Comparing Voltage Adaptation Performance between Replica and In-Situ Timing Monitors

Yutaka Masuda
Osaka University
Japan
masuda.yutaka@osaka-u.ac.jp

Jun Nagayama
Socionext Inc.
Japan

Hirotaka Takeno
Socionext Inc.
Japan

Yoshimasa Ogawa
Socionext Inc.
Japan

Yoichi Momiyama
Socionext Inc.
Japan

Masanori Hashimoto
Osaka University
Japan
hasimoto@osaka-u.ac.jp

ABSTRACT

Adaptive voltage scaling (AVS) is a promising approach to overcome manufacturing variability, dynamic environmental fluctuation, and aging. This paper focuses on timing sensors necessary for AVS implementation and compares in-situ timing error predictive FF (TEP-FF) and critical path replica in terms of how much voltage margin can be reduced. For estimating the theoretical bound of ideal AVS, this work proposes linear programming based minimum supply voltage analysis and discusses the voltage adaptation performance quantitatively by investigating the gap between the lower bound and actual supply voltages. Experimental results show that TEP-FF based AVS and replica based AVS achieve up to 13.3% and 8.9% supply voltage reduction, respectively while satisfying the target MTTF. AVS with TEP-FF tracks the theoretical bound with 2.5 to 5.6 % voltage margin while AVS with replica needs 7.2 to 9.9 % margin.

KEYWORDS

adaptive voltage scaling, timing error predictive FF, critical path replica, voltage margin, mean time to failure

ACM Reference Format:
Yutaka Masuda, Jun Nagayama, Hirotaka Takeno, Yoshimasa Ogawa, Yoichi Momiyama, and Masanori Hashimoto. 2018. Comparing Voltage Adaptation Performance, between Replica and In-Situ Timing Monitors. In *IEEE/ACM INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN (ICCAD '18), November 5–8, 2018, San Diego, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3240765.3240788

1 INTRODUCTION

Aggressive device miniaturization due to technology scaling has been improving the average device performance. Circuits, on the other hand, have become sensitive to static manufacturing variability and dynamic environmental fluctuation. Moreover, device aging, which is another temporal variation and is represented by negative bias temperature instability (NBTI) [1, 2], degrades performance gradually in the field. These static and temporal variations directly lead to circuit reliability degradation. For overcoming variabilities mentioned above, a traditional worst-case (WC) design gives design and operational margins in design time and in field, respectively, for ensuring correct circuit operation. However, as the performance variation becomes significant, such margins tend to be too painful for designers. Therefore, the conventional WC design with guardbanding is becoming less efficient, and an adaptive post-silicon performance compensation is eagerly demanded as a promising countermeasure.

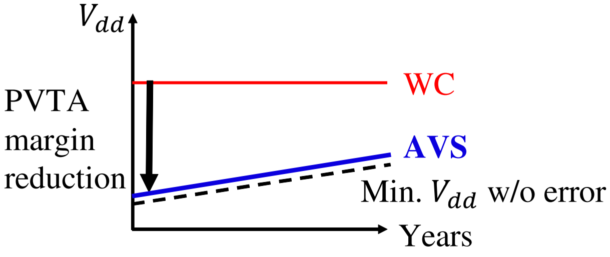

Figure 1: Supply voltages of AVS and conventional WC design in device lifetime. Ideal AVS minimizes PVTA margin of each chip.

The most effective tuning knob for post-silicon compensation is supply voltage control, and adaptive voltage scaling (AVS) is intensively studied [3–7]. AVS is expected to minimize process, voltage, temperature, and aging (PVTA) margin of each chip and allocate only a small margin taking into account the entire lifetime as shown in Fig. 1. The conventional PVTA margins, which are determined by the worst chip across all the variation sources, are excessive in most of the chips, and they can be exploited as the source of power reduction.

There are two AVS strategies in literatures; error detection and recovery based control with, for example, Razor [3], and error prediction and prevention based control with in-situ timing sensors [1] or critical path replica [4, 10, 11]. In both the strategies, sensors are

ICCAD '18, November 5–8, 2018, San Diego, CA, USA

ACM ISBN 978-1-4503-5950-4/18/11…$15.00
https://doi.org/10.1145/3240765.3240788

[1] There are several names for the same structure; canary FF [8], slack monitor [7] and error predictive FF [9]

embedded to detect/predict timing errors, and the supply voltage is controlled according to the sensor outputs. This paper, on the other hand, focuses on the error prediction and prevention strategy since any error recovery mechanisms are not necessary as long as the prediction is appropriate, whereas the error detection and recovery strategy requires a re-execution mechanism to correct timing errors, which is difficult to implement in general sequential circuits.

Once designers decide to introduce AVS for their design, they need to choose a sensor type for AVS and determine where and how many sensors are inserted. When sensors are poorly inserted, the sensors fail to predict timing errors resulting in timing error occurrence. Time to failure (TTF), which is the length of time until a chip starts to cause timing errors, can be a metric to quantitatively evaluate such a misprediction issue. In another case, inadequate sensor insertion cannot reduce design and operation margins. To avoid these unsuccessful AVS designs and eliminate unnecessary margins, the sensor selection and insertion need to be validated in terms of TTF and margin reduction.

This paper discusses the voltage margin reductions achieved by AVS circuits with different sensors; in-situ sensors and replica. We quantitatively evaluate the average supply voltage taking into account manufacturing variability at time zero, subsequent voltage elevation due to aging and dynamic supply noise. Depending on the requirement of TTF, the achievable trade-off between clock period and average supply voltage becomes different. In this work, we give mean time to failure (MTTF) as a design constraint and compare the trade-offs of AVS circuits with different sensors. For such MTTF aware trade-off analysis, we utilize a stochastic framework proposed by Iizuka *et al.* [9], which models AVS circuit behavior under static and dynamic delay variations as a stochastic Markov process and computes average supply voltage and MTTF. In addition, for investigating the remaining margin even with AVS, we derive the lower bound of the average supply voltage. We formulate a problem to derive the lower bound as a linear programming (LP) problem. By comparing the average supply voltages of AVS circuits with the lower bound, we can reveal the remaining margins.

Contributions of this work can be summarized as follows.

- Quantitative comparison of MTTF-aware trade-off between clock period and average supply voltage between AVS circuits with TEP-FF and replica taking into account static and various dynamic delay variations.
- LP based estimation of the lower bound supply voltage, which unveils the remaining margin originating from AVS implementation, to the best of our knowledge, for the first time.

