Design of Power-Optimal Buffers Tunable to Process Variability

Mario Lok[1], Ku He[1], Murari Mani[2] , Constantine Caramanis[1] and Michael Orshansky[1]

[1]Department of Electrical and Computer Engineering, University of Texas, Austin, USA
[2]Advanced Micro Devices, Austin, USA

***Abstract*— In many digital designs, multi-stage tapered buffers are needed to drive large capacitive loads. These buffers contribute a significant percentage of overall power. In this paper, we propose two novel tunable buffer designs that enable power reduction in the presence of process variation. A strategy to derive the optimal buffer size and 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 variation. Using a combination of HSPICE simulations and our optimization algorithm, we show that up to 30% average power reduction can be achieved with the proposed buffer structures.**

Index Terms – Buffer Design, Adaptive Design, Post-Silicon Tuning, Statistical Design, Low Power Design

I. INTRODUCTION

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]. Because of their size, these buffers consume larger area and more power than typical logic gates. In addition, aggressive deployment of buffers in high-performance systems means that they now account for a significant portion of total power consumption of the chip.

With the rise of variability, design-time sizing become challenging as increasingly larger buffers are needed to maintain the parametric yield; however, simply increasing buffer size imposes a power overhead on each instance regardless to its actual performance. To solve this problem, several post-silicon techniques have been proposed, including adaptive body biasing [6][7], dynamic voltage scaling [8], as well as a statistical approach that co-optimizes transistor sizing and body biasing [15]. Compared to the design-time approach, post-silicon tuning has the advantage of allowing the designer to tune each chip individually to help meet performance constraints. Such flexibility would allow designers to use smaller size buffers while meeting the same yield requirement, and hence reduce the overall power.

In this paper, we investigate a specific class of tunable buffer that was first proposed in [9]. It has the capability of switching between a high-speed and a low-power configuration to trade-off energy-delay tradeoffs. Compared to adaptive body biasing and supply voltage scaling, this tunable buffer enables post-manufacture tuning for smaller circuit blocks, as analog blocks such as a on-chip regulator or a voltage-level shifter is not necessary. However, the circuit implementation of buffer chain in [9] 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. To improve on the design and tuning of this tunable buffer, we propose new implementations with reduced area overhead and 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 [9] for different system specifications to analyze the pros and cons of each implementation.

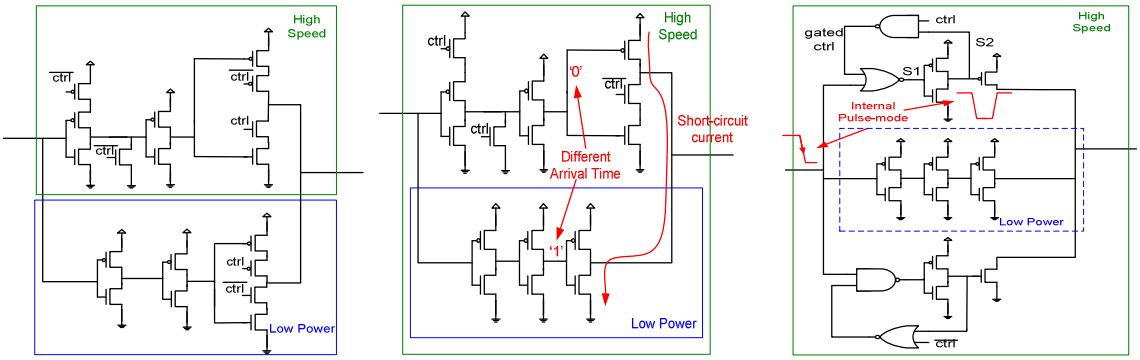
II. BUFFER IMPLEMENTATION

A. Circuit Implementation

An implementation of the size-tunable buffer proposed in [9] is shown in Figure 1a, denoted by B1. The ability to adjust the buffer drive-strength is provided by a control signal that switches the buffer between the high-speed and low-power configurations.

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 1 shows a comparison between a regular non-tunable buffer and the tunable buffer B1. The transistor sizes of both buffers are derived from our proposed statistical sizing algorithm. The details of this minimum-power sizing algorithm are described in section III. The parameters used in the benchmark can be found in Table 2. The average power is obtained with Hspice simulation and consumed area is estimated with total transistor widths. The consumed area is then normalized to the area of a unit-size inverter. As evidenced by



(a) Buffer B1 (b) Buffer B2 (c) Buffer B3

Figure 1. Adaptive buffer design

Table 1. Regular Buffer vs. Adaptive Buffer

	Dynamic Power (μW)	Leakage Power (μW)	Transistor Area (normalized)
Regular buffer	10.31	1.91	21.23
Buffer B1	9.25	1.81	82.81
Buffer B2	9.6	0.972	32.66
Buffer B3	9.01	1.24	39.8

Table 2. Benchmark

parameter	**value**	**parameter**	**value**
supply voltage	1.0V	switching factor α	0.2
timing constraint	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

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%.

