# Adaptive Performance Compensation with On-Chip Variation Monitoring

Masanori Hashimoto[1]     Hiroshi Fuketa[2]
[1]Dept. of Information Systems Engineering, Osaka University & JST, CREST
[2]Institute of Industrial Science, University of Tokyo
hasimoto@ist.osaka-u.ac.jp     fuketa@iis.u-tokyo.ac.jp

***Abstract*—This paper discusses adaptive performance control with two types of on-chip variation sensors. The first sensor aims to extract several device-parameters for performance adaptation from a set of on-chip ring-oscillators with different sensitivities to device-parameters, and the device-parameter decomposition is discussed. The second sensors, which are embedded into functional circuits, predict timing errors due to PVT variations and aging. By controlling circuit performance according to the sensor outputs, PVT worst-case design can be overcome and power dissipation can be reduced while satisfying performance requirements. Measurement results of a subthreshold adder on 65-nm test chips 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 [1]–[4]. 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 performance tuning techniques at various phases, and introduces on-chip variation sensors for shipping test. We also discuss run-time adaptive speed control using on-chip sensors for timing error prediction. Measurement results of a subthreshold circuit on 65-nm test chips demonstrate that the run-time adaptive speed control overcomes PVT variations and eliminates large design margin for guardbanding.

## II. Post-silicon Tuning

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 has been applicable only for high-end products. On the other hand, by using on-chip variation sensor, it would be possible to simplify the testing and reduce the tuning cost. Section III will discuss the variation sensors for such a purpose.

As aging effects become significant, field test that aims to detect gradual performance degradation and wearing-out failures is drawing an attention. An approach to tackle this problem is to carry out a test when a chip is powered on [5]. 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 [6]. 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 [7] 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 speed adaptation requires sensing the timing margin of the circuit. For this purpose, critical path replica [8] has been traditionally used. However, its efficiency is deteriorating because the performance mismatch between the replica and the actual critical path tends to be significant due to increasing within-die variation and aging. To more efficiently sense the timing margin, in-situ techniques have been studied [9]–[12]. Nevertheless, 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.

The run-time adaptation is classified into two groups, error correction approach and error prediction approach. “Razor I” in [9] and “Razor II” [10] are the first approach that detects timing errors with a delayed clock in a processor and corrects

978-1-61284-857-0/11/$26.00 ©2011 IEEE

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" [11] and "Defect Prediction Flip-Flop (DPFF)" [12] 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 IV.

## III. On-Chip Variation Sensors for Device-Parameter Extraction

To adapt the performance efficiently at shipping test phase, it is required to estimate for every chip how device-parameters varied from their typical values during the manufacturing process. Then, estimates of device-parameters are used to obtain an appropriate tuning. For example, when the magnitude of PMOS threshold voltage is high and NMOS threshold voltage is typical, forward body bias should be given to PMOSs, not to NMOSs. Otherwise, large increase in leakage current would be introduced. As an application of this type of sensors, clock skew reduction is investigated in [13].

For such a purpose, RO (Ring-Oscillator)-based sensors have been intensively studied [14]–[17]. They can be easily implemented in a chip and can be used to obtain variability information even after the product shipment, because the oscillating frequencies of ROs can be easily measured with a simple circuit structure. Besides, ROs consisting of ordinary standard cells give the speed variation. However, they are not capable of decomposition of the speed variation into device-parameters, such as threshold voltages and channel lengths of PMOS and NMOS, because the speed sensitivities of the ROs to device-parameters are similar.

To extract not only the speed variation but also device-parameters, sophisticated ROs have been proposed [16], [17]. In these ROs, the sensitivity to a single or two device-parameters is intensified, and the sensitivities to other device-parameters are suppressed. Using a set of these ROs with different sensitivities, device-parameters are estimated.

However, when using such ROs, random variations might not be canceled out. An example of this phenomenon is demonstrated here when using such a highly-sensitive RO to device-parameters in 65-nm process. Random variations of $\Delta V_{th_{n/p}}$ and $\Delta L_{n/p}$, whose averages are 0, are given, where $V_{th_{n/p}}$ means N/PMOS threshold voltage and $L_{n/p}$ is gate length of N/PMOS. Frequency distributions of 101-stage ROs composed of normal inverters or highly-sensitive inverters to NMOS parameters are shown in 1. $\Delta F$ denotes the shift of the oscillation frequency from its nominal value in an RO. When an RO comprises normal inverters, the average of $\Delta F$ is -0.45% and the standard deviation is 0.8%. On the other hand, when an RO consists of sophisticated inverters that have intentionally high sensitivity to device-parameters of $\Delta V_{th_n}$ and $\Delta L_n$, the effect of random variability on $\Delta F$ is more significant. In fact, the random variability changes the average of $\Delta F$ by -3.8% even though the averages of every variation are 0. The standard deviation of $\Delta F$ reaches 2.0% indeed even when the number of RO stage is 101. Therefore, in order to accurately estimate device-parameters using such sophisticated ROs, random variability should be taken into account explicitly. Both the large standard deviation of $\Delta F$ and the shift of the average must be considered in the device-parameter estimation from the measured sensor outputs.

