Table 4.2: Multiple logic-level gate delays with a single INV load at $V_{DDL} = 230mV$ and $V_{DDH} = 300mV$ in PTM $90nm$ CMOS (High PMOS $V_{th} = 0.29V$).

Multiple logic-level gates	Delay Norm. to INV(FO4)
INV	1.3
NAND2	2.3
NAND3	3.1
NOR2	3.9

case [63]. To reduce the delay penalty of level converting, we need to investigate alternative approaches to remove ALCs without topological constraints in the dual-V_{dd} design.

As discussed in the literature, two types of logic gate designs have the capability to handle multiple logic levels. Among these the embedded logic level converting circuit [40] may not be a good choice because the previous ALC structures, when integrated with logic gates, will not reduce the overall delay penalty. A level-shifter free design using dual V_{th} [17] places high V_{th} devices in the pull-up PMOS network of a logic gate to suppress DC static leakage with low input signals, as shown in Figure 4.3. This causes the rise time of the gate to increase, thus the overall level shifting logic gate delay is larger than that of a normal gate (PMOS $V_{th} = 0.21$). As shown in Table 4.2, the delay penalty of these *multiple logic-level gates* is much less than that of standard ALCs in the subthreshold region. Within some range of low input voltages close to V_{dd}, a multiple logic-level INV consumes less leakage power than a standard INV. This leakage power increases as the low input voltage goes down in Figure 4.4. Considering the delay and power overheads, we are compelled to use the multiple logic-level gates instead of ALCs in our dual voltage design.

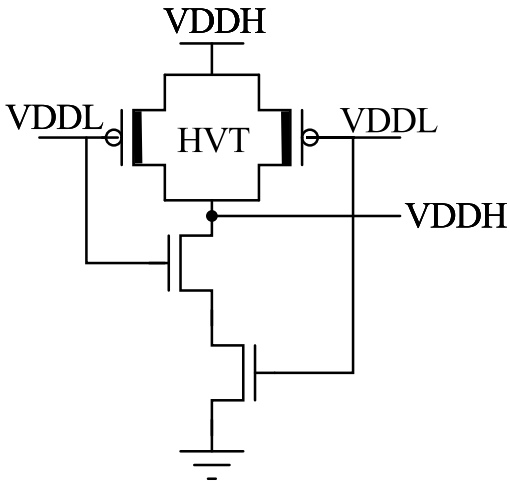


Figure 4.3: Multiple logic-level NAND2 gate [17].

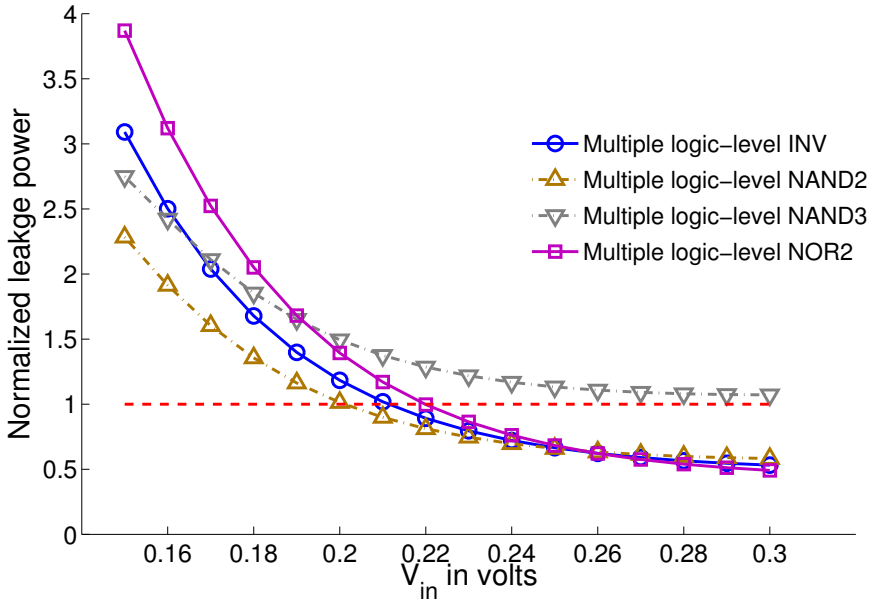


Figure 4.4: Multiple logic-level gate leakage power normalized to a standard INV (V_{dd}=V_{in} = 300mV) in PTM 90nm CMOS.

4.2 MILP for Dual Voltage Design with Multiple Logic-Level Gates

In this section, we design minimum energy circuits with dual-V_{dd} assignments without ALCs using mixed integer linear programing (MILP) [19]. Multiple logic-level logic gates eliminate the use of ALCs and allow V_{DDL} gates to drive V_{DDH} gates with affordable overheads in terms of delay and leakage power in a combinational circuit. First, the performance requirement (critical path delay T_c) of a system is given. Therefore, V_{DDH} is determined to satisfy the system speed (or clock cycle time). The MILP automatically assigns the predetermined V_{DDH} to gates on critical paths to maintain the performance and finds optimal V_{DDL} for gates on non-critical paths to reduce the total energy consumption (i.e., minimum energy per cycle) by a global optimization. Inherently, CVS and ECVS are heuristic algorithms that tend to be non-optimal, because of the backward traversal from primary outputs through gates with time slack for assigning lower supply voltage V_{DDL}.

Assuming that gates become active once per clock cycle, the total energy per cycle (E_{tot}) is given by following equations [76]:

$$\begin{aligned} E_{dyn} &= \alpha_{0\to1} \cdot C_{load} \cdot V_{dd}^2 \\ &= C_{sw} \cdot V_{dd}^2 \\ E_{leak} &= I_{off} \cdot V_{dd} \cdot T_c \\ &= P_{leak} \cdot T_c \\ E_{tot} &= E_{dyn} + E_{leak} \\ &= C_{sw} \cdot V_{dd}^2 + P_{leak} \cdot T_c \end{aligned} \tag{4.1}$$

where $\alpha_{0\to1}$ is the low to high transition activity for the gate output node and C_{load} is the load capacitance of the gate. In (4.1), dynamic energy (E_{dyn}) quadratically depends on scaling the power supply voltage V_{dd} with the total switched capacitance C_{sw} of a circuit, while the leakage energy (E_{leak}) is linearly proportional to leakage power P_{leak} during a clock cycle.

Before we formulate the MILP model of the optimal minimum energy V_{DDL} assignment, all variables and constants used in the MILP model are listed:

- V_v: supply voltage integer variable that is 1 for two selected V_{DDH} and V_{DDL} in a span of scaling supply voltage v.
- $X_{i,v}$: voltage assignment integer variable that is 1 for gate i with supply voltage v.
- $F_{i,v}$: fan-in integer variable that is 1 for gate i having at least one fan-in gate that is powered by supply voltage v.
- $P_{i,v}$: penalty integer variable that is 1 when gate i driven by low input voltage v.
- T_i: latest arrival time variable at gate i output from primary input events.
- α_i: low to high transition activity of gate i.
- $V_{dd,v}$: supply voltage value of v.
- $C_{i,v}$: load capacitance of gate i with supply voltage v.
- $P_{leak,i,v}$: leakage power of gate i with supply voltage v.
- $P_{leako,i,v}$: leakage power overhead of multiple logic-level gate i driven by low input voltage v.
- $td_{i,v}$: gate delay of gate i with supply voltage v.
- $tdo_{i,v}$: gate delay overhead of multiple logic-level gate i driven by low input voltage v.
- N_i: number of inputs for gate i.
- T_c: critical path delay of a circuit.
- G_{tot}: total number of gates in a circuit.
- V_{nom}: nominal supply voltage value (1.2V) for 90nm CMOS.

The optimal V_{DDL} assignment for the minimum energy design is modeled by MILP equations:

$$\text{Minimize}\Bigg[\sum_i \sum_{v\in V} \left(\alpha_i \cdot C_{i,v} \cdot V_{dd,v}^2 + P_{leak,i,v} \cdot T_c\right) \cdot X_{i,v}$$

$$+\sum_i \sum_{v\in V_L} P_{leako,i,v} \cdot T_c \cdot P_{i,v}\Bigg], \quad \forall i \in \text{all gates} \tag{4.2}$$

$$V_{min} \le V \le V_{DDH}, \;\; V_{low} \le V_L < V_{DDH}$$

where V_{min} is the minimum operating voltage for the correct logic function of a gate with subthreshold supply voltage and V_{low} is the lowest input voltage to keep 10% to 90% output voltage swing for a logic gate when V_{DDH} is predetermined. The timing constraints are [55, 56]:

$$T_i \ge T_j + \sum_{v\in V} td_{i,v} \cdot X_{i,v} + \sum_{v\in V_L} tdo_{i,v} \cdot P_{i,v}$$

$$\forall i \in \text{all gates}, \; \forall j \in \text{all fanin gates of gate } i \tag{4.3}$$

$$T_i \le T_c \qquad \forall i \in \text{all primary output gates} \tag{4.4}$$

Penalty condition:

$$\begin{aligned}
&\sum_j X_{j,v} \leq N_i \cdot F_{i,v} \qquad \forall j \in \text{all fanin gates of gate } i \\
&\sum_j X_{j,v} \geq N_i \cdot F_{i,v} - (N_i - 1) \;\; \forall i \in \text{all gates}, \; \forall v \in V_L
\end{aligned} \tag{4.5}$$

$$\begin{aligned}
&F_{i,v} + X_{i,V_{DDH}} \geq 2 \cdot P_{i,v} \qquad \forall i \in \text{all gates} \\
&F_{i,v} + X_{i,V_{DDH}} \leq 2 \cdot P_{i,v} + 1 \quad \forall v \in V_L
\end{aligned} \tag{4.6}$$

$$\begin{aligned}
&\sum_{v \in V} V_{dd,v} \cdot X_{i,v} \leq \sum_{v \in V} V_{dd,v} \cdot X_{j,v} + \sum_{v \in V_L} V_{nom} \cdot P_{i,v} \\
&\forall j \in \text{all fanin gates of gate } i
\end{aligned} \tag{4.7}$$

Dual supply voltages selection:

$$\sum_{v \in V} V_v = 2 \tag{4.8}$$

$$V_{V_{DDH}} = 1 \tag{4.9}$$

$$\sum_{v \in V} X_{i,v} = 1 \qquad \forall i \in \text{all gates} \tag{4.10}$$

$$\sum_i X_{i,v} \leq G_{tot} \cdot V_v \quad \forall i \in \text{all gates}, \; \forall v \in V \tag{4.11}$$

As mentioned before, T_c is given by the performance requirement. Therefore, V_{DDH} is selected from (4.9) in scaling supply voltage span. In dual power supply constraints, MILP only chooses two supply voltages, given V_{DDH} and optimal V_{DDL}, then each gate in the circuit must be assigned to one of them from (4.11); we use a bin-packing technique [1]. Penalty condition tests the existence of a V_{DDH} gate driven by at least one V_{DDL} fan-in gate from (4.5) (Boolean Or) and (4.6) (Boolean AND). The non-linear Boolean functions are expressed as linear constraints. When penalty exists, $P_{i,VDDL}$ becomes 1 and (4.7) allows low voltage inputs to drive a V_{DDH} gate by replacing it with a multiple logic-level gate. When

assigning V_{DDL} to the time slack gate, MILP checks the timing violation against clock time using (4.3) and (4.4) timing constraints. Cost function (4.2) favorably balances both delay and leakage penalties of the multiple logic-level gates.

4.3 Simulation Results

All simulation results are from HSPICE using PTM 90nm CMOS at room temperature (300K). The CMOS device threshold voltages are $V_{th,pmos} = 0.21$V and $V_{th,nmos} = 0.29$V at nominal $V_{dd} = 1.2$V. For simplicity, we use only four types of basic standard cells, namely, INV, NAND2, NAND3, and NOR2, to synthesize ISCAS'85 benchmark circuits. Therefore, only four types of multiple logic-level gates are used with high PMOS threshold voltage assigned to the pull-up PMOS network of basic cells. High PMOS threshold voltage ($V_{th,pmos}$ = 0.29) is selected.

We assume that randomly generated input signals with high input voltage V_{DDH} drive all primary inputs of the circuit. Two subthreshold supply voltages, V_{DDH} and V_{DDL}, can be provided by a voltage scalable DC to DC converter [57]. We also assume that combinational benchmark circuits have no restrictions in primary output voltage level, either of V_{DDH} or V_{DDL}. In reality, level shifting flip-flops (LCFF) [67, 28] can be placed at low voltage primary outputs as the sequential elements of the design.

The MILP algorithm of Section 4.2 is applied to find the optimal V_{DDL} for the benchmark circuits with given performance (i.e., V_{DDH}) in subthreshold region. Table 4.3 shows HSPICE simulation results for single V_{dd} total energy per cycle as a reference and dual V_{dd} optimized energy per cycle with the optimal V_{DDL} selection. Activity α is the average number of low to high transitions at circuit nodes and V_{DDL} is the optimal low voltage supply corresponding to V_{DDH}. Multiple logic-level gates were not required for c432, c499 and c1355, and therefore, there were no V_{DDH} gates driven by V_{DDL} gates in optimized circuits; they were the same as in [35]. From (4.7), the MILP algorithm automatically determines whether or not a multiple logic-level gate is to be used, based upon the benefit of energy saving. The design of c3540

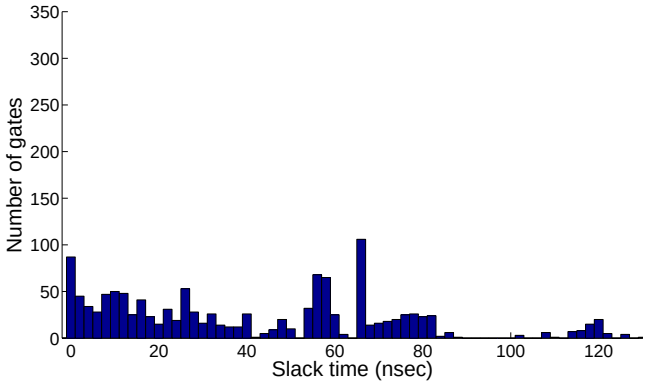
Table 4.3: Total energy per cycle with optimal V_{DDL} for given V_{DDH} and performance of ISCAS'85 benchmark circuits and 32-bit ripple carry adder.

