The Thesis Committee for Mario Chichun Lok
Certifies that this is the approved version of the following thesis:

Process Variation Aware Low Power Buffer Design

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor: _____
Michael Orshansky

Reader: _____
Mark Mcdermott

Process Variation Aware Low Power Buffer Design

by

Mario Chichun Lok

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Masters of Science in Engineering

The University of Texas at Austin

May 2010

Acknowledgements

These two years at the University of Texas has been a wonderful journey for me, and my experience at UT would not be the same without the help of these people to whom I would like to pay tribute.

First, I thank my research advisor, Professor Michael Orshansky, for being such a great advisor and a terrific mentor. He taught me to precisely define the scope, clearly present the key concepts and critically review my results in a research project. He has been very supportive during the ups and downs of my research. Working with him, I have learnt a lot about the bits and pieces of being an independent researcher.

I would also like to thank Professor Mark McDermott for reading my thesis and for extending his help beyond the coursework setting on many occasions. I am grateful to Professor Adan Aziz, Professor Arjang Hassibi and Professor Derek Chiou for their inspiring lectures and great advice.

As a member of the Robust IC Design group, I have had the opportunity to work with many great minds. Murari Mani, Ashish Singh, Ku He, and Kareem Kragab each helped me with different aspects of my research at UT. I really appreciate the numerous interesting discussions with them on technical and non-technical subjects. Special thanks to Murari and Ku are in order for their early contribution to this work and their guidance throughout this project.

I was fortunate to be surrounded by really good friends during my time here. Yi Yuan, Arnab Dutta, Yilin Zhan, Eric Chu helped me adjust to a new environment at Austin and I thank them for their friendship. I will definitely remember them after I leave UT.

Words cannot express my gratitude towards Bonnie Lam, my girlfriend. I want to thank her for always being there for me through difficult times of my life. She has always been really encouraging and has had so much confidence in me. I thank her for being a great friend and companion through my years of college and graduate studies.

Finally, I thank my father and mother for their unconditional love and support. I really appreciate their encouragement and support for me to pursue a life-long dream. I thank them also for bringing me to North America, as they provided me the first of many opportunities that would allow me to become an engineer practicing at the center of innovation. As I complete my thesis today, I hope to give my best wishes to my father for returning to good health soon.

May 7th, 2010

Abstract

Process Variation Aware Low Power Buffer Design

Mario Chichun Lok, MSE

The University of Texas at Austin, 2010

Supervisor: Michael Orshansky

In many digital designs there is a need to use multi-stage tapered buffers to drive large capacitive loads. These buffers contribute a significant percentage of overall power. In this thesis, we propose two novel tunable buffer designs that enable reduction in power in the presence of process variation. A strategy to derive the optimal buffer size and the optimal tuning rule in post-silicon phase is developed. By comparing several tunable buffer circuit topologies, we also demonstrate the tradeoffs in tunable buffer topology selection as a function of switching activity, timing requirements, and the magnitude of process variations. Using HSPICE simulations based on the high performance 32nm ASU Predictive Model, we show that up to 30% average power reduction can be achieved for a SRAM word-line decoder while maintaining the same timing yield.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 INTRODUCTION

1.1 Motivation

Large capacitive loads are ubiquitous in CMOS integrated circuits. Typically, tapered buffers are designed to drive these large capacitances while ensuring that the load placed on previous stages of the signal path is not too large [1]. Buffers are used in the memory access path as word-line drivers [2], to drive large off-chip capacitances in I/O circuits [3], and in clock trees to ensure that skew constraints are satisfied [4]. Moreover, the recent trend of exacerbating wire delays necessitates the insertion of more buffers per unit length of global interconnect to meet delay targets [5]. Aggressive deployment of buffers in high-performance microprocessors means that they now account for a significant portion of total power consumption of the chip.

The growing need for power efficiency in mobile and portable devices, in conjunction with the increase in leakage power with scaling, has inspired the development of techniques for low-power buffer design [6][7]. With the rise of variability, several post-silicon techniques have been proposed to reduce parametric yield loss due to variability, more generally for statistical power optimization [8][9], and in particular for buffer design [10][11]. The fundamental limitation of design-time methods is that they impose an overhead on each instance of the fabricated chip since they intrinsically lack the ability to react to the actual conditions on the chip. An alternate paradigm to design-time optimization is post-silicon adaptivity, which allows the designer to tune chips individually to help meet performance constraints. A number of techniques have been shown to be effective, including adaptive body biasing [12][13],

adaptive supply voltage [14], and adaptive sizing of keepers in dynamic circuits [15]. Recently, methods based on tuning of clock buffers have been proposed [16] to reduce parametric yield loss.

One specific class of tunable buffer chains explored in [17] is to use the capability of switching between high-speed and low-power configurations to exploit their energy-delay tradeoffs. Compared to adaptive body biasing and supply voltage scaling, tuning of the buffer chains in [17] can be implemented with a pure digital-design flow. Since neither a voltage regulator nor a voltage reference is needed, such tunable buffers enable post-manufacture tuning at a much smaller granularity. However, the circuit implementation of buffer chain in [17] has a significant area overhead and exhibits large leakage power. In addition, the strategies used for design time buffer sizing and run time tuning do not take into account the magnitude and characteristics of process variability.

In this work, we present several improvements on the design and tuning of this class of tunable buffer with high-speed and low-power configurations. Two new implementations are proposed to reduce area overhead and conserve leakage power. In addition, we use the framework of adaptable optimization to develop a process-variation-aware strategy to size these tunable buffers for minimum power consumption. Moreover, the two new implementations are compared against the buffer chain in [17] for different system specifications to analyze the pros and cons of each implementation. Finally, the design of a buffer chain is shown in the context of a 64kb SRAM decoder.

1.2 Organization

Section 2 first introduces the basic principle of designing a buffer chain. Then it reviews the extension of the basic principle to account for process variability. It further

compares the strategy that is discussed in this work with other known low power design approaches. The section concludes with a comparison between the supply voltage scalable buffer and a two-configuration sizing-tunable buffer in low power consumption.

Section 3 describes three possible implementations for a tunable buffer with two size configurations. Then it summarizes the power and area overhead to implement the tuning ability.

Section 4 illustrates a two-stage optimization algorithm that would be able to size the tunable buffer for minimum power consumption given the timing requirement and other specifications.

Section 5 compares the three implementations that are described in section 3 and demonstrated the design flow of a tunable buffer using a world-line driver of a 64kb SRAM as an example.

CHAPTER 2 BUFFER DESIGN PRINCIPLES

2.1 Power and Delay

Buffer chains are simple but indispensable circuits required for driving nodes with large capacitances. Because of their size, they consume larger area and more power than typical logic gates. Consider an *N* stage buffer driving capacitive load C_L. By the theory of logical effort [1], the delay is given by

$$D = t_{p0}\sum_{j=1}^{N}(1+\frac{w_{j+1}}{\xi w_j});\ w_{N+1} = C_L \tag{2.1}$$

Where w_j is the sizing factor for the inverter at stage j, defined as the ratio of its input capacitance to the input capacitance of the minimum size inverter C_m. The intrinsic delay of the minimum size inverter is denoted by t_{p0} and ξ is a proportionality factor dependent on technology. Given a capacitive load C_L, the challenge in buffer design is to find the values of the sizing factors such that the path delay requirement is met while a certain objective function is minimized. This objective function can be power consumption of the buffer chain or its energy-delay product [28][29]. Additionally, the taper factor (or fan-out ratio) w_{j+1}/w_j can be kept the same for all stages, or variable taper factors can be used [18]. When variable taper factors are allowed, a more accurate delay equation, which account for the slew rate at each stage, is given by

$$D = \sum_{j=1}^{N}(t_{p0}(1+\frac{w_{j+1}}{\xi w_j}) + c_3\tau_j);\ w_{N+1} = C_L \tag{2.2}$$

$$\tau_{j+1} = c_2\tau_j + c_1\frac{w_{j+1}}{w_j} + c_0 \tag{2.3}$$

The parameter τ_j is the output slew rate of j-th stage. It can be computed as a function of the input slew rate and the taper factor of that stage as shown by equation (2.3). The symbols $c_0, c_1, c_2, \cdots c_N$ represent proportionality constants that are technology-specific.

The power dissipation for a buffer chain consists of dynamic power, leakage power and short-circuits power, denoted by P_{dyn}, P_{leak} and P_{short} respectively. The total power and each of three components can be computed as

$$P = P_{dyn} + P_{leak} + P_{short} \tag{2.4}$$

$$P_{dyn} = \tfrac{1}{2}\alpha C_m V_{dd}^2 f(\sum_{j=1}^{N}(1+\xi)w_j + w_{N+1}) \tag{2.5}$$

$$P_{leak} = (1-\alpha)P_o \sum_{j=1}^{N} w_j \exp(c_4 L + c_5 V_{th} + c_6 V_{dd}) \tag{2.6}$$

$$P_{short} = \alpha P(\tau) \tag{2.7}$$

The switching activity factor α is a measure of the portion of the overall time that the circuit is switching. It specifies percentage of contribution to total power from the three components [1]. More specifically, when calculating the average power, dynamic power and short-circuit power is scaled by α and the leakage power is scaled by $(1-\alpha)$. The dynamic power has a square dependency on supply voltage V_{dd}, and a linear dependency on frequency f and transistor sizes. The leakage power has a linear dependency on transistor sizes, and scales exponentially with V_{dd}, threshold voltage V_{th} and transistor gate length L. A buffer dissipates short-circuit power when both the pull-up and the pull-down network are turned on at the same time. For a regular static CMOS buffer, the time during which the pull-up and pull-down network are both on is very short unless the output slew rate is much smaller than the input slew rate. Thus, the short-circuit power is a function of slew rate at each stage.

