Run-time Performance Adaptation: Opportunities and Challenges

Masanori Hashimoto
Dept. Information Systems Engineering, Osaka University, Japan
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. First, this paper exemplifies the power reduction thanks to run-time performance adaptation with a 65nm test chip. Then, we introduce a stochastic framework to verify and optimize the run-time adaptation system in design time.**

I. Introduction

Device miniaturization due to technology scaling has made parametric performance variation more and more significant. Lower supply voltage makes circuits sensitive to environmental fluctuation, especially to supply voltage. Furthermore, aging effects, such as NBTI (negative bias temperature instability), HCI (hot carrier injection) and TDDB (time dependent dielectric breakdown), cause unexpected timing failures in field. To overcome manufacturing variability, environmental fluctuation and aging, designers set design margin and production tests give operational margin. When the given margin is large enough, timing failures in filed can be avoided. On the other hand, if the given margin is too large, the chip is often operated at a supply voltage higher than necessary and consequently it consumes larger power. Such excessive design and operational margins involve power, area and cost overhead, and/or performance loss, which deteriorates the competitiveness of the chips.

Figure 1 illustrates the operational margin in the chip lifetime. The operational margin at the beginning of the chip lifetime is large, and the margin decreases as the chip ages. If the delay increase due to aging exceeds the timing margin, timing error occurs in the chip. In addition, it is known that some aging effects vary transistor by transistor. For example, the threshold voltage variation due to NBTI is randomly distributed because the location and number of traps in the gate oxide are determined by a stochastic process. Furthermore, the speed of aging process depends on the workload (switching activity), supply voltage and temperature. In industry, the worst-case aging effects, i.e. the worst-case transistor, the worst workload, the highest supply voltage, and the highest temperature are assumed for estimating the impact of the aging effects, and the design and operational margins are often determined according to the worst-case aging effects. For some chips, such large margins could be really necessary. However, for most of the other chips, such large margins are excessive.

To overcome this problem, adaptive speed control system is studied in which each chip self-adjusts its operating condition, such as supply voltage and body bias, accompanied with timing self-test [1], [2]. 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 [3], [4]. 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.

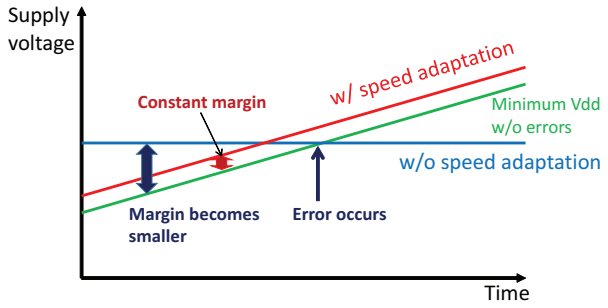

Fig. 1. Margins of circuits with and without adaptive speed control in chip lifetime.

Run-time adaptive speed control minimizes timing margin for power reduction and hence may cause timing errors due to unexpected delay increase, which is applicable to not only voltage over-scaling but also ordinary voltage scaling. 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.

This paper exemplifies how much power reduction can be obtained by reducing margin for manufacturing and environmental variation through run-time adaptation with a 65nm test chip. 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 analysis examples showing how MTTF (mean time to failure) depends on design parameters.

II. Run-time Adaptation Schemes

This section discusses two types of run-time performance adaptation; online test based run-time adaptation and offline test based run-time adaptation.

A. Online Test Based

Online test based run-time adaptation includes two approaches; error detection based approach and error prediction based approach. “Razor I” [3] and “Razor II” [4] detect timing

978-1-4799-8364-3/15/$31.00 ©2015 IEEE

errors in actual paths and correct the errors. Razor techniques require a re-execution mechanism to correct timing errors. In [5], error detection sequential element and tunable replica circuit are comparatively discussed. An advantage of the replica circuit is less intrusiveness to critical path timing, whereas its tuning is difficult. In both the cases, the re-execution is necessary and is performed through architectural replay, which is often integrated in high-performance processors to support branch prediction. However, the feature of such architectural reply is not available for general sequential circuits.

In contrast, references [6]–[8] presented an error predictive sensor embedded into actual paths. This sensor cannot detect timing errors but predict them. [9] proposes to insert a sensor at an intermediate node to detect the late-arriving data transitions for enabling dynamic clock gating. 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. The opportunity of the error prediction approach will be exemplified in Section III.

B. Offline Test Based

Next, we explain the second type of run-time adaptation, which is an adaptive speed control system that repeatedly performs delay test in idle times of the circuit (Fig. 2). 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 successive operations. 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.

III. Opportunities

Vth variation due to manufacturing variability and temperature fluctuation significantly varies speed and power consumption of low voltage circuits. If adding up worst-cases for each variation factor, power dissipation may increase more than 10x in a low voltage circuit. We therefore adopted an adaptive speed control scheme [6] (Fig. 3). The timing error predictive flip-flop (TEP-FF) causes a setup violation earlier than the

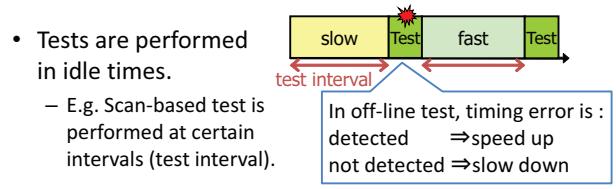

Fig. 2. Run-time adaptive speed control with offline test.