Benchmark Circuit	Total gates	Activity α	V_{DDH} (V)	V_{DDL} (V)	V_{DDL} gates (%)	Multiple logic-level gates	E_{single} (fJ)	E_{dual} (fJ)	Reduc. (%)	Reduc.[35] (%)	Freq. (MHz)
c432	154	0.19	0.25	0.23	5.2	0	7.9	7.8	1.1	1.1	14.4
c499	493	0.21	0.22	0.18	9.7	0	20.2	19.8	2.0	2.0	11.9
c880	360	0.18	0.24	0.19	56.7	23	14.4	10.9	24.5	22.2	13.6
c1355	469	0.21	0.21	0.18	10.2	0	19.5	19.0	2.5	2.5	9.8
c1908	584	0.20	0.24	0.21	27.6	71	26.5	23.2	12.4	5.8	11.8
c2670	901	0.16	0.25	0.19	40.2	41	32.8	26.9	18.1	14.8	17.4
c3540	1270	0.33	0.23	0.16	40.8	69	88.0	70.8	19.5	3.8	7.2
c5315	2077	0.26	0.24	0.19	60.5	62	116.8	92.2	21.1	16.1	9.8
c6288	2407	0.28	0.29	0.19	4.7	20	165.4	159.1	3.8	2.1	9.4
c7552	2823	0.20	0.25	0.21	51.6	201	131.7	112.1	14.9	11.1	13.6
32-bit RCA	352	0.17	0.31	0.18	52.3	11	21.2	14.1	33.5	31.3	16.7
Average					32.7				14.0	10.2	

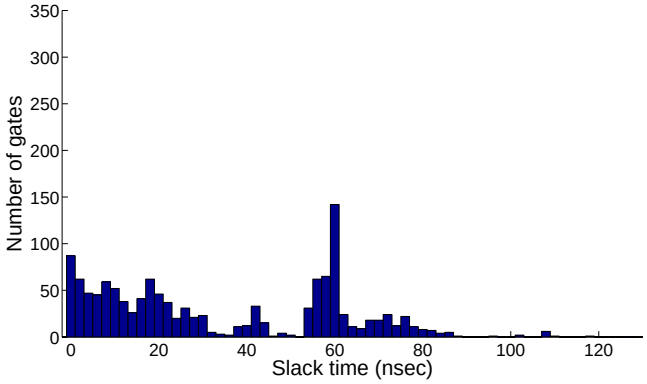
shows that energy saving of the dual-V_{dd} circuit is improved 15.7% more than [35]. It is evident that the optimized circuit with multiple logic-level gates utilizes more time slack as shown in Figure 4.5.

Multiple logic-level gates remove topological constraints and allow V_{DDL} gates to drive V_{DDH} gates. Thus, MILP can assign V_{DDL} to more gates on non-critical paths and further increase energy saving as expected. For the dual-V_{dd} design with multiple logic-level gates, the best case is about 24.5% energy reduction for c880 (an 8-bit ALU). Another circuit, c6288 (a 16×16 multiplier), has only 3.8% reduction. There is little benefit of dual-V_{dd} design for c432, c499, and c1355, where most paths are balanced. The optimized circuits show energy saving of 14.0% on an average, even it includes the energy savings of path balanced circuits. Figure 4.6 shows the gate slack distributions obtained from static timing analysis [33] of the single-V_{dd} and dual-V_{dd} designs of c880. Clearly, it is the large number of gates with large slack in the single-V_{dd} design that allows many low V_{dd} assignments.

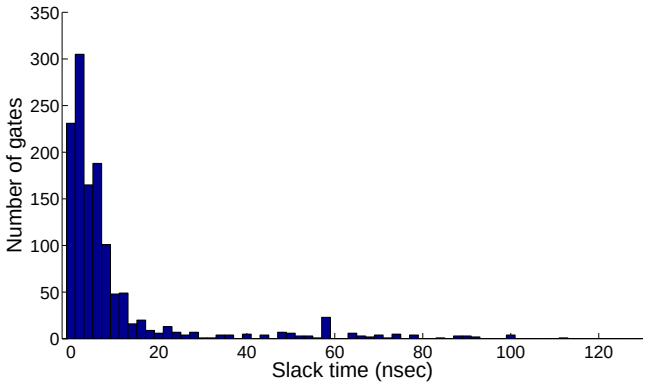
The energy saving from dual voltage design depends on the time slacks of gates. In the subthreshold region it is also affected by the number of V_{DDL} gates driven by V_{DDH} gates. Leakage current of PMOS devices in a V_{DDL} gate is suppressed by high voltage input signal from a V_{DDH} gate, because the source to gate voltage, V_{sg}, in PMOS devices is negative. The leakage energy is comparable to dynamic energy in the subthreshold region. This leakage



(a) Single-V_{dd} design at $V_{dd} = 0.23$V.

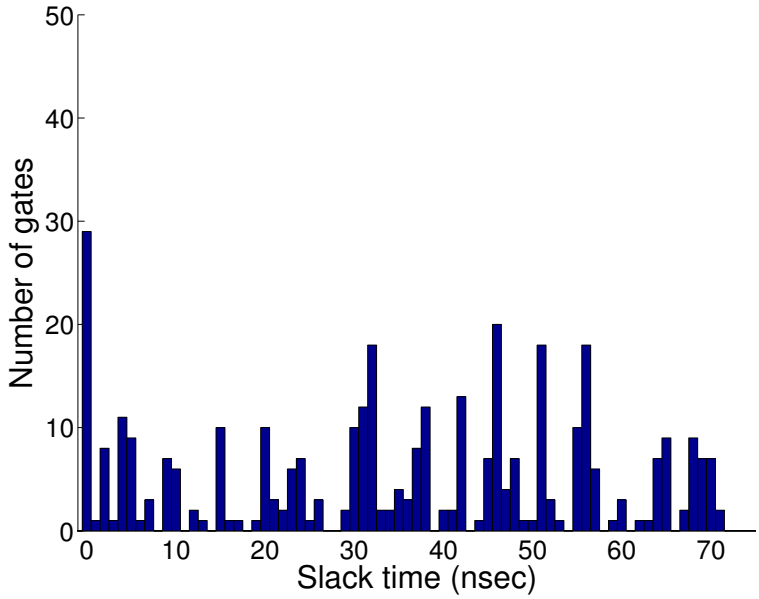


(b) Dual-V_{dd} design without level converters at $V_{DDH} = 0.23$V and $V_{DDL} = 0.14$V.

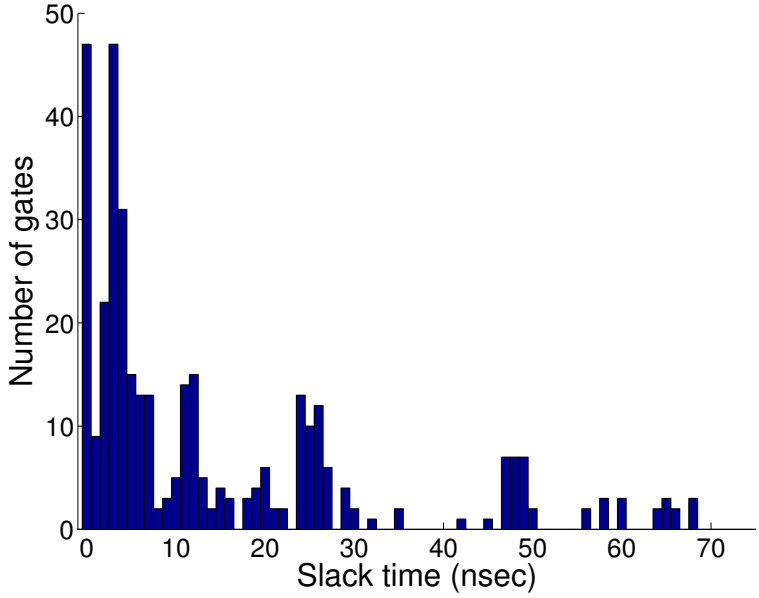


(c) Dual-V_{dd} design with multiple logic-level gates at $V_{DDH} = 0.23$V and $V_{DDL} = 0.16$V.

Figure 4.5: Gate slack distribution for minimum energy per cycle for c3540.



(a) Single-V_{dd} design at $V_{dd} = 0.24$V.



(b) Dual-V_{dd} design at $V_{DDH} = 0.24$V and $V_{DDL} = 0.19$V.

Figure 4.6: Gate slack distribution for minimum energy per cycle for c880.

reduction is another benefit of dual voltage design for low voltage circuits. The dual voltage technique for a nominal voltage circuit is mainly applied for dynamic power saving, while leakage power saving is considered negligible [39].

4.4 Summary

We presented a dual-V_{dd} design in which special multiple logic-level gates are used in the subthreshold regime [34]. This approach is particularly beneficial for subthreshold voltage operation. A new MILP is devised to find an optimal low supply voltage below a given subthreshold supply voltage. The given supply voltage is chosen for the minimum energy per cycle for any single voltage. When paired with the lower voltage from the MILP, the energy is further reduced. The MILP optimally selects the boundaries between the supply voltage domains to position multiple logic-level gates. With this MILP, ISCAS'85 benchmark circuits could save up to 24.5% energy per cycle more than the previous MILP results in Chapter 3. Notably, the energy per cycle for these designs is always less than the absolute minimum energy point for the circuit with single voltage operation. Alternatively, the MILP can trade energy reduction for speed increase without letting the energy rise.

Chapter 5

Process Variation Effect on Minimum Energy Design Using Dual Subthreshold Supply

5.1 Multiple Supply Voltages

Utilizing the time slack for power reduction with multiple supply voltages has been presented with nominal operating circuits in [22]. The theoretical models assume non-crossing parallel signal paths and are developed to determine the effective number of power supply voltages for power saving. The power reduction effect becomes saturated as supply voltages are added to optimize a circuit. There is no reason to use more than three supply voltages for power reduction in above-threshold operating circuits, considering power penalties induced by multiple-V_{dd}.

For subthreshold circuits, we investigate the energy reduction effect from multiple-V_{dd} in a real benchmark circuit, c2670. To verify the energy saving of multiple-V_{dd} design from path slack as [22], we do not consider multiple voltage boundaries within the optimized benchmark circuit. Thus, we eliminate topological constraints in MILP [35] and modify it by allowing multiple-V_{dd} selections during minimizing energy consumption. Figure 5.1 shows gate slack distribution of c2670 at a single $V_{dd} = 0.30$V. After optimizing c2670 using MILP with up to quadruple V_{dd}, we obtained the results of optimal V_{DDL} and energy saving as shown in Table 5.1. Energy reduction effect is more quickly saturated with multiple-V_{dd} for a subthreshold circuit. It is not promising to utilize all of the time slack inside a circuit with multiple-V_{dd}, because gate delay exponentially depends on V_{dd}. Even optimized c2670 with quadruple V_{dd} improves more 4.3% energy saving, compared to the dual-V_{dd} design. The energy saving will be further reduced when we consider energy overhead from level converting devices to solve multiple voltage boundaries in real circuit design. Therefore, we focus on optimizing subthreshold circuits with dual-V_{dd} for minimum energy design.

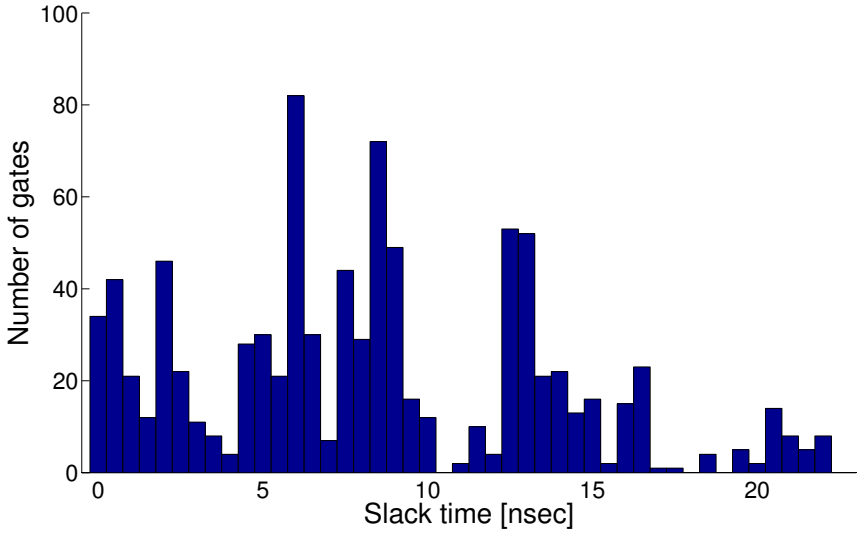


Figure 5.1: Gate slack distribution (number of gates vs. slack) for c2670 at V_{dd} = 0.30V; slacks obtained by static timing analysis using gate delays for PTM 90nm CMOS.