In summary, the variables in the delay and power equations can be categorized in to three groups: specifications, technology-dependent parameters and design parameters, as shown in Table 1.

Table 1. Buffer design parameters

specifications	α, C_L, f
technology-dependent parameters	$V_{Th}, V_{dd}, L, t_{p0}, C_m, a_5, a_4, a_3, a_2, a_1, a_0, \xi$
design parameters	w_j

2.2 Design for Variability

In the presence of variability, the intrinsic delay t_{p0} and capacitance of a unit-size inverter C_m are no longer constants for a given technology, but process dependent variables. Variations in manufacturing process, supply voltage and temperature can all affect the delay and power of a buffer chain. In this work, opportunities to compensate for process variability in post-silicon phase using circuit techniques are explored. Among different process variation parameters, we focus on the two major sources that heavily influence buffer design, L and V_{th}. The variation in L and V_{TH} each has intra-chip and inter-chip components. In the view of a buffer chain instance or a bank of buffer chains, the intra-chip variation also can be sub-divided into systematic variation and random variation. For the purpose of buffer design analysis, we can group the intra-chip systematic variation and inter-chip variation together as global variation while regarding the random component of intra-chip variation a local parameter. By [26] and [33], we can treat the variation in L as a global parameter and the variation in V_{TH} is as a local variant. We further assume that the variation in gate length and threshold voltage both

follow a normal distribution, denoted by $L \sim N(\mu_L, \sigma_L^2)$ and $V_{TH} \sim N(\mu_{Vth}, \sigma_V^2)$. Table 2 summarizes the modified list of buffer design parameters in the presence of process variation.

Table 2. Modified buffer design parameters

specifications	$\alpha,\ C_L, f$
technology-dependent parameters	$\mu_L,\ \sigma_L,\ \mu_{Vth},\ \sigma_{Vth},\ V_{dd},\ c_5,\ c_{4,}\ c_3,\ c_{2,}\ c_1,\ c_0,\ \xi$
process-dependent parameters	$V_{Th},\ L,\ t_{p0},\ C_m$
design parameters	w_j

The traditional approach for designing a buffer chain in the presence of process variability is using worst-case. First, the joint distribution L and V_{th} as well as the yield constraint for the design are defined. Then worst case corners points are chosen from the contour where the joint probability of the enclosed area by the contour is equal the yield requirement. After that, the worst case technology dependent parameters, such as t_{p0} and C_m, are calculated based on the corners. Finally, these technology parameters are used to derive the sizing factors for each buffer stage to meet timing requirements. Using this approach, the buffer chain is sized to be sufficiently large to meet timing requirement even in the slow process corners. However, for most die instances, such a large buffer chain produces a large timing slack and consumes more power than necessary. Moreover, for die instances that are in a fast process corner, the power dissipation of an oversized design may exceed the maximum power consumption specified for any given chip.

2.3 Sample Benchmark

A benchmark is established to evaluate the power consumption for a buffer chain with a propagation delay constraint. The target technology for this work is chosen to be a high-performance 32nm technology. It consists of a set of system specifications and assumptions made about the magnitude of the variation in L and V_{TH}. Table 3 shows a benchmark that will be used consistently throughout this work to compare different buffer implementations.

Table 3. Sample benchmark

Supply Voltage	1.0V	switching factor α	0.2
Timing constraint T	125ps	μ_L	35nm
Frequency f	1GHz	μ_{Vth}	250mV
timing yield γ	0.999	σ_L	2nm
capacitive load C_L	128C_m	σ_V	50mV

In this benchmark, the timing target 125ps is chosen based on the minimum delay achievable by a four-stage buffer with a yield of 0.999. An aggressive timing target is chosen here, as we would like to demonstrate in later sections that the tunable buffer design proposed can also meet the same aggressive delay target at the same yield with lower average power. In addition, a moderate activity factor is chosen because an initial design that is optimized for both dynamic power and static power is desired. Furthermore, the standard distribution of gate length variation is taken to be 2nm, including impacts from both inter-die and intra-die global variation. Finally, the variation in V_{TH} is assumed to have one σ_V value of 50mV for minimum width devices, and it scales with $\sigma_V \propto \frac{1}{\sqrt{WL}}$.

2.4 Low Power Buffer Design Techniques

A post-silicon tunable design allows designers to reduce excess performance for lower power consumption. In the context of buffer design, post silicon tuning can be applied to reduce the excess buffer drive-strength for individual buffers.

Many different approaches have been studied to make the buffer chain tunable in post-silicon phase to adjust the power and performance of a buffer chain. The common techniques are adaptive body biasing [12][13] and adaptive supply voltage scaling [14]. However, as stated in [31], body biasing is becoming less effective in the latest technology as the dependency of threshold voltage on substrate voltage is reduced. In [30], the authors proposed a method to adjust the overdrive voltage of the transistors in a logic gate by skewing both supply voltage and ground node of that gate. Since the effect of adjusting overdrive voltage of a transistor is the same as changing its threshold voltage, this technique can serve as an alternative to body biasing. Another possible optimization is allowing the pMOS and nMOS transistors of a buffer stage to be sized independently. Such a buffer design, named asymmetric-buffer in [25], has been demonstrated to reduce power and area when used for intermediate stages to individually drive the pull-up or pull-down transistors of the last stage. In [34], the adaptive supply voltage scaling technique and the skewed-supply technique are combined to create a novel logic family that is tunable in energy and performance over a wide range, if two independently adjustable supply voltages are available.

The approach of our work is the size-tuning scheme described in [17]. The power consumption, drive strength and the propagation delay are adjustable by varying the size of the buffer chain. In order to keep the area overhead under control the available choice

of buffer size is set to be only two discrete values: a high-speed configuration and a low-power configuration. However, as we will show in later sections, significant power reduction can be achieved despite the simplicity of this scheme. In addition, we develop an algorithm to size the two configurations optimally for average power consumption.

2.5 Supply Scaling Versus Size Tuning

Supply voltage scaling has been behind numerous efforts in low power design. It is an effective approach to reduce power consumption as the dynamic power of a circuit scales with V_{DD}^2. One natural question to ask about the size-tuning strategy is how it compares to scaling the supply voltage. We will try to answer this question by making a comparison with buffer designing in this section.

The initial transistor sizes of the buffer chain are chosen based on the benchmark from section 2.3. In the presence of gate length variation, the performance of a buffer chain degrades with increasing *L*. This buffer chain meets a timing target of 125ps with a yield of 0.999.

If supply voltage scaling is available, the power-delay product of the buffer chain decreases with *L* as shown in Figure 1. At each gate length value, it is possible to make tradeoff between power and propagation delay by setting supply voltage to an appropriate level. Three tradeoff curves of a buffer chain when gate length is at the nominal value or at one σ_L away from its nominal value are plotted in Figure 1. When a delay target is specified, each buffer instance has a different timing slack according to the value of its process parameters. By supply voltage scaling, we can fully absorb the timing slack of any single instance and turn it into power saving.

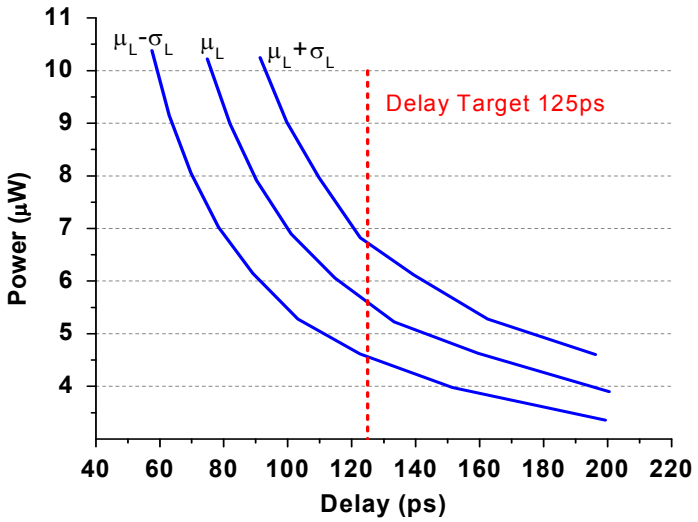


Figure 1. Power-delay tradeoff produced by supply voltage scaling

In Figure 2, the power and delay of a size-tunable buffer at three gate length values is added to the plot. We see that supply-voltage scaling has two main advantages. First, scaling supply-voltage has a much larger tuning range. In addition, size tuning does not have the flexibility of choosing an appropriate size that fully utilizes the timing slack.

