Temporal Performance Degradation under RTN: Evaluation and Mitigation for Nanoscale Circuits

Hong Luo*, Yu Wang*, Yu Cao†, Yuan Xie‡ and Yuchun Ma§ and Huazhong Yang*
*Dept. of E.E., TNList, Tsinghua Univ., Beijing, China
Email: hongluo@tsinghua.edu.cn, yu-wang@tsinghua.edu.cn
†Dept. of ECEE., Arizona State Univ., USA
‡Dept. of CSE, Pennsylvania State Univ., USA
§Dept. of C.S., TNList, Tsinghua Univ., Bejing, China

***Abstract*—Random telegraph noise (RTN) is one of the critical reliability concerns in nanoscale circuit design, and it is important to consider the impact of RTN on the circuits' temporal performance. This paper proposes a framework to evaluate the RTN-induced performance degradation and variation of digital circuits, and the evaluation results show that RTN can result in 54.4% degradation and 59.9% variation on the circuit delay at 16nm technology node. Power supply tuning and gate sizing techniques are investigated to demonstrate the impact of such circuit-level techniques on mitigating the RTN effect.**

***Index Terms*—Random telegraph noise, Performance degradation, Mitigation technique**

I. Introduction

In recent years, a variety of fluctuations, such as V_{th} fluctuation and I_d fluctuation have attracted attention, as the channel length of MOSFET devices continue to shrink into the nano-scale regime. Random telegraph noise (RTN) can cause electrical parameters (such as V_{th} and I_d) to exhibit random fluctuations as a function of time. Recent studies show that the fluctuation due to RTN becomes quite large and can be more significant than the Random Dopant Fluctuation (RDF) under 22nm regime [1]. For example, the drain current fluctuation induced by random telegraph noise (RTN) has already been identified as a large obstacle in both sub-V_{th} and super-V_{th} operation of digital circuits [2]. The variation of I_d due to RTN can be 40% for $30 \times 30\text{nm}$ devices [3].

The physics of RTN has been widely investigated [2]–[5], and the RTN effect on the memories has also been studied [6]–[10]. Though some models which can be integrated into HSPICE analysis are proposed in [11]–[13], the impact of RTN effect on the digital circuits' temporal performance has been rarely studied [14]. Therefore, our contribution in this paper distinguishes itself in the following aspects:

- This paper proposes a framework for evaluating the impact of RTN on the circuits' temporal performance. In this framework, the "**sampling**" and **statistical critical path analysis** techniques are used to estimate the distribution of the delay fluctuation, and the "**grouping**" technique is applied to reduce the complexity of the probability computation, which reduces the complexity from $O(2^N)$ to $O(N)$.
- The impact of RTN on circuit delay degradation and variation is investigated. The experimental results show that RTN will degrade the circuit delay, and increase the delay variation. The average delay degradation is 34.6%, and the variation is 32.1% at 16nm technology node. The results also demonstrate that the performance degradation and variation will grow rapidly with technology and power supply voltage scaling down.
- Two design techniques, **power supply tuning** and **gate sizing**, are applied for RTN mitigation. The impact of these techniques on the circuit degradation and variation are investigated. The simulation results show that both techniques can reduce the temporal degradation and variation.

The rest of this paper is organized as follows. Section II reviews previous work on RTN, and Section III introduces the model used in this paper. Section IV proposes the statistical critical path analysis technique and the evaluation framework. The RTN-induced temporal performance degradation and variation in digital circuits are also evaluated. The impact of design techniques on RTN mitigation is investigated in Section V.

II. Related Work

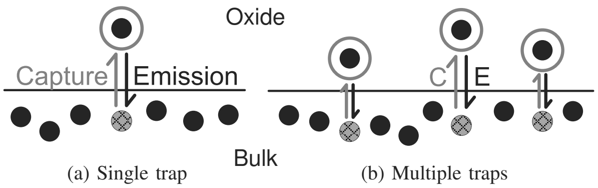
Over the last decade, the studies mainly focused on understanding the physics of RTN. Campbell *et al.* [4] suggested that RTN is originated from the capture and emission of the channel carriers by interface traps. They also conducted a systematic study of the channel length, width, and gate overdrive dependencies for RTN effects [2], and proposed a new method to characterize the oxide traps considering the energy band structure of HK/MG MOSFETs [5].

The RTN effects in both SRAM and FLASH memory technologies have been investigated recently. For example, the RTN in deca-nanometer flash memories was investigated by Ghetti *et al.* [6], and statistical distribution of V_{th} was analyzed. Tega *et al.* [7] estimated the impact of RTN on the scaled-down SRAM, and both read/write margins with or without RTN are simulated. Toh *et al.* [8] has shown that V_{min} degradation due to RTN is 50mV in 45nm SRAM. An accurate computational method for trap-level, non-stationary analysis of RTN in SRAMs is presented by Aadithya *et al.* [9], who also proposed a technique for predicting the impact of RTN on SRAMs/DRAMs in the presence of variability [10]. However, the continuous-time simulation approach used by Aadithya *et al.* in [10] is too complex, and not suitable for circuit performance evaluation.

