IEICE TRANS. FUNDAMENTALS, VOL.E92–A, NO.12 DECEMBER 2009

PAPER Special Section on VLSI Design and CAD Algorithms

Trade-Off Analysis between Timing Error Rate and Power Dissipation for Adaptive Speed Control with Timing Error Prediction

Hiroshi FUKETA[†,††a)], ***Student Member***, **Masanori HASHIMOTO**[†,††], **Yukio MITSUYAMA**[†,††], ***and*** **Takao ONOYE**[†,††], ***Members***

SUMMARY Timing margin of a chip varies chip by chip due to manufacturing variability, and depends on operating environment and aging. Adaptive speed control with timing error prediction is promising to mitigate the timing margin variation, whereas it inherently has a critical risk of timing error occurrence when a circuit is slowed down. This paper presents how to evaluate the relation between timing error rate and power dissipation in self-adaptive circuits with timing error prediction. The discussion is experimentally validated using adders in subthreshold operation in a 90 nm CMOS process. We show a trade-off between timing error rate and power dissipation, and reveal the dependency of the trade-off on design parameters.

key words: *adaptive speed control, subthreshold circuit, timing error prediction, timing margin variability*

1. Introduction

Circuit speed is becoming more sensitive to manufacturing variability, operating environment, such as supply voltage and temperature, and aging due to NBTI (negative bias temperature instability) and HCI (hot carrier injection). Thus, timing margin of a chip varies chip by chip due to manufacturing variability, and it also depends on its operating environment and age. For a certain chip, a large timing margin exists and it is desirable to slow down the chip for reducing power dissipation with dynamic voltage scaling or body-biasing. In an operating condition, the timing margin is not enough and the circuit should be speeded up. The adaptive speed control is believed to be promising.

To sense the timing margin, critical path replica [1] has been traditionally used. However, its efficiency is deteriorating because the performance difference between the replica and the actual critical path is significant due to increasing within-die variation. To more efficiently sense the timing margin, in-situ techniques have been studied [2]–[5]. However, this scheme inherently involves a critical risk of timing error occurrence. When the circuit is slowed down, it is not possible to perfectly predict whether the enough timing margin exists after slowed down.

"Razor I" in [2] and "Razor II" [3] detect timing errors with a delayed clock in a processor and correct the errors using extra recovery logic or re-execution of instructions. They control supply voltage monitoring the timing error rate and reduce power dissipation. The error recovery is performed exploiting a function commonly implemented in processors, and hence it is not easy to apply it to general sequential circuits. In contrast, "Canary Flip-Flop*" [4] and "Defect Prediction Flip-Flop (DPFF)" [5] have been proposed that aim not to detect timing errors but to predict them. When the timing margin is not enough, they capture wrong values, whereas the main flip-flops capture correct values. The difference of captured values gives a timing warning. Timing error prediction is superior to timing error detection in terms of applicability since error recovery mechanism is not necessary as long as a timing warning can be generated before a timing error occurs.

When canary FF is used for adaptive speed control, a timing error can not be completely eliminated, which is believed to be a critical problem that prevents a practical use. When a circuit is slowed down, a timing error could occur before a timing warning emerges. To practically use the adaptive speed control with canary FF, the occurrence of timing errors must be systematically and quantitatively estimated, and designers have to guarantee that the frequency of timing error is lower than the specification. We know an argument that a timing error is definitely unacceptable even though its frequency is extremely low such as once per ten years. However, we believe that when the occurrence of timing error is very low, some systems could accept the errors. For example, video decoding for TV and video recording for security monitoring can accept an error per day, since a small piece of image degradation in a short time is not a problem. Furthermore, strictly speaking, even though fabricated chips are shipped after testing, the timing error occurrence has not been verified to be zero, because the number of test patterns and environmental conditions are limited.

This paper proposes a framework that systematically evaluates the occurrence of timing errors. With the proposed framework, we explore the design space of the adap-

Manuscript received March 19, 2009.
Manuscript revised June 19, 2009.

[†]The authors are with the Department of Information Systems Engineering, Osaka University, Suita-shi, 565-0871 Japan.
[††]The authors are with JST, CREST, Tokyo, 102-0075 Japan.
a) E-mail: fuketa.hiroshi@ist.osaka-u.ac.jp

DOI: 10.1587/transfun.E92.A.3094

*Although a term "Canary Flip-Flop" is also defined in [6], its structure and purpose are different from canary FF referred in this paper. Canary FF in [6] is used to detect how much the supply voltage of FFs can be reduced without losing their states under DVS (Dynamic Voltage Scaling) systems.

tive speed control with canary FF and reveal how the error occurrence depends on design parameters. We also examine the relation between the error occurrence and power dissipation, and demonstrate how much additional power dissipation is necessary to reduce the timing error occurrence. This is a first work that explicitly studies how to evaluate and assure the error occurrence in self-adaptive circuits comprehensively, as far as the authors know. The discussion is experimentally validated using a 32-bit ripple carry adder and a 32-bit Kogge-Stone adder in subthreshold operation in a 90 nm CMOS process. The performance of a subthreshold circuit is sensitive to temperature, and the adaptive speed control for temperature is used for experimental validation in this paper.

The remainder of this paper is organized as follows. Section 2 describes the adaptive circuit delay and power control system with canary FF. In Sect. 3, we discuss the systematic evaluation of power dissipation and timing error rate. Section 4 demonstrates the experimental results, and finally Sect. 5 concludes this paper.

2. Adaptive Speed Control with Canary FF

Figure 1 shows a circuit that adaptively controls the speed and power dissipation using a warning signal generated by a canary FF. The canary FF consists of a normal flip-flop, a delay buffer and a comparator (XOR gate). When the timing margin is gradually decreasing, a timing error occurs at the canary FF before the main FF captures a wrong value due to the delay buffer, which enables us to predict that the timing margin of the main FF is not large enough. A warning signal is generated to predict the timing errors, and it is monitored during a specified period. Once a warning signal is observed, the circuit is controlled to speed up. If no warning signals are observed during the monitoring period, the circuit is slowed down for power reduction. This speed control overcomes the variation of the timing margin which is different chip by chip and varies depending on operating condition and aging.

