Run-Time Adaptive Performance Compensation using On-chip Sensors

Masanori Hashimoto
Dept. of Information Systems Engineering, Osaka University & JST, CREST
hasimoto@ist.osaka-u.ac.jp

Abstract— This paper discusses run-time adaptive performance control with on-chip sensors that predict timing errors. The sensors embedded into functional circuits capture delay variations due to not only die-to-die process variation but also random process variation, environmental fluctuation and aging. By compensating circuit performance according to the sensor outputs, we can overcome PVT worst-case design and reduce power dissipation while satisfying circuit performance. We applied the adaptive speed control to subthreshold circuits that are very sensitive to random variation and environmental fluctuation. Measurement results of a 65nm test chip show that the adaptive speed control can compensate PVT variations and improve energy efficiency by up to 46% compared to the worst-case design and operation with guardbanding.

I. Introduction

As manufacturing technology advances and supply voltage is lowered, 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, 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.

This paper reviews post-silicon tuning techniques with an emphasis on run-time adaptation, and introduces a run-time adaptive speed control using on-chip sensors for timing error prediction. Then, case studies are shown that the run-time adaptive speed control is applied to subthreshold circuits which are very susceptive to manufacturing and environmental variations. Measurement results of 65nm test chips demonstrate that the run-time adaptive speed control overcomes PVT variations and eliminates large design margin for guardbanding.

II. Post-silicon Tuning

This section reviews post-fabrication performance adjustment.

A. Tuning Phase

Post-silicon performance tuning is often carried out in the following four phases.

- Shipping test
- Power-on test
- Off-line (pseudo on-line) test
- Run-time

For high-end microprocessors for super-computers and servers, intensive delay tests are carried out on an LSI tester before the shipment, and the necessary supply voltage is carefully evaluated and recorded using fuse or flush for each chip. This approach requires an expensive test cost, and hence it is applicable only for high-end products.

As aging effects become significant, field test that aims to detect gradual performance degradation and wearing-out failures. An approach to tackle this problem is to carry out a test when a chip is powered on [1]. Good points of this power-on test approach are that the time for test is almost invisible for users and relatively long test patterns can be applied compared to the off-line test. However, the power-on test is not applicable to the chips running continuously without power-off and it does not capture environmental fluctuation.

To overcome the drawbacks of the power-on test, off-line test has been studied. This approach is well matched with multi-/many-core chips, since all the cores are not running all the time and some cores are temporally idle. Exploiting this temporal idol time or forgiving a slight performance degradation due to decrease in the number of cores running, functional and delay tests can be executed [2]. Thus, this approach is called pseudo on-line test as well. In this approach, there is a tradeoff between the idol/down time and test coverage. On the other hand, it is difficult to apply off-line test to uncore circuits and SoCs in general, because hardware redundancy is not usually available, although [3] tests uncore circuits with a special hardware support.

The last one is run-time adaptation that can cope with manufacturing variability, environmental fluctuation and aging. The run-time adaptation is classified into two groups, error correction approach and error prediction approach. These will be discussed in the next subsection.

B. Run-time Timing Sensing and Adaptation

The run-time speed adaptation requires sensing the timing margin of the circuit. For this purpose, critical path replica [4] 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 [5, 6, 7, 8]. However,

978-1-4244-7514-8/11/$26.00 ©2011 IEEE

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 [5] and "Razor II" [6] 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 high-performance processors, and hence it is not easy to apply it to general sequential circuits. In addition, Razor FF requires the timing window of error detection just after the clock edge in order to detect a late-arriving signal as a timing error, which induces severer minimum path delay constraints.

In contrast, "Canary Flip-Flop" [7] and "Defect Prediction Flip-Flop (DPFF)" [8] 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. The adaptive speed control with timing error prediction will be introduced in Section III.

C. Performance Control

Post-silicon performance adjustment is mainly performed by supply voltage scaling or body biasing [9, 10]. Various supply voltage can be provided by an external voltage regulator or an on-chip DC-to-DC converter. On the other hand, low-impedance distribution of multiple supply voltages is burden to physical design. In addition, spatially fine-grained scaling is difficult because level shifters are necessary between different voltage levels. Therefore, core-level or block-level scaling to select from a few variable voltage levels is reasonable.