It is believed that RTN can be also a serious issue in digital circuits. Leyris *et al.* [11] proposed a Shockley-Read-Hall based model to explain the RTN behavior. A methodology to include RTN in circuit analysis is proposed in [12], and the transient analysis is applied on the four-quadrant Chible multiplier circuit. Ye *et al.* [13] proposed a two-stage L-shaped circuit to generate RTN signal which is fully compatible with SPICE. The time-domain delay model was used to simulate and measure the fluctuation in [14], but the approach can only applied to simple circuits such as SRAM cell and ring oscillator

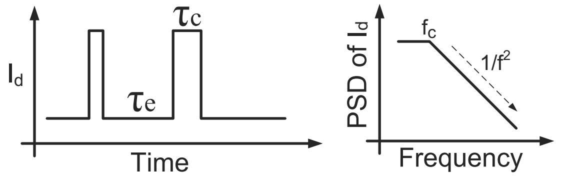
This work was supported by National Science and Technology Major Project (2011ZX01035-001-001-002, 2010ZX01030-001-001-04), National Natural Science Foundation of China (No. 61028006, 60870001, 61076035), and Tsinghua University Initiative Scientific Research Program. The work of Y. Cao was supported in part by SRC.

978-0-7695-4767-1/12 $26.00 © 2012 IEEE
DOI 10.1109/ISVLSI.2012.35



(a) Single trap (b) Multiple traps

Fig. 1. Capture/emission process of RTN



(a) Time domain (b) Frequency domain

Fig. 2. Drain current I_d due to RTN

because of the extraordinary computation complexity. In this paper, the delay characterization of circuit will be investigated, and a fast algorithm will be performed on the circuit-level analysis for RTN.

III. Modeling Random Telegraph Noise

A. Physics of RTN

As shown in Fig. 1(a), the carrier (the black solid circle) is occasionally captured by the trap (the red hollow circle) in the oxide, and it will be emitted back into the channel after a period of time. Multiple capture/emission events can occur at the same time, as shown in Fig. 1(b).The traps in the oxide have two states: the "filled" state which indicates the carrier is captured by the trap, and the "empty" state indicating the carrier is emitted back into the channel. For a given trap, the transition between these two states is inherently random, and the activity of a single trap can be modeled as a two-state time-inhomogeneous Markov chain [9].

At time domain, due to the RTN effect, the drain current I_d shows a fluctuational waveform as shown in Fig. 2(a). The high level of I_d corresponds to the low level of RTN, at which the trap is empty and the carrier is emitted back into the channel. The low level of I_d corresponds to the high level of RTN, at which the trap is filled and the carrier is captured by the trap. When the trap is empty, the carrier is "waiting to be captured", and the time spent in this state is the capture time τ_c. At the other side, when the trap is filled, the carrier is "waiting to be emitted", and the time spent in this state is the emission time τ_e [4]. Both the capture time τ_c and emission time τ_e are time-varying, and depend on the position of the traps, the trap energy level, and the gate overdrive voltage $V_{gs} - V_{th}$ [4], [9]. The typical values of τ_c and τ_e are about $1\text{ms} \sim 1000\text{ms}$ [4].

At frequency domain, the power spectral density (PSD) of the drain current I_d shows a Lorentzian shaped spectrum with the slope of $1/f^2$, as shown in Fig. 2(b) [5]. The cut-off frequency is

$$f_{cut} = \frac{1}{2\pi\tau_{cut}} \tag{1}$$

The time constant τ_{cut} in the above equation is defined as [13]

$$\frac{1}{\tau_{cut}} = \frac{1}{\tau_c} + \frac{1}{\tau_e} \tag{2}$$

B. RTN-induced V_{th} fluctuation in digital circuits

In order to model the impact of RTN on digital circuits, the equivalent circuit technique is used [15], as shown in Fig. 3. The high current state in Fig. 2(a) corresponds to the left device in Fig. 3, and there is no shift in threshold voltage. The right device shows the low current state induced by RTN, which is modeled as the voltage source, and the voltage is given by

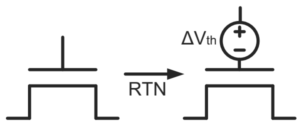


Fig. 3. The equivalent circuit of RTN effect

$$\Delta V_{th} = \frac{N \cdot q}{C_{ox} W L} \tag{3}$$

where N is the number of oxide traps, q is the elementary charge, C_{ox} is the unit area capacitance, while W and L are channel width and channel length respectively [13].