Even when the canary FF is well configured to generate the warning signal, the occurrence of timing error can not be reduced to zero. This is because when critical paths are not activated for a long time in the circuit operation, the circuit might be slowed down too much. If a critical path is activated in this condition, a timing error necessarily happens. To reduce the error occurrence, we have to examine

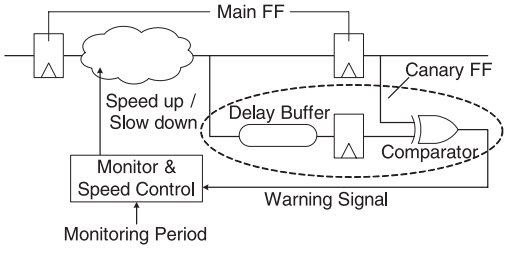


Fig. 1 Adaptive speed control with canary FF.

and tune the following design parameters.

- location where canary FF should be inserted
- delay time of the delay buffer in canary FF
- monitoring period
- fineness of the speed control

In this paper, we examine how the error occurrence depends on the design parameters, and demonstrate that the optimal parameters vary depending on the required error frequency. To do this, the next section discusses how to estimate the timing error occurrence.

3. Systematic Evaluation of Power Dissipation and Timing Error Rate

3.1 Assumed System and Notations

We here assume that the speed is controlled digitally in Fig. 1, because discrete supply voltage and body-bias voltage are often generated and used [1], [7]. We, in this paper, assume that the speed can be changed without additional power dissipation just for simplicity, although its consideration is straightforward in our analysis.

In this paper, a term "speed level" is used to express how fast or slow the circuit is controlled. Let l be the speed level, and higher l means that the circuit is controlled faster. The maximum and minimum levels $l_{\max}$ and $l_{\min}$ are given. The system starts with $l = l_{\max}$, and when no warning signals are observed during the monitoring period, l is decremented by one and the circuit is slowed down. Once a warning signal is observed, l is incremented by one, and the circuit is speeded up.

We define the following design parameters.

- i: the location of the canary FF, where the canary FF is inserted to ith FF. In this paper, we insert only one canary FF.
- D_d: the buffer delay in the canary FF.
- N_{mon}: the monitoring period of the warning signal.

The system requirements are often given by

- $P_{\text{ow, avg}}$: the average power dissipation.
- N_{err}: the average interval (cycles) between the timing errors, which is directly related to the timing error rate.
- T_c: the clock period.

The buffer delay D_d and the clock cycle T_c are represented in seconds. Meanwhile the unit of N_{mon} and N_{err} is the number of cycles.

3.2 Probabilities of Warning Signals and Timing Errors

The timing margin varies depending on operating conditions, such as supply voltage, temperature and aging, and the conditions change with various time span, for example aging is often evaluated by year, and temperature changes in seconds. In this paper, a parameter X denotes the operating condition under consideration.

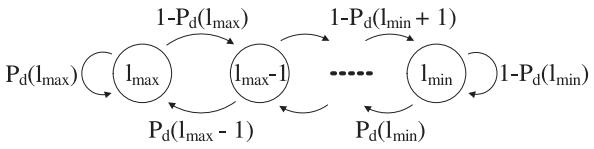


Fig. 2 Speed level transition.

To evaluate the occurrence probabilities of the warning signal and timing error, we introduce path activation probabilities P_i and P_{all}. Let $P_i(t, l, X)$ be a probability that at least one of the paths terminating at the ith FF whose delays are larger than t is activated in a cycle. P_i depends on speed level l and condition X. Let $P_{\text{all}}(t, l, X)$ be a probability that at least one path in a circuit whose delay is larger than t is activated in a cycle. P_{all} also depends on speed level l and condition X. P_i and P_{all} are dependent on the circuit structure, and the following discussion assumes that they are given. In addition, we assume that path activation probabilities have no correlation with those of the other cycles for simplicity in this paper.

When a canary FF is inserted at the ith FF, the occurrence probability of a warning signal at speed level l and condition X, $P_w(l, X)$, can be expressed as

$$P_w(l, X) = P_i(T_c - D_d, l, X) - P_i(T_c, l, X), \tag{1}$$

where D_d is the buffer delay in the canary FF and T_c is the clock period.

$P_{\text{d}}(l, X)$ is a probability that at least one warning signal is detected during monitoring period (cycles) N_{mon} and can be expressed as

$$P_{\text{d}}(l, X) = 1 - (1 - P_w(l, X))^{N_{mon}}. \tag{2}$$

We define $P_{\text{err}}(l, X)$ as a probability that timing errors occur in a cycle at speed level l and condition X when the clock cycle is T_c. $P_{\text{err}}(l, X)$ can be expressed as

$$P_{\text{err}}(l, X) = P_{\text{all}}(T_c, l, X). \tag{3}$$

P_{err} is used for calculating the timing error rate, which will be explained in Sect. 3.3.

3.3 Modeling of the System

From now, we explain how to evaluate the timing error rate and the power dissipation of the adaptive speed control circuits with canary FF. Figure 2 shows the transition of the speed level. Once a warning signal is observed, l is incremented by one. When no warning signals are observed during the monitoring period, l is decremented by one.

The next speed level is determined by the present speed level and by the detection of the warning signal. This means that the speed level transition satisfies Markov property. Then, transition matrix $\mathbf{P}$ can be expressed as

$$\mathbf{P} = \begin{bmatrix} P_{\text{d}}(l_{\max}) & 1 - P_{\text{d}}(l_{\max}) & 0 & \cdots \\ P_{\text{d}}(l_{\max} - 1) & 0 & 1 - P_{\text{d}}(l_{\max} - 1) & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots \end{bmatrix}$$

$$\begin{bmatrix} \cdots & 0 & 0 & 0 \\ \cdots & 0 & 0 & 0 \\ \ddots & \vdots & \vdots & \vdots \\ \cdots & P_{\text{d}}(l_{\min} + 1) & 0 & 1 - P_{\text{d}}(l_{\min} + 1) \\ \cdots & 0 & P_{\text{d}}(l_{\min}) & 1 - P_{\text{d}}(l_{\min}) \end{bmatrix}, \tag{4}$$

where the ith row and column of $\mathbf{P}$ correspond to speed level $l_{\max} - i + 1$.