The rest of this paper is organized as follows. Section 2 describes a strategy for comparing the performance of AVS circuits with TEP-FF and replica. The strategy in Section 2 first explains points for discussion and then defines design optimization for AVS circuits with each sensor. Section 3 proposes an LP based analysis method for estimating lower bound of average supply voltage with AVS under MTTF constraint. Section 4 demonstrates supply voltage reductions of AVS circuits with TEP-FF and replica. Lastly, concluding remarks are given in Section 5.

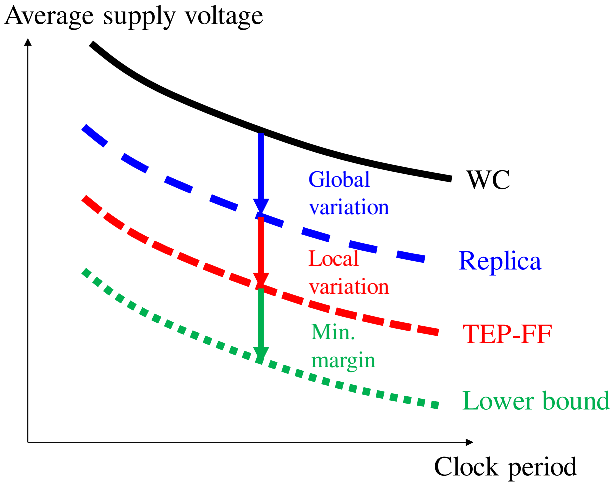

Figure 2: Expected voltage reduction thanks to the AVS with in-situ TEP-FF or replica under MTTF constraint.

2 STRATEGY FOR COMPARING IN-SITU SENSORS AND REPLICA

This section shows a strategy for comparing in-situ sensors and replica. First, Section 2.1 discusses the expected reduction of voltage margin thanks to AVS with in-situ TEP-FF and replica and highlights the points for discussion in this paper. Next, Section 2.2 explains the assumed TEP-FF based AVS and defines its design optimization. Also, Section 2.3 presents the assumed replica based AVS and its optimization problem.

2.1 Points for discussion

Fig. 2 exemplifies the expected V_{dd} reduction effects obtained by TEP-FF based AVS and replica based AVS under MTTF constraint. The top black curve represents the conventional WC design that accumulates timing margins assuming the worst PVTA condition. The second blue curve shows the trade-off curve of replica based AVS, and the elimination of timing margins for global variation is expected to reduce supply voltage. The third red curve corresponds to AVS with TEP-FF, and this AVS is supposed to lower supply voltage further by exploiting design margins for intra-die random variation. The bottom green curve means a lower bound of the trade-off, i.e. an ideal performance of AVS. Compared with the green curve and blue/red one, we can know how much voltage margin remains in the AVS circuit implemented with replica/TEP-FF.

Based on the above expectation, this work addresses the following questions; (1) how much voltage reduction can be achieved by TEP-FF based AVS and replica based AVS from conventional WC design under static and dynamic variations and MTTF constraint, and (2) how much voltage margin each AVS should remain under static and dynamic variations and MTTF constraint. For answering the first question, this paper utilizes a stochastic error rate estimation method [9] and evaluates MTTF and average supply voltage taking into account static manufacturing variability and dynamic variations such as supply noise and aging. For answering the second question, this work proposes to derive the lower bound of average supply voltage using an LP formulation and evaluates the gap between the lower bound and actual supply voltages. The detail of lower bound derivation will be explained in Section 3. The V_{dd} reduction achieved by each AVS and the gap from the lower bound will be experimentally demonstrated for an industrial design, a cipher circuit, and an embedded processor in Section 4.

Let us highlight two important points for discussion in this paper. A crucially important issue in investigating Fig. 2 is that the trade-off analysis must be conducted under the same MTTF constraint. If we accept shorter MTTF, we can aggressively reduce supply voltage and consequently the trade-off curve shifts. References [12, 13] compare critical path replica and in-situ slack monitor and experimentally show that replica fails to capture within-die variations such as random manufacturing variations. For example, [13] reports that in-situ slack monitor needs only 0.9% timing margins whereas replica requires 4.2% margins for ensuring correct operation at nominal PVTA condition. However, conventional works[12, 13] do not explicitly take into account the MTTF constraint and the impact of dynamic delay variations such as supply noise and aging. As mentioned earlier, the margin reduction performance of the AVS circuits having different MTTFs cannot be directly compared. In addition, appropriate margining for dynamic variations are indispensable in actual designs. To derive reliable implications from the comparison, we need to prepare a setup that can fairly compare the performance in practical situations. From this standpoint of view, it is necessary for designers to take into account not only static variation but the MTTF constraint and dynamic variations.

The second notable discussion is the comparison of the lower bound of the average supply voltage. This lower bound, which is plotted as the green curve in Fig. 2, represents the ideal trade-off between the supply voltage and cycle time under the given MTTF constraint. The difference between the AVS trade-off and the lower bound lets designers know how much voltage margins remain in field operation, and such margin information is helpful to examine design quality and provide feedback to design. For this purpose, Chen *et al.* proposed a method in [14] that determines how long AVS should stay at each supply voltage for satisfying a given bound of timing failure probability and discusses how to control a circuit being at each supply voltage for the determined duration. This method provides an exact answer in consideration of process and static temperature variations, but temporal variations, such as supply noise and aging, are not considered. For AVS circuit design aiming at the noise and aging compensation, we need to explicitly consider temporal variation in MTTT aware lower bound estimation of the average supply voltage, which will be discussed in Section 3.

Here, to achieve reliable discussion, we need to design both TEP-FF based AVS and replica based AVS reasonably well. For a fair comparison, we formulate their design problems as similar design optimization problems and compare the solutions. More precisely, we define the same objective function and similar design constraints using identical metrics. Section 2.2 and 2.3 describe TEP-FF based AVS design and replica based AVS design, respectively.

2.2 Designing TEP-FF based AVS

Fig. 3 shows the AVS circuit which is composed of a voltage scaled circuit, voltage control logic, and TEP-FF. TEP-FF consists of a normal flip-flop, delay buffers and a comparator, e.g. XOR gate. When the timing margin is gradually decreasing, a timing error occurs at TEP-FF before the main FF captures a wrong value due to delay buffers, 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. Note that TEP-FF is expected to convert timing margins for intra-die random variations to V_{dd} reduction since it shares main logic and its variation.

