# Representative Critical Reliability Paths for Low-Cost and Accurate On-Chip Aging Evaluation

Shuo Wang, Jifeng Chen, and Mohammad Tehranipoor
Dept of ECE, University of Connecticut
{shuo.wang, jic09003,tehrani}@engr.uconn.edu

## ABSTRACT

Aging of transistors degrades circuit performance and can potentially lead to functional failure in the field. This has become a major reliability concern especially when technology further scales to 45 nm and below. It is thus necessary to design on-chip structures that can provide accurate aging evaluation with no performance penalty. In this paper, we propose a novel methodology to accurately evaluate aging in the field. *Representative Critical Reliability Paths (RCRPs)* are synthesized as a stand-alone circuit to represent the aging of critical reliability paths, which are defined as paths that can potentially become critical at some point in time due to aging. By monitoring the RCRPs, aging of the critical reliability paths can be efficiently and accurately evaluated with no impact on the normal operation of the chip. The aging evaluation results can then be exploited to guide on-chip performance calibration to ensure lifetime reliability. Simulation results demonstrate the efficiency of the proposed structure.

**Categories and Subject Descriptors:** B.8.1 Performance and Reliability: Reliability, Testing, and Fault-Tolerance

**Keywords:** Circuit aging, Representative path, On-chip measurement, Path delay measurement

## 1. INTRODUCTION

As CMOS feature size shrinks down into deep nanometer regime, VLSI circuits are facing increasing challenge of reliability degradation [1] caused by negative/positive bias temperature instability (NBTI/PBTI), hot carrier injection (HCI), and time-dependent dielectric breakdown (TDDB), which cause parametric shifts and eventually device failure. One-time worst-case guardbanding at design stage such as gate sizing is either inadequate or otherwise over-pessimistic that leaves a lot of performance on the table [2], as circuit can age at different rates depending on several factors, especially workload that is unknown at design stage. Thus, it is necessary to monitor the aging on the chip and react dynamically.

Many existing solutions try to directly monitor aging of functional paths by inserting sensors into flip-flops on them [4, 5, 6]. However, methods in [4, 5] are only able to detect serious aging and may therefore miss early opportunities for optimal compensation. In our prior work [6], an aging sensor is proposed to directly measure path delay and degradation during normal operation. However, similar to [4, 5], it requires modification on the flip-flops and has impact on the paths that are being monitored. Another approach [3] evaluates aging based on periodical self-test using patterns pre-stored in off-chip nonvolatile memory. However, interruption to the normal execution as well as non-trivial storage for test patterns incur inevitable performance and area overhead.

Other existing dynamic solutions are based on on-chip structures that monitor aging of a stand-alone circuit, such as ring-oscillators [7] and replica paths [17, 12]. These methods offer small area overhead and do not introduce performance penalty to the functional circuit. However, ring-oscillators do not represent the functional circuits' structure and workload, whereas replica paths can only help monitor a very limited number of selected paths but leaving all the other paths uncovered. As a result, aging estimation and hence the calibration made accordingly may be inaccurate.

In this paper, we propose a novel methodology to accurately evaluate aging on the chip. Instead of inserting sensors on the functional paths, our method tries to synthesize a stand-alone circuit that has minimum impact to the functional circuit. This stand-alone circuit, referred to as *Representative Critical Reliability Paths (RCRPs)*, is not based on either ring-oscillators or a simple replica of a small number of selected paths. Rather, RCRPs are synthesized so that they can accurately represent the "critical reliability paths" in the functional circuit that are defined to have the potential to become critical at some point in time due to aging in the field. The key for synthesizing such RCRPs is not to predict what paths are critical at time 0 or at time $t$. Rather, it is to exploit the topology and the workload of the critical reliability paths so that a small number of RCRPs can represent and closely track what is going on in terms of aging in the functional circuit.

First, topology of the critical reliability paths are captured by identifying important delay segments that are shared among many critical reliability paths and therefore their degradation has potentially huge impact to the overall reliability of the circuit. Thus, a very small number of RCRPs can be synthesized based on these important delay segments and the critical reliability paths that contain them. Second, workload of these important delay segments can also be selectively sampled to help the RCRPs even more accurately represent the critical reliability paths, if it is needed. To


*IEEE/ACM International Conference on Computer-Aided Design (ICCAD) 2012,* November 5-8, 2012, San Jose, California, USA.

minimize performance impact, workload is only indirectly sampled from non-critical paths that also share the workload of these important delay segments.

As a result, by monitoring delay degradation of the stand-alone RCRPs, aging, including recovery effects (e.g., in sleep mode), of the critical reliability paths in the functional circuit can be efficiently and accurately evaluated with minimum impact on the normal operation. Note that RCRP is different from "representative critical path" in the literature (e.g., [14, 15]) for post-silicon delay prediction. Rather than trying to accurately estimate as many target paths as possible only at time zero, RCRPs aim at always being able to track the largest delay among the critical reliability paths throughout the lifetime of operation.