Let $\pi(n)$ be a state probability distribution vector in n-th time step,

$$\pi(n) = \pi(n - 1)\mathbf{P}. \tag{5}$$

We define π^∞ as a steady state distribution obtained by $n \to \infty$ and define $\pi_l(X)$ as a steady state probability of being at speed level l at condition X. π^∞ can be expressed as

$$\pi^\infty = \pi^\infty \mathbf{P}, \tag{6}$$

where

$$\pi^\infty = \begin{bmatrix} \pi_{l_{\max}}(X) & \pi_{l_{\max}-1}(X) & \cdots \pi_{l_{\min}}(X) \end{bmatrix}. \tag{7}$$

π^∞ can be obtained with (6) and the relation below.

$$\sum_{j=l_{\min}}^{l_{\max}} \pi_j(X) = 1. \tag{8}$$

Periods (# cycles) of being at a certain speed level are not always equal because the speed level changes immediately once a warning signal is observed. This means that π_l is not directly related to actual time. Hence, the duration at each level, which is suitable to evaluate the timing error rate and the power dissipation, must be computed from π_l. We here introduce $N_{rem}(l, X)$, which means the average cycle of a single stay at level l. $N_{rem}(l, X)$ is represented in the number of cycles, and it can be expressed as

$$\begin{aligned} N_{rem}(l, X) \;=\; & P_w(l, X) \sum_{j=0}^{N_{mon}-1} (j + 1)(1 - P_w(l, X))^j \\ & + N_{mon} \cdot (1 - P_w(l, X))^{N_{mon}}. \end{aligned} \tag{9}$$

The probability that no timing errors occur during j cycle(s) and a timing error occurs just at the $(j + 1)$-th cycle can be expressed as $P_w(l, X)(1 - P_w(l, X))^j$. Hence, the first term in (9) represents the average cycle of a single stay at the speed level l in case that at least one timing error occurs during the monitoring period N_{mon}. The second term in (9) corresponds to the case that no timing errors occur during N_{mon}.

In order to evaluate the timing error rate and the power dissipation, we define $P_{\text{time}}(l, X)$ as a time-based probability of being at speed level l at condition X. It can be expressed as

$$P_{\text{time}}(l, X) = \frac{N_{rem}(l, X) \cdot \pi_l(X)}{\sum_{j=l_{\min}}^{l_{\max}} N_{rem}(j, X) \cdot \pi_j(X)}, \tag{10}$$

where N_{rem} represents the average cycle of a single stay at

the speed level l, and π_l represents the state probability of being at l. Thus, the numerator in (10) is the average stay time (in the number of cycles) of being at l. The denominator is the sum of the average stay time of being at all speed levels.

The expected power dissipation of the system with canary FF is

$$P_{\text{ow, avg}}(X) = \sum_{j=l_{\min}}^{l_{\max}} P_{\text{ow}}(j, X) \cdot P_{\text{time}}(j, X), \tag{11}$$

where $P_{\text{ow}}(l, X)$ is the power dissipation at speed level l and condition X.

As a metric of timing error rate, we introduce average interval between timing errors, N_{err}, which is similarly defined to MTBF (Mean Time Between Failure). MTBF is defined as

$$\text{MTBF} = \frac{\text{Operating time}}{\text{Number of failures}}. \tag{12}$$

According to the MTBF definition, N_{err} can be expressed as

$$N_{err}(X) = \frac{\sum_{j=l_{\min}}^{l_{\max}} N_{rem}(j, X) \cdot \pi_j(X)}{\sum_{j=l_{\min}}^{l_{\max}} N_{rem}(j, X) \cdot \pi_j(X) \cdot P_{\text{err}}(j, X)}, \tag{13}$$

where P_{err} is a probability that timing errors occur in a cycle (See Eq. (3)). The numerator in (13) represents an expected stay time in a state and the denominator is the number of the timing errors occurring during the time.

From the above discussion, we can calculate average power dissipation $P_{\text{ow, avg}}$ and average interval between timing errors N_{err} from given path activation probabilities P_i, P_{all} and power dissipation P_{ow} at each speed level and condition.

4. Experimental Results

This section experimentally validates the discussion in Sect. 3. We use a 32-bit ripple carry adder (RCA) and a 32-bit Kogge-Stone adder (KSA) in subthreshold operation in a 90 nm CMOS process for experiments. We denote the outputs of RCA and KSA by S[0] – S[32], where S[32] is the most significant bit. The adders operate at V_{DD} = 300 mV and the speed control is implemented by body-biasing. Speed level $l = 0$ indicates zero body-bias, and both forward and reverse biasing are considered. The performance of a subthreshold circuit is sensitive to temperature, and we focus on the adaptive speed control for temperature in this experiment. From now, we use temperature $Temp$ as condition X described in Sect. 3, and we consider a temperature variation from 0℃ to 80℃. The overhead in time and energy to change the speed level is not considered for simplicity.

4.1 Evaluation of RCA

First we will evaluate a trade-off between the timing error rate and power dissipation of the RCA. We assume that the RCA consists of series-connected 32 full adders (FAs) and the clock period T_c is 100 ns (10 MHz) in this experiment. The RCA contains 96 cells and the critical path consists of 33 cells (including 31 complex gates).

4.1.1 Model of P_i, P_{all} and P_{ow}

The analysis of the timing error occurrence in Sect. 3 requires P_i, P_{all} and P_{ow}, and we here assume that they are given as closed-form expressions below. The expressions are derived by numerical fitting based on circuit simulations with a 90 nm CMOS technology. As for the appropriateness of the expressions, please see Appendix. Note that P_i, P_{all} and P_{ow} do not have to be expressed analytically and/or continuously, and the analysis can be carried out with histograms or piece-wise linear expressions.

$$P_i(t, l, Temp) = \begin{cases} \left(\frac{1}{2}\right)^{\frac{t}{D_c(l,Temp)}} & (t \le i \cdot D_c) \\ 0 & (t > i \cdot D_c), \end{cases} \tag{14}$$

$$P_{\text{all}}(t, l, Temp) = \begin{cases} (32 - \frac{t}{D_c(l,Temp)}) \\ \quad \times P_{32}(t, l, Temp) & (t \le 31 D_c) \\ P_{32}(t, l, Temp) & (t > 31 D_c) \end{cases}, \tag{15}$$

where D_c is the delay from carry-in to carry-out of a single FA. D_c and buffer delay D_d are dependent on the speed level and the temperature as well.

$$D_c(l, Temp) = D_{c0} \cdot \gamma^{\,l} \cdot 0.85^{\frac{Temp-25}{10}}, \tag{16}$$

$$D_d(l, Temp) = D_{d0} \cdot \gamma^{\,l} \cdot 0.85^{\frac{Temp-25}{10}}, \tag{17}$$

where D_{c0} and D_{d0} are the delays at $l = 0$ and $Temp = 25$℃. γ means the delay become γ (< 1) times shorter when speed level l is incremented by one.

