2014 IEEE 23rd Asian Test Symposium

Opportunities and Verification Challenges of Run-time Performance Adaptation

Masanori Hashimoto
Dept. Information Systems Engineering, Osaka University
Email: hasimoto@ist.osaka-u.ac.jp

***Abstract*—Run-time performance adaptation with field delay testing is a promising approach for minimizing design margin while sustaining necessary operational margin in the field. However, run-time performance adaptation has not been adopted in industrial designs since a serious concern on timing error occurrence exists. For putting the run-time performance adaptation in a practical use, we need to verify and optimize the run-time adaptation system in design time, but a straightforward verification with logic simulation could need billion years and is totally insufficient. For this problem, we have developed a stochastic framework for error rate estimation that models adaptive speed control as a continuous-time Markov process. This paper first exemplifies the power reduction thanks to run-time performance adaptation with a 65nm test chip. Then, the proposed stochastic framework is introduced. With this framework, we evaluate MTTF of an embedded processor whose performance is adaptively controlled with online testing and offline testing. This evaluation shows how design parameters affect MTTF as an example.**

***Keywords*-run-time performance adaptation; online test; offline test; error rate; MTTF**

I. Introduction

Circuit delay fluctuation due to PVT (process, voltage and temperature) variation is becoming more and more significant. In addition, unexpected timing error can occur in field due to aging effects, such as NBTI (negative bias temperature instability), HCI (hot carrier injection) and TDDB (time dependent dielectric breakdown). To avoid timing errors, circuits are usually designed with guard-banding. However, large timing margin makes timing closure difficult and involves an increase in area and power dissipation. Moreover, power supply voltage which is higher than the necessary and sufficient voltage becomes necessary, which results in wasteful power dissipation.

To overcome this problem, adaptive speed control system has been studied in which each chip self-adjusts its operating condition, such as supply voltage and body bias. Traditionally, replica circuits have been used for performance monitoring. Adaptive control techniques with a critical path replica have been presented in [1]–[3]. However, the critical path replica is losing its effectiveness, since the delay mismatch between the replica and the actual critical path is increasing due to within-die variation and aging. Consequently, performance adaptation methods recently proposed are accompanied with timing self-test [4], [5]. Generally, adaptive speed control is performed so that no paths have timing violations. On the other hand, voltage over-scaling, which accepts rare timing errors for pursuing aggressive power reduction, is also studied [6], [7]. Path activation probability heavily depends on the running program on a processor, and in some cases, significant power reduction can be achieved by exploiting this property.

To adopt run-time adaptive speed control, each circuit needs to regularly perform online or offline test to check whether the current circuit performance satisfies the speed specification. On the other hand, run-time adaptive speed control cannot completely eliminate timing errors due to unexpected delay increase, which is applicable to not only voltage over-scaling but also ordinary voltage scaling. In addition, a biased circuit operation might mislead speed control in case of online test, and limited number of test patterns for offline test could miss timing errors depending on manufacturing variability and aging progress. Meanwhile, the occurrence frequency of timing errors can be changed by design parameter modification, and long MTTF (mean time to failure), such as ten years, is supposed to be obtained via parameter optimization. However, this timing error occurrence is very difficult to evaluate in design time, since simulation is too slow for rare errors. For enabling such design parameter optimization, we have developed a stochastic error rate estimation method [11]. The necessary computation time was reduced by twelve orders of magnitude, which can guide design optimization of run-time adaptive system.

This paper exemplifies how much power reduction can be obtained by reducing margin for manufacturing and environmental variation through run-time adaptation with silicon measurement results. Next, a serious design issue that prevents run-time performance adaptation from being adopted for a practical use is explained. A challenge is how we verify the run-time performance adaptation system in design time. We then introduce our stochastic error rate estimation method and give experimental results showing how MTTF depends on design parameters.

II. Online Test Based Run-time Adaptation

This paper discusses two types of run-time performance adaptation; online test based run-time adaptation and offline test based run-time adaptation. This section describes online test based run-time adaptation. Offline test based run-time adaptation will be explained in the next section.