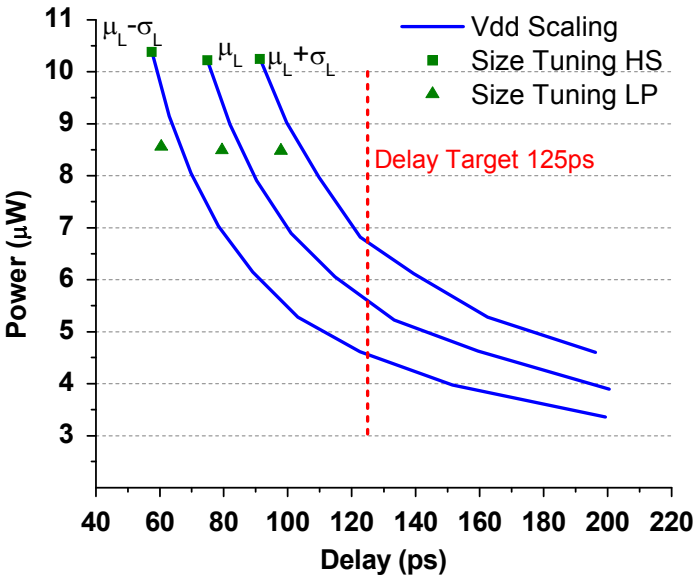


Figure 2. Supply voltage scaling vs. transistor sizing tuning

One way to compare supply-scaling and sizing-tuning under the influence of process variation is by evaluating the statistical average power of either approach. In the presence of gate length variation, the average power can be computed using equation (2.8), where $\rho(L)$ is the probability density function for gate length variation and $P(L)$ is the power consumption of the buffer at each value of L.

$$\mathrm{E}[P] = \int_{-\infty}^{\infty} P(L)\rho(L)dL \qquad (2.8)$$

$P(L)$ for supply scaling can be found by taking the minimum power needed to meet the delay constraint. For example, as shown in Figure 2, for a buffer instance with its gate length equal to μ_L at 35nm, the minimum power needed to meet 125ps delay target is 5.5μW. For the two-configuration size-tuning approach, at each gate length, $P(L)$ is equal to the power consumption of the buffer's low-power configuration if timing constraint is met with the low-power configuration. Otherwise, the power consumption of the high-speed configuration is used for $P(L)$.

The function $P(L)$ for supply-voltage scaling and sizing tuning is plotted in Figure 3. Initially, we assume the voltage conversion necessary for supply voltage scaling is 100% efficient. By using the data in Figure 3 to evaluate equation (2.8), we find the average power achieved by using supply-voltage scaling is 6.02μW. For reasons mentioned above, the average power obtained by employing the sizing tuning is scheme is higher, at 8.2μW. However, in reality, the efficiency of voltage conversion highly depends on the size of the load current. For fine grain power-performance tuning, efficiency of a state-of-the-art voltage regulator, designed for a load current <1mA, is reported to be 70% [27]. Since the load current of a buffer chain is on the order of 10μA, it is reasonable to assume that the voltage scaling for a buffer and buffer bank is 70% efficient. After the efficiency of voltage conversion is taken into account, the average

power achieved for supply scaling is 8.6μW. This result indicates that the size-tuning approach has a more significant power reduction for power-performance tuning of small size blocks because of the moderate efficiency in scaling voltage for small load.

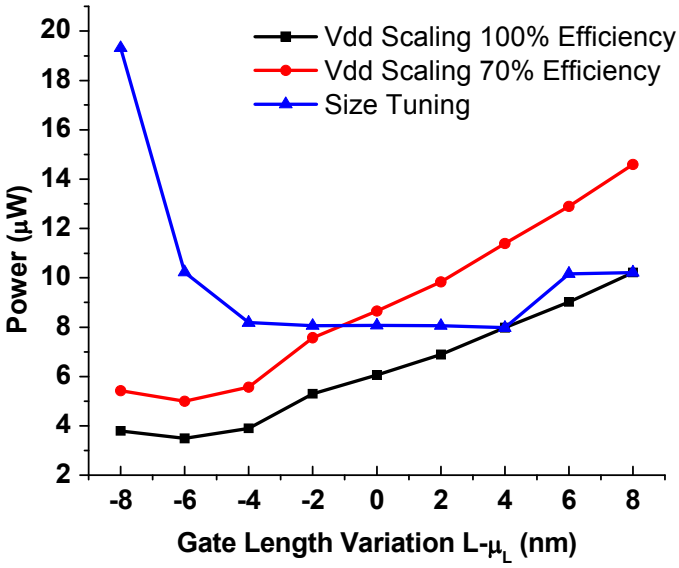


Figure 3. Power consumption with supply voltage and sizing tuning

CHAPTER 3 TUNABLE BUFFER DESIGN

3.1 Existing Implementation

The analysis in section 2.5 demonstrated that the sizing tuning strategy has some advantages over supply voltage scaling in providing fine grain power-performance tuning. The implementation of a size-tunable buffer proposed in [17] is shown in Figure 4, denoted by B1. The ability to adjust the buffer drive-strength is made available by having a control signal to switch the buffer between a high-speed and a low-power configuration.

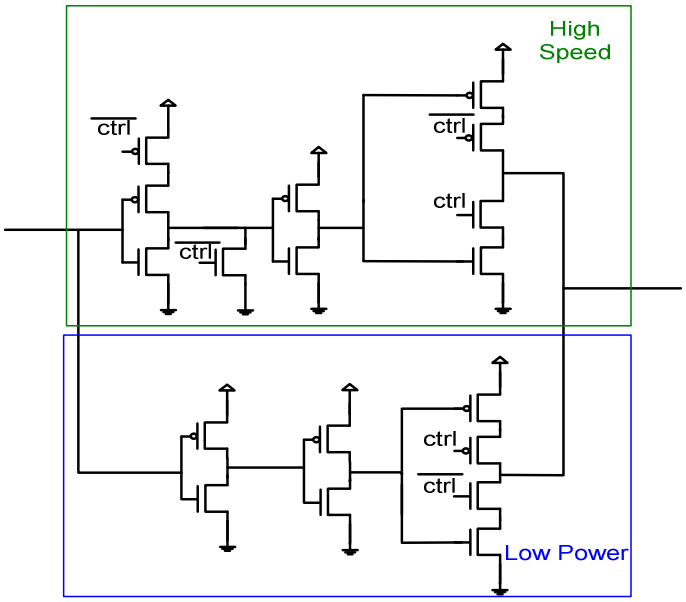


Figure 4. Adaptive buffer design (B1)

This tunable buffer is designed to replace the last three stages of a regular buffer chain and sized to meet the same timing and yield constraints. When the control signal *ctrl* is asserted, the input ripples through the high-speed branch. Then, when *ctrl* is de-

asserted, the high-speed branch is deactivated while the input goes through the low power branch. Since the transistor sizes are different in the high-speed branch and the low-power branch, the drive-strength of the last stage and its power consumption are unequal between the two branches. Depending on realization of process parameters, the appropriate configuration can be selected by setting the signal *ctrl* .

Table 4 compares this tunable buffer chain with a regular non-tunable buffer using the benchmark described in section 2.3. As a sizing strategy is not provided in [17], the transistor sizes of the tunable buffer are derived from a statistical sizing algorithm described in section 3.2. The transistor area consumed by each buffer is normalized to the area of a unit-size inverter. As evidenced by the benchmark, the tunable buffer B1 reduces the average dynamic power by 5.5% and average static power by 5% while imposing an area penalty of 300%.

Table 4. Benchmark regular vs. B1

	Dynamic Power (μW)	Leakage Power (μW)	Transistor Area (normalized)
Regular buffer	10.31	1.91	21.23
Tunable buffer B1	9.25	1.81	82.81