The power dissipation of the RCA, $P_{\text{ow,rca}}(l, Temp)$, is

$$P_{\text{ow, rca}}(l, Temp) = P_{\text{ow0, rca}} \times 0.5 \times (1 + 1.35^{\frac{Temp-25}{10}} \cdot \beta^{l}), \tag{18}$$

where $P_{\text{ow0, rca}}$ is the power dissipation at $l = 0$ and $Temp = 25$℃. β means that the power dissipation become β (> 1) times higher when speed level l is incremented by one. The power dissipated by the delay buffer is assumed to linearly increase according to the delay time. When the delay is 1 ns at $l = 0$ and $Temp = 25$℃, the power overhead is 0.2% of the RCA. Thus, P_{ow} is expressed as

$$P_{\text{ow}}(l, Temp) = P_{\text{ow, rca}}(l, Temp) \times (1 + D_{d0} \times 10^9 \times 0.002). \tag{19}$$

4.1.2 Evaluation Setup

We show a trade-off between the average interval between the timing errors $N_{err}(Temp)$ and the power dissipation $P_{\text{ow,avg}}(Temp)$, and reveal the dependency of the trade-off

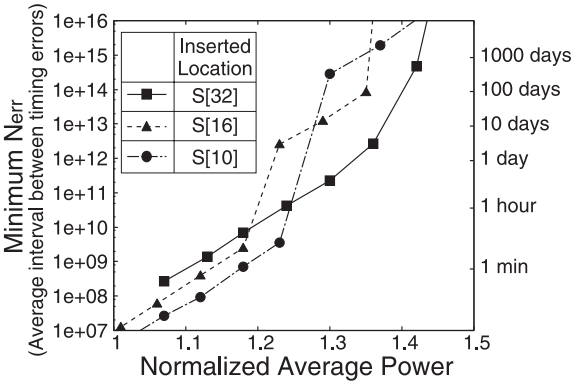


Fig. 3 Ave($P_{\mathrm{ow,avg}}$) versus min(N_{err}) with various buffer delays D_{d0}. Each dot corresponds to different configuration of buffer delay (N_{mon} = 10^8, $\gamma = 0.85$, $\beta = 1.25$). Ave($P_{\mathrm{ow,avg}}$) is normalized by that at $l = 0$ and $Temp = 25$℃.

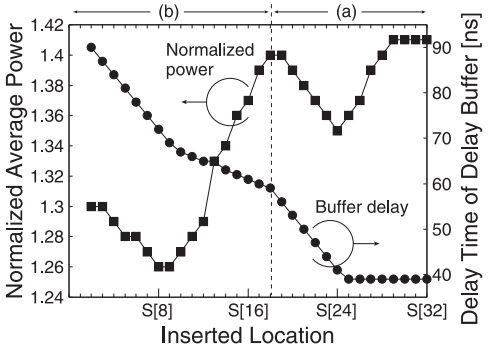
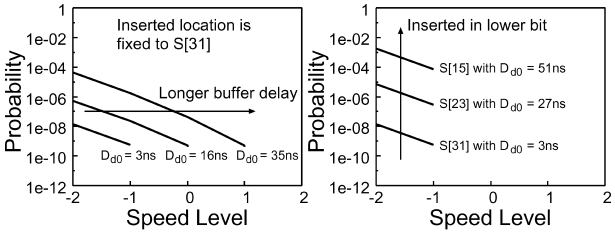


Fig. 4 Minimum buffer delay D_{d0} and ave($P_{\mathrm{ow,avg}}$). A constraint that min(N_{err}) must be larger than 10^{14} cycles is given. Ave($P_{\mathrm{ow,avg}}$) is normalized by that at $l = 0$ and $Temp = 25$℃.



(a) P_w v.s. buffer delay (b) P_w v.s. inserted location

Fig. 5 The probability of warning occurrence P_w as functions of the speed level ($\gamma = 0.85$, $Temp = 30$℃).

on design parameters. Both $N_{err}(Temp)$ and $P_{\mathrm{ow,avg}}(Temp)$ vary depending on the temperature. To conservatively evaluate the error rate, we sweep the temperature from 0℃ to 80℃ by 1℃, and the worst N_{err}, i.e. min(N_{err}), is evaluated. As for power dissipation, we evaluate the average of $P_{\mathrm{ow,avg}}$ from 0℃ to 80℃, ave($P_{\mathrm{ow,\,avg}}$).

We evaluate the dependency of the trade-off between min(N_{err}) and ave($P_{\mathrm{ow,avg}}$) on design parameters; canary FF position i, buffer delay D_{d0}, monitoring period N_{mon}, and speed control fineness γ and β.

Larger N_{mon} could deteriorate the adjustment response to the temperature change, whereas the timing error is less likely to happen. In this paper, considering the speed of temperature change, we choose N_{mon} from 10^7 cycles (1 second) to 10^9 cycles (100 seconds). As for the resolution of speed control, we use two parameter sets $\gamma = 0.85$, $\beta = 1.25$ (equivalent to 50 mV step body-biasing) and $\gamma = 0.96$, $\beta = 1.06$ (equivalent to 25 mV step body-biasing), where γ and β that are closer to 1 mean finer speed control.