1081-7735/14 $31.00 © 2014 IEEE
DOI 10.1109/ATS.2014.53

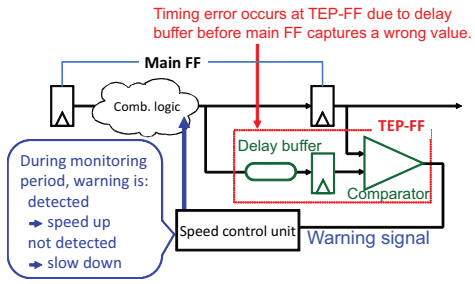

Figure 1. Run-time adaptive speed control with TEP-FF.

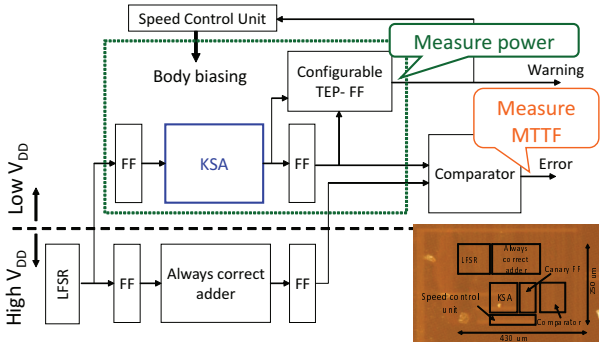

Figure 2. Block diagram of test circuit. 32-bit Kogge-Stone adder (KSA) is controlled adaptively with configurable TEP-FF.

Online test based run-time adaptation includes two approaches; error detection based approach and error prediction based approach. "Razor I" [6] and "Razor II" [7] detect timing errors in actual paths and correct the errors. In contrast, references [8]–[10] presented an error predictive sensor embedded into actual paths. This sensor cannot detect timing errors but predict them. Error prediction approach has two distinct advantages; it does not require error recovery system and then it can be applied to any sequential circuits, and it does not involve short-path problem. In the following, error prediction based approach is mainly discussed.

Figure 1 shows a circuit that adaptively controls the speed and power dissipation using a warning signal generated by a timing-error predictive (TEP) FF [8]. The TEP-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 TEP-FF before the main FF captures a wrong value due to the delay buffer, which enables us to know 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. Note that timing errors are predicted, not detected, which is a distinct difference from Razor [6]. Once a warning signal is observed, the circuit is controlled to speed up, in other words, the circuit delay is reduced by voltage scaling and/or body biasing. Clock frequency is supposed to be fixed throughout this paper. If no warning signals are observed during the monitoring period, the circuit is slowed down for power reduction. This proactive speed control overcomes the variation of the timing margin which is different chip by chip and varies depending on operating condition and aging.

A. Silicon Results

We designed and fabricated a test circuit to validate the adaptive speed control with TEP-FF in a 65 nm CMOS process [8]. Measurement results are shown in this subsection. Here, the circuit was operated in subthreshold region, since the subthreshold operation is sensitive to PVT variation and resembles the circuit operation in the future.

The structure of the test circuit and the micrograph are shown in Fig. 2. A 32-bit Kogge-Stone adder (KSA) was adopted as a circuit whose performance was controlled adaptively. S[32]-S[0] denote the outputs of the KSA, and S[32] is the most significant bit.

Input patterns are generated by a linear feedback shift register (LFSR). The KSA outputs are compared to the answer to check if a timing error occurs. The answer is generated by "always correct" adder operating at higher supply voltage.

The speed control unit alters by body-biasing the speed of the KSA, main FFs and TEP-FFs at inputs and outputs of the KSA. Four speed levels can be provided by applying four pairs of body-bias voltage. The body voltages are selected according to the speed level stored in a two-bit register. When the timer signal is asserted, the speed control unit immediately decrements the speed level by one and the circuit is controlled to slow down. In contrast, when the warning signal is asserted, the speed control unit increments the speed level by one.