Because the magnitude of single trap RTN sharply goes up as device shrinks [13], this paper targets at the single-trap RTN fluctuation as shown in Fig. 1(a). Eq. (3) indicates that RTN depends on the area of the device, but experiments show that the gate overdrive $V_{gs} - V_{th}$ can also affect the RTN amplitude [2]–[4]. Therefore, the maximum shift in threshold voltage can be extracted from the experimental data, and is given by

$$\Delta V_{th} = \frac{\beta - \lambda V_{ov}}{W \cdot L} \tag{4}$$

where $V_{ov} = V_{gs} - V_{th}$ is the gate overdrive voltage, β and λ can be fitted by experimental data. Using PTM device library, the above model shows that **$\boldsymbol{\Delta V_{th}}$ of a 16nm device can be as much as 130mV.**

IV. RTN Evaluation in Digital Circuits

As described in section II, the capture time τ_c and emission time τ_e are both at the milli-second order [4], while the operation time of a digital circuit is at the nano-second order. The operation of the digital circuit is much faster than the transition between high and low current states, thus during the operation time of the digital circuit $[t, t+\Delta t)$, all the traps are considered to keep their filled/empty states. Therefore, the "sampling" technique can be applied to evaluate the impact of RTN on the digital circuits as shown in Fig. 4: the trap states at time t are sampled to evaluate the RTN-induced temporal performance of the digital circuit at t.

The trap state of the MOSFET at time t can be described by a random variable $\boldsymbol{S}$, which has two values: 0 corresponding to empty state and 1 corresponding to filled state. Thus, the threshold voltage of this MOSFET is expressed as

$$V_{th} = V_{th0} + \boldsymbol{S} \cdot V_R \tag{5}$$

where V_R is the maximum shift in threshold voltage described by Eq. (3) and (4).

The probability distribution of the random variable $\boldsymbol{S}$ is determined by the capture time and emission time, which is given by

$$\begin{cases} P(\boldsymbol{S}=0) = \dfrac{\tau_c}{\tau_c + \tau_e} = \dfrac{1}{1+r} \\ P(\boldsymbol{S}=1) = \dfrac{\tau_e}{\tau_c + \tau_e} = \dfrac{r}{1+r} \end{cases} \tag{6}$$

where the constant $r = \tau_e/\tau_c$ is the ratio of the emission time to the capture time, which is a constant only depending on trap energy level and Fermi level [13].

When the circuit are "sampled" at time t, the threshold voltage of a given MOSFET is

$$V_{th}(t) = V_{th0} + S(t) \cdot V_R \tag{7}$$

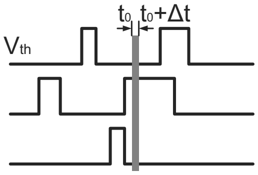


Fig. 4. Sampling the high and low states of devices induced by RTN

where $S(t)$ is a fixed value: 0 or 1.

Because the traps in the devices are independent, all $\boldsymbol{S_i}$ are independent. Therefore, by the "sampling" technique, Monte-Carlo simulations can be used to evaluate the circuit performance under RTN. In Monte-Carlo simulations, one simulation can be considered as a "sample" of the given circuit, and the value of $\boldsymbol{S}$ can be generated by randomly choosing 0 or 1. Then, traditional STA tools such as "PathMill" and "PrimeTime" can be used for subsequent simulation. However, the Monte-Carlo simulations are time-consuming. Thus, new technique will be proposed in the following section.

A. Statistical critical path analysis technique

In this section, the statistical critical path analysis technique is proposed to evaluate the impact of RTN effect on the temporal performance of digital circuits.

The circuit delay is determined by a set of critical paths in the circuit, which is described by

$$\tau_c = max\{\tau_{cp,i}\} \tag{8}$$

where τ_c is the circuit delay, and $\tau_{cp,i}$ is the delay of the i-th critical path.

The delay of a given critical path is

$$\tau_{cp} = \sum_j \tau_j \tag{9}$$

where τ_j is the delay of k-th gate in this critical path.

The propagation delay of a given logic gate is given by

$$\tau = \frac{K \cdot V_{dd}}{A \cdot (V_{dd} - V_{th})^{\alpha}} \tag{10}$$

where K is a coefficient related with load capacitance, mobility and other device parameters, A is the equivalent area of the logic gate, and α is the velocity saturation index.