4.1.3 Results and Discussions

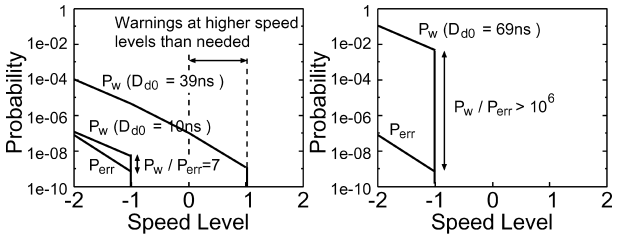
Figure 3 shows the relation between ave($P_{\mathrm{ow,\,avg}}$) and min(N_{err}) when $\gamma = 0.85$, $\beta = 1.25$, and $N_{mon} = 10^8$ cycles. At each canary FF position, we changed buffer delay D_{d0} with 5 ns step, and evaluated ave($P_{\mathrm{ow,\,avg}}$) and min(N_{err}). The Y axis on the right side indicates the actual time which is computed from min(N_{err}) represented in Y axis on the left side. For example, 10^9 cycles in the Y axis on the left side indicates 100 seconds in the Y axis on the right side at 10 MHz (1 cycle = 100 ns) operation. The power dissipation is normalized by that at $l = 0$ and $Temp = 25$℃. Figure 3 indicates that inserted location S[i] and buffer delay D_{d0} affect min(N_{err}) significantly, which means the optimal design parameters vary depending on the required error rate.

Longer N_{err} means that the timing error rate is lower. Figure 3 shows that larger power dissipation is required if the timing error rate is kept lower, that is N_{err} is kept larger, whereas smaller power dissipation is demanded if higher timing error rate, i.e. smaller N_{err} is acceptable. This relation indicates that there is a trade-off between the timing error rate and power dissipation.

Figure 4 shows buffer delay D_{d0} and power dissipation ave($P_{\mathrm{ow,avg}}$) when a canary FF is inserted at S[2] to S[32]. Minimum buffer delay that makes min(N_{err}) larger than 10^{14} cycles is computed at each inserted location. In this case, S[8] and S[9] archive the minimum power dissipation.

Let us explain why the most power-efficient location is in lower bits using an example. We first review the dependence of P_w on the buffer delay and the inserted location of canary FF. Figure 5 shows P_w at each speed level l when $Temp = 30$℃. Figure 5(a) indicates that by increasing the buffer delay with the fixed inserted location of canary FF, the probability of warning occurrence can be increased, but warnings can occur at higher speed level at the same time. On the other hand, by inserting canary FF in the lower bits with the appropriate buffer delay (Fig. 5(b)), the probability of warning occurrence can be increased without warning occurrence at higher speed level because the critical paths to the lower bits are more likely activated.

Figure 6 shows P_w and P_{err} at each speed level l when $Temp = 30$℃. In this example, the timing error probability at $l = 0$ is zero, and hence it is appropriate to assign the speed level to 0. On the other hand, a timing error hardly occurs at speed level $l = -1$ as long as the probability of warning occurrence is much higher than the error occurrence probability, i.e. $P_w(-1, Temp) \gg P_{\mathrm{err}}(-1, Temp)$. In this case, it is acceptable to change the speed level to -1,



(a) Inserted location: S[32] (b) Inserted location: S[9]

Fig. 6 The probability of warning occurrence P_w and the error occurrence probability $P_{\rm err}$ (γ = 0.85, $Temp$ = 30℃).

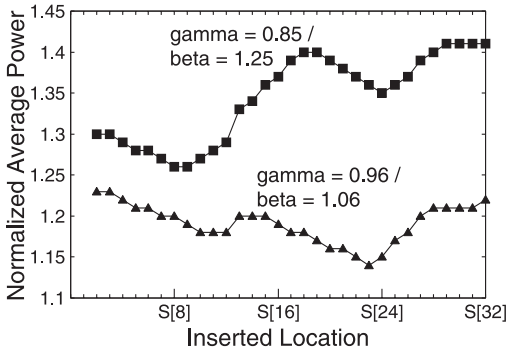


Fig. 7 Dependency of ave($P_{\rm ow,avg}$) on speed control fineness γ (min(N_{err}) > 10^{14}, N_{mon} = 10^8 cycles). Ave($P_{\rm ow,avg}$) is normalized by that at l = 0 and $Temp$ = 25℃.

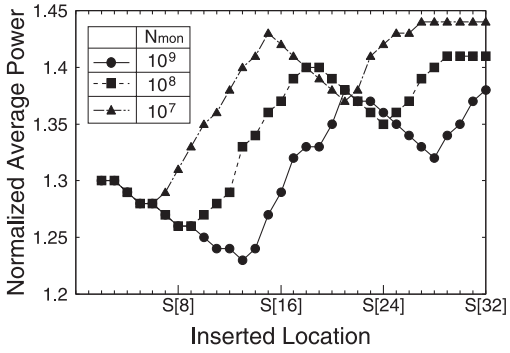


Fig. 8 Dependency of ave($P_{\rm ow,avg}$) on monitoring period N_{mon} (min(N_{err})> 10^{14}, γ = 0.85, β = 1.25). Ave($P_{\rm ow,avg}$) is normalized by that at l = 0 and $Temp$ = 25℃.

which enables further power reduction while keeping the average interval between timing errors N_{err} high enough.

Suppose a canary FF is inserted to S[32], which is the output of the critical path (Fig. 6(a)). In this case, longer buffer delay is required to maintain the high ratio of P_w to $P_{\rm err}$ at speed level $l = -1$. For example, when buffer delay D_{d0} is 10 ns, P_w becomes zero at $l = 0$ and speed level l can be -1. However, the ratio of P_w to $P_{\rm err}$ at speed level $l = -1$ is small and is 7. Thus, N_{err} is only 8×10^9 cycles. When we increase buffer delay D_{d0} to 39 ns, N_{err} can be increased to above 10^{14} cycles. However, in this case, warning signals are generated at $l = 0$ and $l = 1$, which means the speed level l can be incremented to 2. The RCA likely operates at higher speed levels (1 and 2) than needed (0), which results in increase in power dissipation.

When a canary FF is inserted at S[9] (Fig. 6(b)), which is the optimum location obtained in Fig. 4, the speed level is mostly controlled to -1 and 0, and the upper levels are never used, because P_w is zero at $l = 0$. An important point is that the ratio of P_w to $P_{\rm err}$ at $l = -1$ is very high and is $> 10^6$. Thanks to this high ratio, N_{err} becomes larger than 10^{14} cycles.

When a canary FF is inserted in upper bits, the speed level tends to be higher than needed, because the probability of warning occurrence becomes non-zero at the higher speed level. Thus, the power dissipation increases in region (a) in Fig. 4. On the other hand, when a canary FF is inserted in lower bits ((b) in Fig. 4), the speed level is controlled more appropriately, and the power dissipation decreases. With the power increase due to longer buffer delay, the power becomes minimum at S[9] in this case. Consequently, the most power-efficient insertion location is not the output of the critical path, but lower output bits. In addition, the power dissipation increases in S[17]–S[24] in Fig. 4. This is because longer buffer delay involves larger power dissipation of the buffer.