We define the design optimization problem for TEP-FF based AVS as follows.

- Objective
  - Minimize : V_{dd}
- Variables
  - $B_{TEP_i} (1 \leq i \leq N_{FF})$
- Constraints
  - $MTTF \geq MTTF_{const}$
  - $N_{TEP} (= \sum_{i=1}^{N_{FF}} B_{TEP_i}) \leq N_{TEP}^{\max}$

The objective of this problem is to minimize V_{dd} aiming at power minimization. The variable for optimization is B_{TEP_i}. B_{TEP_i} is a binary variable, and it becomes 1 when i-th FF is replaced to TEP-FF. The primary constraint is MTTF, and the lower bound of MTTF ($MTTF_{const}$) is given as a constraint. The second constraint gives the upper bound of the number of TEP-FF ($N_{TEP}^{\max}$), and this limits the area increase due to TEP-FF insertion. To make AVS work well, TEP-FF should monitor timing margins of paths that have a higher probability of timing error occurrence and output warning signals to prevent the error occurrence. For this purpose, [15] proposed a timing failure probability aware sensor insertion method. This method inserts TEP-FFs to voltage-scaled circuits using the timing failure probability, which is a joint probability of timing violation probability and activation probability, as a metric. In other words, the inserted sensors check timing margins of critical paths more frequently, and thus it enables temporally fine voltage control and helps to avoid timing error occurrence. Therefore, this paper focuses on timing failure probability and inserts TEP-FF referring to [15].

2.3 Designing replica based AVS

Fig. 4 shows the AVS circuit which is composed of voltage scaled circuit, voltage control logic, and critical path replica. Critical path replica includes replicated logic, delay buffers, and an edge detector. The edge detector checks the edge timing for every clock cycle and generates a warning signal when the edge is too late. Therefore, the timing margin can be measured much more frequently compared with TEP-FF.

For attaining the same sensitivity of the replica to variations with that of the voltage scaled circuit, the replica should include many paths in the voltage scaled circuit. However, it requires a large area cost, and hence Kim *et al.* designed a compact replica with comprehensive sensitivity analysis in design time[16]. In this paper, for pursuing a discussion that is independent of replica implementation methods, we assume that the inserted replica can perfectly reproduce the delay characteristics of the paths that are selected for monitoring. With this setup, the accuracy of critical

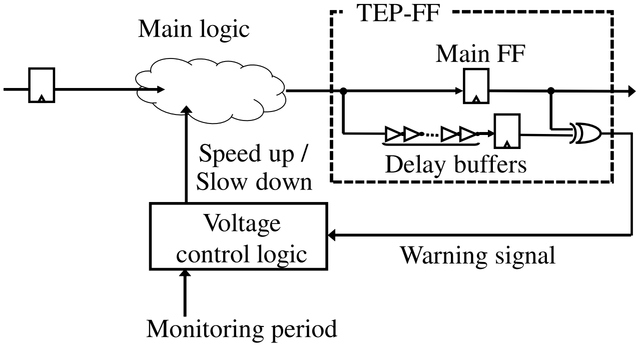

Figure 3: Assumed AVS with TEP-FF.

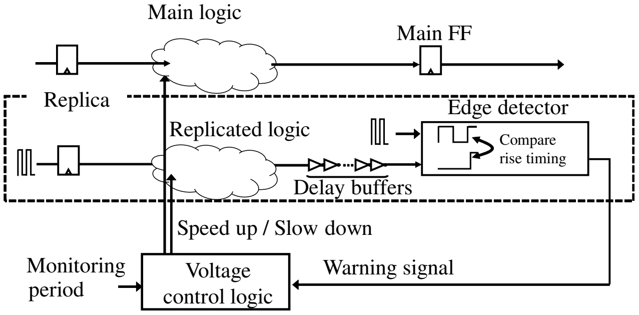

Figure 4: Assumed AVS with replica.

path delay measurement degrades only due to within-die variation, which is considered in our analysis.

On the other hand, Tschanz *et al.* proposed to integrate a tunable replica and tune it after fabrication[17]. Similarly, TEP-FF can be tuned if it is designed with tunable buffer. However, post-fabrication tuning during chip test is expensive for most of the products, and hence it is not considered in this study.

Similarly to Section 2.2, we formulate the design optimization of AVS with replica as follows.

- Objective
  - Minimize : V_{dd}
- Variables
  - $B_{replica_j} (1 \leq j \leq N_{path})$
- Constraints
  - $MTTF \geq MTTF_{const}$
  - $N_{replica} (= \sum_{j=1}^{N_{path}} B_{replica_j}) \leq N_{replica}^{\max}$

The objective of this problem is identical with that of Section 2.2. The variable for optimization is $B_{replica_j}$. $B_{replica_j}$ is a binary variable, and it becomes 1 when the j-th path is replicated. As the primary constraint, the lower bound of MTTF ($MTTF_{const}$) is given, which is the same as the first constraint in Section 2.2. The second gives the upper bound of the number of replicated paths ($N_{replica}^{\max}$), and this limits the area increase due to replica implementation. Note that the constraint of area overhead by AVS with TEP-FF and AVS with replica will be set identically for keeping fairness. Similarly to Section 2.2, we focus on timing failure probability and insert replica for sensing paths whose timing failure probabilities are high.

3 LOWER BOUND VDD ESTIMATION

This section proposes a lower bound estimation method of the average supply voltage that satisfies $MTTF_{const}$. Remind that this lower bound of the average supply voltage is referred as an ideal average supply voltage that exploits all the design margins to V_{dd} reduction. The proposed method derives the lower bound via an LP based optimization problem. Section 3.1 formulates the optimization problem and Section 3.2 shows a simple example to help intuitive understanding.

3.1 LP-based estimation

First, let us define parameters. The number of available supply voltages is N_v, and V_i ($1 \leq i \leq N_v$) denotes the i-th supply voltage. The aging process is a continuous process, but for the sake of computation compatibility with [9], the aging states are discretized, and the number of aging states is N_{age}. For each pair of supply voltage and aging state, we can define the duration $t_{i,j}$ in which the circuit operates at V_i in the j-th aging state. With these notations, the total duration in which the circuit operates at V_i is expressed as $\sum_{j=1}^{N_{age}} t_{i,j}$. Therefore, the average supply voltage becomes $\frac{\sum_{i=1}^{N_v} V_i \times (\sum_{j=1}^{N_{age}} t_{i,j})}{\sum_{i=1}^{N_v} \sum_{j=1}^{N_{age}} t_{i,j}}$. We want to minimize this average supply voltage under the constraint of $MTTF_{const}$, and we formulate this problem as an LP problem as follows.