Once aging is evaluated by RCRPs, adaptive body bias (ABB) and supply voltage (ASV) [16], or dynamic frequency scaling (DFS) [8] can be used to calibrate the circuit in order to ensure reliability throughout the lifetime based on the observed aging. Note that RCRP circuit will not impose any constraint on physical design and can be placed anywhere in the circuit layout. We acknowledge that RCRPs would better represent critical reliability paths if they are placed at locations where the voltage and temperature are similar to what the critical reliability paths are experiencing.

The rest of the paper is organized as follows. Section 2 presents the basic concept of RCRP and aging evaluation. In Sections 3 and 4, problem formulation and RCRP synthesis are presented, respectively. Section 5 discusses the simulation results and Section 6 concludes this paper.

## 2. CONCEPT OF RCRP

We use Fig. 1 to demonstrate the basic concept of RCRP. The objective is to use a small number of stand-alone paths to statistically estimate the delay and aging of a large number of critical reliability paths in the chip. More importantly, the largest delay among the critical reliability paths at any time point during the entire lifetime can be accurately estimated from the measurements of these stand-alone paths. Suppose that, among many functional paths, paths $p_1$, $p_2$, and $p_3$ stand out as three critical reliability paths at different time points, e.g., path $p_1$ has the largest delay from time 0 up to time $t_1$, path $p_2$ has the largest delay during $[t_1, t_2]$, and path $p_3$ becomes the most critical after time $t_2$. This phenomenon results from their different aging rates as they may have different structures (gates, interconnects, load capacitance, etc.) and process variations, and experience different workloads. As the workload is usually unknown at design stage, it is extremely difficult to predict the circuit aging and criticality.

We design RCRPs in a way that, by measuring RCRPs' delay using on-chip structures such as [10, 11, 17], estimation on the largest delay can always closely track that of the critical reliability paths. This enables adaptive calibration methods to be performed in a timely manner for ensuring lifetime reliability. It is a cost-efficient solution compared with inserting monitors all over the critical reliability paths, as only a small number of RCRPs need to be implemented and monitored, plus the impact on the critical reliability paths is minimized.

We acknowledge that RCRPs cannot capture effects such as crosstalk that critical paths experience but temperature and power supply noise in the circuit can be taken into account by the RCRPs. Analyzing the impact of these two effects on RCRPs is part of our future work.

## 3. PROBLEM FORMULATION

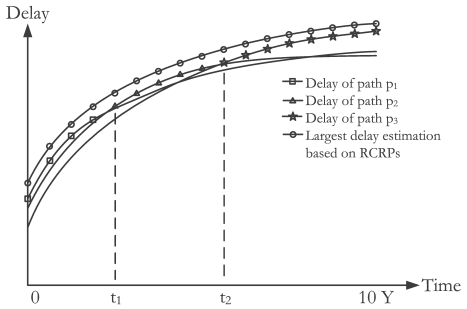


**Figure 1: Demonstration of the RCRP concept.**

Suppose there are $m$ critical reliability paths, each of which consists of delay segments from on-path gate input to output plus interconnect to the next gate. If there are totally $n$ unique segments on these critical reliability paths, we can then use an $m \times n$ matrix $P = \{p_1, p_2, ..., p_m\}^T$ to denote the paths, where each row vector $p_i$ refers to a specific path among the group and is in the form of a series of 1s and 0s, indicating whether or not a specific segment is on the path. For example, if $p_i = [0, 0, 1, ...]$, it means segments 1 and 2 are not on path $p_i$ while segment 3 is. If the delay of segment $j$ is $x_j$, we can use a vector $X = [x_1, x_2, ..., x_n]^T$ to denote the segments delay. Thus, if the delay of each path $p_i$ is $d_i$, delay of the $m$ paths can be expressed as $d = [d_1, d_2, ..., d_m]^T = PX$.

We then try to build $r$ RCRPs based on the $m$ critical reliability paths in $P$ ($r << m$). That is, $P_R = \{p'_1, p'_2, , p'_r\}^T$, where $p'_j \in P$. Similarly, the delay measurement of each RCRP can be expressed in the form of $d_R = [d'_1, d'_2, ..., d'_r]^T = P_R X$. Thus, we can estimate the delay $d$ based on $d_R$. The delay estimation $\widehat{d}$ and estimation error $Error$ can be calculated respectively as below:

$$\widehat{d} = PP_R^T(P_R P_R^T)^{-1} d_R, \tag{1}$$

$$Error = \widehat{d} - d = P[P_R^T(P_R P_R^T)^{-1} P_R - I]X. \tag{2}$$

where $()^{-1}$ denotes the inverse matrix and $I$ denotes the unit matrix that has ones on the main diagonal and zeros elsewhere. For the purpose of accurate aging evaluation, the objective is to minimize the estimation error at all time points $t$ under various work conditions (e.g., workload, temperature, etc.) for the given area budget.