Figure 7 shows the dependency on design parameters of speed control fineness γ and β. When γ and β are closer to 1, the speed can be controlled finely and the power dissipation decreases. Compared to γ = 0.85, the optimum inserted location moves to the upper bit in the case of γ = 0.96, because the power penalty by being controlled to the higher speed level than needed is small.

Figure 8 shows the dependency on monitoring period N_{mon}. N_{mon} is changed to 10^7 cycles (1 second), 10^8 cycles (10 seconds) and 10^9 cycles (100 seconds), and minimum buffer delay that makes min(N_{err}) larger than 10^{14} cycles is derived. The total power dissipation ave($P_{\rm ow,\,avg}$) is shown in Fig. 8. Figure 8 indicates that the power dissipation can be reduced by lengthening N_{mon}. This is because the longer N_{mon} is, the smaller the possibility that no waning signals are generated during the monitoring period is. On the other hand, too large N_{mon} could deteriorate the adjustment response to the temperature change, which is not shown in Fig. 8.

Figure 9 shows trade-off relations between ave($P_{\rm ow,avg}$) and min(N_{err}) with the following two cases – 1) the buffer delay and the inserted position are freely selected such that the power dissipation is minimized, 2) inserted position is fixed to S[32] which is the output bit of the critical path. We can see that the power dissipation can be reduced by optimally selecting the inserted position as well as the buffer delay. Assuming that a constraint of min(N_{err})> 10^{14} is given, inserting a canary FF at S[13] and adjusting the buffer delay reduce the power dissipation by 10% in comparison to inserting canary FF at S[32] on the critical path fixedly. In this example, we can see that the inserted location affects min(N_{err}) exponentially.

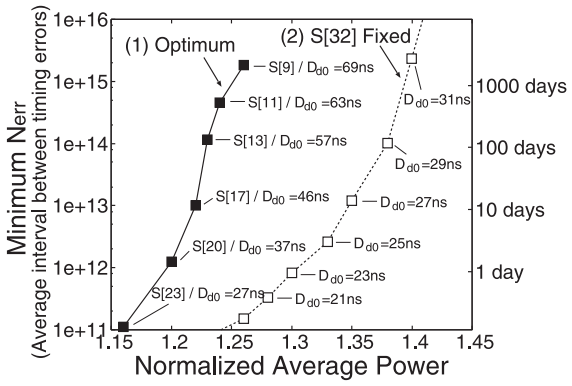


Fig. 9 Comparison between two cases; (1) both inserted location and buffer delay are optimized and (2) insertion location is fixed to S[32] (γ = 0.85, β = 1.25, $N_{mon} = 10^9$ cycles). Ave($P_{\mathrm{ow,avg}}$) is normalized by that at $l = 0$ and $Temp = 25$℃. Ave($P_{\mathrm{ow,avg}}$) is normalized by that at $l = 0$ and $Temp = 25$℃. The Y axis on the right side indicates the actual time which is computed from min(N_{err}) represented in Y axis on the left side.

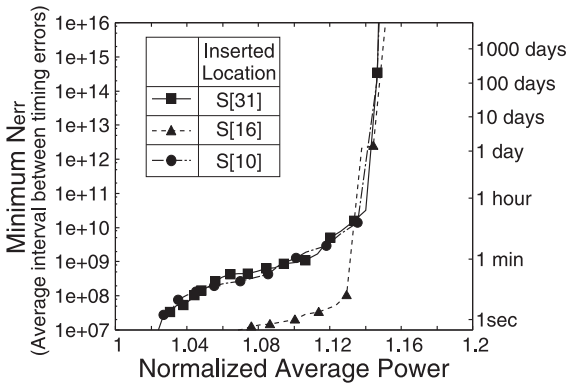


Fig. 10 Ave($P_{\mathrm{ow,avg}}$) versus min(N_{err}) with various buffer delays D_{d0}. Each dot corresponds to different configuration of buffer delay ($N_{mon} = 10^8$). Ave($P_{\mathrm{ow,avg}}$) is normalized by that at $l = 0$ and $Temp = 25$℃. The Y axis on the right side indicates the actual time which is computed from min(N_{err}) represented in Y axis on the left side.

4.2 Evaluation of KSA

Next we will evaluate a trade-off relation between the timing error rate and the power dissipation of the KSA.

4.2.1 Experimental Setup

In contrast to RCA, it is difficult to express the path activation probabilities P_i and P_{all} of the KSA as closed-form expressions. In this evaluation, we thus express P_i and P_{all} as histograms. Here they are derived with circuit simulations by the following procedure:

1. We perform circuit simulations with a substantial number of random input vectors (10^8 vectors in this experiment) and observe toggles at each output bit S[0] – S[32] and their delays from primary inputs.
2. $P_i(t)$ is derived by dividing the number of the toggles at S[i] whose delays are larger than t by the number of the input vectors. $P_{\mathrm{all}}(t)$ is derived as $P_{\mathrm{all}}(t) = P_0(t) \cup P_1(t) \cup \cdots \cup P_{32}(t)$.
3. We repeat procedures 1. and 2. at each speed level and temperature, and consequently $P_i(t, l, Temp)$ and $P_{\mathrm{all}}(t, l, Temp)$ can be derived.

In this experiment, we simplified procedure 3. We derive $P_i(t, l)$ and $P_{\mathrm{all}}(t, l)$ at each speed level l when temperature $Temp$ is 25℃, and we estimate P_i and P_{all} in other temperature conditions as follows:

$$P_i(t, l, Temp) = P_i\left(\frac{t_{c,25}}{t_{c,Temp}}t, l, Temp = 25℃\right), \quad (20)$$

$$P_{\mathrm{all}}(t, l, Temp) = P_{\mathrm{all}}\left(\frac{t_{c,25}}{t_{c,Temp}}t, l, Temp = 25℃\right), \quad (21)$$

where $t_{c,Temp}$ is the delay of the critical path at temperature $Temp$ and $t_{c,25}$ is the delay of the critical path at $Temp$ = 25℃.