- Objective
  - Minimize : $\sum_{i=1}^{N_v} V_i \times (\sum_{j=1}^{N_{age}} t_{i,j})$
- Variables
  - $t_{i,j}$
- Constraints
  - $\sum_{i=1}^{N_v} \sum_{j=1}^{N_{age}} (F_{i,j} \times t_{i,j}) \leq 0.5$
  - $\sum_{i=1}^{N_v} \sum_{j=1}^{N_{age}} t_{i,j} = MTTF_{const}$
  - for each j : $\sum_{i=1}^{N_v} (t_{i,j} \times a_{i,j}) \leq 1$

Next, the constraints are explained. The first constraint is given to satisfy $MTTF_{const}$, where $F_{i,j}$ is the timing failure probability at the i-th supply voltage in the j-th aging state. This constraint expression is derived with Maclaurin expansion from the following equation.

$$\prod_{i=1}^{N_v} \prod_{j=1}^{N_{age}} (1 - F_{i,j})^{t_{i,j}} \geqq 0.5. \qquad (1)$$

$(1 - F_{i,j})$ represents the probability that no errors occur during a unit time, where the unit time is defined as a clock period in this paper. Note that this probability computation takes into account manufacturing variability. Therefore, $(1 - F_{i,j})^{t_{i,j}}$ is the probability that no errors occur during time $t_{i,j}$. By multiplying this probability for all the combinations of supply voltages and aging states, we can calculate the overall probability that no errors occur as the left-hand side of Eq. (1). On the other hand, supposing a time-invariant timing failure probability per a unit time, it should be smaller than $0.5/MTTF_{const}$. When the left-hand side of Eq. (1) is larger than the 0.5, the circuit achieves the MTTF equal to or longer than $MTTF_{const}$. The second constraint is given to set the total operating time of AVS to $MTTF_{const}$.

The third constraint is given to control the transition speed between aging states. $a_{i,j}$ represents the within-state aging progress per a unit time in j-th aging state. The product term $t_{i,j} \times a_{i,j}$ expresses the accumulated aging progress within j-th aging state at i-th voltage in time $t_{i,j}$. $\sum_{i=1}^{N_v}(t_{i,j} \times a_{i,j})$ means the overall aging progress within j-th aging state. If this $\sum_{i=1}^{N_v}(t_{i,j} \times a_{i,j})$ is equal to 1, the aging proceeds from j-th state to $(j+1)$-th state. We note that $F_{i,j}$ and $a_{i,j}$ can be obtained from [9], and hence the variable is only $t_{i,j}$ in the above LP problem.

3.2 Example

Let us exemplify the lower bound estimation of the average supply voltage. The example in Fig. 5 supposes AVS with four discrete states, which are composed of combinations of two V_{dd} levels of 1.2 V and 1.0 V, and two ΔV_{thp} aging levels of 0 mV and 5 mV and one state for representing the circuit operation fails. In this example, there are six parameters under consideration; $MTTF_{const}$, $t_{1.2V,0mV}$, $t_{1.0V,0mV}$, $t_{1.2V,5mV}$, $t_{1.2V,5mV}$, and V_{ave}, where the first parameter is the target MTTF and set to 3.0×10^{15} cycles, the second to fifth parameters are elapsed times in which AVS operates in the corresponding combination of V_{dd} and ΔV_{thp}, and the last parameter is the average supply voltage until the operating time reaches $MTTF_{const}$. Here, we want to minimize the average supply voltage, and hence we need to maximize the sum of $t_{1.0V,0mV}$ and $t_{1.0V,5mV}$ while satisfying the target MTTF. Here, MTTF is defined as the total operation time such that the accumulated failure probability reaches 0.5, and thus we discuss $MTTF$ by examining the accumulated failure probability. Firstly, we utilize stochastic error estimation method[9] and represent the circuit behavior as the Markov model as shown in Fig. 5(a). This process gives us the aging speed, i.e. green arrows, and failure probability, i.e. red arrows, for each pair of supply voltage and aging states. Then, we extract $a_{i,j}$ and $F_{i,j}$ from the constructed Markov model as shown in Fig. 5(b). We can obtain the parameters $a_{i,j}$ and $F_{i,j}$ as follows; $a_{1.0V,0mV} = 1.0 \times 10^{-15}$ [times/clock cycle], $a_{1.0V,5mV} = 0$ [times/clock cycle], $a_{1.2V,0mV} = 1.0 \times 10^{-14}$ [times/clock cycle], $a_{1.2V,5mV} = 0$ [times/clock cycle], $F_{1.0V,0mV} = 0$ [1/clock cycle], $F_{1.0V,5mV} = 0.5 \times 10^{-15}$ [1/clock cycle], $F_{1.2V,0mV} = 0$ [1/clock cycle], $F_{1.2V,5mV} = 0.5 \times 10^{-16}$ [1/clock cycle]. Note that when the failure probability is time-invariant in the same state, the accumulated failure probability increases linearly in time. Therefore, we assume the accumulated failure probability and threshold voltage degradation increase linearly in time. Note that the linearity assumption in aging degradation may have a computational error, but we can control the accuracy similarly with Markov model in [9]. For example, if the Markov model prepares the larger number of discrete aging states, our model can increase the number of discrete states and thus can improve the accuracy. When ΔV_{thp} is less than 5 mV, the accumulated failure probability does not increase and hence, AVS can stay at 1.0 V as long as ΔV_{thp} is less than 5 mV, then $t_{1.2V,0mV} = 0$ and $t_{1.0V,0mV} = 10^{15}$. If the supply voltage is fixed to 1.0V, the accumulated failure probability becomes 0.5 when 2.0×10^{15} cycles elapses. Therefore, after ΔV_{thp} reaches at 5 mV, we need to change V_{dd} to satisfy target $MTTF_{const}$ so that the accumulated failure probability is lower than 0.5 when the operating time is equal to $MTTF_{const}$. From the above, we can construct the following simultaneous equations.