We build the RCRPs based on factorization of the path matrix $P$. Specifically, after performing singular value decomposition, we can obtain $P = U \times S \times V$, where $U$ and $V$ are $n \times n$ and $m \times m$ orthogonal matrices, respectively, and $S$ is an $n \times m$ diagonal matrix consisting of the eigenvalues $\{\lambda_i\}$ of $P$ ranked in the descending order, i.e., $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_i \geq \lambda_{i+1} \geq ... \geq 0$. The $r$ most significant eigenvalues $\{\lambda_i\}_{i=1}^r$ identifies the $r$ most important row vectors (i.e., paths) that can be used to represent the path matrix $P$ (i.e., the entire circuit structure). Therefore, we can use QR decomposition with column pivoting to select the corresponding row vectors, and in other words, the paths $p'_1, p'_2, ..., p'_r$, as RCRPs to represent the original paths in matrix $P$. This mathematic tool is also used in the related work [13, 15] and many other applications in signal processing and statistics. According to [13], the estimation error is bounded by a function of $\eta = 1 - \frac{\sum_1^r \lambda_i}{\sum \lambda_i}$ [13]. The selected $r$ RCRPs are then implemented as stand-alone circuit and are used to monitor circuit aging in the field.

Delay measurements on RCRPs are performed at time 0

and at any desired time point $t$ in the field. Let the estimation error at time 0 and time $t$ be $Error|_0$ and $Error|_t$, respectively. According to (2),

$$Error|_0 = P[P_R^T(P_R P_R^T)^{-1}P_R - I]X|_0, \quad (3)$$

$$Error|_t = P[P_R^T(P_R P_R^T)^{-1}P_R - I]X|_t, \quad (4)$$

$$X|_t = A|_t X|_0, \quad (5)$$

where $A|_t$ is a $m \times m$ diagonal matrix. Each element on the diagonal $a_i|_t$ is greater than 1 indicating the aging of segment $i$ at time $t$ compared with time 0. The delay difference between RCRPs estimation and the actual delay of critical reliability paths at time 0 is used to compensate for the systematic mismatch as a combined effect of non-ideal matrix factorization (when $\eta < 1$), process variations, and the delay difference between the interconnects in the RCRPs and the counterpart in the critical reliability paths. As these effects are unlikely to change over time due to aging, error at time 0, i.e., $Error|_0$, can be used to compensate for the systematic error in future time points, so that the adjusted error at time $t$, i.e., $adjError|_t$, can be calculated as

$$\begin{aligned} adjError|_t &= Error|_t - Error|_0, \\ &= P[P_R^T(P_R P_R^T)^{-1}P_R - I][A|_t - I]X|_0. \end{aligned} \quad (6)$$

Compared with $Error|_t$ in (4), $adjError|_t$ is clearly expected to be smaller as each diagonal element in matrix $[A|_t - I]$ is smaller than 1. For example, if $a_i|_t = 1.2$, which indicates a 20% degradation, $a_i|_t - 1 = 0.2 < 1$. Thus, even if the error at time 0 is relatively large when only a small number of $r$ RCRPs is allowed due to area budget, adjusted error at time $t$ can be largely reduced.

Note that the above-discussed error compensation method is based on the assumption that delay of the functional paths at time 0, and hence $Error|_0$, can be measured during manufacturing test. When only the largest delay is obtained during speed binning process, the adjusted error will not be the same as the ideal form in Equation (6). However, the adjusted error can still be improved for estimating the largest delay among the critical reliability paths. In this paper, we assume that only the largest delay is obtained at time 0 for the error compensation. Thus, no extra constraint is imposed on the functional path delay measurement, as the major objective of RCRPs is to estimate the largest delay on the chip.

In summary, the problem of building $r$ RCRPs for reliability monitoring can therefore be modeled as finding the optimal $r$ with area budget to minimize the mean square error (MSE) of adjusted error, especially for the largest delay, under various workload, temperature, and at different time $t$. That is:

$$\textbf{Minimize: } MSE(adjError|_t), max(adjError|_t) \quad (7)$$
$$\forall p_i, t, workload, temperature,$$

$$\textbf{Subject to: } \textbf{Area}_{RCRP} \leq \textbf{Area}_{budget}. \quad (8)$$

# 4. RCRP SYNTHESIS

The RCRPs' synthesis flow shown in Algorithm 1 can be performed to synthesize the RCRPs according to (7) and (8). The complexity mainly comes from the aging-aware delay analysis in line 11 to compare RCRPs' estimation and actual delay from the functional circuit, the singular value decomposition in line 4, and the QR-decomposition in line 6.