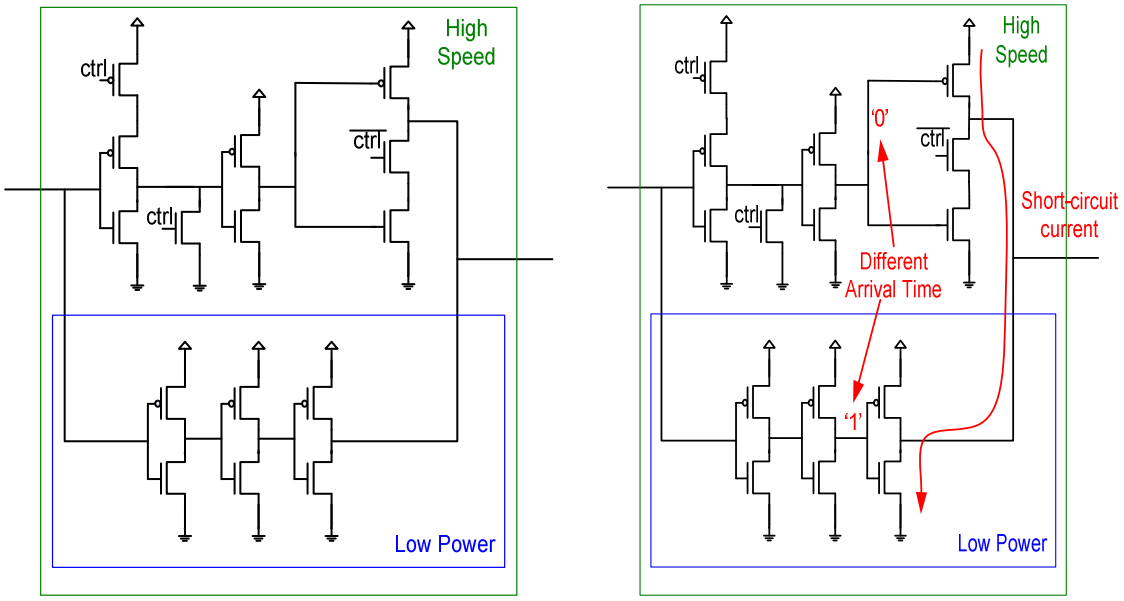
3.2 Alternative Design

An alternative implementation labeled B2 is shown in Figure 2a. It has the ability to switch between a low-power and a high-speed configuration similar to B1. Unlike B1, however, an input signal propagates through both branches when B2 is in the high-speed configuration. The effective drive-strength of the last stage is the combined strength of the two branches. In comparison to B1 under the same benchmark, B2 significantly reduces total transistor area needed and lowers leakage power.

Table 5. Benchmark – design B1 vs. B2

	Dynamic Power and Short- Circuits Power (µW)	Nominal Leakage Power (µW)	Transistor Area (normalized)
Tunable buffer B1	9.25	1.81	82.81
Tunable buffer B2	9.6	0.972	32.66

Nevertheless, the challenge becomes synchronizing the arrival times of the two branches. When the arrival times are different, the last stage of the two branches will create a direct path for DC current from the supply to ground, as shown in Figure 5b). Consequently, the buffer circuit may dissipate a significant amount of short-circuit power.



(a) (b)

Figure 5. Alternative adaptive buffer design(B2)

Design B2 can be made immune to delay offset that is due to global variation by enforcing the fanout ratio at each stage to be the same for both branches. However, due

local variation, the propagation delay through the two branches cannot be made completely correlated. The arrival time difference of the two branches is a function of local variation. Using the sample benchmark, the arrival time difference has zero mean and a standard deviation of 4.3ps. In addition, the active power measured as a function the arrival time difference is shown in Figure 6. As timing offset increases, the short-circuit power is taking a larger percentage of the total power.

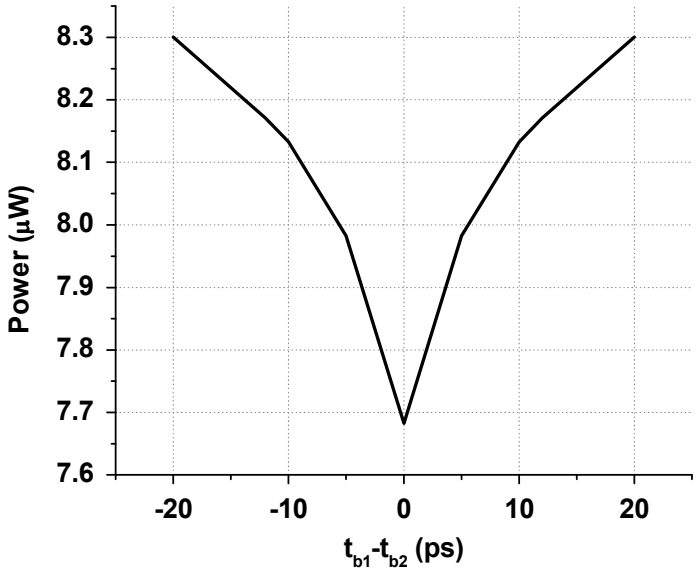


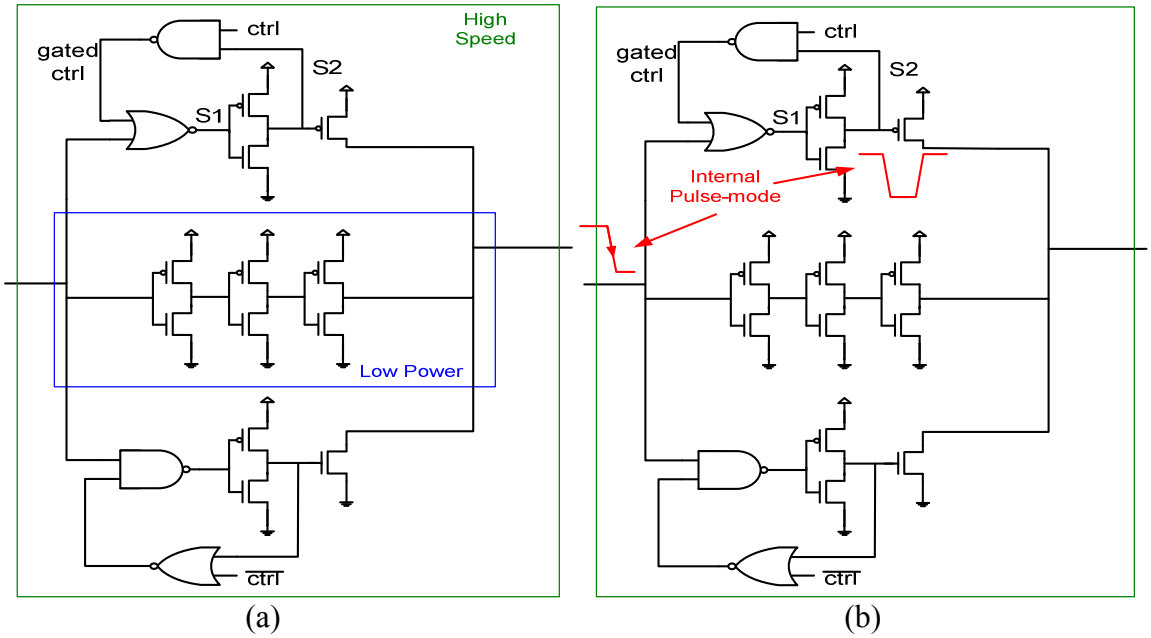
Figure 6. Active power vs. delay offset for B2

A third implementation labeled B3 is shown in Figure 7a. The design B3 consists of a low power branch, and two extra branches—a pull-up branch and a pull-down branch. When the control signal ctrl is asserted, the buffer chain is in the high-speed configuration with the two extra branches enabled. These extra branches in B3 use a pulse-mode implementation, in which an input transition is propagated through the extra branches in the form of a pulse signal internally, as shown in Figure 7b. When the input toggles 0→1, the internal pulse signal at *S2* turns on the last stage pull-up transistor for

the duration of the pulse. Within this time window, the drive strength of last stage is equivalent to combined strength of the extra branch and the low power branch while output is being driven to *Vdd*. At the end of the pulse, the pull-up transistor is turned off and the state of the output is maintained by the last stage of the low branch. This internal pulse-mode implementation helps reducing the leakage power consumption by the extra branches since both the pull-up and pull-down of the extra branches are off when the input is steady. Table 6 shows the benchmark result of design B3 in terms of power and transistor area.

Table 6. Benchmark – design B3

	Dynamic Power+ Short- Circuits Power (µW)	Nominal Leakage Power (µW)	Transistor Area (normalized)
Tunable buffer B3	9.01	1.24	39.8



(a) (b)

Figure 7. Alternative adaptive buffer design (B3)

The operation of the extra branch can be illustrated in more details with a timing diagram shown in Figure 8. The signal "gated ctrl", "S1" and "S2" are shown in the circuit diagram in Figure 7a. Initially, if control is asserted to select the high speed configuration, gated_ctrl will be at 0. When the input transitions from 1→0, signal *S1* and *S2* on pull-up branch will be triggered. As *S2* cause the output node to transition from 0→1, *S2* is also fed back and as a result gated_ctrl switches 0→1. Assertion of gated_ctrl will also toggle *S1* and *S2* to turn off the pull-up assisting branch. When input switches from 0→1, the pull-up branch will not toggle, and only the pull-down assisting branch will be activated. Same as the pull-up assisting branch, the pull-down assisting branch will be turned off through a feedback signal after output has completed the transition 1→0.