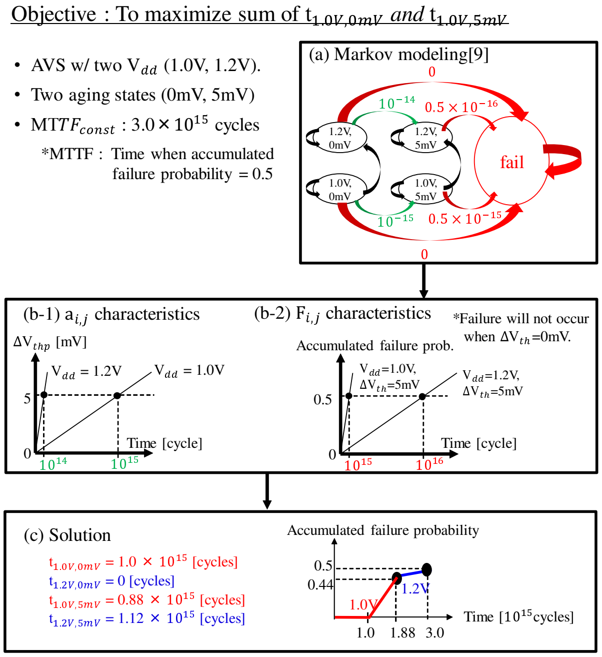
Objective : To maximize sum of $t_{1.0V,0mV}$ and $t_{1.0V,5mV}$
AVS w/ two V_{dd} (1.0V, 1.2V).
Two aging states (0mV, 5mV)
$MTTF_{const}$: 3.0×10^{15} cycles
*MTTF : Time when accumulated failure probability = 0.5
(a) Markov modeling[9]
0
10^{-14}
0.5×10^{-16}
1.2V, 0mV
1.2V, 5mV
fail
1.0V, 0mV
1.0V, 5mV
10^{-15}
0.5×10^{-15}
0
(b-1) $a_{i,j}$ characteristics
ΔV_{thp} [mV]
V_{dd} = 1.2V
V_{dd} = 1.0V
5
0
Time [cycle]
10^{14}
10^{15}
(b-2) $F_{i,j}$ characteristics
*Failure will not occur when ΔV_{th}=0mV.
Accumulated failure prob.
V_{dd}=1.0V, ΔV_{th}=5mV
V_{dd}=1.2V, ΔV_{th}=5mV
0.5
0
Time [cycle]
10^{15}
10^{16}
(c) Solution
$t_{1.0V,0mV} = 1.0 \times 10^{15}$ [cycles]
$t_{1.2V,0mV} = 0$ [cycles]
$t_{1.0V,5mV} = 0.88 \times 10^{15}$ [cycles]
$t_{1.2V,5mV} = 1.12 \times 10^{15}$ [cycles]
Accumulated failure probability
0.5
0.44
1.0V
1.2V
Time [10^{15}cycles]
1.0
1.88
3.0

Figure 5: An example estimating lower bound of average supply voltage while satisfying target MTTF. (a)Markov model with [9], (b) aging and accumulated failure probability characteristics, and (c) solution.

$$\left\{\begin{array}{lr}
t_{1.0V,5mV} \times F_{1.0V,5mV} + t_{1.2V,5mV} \times F_{1.2V,5mV} \le 0.5. & (2)\\
t_{1.0V,0mV} + t_{1.0V,5mV} + t_{1.2V,0mV} + t_{1.2V,5mV} = MTTF_{const}. & \\
 & (3)\\
t_{1.0V,0mV} \times a_{1.0V,0mV} + t_{1.2V,0mV} \times a_{1.2V,0mV} \le 1. & (4)\\
a_{1.0V,0mV} = 1.0 \times 10^{-15}, a_{1.0V,5mV} = 0, & \\
a_{1.2V,0mV} = 1.0 \times 10^{-14}, a_{1.2V,5mV} = 0. & (5)\\
F_{1.0V,0mV} = 0, F_{1.0V,5mV} = 0.5 \times 10^{-15}, & \\
F_{1.2V,0mV} = 0, F_{1.2V,5mV} = 0.5 \times 10^{-16}. & (6)\\
V_{ave} = \dfrac{(t_{1.0V,0mV} + t_{1.0V,5mV}) \times 1.0}{MTTF_{const}} + & \\
\qquad \dfrac{(t_{1.2V,0mV} + t_{1.2V,5mV}) \times 1.2}{MTTF_{const}}. & \\
 & (7)
\end{array}\right.$$

Eq. (2) constrains the accumulated failure probability to satisfy $MTTF_{const}$ and corresponds to the first constraint in Section 3.1. Eq. (3) represents that the sum of operation time is equals to $MTTF_{const}$, i.e. 3.0×10^{15} cycles, and corresponds to the second constraint in Section 3.1. Eq. (4) controls aging progress and corresponds to the third constraint in Section 3.1. Eqs. (5) and (6) can be constructed by

Fig. 5(a) and (b), and Eq. (7) calculates the average supply voltage and corresponds to the objective function in Section 3.1.

By minimizing V_{ave} with Eqs. (2) to (6), $t_{1.2V,0mV} = 0$ cycles, $t_{1.0V,0mV} = 1.0 \times 10^{15}$ cycles, $t_{1.2V,5mV} = 1.12 \times 10^{15}$ cycles, $t_{1.0V,5mV} = 0.88 \times 10^{15}$ cycles and $V_{ave} = 1.07$ V are obtained.

4 EVALUATION

This section experimentally evaluates supply voltage reduction of AVS with in-situ TEP-FF and replica from conventional WC design. First, Section 4.1 explains the evaluation setup, and Section 4.2 demonstrates the average supply voltage of each AVS and lower bound. Then, Section 4.3 discusses the performance difference between AVS circuits with TEP-FF and replica.

4.1 Evaluation setup

In this work, we used an industrial image signal processor (ISP), an advanced encryption standard (AES) circuit and an OR1200 OpenRISC processor, which is a 32-bit RISC microprocessor with five pipeline stages, as target circuits. ISP was designed by a commercial place and route tool with a 28 nm Socionext standard cell library and AES and OpenRISC were laid out with a 45 nm Nangate standard cell library. Also, standard cell memories[18, 19] were used as SRAMs in OpenRISC processor. The post-layout circuits include 3,133,640 combinational logic cells, 16,870 latches, and 374,880 FFs in ISP, 1,276,989 combinational logic cells, 589,890 latches, and 2,504 FFs in OpenRISC, and 17,948 combinational logic cells and 530 FFs in AES, respectively.