We express buffer delay D_d and the power dissipation P_{ow} as closed-form expressions, which are derived by numerical fitting based on circuit simulations. We assume that the speed level is implemented by body-biasing and the fineness of the speed control is 100 mV step (for example, $l = 1$ means 100 mV forward body-bias). We also assume that the clock period T_c is 50 ns (20 MHz), and the monitoring period N_{mon} is 10^8 cycles (5 seconds). We evaluate min(N_{err}) and ave($P_{\mathrm{ow,avg}}$) in the temperature range from 0℃ to 80℃.

The KSA consists of 467 cells and the critical path is composed of 29 cells.

4.2.2 Results

Figure 10 shows the relation between ave($P_{\mathrm{ow,\,avg}}$) and min(N_{err}). At each canary FF position, we changed buffer delay D_{d0} with 1 ns step, and evaluated ave($P_{\mathrm{ow,\,avg}}$) and min(N_{err}). D_{d0} is the delay of the delay buffer of canary FF at $l = 0$ and $Temp = 25$℃. The Y axis on the right side indicates the actual time at 20 MHz (1 cycle = 50 ns) operation. The power dissipation is normalized by that at $l = 0$ and $Temp = 25$℃. This figure indicates that when min(N_{err}) ($> 10^{10}$ cycles) is required, the trade-off between the timing error rate and the power dissipation is less dependent on the location of Canary FF, whereas below 10^9 cycles, power dissipation can be reduced by choosing an appropriate location.

Figure 11 shows buffer delay D_{d0} and power dissipation ave($P_{\mathrm{ow,avg}}$) when a canary FF is inserted at S[4] to S[32]. Minimum buffer delay is computed such that the constraint min(N_{err}) $> 10^{14}$ cycles is satisfied at each inserted location. S[31] is the output bit of the critical path of the KSA and the buffer delay is the smallest when canary FF is inserted in S[31]. This figure shows that the power dissipation of the KSA is less sensitive to the location of Canary FF in comparison to that of the RCA (Fig. 4). We will discuss the reason below.

Figures 12 and 13 show P_w and P_{err} at each speed level l when $Temp$ = 30℃. Figure 12 depicts P_w and P_{err} in two

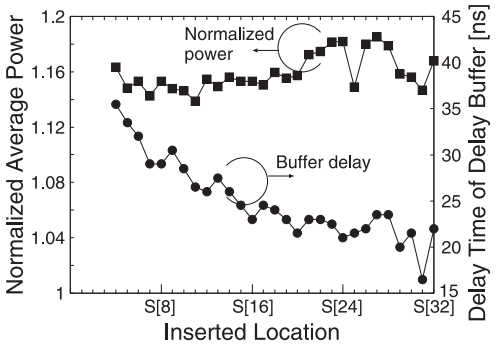
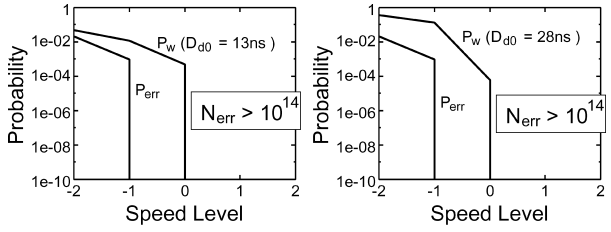
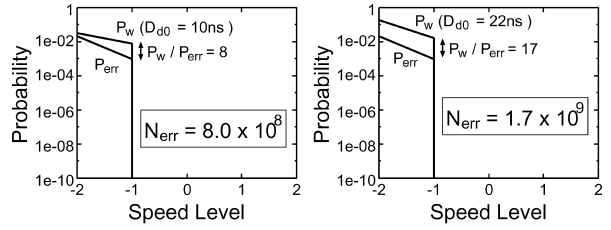


Fig. 11 Minimum buffer delay D_{d0} and ave($P_{\mathrm{ow,avg}}$). A constraint that min(N_{err}) must be larger than 10^{14} cycles is given. Ave($P_{\mathrm{ow,avg}}$) is normalized by that at $l = 0$ and $Temp = 25$℃.



(a) Inserted location: S[31] (b) Inserted location: S[9]

Fig. 12 The probability of warning occurrence P_w and the error occurrence probability P_{err} ($Temp = 30$℃). N_{err}s are larger than 10^{14} cycles in both inserted locations.



(a) Inserted location: S[31] (b) Inserted location: S[9]

Fig. 13 The probability of warning occurrence P_w and the error occurrence probability P_{err} with smaller buffer delays. ($Temp = 30$℃).

cases; $D_{d0} = 13$ ns at S[31] and $D_{d0} = 28$ ns at S[9]. Irrelevant to inserted locations, the KSA is mainly controlled to the speed level $l = 0$ and $l = 1$, and is hardly controlled to $l = -1$ where timing errors may occur. Consequently, N_{err} can be larger than 10^{14} cycles. If we decrease the buffer delay such that the KSA is controlled to $l = -1$ as described in Fig. 13, the power dissipation of the KSA can be reduced. However N_{err} is dropped sharply to 10^8–10^9 cycles (several seconds – several minutes). This is because P_{err} is much higher in comparison to RCA (Fig. 6) and it is difficult to maintain the high ratio of P_w to P_{err} at $l = -1$ even when canary FF is inserted at a lower output bit such as S[9]. In order to sustain $N_{err} > 10^{14}$ cycles, we need to set the buffer delay such that the KSA is hardly controlled to the speed level where timing errors may occur ($l = -1$ in this example) regardless of the inserted location. This makes the power dissipation of the KSA is less sensitive to the location of Canary FF when keeping $N_{err} > 10^{14}$ cycles.

We finally examine the reason why min(N_{err}) increases sharply near ave($P_{\mathrm{ow,\,avg}}$) = 1.14 in Fig. 10. The critical paths of the KSA are more likely activated than the RCA. Therefore when the KSA is controlled to the speed level where timing errors occur, N_{err} becomes small as mentioned above. For example, when canary FF is inserted at S[31] and the buffer delay D_{d0} is 10 ns (Fig. 13(a)), the KSA is mainly controlled to $l = -1$ because the warnings are not detected at $l \geq 0$. The error occurrence probability P_{err} at $l = -1$ is significantly high, hence N_{err} is small and is 8×10^8 cycles. On the other hand, when D_{d0} is 13 ns (Fig. 12(a)), the KSA is hardly controlled to $l = -1$ due to the high probability of warning occurrence at $l = 0$ and mainly controlled to $l = 0$ and $l = 1$ where no timing errors occur. Consequently N_{err} increases sharply to more than 10^{14} cycles. At this time the normalized power dissipation is 1.15. This is the reason why min(N_{err}) increases sharply near ave($P_{\mathrm{ow,\,avg}}$) = 1.14 in Fig. 10.