1) Adaptive Compensation of Environmental Variability: Figure 3 shows the power dissipation at various temperature conditions (25–70 °C) when the operation frequency was set to 3 MHz in the following cases;

(a): the circuit was controlled adaptively with a TEP-FF,
(b): 200-mV FBB, which was the minimum body-bias for a 3-MHz operation at 25 °C, was fixedly applied,
(c): the minimum FBB voltage required for a 3-MHz operation at each temperature was applied.

In (a), a TEP-FF at S[20] was enabled. The power dissipation includes those of the KSA, main FFs, speed control unit, and TEP-FF. The power overhead of the TEP-FF was estimated to be around 2% by circuit simulation.

Figure 3 indicates that the power dissipation of (a) is very close to that of (c), which means optimal body-bias voltages were selected adaptively at each temperature. On the other hand, when the 200-mV FBB was fixedly applied ((b)), the power dissipation at 70 °C was much larger. Compared to ((b)), adaptive speed control of (a) reduced power dissipation by 39%.

This result indicates that the adaptive speed control with TEP-FF can well compensate delay fluctuation due to temperature shift.

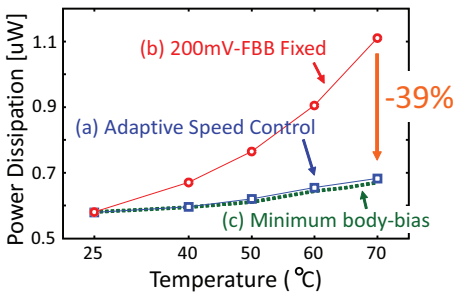

Figure 3. Power dissipation at the various temperature conditions (3 MHz @ $V_{DD} = 0.35$ V). Circuit operates (a) adaptively, (b) with 200-mV FBB fixedly, and (c) with minimum body-bias required for 3-MHz operation at each temperature.

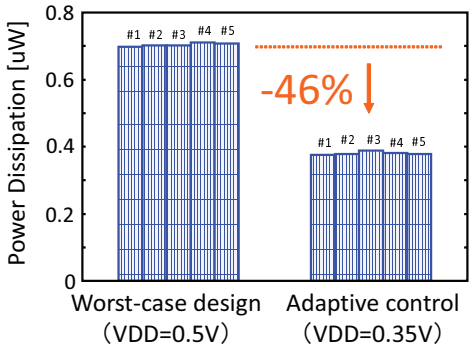

Figure 4. Power dissipation when operation frequency is 2 MHz in the following cases; (worst-case design) all chips operate at $V_{DD} = 0.5$ V, (adaptive speed control) all chips operate with adaptive control at $V_{DD} = 0.35$ V.

2) Comparison to Operation Considering Worst-case: We next demonstrate how inefficient the worst-case design for process variation is for subthreshold circuits, and clarify how beneficial the adaptive performance control is.

We here discuss the worst-case design in terms of manufacturing variability. Assuming 2-MHz operation, the supply voltage must be 0.5 V or higher for a chip at the SS device corner, for example. In this case, all chips should operate at $V_{DD} = 0.5$ V when the traditional worst-case design with guardbanding is adopted. Figure 4 shows the power dissipation of five chips in the following cases;

worst-case design: all chips operated at V_{DD}=0.5 V, which was the minimum V_{DD} for a chip at the SS device corner,

adaptive control: all chips operated with adaptive control at $V_{DD} = 0.35$ V.

One TEP-FF was enabled. The power dissipation with the adaptive control was smaller than that with guardbanding (worst-case design) by 46%, because of lower supply voltage. Breaking away from "worst-case design" halved power dissipation thanks to reduced margin for manufacturing variability.

B. Problem

In the previous subsection, the advantage of the run-time performance adaptation was demonstrated. However, there is a serious design concern that prevents the run-time adaptation from being used for a practical use. This issue is that a timing error could occur depending on the circuit operation. Even when the TEP-FF is well configured to generate the warning signal, the error occurrence cannot 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 excessively. If a critical path is activated in this condition, a timing error happens.

In applying the run-time adaptive speed control with TEP-FF to a circuit, there are following four major design parameters to control the rate of timing errors, power dissipation, area overhead and response speed to temporal fluctuation.