For calculating meaningful MTTF, practical delay variations should be considered. Our evaluation took into account the following variations.

- Dynamic supply noise, which is assumed to temporally fluctuate between -90 mV and 70 mV in ISP and between -50 mV and 50 mV in AES and OpenRISC.
- Manufacturing variability, which is assumed to consist of intra-die random variation and inter-die variation. In ISP, the inter-die variation is extracted from the difference of delay characteristics between TT, i.e. typical-typical, library and SS, i.e. slow-slow, global library and the intra-die variation is calculated with on-chip variation coefficient defined in the 28 nm standard cell library. In AES and OpenRISC, both the intra-die random variation and inter-die variation include NMOS and PMOS threshold voltage variation of $\sigma = 30$ mV and gate length variation of $\sigma = 1$ nm, respectively.
- NBTI aging, whose model was obtained by fitting a trapping/detrapping model [20] to the measured data in [21]. Note that, in ISP, this NBTI model is not used since the on-chip variation coefficient in the 28 nm standard cell library already includes aging-induced delay variation. In AES and OpenRISC, six degradation states of 0 mV, 0.5 mV, 1 mV, 5 mV, 10 mV and 15 mV are prepared. Note that [21] measures the NBTI degradation with stress probability of 100%, and thus the NBTI model used in our experiment does not consider recovery situation. Our future work includes to investigate the adequacy of the degradation status assignment and consider the relationship between degradation and activation probability.
- Temperature gradation, which is assumed to temporally fluctuate between −10°C and 110°C in ISP. Note that this temperature gradation is not taken into account in AES and OpenRISC.

For performing SSTA, we generate probability density functions of gate delay variability according to the assumed variations, execute sensitivity-based SSTA (such as [22] and [23]) to obtain the canonical-form expression of the timing violation probability, and calculate the timing violation probability by integrating the canonical-form expression with MATLAB 2016b.

As for workload, we selected one for ISP aiming to maximize power consumption. In OpenRISC, we chose three benchmark programs (CRC32, SHA1, and Dijkstra) from MIBenchmark [24]. For each program, 30 sets of input data were prepared for MTTF estimation. Totally, we used 90 (= 3 × 30) workloads. In AES, 1,000 random test patterns were used.

We prepared eight supply voltages from 0.90 V to 0.76 V with a 20 mV interval in ISP and six supply voltages from 1.20 V to 0.95 V with a 50 mV interval in AES and OpenRISC. We set MTTF of 1.00×10^{17} cycles, i.e. 10.5 years in ISP, 1.6 years in AES, and 13.7 years in OpenRISC, as $MTTF_{const}$. Note that the above $MTTF_{const}$ is just an example, and we can cope with other constraints of $MTTF_{const}$ similarly. With this setup, we inserted several TEP-FF or replica circuits to the voltage-scaled circuits. The constraints of area overhead by TEP-FF or replica circuits are set to 0.1% for ISP and OpenRISC and 1.0% for AES, respectively. In other words, the upper bound of the number of TEP-FF and replica paths, i.e. $N_{TEP}^{\max}$ and $N_{replica}^{\max}$ are 483 and 69 in ISP, 30 and 9 in AES, 50 and 11 in OpenRISC, respectively. In this work, we inserted the delay buffers whose delay were comparable to the delay variation caused by 20 mV supply noise in ISP and 50 mV one in AES and OpenRISC, where these numbers of 20 mV and 50 mV correspond to one level decrement of the supply voltage. We note that, in our evaluation, TEP-FF and replica circuits are virtually inserted to voltage-scaled circuits for simplicity. In other words, we calculated the MTTF from delay characteristics of laid out voltage scaled circuits and the nominal delay and variation of logic cells in TEP-FF and replica. Therefore, the area overhead by replica is denoted as the sum of the cell area of the target monitoring data path and delay buffers. Similarly, the area overhead by TEP-FF is denoted as the sum of the cell area of duplicated FF, comparator, and delay buffers. One of our future works includes to take into account the control circuit and its wirings to TEP-FF and replica.

MTTF and average supply voltage under PVTA variation are evaluated by a stochastic MTTF estimation framework proposed in [9]. In MTTF evaluation, we swept the clock period from 2,500 ps to 3,300 ps in ISP, 450 ps to 550 ps in AES, and from 4,000 ps to 5,500 ps in OpenRISC. For each clock period, AVS dynamically adjusts the supply voltage. In our experiment, the monitor period for AVS is varied from 10^6 cycles to 10^{15} cycles. Here, the monitor period of 10^6 cycles means, if no error prediction signals are outputted for 10^6 cycles, the supply voltage is decreased. The minimum monitor period, i.e. 10^6 cycles, is about 3.3 ms in ISP, 0.5 ms in AES, and 4.3 ms in OpenRISC, respectively, and it is longer than the response time of the fast transient voltage regulator, e.g. 1.6 μs in [25].

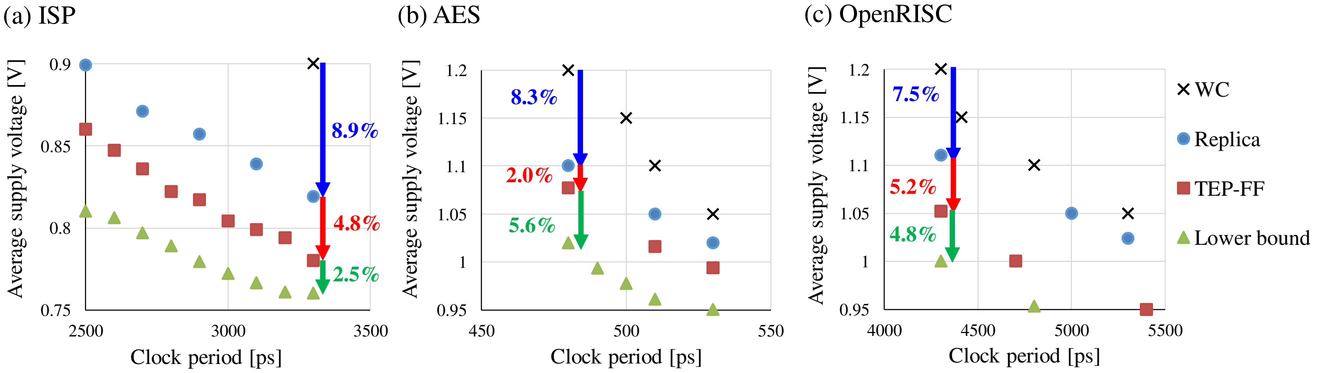

Figure 6: Trade-off comparison between conventional WC design, AVS with TEP-FF, AVS with replica, and lower bound. (a)ISP, (b)AES, (c)OpenRISC.