Combined with Eq. (5), the shift in gate delay considering RTN effect can be approximately described as

$$\frac{\boldsymbol{\Delta\tau}}{\tau} = \frac{\alpha \cdot \boldsymbol{S} \cdot V_R}{V_{dd} - V_{th0}} \tag{11}$$

where $\boldsymbol{\Delta\tau}$ is also a random variable, and has the similar probability distribution as $\boldsymbol{S}$, which is given by

$$\begin{cases} P(\boldsymbol{\Delta\tau} = 0) = \dfrac{1}{1+r} \\ P(\boldsymbol{\Delta\tau} = \dfrac{\alpha\tau \cdot V_R}{V_{dd} - V_{th0}}) = \dfrac{r}{1+r} \end{cases} \tag{12}$$

Therefore, the delay shift in a given critical path is also a random variable

$$\boldsymbol{\Delta\tau_{cp}} = \sum_j \boldsymbol{\Delta\tau_j} \tag{13}$$

where $\boldsymbol{\Delta\tau_{cp}}$ varies from 0 to $\sum_j \left(\dfrac{\alpha\tau_j \cdot V_{R,j}}{V_{dd} - V_{th0,j}} \right)$.

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, the distribution of $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{Y}$ can be calculated by the convolution of the distribute function of $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$p_Z(z) = \sum_k (p_X(k) \cdot p_Y(z-k)) \tag{14}$$

Thus, the distribution of $\boldsymbol{\Delta\tau_{cp}}$ can be calculated by using Eq. (14) recursively, which means $\boldsymbol{\Delta\tau_1} + \boldsymbol{\Delta\tau_2}$ is calculated firstly, then $\boldsymbol{\Delta\tau_3}$ is added, and finally all $\boldsymbol{\Delta\tau_j}$ are sum up.

The distribution of $\boldsymbol{\tau_{cp}}$ is also determined, since $\boldsymbol{\tau_{cp}} = \tau_{cp} + \boldsymbol{\Delta\tau_{cp}}$, and τ_{cp} is a fixed value.

We should notice that the circuit delay $\boldsymbol{\tau_c} \leq x$ if and only if all $\boldsymbol{\tau_{cp,i}} \leq x$. Hence, the distribution of $\boldsymbol{\tau_c}$ can be described the cumulative distribution function (CDF)

$$F(x) = \prod_i F_i(x) \tag{15}$$

where $F(x)$ is the CDF of $\boldsymbol{\tau_c}$, and $F_i(x)$ is the CDF of $\boldsymbol{\tau_{cp,i}}$.

B. Algorithm for calculating critical path delay distribution

From Eq. (12) to (13), we can conclude that the total number of multiplication is $2^{N+1}-4$, and N is the number of gates in the critical path. Thus, the computation complexity is $O(2^N)$, which runs very slow for the long path.

In order to reduce the complexity, we use the "grouping" technique to construct the approximate distribution of the partial sum $\boldsymbol{\phi_L} = \sum_{j=1}^{L<N} \boldsymbol{\Delta\tau_j}$. This "grouping" technique is described as following.

First, we construct a new random variable $\boldsymbol{\Phi}$, whose distribution is defined by

$$P(m\delta < \boldsymbol{\Phi} \leq (m+1)\delta) = \sum_{m\delta < x \leq (m+1)\delta} p_L(x) \tag{16}$$

where $m = 0 \ldots M-1$, $\delta = \left(\sum_{j=1}^{L} \dfrac{\alpha\tau_j \cdot V_{R,j}}{V_{dd} - V_{th0,j}} \right) / M$, and $p_L(x)$ is the probability density function (PDF) of $\boldsymbol{\phi_L}$. Here, M is a user-defined parameter, and larger value of M leads to better approximation.

Second, the probability distribution of $\boldsymbol{\Phi}$ is denoted by the probability of M discrete values, which is given by

$$p_{\Phi}((m+0.5)\delta) = P(m\delta < \boldsymbol{\Phi} \leq (m+1)\delta) \tag{17}$$

where 2^L value discrete distribution is "grouped" to M value discrete distribution. Our experiments show that $M = 20$ is often enough.

Obviously, by using the "grouping" technique, the number of multiplication is less than $2MN$. Since M is a constant which can be specified before the algorithm, the computation complexity is reduced to $O(N)$. This algorithm is described in Fig. 5.

C. RTN evaluation framework

Based on the statistical critical path analysis technique, we propose the framework of evaluating the impact of RTN effect on digital circuits, as shown in Fig. 6.

The "STA Tool" which can be invoked by our framework generates the critical paths through the circuit netlist and gate library, which is created by the Hspice simulation based on the PTM model. The "PathMill" tool is invoked in this framework currently. Besides, the "RTN Vth Calculator" calculates the V_R value of all logic gates in the library, and these values are stored for future analysis.