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

We can easily understand that longer buffer delay time reduces the number of timing errors but increases power dissipation, because the circuit tends to be speeded up. As the monitoring period becomes longer, the number of timing errors decreases, but the response to temporal environmental fluctuation degrades. Finer speed control decreases timing error, but it requires larger implementation overhead. As for the location, we intuitively think that the critical path is the best position. However, it is not true, because the probability of the path activation is significantly influential on timing error in addition to the path delay, which will be shown in the following. The number of TEP-FFs trades the rate of timing errors and area overhead.

Let us show some examples of analyzed results [12]. For experiments, we used a 32-bit ripple carry adder (RCA)in subthreshold operation in a 90nm CMOS process. The outputs of RCA are denoted 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. The performance of a subthreshold circuit is sensitive to temperature, and we here focus on the adaptive speed control for temperature (0 ˚C to 80 ˚C).

Figure 5 shows an example of the relation between average power dissipation and mean time to failures (MTTF). Here, a TEP-FF is inserted to S[32], S[16] or S[10] and its buffer delay is changed. The Y axis on the right side indicates the actual time which is computed from MTTF assuming 10MHz operation. Figure 5 indicates that inserted location S[i] and buffer delay affect MTTF significantly, which means the optimal design parameters vary depending on the required error rate.

Longer MTTF means that the timing error rate is lower. Figure 5 shows that larger power dissipation is required

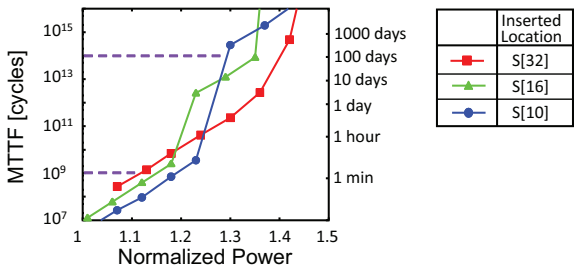

Figure 5. Average power dissipation versus mean time to failures (MTTF) with various buffer delays in RCA. Each dot corresponds to different configuration of buffer delay. Monitoring period is 10^8 cycles.

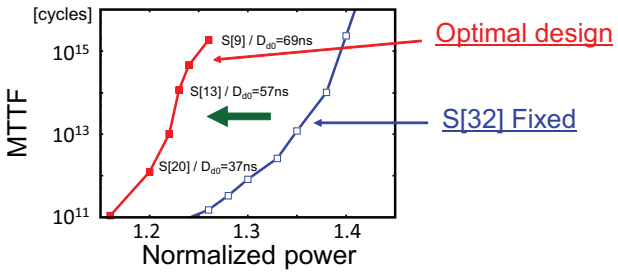

Figure 6. Comparison between two cases in RCA; (1) both inserted location and buffer delay are optimized and (2) insertion location is fixed to S[32]. Monitoring period is 10^9 cycles.

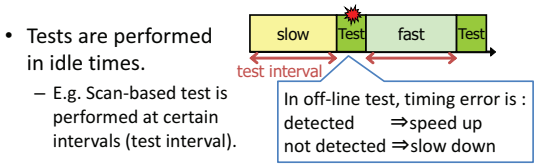

Figure 7. Run-time adaptive speed control with offline test.

if the timing error rate is kept lower, that is MTTF is kept larger, whereas smaller power dissipation is possible if higher timing error rate, i.e. smaller MTTF is acceptable. This relation indicates that there is a trade-off between the timing error rate and power dissipation.

Figure 6 shows trade-off relations between average power dissipation and MTTF in 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 MTTF $> 10^{14}$ is given, inserting a TEP-FF at S[13] and adjusting the buffer delay reduce the power dissipation by 10 % in comparison to inserting TEP-FF at S[32] on the critical path fixedly. In this example, we can see that the inserted location affects MTTF by more than four orders of magnitude.