Table 5.1: The optimal V_{DDL} and energy saving of c2670 at V_{DDH} = 0.30V from MILP solutions [35] for multiple-V_{dd} design without topological constraints in PTM 90nm CMOS.

Multiple V_{dd}	Optimal V_{DDL}	Energy Saving (%)
Dual	0.24V	19.6
Triple	0.25V, 0.21V	22.8
Quadruple	0.26V, 0.22V, 0.17V	23.9

5.2 Technology Scaling

When performance is not a concern for energy constrained applications, a circuit can operate at the energy optimal voltage (V_{opt}) to achieve the minimum energy per cycle (E_{min}) by scaling V_{dd}. V_{opt} is theoretically independent of V_{th}, as reduced delay by V_{th} offsets increased leakage current in E_{leak}. The relative significance of E_{dyn} and E_{leak} determines V_{opt} when scaling V_{dd} [76]. When E_{leak} is larger than E_{dyn} in E_{tot}, then it causes V_{opt} value to move up to suppress E_{leak}. Conversely, larger E_{dyn} results in lower V_{opt} value. Thus, E_{dyn} and E_{leak} are quite close to the same value at V_{opt}.

For technology scaling, V_{opt} is proportional to S, which is dependent on the scaling [81, 7]. Without considering the slope of input signals, V_{opt} can be expressed as $K_{opt} \cdot S$, where K_{opt} is a dependent parameter of the circuit structure and independent of the scaling effect [23]. Using V_{opt}=$K_{opt} \cdot S$, total energy components, E_{dyn} and E_{leak}, in (3.6) are presented at the minimum energy point as [23]

$$\begin{aligned} E_{min,dyn} &= \alpha_{0\to1} \cdot C_L \cdot V_{opt}^2 \\ &= (\alpha_{0\to1} \cdot K_{opt}^2) \cdot C_L \cdot S^2 \\ E_{min,leak} &= K \cdot C_L \cdot V_{opt}^2 \cdot 10^{\frac{-V_{opt}}{S}} \\ &= (K \cdot 10^{-K_{opt}} \cdot K_{opt}^2) \cdot C_L \cdot S^2 \end{aligned} \tag{5.1}$$

where S increases and C_L decreases with technology scaling. Figure 5.2 shows the scaling trends of E_{min} and V_{opt} for a 32-bit RCA in PTM CMOS technology. Technology scaling apparently raises V_{opt} and reduces $C_L \cdot S^2$. Thus, minimum energy of a circuit is reduced and its performance may improve at V_{opt} on the device scaling.

Before investigating technology scaling effect on the energy saving of dual V_{dd} design for a subthreshold circuit, we derive the energy consumption ratio of dual-V_{dd} design to single V_{dd} reference in terms of E_{dyn} and E_{leak}. The dynamic energy ratio is given from (3.6)

$$\begin{aligned} \frac{E_{dyn,dual}}{E_{dyn,single}} &= \frac{\alpha_{0\to1} \cdot (C_{VL} \cdot V_{DDL}^2 + C_{VH} \cdot V_{DDH}^2)}{\alpha_{0\to1} \cdot C_{tot} \cdot V_{DDH}^2} \\ &= 1 - \frac{C_{VL}}{C_{tot}} \cdot \left(1 - \left(\frac{V_{DDL}}{V_{DDH}}\right)^2\right) \end{aligned} \tag{5.2}$$

where C_{VL} is the sum of load capacitances in V_{DDL} cells and C_{VH} is the sum of those in V_{DDH} cells. C_{tot} (= C_{VL}+C_{VH}) is total load capacitance of a circuit. V_{DDH} is equal to a single V_{dd} of the reference circuit. From (3.4), (3.6) and (3.7), the leakage energy ratio is

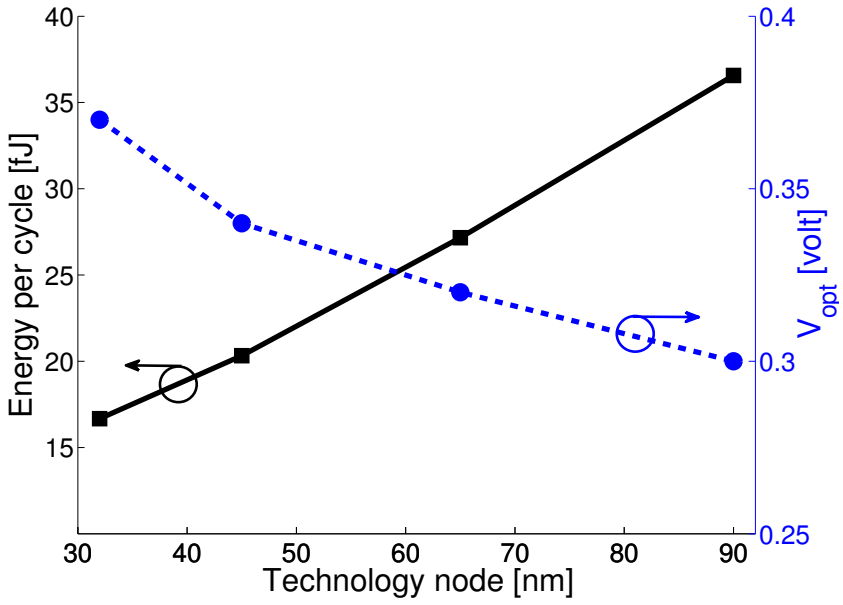


Figure 5.2: HSPICE simulation results of minimum energy per cycle and energy optimal voltage for a 32-bit RCA for a single-V_{dd} in PTM CMOS technology ($\alpha = 0.30$).

given as follows

$$\begin{aligned}\frac{E_{leak,dual}}{E_{leak,single}} &= \frac{I_{off,VL}\cdot V_{DDL}\cdot T_c + I_{off,VH}\cdot V_{DDH}\cdot T_c}{I_{off,tot}\cdot V_{DDH}\cdot T_c} \\ &= \frac{W_{VL}\cdot 10^{\frac{\eta V_{DDL}}{S}}\cdot V_{DDL} + W_{VH}\cdot 10^{\frac{\eta V_{DDH}}{S}}\cdot V_{DDH}}{W_{tot}\cdot 10^{\frac{\eta V_{DDH}}{S}}\cdot V_{DDH}} \\ &= 1 - \frac{W_{VL}}{W_{tot}}\cdot\left(1 - \frac{V_{DDL}}{V_{DDH}}\cdot 10^{\frac{-\eta(V_{DDH}-V_{DDL})}{S}}\right)\end{aligned} \tag{5.3}$$

where V_{VL} is the sum of device widths in V_{DDL} cells and V_{VH} is the sum of those in V_{DDH} cells. W_{tot} ($= W_{VL}$+W_{VL}) is the total device width of a circuit. T_c is a critical path delay for a circuit.

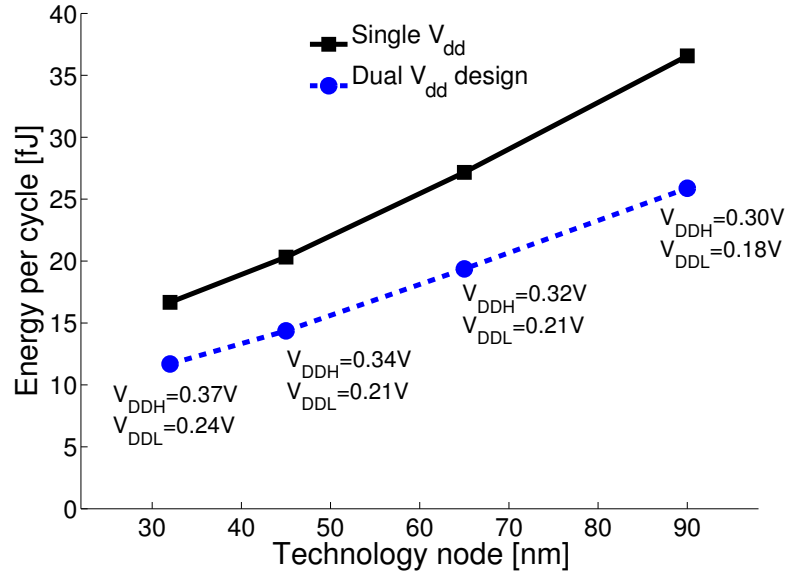
Applying a dual-V_{dd} technique for a subthreshold logic circuit on E_{min} (E_{dyn}≈E_{leak}), we need to find the optimal V_{DDL} for minimum energy consumption with given V_{DDH}=V_{opt}. We use the MILP algorithm [35] for dual-V_{dd} design to optimize a 32-bit RCA operating at V_{opt}. The MILP model does not allow a V_{DDL} cell to drive a V_{DDH} cell as its fanout gate on

account of topological constraint (similar to CVS). The results are shown in Figure 5.3(a). The minimum energy per cycle for the dual-V_{dd} circuit is further reduced from its minimum energy operation, while performance remains constant.

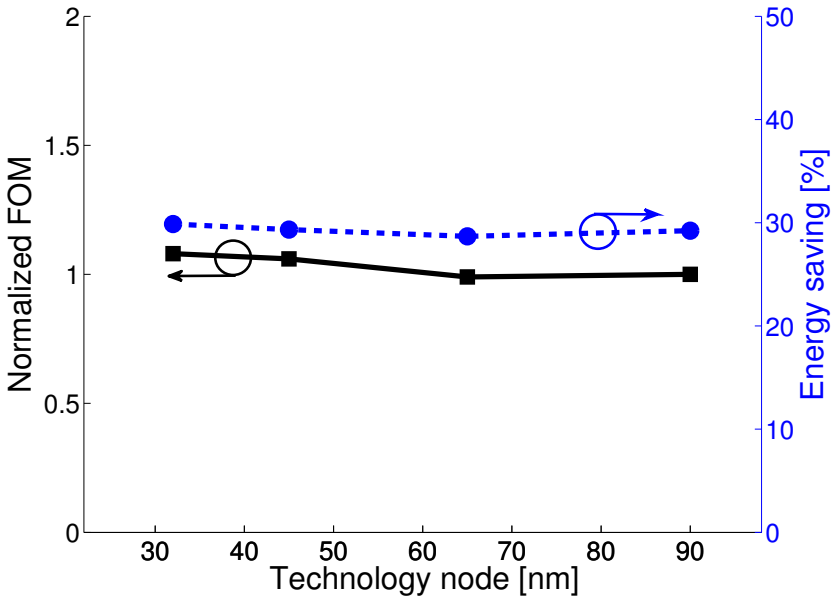
We introduce the figure-of-merit (FOM) of energy saving as *Number of* V_{DDL} *gates* times $(V_{DDH} - V_{DDL})$. The FOM is well matched with the energy saving of dual-V_{dd} design. Although V_{opt} moves to slightly higher value with technology scaling, device scaling does not considerably affect the energy saving in Figure 5.3(b). As seen in (5.2), the dynamic energy ratio is independent of technology scaling parameters and the scaling of load capacitance does not affect its ratio. The leakage energy ratio has S and η as technology parameters in (5.3), but both parameters increase together with device scaling [3]. Thus, the term of $\frac{\eta}{S} \cdot (V_{DDH} - V_{DDL})$ does not affect significantly the leakage energy saving, where the optimal V_{DDL} is close to V_{DDH} from exponential delay characteristic of a subthreshold logic circuit on scaling V_{dd}. Therefore, the amount of total energy saving comes from circuit structure, rather than technology choice. It means that the distribution of time slack in a circuit structure is not changed by device scaling. For each technology, only small variation of FOM and energy saving in Figure 5.3(b) may come from relatively different delay increments of logic gates on scaling V_{dd} [81]. But, it does not alter considerably the time slack distribution of a subthreshold logic circuit.

5.3 Process Variation

Subthreshold circuits are highly sensitive to V_{th} variation, which exponentially affects I_{on} and delay. V_{th} variation also causes different relative strength of PMOS and NMOS devices and thus affects functional failure of logic gates [42]. Variability of V_{th} comes from global (inter-die) and local (intra-die) process variations [3]. Global variation of V_{th} is induced by manufacturing process and temporal variation, but it can be compensated through the adaptive body biasing (ABB) technique [24]. Random dopant fluctuation (RDF) is the dominant source of local V_{th} variation compared to geometric variations such as L_{eff} in the



(a) Minimum energy per cycle at $V_{dd} = V_{opt} = V_{DDH}$.



(b) Normalized FOM and energy saving.

Figure 5.3: The optimal V_{DDL} from MILP [35] algorithm and total energy per cycle from HSPICE simulation of dual-V_{dd} design for a 32-bit RCA (Fig. 5.2) in PTM CMOS Technology. The relationship of figure of merit (FOM) to energy saving is shown for technology scaling trend.

subthreshold region [82]. RDF variations have independent nature and inverse dependence on $(WL)^{-\frac{1}{2}}$. Therefore, local V_{th} variation can be reduced by the gate sizing and logic depth choice through averaging variability [53, 82].