The "Delay Calculator" calculates the delay distribution of each critical path based on the algorithm described in Fig. 5, and the delay distribution of the circuit based on Eq. (15).

```
Input: Distribution of $\boldsymbol{\Delta\tau_j}$ , $j = 1 \dots N$
    Maximum number of discrete values $M$
Output: Distribution of $\boldsymbol{\Delta\tau_{cp}}$
$L \leftarrow 1$
$\boldsymbol{\phi_L} = \boldsymbol{\Delta\tau_1}$
for $L = 2 \rightarrow N$ do
    calculate $\boldsymbol{\phi_L}$ using Eq. (14)
    if $M < 2^L$ then
        construct $\boldsymbol{\Phi}$ by grouping technique using Eq. (16) and (17)
    end if
end for
return distribution of $\boldsymbol{\Phi}$
```

Fig. 5. Algorithm for calculating critical path delay distribution

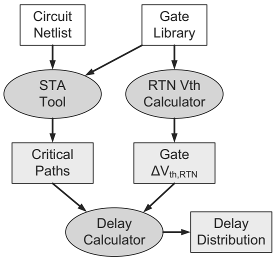


Fig. 6. RTN evaluation framework

D. Compared with Monte-Carlo simulation

This section will compare the result from the proposed statistical critical path technique with the Monte-Carlo simulation. The benchmark circuit is "kogge16", 16nm PTM model is used, and r is set to 1. The Monte-Carlo simulation result is shown in Fig. 7(a). The left vertical line represents for the circuit delay without RTN effect τ_0, which is about 1.39ns and can be considered as the design specification of this circuit. The histogram in Fig. 7(a) represents for the Monte-Carlo simulation results with RTN. The minimum delay τ_{min} is 1.55ns, and the maximum delay τ_{max} is 2.0ns. The average delay τ_{avg} is 1.8ns as shown by the right vertical line in Fig. 7(a), which is 29.3% larger than τ_0.

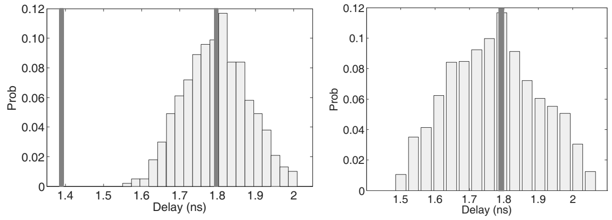
The result from the statistical critical path technique is shown in Fig. 7(b), which shows a similar histogram compared with Fig. 7(a). The minimum delay τ_{min} is 1.5ns, the maximum delay τ_{max} is 2.1ns, and $\tau_{avg} = 1.8\text{ns}$, which is the same as the Monte-Carlo simulation.

The above results demonstrate that RTN will be a very serious obstacle in digital circuit design under deca-nanometer regime, which exhibits in the following two aspects:

1) The RTN can cause significant circuit performance degradation, resulting a serious timing violation. The possible minimum delay as shown in Fig. (7) is still lager than τ_0. Thus, RTN must be considered in the circuit simulation and corresponding design flow.
2) Even the delay degradation is considered, there still exists another problem: delay variation. As shown in Fig. 7(b), the variation is 33.3%. Hence, statistical analysis should be considered in RTN evaluation.

E. Circuit delay fluctuation analysis

The ISCAS85 and ALU circuits are evaluated using our proposed framework, while the technology node is 16nm, and $r = 1$. The simulation results are shown in Table I. The second column shows the number of gates, and the third column shows the original delay of the circuits τ_0. The fourth column shows the average delay degradation $\Delta\tau_{avg} = (\tau_{avg} - \tau_0)/\tau_0$, and the fifth column shows the delay variation $\Delta\tau_{var} = (\tau_{max} - \tau_{min})/\tau_{avg}$. The last column shows the run time of each benchmark.



(a) Monte-Carlo (b) Statistical critical path

Fig. 7. Kogge16 delay fluctuation due to RTN

TABLE I
DELAY DEGRADATION AND VARIATION DUE TO RTN AT 16nm

Circuit	# gates	τ_0 (ns)	$\Delta\tau_{avg}$ (%)	$\Delta\tau_{var}$ (%)	Time (s)
c432	169	2.273	31.3	19.0	11
c499	204	1.506	49.6	15.2	9
c880	383	1.022	37.6	17.7	3
c1355	548	1.366	21.9	21.6	15
c1908	911	1.478	47.3	32.4	22
c2670	1279	1.854	38.1	17.5	10
c3540	1699	2.123	32.9	34.7	48
c5315	2329	1.723	32.6	35.5	21
c6288	2447	5.399	20.6	30.7	402
c7552	3566	1.603	37.5	43.1	41
array4	69	1.826	40.7	59.6	1
array8	375	3.674	54.4	59.9	47
booth9	412	1.953	34.1	29.3	10
bkung16	81	1.389	28.2	21.2	1
bkung32	165	2.269	27.5	17.4	3
kogge16	81	1.389	28.7	22.7	1
kogge32	164	2.307	27.8	14.8	2
log32	371	1.472	52.0	47.2	1
log64	862	2.052	35.2	59.4	3
Pmult8	356	2.340	35.8	29.0	12
Pmult16	1672	3.880	21.3	36.8	213
Pmult32	6814	7.113	26.1	40.9	2400
Avg.			34.6	32.1	148.9

According to the results in Table I, the average delay degradation is 34.6%, and the average delay variation is 32.1%. Meanwhile, the maximum delay degradation can reach up to 54.4%, and the maximum delay variation is 59.9%.