Let us explain why the most power-efficient location is in lower bits. In RCA, the critical path S[32] is less probable to be activated, since the carry signal must be propagated through all the full adders. This means the probability of warning signal generation is very low, which often results in slowing down excessively. To prevent it, a longer buffer delay is necessary and it increases power dissipation due to circuit operation at higher speed level. On the other hand, by inserting TEP-FF in the lower bits with the appropriate buffer delay, 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.

III. Offline Test Based Run-time Adaptation

We next explain the second type of run-time adaptation, that is an adaptive speed control system that repeatedly performs delay test in idle times of the circuit (Fig. 7). While the circuit is idle, test patterns, which can be for scan test or SBST (software-based self-test), that were prepared beforehand and stored in an internal or external memory are loaded and it is checked if the circuit includes timing-violating paths or not. When a timing-violating path is detected, the minimum speed level that includes no timing-violating paths is selected for the operation in the following. Otherwise, the speed level is decremented. The scan test has higher freedom of applicable test patterns, and hence accurate error detection, in other words, lower missing rate of timing-violating paths can be expected.

Here, there are two strategies for test execution. One strategy forces the circuit to be idle with a fixed time interval, which can guarantee the time interval between the delay tests. This strategy is helpful to make the timing error rate predictable in addition to mitigating the error rate. A drawback is the performance degradation due to the test, and in some real-time systems, this strategy could be difficult to adopt. The other strategy is to perform offline tests only in true idle time. While the performance degradation does not arise, the test interval is less predictable and consequently the error rate tends to be higher.

In this offline test based adaptation, the test interval is a key parameter to determine the rate of timing error occurrence. If it is set to be long, the delay fluctuation in the duration of successive delay tests is likely to be large enough to cause timing violations. For mitigating the timing errors, the test interval should be short. On the other hand, frequent tests increase performance overhead and degrade the system throughput. To reduce the error occurrence while coping with the overhead, we need to carefully tune the test interval and the number of test patterns.

IV. Stochastic Verification

As discussed in the previous section, timing errors cannot be completely eliminated in the circuits with adaptive speed control. Researchers working for any types of adaptive speed control claim that by tuning some design parameters the pos-

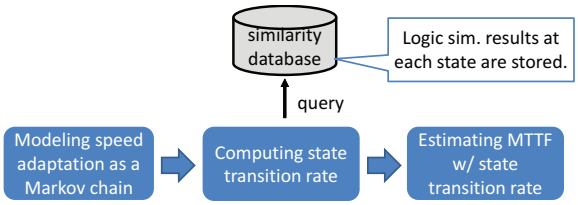

Figure 8. Overview of stochastic error rate estimation.

sibility of timing error occurrence can be reduced to almost zero and the mean time to failure (MTTF) over years can be easily attained with some overhead. For example, delay test should be more frequently carried out, or earlier error prediction should be enforced. However, it is challenging to quantitatively estimate such long MTTF and extremely low probability of error occurrence. A naive simulation is totally impractical since one year operation of a processor, for example, includes 3×10^{16} cycles, and to get 10,000 samples, 3×10^{20} cycles must be simulated. With a logic simulator processing 3×10^{3} cycles per second, it takes 3×10^{9} years, and hence another approach instead of naive simulation is indispensable.

For such a purpose, we have developed a stochastic estimation method of timing errors instead of simulation [11] (Fig. 8). The proposed method models adaptive speed control under dynamic delay variation as a continuous-time Markov process. Markov process is a stochastic process having a Markov property that the next state is determined by only the current state and is independent of the previous states. Especially, continuous-time Markov process is a special Markov process whose time parameter is continuous.

We assign states as follows. The circuit delay temporally fluctuates due to unintentional temperature change, power supply noise and aging. By sensing such temporal delay fluctuation with online/offline delay testing, the performance of the circuit under adaptive speed control is intentionally tuned by supply voltage scaling and/or body biasing. We define states in Markov process such that each state is associated with a pair of unintentional delay variation and levels of intentional speed control. We often prepare several discrete values for supply voltage scaling and body biasing. On the other hand, the unintentional delay variation is continuous in nature, but for the model simplicity, we discretize the unintentional delay variation into several representative values. We call these states as normal states. On the other hand, we add one more failure state meaning that a timing error happened in the past.

