Statistical Analysis of Random Telegraph Noise in Digital Circuits

Xiaoming Chen†, Yu Wang†, Yu Cao§, Huazhong Yang†
†Department of Electronic Engineering, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China
§Department of ECEE, Arizona State University, Tempe, AZ, USA
Email: chenxm05@mails.tsinghua.edu.cn, yu-wang@tsinghua.edu.cn, yu.cao@asu.edu, yanghz@tsinghua.edu.cn

***Abstract*—Random telegraph noise (RTN) has become an important reliability issue at the sub-65nm technology node. Existing RTN simulation approaches mainly focus on single trap induced RTN and transient response of RTN, which are usually time-consuming for circuit-level simulation. This paper proposes a statistical algorithm to study multiple traps induced RTN in digital circuits, to show the temporal distribution of circuit delay under RTN. Based on the simulation results we show how to protect circuit from RTN. Bias dependence of RTN is also discussed.**

***Keywords*—Random telegraph noise; Statistical analysis; Reliability**

I. Introduction

As CMOS technology scales, many reliability mechanisms that affect circuits' reliability are becoming more serious. These issues must be evaluated and well addressed during the design time. In recent years, random telegraph noise (RTN) has attracted researchers' attention. RTN can cause random fluctuations in electrical parameters (such as V_{th} and I_d) [1]. RTN-induced I_d variation can be up to 40% in 30×30nm devices [2], and the V_{th} variation can be larger than 70mV for the smallest devices at 22nm technology node [3]. The RTN effect increases superlinearly with the scaling down of the device's size [4]. RTN is also a serious concern in CMOS logic circuits [5].

The physical mechanism of RTN has been studied for many years [1], [6]–[8]. The impact of single trap induced RTN on memories was widely studied [4], [9]–[15], some circuit-level simulation approaches were also proposed [16]–[19]. Most of them evaluate single trap induced RTN. Multiple traps induced RTN has been rarely studied [11]. However, there should be 2∼3 detectable traps in each device, and the number of traps follows Poisson distribution [20]. The multi-trap problem is actually a statistical problem, which is more complex than the single-trap problem. This paper will evaluate circuit performance under multiple traps induced RTN.

On the other hand, many existing researches used time-consuming SPICE simulation to obtain the transient response under RTN [14]–[18]. This method can be used for predicting read/write failures in memories, but it is useless for evaluating the timing of CMOS logic circuits. To actually understand the impact of RTN on logic circuits, a temporal distribution of circuit delay should be used to obtain the statistical information. In other words, RTN should be integrated into timing analysis to see the real impact of RTN on logic circuits.

This work was supported by National Natural Science Foundation of China (No. 61373026, No.61261160501, No. 61028006), 973 project (2013CB329000), National Science and Technology Major Project (2011ZX01035-001-001-002), and the youth talent development plan of Beijing (YETP0099).

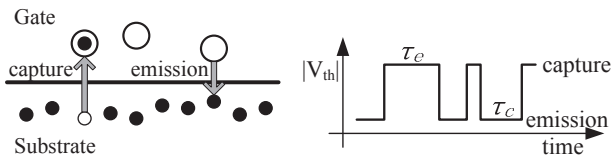
The contributions of this paper are summarised as follows.

- This paper integrates RTN into timing analysis to estimate the impact of RTN on logic circuits. A fast estimation framework which is based on a statistical algorithm is proposed to obtain the temporal distribution of circuit delay without SPICE or Monte-Carlo.
- This is the first time to analyze the impact of multiple traps induced RTN on logic circuits.
- We find that it is not practical to protect circuit from the maximum possible delay under RTN, but we should only ensure that circuit functions correctly with a certain probability during the whole lifetime, such that timing violation can hardly happen. By ensuring that circuit functions correctly with $1-10^{-9}$ probability, circuit should be protected from the degraded delay which is about 7% to 40% larger than the intrinsic delay.
- The impact of gate voltage on RTN effect is investigated. We show that the impact of V_{gs} on RTN has two trends: increasing and saturation. A simple V_{dd} tuning approach can effectively protect circuit from the effect of RTN.

The rest of the paper is organized as follows. Section II gives backgrounds and modeling of RTN. We introduce the proposed algorithm in Section III. The experimental results are shown in Section IV. Finally Section V concludes the paper.

II. Modeling Random Telegraph Noise

A. Physics of RTN



(a) Capture/emission of traps. (b) V_{th} fluctuation.