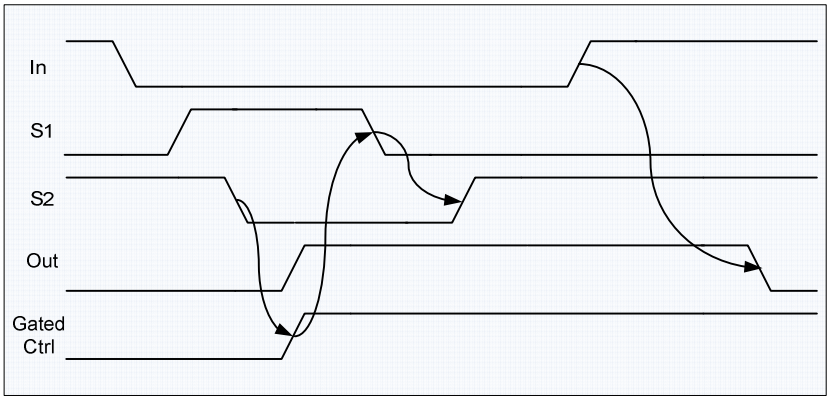


Figure 8. Timing diagram for B3

Similar to design B2, design B3 also suffers from short-circuit power in the last stage if there is a timing offset between the extra branch and the low power branch. However, since for any transition only one of the pull-up or the pull-down propagates the input signal, there would be short-circuit power concern only when the extra branch signal arrives earlier than the low-power branch. If we take the propagation delay

difference as $t_{b1} - t_{b2}$, then the last stage dissipates short-circuit power when $t_{b1} - t_{b2} > 0$.

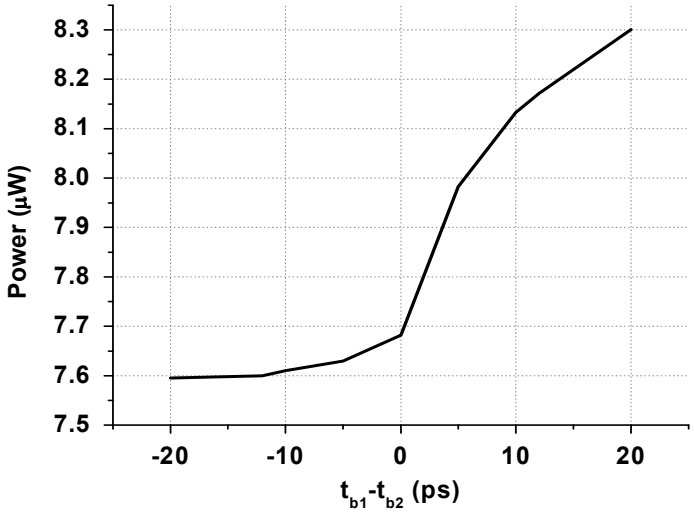


Figure 9. Active power vs. delay offset for B2

3.3 Implementation Overhead

The ability to switch between a high-speed and a low-power configuration comes with some penalties. When the buffer is in its low-power configuration, the high-speed branch in design B1 and the extra branches in design B2 and B3 are deactivated by setting *ctrl* to '0'. However, these deactivated branches still dissipate leakage power, adding to the total power of the buffer. Furthermore, the gate capacitance at the input of the deactivated branches, as well as the drain capacitance at the output, also appear as parasitic capacitance to the low power branch. Moreover, the control transistors, driven by *ctrl* and $\overline{ctrl}$, are used to configure the tunable buffer. Nonetheless, the buffer stages with the control transistors have larger logical effort. In other words, for the same drive strength, these buffer stages have larger input capacitance. As mentioned earlier, for design B2 and B3, synchronization of different branches is important, and the output

stage will be short-circuited when the last stage pull-up of one branch is on at the same time as the last stage pull-down of another branch.

Fortunately, these costs are different for the three implementations described earlier. The low-power configuration of these implementations sees different amount of parasitic capacitance. The number of control transistors used is also different for each implementation. Furthermore, the last stage short-circuit power problem is more severe for B2 than for B3, while design B1 does not exhibit short-circuit power in the last stage when there is significant local variation. In section 5.1, we will pros and cons of these three buffers and provide guidelines for choosing the best implementation for a given system specification.

CHAPTER 4 BUFFER DESIGN USING OPTIMIZATION

4.1 VARIATION-AWARE OPTIMIZATION

In this section, an optimal sizing strategy for the tunable buffer shown earlier is described in details. The objective is to size the tunable buffer such that the average power consumption is minimized. We propose to formulate this problem as a two-stage optimization, using the conceptual framework of adaptable optimization developed in [20]. While some variables are solved in the first stage, some are left undermined until the targeted uncertain data is revealed in the second stage. In order to make the best decision in the first stage, we need to make the first stage depend on the statistics of uncertainty as well as the range for stage-2 variables.

An application of adaptable optimization to dynamic voltage scaling and adaptive body biasing is shown in [34]. In this thesis, we adapt the algorithm in [34] to design a two-configuration size-tunable buffer. The two stages in optimization are the design-time sizing and post-manufacture-time tuning while the target uncertain parameter is gate length L. The only stage-two variable here is the choice between a high-speed configuration or a low-power configuration, denoted by a selector variable $x \in \{x_{LP}, x_{HS}\}$. The statistical average power consumption can be expressed as

$$\mathrm{E}[P_{tot}] = p(x = x_{LP})E[P(x = x_{LP})] + p(x = x_{HS})E[P(x = x_{HS})] \qquad (3.1)$$

$p(x = x_{LP})$ and $p(x = x_{HS})$ represent the probability of using the high-speed and the probability of using the low-power configuration while $E[P(x = x_{LP})]$ and $E[P(x = x_{LP})]$ are the average power for the two configurations.

In order to account for process variation and post-silicon tuning in design-time, the post-silicon tuning rule is chosen to be based on the realization of process parameter *L*. The low-power configuration is used when gate length $L \in (-\infty, l_0]$, and the high-speed buffer is used for $L \in [l_0, \infty)$. With this choice, the variation distribution of *L* is partitioned at the point l_0, as shown in Figure 10.

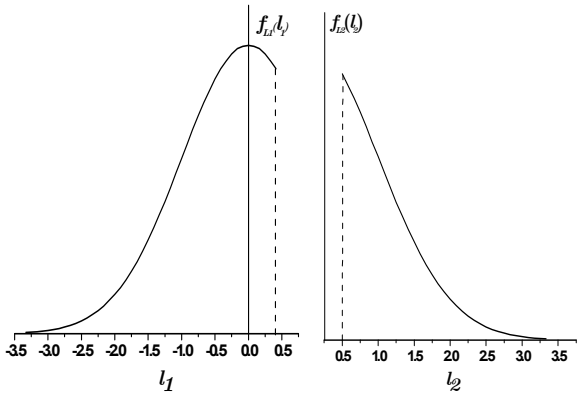


Figure 10. Strategy for truncating the normal distribution

By this partition, the probability of using either configuration can be computed with equation (3.2). Both the statistics of the process variation and the post-silicon tuning strategy are captured in the computation of these probability values. By substituting these two probability values back to the objective function in equation (3.1), the design-time optimization is now dependent on the uncertainty and stage-two decisions.

$$p(x = x_{LP}) = p(L \le l_0) = \phi(\frac{l_0 - \mu_L}{\sigma_L}) \qquad (3.2)$$
$$p(x = x_{HS}) = \gamma - p(x = x_{LP})$$

In the equation, ϕ is the *cdf* of $N(0,1)$ and γ is the timing yield.

The design-time sizing problem can then be described by the following summary given in Table 7, including the optimization variables as well as the constraints on minimizing average power and meeting timing delay.

Table 7. Optimization summary

Variables	Transistor size w_j , truncation point l_0
Constraints	$\min \ \mathrm{E}[P_{tot}] s.t. p(D \le T) \ge \gamma$ (3.3)

To solve this optimization problem, the objective function *E[P]* and the delay constraint $D \le T$ need to be written as functions of the optimization variables. The calculation of the probability in equation (4) has been shown earlier. The average power of either configuration $\mathrm{E}[P_{LP}]$ and $\mathrm{E}[P_{HS}]$, can be computed with the following integrals

$$\mathrm{E}[P_{LP}] = \int_{-\infty}^{l_0} P(L, w_{LP1}, ..., w_{LPN}) \rho(L) dL$$
$$\mathrm{E}[P_{HS}] = \int_{l_0}^{\infty} P(L, w_{HS1}, ..., w_{HSN}) \rho(L) dL \quad (3.4)$$

The integrant $P(L,w_1 \ldots w_N)$ can be calculated with equation (2.4) and $\rho(L)$ is the probability density function of gate length with $L \sim N(\mu_L, \sigma_L^2)$.

The delay constraint can be written as a function of the variables in the following manner:

$$l_Y = \mu_L + \phi^{-1}(\gamma) * \sigma_L$$
$$D_{LP} = D(l_0, w_{LP1}, ..., w_{LPN}) \le T \quad (3.5)$$
$$D_{HS} = D(l_Y, w_{HS1}, ..., w_{HSN}) \le T$$

Here, $D(L,w_1 \ldots w_N)$ can be calculated with equation (2.2).