To investigate the effect of V_{th} variability on dual-V_{dd} design for a subthreshold logic circuit, we normally randomize the *vth0* parameter in the BSIM4 model card of PTM CMOS technology [85] in the Monte Carlo simulation. For the global variations, we characterize the standard deviation (σ_{vth0}) as 5% variation relative to its original *vth0* value for both PMOS and NMOS devices. This presents samples of logic gates through multiple dies as inter-die process variation. As the local variation, RDF is modeled from an empirical expression [2, 3] through normally distributed *vth0* with

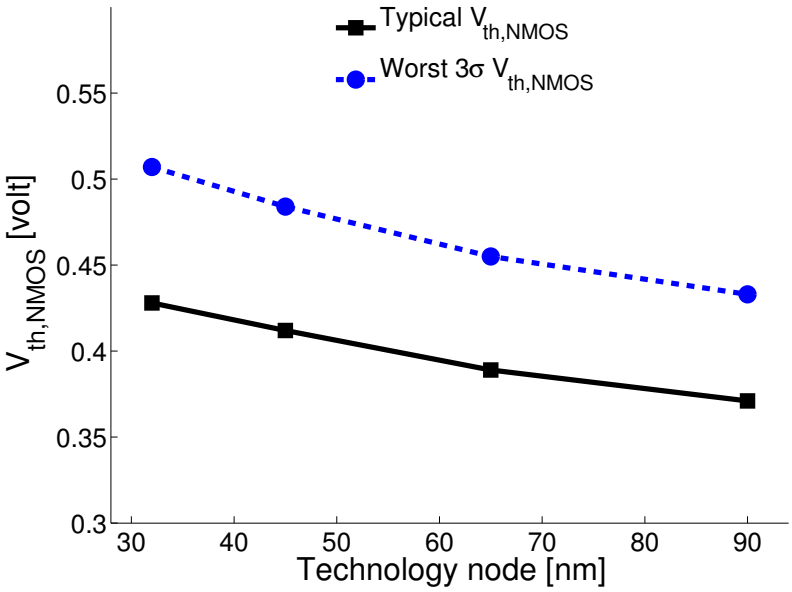
$$\sigma_{vth0,RDF} = 3.19 \times 10^{-8} \frac{T_{ox} \cdot N_{ch}^{0.4}}{\sqrt{W_{eff} \cdot L_{eff}}} \tag{5.4}$$

where T_{ox} is the gate equivalent oxide thickness and N_{ch} is the channel doping concentration. L_{eff} and W_{eff} are the effective channel length and width of device, respectively. Both σ_{vth0} and $\sigma_{vth0,RDF}$ demonstrate entire V_{th} variation of a subthreshold circuit, which is still normally distributed.

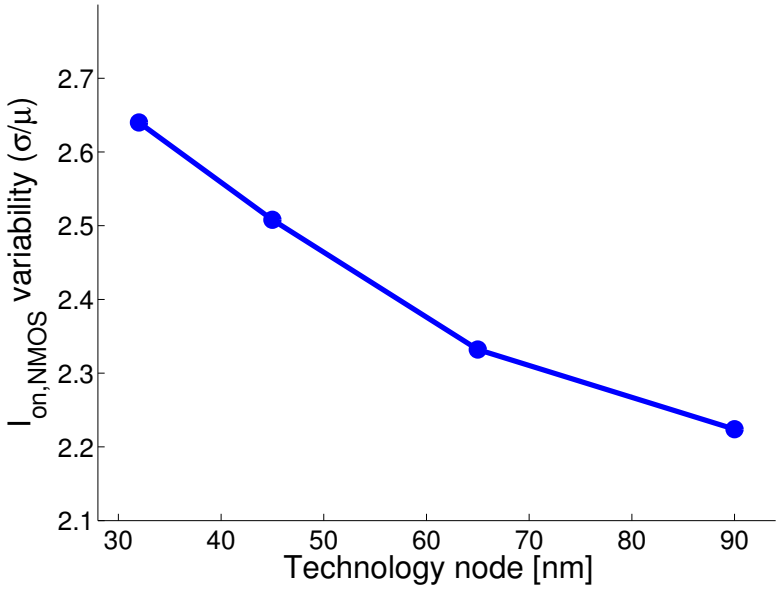
We ran a 1k-point Monte Carlo simulation using HSPICE simulator [27] with global and local *vth0* variations. Figure 5.4(a) shows the simulation result of NMOS V_{th} variation for technology scaling. For subthreshold supply voltage, V_{dd}=0.30V, the worst 3σ V_{th} value is as high as 79mV than the typical V_{th} value in PTM 32nm NMOS compared to 62mV in PTM 90nm NMOS. Therefore, V_{th} variation is higher with small feature size.

Under normally distributed V_{th} variation in the subthreshold region, active current I_{on} variability can be modeled as lognormal random variable and exhibits lognormal distribution [82, 42] with

$$\frac{\sigma_{I_{on}}}{\mu_{I_{on}}} = \sqrt{e^{(\frac{\sigma_{V_{th}}}{mV_T})^2} - 1} \tag{5.5}$$



(a) NMOS threshold voltage variation.



(b) Active current $I_{on,\mathrm{NMOS}}$ variability.

Figure 5.4: HSPICE simulation results of NMOS V_{th} variation and active current I_{on} variability at $V_{dd} = 0.30$V from a 1k-point Monte Carlo simulation with normally distributed vth0 parameter in PTM CMOS technology.

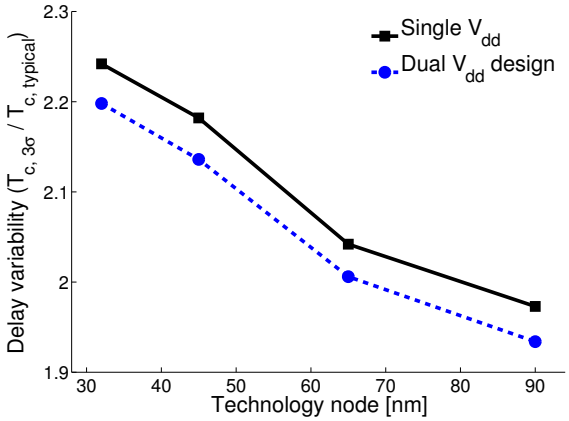
where the subthreshold slope coefficient m decreases as V_{dd} reduces. It causes an increase in I_{on} variability in low voltage operation. As shown in Figure 5.4(b), I_{on} variability is up to 2.64X from the mean value in PTM 32nm CMOS. Since I_{on} exponentially depends on V_{th}, I_{on} variability is higher than V_{th} variation. It also induces the delay variability of a subthreshold circuit from (3.2).

As mentioned before, the gate sizing and logic depth choice of a subthreshold circuit reduce independent local V_{th} variation through averaging. Figure 5.5(a) shows the worst case critical path delays of the single-V_{dd} and dual-V_{dd} 32-bit RCA in Figure 5.3(a) from 1k-point Monte Carlo simulation. For single-V_{dd} design, worst 3σ critical delay is reduced through averaging compared to I_{on} variability in Figure 5.4(b).

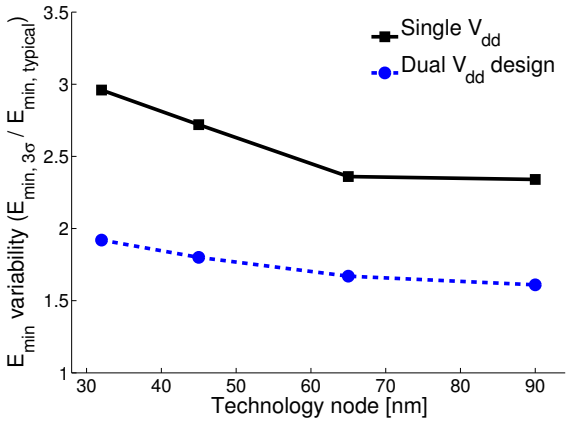
Dual-V_{dd} design uses two supply voltages which can provide a chance for reducing critical path delay in a circuit. In the subthreshold region, the gate capacitance of the MOS device may reduce when V_{dd} goes down [26]. The critical path delay reduces when V_{DDH} gates on the critical path drive V_{DDL} gates as fanout.As shown in Figure 5.6(a), an inverter ($V_{DDH} = 0.30$V) driving four inverters ($V_{DDL} = 0.18$V) reduces its output capacitance load. Figure 5.6(b) shows the delay of the inverter reduces about 8% from the reduced output capacitance.

From this aspect, the worst critical delay of dual-V_{dd} 32-bit RCA is less than that of a single-V_{dd} 32-bit RCA. The worst critical delay depends on V_{DDL} assignment to the fanout gates of V_{DDH} gates on the critical path.

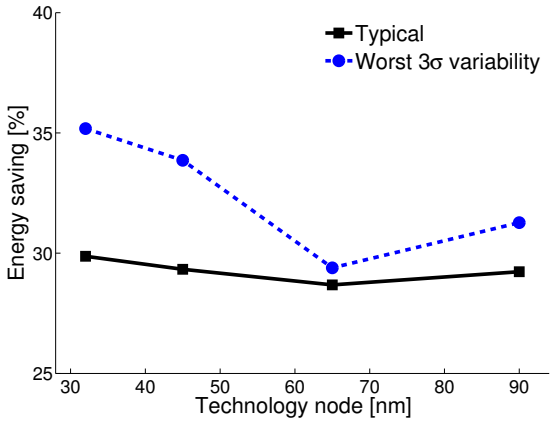
We also measure minimum energy variability for the single and a dual-V_{dd} 32-bit RCA using each 3σ critical delay with V_{th} variation as shown in Figure 5.5(b). Compared to typical E_{min} with a single-V_{dd}, both minimum energies increase with delay variability, which induces more leakage energy from the extended operation time. In PTM 32nm CMOS, the worst case of E_{min} for dual-V_{dd} design is 1.92 times typical E_{min}, while that of E_{min} for a single-V_{dd} is 2.96 times. It means that the worst 3σ E_{min} of dual-V_{dd} design is reduced 35.2% from the worst case E_{min} with a single-V_{dd}. Dual-V_{dd} design for subthreshold circuits is more effective



(a) Worst 3σ critical path delay variability.

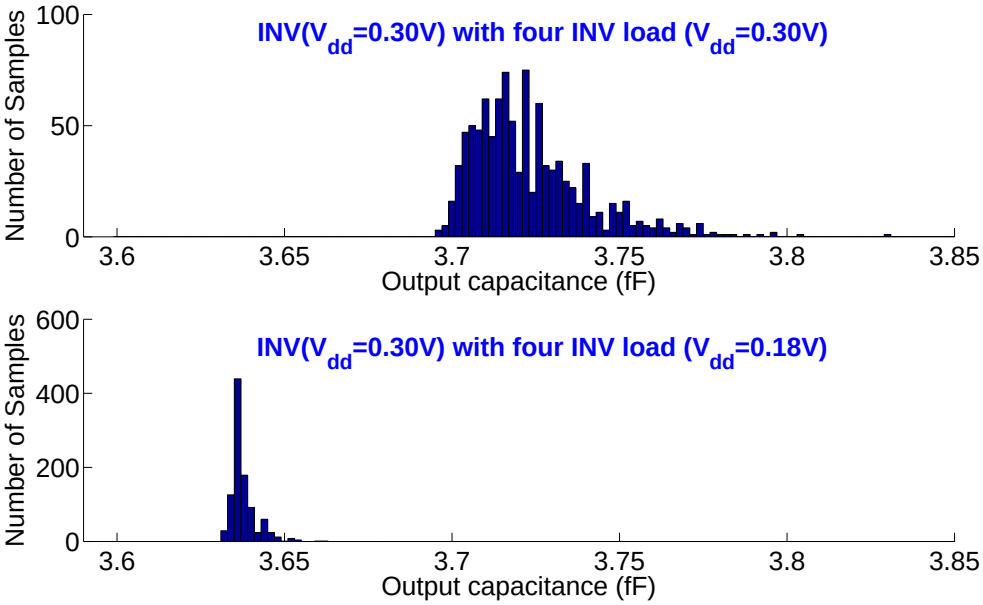


(b) Worst 3σ minimum energy variability.

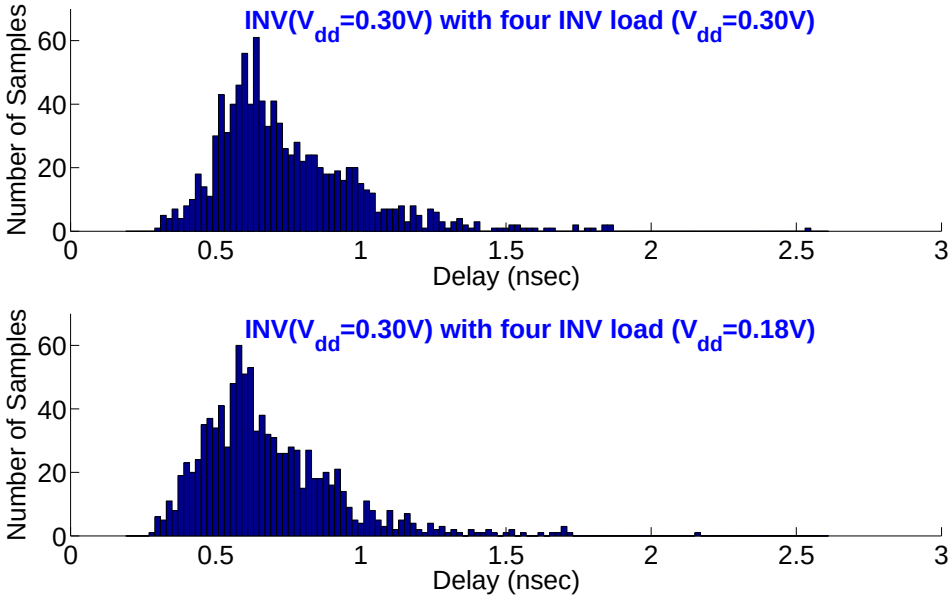


(c) Energy saving with worst 3σ variability.

Figure 5.5: HSPICE simulation results of critical path delay and minimum energy for a 32-bit RCA (Fig. 5.3(a)) from a 1k-point Monte Carlo simulation in PTM CMOS technology.