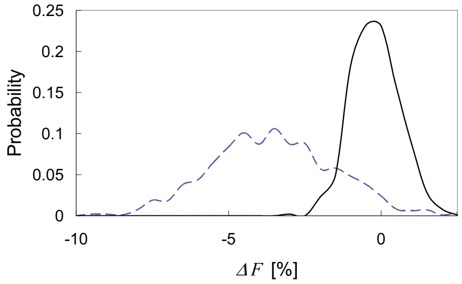

Fig. 1. Frequency-distributions with only random variations (without global variations). Supply voltage is 0.9 V. Solid line denotes a distribution with an RO constructed by normal inverters and dashed line denotes that with an RO constructed by inverters that have high sensitivity to NMOS $L$ and $V_{th}$. The number of stages in both ROs is 101.

To deal with the above fact, we have proposed a device-parameter extraction method considering random variations explicitly and demonstrated that the proposed method can accurately estimate both of global and random variations [18]. This method is based on MLE (Maximum Likelihood Estimation) and extracts not only D2D parameter variations but also standard deviations of random variations.

Let us shows an example of device-parameter extraction result in 65-nm process. This evaluation was experimentally performed using virtually-fabricated chips (simulated data). It is assumed that variation sources to be extracted are $V_{th_{n/p}}$ and $L_{n/p}$, and $\sigma_{\Delta G/R_{Vth_{n/p}}} = 20$ mV and $\sigma_{\Delta G/R_{L_{n/p}}} = 2$ nm. Here, $\Delta G_x$ denotes global variability, and $\Delta R_x$ corresponds to random variation of parameter $x$. Please see [18] for details including sensor structures. Table I lists the averages of the absolute estimation errors of $\boldsymbol{\Delta G_x}$ from the given variations. The proposed method reduces the estimate error by 11.1%–73.4% and provides more accurate estimation thanks to the consideration of random variations.

TABLE I
Absolute values of average estimate errors of $\boldsymbol{\Delta G_x}$ under global and random variations.

| Method | $\Delta G_{Vth_n}$ [mV] | $\Delta G_{Vth_p}$ [mV] | $\Delta G_{L_n}$ [nm] | $\Delta G_{L_p}$ [nm] |
|---|---|---|---|---|
| Proposed | 1.35 | 1.66 | 0.38 | 0.40 |
| Conventional | 4.01 | 6.23 | 0.57 | 0.45 |
| Error reduction | 66.3% | 73.4% | 33.3% | 11.1% |

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

### *A. Circuit Operation and Design Parameters*

Figure 2 shows a circuit that adaptively controls the speed and power dissipation using a warning signal generated by a timing-error predictive (TEP) FF [19]–[21]. 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 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 TEP-FF is well configured to generate the warning signal, the occurrence of timing error 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 necessarily happens, which is believed to be a critical problem that prevents a practical use. To reduce and manage the error occurrence, we have to examine and tune the following design parameters on the basis of systematically estimated error rate.

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

For this purpose, we developed a framework that uses path activation probabilities to estimate the timing error rate and power dissipation. Please see [19], [20], [22] for the details.

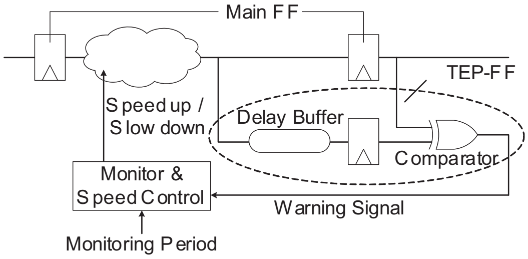

Fig. 2. Run-time adaptive speed control with TEP-FF.

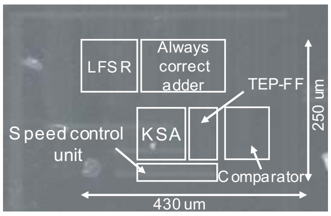

Fig. 3. Micrograph of test chip.

### *B. Silicon Results*

We designed and fabricated a test circuit to validate the adaptive speed control with TEP-FF in a 65-nm CMOS process [21]. Measurement results are shown in this section. Here, the run-time adaptive speed control is applied to subthreshold circuits which are very susceptive to manufacturing and environmental variations.

*1) Circuit:* A 32-bit Kogge-Stone adder (KSA) was adopted as a circuit whose performance was controlled adaptively. The micrograph is shown in Fig. 3. 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.