4.2 ALGORITHM IMPLEMENTATION

Figure 11 shows the pseudo-code for one possible implementation of the optimization algorithm. After obtaining all specifications for the buffer chain in first step, a regular buffer chain is designed in the second step. Then, in the final step, an adaptive buffer is designed with the same number of stages as the basic buffer design.

```
1. get specs
2. while ( N < ln C_L )
     solve (3.3)
     if ( feasible )
        save P_0
        goto step 3
     endif
     set N = N + 1 and goto step 2
3. adapt_buffer ( N )

adapt_buffer ( N )
     1. init P_MIN = ∞ ,  trunc_point = 0
     2.  for l_0 = from (μ_L − 3σ_L) to (μ_L + 3σ_L)
        solve (3.3)
        if ( E[P(l_0)] < P_MIN )
           P_MIN = E[P(l_0)]
           trunc_point = l_0
        endif
```

Figure 11. Pseudo-code for the adaptive buffer design problem

Up to this point, the objective function and constraints in (3.3) have been shown to be functions of variable transistor sizes w_j and truncation point l_0 using equation (3.4) and (3.5). The average power and the buffer chain delay are posynomial functions of transistor sizes but not the truncation point. If all the objective functions and constraints were posynomial functions of the optimization variables, then the optimization problem could be solved as a Geometric Programming problem, and global optimum could be found efficiently.

As a solution to this, we formulate a sub-optimization problem, in which the transistor sizes are solved for with a standard Geometric Programming solver treating l_0 as a known constant. This sub-optimization problem then is solved iteratively for l_0 from $\mu_L - 3\sigma_L$ to $\mu_L + 3\sigma_L$, which covers 99.8% of the possible truncation point values.

Finally, the global optimum of truncation point and transistor sizes can be found by comparing the minimum power consumption at each value of the truncation point l_0.

Using this algorithm, the optimal combination of truncation point and transistors sizes for either configuration is found. The next step is to apply these optimal transistor sizes to the three buffer designs from section 2.

For design B1, the transistor size for the high-speed configuration and the low-power configuration can be applied directly. For design B2 and B3, the sizes for the extra branch or branches are obtained by subtracting low-power transistor sizes from the high-speed branch size at each buffer stage. Due to layout resolution, the transistor sizes needed to be rounded to values that are supported by the technology. Moreover, the transistor sizes of the extra branch have to be larger than the minimum width devices, which places an additional constraint for the design B2 and B3.

The intermediate output of an optimization based on the benchmark described in 2.3 is shown in Figure 12. The optimization algorithm derives the optimal transistor sizes for B1, B2 and B3 at each truncation point, and the power consumption of each implementation is plotted as a function of the truncation point. The optimal truncation point and the minimum power consumption of one implementation can be obtained from the x and the y value of the minimum point on its corresponding curve. As indicated by the results, the minimum power consumption of the three implementations can be ranked as $P_3 < P_2 < P_1$ for this set of specifications. The power consumption of these three implementations under different specifications will be studied in more details in section 5.1. Another notable observation here is that for each of the three buffer designs has a different optimal truncation point.

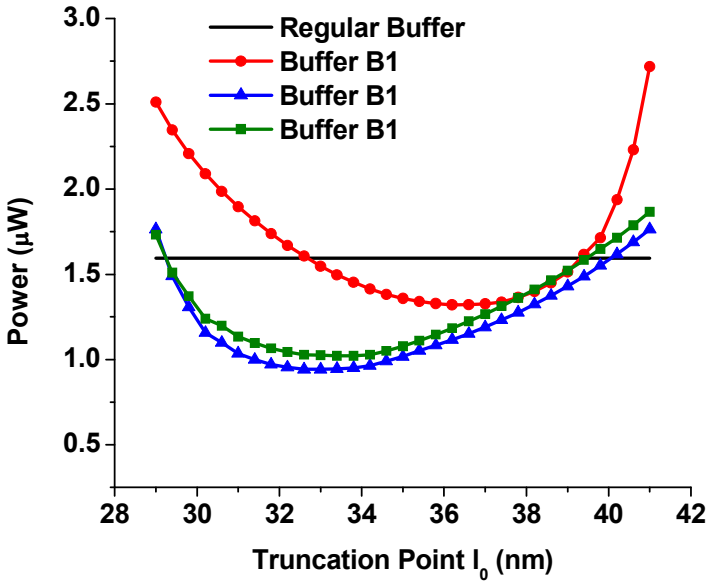


Figure 12. Location of the optimum truncation point

CHAPTER 5 RESULTS AND ANALYSIS

5.1 Implementation Verse Specification

Section 3.3 described the extra capacitance required to implement the ability to switch between a high-speed and a low-power configuration. In addition, the leakage power consumption from deactivated branches is non-negligible. Moreover, when an extra branch or extra branches are used, the circuit could dissipate a considerable amount of short-circuit power. We also provided an explanation on the difference in overhead for the three tunable buffer implementations. Because of this difference, the power-optimal implementation may depend on the specifications. Thus, in this section, we study the relative power consumption of the three implementations while altering the specifications.

Three tests are conducted based on a variation from the benchmark described in section 2.3. In each test, one specification takes on a wide range of values while the others remain the same as the benchmark. The power consumption of the three tunable buffers are measured and plotted as a function of the varying specification. Table 8 summarizes the three tests and the varying specification.

Table 8. Test summary

Test	Varying Specification
1	switching factor α **(1 ~ 10^{-5})**
2	timing constraint T **(125ps~260ps)**
3	standard deviation of threshold voltage variation σ_V **(5mV ~ 120mV)**

At various activity levels, three tunable buffers of different implementation are first sized optimally using the algorithm from section 3.2. Then power consumption of the three tunable buffers is derived from the optimization, and normalized to the power consumption of a regular buffer. As indicated by Figure 13, the result of the test No.1 shows that design B2 is the optimal tunable buffer for low switching activity. This result confirms that design B2 has the lowest leakage power. Looking from the circuit topology of design B2, it has the lowest leakage power because it uses an extra branch that has smaller transistor size than the high-speed branch used in design B1 and uses fewer gates than design B3. However, because of the potential branch timing offset, the design B2 exhibits high short-circuit power. In contrast, design B3 has better dynamic power and short-circuit power characteristics, and it is shown by the test to be the best implementation when switching activity for the buffer is high.

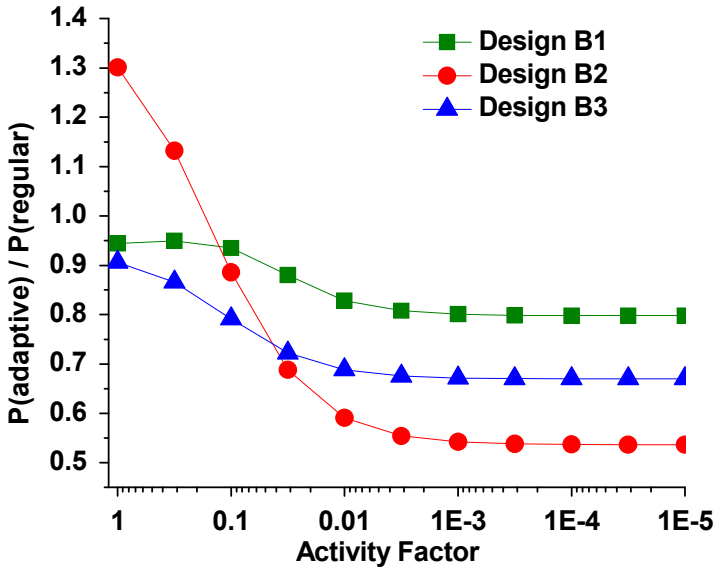


Figure 13. Power of tunable buffer vs. activity factor

In the second test, the normalized power of the three buffer implementations are observed as the delay constraint relaxes. It is foreseeable that as delay constraint relaxes, the transistor sizes are becoming smaller. As a result, the overhead in implementing the ability to switch between two configurations takes up a larger percentage of area and power. Figure 14 shows that buffer 3 is the preferred implementation when timing requirement is tight. However, as the timing requirement is relaxed, the power reduction benefits of using a tunable buffer diminish. For a path delay requirement larger than 250ps, as the normalized power of the three buffers is greater than 1, it is no longer beneficial to use a tunable buffer. When the delay target is less than 180ps, or no more than approximately 45% over the minimum delay at 125ps, the achievable power reduction by design B3 is at least 20%.

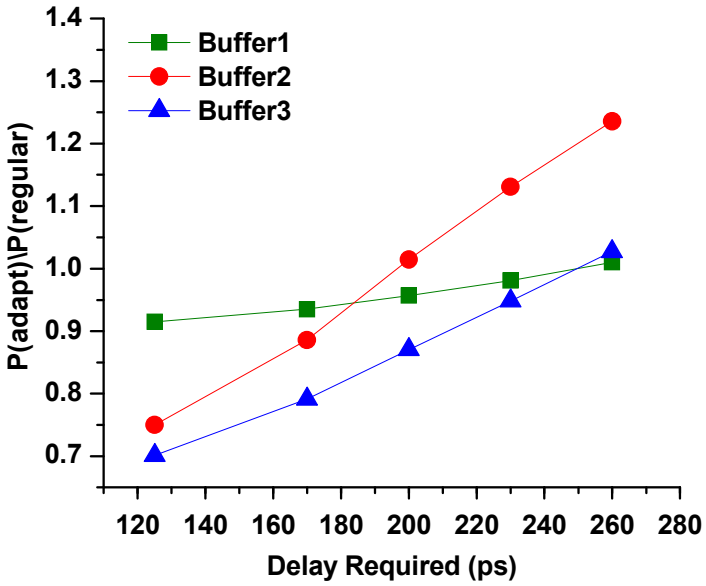


Figure 14. Power of adaptive buffer vs. timing requirement

The synchronization of two branches is important for design B2 and B3. However, due to local variation, it is highly unlikely that timing offset between the two branches is zero. In the third test, the normalized power of the three implementations is plotted against the magnitude of local variations. A larger σ_V would cause a larger statistical average timing offset. As shown in Figure 15, both design B2 and design B3 have a notable increase in power when average timing offset increases with local variations. However, design B3 is still the implementation with the lowest average power when all other specifications in the benchmark are unchanged.