(a) Output capacitance.



(b) Delay variability: $t_{d,3\sigma} = 1.51$ns (0.30V,0.30V), $t_{d,3\sigma} = 1.39$ns (0.30V,0.18V).

Figure 5.6: Distribution of the output capacitance and delay variability for an inverter with fanout of four from a 1k-point Monte Carlo simulation with normally distributed vth0 parameter in PTM CMOS technology.

to mitigate increment of minimum energy with process variation in small feature sizes. Thus, we expect more energy saving when variability is more of a concern. Figure 5.5(c) shows energy savings of dual-V_{dd} 32-bit RCA with and without process variation for technology scaling.

5.4 Summary

A subthreshold circuit is susceptible to process variation, which affects the delay of gates. Dual-V_{dd} design may mitigate the delay variability of a circuit in the subthreshold region, when V_{DDL} is assigned to more fanout gates of V_{DDH} gates on the critical path. The worst delay reduction comes from the reduced gate capacitance of V_{DDL} fanout gates. Thus, we expect more energy saving when process variation is more concerned. Dual-V_{dd} technique is valid and beneficial for minimum energy design.

A recent study has investigated the process variation in 45nm bulk and high-k CMOS technologies [70, 71]. As pointed out in that study, there may be some advantages for subthreshold circuits in the 45nm high-k technology but more detailed work is needed.

Chapter 6

Dual Voltage Design for Minimum Energy Using Gate Slack

In this chapter, we present a new slack-time based algorithm for dual-V_{dd} design with linear-time complexity. Although a global optimum is sought, computation time is kept low. The slack of a gate is defined as the difference between the critical path delay for the circuit and the delay of the longest path through that gate. Positive non-zero slack gates are classified into two groups, one in which all gates can be unconditionally assigned low voltage and the other where only a selected subset can be assigned low voltage without violating the positive non-zero slack requirement. Multiple voltage boundaries are given special consideration to avoid the use of level shifting devices. The overall complexity of this power optimization algorithm is linear in number of gates as compared to a previously published exponential-time exact algorithm using mixed integer linear program (MILP).

Two heuristic algorithms, CVS and ECVS, for dual-V_{dd} design have theoretical run-time complexity $O(n^2)$, where n is total number of gates in a circuit [13]. Most research in this field has focused on improving power saving by implementing their own greedy algorithms [11, 12, 39]. These are still heuristic approaches and provide a suboptimal solution for dual-V_{dd} assignment. Mixed integer linear programs (MILP) [19] are widely used to optimize a circuit for minimizing power or energy consumption using sizing, multiple V_{dd}, multiple threshold voltage (V_{th}) and combinations of those [15, 34, 35, 64]. MILP searches for a global optimal solution for an objective function, which is designed to minimize power, considering the entire design space. Thus, it may take huge time to optimize large circuits used in modern VLSI systems. The time complexity of MILP optimization may not be acceptable in practice.

For dual-V_{dd} design, we need to find the optimal V_{DDL} and its assignments to positive slack gates in a circuit for minimum power. If we can quickly find all positive slack gates

that can be assigned to V_{DDL}, it reduces much optimization work of dual-V_{dd} design and saves computation time.

6.1 MILP for Optimal V_{DDL} and Dual V_{dd} Assignment

There are two ways to find the optimal lower supply voltage V_{DDL} and its assignments for dual-V_{dd} design in the literature. First, the optimal V_{DDL} is searched by applying a V_{DDL} assignment algorithm to a circuit with different V_{DDL} values, then it selects a pair of the optimal V_{DDL} and its assignment for minimum power consumption [12, 67, 68]. Otherwise, theoretical path delay model is developed to determine the optimal V_{DDL} for maximum power saving, then V_{DDL} assignments are executed to achieve lowest power consumption considering multiple voltage boundaries [22, 39]. Most dual-V_{dd} techniques are based on heuristic greedy algorithms and applied to nominal operating circuits for lowering power consumption.

For energy constrained applications, the dual-V_{dd} technique is applied to a subthreshold logic circuit for further reducing the minimum energy operating point [35], where the MILP models similar to CVS are formulated to find the best optimal V_{DDL} and its assignments for dual-V_{dd} design. This global optimum algorithm is applicable to a circuit operating at both subthreshold and nominal supply voltage, but multiple runs are needed to consider all available V_{DDL} to given V_{DDH} for searching the optimal V_{DDL}. Now, we extend the MILP models to select automatically the optimal V_{DDL} and its assignments by introducing new variables for one-time run. We briefly explain the new variables and parameters here before presenting the MILP models.

- $X_{i,v}$: supply voltage assignment integer variable that is 1 for gate i with power supply voltage v.
- V_v: supply voltage integer variable that is 1 for two selected V_{DDH} and V_{DDL} in available power supply voltage v.

- $td_{i,v}$: gate delay for gate i with supply voltage v.
- $V_{dd,v}$: power supply voltage value for v.
- G_{tot}: total number of gates in a circuit.

MILP models are reformulated from [35]:

$$\begin{aligned} &\text{Minimize } \sum_i \sum_v E_{tot,i,v} \cdot X_{i,v} \\ &\forall i \in \text{all gates and } \forall v \in \text{power supply voltage domain } V \end{aligned} \tag{6.1}$$

$$E_{tot,i,v} = \alpha_i \cdot C_{L,i,v} \cdot V_{dd,v}^2 + P_{leak,i,v} \cdot T_c \tag{6.2}$$

Subject to timing constraints:

$$T_i \geq T_j + t_{d,i,v} \cdot X_{i,v} \quad \forall j \in \text{all fanin gates of gate } i \tag{6.3}$$

$$T_i \leq T_c \qquad \forall i \in \text{all primary output gates} \tag{6.4}$$

Subject to topological constraints:

$$\begin{aligned} &\sum_{v \in V} V_{dd,v} \cdot X_{i,v} \leq \sum_{v \in V} V_{dd,v} \cdot X_{j,v} \\ &\forall j \in \text{all fanin gates of gate } i \end{aligned} \tag{6.5}$$

Subject to dual supply voltages selection:

$$\sum_{v \in V} V_v = 2 \tag{6.6}$$

$$V_{V_{DDH}} = 1 \tag{6.7}$$

$$\sum_{v \in V} X_{i,v} = 1 \qquad \forall i \in \text{all gates} \tag{6.8}$$

$$\sum_i X_{i,v} \leq G_{tot} \cdot V_v \quad \forall i \in \text{all gates}, \ \forall v \in V \tag{6.9}$$

The main difference of MILP models from [35] is dual-V_{dd} selection conditions. T_c is critical path delay and given by the performance requirement. V_{DDH} is selected to hold T_c from (6.7) in power supply domain V. Using a bin-packing technique [1] all gates must be assigned to one of the power supply voltages in V from (6.8) and (6.9).

MILP always guarantees that a dual-V_{dd} circuit with the optimal V_{DDL} and it assignments achieve minimum energy consumption at the same performance. We use absolute optimal results of MILP as a reference to check the accuracy of our slack-time based algorithm that is presented in the next section.

6.2 New Slack-Time Based Algorithm for Dual-V_{dd} Design

In this section, we propose a new slack-time based algorithm that finds the optimal V_{DDL} and its assignments for dual-V_{dd} design. The energy saving is as much as the optimal solution from the MILP model.

First, our algorithm generates slack time distribution for a given circuit. We have developed an expanded version of static timing analysis (STA) [25]. For the output of gate i, let $T_{PI}(i)$ be the longest time for an event to arrive from a PI and $T_{PO}(i)$ be the longest time for an event to reach a PO. The delay of the longest path [45, 46] through gate i is given by,

$$D_{p,i} = T_{PI}(i) + T_{PO}(i) \tag{6.10}$$

The critical path delay for the circuit is,

$$T_c = \text{Max}\{D_{p,j}\} \quad \forall \text{ gate } j \tag{6.11}$$

Slack time for gate i is found as follows:

$$S_i = T_c - D_{p,i} \tag{6.12}$$

The time for calculating slack time for all gates of a circuit is $O(n)$, where n is total number of gates. Figure 6.1(a) shows the slack time distribution for c2670 in ISCAS'85 benchmark circuits in PTM 90nm CMOS technology [85].

To quickly identify the possible V_{DDL} gates on non-critical paths, we introduce an upper slack time (S_u) that guarantees that any gate with slack time larger than S_u will be free from timing violation, i.e., *negative slack*, irrespective of the voltage assignment for other gates. The slack time of a V_{DDH} gate that is equal to S_u becomes zero after assigning V_{DDL} to all gates on the longest path through it. We find S_u using (6.12). Let S'_i be the slack time of gate i after assigning V_{DDL} to all gates on the longest path through it. Now, $D'_{p,i}$ is the longest path delay through the gate i.

$$\begin{aligned} S'_i &= T_c - D'_{p,i} \\ &= T_c - \beta \cdot D_{p,i} \\ &= T_c - \beta \cdot (T_c - S_i) \end{aligned} \tag{6.13}$$

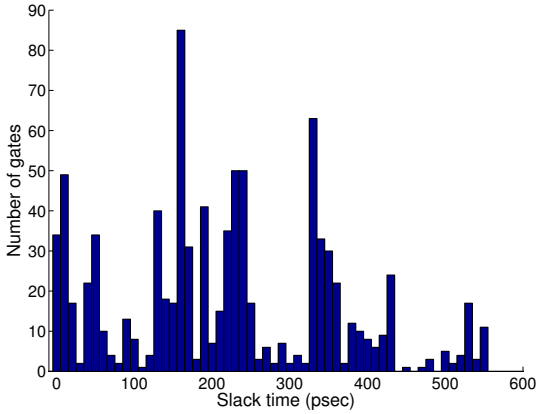
Where β is the ratio of $D_{p,i}$ to $D'_{p,i}$. It is approximated by

$$\beta = \frac{D'_{p,i}}{D_{p,i}} \approx \frac{T'_c}{T_c} \tag{6.14}$$

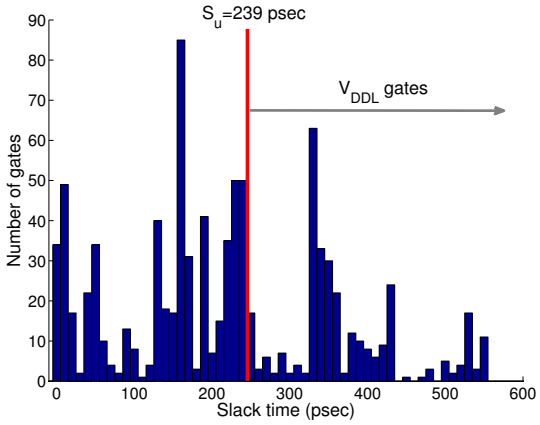
T'_c is the critical path delay when V_{DDL} is supplied to the entire circuit. It is determined by the static timing analysis in the same way as [25]. By substituting S_u for S_i in (6.13), S'_i become zero. Thus, S_u is obtained as:

$$S_u = \frac{\beta - 1}{\beta} \cdot T_c \tag{6.15}$$

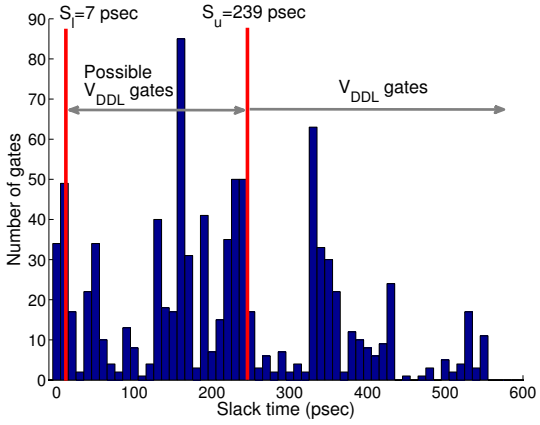
In Figure 6.1(b), any gate that has a positive slack time larger than S_u, i.e., in the range covered by the right arrow, is safely assigned to V_{DDL} without timing violation. S_u serves



(a) Slack time distribution for a single nominal $V_{dd} = 1.2V$.



(b) Upper slack time S_u at $V_{DDH} = 1.2$V and $V_{DDL} = 0.69V$.



(c) Lower slack time S_l at $V_{DDH} = 1.2V$ and $V_{DDL} = 0.69V$.

Figure 6.1: Procedure of slack-time based algorithm for ISCAS'85 benchmark circuit c2670 in PTM 90nm CMOS.

as a slack threshold. Any gate with slack above this threshold is unconditionally assigned to V_{DDL} irrespective of voltages of other gates on paths passing through it.

The slack time of all gates on critical paths is zero. Hence, there is no room to assign V_{DDL} to those gates. But, if there is a gate with a positive slack time that is close to zero, it may be possible to assign V_{DDL}, provided other gates on paths through it remain with V_{DDH}, such that no path delay exceeds T_c.

Let t_d be a gate delay in the circuit. After assigning V_{DDL}, t_d is increased. Suppose, it becomes t'_d. The amount $t'_d - t_d$ is the increase in path delay through the gate. This is also the reduction in the slack of other gates on paths through the V_{DDL} gate. Therefore, a gate that has slack time larger than $t'_d - t_d$ can be assigned to V_{DDL}. Let us call this slack time the lower slack time (S_l). Because each logic gate has a different value of $t'_d - t_d$, the minimum value of $t'_d - t_d$ is used to define S_l.

$$\begin{aligned} S_l &= \text{Min}\left[(t'_d - t_d)_{\text{gates } j}\right] \\ &= \text{Min}\left[(\beta - 1) \cdot t_{d\,\text{gates } j}\right] \quad \forall j \in \text{all gates} \\ &\text{assume } \frac{t'_{d,j}}{t_{d,j}} \approx \frac{D'_{p,j}}{D_{p,j}} = \beta \end{aligned} \tag{6.16}$$

For simplicity, we assume that path delay is proportional to the delay of a gate on it. Timing violations from this assumption are checked later when V_{DDL} gates are chosen, finally.