The average computation time of our RTN evaluation framework is 148.9s. Actually, this run time is dominated by the time of searching critical paths, and only single search is executed in our framework. On the other hand, multiple critical-path searching are executed in Monte-Carlo simulation. Therefore, our proposed framework can be used for RTN evaluation in large-scale circuits.

F. Technology scaling analysis

The scaling down of the technology node shows two trends: the area scaling down and power supply voltage scaling down. From Eq. (4), the RTN becomes worse with both the trends. This section

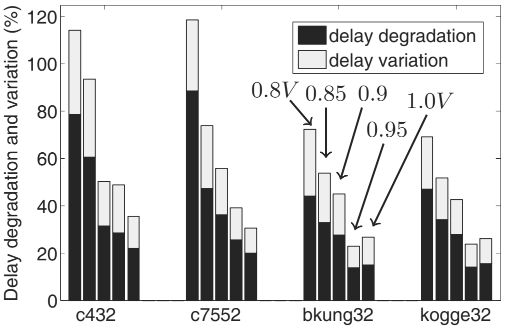


Fig. 9. Delay fluctuation with different V_{dd} (0.8 ∼ 1.0V)

investigates the technology dependence of RTN impact on the circuit temporal performance.

Fig. 8 shows the simulation results of temporal performance degradation and variation against technology scaling. The bar for delay degradation (at the downside) and the bar for delay variation (at the upside) are stacked as one bar. Each circuit corresponds to four bars, and the leftmost bar to the rightmost one represents for the results under 16nm, 22nm, 32nm and 45nm respectively. The results in Fig. 8 show that with technology scaling down, both the temporal performance degradation and variation increase.

G. Power supply scaling analysis

Eq. (4) shows that the RTN impact on the circuit delay can be affected by the power supply voltage, and the scaling down of power supply voltage increases the RTN effect.

The performance degradation and variation against the power supply voltage under 16nm are shown in Fig. 9. Each circuit corresponds to five bars, and the leftmost bar to the rightmost one represents for the results under 0.8V, 0.85V, 0.9V, 0.95V and 1.0V respectively. The results show that with power supply voltage scaling down, both the temporal performance degradation and variation increase.

V. RTN Mitigation in Digital Circuits

In this section, we apply power supply tuning and gate sizing techniques on the digital circuits, and simply demonstrate the efficiency of such techniques on mitigating the RTN-induced delay degradation and variation.

A. Power supply tuning

Previous section shows that increasing power supply V_{dd} will reduce threshold voltage degradation. Thus, we could reduce the circuit delay by tuning V_{dd}.

As shown in Fig. 10, we take "c432" circuit under 16nm as an example. The rectangles represent for the possible circuit delay $[\tau_{min}, \tau_{max}]$ with RTN effect, while the triangle points represent for the original circuit delay τ_0 without RTN. We choose τ_0 at 0.9V as the design specification (the horizontal line in Fig. 10). With no V_{dd} tuning, the average delay degradation is 31.3%, and the variation is 19.0% as shown in Table I. If the power supply voltage increases to 0.95V, we can find that the rectangle is almost below the specification line, which means V_{dd} tuning can make the circuit delay meet the design specification, and the variation is reduced by 26.3%. On the other hand, the power consumption will increase by 11.4%.

B. Gate sizing and replacement

From Eq. (4), we know that RTN strongly depends on the area of the MOSFET device. Thus, this section investigates the effect of gate sizing and replacement technique on mitigating RTN effect.

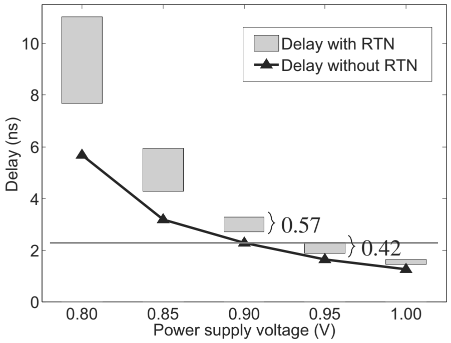


Fig. 10. Delay degradation and variation using power supply tuning

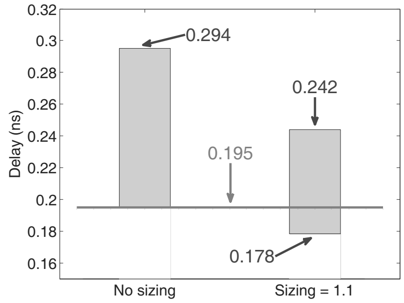


Fig. 11. Gate sizing for NAND2

Assuming that the area of a given gate in Eq. (10) becomes γA, and according to Eq. (4), the RTN-induced delay of this gate after sizing can be expressed as

$$\boldsymbol{\tau_s} = \frac{K \cdot V_{dd}}{\gamma A \cdot (V_{dd} - V_{th0} - \boldsymbol{S}V_R/\gamma)^{\alpha}} \tag{18}$$

where the sizing coefficient $\gamma > 1$.