Fig. 1: Physics of RTN and its impact on V_{th}.

As shown in Fig. 1a, RTN is caused by the capture/emission process of charge carriers by the oxide traps (defects) [1], [8]. A carrier in the channel is occasionally captured by a trap in the oxide, and the carrier will be emitted back after a period of time. The capture/emission process of a given trap can be described by a two-state Markov chain [16]. With reference to Fig. 1b, the high V_{th} state occurs when the carrier is captured by the trap (i.e. the trap is filled), and the low V_{th} state occurs when the carrier is emitted back (i.e. the trap is

978-1-4799-2816-3/14/$31.00 ©2014 IEEE

empty). The time spent in the high V_{th} state is emission time τ_e, which means "time to be emitted back", and the time spent in the low V_{th} state is capture time τ_c, which means "time to be captured". τ_e and τ_c strongly depend on gate overdrive ($\tau_c = 10^{-3}$s to 10^{-2}s and $\tau_e = 10^{-1}$s to 10^{+1}s [1]).

B. Dependence of Switched Bias Conditions

The capture and emission time constants also depend on the switched bias conditions. When a transistor is on, the time constants are $\tau_c^{(on)}$ and $\tau_e^{(on)}$; when the transistor is off, the time constants can be expressed as $\tau_c^{(off)} = \tau_c^{(on)} \times m_c$ and $\tau_e^{(off)} = \tau_e^{(on)}/m_e$ [21]. Consider a transistor with duty cycle SP (the probability of on is SP), its average time constants are

$$\begin{aligned} \tau_c &= SP \times \tau_c^{(on)} + (1 - SP) \times \tau_c^{(off)} \\ \tau_e &= SP \times \tau_e^{(on)} + (1 - SP) \times \tau_e^{(off)} \end{aligned} \tag{1}$$

A logic simulator is used to calculate the signal probability of all internal nodes in a circuit.

C. RTN-induced V_{th} Shift

Single trap induced V_{th} shift is given by [10]

$$\Delta V_{th} = \frac{q}{C_{ox} W L} \tag{2}$$

where q is the elementary charge, C_{ox} is the unit area capacitance, W and L are channel width and length respectively.

It is shown that the number of traps in each device follows Poisson distribution: $N_{tr} \sim Pois(\lambda)$ (N_{tr} is not the number of existing traps, which is constant, but that of detectable traps) [20], in which λ is the average number of traps. Based on the measured data plotted in [10], [11], [20], multiple traps induced maximum V_{th} shift can be approximately expressed as the sum of each individual trap induced V_{th} shift, so V_{th} shift caused by N_{fil} filled traps is

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

D. RTN-induced Gate Delay Modeling

The propagation delay of a logic gate i is given by

$$D(i) = \frac{K_i C_{L,i} V_{dd}}{(V_{gs} - V_{th,i})^{\alpha}} \tag{4}$$

where K_i is a coefficient related with device physical parameters, $C_{L,i}$ is the load capacitance, and α is the velocity saturation index. The delay shift caused by V_{th} shift is

$$\Delta D(i) \approx \frac{\alpha \Delta V_{th,i}}{V_{gs} - V_{th0}} \times D(i) \tag{5}$$

For a given trap, its state can be described by a two-valued random variable X: 0 corresponding to empty state and 1 corresponding to filled state. When the capture/emission process is stationary, the two states have a stationary distribution, which is given by

$$P(X = 0) = \frac{\tau_c}{\tau_e + \tau_c}, P(X = 1) = \frac{\tau_e}{\tau_e + \tau_c} \tag{6}$$

Note that our target is statistical analysis so only stationary state is considered, the detailed capture/emission sequence is not considered. Combining Eq. (3), (5) and (6), delay shift of gate i caused by multiple traps is described by a compound Poisson distribution:

$$\Delta D(i) = \sum_{i=1}^{N_{eff}} \Delta D_O(i), N_{eff} \sim Pois(r\lambda) \tag{7}$$

where $r = \frac{\tau_e}{\tau_e + \tau_c}$, and $\Delta D_O(i)$ is the delay shift caused by a single filled trap, which is calculated by Eq. (2) and (5).