As shown in Figure 6.1(c), S_l can be used to search possible V_{DDL} gates between S_l and S_u, the range shown by a double arrow. The gates with positive slack time less than S_l are unconditionally assigned to V_{DDH} and are located near or on critical paths.

Until now, we have demonstrated how to select gates that can be assigned to V_{DDL} using simple two slack times, S_u and S_l. A gate with slack time larger than S_u is assigned to V_{DDL}, while a gate with slack time less than S_l is assigned to V_{DDH}. For a gate with slack time between S_l and S_u, we need to carefully select the power supply voltage. V_{DDL} assignment for these gates affects the assignment of other gates on paths if we have to hold the path

delay within T_c. The order of V_{DDL} assignment to these gates affects the energy saving of the dual-V_{dd} design, when we consider multiple voltage boundaries. Thus, we need to use a greedy approach depending on the type of dual-V_{dd} design. If we allow V_{DDL} gates to drive V_{DDH} gates like ECVS, the selection order should minimize the use of level converters to maximize energy saving. Because CVS does not use level converters, there exists topological constraints that prevent a V_{DDL} gate from driving a V_{DDH} gate. Therefore, the selection order is chosen to maximize V_{DDL} assignment to gates with this topological constraint.

In this chapter, we use the slack time distribution to implement a dual-V_{dd} algorithm like CVS. The result of the algorithm is compared to MILP solution in terms of energy saving and run-time. To maximize V_{DDL} assignment with topological constraints, first, higher logic depth gates between S_l and S_u should be assigned to V_{DDL}. This priority reflects the fact that V_{DDL} gates do not feed into V_{DDH} gates directly. The timing violation should be checked when a gate between S_l and S_u is assigned to V_{DDL}. We find all V_{DDL} gates, which do not violate the critical path timing constraint T_c. Additionally, checking topological constraints for these V_{DDL} gates, we ascertain that all V_{DDL} gates satisfy both timing and topological constraints.

The final stage of the algorithm searches for the optimal V_{DDL} value to give maximum energy saving. We already know all V_{DDL} gates for each available V_{DDL} value from previous procedures. Thus, we simply calculate the energy saving from V_{DDL} gates, then select the optimal V_{DDL} to meet best energy saving. Figure 6.2 shows the slack time distribution of an optimized c2670 circuit that has the optimal $V_{DDL} = 0.69$V from our algorithm. In next section, we show the results of optimization from the slack-time based algorithm for ISCAS'85 benchmark circuits, which operate in either subthreshold or nominal supply voltage.

6.3 Simulation Results

As example circuits, ISCAS'85 benchmark circuits are synthesized with four types of basic standard cells, namely, INV, NAND2, NAND3, and NOR2. Average activity of a

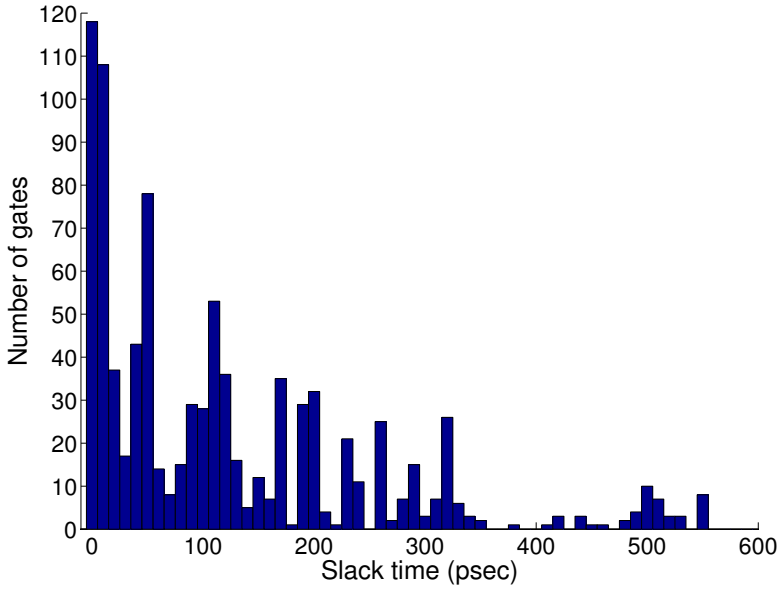


Figure 6.2: Slack time distribution of an optimized c2670 with $V_{DDH} = 1.2V$ and $V_{DDL} = 0.69V$.

synthesized circuit is found from logic simulation with randomly generated input vectors. We extract gate delay, capacitance and leakage power of basic standard cells through HSPICE simulation by varying power supply voltage from 0.1V to 1.2V in 10mV steps. All HSPICE simulations were run for room temperature (300K) using PTM 90nm CMOS process, where CMOS device threshold voltages are $V_{th,pmos} = 0.21$V and $V_{th,nmos} = 0.29$V at nominal $V_{dd} = 1.2$V.

For comparing the algorithm of Section 6.2 with MILP of Section 6.1, we measure the energy consumption of benchmark circuits using HSPICE simulation [27] for a single-V_{dd} as a reference. Random input vectors for each circuit in HSPICE simulation are the same as those used in logic simulation to measure the average activity. To find the optimal V_{DDL} and its assignments for maximum energy saving, the MILP algorithm is applied to a synthesized circuit. With MILP solution, the SPICE netlist of an optimized circuit is generated, where each gate has its voltage assignment either as the given V_{DDH} or an optimal V_{DDL}. HSPICE simulation runs with this netlist to measure energy consumption of the optimized dual-V_{dd}

Table 6.1: Energy saving and optimal V_{DDL} from MILP [35] or slack-time based algorithm for given V_{DDH} in ISCAS'85 benchmark circuits in subthreshold region in PTM 90nm CMOS. Both algorithms produced identical result.

Benchmark circuit	Total gates	Activity α	V_{DDH} (V)	V_{DDL} (V)	V_{DDL} gates (%)	E_{single} (fJ)	E_{dual} (fJ)	$E_{reduc.}$ (%)	Freq. (MHz)	MILP CPU time(s)*	Slack CPU time(s)*
c432	154	0.19	0.25	0.23	5.2	7.9	7.8	1.1	14.4	0.3	2.5
c499	493	0.21	0.22	0.18	9.7	20.2	19.8	2.0	11.9	0.3	19.2
c880	360	0.18	0.24	0.18	46.4	14.4	11.2	22.2	13.6	5.8	17.9
c1355	469	0.21	0.21	0.18	10.2	19.5	19.0	2.5	9.8	0.2	13.3
c1908	584	0.20	0.24	0.21	24.3	26.5	25.0	5.8	11.8	3.2	47.6
c2670	901	0.16	0.25	0.21	46.4	32.8	28.0	14.8	17.4	35.9	134.4
c3540	1270	0.33	0.23	0.14	7.0	88.0	84.6	3.8	7.2	3.2	256.5
c5315	2077	0.26	0.24	0.19	47.1	116.8	98.0	16.1	9.8	852.3	692.0
c6288	2407	0.28	0.29	0.18	2.7	165.4	162.0	2.1	9.4	2.6	1293.7
c7552	2823	0.20	0.25	0.21	42.3	131.7	117.1	11.1	13.6	1452.2	1408.3
Average					24.1			8.2			

*Intel Core 2 Duo 3.06GHz, 4GB RAM.

Table 6.2: Energy saving and optimal V_{DDL} from MILP [35] and slack-time based algorithm for ISCAS'85 benchmark circuit operating in nominal V_{dd} in PTM 90nm CMOS.

	Single V_{dd}			Dual V_{dd}									
				MILP					Slack-time based algorithm				
Benchmark circuit	V_{DDH} (V)	E_{single} (fJ)	Freq. (GHz)	V_{DDL} (V)	V_{DDL} gate (%)	E_{dual} (fJ)	$E_{reduc.}$ (%)	CPU (s)*	V_{DDL} (V)	V_{DDL} gate(%)	E_{dual} (fJ)	$E_{reduc.}$ (%)	CPU (s)*
c432	1.20	160.1	1.7	0.75	5.2	153.9	3.9	0.6	0.75	5.2	153.9	3.9	15.8
c499	1.20	460.6	2.3	0.79	19.5	433.4	5.9	403.8	0.79	19.5	433.4	5.9	194.4
c880	1.20	277.6	2.0	0.59	56.9	136.1	51.0	455.0	0.60	57.5	136.6	50.8	62.1
c1355	1.20	453.0	2.3	0.69	13.6	433.6	4.3	340.2	0.69	13.6	433.6	4.3	132.0
c1908	1.20	496.5	1.5	0.67	26.9	402.4	19.0	2146.9	0.67	26.9	402.4	19.0	247.8
c2670	1.20	647.6	1.8	0.69	57.9	337.9	47.8	20848.9	0.69	57.9	337.9	47.8	480.7
c3540	1.20	1844.0	1.1	0.70	11.6	1667.0	9.6	601.0	0.70	11.6	1667.0	9.6	1243.5
c6288	1.20	3066.0	0.5	1.18	53.1	2976.0	2.9	10523.7	0.47	2.9	2985.0	2.6	6128.0
Average					30.6		18.0			24.4		18.0	

*Intel Core 2 Duo 3.06GHz, 4GB RAM.

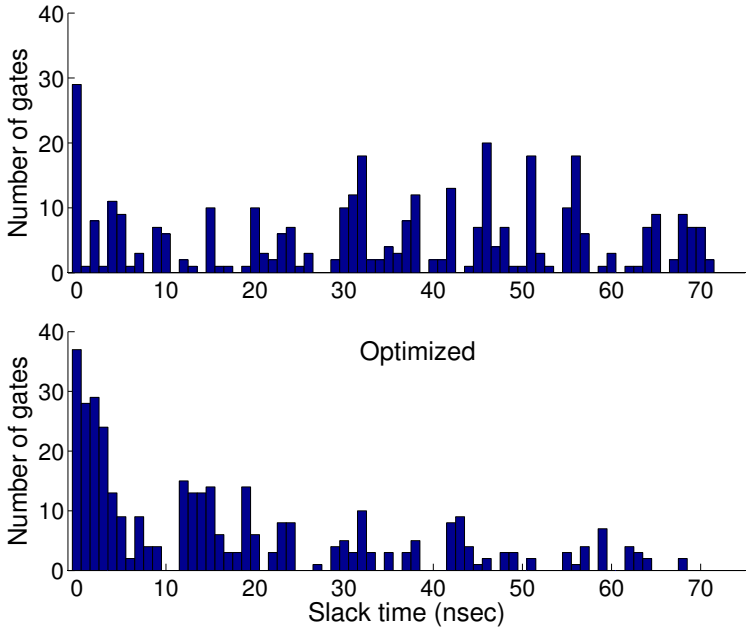
circuit. The same procedure is repeated for the design obtained by the slack-time based algorithm.

First, we apply both algorithms to benchmark circuits operating in the subthreshold region. We assume that V_{DDH} at minimum energy is given with the corresponding speed for each benchmark circuit. Table 6.1 shows HSPICE simulation results from the two algorithms. The results of the two algorithms exactly match each other. Using dual-V_{dd} design, total energy saving for c880 (8-bit ALU) is 22.2% as the best case.

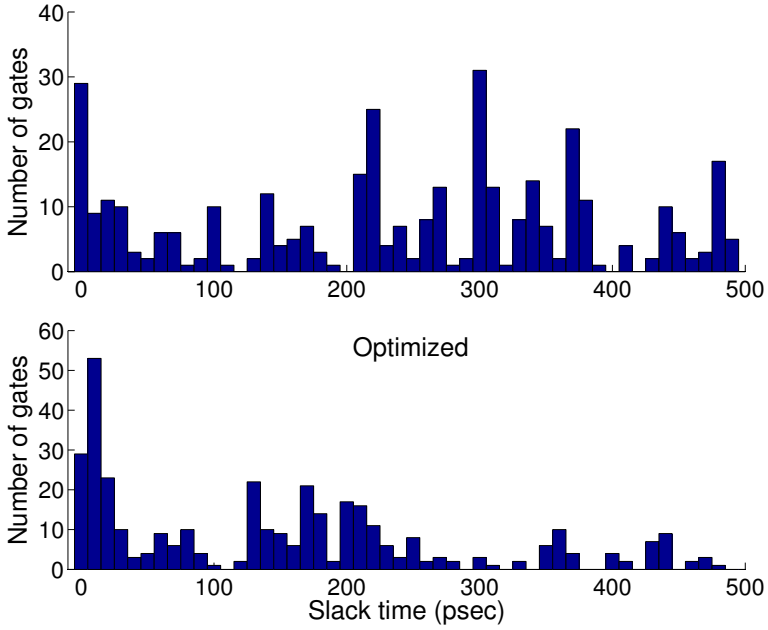
In Table 6.2, both algorithms are applied to optimize benchmark circuits operating with the nominal supply voltage. We set 1.2V as a nominal power supply voltage for PTM 90nm CMOS by referring to the industry standard 90nm CMOS technology. The results from from the two algorithms do not match for c880 and c6288, but energy savings are very close. Evidently, the result of the slack-time based algorithm is very close to the global optimization, even though it uses a greedy heuristic to select the best V_{DDL} gates from all V_{DDL} gates that pass timing constraint. For c880, energy savings from MILP and our algorithm are 51.0% and 50.8%, respectively. Compared to the energy savings in the subthreshold region, these energy savings are much larger. This is because lower supply voltage increases the gate delay exponentially in the subthreshold region, while the gate delay increase for the nominal voltage operation is polynomial according to the alpha-power law model [60, 76]. It means that positive slack of gates in a circuit is reduced quicker by assigning V_{DDL} in subthreshold region. Thus, we obtain an optimal V_{DDL} that is closer to V_{DDH} and there are fewer V_{DDL} gates as well. Figure 6.3 shows slack time distributions before and after optimization by our algorithm applied to c880 for both subthreshold and nominal voltage operations.