5. Conclusion

In this paper, we discussed how to evaluate the relation between the timing error rate and the power dissipation in self-adaptive circuits with timing error prediction. In the experiments using 32-bit adders in subthreshold operation, we demonstrated a trade-off between the timing error rate and the power dissipation. We also revealed that the trade-off depends on design parameters and the optimal design parameters vary depending on the required error rate and speed control fineness.

Acknowledgments

This work is supported in part by New Energy and Industrial Technology Development Organization (NEDO) and VLSI Design and Education Center (VDEC).

References

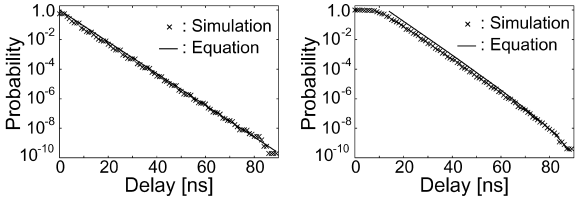
[1] J. Kao, M. Miyazaki, and A. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," IEEE J. Solid-State Circuits, vol.37, no.11, pp.1545–1554, Nov. 2002.
[2] S. Das, D. Roberts, L. Seokwoo, S. Pant, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "A self-tuning DVS processor using delay-error detection and correction," IEEE J. Solid-State Circuits, vol.41, no.4, pp.792–804, April 2006.
[3] D. Blaauw, S. Kalaiselvan, K. Lai, M. Wei-Hsiang, S. Pant, C. Tokunaga, S. Das, and D. Bull, "Razor II: In situ error detection and correction for PVT and SER tolerance," Int. Solid-State Circuits Conference Dig. Tech. Papers, pp.400–401, Feb. 2008.
[4] T. Sato and Y. Kunitake, "A simple flip-flop circuit for typical-case designs for DFM," Proc. Int. Symp. Quality Electronic Design, pp.539–544, March 2007.
[5] T. Nakura, K. Nose, and M. Mizuno, "Fine-grain redundant logic using defect-prediction flip-flops," Int. Solid-State Circuits Conference Dig. Tech. Papers, pp.402–403, Feb. 2007.
[6] B.H. Calhoun and A. Chandrakasan "Standby power reduction using dynamic voltage scaling and canary flip-flop structures," IEEE J.

Solid-State Circuits, vol.39, no.9, pp.1504–1511, Sept. 2004.

[7] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," IEEE J. Solid-State Circuits, vol.37, no.11, pp.1396–1402, Nov. 2002.

Appendix

This appendix validates the closed-form expressions used in the experiments (Eqs. (14) to (18)). We first verify $P_{\rm i}$ and $P_{\rm all}$. The points plotted in Fig. A· 1 correspond the probabilities obtained by logic simulation when five billions of random vectors are given. The lines of Eqs. (14) and (15) are well correlated with the simulation results. Figure A· 2 shows the delay and power dissipation when the speed level and temperature are changed. The open symbols are circuit simulation results, and the closed symbols correspond to Eqs. (16) and (18). At each temperature and speed level, the error is acceptable.



(a) $P_{32}(t, l = 0, Temp = 25\ ^\circ\mathrm{C})$ (b) $P_{\rm all}(t, l = 0, Temp = 25\ ^\circ\mathrm{C})$

Fig. A· 1 Correlation between simulation results and equations of $P_{\rm i}$ and $P_{\rm all}$ ($D_c(l = 0, Temp = 25^\circ\mathrm{C}) = 2.8$ ns).

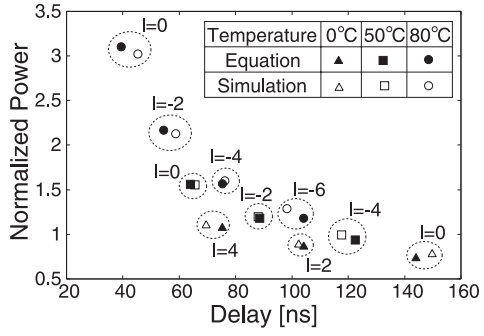


Fig. A· 2 Delay and power dissipation of RCA at various speed levels and temperature ($\gamma = 0.85, \beta = 1.25$).

Hiroshi Fuketa received the B.E. degree from Kyoto University in 2002 and the M.E. degree in information systems engineering from Osaka University in 2008. He is currently a doctoral student of the Graduate School of Information Science and Technology, Osaka University. His research interests include ultra-low-power circuit design and variation modeling. He is a student member of IEEE.

Masanori Hashimoto received the B.E., M.E., and Ph.D. degrees in communications and computer engineering from Kyoto University, Kyoto, Japan, in 1997, 1999, and 2001, respectively. Since 2004, he has been an Associate Professor in the Department of Information Systems Engineering, Osaka University, Osaka, Japan. His research interests include computer-aided-design for digital integrated circuits, and high-speed circuit design. Dr. Hashimoto served on the technical program committees for international conferences including DAC, ICCAD, ASP-DAC, ICCD and ISQED. He received the Best Paper Award at ASP-DAC 2004. He is a member of IEEE and IPSJ.

Yukio Mitsuyama received B.E. and M.E. degrees in Information Systems Engineering from Osaka University, Japan, in 1998 and 2000, respectively. He is currently an assistant professor in Graduate School of Engineering, Osaka University. His research interests include reconfigurable architecture and its VLSI design. He is a member of IEEE and IPSJ.

Takao Onoye received B.E. and M.E. degrees in Electronic Engineering, and Dr.Eng. degree in Information Systems Engineering all from Osaka University, Japan, in 1991, 1993, and 1997, respectively. He is currently a professor in the Department of Information Systems Engineering, Osaka University. His research interests include media-centric low-power architecture and its SoC implementation. He is a member of IEEE, IPSJ, and ITE-J.