As aging-aware Spice simulation is extremely time consuming, we developed an in-house tool for aging-aware path delay analysis at design time [18]. This tool generates a lookup-table (LUT) to store delay degradation information for each gate type of a standard cell under various operation conditions. Thus, time-consuming Spice simulation is only run once for a given technology library. The LUT is then reused to quickly calculate path delay under aging. As a result, the tool provides greater than 200X speedup compared to HSPICE and still maintains a high accuracy of over 99%.

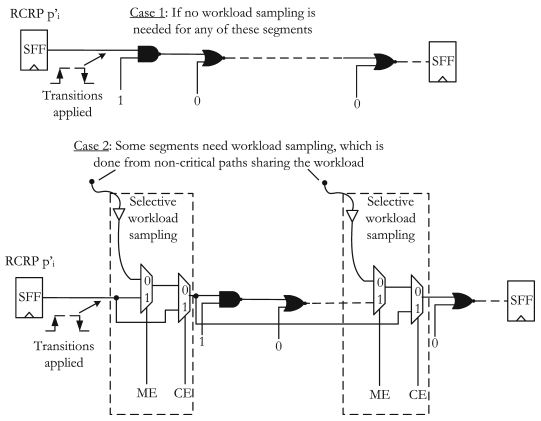


**Figure 2: Two cases of RCRP implementation.**

```
    Input: P, Area_budget
    Output: r, P_R
 1  begin
 2      Initialize r to a small value
 3      while Area_RCRP ≤ Area_budget do
 4          Singular value decomposition: [U, S, V] = svd(P)
 5          Use the first r columns in U to form U_r = U(:, 1 : r)
 6          QR-decomposition, i.e., [Q, R, K] = qr(U_r)
 7          P_n = K^T P
 8          Use the first r rows in P_n to form P_R = P_n(1 : r, :)
 9          Synthesize RCRPs according to P_R
10          Obtain Area_RCRP
11          Perform aging-aware delay analysis to evaluate
            adjError|_t and MSE(adjError|_t)
12          if either max(adjError|_t) or MSE(adjError|_t)
            gets improved then
13              r = r + 1
14          else
15              break out of the while loop
16          end
17      end
18      return the optimal r and P_R for eventual RCRPs
19  end
```

**Algorithm 1:** RCRPs synthesis flow.

Complexity of the synthesis flow mainly comes from the singular value decomposition for matrix $P$. A "divide-and-conquer" approach can be used to reduce the computational overhead for industrial designs that have a huge pool of critical reliability paths by dividing them into multiple groups, applying the RCRPs synthesis flow on each group in parallel, and eventually identifying and synthesizing RCRPs for all the critical reliability paths.

Once the eventual $P_R$ is obtained, RCRPs can be implemented. Two cases of the implementation are shown in Fig. 2 for demonstration. In Case 1, the designers decide that no workload sampling is needed (how to decide whether workload sampling is needed will be discussed in Section 5). Thus, segments on the RCRP can be directly connected with each other. Transitions generated at the input (e.g., at a duty ratio of 50%) can propagate through all the segments in

order to age the RCRP and to measure its delay. In Case 2, if it is decided that some of the segments require workload sampling from the functional circuit, minimum-sized buffers and MUXes are inserted in front of these segments, so that the stand-alone RCRP can share the same workload that the corresponding segments in the functional circuit are experiencing in the field. Specifically, minimum-sized buffers are used to minimize capacitive load effect on the functional circuit, and to avoid impact on the most critical reliability paths selective workload is only indirectly sampled from non-critical reliability paths that share the same workload of the important segment. MUXes inserted in Case 2 shown in Fig. 2 provide three modes of operations: (i) Stress mode (ME=0 and CE=0), where RCRP $p_i'$ is stressed under similar workload that functional path is experiencing; (ii) Measurement mode (ME=1 and CE=0), where transitions are applied at the input of $p_i'$ so that the delay of this RCRP, including the MUXes' delay, can be measured; and (iii) Calibration mode (ME=1 and CE=1), where transitions are applied and delay of the MUXes can be obtained. Thus, by subtracting measurement results in (iii) from that in (ii), delay and aging of the MUXes will be excluded and delay of RCRP can then be accurately obtained. Next, aging of all the critical reliability paths can be estimated according to Equation (1).

In both cases, scan flip-flops are used at the beginning and end of the RCRPs to ensure their testability at time 0. Measurement circuit will also be implemented (not shown in Fig. 2) to measure the path delay and its degradation of the RCRPs at different time points. Among many options for the measurement circuit, Vernier delay line and its variations (e.g. [10]) or time-to-digital converter (e.g., [17]) can be implemented. Besides, the RCRPs can also be reconfigured into ring-oscillators during measurement [11]. These options have different area overheads and measurement resolutions. Specific design choice is beyond the scope of this paper.