We measured the run-time of two algorithms based on CPU time in seconds. Our algorithm is written in the Perl script language. Thus, it has inherently slower execution than a program in the C language. The run time for MILP depends on the number of integer variables, the complexity of inequalities that specify the linear constraints, and the size of optimization space. From Table 6.1, MILP is mostly faster than our algorithm except for c5315 and c7552. Both circuits have large slacks and have larger optimization spaces to be searched. Also, available V_{DDL} as an integer variable in MILP is limited by minimum operating voltage that guarantees correct logic function for the lower supply voltage. It is 0.1V below the point at which the circuit function fails. This limitation reduces the size of the optimization space for the MILP algorithm.

In Table 6.2, the run time of our algorithm is $\sim$43X faster than MILP for c2670, because a larger range for V_{DDL} in nominal operation needs to be searched by MILP. For available



(a) Subthreshold: $V_{DDH} = 0.24V$ and $V_{DDL} = 0.18V$.



(b) Nominal: $V_{DDH} = 1.2V$ and $V_{DDL} = 0.60V$.

Figure 6.3: Slack time distribution before and after optimization of slack-time based algorithm for c880.

V_{DDL} from power supply domain, our algorithm has linear time complexity $O(n)$ for finding the best energy saving by reducing time of searching for V_{DDL} gates using the thresholds S_l and S_u. MILP recursively recursively V_{DDL} gates from all gates inside the circuit to obtain the best energy saving. Thus, MILP displays an exponential time complexity for some benchmark circuits. Therefore, we can use our algorithm to optimize large circuits for dual V_{dd} design within reasonable time instead of using the exponential complexity MILP.

6.4 Summary

We present here a new slack-time based algorithm for dual-V_{dd} design [33]. Emphasis is on saving computation time and effort for maximizing energy saving in a given circuit. In a dual-V_{dd} design, the given performance for a circuit determines the higher supply voltage V_{DDH}. The method of selecting a lower supply voltage V_{DDL} and the use of positive slack gates are the main ideas presented in this paper. The proposed algorithm classifies all positive slack gates into V_{DDH}, possible V_{DDL}, and V_{DDL} groups, respectively, based on the slack time of gates. After classification, the algorithm only investigates the "possible V_{DDL} gates" for available V_{DDL} considering multiple voltage boundaries in the energy optimization procedure. This reduces the complexity of the energy optimization process and the computation time remains tolerable for large circuits compared to the other available MILP methods. HSPICE simulations for ISCAS'85 benchmark circuits show energy savings up to 22.2% in subthreshold operation and 50.8% in nominal operation, which are the same as were obtained by the higher-complexity MILP method [35]. Computation time is reduced up to 43X compared to MILP. Our proposed algorithm has linear time complexity of $O(n)$ with n being the number of gates in the circuit. This novel slack-time based algorithm is useful because the MILP method is limited by its exponential run time cost.

Chapter 7

Conclusion and Future Work

This chapter provides the summary of our contribution, the conclusions of this work, and some suggestions for the future work.

7.1 Conclusion

With rigid energy budget in energy constrained systems, subthreshold circuit design has become a predominant technique in recent years. The battery life of remote or portable devices may not be affordable to the system demands. In an extreme case, micro-sensor networks may require very little energy consumption to be supplied by electrical energy converted from the ambient energy, such as energy harvesting or energy scavenging. These challenges are solved by designing the systems with respect to a very low supply voltage below V_{th}, but performance penalty still remains for subthreshold circuits. Without the performance requirement, we can focus on minimum energy operation as a primary goal. On the other hand, some energy efficient systems have a wide range of speed requirements, therefore the operation of systems may occur at a non-minimum energy point. The contribution of this dissertation utilizes the time slack using dual-V_{dd} to further lower energy budget for energy constrained systems that have speed requirement or not. Using dual voltage design for subthreshold circuits, minimum energy is always less than the absolute minimum energy point for single voltage design when the system does not require a certain speed. Alternatively, using dual-V_{dd} the energy constrained systems can operate several times faster than single-V_{dd} operation without increasing its energy consumption.

We proposed the MILP algorithm of dual voltage design for minimum energy design without level converting devices in Chapter 3. The MILP determines globally the energy

optimized circuit by assigning an extra supply voltage V_{DDL} to gates on non-critical paths. The topological constraints eliminate lever converters that have unacceptable delay overhead in subthreshold regimes.

In Chapter 4, we proposed another MILP algorithm for subthreshold circuits using dual subthreshold supplies in which level converters are eliminated and special multiple logic-level gates are used instead. The MILP optimally substitutes multiple logic-level gates into V_{DDH} gates at the places where V_{DDL} gates feed into V_{DDH} gates considering the benefit for energy saving. From eliminating topological constraints by multiple logic-level gates, this MILP improves energy saving up to 15.7% for ISCAS'85 benchmark circuits compared to the previous proposed MILP.

We investigated validation of dual-V_{dd} design for subthreshold circuits with process variation and technology scaling in Chapter 5. Subthreshold circuits are susceptible to V_{th} variation that exponentially affects delay. A subthreshold circuit using dual-V_{dd} is more immune to the delay variation induced by V_{th} variation, where worst delay variability is reduced by lower gate capacitance of V_{DDL} gates as load capacitance for V_{DDH} gates on critical paths. Technology trends with smaller feature size improve the speed of subthreshold circuits, but energy saving is not solely affected by technology choice. Only the leakage energy saving component in total energy saving is dependant on technology parameters, the ratio of DIBL coefficient η and subthreshold swing S. These two parameters simultaneously increase with technology scaling, thus total energy saving eventually remains quite similar. The amount of time slack inside a circuit determines dominantly total energy saving.

Applying the proposed framework for dual-V_{dd} techniques to subthreshold circuits, we can extend the eligibility of subthreshold circuit design to more energy constrained applications in future markets.

In Chapter 6, we proposed a linear-time algorithm for dual-V_{dd} design using gate slack. For an n-gate circuit, previous heuristic algorithms have theoretical time complexity $O(n^2)$ to utilize time slack for low power consumption, where static timing analysis takes $O(n)$ time to

check timing violations for each gate. Using two slack times, upper slack time (S_u) and lower slack time (S_l), we can unconditionally classify all gates into three groups, V_{DDL}, possible V_{DDL}, and V_{DDH} groups, for dual-V_{dd} techniques. The optimization procedure only makes an effort to search V_{DDL} gates in the possible V_{DDL} group for minimum power or energy. By reducing the search space, the time of optimization is drastically reduced for modern VLSI circuits. We compared our slack-time based algorithm and the proposed MILP in Chapter 3 for computation run-time and energy saving. The computation run-time for our algorithm using gate slack is up to 43 times faster than the MILP for ISCAS'85 benchmark circuits. Also, the energy saving from our algorithm is close to the global optimal solution from MILP. The method of gate slack analysis can be applicable for low power design that utilizes positive slack time inside a circuit.

7.2 Future Work

7.2.1 Minimum Energy Design with Process Variations Using Dual-V_{dd}

In the proposed MILP algorithms, we do not take into account process variations. As mentioned before, subthreshold circuits are highly sensitive to V_{th} variation. The gate delay and leakage current exponentially depend on V_{th} in the subthreshold region. The proposed MILP algorithms utilize positive time slack based on the deterministic gate delay using dual-V_{dd} and find a minimum energy point considering the deterministic leakage energy. If we consider process variations, the gate delay and leakage current should be characterized statistically during the optimization process. The MILP with process variations will give more reliable global solutions for minimum energy design in newer CMOS technologies with smaller feature sizes.

7.2.2 Level Converter for Multi-V_{dd} Design in Subthreshold Regime

Present level converters in industrial standard cell libraries do not show suitable choice for multi-V_{dd} design in the subthreshold region. The main problem with these level shifting

devices is not the output voltage level for logic high "one", but unacceptable performance overhead compared to the delay overhead in nominal operation. This huge delay of level converters prevents inserting them on positive slack paths for efficient energy saving. Without proper level converters for subthreshold design, we introduced topological constraints or multiple logic-level gates to remove use of level converts in our work. In chip design industries, standard cell libraries are well characterized with a clean and fast input that goes fully rail to rail [31]. Without proper level converter cells, signals may experience significant rise and fall time degradation between the driver and receiver cells in different voltage domains. These cause timing closure problems in chip design procedures. To solve these problems, new level converter cells should be designed for subthreshold circuit blocks in multi-V_{dd} domains.

7.2.3 A New Hybrid (MILP + Gate Slack Analysis) Linear-Time Algorithm for Low Power Design Using Multi-V_{dd}

The proposed slack-time based algorithm in Chapter 7 has linear-time complexity $O(n)$ to optimize a given n-gate circuit for minimum energy. This algorithm uses gate slack analysis to group all gates into three groups in a simple and fast way for dual-V_{dd} design and finds the best solution close to the global optimum. But, gates in a possible V_{DDL} group are tested and then assigned to V_{DDL} based on the heuristic priority chosen by higher logic depth for CVS structure. V_{DDH} gates always feed into V_{DDL} gates in CVS, thus the heuristic algorithm is very simple and straightforward for implementation. For the ECVS structure, we should consider the power and delay overheads of level converters during the optimization process for low power. Heuristic algorithms may not be affordable to find nearly global optimal solutions. MILP algorithm always guarantees the global optimum for low power, but can not handle very large circuits due to exponential run-time. If we reduce the optimization space using gate slack analysis, MILP can drastically reduce its exponential run-time and find the global optimal solution. Using both benefits from MILP and gate slack analysis, we can efficiently and accurately solve the optimization problem for multi-V_{dd} design.

Bibliography

[1] M. Anis, S. Areibi, M. Mahmoud, and M. Elmasry, "Dynamic and Leakage Power Reduction in MTCMOS Circuits using an Automated Efficient Gate Clustering Technique," in *Proceedings of 39th Design Automation Conference*, 2002, pp. 480–485.

[2] A. Asenov, A. Brown, J. Davies, S. Kaya, and G. Slavcheva, "Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs," *IEEE Trans. on Electron Devices*, vol. 50, no. 9, pp. 1837–1852, Sept. 2003.

[3] D. Bol, R. Ambroise, D. Flandre, and J.-D. Legat, "Interests and Limitations of Technology Scaling for Subthreshold Logic," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 10, pp. 1508–1519, Oct. 2009.

[4] D. Bol, D. Flandre, and J.-D. Legat, "Technology Flavor Selection and Adaptive Techniques for Timing-Constrained 45nm Subthreshold Circuits," in *Proceedings of 14th ACM/IEEE International Symposium on Low Power Electronics and Design*, 2009, pp. 21–26.

[5] D. Bol, D. Kamel, D. Flandre, and J. D. Legat, "Nanometer MOSFET Effects on the Minimum-Energy Point of 45nm Subthreshold Logic," in *Proceedings of 14th International Symposium on Low Power Electronics and Design*, 2009, pp. 3–8.

[6] A. Bryant, J. Brown, P. Cottrell, M. Ketchen, J. Ellis-Monaghan, and E. J. Nowak, "Low-Power CMOS at Vdd = 4kT/q," in *Proceedings of Device Research Conference*, 2001, pp. 22–23.

[7] B. H. Calhoun and A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," in *Proceedings of International Symposium on Low Power Electronics and Design*, 2004, pp. 90–95.

[8] B. H. Calhoun and A. P. Chandrakasan, "Ultra-Dynamic Voltage Scaling (UDVS) Using Sub-Threshold Operation and Local Voltage Dithering," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 238–245, 2006.

[9] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 9, pp. 1778–1786, Sept. 2005.

[10] B. H. Calhoun, A. Wang, N. Verma, and A. Chandrakasan, "Sub-Threshold Design: The Challenges of Minimizing Circuit Energy," in *Proceedings of International Symposium on Low power Electronics and Design*, Oct. 2006, pp. 366–368.

[11] C. Chen, A. Srivastava, and M. Sarrafzadeh, "On Gate Level Power Optimization Using Dual-Supply Voltages," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 5, pp. 616–629, 2001.

[12] J. C. Chi, H. H. Lee, S. H. Tsai, and M. C. Chi, "Gate Level Multiple Supply Voltage Assignment Algorithm for Power Optimization Under Timing Constraint," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 6, pp. 637–648, 2007.

[13] D. Chinnery and K. Keutzer, *Closing the Power Gap Between ASIC & Custom: Tools and Techniques for Low Power Design*. Springer, 2007.

[14] D. G. Chinnery and K. Keutzer, "Closing the Gap Between ASIC and Custom: An ASIC Perspective," in *Proceedings of 37th Design Automation Conference*, 2000, pp. 637–642.

[15] D. G. Chinnery and K. Keutzer, "Linear Programming for Sizing, Vth and Vdd Assignment," in *Proceedings of International Symposium on Low power Electronics and Design*, 2005, pp. 149–154.