Noting the large area penalty due to having dedicated high-speed branch and low power branch, we propose an alternative implementation labeled B2 is shown in Figure 1b. When B2 is inthe high-speed configuration, the low-power branch is re-used and the input signal propagates through both of the low branch and an extra branch. 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.

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 annotated in Figure 1b. Consequently, the buffer circuit may dissipate a significant amount of short-circuit power. This short circuit power is primarily a function random local variation, as global variation can be compensated by inducing high correlation between two branches with layout techniques.

A third implementation labeled B3 is shown in Figure 1c. 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. As annotated in Figure 1c, 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. 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. As shown in Table 2, design B2 has the minimum static power while design B3 has the minimum total power.

B. *Implementation Overhead*

To best utilize the advantage of the tunable buffer, we also need to understand the cost to implement the control switch. First of all, the gate capacitance at the branch off point, as well as the drain capacitance at the output is larger than a regular buffer. Secondly, when the buffer is in its low-power configuration, the high-speed branch in design B1 or 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. Moreover, the buffer stages with the control transistors have larger logical effort.

Similar to design B2, synchronization of different branches also is important for B3, and the output stage will be short-circuited when the last stage pull-up of one branch is conducting current 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 V, we will analyze the pros and cons of these three buffers and provide guidelines for choosing the best implementation for a given system specification.

III. Variation-Aware Buffer Design

A. Algorithm

In this section, we discuss the optimal buffer sizing algorithm when post-silicon tuning is available. With the tunable buffer topologies described in section II, we can account for the magnitude and statistics of process variation during transistor sizing to minimize average power. We assume that two main source of variation are gate length (L) and threshold voltage (V_{TH}), and both follow a normal distribution, denoted by $L \sim N(\mu_L, \sigma_L^2)$ and $V_{TH} \sim N(\mu_{Vth}, \sigma_V^2)$. As shown in [13] and [14], the variation in L is mostly a global parameter while the variation in V_{TH} is a local variant. In this algorithm, we restrict our attention to the global variation in gate length L, and consider the impact of local variation in section IV.

We formulate the tunable buffer sizing problem as a two-stage optimization, using the conceptual framework of adaptable optimization developed in [10]. While some variables are solved in the first stage, some are left undermined until the 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 dependent on the statistics of uncertainty as well as the range of stage-2 variables.

The two stages in optimization are the design-time sizing and post-manufacture-time tuning while the uncertain parameter is gate length L. The stage-two variable here is the choice between the high-speed configuration and the low-power configuration. The statistical power consumption can be expressed as

$$\mathrm{E}[P_{tot}] = p_{LP} E[P_{LP}] + p_{HS} E[P_{HS}] \quad (1)$$

Here, p represents the probability of using either configuration while $E[P_{LP}]$ and $E[P_{HS}]$ are the statistical 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 2.

By this partition, the probability of using either configuration can be computed with equation (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 (1), the design-time optimization is now dependent on the uncertainty and stage-two decisions.

$$p_{LP} = p(L \le l_0) = \phi\left(\frac{l_0 - \mu_L}{\sigma_L}\right)$$
$$p_{HS} = \gamma - p_{LP} \quad (2)$$

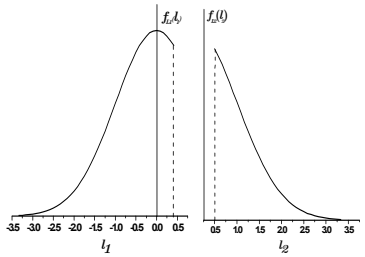


Figure 2. Strategy for truncating the normal distribution

In the equation, ϕ is the *cdf* of $N(0,1)$ and γ is the timing yield. The sizing problem can be described by equation (3) including the optimization variables as well as the constraints on minimizing average power and meeting timing delay:

$$\min_{w_j, l_0} \mathbf{E}[P_{tot}] \quad s.t. p(D \le T) \ge \gamma \quad (3)$$

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

$$\mathrm{E}[P_{LP}] = \int_{-\infty}^{l_0} P(L, w_{LP1}, ..., w_{LPN}) \rho(L) dL$$
$$\mathrm{E}[P_{HS}] = \int_{l_0}^{\infty} P(L, w_{HS1}, ..., w_{HSN}) \rho(L) dL \quad (4)$$