On the other hand, generation and distribution of body voltages are relatively easier compared to those of supply voltage, because the flowing current is very small. However, still distribution of multiple voltages is expensive. On the other hand, finer-grained body biasing is possible as long as well separation overhead is acceptable, because no level converters are necessary. This makes gate clustering for body biasing possible [11, 12], and multiple speed levels can be provided for each block even when the number of provided bias voltages is two [11].

III. Run-Time Adaptive Performance Compensation with Timing Error Prediction using On-chip Sensors

Figure 1 shows a circuit that adaptively controls the speed and power dissipation using a warning signal generated by a canary FF [13, 14, 15]. 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

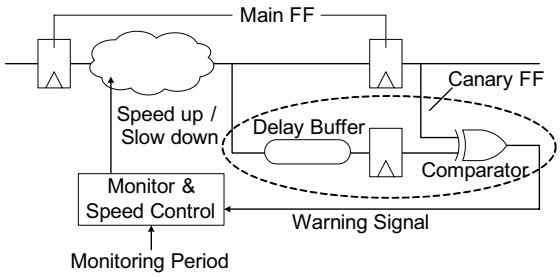

Fig. 1. Run-time adaptive speed control with canary FF.

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, which is believed to be a critical problem that prevents a practical use. 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, which will be discussed in the next section.

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. Similar discussion is applicable to power-on test and off-line test as well.

IV. Systematic Estimation of Timing Error and Its Dependence on Design Parameters

In applying the run-time adaptive speed control with canary 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 canary FF should be inserted
- delay time of the delay buffer in canary FF

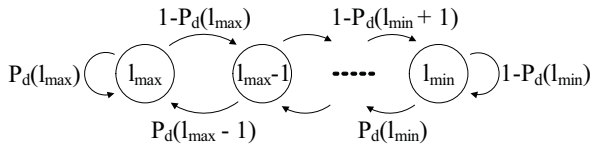

Fig. 2. Speed level transition modeled with Markov chain.

- 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 canary FFs trades the rate of timing errors and area overhead.

To quantitatively discuss the dependency of timing error on the design parameters, we briefly introduce a framework presented in [13, 14] (see [13, 14] for details). The framework exploits the path activation probabilities to estimate the timing error rate and power dissipation. The occurrence probabilities of warning signals and timing errors are derived from the path activation probabilities. The speed-level transition satisfies Markov property, because the next state (speed level l) is derived from only the current state and the occurrence probability of warnings. Therefore, the state transition is modeled using Markov chain (Fig. 2). The state (speed level) transition probabilities $Pd(l)$, $(1 - Pd(l))$ and the state probability of being at each speed level are calculated from the occurrence probability of warnings. Based on the state probability and the occurrence probability of timing errors, the timing error rate and power dissipation are obtained.

Let us show some examples of analyzed results. For experiments, we used a 32-bit ripple carry adder (RCA) and a 32-bit Kogge-Stone adder (KSA) in subthreshold operation in a 90nm CMOS process. The outputs of RCA and KSA 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. 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 here focus on the adaptive speed control for temperature (0 ℃ to 80 ℃).

Figure 3 shows an example of the relation between average power dissipation and mean time between failures (MTBF). Here, a canary 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 MTBF assuming 10MHz operation. Figure 3 indicates that inserted location S[i] and buffer delay affect MTBF significantly, which means

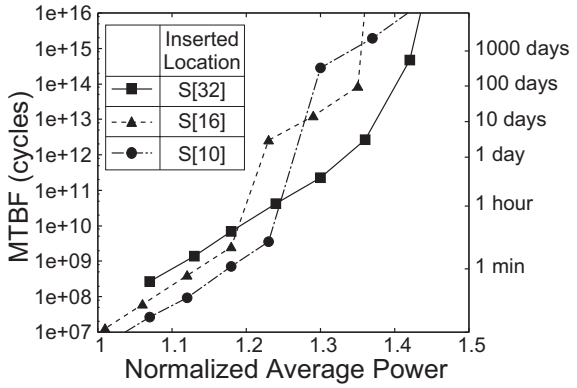

Fig. 3. Average power dissipation versus mean time between failures (MTBF) with various buffer delays in RCA. Each dot corresponds to different configuration of buffer delay. Monitoring period is 10^8 cycles.

the optimal design parameters vary depending on the required error rate.

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

Figure 4 shows trade-off relations between average power dissipation and MTBF 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 MTBF > 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 MTBF exponentially.