To determine whether to sample actual workload for a specific segment, we count among how many functional paths ($S$) it is shared and set a threshold value $S_{TH}$. If $S \geq S_{TH}$, this segment is likely to be important, as its aging affects many critical reliability paths at the same time. Thus, actual workload can be extracted for RCRP as shown in Fig. 2 indirectly from non-critical paths. Otherwise, workload is not extracted for complexity and area overhead consideration. Note that, as observed from our results shown in Section 5, selective workload sampling is not always necessary for every circuit especially if the simulation and analysis indicates so in the design stage.

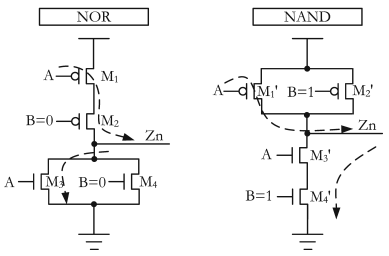


**Figure 3: Off-path workload analysis.**

Meanwhile, non-controlling values are applied to off-path inputs to simplify workload controlling for the RCRPs. This also provides an additional benefit in that the worst-case bias temperature instability (BTI) stress is applied to the off-path inputs. For example, as shown in Fig. 3, a 0 (1) for NOR-gate (NAND-gate) provides the worst-case NBTI (PBTI) stress on the PMOS transistor $M_2$ (NMOS transistor $M_4'$) for the off-path pin $B$, which then exacerbates performance degradation of this NOR-gate (NAND-gate) on top of the contribution from on-path pin. This gives a relatively conservative estimation on the largest delay among the critical reliability paths as a result. It is noteworthy that, although a 1 (0) for NAND-gate (NOR-gate) gives the minimum NBTI (PBTI) stress on the PMOS transistor $M_2'$ (NMOS transistor $M_4$) for the off-path pin, $M_2'$ ($M_4$) does not affect the propagation delay from input $A$ to output $Z_n$.

# 5. RESULTS AND ANALYSIS

In this section, we evaluate the effectiveness of the proposed method in circuit simulation. The simulation is performed on 45-nm technology using the open-source Nangate library. Aging-aware LUT [18] is generated using commercial tool HSPICE's MOS Reliability Analysis (MOSRA), which modeled the NBTI, PBTI, and HCI effects, and is then reused to facilitate delay and aging analysis. Several benchmark circuits are used in the experiments.

The optimal RCRPs are generated using the method discussed in Section 4 according to the area budget. We first constrain the area overhead of RCRPs (measurement circuit is not included) to be 1% of the design and will later use different constraints to see its impact. Different workload scenarios and temperatures are applied. Note that process variations, crosstalk, and power supply noise are not included in the simulations in order to focus on the aging variations. However, as discussed earlier, time 0 measurement can largely compensate for process variations and impact of interconnects. Besides, a design margin can be introduced to account for other effects that RCRPs may not capture. Thus, we use mainly the adjusted error and the mean square error (MSE) to evaluate the accuracy of the RCRP-based aging evaluation method. The adjusted error is obtained by comparing the actual largest delay among the critical reliability paths with the RCRPs' estimation that is adjusted using the time 0 measurement results. Meanwhile, MSE is calculated over all the critical reliability paths, time points of measurement in a 10-year span, and various temperature and workload conditions. Therefore, adjusted error focuses on the RCRP's accuracy in tracking the largest delay emerged from the functional circuit; whereas MSE focuses on the RCRP's capability of representing/covering all the critical reliability paths.

We implement RCRPs for six benchmark circuits $s5378$, $s9234$, $s13207$, $s15850$, $s38417$, and $b15$. Different workload and temperature scenarios are generated for each benchmark circuit. Specifically, workload $(WL) = 25\%$, $50\%$, and $75\%$ are generated as three different workload scenarios, where the value of $WL$ is defined as the percentage of being 1 at the primary inputs of the circuits. Two different temperature settings are $75^oC$ and $125^oC$. RCRPs along with the functional circuit are aged for 10 years and measurements are taken for the RCRPs every 2 years ($t_i = 0, 2, 4, 6, 8$, and 10 years). Note that any other measurement time step can also be considered. The CPU runtime for building and evaluating RCRPs on these benchmark circuits ranges from a couple of minutes (e.g., for $s5378$) to about 10 minutes for $s38417$ and about 35 minutes for $b15$ on a computer equipped with a 2.3-GHz quad-core processor. Among these benchmark circuits, $b15$ is a subset of 80386 processor and has over 6873 critical reliability paths, CPU runtime of which shows that the proposed RCRP technique can be applied to large circuits.