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

*2) Adaptive Compensation of Environmental Variability:* Figure 4 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 TEP-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 TEP-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 TEP-FF. The power overhead of the TEP-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 $1.8 \times 10^9$ cycles at all temperature conditions.

Figure 4 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 TEP-FF can well compensate delay fluctuation due to temperature shift.

*3) Comparison to Worst-case Design:* 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

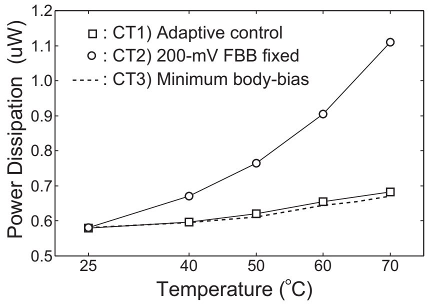

Fig. 4. Power dissipation at the various temperature conditions (3 MHz @ $V_{DD}$ = 0.35 V). Circuit operates CT1) adaptively, CT2) with 200-mV FBB fixedly, and CT3) with minimum body-bias required for 3-MHz operation at each temperature.

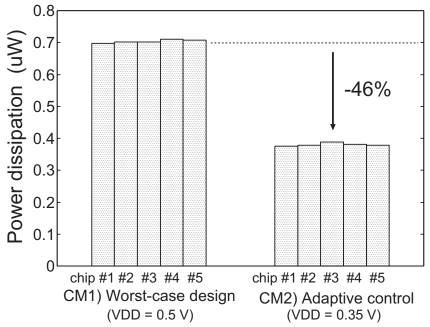

Fig. 5. Power dissipation when operation frequency is 2 MHz in the following cases; CM1) all chips operate at $V_{DD}$ = 0.5 V, CM2) all chips operate with adaptive control at $V_{DD}$ = 0.35 V.

$V_{DD}$ = 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;

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

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

One TEP-FF was enabled, and its location and buffer delay were determined such that no timing errors occurred during $1.2 \times 10^9$ 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.

## V. Conclusions

This paper reviewed post-silicon tuning for adaptive performance compensation, and first introduced on-chip variation sensors for device-parameter extraction. By explicitly considering random variations, threshold voltages and channel lengths of PMOS and NMOS were successfully extracted from oscillating frequencies of sophisticated ring-oscillators. We also presented a self-adaptive compensation technique using TEP-FF as a run-time performance adaptation technique that can overcome manufacturing variability, environmental fluctuation and aging. 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.

## Acknowledgments

This work is supported in part by New Energy and Industrial Technology Development Organization (NEDO).

## References

[1] 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.

[2] 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.

[3] K. Hamamoto, M. Hashimoto, Y. Mitsuyama, and T. Onoye, "Tuning-Friendly Body Bias Clustering for Compensating Random Variability in Subthreshold Circuits," *Proc. ISLPED*, pp. 51–56, 2009.

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

[5] Y. Hyunbean, T. Yoneda, M. Inoue, Y. Sato, S. Kajihara, H. Fujiwara, "Aging test strategy and adaptive test scheduling for SoC failure prediction,", *Proc. IOLTS*, pp.21–26, 2010.

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

[7] Y. Li, O. Mutlu, D. S. Gardner, S. Mitra, "Concurrent autonomous self-test for uncore components in system-on-chips," *Proc. VTS*, pp.232–237, 2010.

[8] 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.

[9] 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.

[10] 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," *Dig.. ISSCC*, pp.400–401, 2008.

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

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

[13] S. Abe, K. Shinkai, M. Hashimoto, and T. Onoye, "Clock Skew Reduction by Self-Compensating Manufacturing Variability with On-Chip Sensors," *Proc. GLSVLSI*, pp. 197–202, 2010.

[14] M. Bhushan, M. Ketchen, S. Polonsky, and A. Gattiker, "Ring oscillator based technique for measuring variability statistics," *Proc. ICMTS*, pp.87–92, 2006.

[15] L. Pang and B. Nikolic, "Measurements and Analysis of Process Variability in 90 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 5, pp. 1655–1663, 2009.

[16] I. A. K. M. Mahfuzul, A. Tsuchiya, K. Kobayashi, and H. Onodera, "Variation-sensitive Monitor Circuits for Estimation of Die-to-Die Process Variation," *Proc. ICMTS* 2011.

[17] B. Wan, J. Wang, G. Keskin, and L. T. Pileggi, "Ring Oscillators for Single Process-Parameter Monitoring," in *Proc. Workshop on Test Structure Design for Variability Characterization*, 2008.

[18] K. Shinkai and M. Hashimoto, "Device-Parameter Estimation with On-Chip Variation Sensors Considering Random Variability," *Proc. ASP-DAC*, pp. 683–688, 2011.

[19] 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," *Proc. ASP-DAC*, pp. 266–271, 2009.

[20] 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.

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

[22] 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*, in press.