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. 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 canary 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.

Figure 5 shows the relation between average power dissipation and MTBF in the case of KSA, where buffer delay is changed at each canary FF position. This figure indicates that when MTBF (> 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, which is different from RCA case, whereas below 10^9 cycles, power dissipation can be reduced by choosing an appropriate location. Thus, the appropriate design

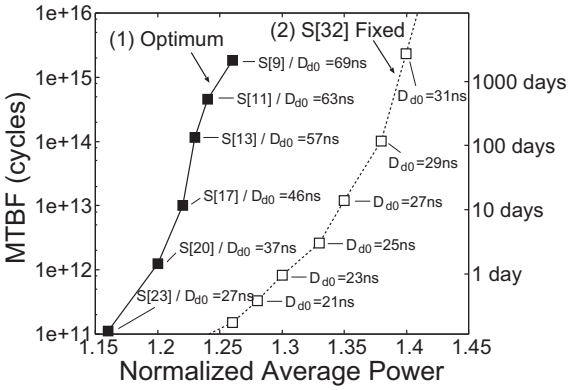

Fig. 4. Comparison between two cases in RCA; (1) both inserted location and buffer delay are optimized and (2) insertion location is fixed to S[32]. Monitoring period is 10^9 cycles.

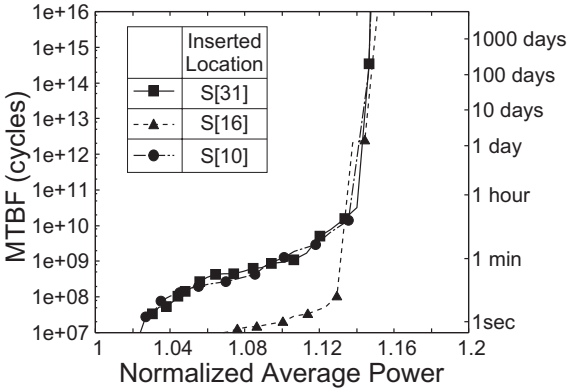

Fig. 5. Average power dissipation versus MTBF with various buffer delays in KSA. Each dot corresponds to different configuration of buffer delay. Monitoring period is 10^9 cycles and operating frequency is 20MHz.

parameters are circuit dependent.

V. Silicon Results

We designed and fabricated a test circuit to validate the adaptive speed control with canary FF in a 65 nm CMOS process [15]. Measurement results are shown in this section.

A. Circuit

The structure of the test circuit is depicted in Fig. 6 and the micrograph is shown in Fig. 7. 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. A timer signal is asserted when the monitoring period of the warning signal is elapsed.

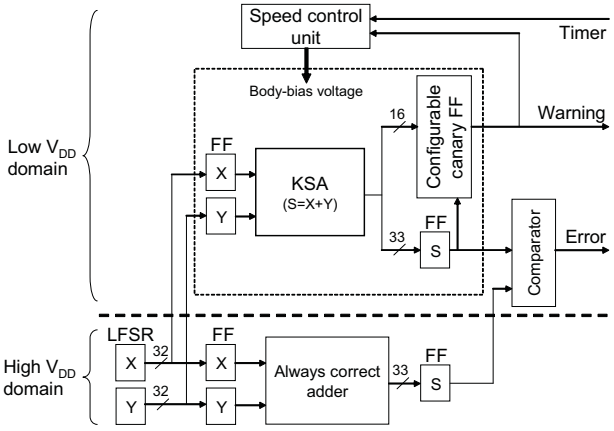

Fig. 6. Block diagram of test circuit. 32-bit Kogge-Stone adder (KSA) is controlled adaptively with configurable canary FF.

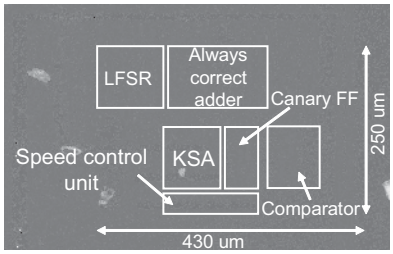

Fig. 7. Micrograph of test chip.

The speed control unit alters by body-biasing the speed of the KSA, main FFs and canary 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.

We implemented the "configurable" canary FF such that the inserted location and the buffer delay can be configured. The configurable canary FF is composed of 16 canary FFs with variable delay buffer. Each canary FF inserted at S[17]-S[32] can be enabled or disabled individually.