We first disable the selective workload sampling described in Section 4 and look at the adjusted error $adjError|_t$ at

**Table 1: Area overhead and estimation accuracy of RCRPs when $Area_{budget} = 1\%$.**

| Circuit | m | r | Gate Cnt. | Actual Area O/H (%) | Range of adjError (%) |
|---|---|---|---|---|---|
| s5378 | 14 | 1 | 16 | 0.6 | -2.0 ∼ 0 |
| s9234 | 459 | 1 | 16 | 0.7 | 0 ∼ 7.0 |
| s13207 | 28 | 1 | 21 | 0.6 | 0 ∼ 4.0 |
| s15850 | 962 | 2 | 55 | 0.7 | 0 ∼ 3.4 |
| s38417 | 4726 | 7 | 116 | 0.8 | -0.5 ∼ 0.4 |
| b15 | 6873 | 1 | 67 | 0.5 | 0 ∼ 1.8 |

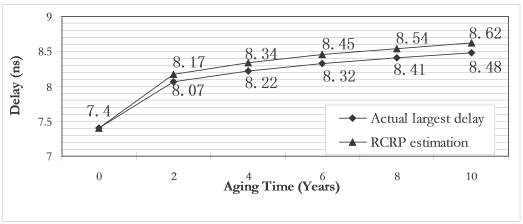


**Figure 4: Actual largest delay vs. RCRP estimation in b15 when WL=50% and temperature is $75^oC$.**

different time points for each benchmark circuits' RCRPs. In the basic setting, maximum area budget for the RCRPs is only allowed to be 1% of the entire circuit.

Table 1 shows the area overhead and estimation accuracy for RCRPs in the 6 selected benchmark circuits. $m$ in Column 2 shows the number of critical reliability paths in the benchmarks. They are obtained using our in-house tool by selecting paths whose delay if degraded by 20% during the course of aging would possibly cause failure (as path delay could degrade as much as 20% over a period of ten years [9]). The optimal $r$ obtained for RCRPs using the flow in Algorithm 1 is listed in Column 3. The resulting gate count and area overhead are shown in Columns 4-5, respectively. The percentage range of adjusted error for the largest delay are then calculated and listed in Columns 6.

As can be observed in Table 1, the estimation accuracy of RCRPs is not equal for different designs. However, the adjusted error that evaluates how accurate RCRPs are tracking the largest delay is generally small for these designs. In particular, for $s38417$, which has over 10000 gates and 1400 flip-flops, 7 RCRPs are allowed according to the area budget. This leads to only a $-0.5\% \sim 0.4\%$ adjusted error at any time point and under any work condition. Another example is $b15$, which has a relatively large number of critical reliability paths. Due to the tight area budget, only one RCRP ($r = 1$) is selected to represent 6873 critical reliability paths. Still, under all the workload and temperature conditions, the adjusted error is found to be within 1.8%. Even in the worst case of $s9234$, the adjusted error is within 7% even without selective workload sampling (and hence without the extra buffers, routing, and MUXes shown in Fig. 2). It can be observed from results shown later that selective workload sampling can further reduce the adjusted error of $s9234$ to as low as 3.1%. Therefore, RCRPs indeed can accurately track the largest delay among the critical reliability paths.

Note that the ranges of adjusted error in Table 1 are obtained for all time points ($t_i = 0, 2, 4, 6, 8$, and 10 years) under various work conditions ($WL = 25\%, 50\%, 75\%$ and $temperature = 75^oC, 125^oC$). A specific example of $b15$ is also shown in Fig. 4, where the workload is $WL = 50\%$ and temperature is $75^oC$. Actual largest delay among the critical reliability paths and that based on RCRP estimation (after being adjusted for time 0 mismatch) are shown at different time points in the 10-year span. It can be seen that the RCRPs consistently track the largest delay of the critical reliability paths at all time points. The same trend was also

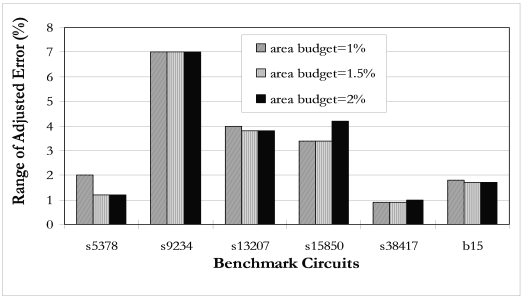


**Figure 5: Range of adjusted error under different area budgets.**

found in different workload conditions and temperatures as well as in other benchmark circuits.

Next, we further study RCRPs under various area budget requirements. Three different area budget requirements are used: 1%, 1.5%, and 2%. In each scenario, the optimal RCRPs are synthesized using the flow in Algorithm 1.