[16] R. Corless, G. Gonnet, D. Hare, D. Jeifrey, and D. Knuth, "On the Lambert W Function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[17] A. U. Diril, Y. S. Dhillon, A. Chatterjee, and A. D. Singh, "Level-Shifter Free Design of Low Power Dual Supply Voltage CMOS Circuits Using Dual Threshold Voltages," *IEEE Trans. on VLSI Systems*, vol. 13, no. 9, pp. 1103–1107, Sept. 2005.

[18] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits," *Proceedings of IEEE*, vol. 98, no. 2, pp. 253–266, Feb. 2010.

[19] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Mathematical Programming Language*. Brooks/Cole-Thomson Learning, 2003.

[20] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Transistor Variability Modeling and its Validation With Ring-Oscillation Frequencies for Body-Biased Subthreshold Circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 7, pp. 1118–1129, jul 2010.

[21] R. Graybill and R. Melhem, *Power Aware Computing*. Kluwer Academic/Plenum Publishers, 2002.

[22] M. Hamada, Y. Ootaguro, and T. Kuroda, "Utilizing Surplus Timing for Power Reduction," in *Proceedings of IEEE Conference on Custom Integrated Circuits*, 2001, pp. 89–92.

[23] S. Hanson, M. Seok, D. Sylvester, and D. Blaauw, "Nanometer Device Scaling in Subthreshold Logic and SRAM," *IEEE Trans. on Electron Devices*, vol. 55, no. 1, pp. 175–185, 2008.

[24] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, "Exploring Variability and Performance in a Sub-200-mV Processor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 881–891, Apr. 2008.

[25] R. B. Hitchcock, Sr., "Timing Verification and the Timing Analysis Program," in *Proceedings of 19th Design Automation Conference*, 1982, pp. 594–604.

[26] http://www device.eecs.berkeley.edu. BSIM4.6.1 MOSFET Model.

[27] http://www.synopsys.com. HSPICE User Guide: Simulation and Analysis.

[28] F. Ishihara, F. Sheikh, and B. Nikolic, "Level Conversion for Dual-Supply Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 2, pp. 185–195, feb 2004.

[29] M. Jamal Deen, M. H. Kazemeini, and S. Naseh, "Ultra-Low Power VCOs - Performance Characteristics and Modeling (invited)," in *Proceedings of the Fourth IEEE International Caracas Conference on Devices, Circuits and Systems*, 2002, pp. C033–1–C033–8.

[30] M. R. Kakoee, A. Sathanur, A. Pullini, J. Huisken, and L. Benini, "Automatic Synthesis of Near-Threshold Circuits with Fine-Grained Performance Tunability," in *Proceedings of International Symposium on Low Power Electronics and Design*, aug 2010, pp. 401–406.

[31] M. Keating, D. Flynn, R. Aitken, A. Gibbons, and K. Shi, *Low Power Methodology Manual for System-on-Chip Design*. Springer, 2007.

[32] C. H. I. Kim, H. Soeleman, and K. Roy, "Ultra-Low-Power DLMS Adaptive Filter for Hearing Aid Applications," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 6, pp. 1058–1067, 2003.

[33] K. Kim and V. D. Agrawal, "Dual Voltage Design for Minimum Energy Using Gate Slack," in *Proceedings of IEEE International Conference on Industrial Technology & 43rd IEEE Southeastern Symposium on System Theory*, Mar. 2011, pp. 405–410.

[34] K. Kim and V. D. Agrawal, "Minimum Energy CMOS Design with Dual Subthreshold Supply and Multiple Logic-Level Gates," in *Proceedings of 12th International Symposium on Quality Electronic Design*, Mar. 2011, pp. 689–694.

[35] K. Kim and V. D. Agrawal, "True Minimum Energy Design Using Dual Below-Threshold Supply Voltages," in *Proceedings of 24th International Conference on VLSI Design*, Jan. 2011.

[36] M. Kulkarni, "A Reduced Constraint Set Linear Program for Low-Power Design of Digital Circuits," Master's thesis, Auburn University, Dept. of ECE, Auburn, Alabama, Dec. 2010.

[37] M. Kulkarni and V. D. Agrawal, "A Tutorial on Battery Simulation - Matching Power Source to Electronic System," in *Proceedings of 14th IEEE VLSI Design and Test Symposium*, July 2010.

[38] M. Kulkarni and V. D. Agrawal, "Energy Source Lifetime Optimization for a Digital System through Power Management," in *Proceedings of 43rd IEEE Southeastern Symposium on System Theory*, Mar. 2011, pp. 75–80.

[39] S. H. Kulkarni, A. N. Srivastava, and D. Sylvester, "A New Algorithm for Improved VDD Assignment in Low Power Dual VDD Systems," in *Proceedings of International Symposium on Low Power Electronics and Design*, 2004, pp. 200–205.

[40] S. H. Kulkarni and D. Sylvester, "High Performance Level Conversion for Dual V_{DD} Design," *IEEE Transactions on VLSI Systems*, vol. 12, no. 9, pp. 926–936, 2004.

[41] V. Kursun and E. G. Friedman, *Multi-Voltage CMOS Circuit Design*. Wiley, 2006.

[42] J. Kwong and A. Chandrakasan, "Variation-Driven Device Sizing for Minimum Energy Subthreshold Circuits," in *Proceedings of International Symposium on Low Power Electronics and Design*, Oct. 2006, pp. 8–13.

[43] J. Kwong, Y. K. Ramadass, N. Verma, and A. P. Chandrakasan, "A 65 nm Sub-V_t Microcontroller With Integrated SRAM and Switched Capacitor DC-DC Converter," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 115–126, Jan. 2009.

[44] R. Lyon and C. Mead, "An Analog Electronic Cochlea," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 7, pp. 1119–1134, July 1988.

[45] A. K. Majhi, V. D. Agrawal, J. Jacob, and L. M. Patnaik, "Line Coverage of Path Delay Faults," *IEEE Trans. VLSI Systems*, vol. 8, pp. 610–614, Oct. 2000.

[46] A. K. Majhi, J. Jacob, L. M. Patnaik, and V. D. Agrawal, "On Test Coverage of Path Delay Faults," in *Proceedings of 9th International Conference on VLSI Design*, Jan. 1996.

[47] D. Markovic, C. Wang, L. Alarcon, T.-T. Liu, and J. Rabaey, "Ultralow-Power Design in Near-Threshold Region," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 237–252, feb 2010.

[48] J. Meindl and J. Davis, "The Fundamental Limit on Binary Switching Energy for Terascale Integration (TSI)," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 10, pp. 1515–1516, 2000.

[49] J. D. Meindl and R. N. Swanson, "Potential Improvements in Power-Speed Performance of Digital Circuits," *Proceedings of the IEEE*, vol. 59, no. 5, pp. 815–816, May 1971.

[50] D. Nguyen, A. Davare, M. Orshansky, D. Chinnery, B. Thompson, and K. Keutzer, "Minimization of Dynamic and Static Power Through Joint Assignment of Threshold Voltages and Sizing Optimization," in *Proceedings of International Symposium on Low Power Electronics and Design*, 2003, pp. 158–163.

[51] K. Nose and T. Sakurai, "Optimization of VDD and VTH for Low-Power and High-Speed Applications," in *Proceedings of ACM/IEEE Design Automation Conference*, 2000, pp. 469–474.

[52] G. Ono and M. Miyazaki, "Threshold-Voltage Balance for Minimum Supply Operation," in *Symposium on VLSI Circuits Digest of Technical Papers*, 2002, pp. 206–209.

[53] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching Properties of MOS Transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

[54] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits*. Prentice-Hall, second edition, 2003.

[55] T. Raja, "A Reduced Constraint Set Linear Program for Low-Power Design of Digital Circuits," Master's thesis, Rutgers University, Dept. of ECE, New Brunswick, New Jersey, Mar. 2002.

[56] T. Raja, V. D. Agrawal, and M. L. Bushnell, "Minimum Dynamic Power CMOS Circuit Design by a Reduced Constraint Set Linear Program," in *Proceedings of 16th International Conference on VLSI Design*, Jan. 2003, pp. 527–532.

[57] Y. Ramadass and A. Chandrakasan, "Voltage scalable switched capacitor dc-dc converter for ultra-low-power on-chip applications," in *Proceedings of Power Electronics Specialists Conference*, 2007, pp. 2353–2359.

[58] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.

[59] K. Roy, L. Wei, and Z. Chen, "Multiple-Vdd Multiple-Vth CMOS (MVCMOS) for Low Power Applications," in *Proceedings of IEEE Int. Symposium on on Circuits and Systems*, 1999, pp. 366–370.

[60] T. Sakurai and A. Newton, "Alpha-Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.

[61] G. Schrom and S. Selberherr, "Ultra-Low-Power CMOS Technologies," in *Proceedings of International Semiconductor Conference*, volume 1, Oct. 1996, pp. 237–246 vol.1.

[62] M. Seok, S. Hanson, Y. S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "The Phoenix Processor: a 30pW Platform for Sensor Applications," in *Proceedings of IEEE Symposium on VLSI Circuits*, 2008, pp. 188–189.

[63] M. Seok, D. Sylvester, and D. Blaauw, "Optimal Technology Selection for Minimizing Energy and Variability in Low Voltage Applications," in *Proceedings of 13th International Symposium on on Low Power Electronics and Design*, 2008, pp. 9–14.

[64] A. Srivastava, D. Sylvester, and D. Blaauw, "Power Minimization using Simultaneous Gate Sizing, Dual-Vdd and Dual-Vth Assignment," in *Proceedings of 41st Design Automation Conference*, 2004, pp. 783–787.

[65] V. Sundararajan and K. K. Parhi, "Synthesis of Low Power CMOS VLSI Circuits Using Dual Supply Voltages," in *Proceedings of 36th Design Automation Conference*, 1999, pp. 72–75.

[66] R. Swanson and J. Meindl, "Ion-Implanted Complementary MOS Transistors in Low-Voltage Circuits," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 1972, pp. 192–193.

[67] K. Usami and M. Horowitz, "Clustered Voltage Scaling Technique for Low-Power Design," in *Proceedings of International Symposium on Low Power Design*, 1995, pp. 3–8.

[68] K. Usami, M. Igarashi, F. Minami, T. Ishikawa, M. Kanzawa, M. Ichida, and K. Nogami, "Automated Low-Power Technique Exploiting Multiple Supply Voltages Applied to a Media Processor," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 3, pp. 463–472, 1998.

[69] R. Vaddi, S. Dasgupta, and R. P. Agarwal, "Device and Circuit Design Challenges in the Digital Subthreshold Region for Ultralow-Power Applications," *VLSI Design*, vol. 2009, pp. 1–14, Jan. 2009.

[70] M. Venkatasubramanian, "Energy Effciency and Process Variation Tolerance of 45 nm Bulk and High-k CMOS Devices," Master's thesis, Auburn University, Dept. of ECE, Auburn, Alabama, May 2011.

[71] M. Venkatasubramanian and V. D. Agrawal, "Subthreshold Voltage High-k CMOS Devices Have Lowest Energy and High Process Tolerance," in *Proceedings of 43rd IEEE Southeastern Symposium on System Theory*, Mar. 2011, pp. 100–105.

[72] N. Verma, J. Kwong, and A. Chandrakasan, "Nanometer MOSFET Variation in Minimum Energy Subthreshold Circuits," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 163–174, jan 2008.

[73] E. Vittoz and J. Fellrath, "CMOS Analog Integrated Circuits Based on Weak Inversion Operations," *IEEE Journal of Solid-State Circuits*, vol. 12, no. 3, pp. 224–231, June 1977.

[74] E. Vittoz, B. Gerber, and F. Leuenberger, "Silicon-Gate CMOS Frequency Divider for Electronic Wrist Watch," *IEEE Journal of Solid-State Circuits*, vol. 7, no. 2, pp. 100–104, Apr. 1972.

[75] E. A. Vittoz, "The Electronic Watch and Low-Power Circuits," *IEEE Solid-State Circuits Newsletter*, vol. 13, no. 3, pp. 7–23, 2008.

[76] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, *Sub-Threshold Design for Ultra Low-Power Systems.* Springer, 2006.

[77] A. Wang and A. Chandrakasan, "A 180mV FFT Processor Using Subthreshold Circuit Techniques," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2004, pp. 292–529.

[78] A. Wang, A. P. Chandrakasan, and S. V. Kosonocky, "Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits," in *IEEE Computer Society Annual Symposium on VLSI*, 2002, pp. 5–9.

[79] L. Wei, K. Roy, and C.-K. Koh, "Power Minimization by Simultaneous Dual Vth Assignment and Gate-Sizing," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, 2000, pp. 413–416.

[80] N. H. E. Weste and D. M. Harris, *CMOS VLSI Design.* Boston: Addison-Wesley, fourth edition, 2009.

[81] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," in *Proceedings of 41st Design Automation Conference*, 2004, pp. 868–873.

[82] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proceedings of International Symposium on Low Power Electronics and Design*, Aug. 2005, pp. 20–25.

[83] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "A Variation-Tolerant Sub-200 mV 6-T Subthreshold SRAM," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 10, pp. 2338–2348, Oct. 2008.

[84] B. Zhai, S. Pant, L. Nazhandali, S. Hanson, J. Olson, A. Reeves, M. Minuth, R. Helfand, T. Austin, D. Sylvester, and D. Blaauw, "Energy-Efficient Subthreshold Processor Design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 8, pp. 1127–1137, aug 2009.

[85] W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Early Design Exploration," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2816–2823, 2006.