Figure 9 illustrates an example of state assignment and a series of state transitions falling into the failure state. In this example, the circuit starts to operate at speed control level of 0 with 0ps delay fluctuation. Then, both the speed control level and delay fluctuation are varying dynamically. At a certain time, a timing error happens at speed control level of 0 with 30ps delay fluctuation, and the state falls into the failure state.

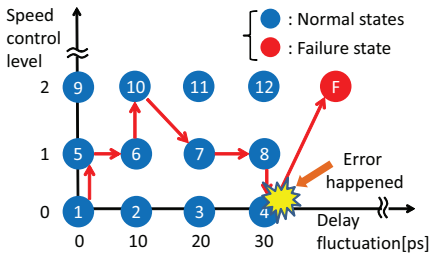

Figure 9. State assignment and transition.

In a continuous-time Markov process, transition rate of going from state i to state j, $q_{i,j}$ is the key parameter that characterizes the process behavior. Given a matrix of the transition rates, we can obtain closed-form expressions of state probability as a function of time t. This means that once the matrix of transition rates is given, the MTTF computation can be carried out with a constant time, and the computation time is independent of the timing error rate and MTTF of the circuit under evaluation. Note that the above computation is applicable to any types of adaptive speed control, since the state assignment explained above is independent of the implementation of adaptive speed control. To construct the transition rate matrix, we developed a similarity database and a direct derivation method of the matrix using the database (Fig. 8). Thanks to this development, the proposed method computes MTTF 10^{12} times faster than a logic simulator in a test case [11].

V. Analysis Example

This section shows an analysis example with the proposed stochastic error rate estimation method.

A. Experimental Setup

Here, the adaptive speed control is applied to MIPS R3000 microprocessor. R3000 is a 32-bit RISC microprocessor and implemented with five pipeline stages. The processor was designed such that RTL hardware description was synthesized by a commercial logic synthesizer with a 65nm industrial standard cell library. The number of standard cells is 6,813. The maximum clock frequency at 1.2V and 25°C is 147MHz, which corresponds to the critical path of 6.8ns.

In constructing similarity database, we selected four benchmark programs (CRC32, SHA1, Dijkstra and Quicksort) from MIBenchmark and 30 sets of input data for each program. The database for scan-test was constructed using patterns for path delay tests generated by a commercial ATPG tool. Launch on capture (LoC) scheme is adopted. Ten speed levels, i.e. ten supply voltages (1.2V, 1.1V, 1.0V, 0.90V, 0.85V, 0.80V, 0.75V, 0.70V, 0.65 and 0.60V) were prepared. For simplicity, all the cells have the same delay variation at each supply voltage, while any technical limitation is not given by the proposed framework. As for dynamic delay variation per gate due to, such as, environmental fluctuation and aging, 0 to 260 ps delay increases with 10 ps

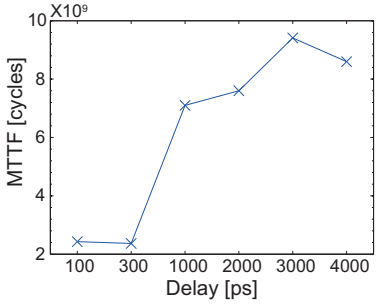

Figure 10. MTTF versus delay of delay buffer in TEP-FF.

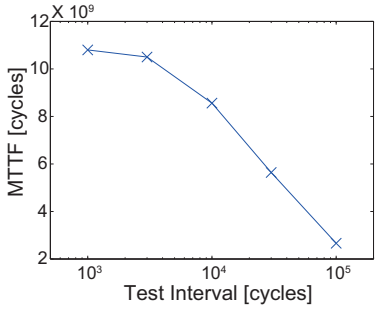

Figure 11. MTTF versus scan-test interval.

step were evaluated. With this database setup, the maximum number of states that can be analyzed with the proposed framework is 10 (speed levels) $\times$ 27 (delay fluctuation) + 1 (failure) = 271, and it was adopted for the experiments.