We used Gurobi Optimizer 7.0 to solve the LP problem defined in Section 3.1. The solver was executed on a 2.4 GHz Xeon CPU machine under the Red Hat Enterprise Linux 6 operating system with 1 TB memory. The required CPU times for solving the proposed ILP problem with Gurobi optimizer were at most 0.01 seconds in all AES, OpenRISC, and ISP. Remind that, in ISP, on-chip variation coefficient in 28 nm standard cell library already includes the delay variation by aging. In ISP, we derived aging-aware timing failure probability, which is the average of timing failure probability in consideration of aging, for each supply voltage, set the total number of aging states to 1, and solved the formulated LP.

4.2 V_{dd} reduction by TEP-FF and replica

Fig. 6 shows the trade-off curves between the minimum average supply voltage and the clock period under the MTTF constraint of 10^{17} cycles, where (a) in ISP, (b) in AES, and (c) in OpenRISC, respectively. The black cross plots represent the conventional WC design with guard-banding for PVTA variation. The blue circular plots and red square plots correspond to AVS circuits with replica and with TEP-FF, respectively. The green triangular plots are the lower bound under the given MTTF constraint.

In this section, we examine our evaluation results from the following two aspects; (1) V_{dd} reduction effect thanks to AVS with TEP-FF and replica, and (2) performance difference between AVS with TEP-FF, AVS with replica, and the lower bound.

First, we compare the black and blue/red plots for clarifying the performance improvement thanks to AVS with replica/TEP-FF. Fig. 6 shows that both replica based AVS and TEP-FF based AVS reduce average supply voltage from conventional WC design while keeping the target MTTF. For example, in Fig. 6(a), at a clock period of 3,300 ps, AVS with replica achieved the target MTTF at an average supply voltage of 0.82 V, whereas the conventional WC design required 0.90 V operation. In other words, replica based AVS achieved 9.0% V_{dd} reduction from 0.90 V to 0.82 V. Similarly, in Fig. 6(b) and Fig. 6(c), AVS with replica achieved 8.3% V_{dd} reduction from 1.20 V to 1.10 V at clock period of 480 ps and 7.5% V_{dd} reduction from 1.20 V to 1.11 V at clock period of 4,300 ps, respectively. As for AVS with TEP-FF, it achieved 13.3% V_{dd} reduction from 0.90 V to 0.78 V in ISP (Fig. 6(a)), 10.9% V_{dd} reduction from 1.20 V to 1.08 V in AES (Fig. 6(b)), and 12.5% V_{dd} reduction from 1.20 V to 1.05 V in OpenRISC (Fig. 6(c)), respectively. We experimentally confirmed that AVS with replica and TEP-FF made the significant voltage margin reduction both in ISP, AES, and OpenRISC at the cost of 0.1% area increase in ISP and OpenRISC and 1.0% in AES.

Next, we compare AVS with replica, AVS with TEP-FF, and the lower bound. Fig. 6 shows that AVS with TEP-FF further voltage reduction from AVS with replica. For example, AVS with TEP-FF achieved 4.8% V_{dd} reduction from 0.82 V to 0.78 V at the clock period of 3,300 ps in ISP, 2.0% V_{dd} reduction from 1.10 V to 1.08 V at 480 ps in AES, and 5.4% V_{dd} reduction from 1.11 V to 1.05 V at 4,300 ps in OpenRISC. This voltage reduction reveals that TEP-FF helps to exploit more timing margin than replica. TEP-FF converts the timing margin of intra-die random variation to V_{dd} reduction whereas replica needs to keep this margin, which will be discussed in Section 4.3. As for the comparison between AVS with TEP-FF and the lower bound, there is 2.5% difference between 0.78 V and 0.76 V in ISA, 5.6% difference between 1.08 V and 1.02 V in AES, and 4.8% difference between 1.05 V and 1.0 V in OpenRISC. These differences are well matched with the buffer delay in TEP-FF, e.g. it is comparable to the delay variation caused by 20 mV supply noise in ISP. From the above, we experimentally confirmed that TEP-FF kept the minimum timing margin in field operation and satisfied the MTTF constraint.

4.3 Discussion

This section discusses the difference between AVS with replica and AVS with TEP-FF investigating the impact of intra-die random variation on MTTF. We evaluate MTTF in a case that an identical set of paths are monitored by TEP-FF and replica. The difference of MTTF in this experiment is supposed to originate from how much the intra-die random variation can be considered by each sensor. As mentioned earlier, TEP-FF shares the intra-die variation with the main logic and hence it exploits the timing margin for the intra-die variation.

Fig. 7 shows the MTTF comparison between TEP-FF based AVS and replica based AVS in ISP. Note that in the MTTF calculation, we met cases where no timing errors occurred, i.e. MTTF is ∞. In the figure, we plotted the infinity MTTF as 10^{20} cycles to include it

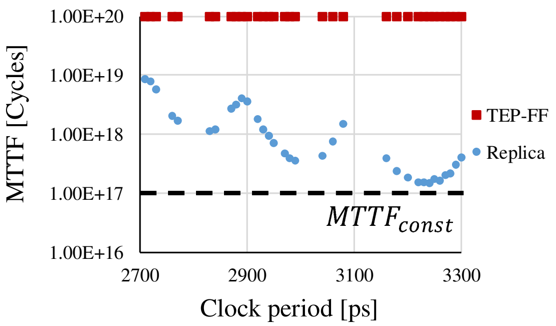

Figure 7: MTTF comparison between TEP-FF and replica (ISP).

in the figure. The number of inserted TEP-FFs is set to 483, and the inserted TEP-FFs sense 15,285 activated paths. Hence, the number of replicated paths is set to 15,285. Fig. 7 shows that the MTTF of TEP-FF based AVS is longer than that of replica based AVS even though the identical set of paths are monitored. Thus, we experimentally confirmed that TEP-FF more exploited the timing margin for the intra-die random variation to MTTF extension resulting in larger voltage reduction observed in the previous subsection.