Thus the delay will degrade by

$$\boldsymbol{\Delta\tau_s} = \left[\frac{1}{\gamma} - 1 + \frac{\alpha \cdot \boldsymbol{S}V_R}{\gamma^2(V_{dd} - V_{th0})}\right]\tau_i \tag{19}$$

Compared to Eq. (11), we can get that sizing can mitigate the delay degradation due to RTN. Meanwhile, the term $1/\gamma^2$ indicates that the delay variation can be also reduced, which is similar to the variation induced by random dopant fluctuation (RDF) [16].

The gate sizing technology on "NAND2" gate is shown in Fig. 11. The original delay with no RTN effect is 0.195ns. The left bar is the possible delay with no sizing, which varies from 0.195ns to 0.294ns. The right bar represents for the delay with the sizing coefficient $\gamma = 1.1$, which varies from 0.178ns to 0.242ns. The delay improvement is 26.7%, and the delay variation can be reduced by 35.4%.

The above results show that a larger gate has smaller RTN-induced delay degradation and variation, thus in the standard-cell design flow, the original logic gates can be replaced by the corresponding larger gates in the library. Two replacement strategies may be applied: "full" replacement (replace all the gates with the larger ones) or "critical" replacement (only replace the gates along the critical path).

The "kogge16" circuit under 16nm is used for assessment of the impact of gate sizing and replacement technique. The amount of 38% gates are replaced with critical replacement strategy. Thus, the area overhead of full replacement is $\gamma - 1$, while the area overhead of critical replacement is $0.38(\gamma - 1)$. The simulation results are shown in Fig. 12. Both τ_{min} and τ_{max} decrease with the sizing coefficient γ increases. The full and critical replacement have similar effect on the minimum delay τ_{min}, while the maximum delay τ_{max} of full

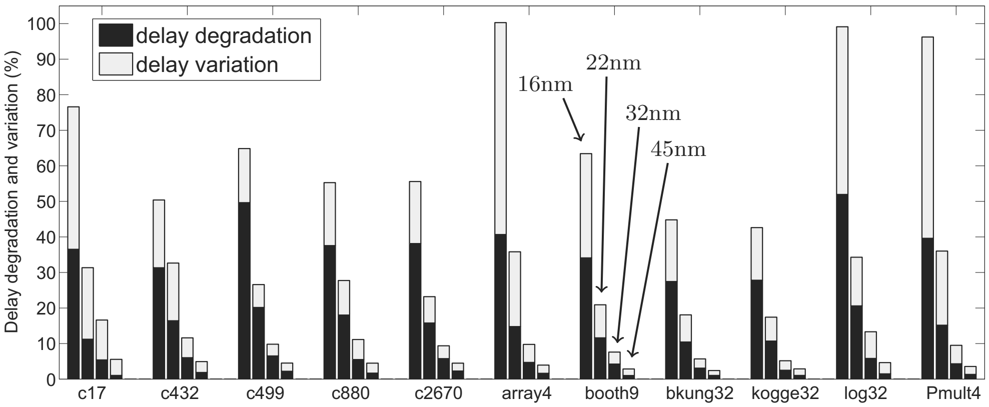


Fig. 8. Delay degradation and variation with technology scaling (16nm $\sim$ 45nm)

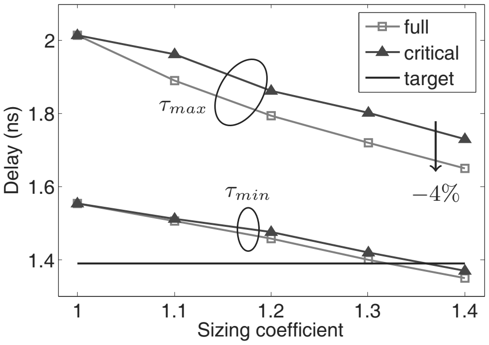


Fig. 12. Gate sizing and replacement for kogge16 circuit

replacement is 4% less than critical replacement. In practice, the critical replacement should be used because the area overhead is 62% smaller with the same γ. The results show that τ_{max} can be reduced by 14% with 15% area overhead ($\gamma = 1.4$) using critical replacement, while the delay variation can be reduced by 22%.