B. Operation Example

Figure 8 shows an operation example with a measured timing error, warning signals, and speed level transitions when the circuit was controlled adaptively with a canary FF. The operation frequency and V_{DD} were 2 MHz and 350 mV. The step of body-biasing levels was set to 30 mV, which means speed level 1 corresponds to a 30-mV forward body bias (FBB) when speed level 0 is zero body bias (ZBB). The speed level was altered according to the warning signal. A timing error occurred in this example.

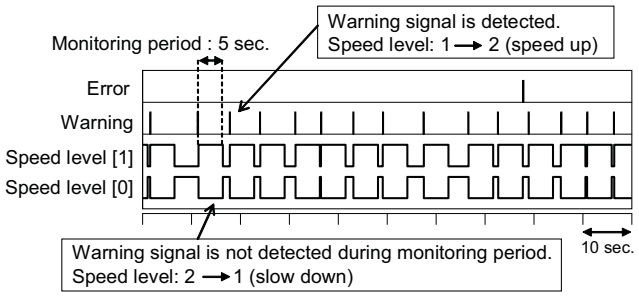

Fig. 8. Measured timing error, warning signals, and transitions of speed level (2 MHz @ V_{DD} = 0.35 V).

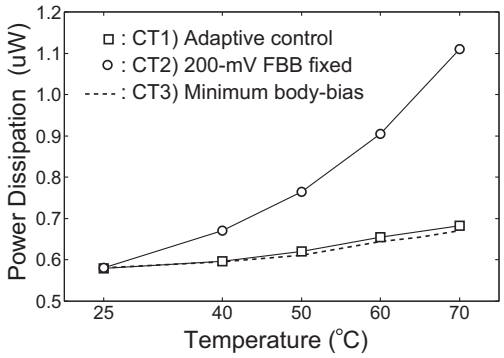

Fig. 9. Power dissipation at the various temperature conditions (3 MHz @ V_{DD} = 0.35 V). Circuit operates CT1) adaptively, CT2) with 200-mV FBB fixedly, and CT3) with minimum body-bias required for 3-MHz operation at each temperature.

C. Adaptive Compensation of Environmental Variability

Figure 9 shows the power dissipation at various temperature conditions (25–70℃) when the operation frequency was set to 3 MHz in the following cases;

CT1: the circuit was controlled adaptively with a canary FF,

CT2: 200-mV FBB, which was the minimum body-bias for a 3-MHz operation at 25℃, was fixedly applied,

CT3: the minimum FBB voltage required for a 3-MHz operation at each temperature was applied.

In CT1, a canary FF at S[20] was enabled and its buffer delay was 130 ns at ZBB and 25℃. The power dissipation includes those of the KSA, main FFs, speed control unit, and canary FF. The power overhead of the canary FF was estimated to be around 2% by circuit simulation. This measurement set four speed levels out of seven speed levels (ZBB – 180-mV FBB) at each temperature. No timing errors were observed during 1.8×10^9 cycles at all temperature conditions.

Figure 9 indicates that the power dissipation of CT1 is very close to that of CT3, which means optimal body-bias voltages were selected adaptively at each temperature. On the other hand, when the 200-mV FBB was fixedly applied (CT2), the power dissipation at 70℃ was 63% larger than that of CT1.

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

D. Comparison to Operation Considering Worst-case

We next demonstrate how inefficient the worst-case design for process variation is for subthreshold circuits, and clarifies 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 10 shows the power dissipation of five chips in the following cases;

CM1: all chips operated at V_{DD} = 0.5 V, which was the minimum V_{DD} for a chip at the SS device corner,

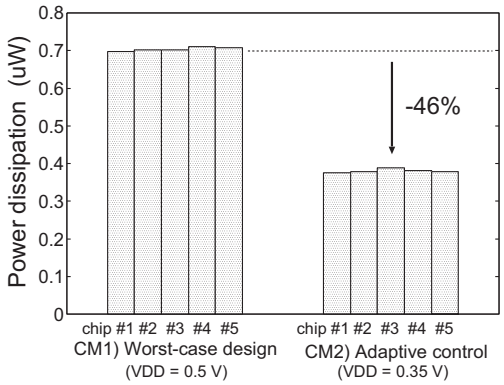

Fig. 10. Power dissipation when operation frequency is 2 MHz in the following cases; CM1) all chips operate at V_{DD} = 0.5 V, CM2) all chips operate with adaptive control at V_{DD} = 0.35 V.