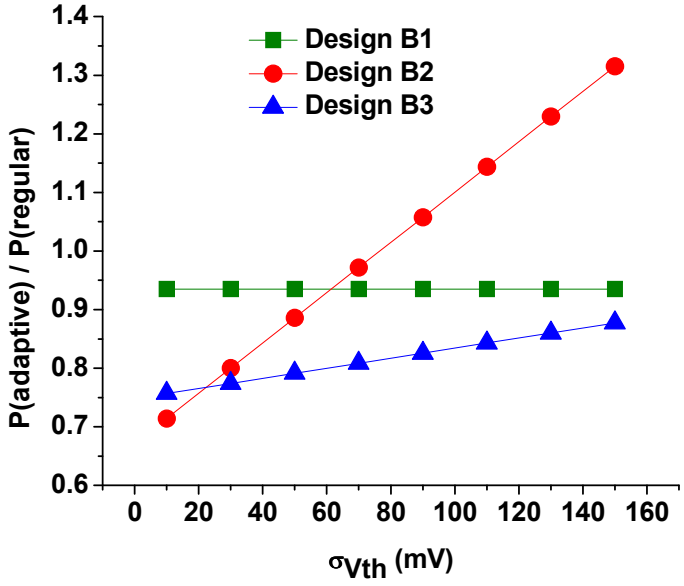


Figure 15. Power of adaptive buffer vs. magnitude of local variation

5.2 Design Example

The result of 4.1 indicates that employing the proposed tunable design B2 and B3 can bring substantial power reduction given that the timing constraint is tight and activity factor of the block is low. The question becomes whether such specification would exist in a real design.

In this section, we will derive the design specification of the word-line drivers of a 64kb SRAM. The SRAM is implemented as a small-signal array, and it is organized as two 128x256 banks of 6T bit cells. It consists of 128 word line drivers and 64 sense-amplifiers, and it is designed for a L1-D Cache running at 1GHz. The block activity factor estimated to be 0.5 according to the percentage of "load" and "store" instructions reported in [32]. The architecture of the memory is shown in the Figure 16.

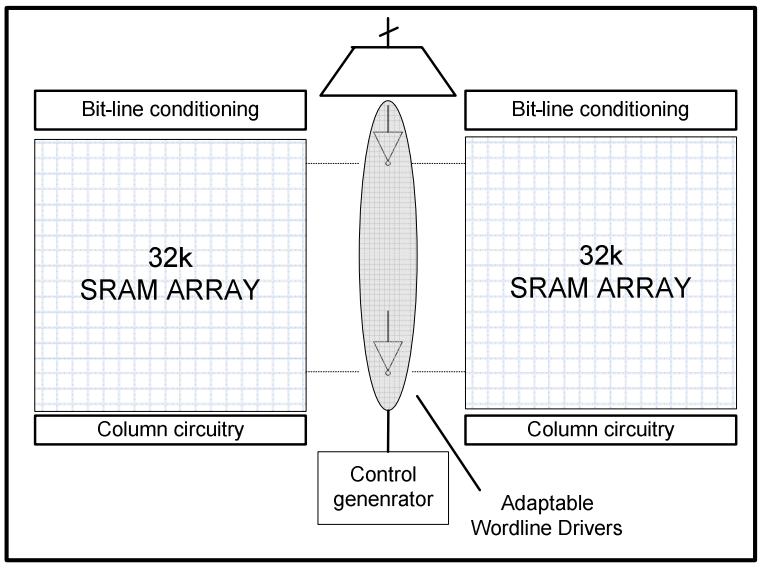


Figure 16. Tunable word-line drivers in a SRAM

The specification for the word-line driver is provided in Table 9 using the same format as the benchmark described in section 2.3, with a few requirements that are unique to a word-line drivers design. Firstly the load on word-line is not purely capacitive because of contributions from wire resistance. Secondly, we can obtain the delay requirement for the word-line drivers by performing a simple critical-path analysis. During a read operation, the bit-line is pre-charged to Vdd in the first half of the clock cycle. In second half-cycle, the critical path delay is the sum of word-line delay, bit-line delay and sense-amplifier delay. Thus, the word-line delay requirement for driving the word-line active (low-to-high) is taken to be 170ps, approximately one third of a half-cycle. However, the output high-to-low transition is not path of the critical path, the delay requirement is relaxed to 250ps. Finally, the switching activity for one world-line driver is computed as

$$\alpha_{swicth} = \tfrac{1}{128} \times \alpha_{block} \times \frac{t_{delay}}{t_{cycle}} \qquad (4.1)$$

The activity factor is scaled by $\frac{1}{128}$ because at any time, only one word-line is activated to turn on the world line, assuming the SRAM is single-ported. According to this calculation, the activity factor is approximately 0.0013.

Table 9. Word-line driver specification

Supply Voltage	1.0V	**switching factor** α	0.0013
Timing constraint T	Rise= 170ps Fall = 250ps	μ_L	35nm
Frequency f	1GHz	μ_{Vth}	250mV
timing yield γ	0.999	σ_L	2nm
capacitive load C_L	$512C_m+R_{wire}$	σ_V	50mV

Taking advantage of different timing requirements for 0→1and 1→0 transitions, we propose to implement the word-line driver with a modified version of the buffer chain design 3. The extra pull-down branch is removed from the design because timing for output transition 1→0 is not critical. This modification allows us to further reduce leakage power and area overhead of tunable buffer. The schematic of modified design 3 is shown in Figure 17.

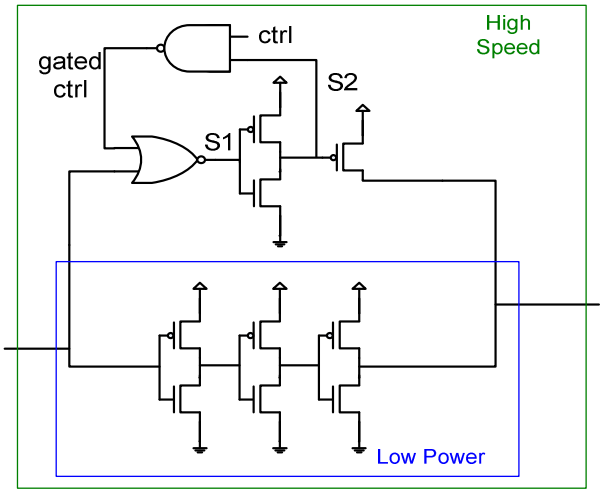


Figure 17. Adaptive word-line driver (modified B3)

A possible implementation for the control signal generator is shown below in Figure 18. The oscillating period of a ring oscillator is a function of the global variation while impact of the local variation tends to average out in the inverter chain. This oscillating frequency can be detected with a counter counting the number of pulses generated for a given amount of time. The resolutions of this sensor can be adjusted by changing the size of the counter or by increasing the number of stages in the ring oscillator.

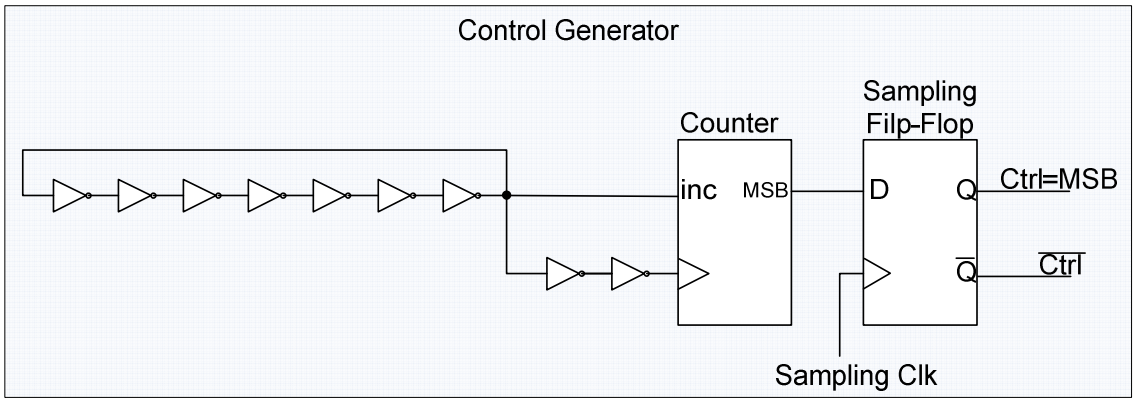


Figure 18. Tunable word-line drivers in a SRAM

As mentioned earlier, the 64kb SRAM consists of 128 word-line drivers. The average power and total area of these drivers implemented using regular buffers and adaptive buffers are shown in Table 10. The average power is obtained from an Hspice simulation using 32nm high performance predictive model, and the area is estimated with NCSU 45nm free PDK. Compared to the regular buffers, the adaptive buffer approach achieves an average power reduction of 30%. The area overhead due to the implementation the switching capability and the generation of the control signal is estimated to be 40% compared to the regular buffers.