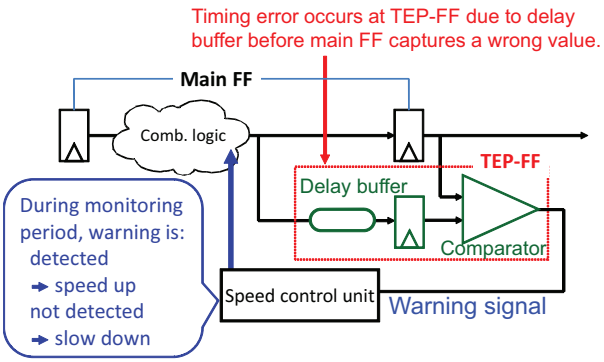

Fig. 3. Run-time adaptive speed control with TEP-FF [6].

main flip-flop due to the inserted delay element. This error signal is used as a warning signal indicating a shortage of timing slack, and the circuit is speeded up or down according to this signal. As mentioned earlier, this error prediction approach has two 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.

This adaptive speed control was applied to a 32-bit Kogge-Stone adder. A test chip was fabricated in 65nm process. Figure 4 shows a measurement result under temperature variation. The supply voltage is 0.35V. In this test chip, the circuit speed is adjusted by body-biasing. (a) corresponds to the proposed speed control, (b) is the power dissipation when 200 mV forward body-bias is given to satisfy the speed requirement at 25°C, and (c) is the power dissipation when the minimum body-bias is given at each temperature. This result shows that the power dissipation of the proposed speed control is close to (c) and the speed control is well working. Compared to conventional adaption of (b), the power dissipation is reduced by 40%.

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 5 shows the power dissipation of five chips in the following cases;

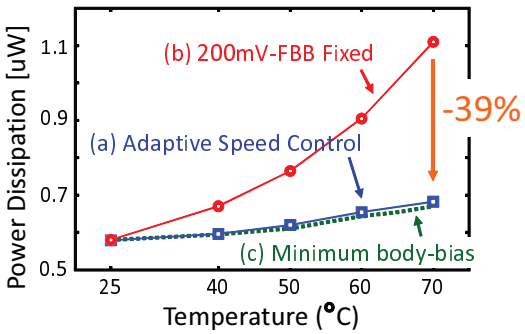

Fig. 4. Measurement result of speed adaptation (3MHz, 0.35V) [6].

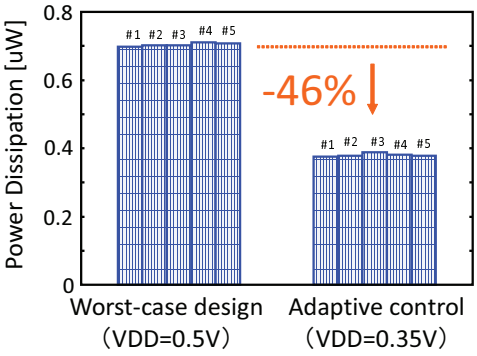

Fig. 5. Power dissipation when operation frequency is 2 MHz in the following cases; (worst-case design) all chips operate at $V_{DD} = 0.5$ V, (adaptive speed control) all chips operate with adaptive control at $V_{DD} = 0.35$ V [6].

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.

IV. Challenges

Run-time adaptive speed control, on the other hand, cannot completely eliminate timing errors due to unexpected delay increase. Such a timing error occurs not only for voltage over-scaling but also for ordinary voltage scaling. In addition, 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. Meanwhile, the occurrence frequency of timing errors can be changed by design modification, and long MTTF, such as a year, is supposed to be obtained via design optimization. For example, delay test should be more frequently carried out, or earlier error prediction should be enforced. However, it is challenging to quantitatively estimate

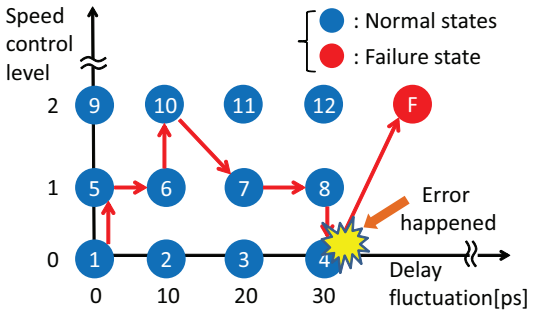

Fig. 6. State assignment and transition.

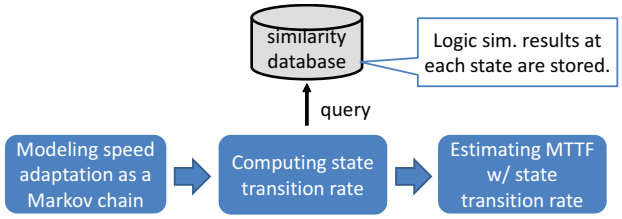

Fig. 7. Overview of Stochastic Error Rate Estimation.