The range of adjusted error (i.e., the difference between the maximum and the minimum adjusted errors) is shown in Fig. 5. In addition, normalized MSE for RCRPs under different area budgets is shown in Fig. 6, where the values are normalized using MSE when area budget is 1%. From the results we can still see the general trend for each individual circuit that, when the area budget increases, more RCRPs are allowed; therefore, estimation tends to be more accurate. It is noteworthy that, in the case of $s5378$, $s9234$, and $s13207$, when budget increases from 1.5% to 2%, neither MSE nor adjusted error can be improved simply by implementing one more RCRP. Thus, the optimal RCRPs remain the same for the two area budget requirements in these cases. In other words, even when area budget raises to 2%, the area overhead of actual RCRPs is still within 1.5%. Meanwhile, when area budget increases from 1.5% to 2% for $s15850$ and $s38417$, one more RCRP is implemented. As a result, adjusted error slightly increases. This indicates that the additional RCRP implemented does not help improve the accuracy on tracking the largest delay in the functional paths. However, it is still considered worthwhile to implement this additional RCRP, as the MSE improves significantly in both designs. In other words, RCRPs become more capable of representing the entire group of critical reliability paths, thereby largely improving the estimation confidence.

Furthermore, we study why RCRPs behave differently among some of the benchmark circuits. We define *critical reliability paths coverage* as the percentage of critical reliability paths that are directly or indirectly represented by the RCRPs. We find, in fact, that the critical reliability paths coverage is different across these benchmarks because of their different topology: If there are too few overlaps between the critical reliability paths and the area budget is tight, there is a good chance that not all critical reliability paths are actually covered (directly/indirectly represented) by RCRPs. As shown in Fig. 7, RCRPs for the other four benchmark circuits can easily be close to or at 100%, which explains why the adjusted error and MSE are both relatively smaller. However, the critical reliability paths coverage ratios for $s9234$ and $s13207$ are below 80%, making it difficult for accurate estimation from too few RCRPs. In large designs, we expect that more RCRPs are allowed and hence may have a better chance of achieving higher coverage and estimation accuracy.

Lastly, we study the impact of selective workload sampling. We take $s9234$ and $s13207$ as two examples. As ex-

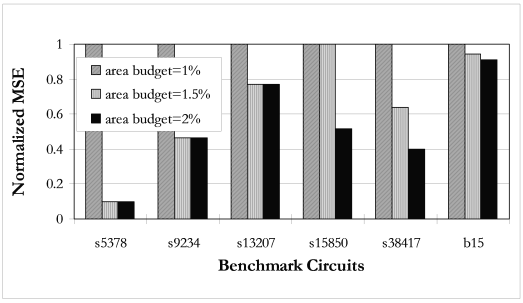


**Figure 6: Normalized MSE under different budgets.**

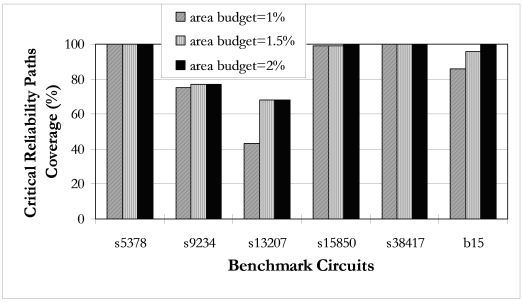


**Figure 7: Critical reliability paths coverage under different area budgets.**

plained in Section 4, selective workload sampling is implemented for important segments based on different threshold level $S_{TH}$. The results are shown in Table 2. When there is no workload sampling, i.e., sampling count in Column 5 is 0, the area overhead of RCRPs is within 2% of the total area of the design. When $S_{TH}$ is 3 (Column 4) in the case of $s9234$ for example, actual workload from functional paths can be sampled for totally just 3 gates on the RCRPs (Column 5). This leads to slight improvement on the adjusted error shown in Column 6 (reduced from 7.6% to 6.9%). When $S_{TH}$ reduces, more workload can be extracted, thereby further reducing the adjusted error. When $S_{TH} = 1$, 32 out of 46 gates on the RCRPs can extract actual workload. This reduces the adjusted error to 3.1%. Yet, as selective workload sampling only takes place in the non-critical paths, its impact on circuit performance is minimized.

In comparison, we observed that without selective workload sampling, even if 10 more RCRPs are implemented for this design, it only improves the MSE while the improvement on adjusted error is marginal. However, selective workload sampling will increase the area overhead by introducing MUXes on the RCRPs as shown in Fig. 2. Nevertheless, this approach provides an option for high estimation accuracy.