5 CONCLUSION

This paper focused on timing sensors necessary for AVS implementation and compared in-situ timing error predictive FF (TEP-FF) and critical path replica in terms of achievable voltage margin reduction. For estimating the theoretical bound of ideal AVS, this work proposed linear programming based minimum voltage estimation and discussed the voltage adaptation performance quantitatively by evaluating the gap between the lower bound and actual average supply voltages. Experimental results showed that TEP-FF based AVS and replica based AVS achieved up to 13.3% and 8.9% supply voltage reduction, respectively, while satisfying target MTTF. Also, we experimentally confirmed that AVS with TEP-FF tracked the theoretical bound with 2.5 to 5.6 % voltage margin while AVS with replica needed 7.2 to 9.9 % margin.

ACKNOWLEDGEMENT

This work was partly supported by JSPS KAKENHI Grant Number JP18J12044.

REFERENCES

[1] B. Zhang and M. Orshansky, “Modeling of NBTI-induced pmos degradation under arbitrary dynamic temperature variation,” in *Proc. ISQED*, pp. 774–779, 2008.

[2] T. Wang and Q. Xu, “On the simulation of NBTI-induced performance degradation considering arbitrary temperature and voltage variations,” in *Proc. DAC*, pp. 1–6, 2014.

[3] S. Das, D. Roberts, S. Lee, S. Pant, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, “A self-tuning DVS processor using delay-error detection and correction,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 4, pp. 792–804, 2006.

[4] 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 45nm resilient microprocessor core for dynamic variation tolerance,” *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 194–208, 2011.

[5] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D. M. Harris, D. Blaauw, and D. Sylvester, “Bubble Razor: Eliminating timing margins in an ARM Cortex-M3 processor in 45 nm CMOS using architecturally independent error detection and correction,” *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 66–81, 2013.

[6] S. Kim and M. Seok, “Variation-tolerant, ultra-low-voltage microprocessor with a low-overhead, within-a-cycle in-situ timing-error detection and correction technique,” *IEEE Journal of Solid-State Circuits*, vol. 50, no. 6, pp. 1478–1490, 2015.

[7] A. Benhassain, F. Cacho, V. Huard, M. Saliva, L. Anghel, C. Parthasarathy, A. Jain, and F. Giner, “Timing in-situ monitors: Implementation strategy and applications results,” in *Proc. CICC*, pp. 1–4, 2015.

[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] S. Iizuka, Y. Masuda, M. Hashimoto, and T. Onoye, “Stochastic timing error rate estimation under process and temporal variations,” in *Proc. ITC*, 2015.

[10] A. Drake, R. Senger, H. Deogun, G. Carpenter, S. Ghiasi, T. Nguyen, N. James, M. Floyd, and V. Pokala, “A distributed critical-path timing monitor for a 65nm high-performance microprocessor,” in *Proc. ISSCC*, pp. 398–399, 2007.

[11] K. Chae and S. Mukhopadhyay, “All-digital adaptive clocking to tolerate transient supply noise in a low-voltage operation,” *IEEE Trans. CAS-II*, vol. 59, no. 12, pp. 893–897, 2012.

[12] S. Mhira, V. Huard, A. Benhassain, F. Cacho, S. Naudet, A. Jain, C. Parthasarathy, and A. Bravaix, “Dynamic adaptive voltage scaling in automotive environment,” in *Proc. IRPS*, pp. 3A-4.1-3A-4.7, 2017.

[13] V. Huard, F. Cacho, A. Benhassain, and C. Parthasarathy, “Aging-aware adaptive voltage scaling of product blocks in 28nm nodes,” in *Proc. IRPS*, pp. 7C-2-1-7C-2-7, 2016.

[14] Y. -G. Chen, T. Wang, K. -Y. Lai, W. -Y. Wen, Y. Shi, and S. -C. Chang, “Critical path monitor enabled dynamic voltage scaling for graceful degradation in sub-threshold designs,” in *Proc. DAC*, pp. 1–6, 2014.

[15] Y. Masuda and M. Hashimoto, “MTTF-aware design methodology of error prediction based adaptively voltage-scaled circuits,” in *Proc. ASP-DAC*, pp. 159–165, 2018.

[16] J. Kim, K. Choi, Y. Kim, W. Kim, K. Do, and J. Choi, “Delay monitoring system with multiple generic monitors for wide voltage range operation,” *IEEE Trans. VLSI Systems*, vol. 26, no. 1, pp. 37-49, 2018.

[17] J. Tschanz, K. Bowman, S. Walstra, M. Agostinelli, T. Karnik, and V. De, “Tunable replica circuits and adaptive voltage-frequency techniques for dynamic voltage, temperature, and aging variation tolerance,” in *Proc. VLSI symp.*, pp. 112-113, 2009.

[18] A. Teman, D. Rossi, P. Meinerzhagen, L. Benini, and A. Burg, “Controlled placement of standard cell memory arrays for high density and low power in 28nm FD-SOI,” in *Proc. ASP-DAC*, pp. 81–86, 2015.

[19] J. Shiomi, T. Ishihara, and H. Onodera, “Fully digital on-chip memory using minimum height standard cells for near-threshold voltage computing,” in *Proc. PATMOS*, pp. 44–49, 2016.

[20] J. B. Velamala, K. B. Sutaria, H. Shimizu, H. Awano, T. Sato, G. Wirth, and Y. Cao, “Compact modeling of statistical BTI under trapping/detrapping,” *IEEE Trans. Electron Devices*, vol. 60, no. 11, pp. 3645–3654, 2013.

[21] H. Awano, M. Hiromoto, and T. Sato, “Variability in device degradations: Statistical observation of NBTI for 3996 transistors,” in *Proc. ESSDERC*, pp. 218–221, 2014.

[22] H. Chang and S. S. Sapatnekar, “Statistical timing analysis under spatial correlations,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 9, pp. 1467–1482, 2005.

[23] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan, “First-order incremental block-based statistical timing analysis,” in *Proc. DAC*, pp. 331–336, 2004.

[24] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown, “Mibench: A free, commercially representative embedded benchmark suite,” in *Proc. IEEE Workshop on Workload Characterization*, pp. 3–14, 2001.

[25] Y. Li, X. Zhang, Z. Zhang, and Y. Lian, “A 0.45-to-1.2-V fully digital low-dropout voltage regulator with fast-transient controller for near/subthreshold circuits,” *IEEE Trans. Power Electronics*, vol. 31, no. 9, pp. 6341–6350, 2016.