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 is indispensable.

For such a purpose, we have developed a stochastic estimation method of timing errors instead of simulation [10]. The proposed method models the adaptive speed control under dynamic delay variation as a continuous-time Markov process, and stochastically estimates MTTF. Figure 6 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 7 shows the overview of the proposed method. First, the state assignment explained above is performed. Once a matrix of transition rates between states is given, the MTTF can be calculated via matrix computations and its calculation time is independent of how long MTTFs are, which is an excellent property for evaluating a long-MTTF circuit operation. To construct the transition rate matrix, we developed a similarity database and a direct derivation method of the matrix using the database. Thanks to this development, the proposed method computes MTTF 10^{12} times faster than a logic simulator in a test case.

Here, let us show how MTTF and power consumption depends on design parameters as an example. We used an MIPS

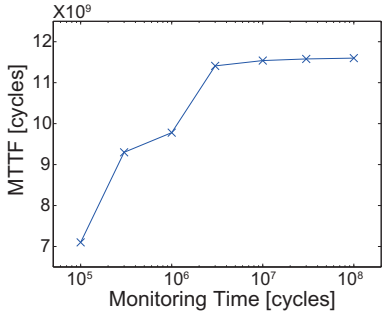

Fig. 8. Relation between MTTF and monitoring time.

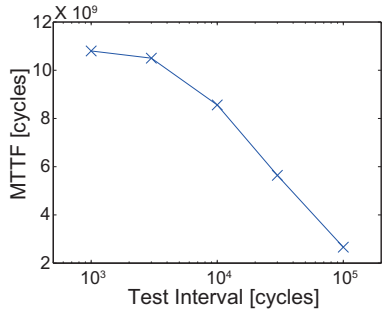

Fig. 9. Relation between MTTF and scan test interval.

R3000 microprocessor, which had 5-stage pipeline and 32-bit RISC instruction set, as a target of adaptive speed control. We synthesized an RTL description into a gate-level netlist with a commercial logic synthesizer and an industrial 65nm standard cell library. Ten speed levels, i.e. ten supply voltages (1.2, 1.1, 1.0, 0.90, 0.85, 0.80, 0.75, 0.70, 0.65 and 0.60V) could be selected in the adaptive speed control. Offline test patterns were prepared for path delay fault. The paths under test were 20,000 most timing-critical paths. A commercial ATPG tool was used for scan test pattern generation.

Figure 8 shows the relation between MTTF and the monitoring time before voltage downscaling. Here, on-line adaptation with TEP-FF is assumed. In this scheme, the circuit is slowed down for power reduction in case that no warning signals are observed during the monitoring period. The monitoring time was varied from 100k to 100M cycles. As the monitoring time becomes longer, MTTF is extended. Depending on the time constant of delay variation in the field and the required MTTF, the monitoring time needs to be determined.

We next evaluated MTTF of adaptive speed control with offline scan test. Figure 9 shows MTTF in case that the time interval of scan test was varied from 1k to 100k cycles. We can clearly see that MTTF was improved as offline scan test was performed more frequently, and this tendency was more significant beyond 10k cycles. This is reasonable since the delay fluctuation occurred within the test interval tends to be small and timing errors become less likely to occur.

V. 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.

The continuous-time Markov process modeling enabled the fast estimation of MTTF of adaptive speed controlled circuit. However, there remain many factors of delay fluctuation that are not taken into consideration, such as inter-die and within-die process variation. Ignorance of the process variation means that the estimated MTTF is only valid for a particular chip and it is not valid for other chips with different inter-die and within-die process variations. Our future work includes the extension of the continuous-time Markov process modeling to cover various factors of delay fluctuation. Also, we need to clarify the advantage against aging and verify the hardware-estimation correlation.

Acknowledgements

This work was partly supported by NEDO and STARC.

References

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

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

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

[4] S. Das, S. Das, C. Tokunaga, S. Pant, W.H. Ma, S. Kalaiselvan, K. Lai, D.M. Bull, and D. Blaauw, “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.

[5] K. A. Bowman, J. W. Tschanz, S. L. Lu, P. A. Aseron, M. M. Khellah, A. Raychowdhury, B. M. Geuskens, C. Tokunaga, C. B. Wilkerson, T. Karnik, and V. K. De, “A 45 nm Resilient Microprocessor Core for Dynamic Variation Tolerance,” *IEEE J. Solid-State Circuits*, Vol. 46 , No. 1, pp.194 –208, Jan. 2011.

[6] 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.

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

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

[9] J. Zhou, X. Liu, Y.-H. Lam, C. Wang, K.-H. Chang, J. Lan, M. Je, “HEPP: A new in-situ timing-error prediction and prevention technique for variation-tolerant ultra-low-voltage designs, ” in *Proc. A-SSCC*, pp. 129–132, 2013.

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