It should be noted that selective workload sampling does not always improve adjusted error as can be seen from the results of $s13207$, where adjusted error stays the same even if 19 gates on the RCRPs are getting actual workload. If this is observed during RCRP synthesis and analysis at design stage, the designer can simply remove the unnecessary sampling from the synthesis.

# 6. CONCLUSIONS AND FUTURE WORK

The proposed representative critical reliability paths methodology provides an efficient in-the-field approach to evaluate aging on the chip throughout the lifetime with no negative impact on the functional circuit. Simulation results demonstrate the effectiveness of this method. Our future work will be directed towards a more comprehensive solution so that effects such as crosstalk and power supply noise can also be taken into account for more accurate estimation. Finally, we plan to fabricate a test chip to include RCRP for the circuit and perform silicon data collection and aging analysis under various workload and temperature.

**Table 2: Examples of selective workload sampling.**

| Circuit | $r$ | RCRPs' Gate Cnt. | $S_{TH}$ | Sampling Cnt. | Range of adjError (%) |
|---|---|---|---|---|---|
| $s9234$ | 3 | 46 | $\geq 4$ | 0 | 7.6 |
| | | | 3 | 3 | 6.9 |
| | | | 2 | 11 | 5.0 |
| | | | 1 | 32 | 3.1 |
| $s13207$ | 3 | 65 | $\geq 2$ | 0 | 3.8 |
| | | | 1 | 19 | 3.8 |

# 7. ACKNOWLEDGEMENTS

This work is supported in part by Semiconductor Research Corporation (SRC) under grants 2053 and 2094, and a gift from Cisco.

# 8. REFERENCES

[1] J. W. McPherson, "Reliability challenges for 45nm and beyond," *Design Automation Conference*, pp. 176–181, 2006.

[2] E. Mintarno et al., "Optimized self-tuning for circuit aging," *Design, Automation and Test in Europe Conference*, pp. 586–591, 2010.

[3] Y. Li, S. Makar, and S. Mitra, "CASP: concurrent autonomous chip self-test using stored test patterns," *Design, Automation and Test in Europe*, pp. 885–890, 2008.

[4] M. Agarwal, B. C. Paul, M. Zhang, and S. Mitra, "Circuit failure prediction and its application to transistor aging," *IEEE VLSI Test Symposium*, pp. 277–286, 2007.

[5] J. C. Vazquez et al., "Low-sensitivity to process variations aging sensor for automotive safety-critical applications," *IEEE VLSI Test Symposium*, pp. 238–243, 2010.

[6] S. Wang, M. Tehranipoor, and LR Winemberg, "In-field aging measurement and calibration for power-performance optimization," *Design Automation Conference*, pp. 706–711, 2011.

[7] T.-H. Kim, R. Persaud, and C. H. Kim, "Silicon odometer: an on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 874–880, 2008.

[8] S. Herbert and D. Marculescu, "Analysis of dynamic voltage/frequency scaling in chip-multiprocessors," *International Symposium on Low Power Electronics and Design*, 2007.

[9] W. Wang et al., "The impact of NBTI on the performance of combinational and sequential circuits," *Design Automation Conference*, pp. 364–369, 2007.

[10] E. Saneyoshi, K. Nose, and M. Mizuno, "A precise-tracking NBTI-degradation monitor independent of NBTI recovery effect," *IEEE International Solid-State Circuits Conference*, pp. 192–193, 2010.

[11] X. Wang, M. Tehranipoor, and R. Datta, "A novel architecture for on-chip path delay measurement," *International Test Conference*, pp. 1–10, 2009.

[12] J. Tschanz et al., "Tunable replica circuits and adaptive voltage-frequency techniques for dynamic voltage, temperature, and aging variation tolerance," *Symposium on VLSI Circuits*, pp. 112–113, 2009.

[13] D. Chua, E. Kolaczyk, and M. Crovella, "Network kriging," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 12, pp. 2263–2272, 2006.

[14] Q. Liu and S. S. Sapatnekar, "Synthesizing a representative critical path for post-silicon delay prediction," *International Symposium on Physical Design*, pp. 183–190, 2009.

[15] L. Xie and A. Davoodi, "Representative path selection for post-silicon timing prediction under variability," *Design Automation Conference*, pp. 386–391, 2010.

[16] J. Tschanz, S. Narendra, R. Nair, and V. De, "Effectiveness of adaptive supply voltage and body bias for reducing the impact of parameter variations in low power and high performance microprocessors," *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 826–829, 2003.

[17] M. Chen et al., "A TDC-based test platform for dynamic circuit aging characterization," *IEEE Internaltional Reliability Physics Symposium*, pp. 2B.2.1–2B.2.5, 2007.

[18] J. Chen, S. Wang, M. Tehranipoor, "Efficient selection and analysis of critical-reliability paths and gates," *ACM Great Lakes Symposium on VLSI*, pp. 45–50, 2012.