CM2: all chips operated with adaptive control at V_{DD} = 0.35 V.

One canary FF was enabled, and its location and buffer delay were determined such that no timing errors occurred during 1.2×10^9 cycles (10 minutes). The power dissipation with the adaptive control (CM2) was smaller than that with guardbanding (CM1) by 46%, because of lower supply voltage.

VI. Conclusions

We presented a self-adaptive compensation technique using canary FF as a run-time performance adaptation technique that can overcome manufacturing variability, environmental fluctuation and aging. Timing error occurrence inherently involved in the error-prediction approach is quantitatively discussed, and the dependency of timing error and power dissipation on design parameters is shown. We applied the self-adaptive speed control to a 32-bit KSA, whose performance was controlled by body-biasing, and fabricated a test chip in a 65-nm CMOS process. Measurement results showed that the adaptive control compensated manufacturing and environmental variability and reduced power dissipation by 46% compared to traditional worst-case design.

Future work includes improvement of design methodology, an application to larger circuits, such as processors, and effectiveness validation on silicon.

Acknowledgments

The author thanks Professor Hiroshi Fuketa of University of Tokyo for his contributions to this work. This work is supported in part by New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Y. Hyunbean, T. Yoneda, M. Inoue, Y. Sato, S. Kajihara, H. Fujiwara, "Aging test strategy and adaptive test scheduling for SoC failure prediction,", in *Proc. International On-Line Testing Symposium (IOLTS)*, pp.21–26, 2010.

[2] L. Yanjing, S. Makar, S. Mitra, "CASP: Concurrent Autonomous Chip Self-Test Using Stored Test Patterns," in *Proc. Design, Automation and Test in Europe*, pp.885–890, 2008.

[3] L. Yanjing, O. Mutlu, D. S. Gardner, S. Mitra, "Concurrent autonomous self-test for uncore components in system-on-chips," in *Proc. VLSI Test Symposium (VTS)*, pp.232–237, 2010.

[4] J. Kao, M. Miyazaki, and A. Chandrakasan, "A 175-mV Multiply-Accumulate Unit Using an Adaptive Supply Voltage and Body Bias Architecture," *IEEE Journal of Solid-State Circuits*, vol.37, pp.1545–1554, Nov. 2002.

[5] 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 Journal Solid-State Circuits*, vol.41, pp.792–804, Apr. 2006.

[6] 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," in *International Solid-State Circuits Conference Dig. Tech. Papers*, pp.400–401, Feb. 2008.

[7] T. Sato and Y. Kunitake, "A Simple Flip-Flop Circuit for Typical-Case Designs for DFM," in *Proc. International Symposium on Quality Electronic Design*, pp.539–544, Mar. 2007.

[8] T. Nakura, K. Nose, and M. Mizuno, "Fine-Grain Redundant Logic Using Defect-Prediction Flip-Flops," in *International Solid-State Circuits Conference Dig. Tech. Papers*, pp.402–403, Feb. 2007.

[9] B. H. Calhoun and A. Chandrakasan "Standby power reduction using dynamic voltage scaling and canary flip-flop structures," *IEEE Journal Solid-State Circuits*, vol.39, pp.1504–1511, Sep. 2004.

[10] 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 Journal Solid-State Circuits*, vol.37, pp.1396–1402, Nov. 2002.

[11] K. Hamamoto, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Tuning-Friendly Body Bias Clustering for Compensating Random Variability in Subthreshold Circuits," in *Proc. International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 51–56, 2009.

[12] S. H. Kulkarni, D. M. Sylvester and D. T. Blaauw, "Design-Time Optimization of Post-Silicon Tuned Circuits Using Adaptive Body Bias," *IEEE Trans. on CAD*, vol. 27, No. 3, pp. 481–494, March 2008.

[13] 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," in *Proc. Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2009, pp. 266–271.

[14] 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. on Fundamentals*, vol. E92-A, no. 12, pp. 3094–3102, Dec. 2009.

[15] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Adaptive Performance Compensation with In-Situ Timing Error Prediction for Subthreshold Circuits," in *Proc. Custom Integrated Circuits Conference (CICC)*, 2009, pp. 215–218.