Table 10. Word-line drivers power and area

	Average Power (μW)	Transistor Area (normalized)	Extra Interconnect Area (normalized)	Control Generator Area (normalized)
Regular buffer	53.4	2368	0	0
Tunable buffer	37.1	2672	426	95

CHAPTER 6 CONCLUSIONS

Buffers chains are heavily used in a modern integrated system to drive large capacitance. These buffer chains can be found in important circuit building blocks such as SRAM, interconnect network, and clock trees. Therefore, the power consumption of these buffers is a critical component to the total power dissipation of a system. In the presence of process variations, a post-silicon tunable buffer can be used to greatly reduce the average power while still meeting the same timing yield as a regular buffer. This tunable buffer can be designed using dynamic voltage scaling and adaptive body-biasing. Alternatively, an adaptive buffer can employ a size-tunable scheme in which the buffer has the ability to switch between a high-speed configuration and a low-power configuration. According to a comparison made in this work, the size-tunable approach out-performs better than supply-voltage scaling for fine grain tuning.

After reviewing an existing buffer topology to implement the size-tuning scheme, two new buffer structures were proposed to reduce the leakage power and area overhead. Furthermore, a post-fabrication and design-time co-optimization algorithm was developed to find the optimum buffer sizes and tuning strategy given the objective of minimizing power consumption.

By studying different possible system specifications, it is found that this class of tunable buffer works best when the buffer has a timing target that is 45% or less bigger than the achievable minimum delay and when activity factor is low. Finally, a SRAM word-line driver design example is presented to demonstrate that such specifications are indeed encountered in real systems. By Hspice simulation with 32nm high performance predictive model, a 30% reduction in average power is possible with an estimated 40%

area overhead. With no performance penalty, this tunable buffer approach is an attractive option for reducing power consumption in systems where area consumption is not a critical factor.

REFERENCES

[1] J. M. Rabaey, et. al., Digital Integrated Circuits: A Design Perspective, 2nd ed., Pearson Education Int., 2003.

[2] B. S. Amrutur and M. A. Horowitz, "Speed and power scaling of SRAM's," IEEE J. Solid-State Circuits, vol. 35, no. 2, pp. 175–185, Feb. 2000.

[3] S. Chen, et. al. "A new output buffer for 3.3-V PCI-X application in a 0.13-/spl mu/m 1/2.5-V CMOS process," Proc. Asia-Pacific Conference on Advanced System Integrated Circuits, 2004, pp. 112-115.

[4] S. Tam, et. al., "Clock generation and distribution for the first IA-64 microprocessor," Solid-State Circuits, IEEE Journal of, vol.35, no.11, pp.1545-1552, Nov 2000.

[5] R. Ho, et. al.,"The future of wires," Proceedings of the IEEE , vol.89, no.4, pp.490-504, Apr 2001.

[6] H. Kaul et. al., "A novel buffer circuit for energy efficient signaling in dual-VDD systems," Proc of GLSVLSI, 2005, pp. 462-467.

[7] G. Villar, et. al., "Energy optimization of tapered buffers for CMOS on-chip switching power converters," Proc of ISCAS, 2005, pp. 4453-4456.

[8] A. Srivastava et al., "Statistical optimization of leakage power considering process variations using dual-Vth and sizing," in Proc. of DAC, 2004, pp. 773 – 778.

[9] M. Mani, et. al., "An Efficient Algorithm for Statistical Minimization of Total Power under Timing Yield Constraints," in Proc. of DAC, 2005, pp. 309-314.

[10] J. Xiong, et. al., "Buffer Insertion Considering Process Variation," Proc. of DATE, 2005, pp. 970-975.

[11] V. Khandelwal, et al., "A probabilistic approach to buffer insertion," Proc. of ICCAD, 2003, pp. 560-567.

[12] J. Tschanz et al., “Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage,” ISSCC Tech. Dig., pp. 422-423, 2002.

[13] S. Narendra et al., “Impact of using adaptive body bias to compensate die-to-die Vt variation on within-die Vt variation”, in Proc. of ISLPED, 1999, pp. 229-232.

[14] S. Martin et. al., “Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads,” in Proc. of ICCAD, 2002, pp. 721-725.

[15] C. H. Kim, et. al., “A process variation compensating technique for sub-90 nm dynamic circuits,” in VLSI Circuits Symp. Tech. Dig., 2003, pp. 205–206.

[16] V. Khandelwal and A. Srivastava, “Variability-driven formulation for simultaneous gate sizing and post-silicon tunability allocation,” in Proc. of ISPD, 2007, pp. 11-18.

[17] H. Wang et al., “Variable tapered pareto buffer design and implementation allowing run-time configuration for low-power embedded SRAMs,” IEEE Trans. on VLSI, Vol. 13, Oct. 2005.

[18] J.-S. Choi and K. Lee, “Design of CMOS tapered buffer for minimum power-delay product,” IEEE J. Solid-State Circuits, vol. 29, no. 9, pp. 1142–1145, Sep. 1994.

[19] R. Rao et al., “Statistical analysis of subthreshold leakage current for VLSI circuits,” IEEE Trans. on VLSI Systems, 12(2), pp. 131-139, February 2004.

[20] C. Caramanis, Adaptable Optimization: Theory and Algorithms, PhD dissertation, Massachusetts Institute of Technology, June 2006.

[21] http://www.stanford.edu/~boyd/ggplab/

[22] Y. Cao et al., “New paradigm of predictive MOSFET and interconnect modeling for early circuit design,” in Proc. of IEEE CICC, 2000, pp. 201-204.

[23] C-W. Eng et. al., “An Improved Shift-and-Ratio Effective Channel Length Extraction Method for Metal Oxide Silicon Transistors with Halo/Pocket Implants,” Jpn. J. Appl. Phys., pp. 2621-2627

[24] Q. Ye and S. Biesemans, “Leff extraction for sub-100 nm MOSFET devices,” Solid-State Electronics, Volume 48, Issue 1, , January 2004, pp. 163-166.

[25] F. Hamzaoglu, et. al., "Split-Path Skewed (SPS) CMOS Buffer for High Performance and Low Power Applications", IEEE Transactions on Circuits and Systems II, vol. 48, no. 10, pp. 998-1002, Oct. 2001.

[26] M. Orshansky et al., “Impact of systematic spatial intra-chip gate length variability on performance of high-speed digital circuits,” in Proc. ICCAD, 2000, pp. 62–67.

[27] Y. K. Ramadass and A. P. Chandrakasan, “Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications,”in *Proc. Power Electronics Specialists Conf.*, 2007, pp. 2353–2359.

[28] A. P. Chandrakasan et al., “Low-power CMOS digital design,” IEEE J. Solid-State Circuits, vol. 27, no. 4, pp.473–483, Apr. 1992.

[29] M. Horowitz, T. Indermaur, and R. Gonzalez, “Low-power digital design,” in *Symp. Low Power Electr.,* Oct. 1994, pp. 8–11.

[30] Kohno, I.; Sano, T.; Katoh, N.; Yano, K., "Threshold cancelling logic (TCL): a post-CMOS logic family scalable down to 0.02 μm," *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International* ,pp.218-219, 459, 2000.

[31] Neau, C.; Roy, K., "Optimal body bias selection for leakage improvement and process compensation over different technology generations," *Low Power Electronics and Design, 2003. ISLPED '03. Proceedings of the 2003 International Symposium on* , pp. 116-121, 25-27 Aug. 2003.

[32] S. Birdj, et al., “Performance Characterization of SPEC CPU Benchmarks on Intel’s Core Microarchitecture based processor,”*Proc. of 2007 SPEC Benchmark workshop*, Jan., 2007.

[33] T. Mizuno, J. Okamura, and A. Toriumi, “Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFETs,” *IEEE Trans. Electron Devices*, vol. 41, pp. 2216–2221, Nov. 1994.

[34] M. Mani, A.K. Singh, M. Orshansky, “Joint design-time and post-silicon minimization of parametric yield loss using adjustable robust optimization”, in Proc. of ICCAD, 2006.

Vita

Mario Chichun Lok was born in Guang Zhou, China. He attended Affiliated High of South China Normal University from grade 7 to grade 11 and completed his high school at Charles E London Secondary in Vancouver upon moving to Canada. He then pursued his undergraduate studies in Engineering Physics at the University of British Columbia from where he graduated in May 2008. In August 2008, he was admitted to the University of Texas at Austin for graduate studies in Electrical Engineering, where he is currently pursuing his Master degree.

Email: mario.c.lok@gmail.com

This thesis was typed by the author.