B. Analysis Example

Now, we can quickly estimate MTTFs of adaptive speed control systems of which operation parameters are changed. This subsection shows two examples illustrating the dependence of MTTF on the operation parameters.

Figure 10 shows MTTF when the delay time of the delay buffer in each TEP-FF is varied from 100ps to 4ns. We can see that MTTF becomes significantly short when the delay is less than 300ps. On the other hand, when the delay is over 3000ps, MTTF does not improve. This result suggests that the delay shorter than 300ps cannot well predict the timing errors and the MTTF of 10 $\times 10^9$ cycles cannot be obtained only adjusting the delay buffer since the delay fluctuation given here was significant.

It should be noted that these analyses can be executed without reconstructing the similarity database. By using this property, we can explore and design an adaptive speed control system satisfying given specifications.

Next, we demonstrate MTTF for the adaptive speed control based on off-line scan-test. Figure 11 plots MTTF when the interval of scan-test is changed from 1k to 100k cycles. We can see that more frequent scan-test contributes to longer MTTF and this becomes significant in the case that the scan-test interval is over 10k cycles. The proposed framework quantitatively tells us the MTTF tendency, which is helpful for system design and validation.

VI. Conclusion

This paper demonstrated how much power reduction could be obtained by minimizing margin for PVT variation. A case study with a 65nm fabricated subthreshold circuit showed that 46% power reduction was possible by breaking away from traditional worst-case design for manufacturing variability. However, run-time performance adaptation involved a serious issue on timing error occurrence. For putting run-time performance for a practical use, we need to verify run-time performance adaptation in design time in terms of performance and timing error rare. As a mean of verification, we introduced a stochastic error rate estimation framework and showed analysis examples. Future works include clarifying the advantage against aging and verifying the hardware-estimation correlation.

Acknowledgements

This work was partly supposed by NEDO and STARC.

References

[1] T. Kuroda et al., "Variable Supply-Voltage Scheme for Low-Power High-Speed CMOS Digital Design," *IEEE J. Solid-State Circuits*, vol. 33, no. 3, pp. 454–462, Mar. 1998.

[2] J. W. Tschanz et al., "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.

[3] J. T. Kao, M. Miyazaki, and A. P. 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.

[4] M. Agarwal, B. C. Paul, Z. Ming, and S. Mitra, "Circuit Failure Prediction and Its Application to Transistor Aging," *Proc. VTS*, pp.277–286, 2007.

[5] Y. Li, S. Makar, and S. Mitra, "CASP: Concurrent Autonomous Chip Self-Test Using Stored Test Patterns," *Proc. DATE*, pp.885–890, 2008.

[6] S. Das, et.al., "A self-tuning DVS processor using delay-error detection and correction," *J. Solid-State Circuits*, vol.41, pp.792–804, Apr. 2006.

[7] S. Das et al., "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan. 2009.

[8] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Adaptive Performance Compensation with In-Situ Timing Error Predictive Sensors for Subthreshold Circuits," *IEEE Trans. VLSI Systems*, vol. 20, no. 2, pp. 333–343, Feb. 2012.

[9] T. Nakura, K. Nose, and M. Mizuno, "Fine-Grain Redundant Logic Using Defect-Prediction Flip-Flops," *ISSCC Dig. Tech. Papers*, pp. 402–403, 2007.

[10] T. Sato and Y. Kunitake, "A Simple Flip-Flop Circuit for Typical-Case Designs for DFM," *Proc. ISQED*, pp. 539–544, 2007.

[11] S. Iizuka, M. Mizuno, D. Kuroda, M. Hashimoto and T. Onoye, "Stochastic Error Rate Estimation for Adaptive Speed Control with Field Delay Testing," *Proc. ICCAD*, 2013.

[12] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Trade-Off Analysis between Timing Error Rate and Power Dissipation for Adaptive Speed Control with Timing Error Prediction," *IEICE Trans. Fundamentals*, vol. E92-A, no. 12, pp. 3094–3102, Dec. 2009.