The integrand $P(L,w_1...w_N)$ is the power consumption of the buffer chain as a function of transistor sizes 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

$$l_Y = \mu_L + \phi^{-1}(\gamma) \cdot \sigma_L$$
$$D_{LP} = D(l_0, w_{LP1}, ..., w_{LPN}) \le T$$
$$D_{HS} = D(l_Y, w_{HS1}, ..., w_{HSN}) \le T \quad (5)$$

Both the power and the delay of a buffer chain can be derived as a posynomial function of transistor dimensions L and W [1]. This allows us to implement our optimization algorithm using geometry programming with which the global optimum can be computed efficiently. To enhance the accuracy of this algorithm, empirical posynomial models of power and delay are developed through Hpsice simulation in place of theoretical models. On top of that, the algorithm is modified to reflect the circuit implementation overhead for each of the three tunable buffers described in section II.

IV. Result and Analysis

A. Implementation Versus Specification

Section II described the different overhead associated with the three buffer implementations. As these overheads vary for different block activity factor, delay constraints and magnitude of local variations, the power-optimal implementation may also hinge on these specifications. Thus, in this section, we study the relative power consumption of the three implementations while altering the specifications.

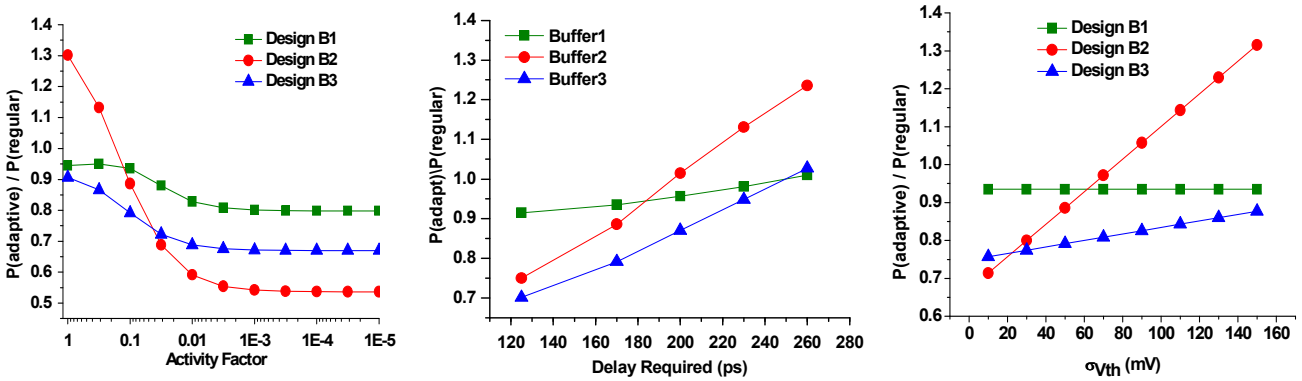


Figure 3. Power of adaptive buffer vs. switching activity, timing requirement and magnitude of local variation

Three tests are conducted based on a variation from the benchmark described in section II. In each test, one parameter 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.

The three tunable buffers are first sized optimally using the algorithm from section III. 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 3, the first test shows that design B2 has the lowest leakage power. Reviewing 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, due to the potential branch timing offset caused by local variation, design B2 exhibits high short-circuit power and is not power efficient for a design with high activity factor

In contrast, design B3 is shown by the first test to be the best implementation when switching activity for the buffer is high. Furthermore, design B3 has a less severe problem in short-circuit power. When the σ_V is swept in third test, a larger σ_V would cause a larger statistical average timing offset. Although both design B2 and design B3 have a notable increase in power when average timing offset increases with local variations, design B3 is still the implementation with the lowest average power over a wide range of σ_V.

Finally, 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 consumes a larger percentage of power. Figure 3 shows that buffer B3 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 45% over the minimum delay at 125ps, the achievable power reduction by design B3 is at least 20%.

V. CONCLUSIONS

In this paper, two new buffer structures were proposed to reduce the leakage power and area overhead of a size-tunable adaptive buffer. 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 the proposed tunable buffers work best when the buffer has a timing target that is less than 1.5 times the achievable minimum delay and when activity factor is low.

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,” JSSC, 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] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] Neau, C.; Roy, K., "Optimal body bias selection for leakage improvement and process compensation over different technology generations," ISLPED '03. pp. 116-121, 25-27 Aug. 2003.