VI. Discussion and Conclusions

In this paper, we proposes a framework for evaluating both the digital circuits' temporal performance degradation and variation induced by RTN for the first time. The evaluation results show that the average degradation and variation under 16nm can be 34.6% and 32.1% respectively. Two design techniques, power supply tuning and gate sizing, are applied to mitigate the RTN effect in digital circuits, and simulation results show that these techniques have limited effects, where the degradation and variation cannot be eliminated completely.

The RTN-induced fluctuations are independent in all the devices, which causes very random performance distribution in digital circuits. Design techniques, such as power supply and gate sizing investigated in our paper, are not effective enough to mitigate RTN effect. Enough performance margin should be reserved in design flow to compensate the impact of RTN. Therefore, more efficient circuit-level and architectural-level techniques should be investigated in our future work.

References

[1] N. Tega, H. Miki, F. Pagette, D. Frank, A. Ray, M. Rooks, W. Haensch, and K. Torii, "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm," in *Symposium on VLSI Technology*, 2009, pp. 50 –51.

[2] J. Campbell, L. Yu, K. Cheung, J. Qin, J. Suehle, A. Oates, and K. Sheng, "Large random telegraph noise in sub-threshold operation of nano-scale nMOSFETs," in *ICICDT*, May 2009, pp. 17 –20.

[3] A. Lee, A. R. Brown, A. Asenov, and S. Roy, "Random telegraph signal noise simulation of decanano MOSFETs subject to atomic scale structure variation," *Superlattices and Microstructures*, vol. 34, no. 3-6, pp. 293 – 300, 2003.

[4] J. Campbell, J. Qin, K. Cheungl, L. Yu, J. Suehlel, A. Oates, and K. Sheng, "The origins of random telegraph noise in highly scaled SiON nMOSFETs," in *IEEE International Integrated Reliability Workshop Final Report (IRW)*, Oct. 2008, pp. 105–109.

[5] J. Campbell, J. Qin, K. Cheung, L. Yu, J. Suehle, A. Oates, and K. Sheng, "Random telegraph noise in highly scaled nMOSFETs," in *IRPS*, 2009, pp. 382 –388.

[6] A. Ghetti, C. Compagnoni, A. Spinelli, and A. Visconti, "Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer flash memories," *IEEE Transactions on Electron Devices*, vol. 56, no. 8, pp. 1746 –1752, 2009.

[7] N. Tega, H. Miki, M. Yamaoka, H. Kume, T. Mine, T. Ishida, Y. Mori, R. Yamada, and K. Torii, "Impact of threshold voltage fluctuation due to random telegraph noise on scaled-down SRAM," in *IEEE International Reliability Physics Symposium (IRPS)*, 2008, pp. 541–546.

[8] S. O. Toh, Y. Tsukamoto, Z. Guo, L. Jones, T.-J. K. Liu, and B. Nikolic, "Impact of random telegraph signals on Vmin in 45nm SRAM," in *IEDM*, 2009, pp. 1 –4.

[9] K. Aadithya, A. Demir, S. Venugopalan, and J. Roychowdhury, "SAMURAI: An accurate method for modelling and simulating non-stationary random telegraph noise in SRAMs," in *DATE*, Mar. 2011, pp. 1 –6.

[10] K. Aadithya, S. Venogopalan, A. Demir, and J. Roychowdhury, "MUSTARD: A coupled, stochastic/deterministic, discrete/continuous technique for predicting the impact of random telegraph noise on SRAMs and DRAMs," in *DAC*, Jun. 2011, pp. 292 –297.

[11] C. Leyris, S. Pilorget, M. Marin, M. Minondo, and H. Jaouen, "Random telegraph signal noise spice modeling for circuit simulators," in *European Solid State Device Research Conference*, 2007, pp. 187 –190.

[12] T. B. Tang and A. Murray, "Integrating RTS noise into circuit analysis," in *ISCAS*, May 2009, pp. 585 –588.

[13] Y. Ye, C.-C. Wang, and Y. Cao, "Simulation of random telegraph noise with 2-stage equivalent circuit," in *ICCAD*, 2010, pp. 709 –713.

[14] K. Ito, T. Matsumoto, S. Nishizawa, H. Sunagawa, K. Kobayashi, and H. Onodera, "Modeling of random telegraph noise under circuit operation simulation and measurement of RTN-induced delay fluctuation," in *ISQED*, march 2011, pp. 1 –6.

[15] M. Tanizawa, S. Ohbayashi, T. Okagaki, K. Sonoda, K. Eikyu, Y. Hirano, K. Ishikawa, O. Tsuchiya, and Y. Inoue, "Application of a statistical compact model for random telegraph noise to scaled-SRAM Vmin analysis," in *Symposium on VLSI Technology*, 2010, pp. 95–96.

[16] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1μm MOSFET's: A 3-D "atomistic" simulation study," *IEEE Transactions on Electron Devices*, vol. 45, no. 12, pp. 2505 –2513, Dec. 1998.