III. Statistical Algorithm for RTN Simulation

This section proposes an RTN simulation framework based on a statistical algorithm. The proposed algorithm can fast obtain the temporal distribution of circuit delay without Monte-Carlo or SPICE. The algorithm is similar to the idea of the well-known statistical static timing analysis (SSTA) algorithm [22], but they are different. SSTA is used to evaluate the impact of process variations on the yield. For each individual chip they are assumed to be constant, and SSTA predicts that certain percent of manufactured circuits functions correctly. Unlike that RTN affects gate delays differently in different time moments. Therefore, delay value varies in time randomly. Because of that if chip operates long enough path delay can always get its worst value. So we are more interested in the maximum circuit delay. In addition, SSTA assumes that any delay can be described by a canonical form, but according to Eq. (7), MAX of two compound Poisson distributions is not a compound Poisson distribution, so canonical forms cannot be used in RTN analysis.

A. Terms and Definitions

Several terms are defined as follows. The *arrival/leaving time* of gate i ($AT(i)$/$LT(i)$) are the maximum propagation delay from circuit primary inputs (PI) to the inputs/output of gate i. $LT(i) = AT(i) + D(i)$. The propagation delay from gate i's output to gate j's output is $PD(i, j)$. An example of AT, LT and PD are illustrated in Fig. 2. An *input path (IP)* of gate i is a path from PIs to gate i, and $IP(i)$ is the set of all input paths of gate i. The *critical input paths (CIP)* are the paths that have longest delay in $IP(i)$, denoted as $CIP(i)$. In the following contents, we use f_X and F_X to represent the probability density function (PDF) and cumulative distribution function (CDF) of random variable X. Φ and ϕ are the CDF and PDF of the standard normal distribution. Φ_2 and ϕ_2 are the CDF and PDF of the standard bivariate normal distribution. μ_X and σ_X are mean and standard deviation of random variable X.

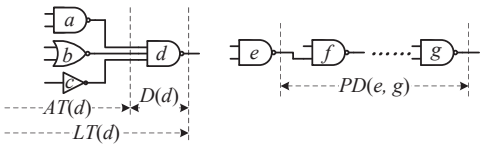


Fig. 2: Example to illustrate $AT(d)$, $LT(d)$ and $PD(e, g)$.

B. Simulation Framework

The RTN simulation framework is shown in Fig. 3. HSPICE is used to create a gate library which includes gate intrinsic delay and oxide capacitance of each gate type (i.e. INVX1, INVX4, NAND2X1, etc), based on the predictive technology model (PTM) [23]. STA tools are used to obtain the critical path information, which is used for correlation calculation.

A logic simulator is used to calculate the probability of all internal nodes. In STA, if $RT(i) - LT(i) <= \varepsilon$ (RT means "required time"), gate i is critical. $\varepsilon > 0$ to ensure that the criticality obtained in STA is also accurate in statistical analysis. Finally, the delay distribution of the circuit is calculated by the proposed statistical algorithm.

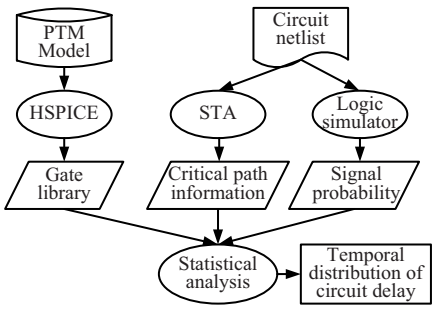


Fig. 3: RTN simulation framework.

C. Correlation of Gates

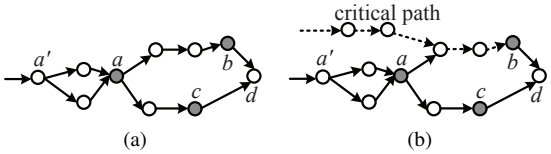


Fig. 4: Example to illustrate the correlation of gates.

In a traditional SSTA, the correlation is spatial correlation, i.e. global/local sources of variation. However, based on the delay model in Section II, there is no spatial correlation in RTN simulation. There is still correlation among gates, but the correlation is caused by path delay. The correlation comes from the case that the IPs of two gates have common gates, as illustrated in Fig. 4a, where gate a and a' are the common gates of $IP(b)$ and $IP(c)$. We are interested in the correlation of gate b and c, and we have

$$\rho_{LT(b),LT(c)} = \frac{E(LT(b)LT(c)) - \mu_{LT(b)}\mu_{LT(c)}}{\sigma_{LT(b)}\sigma_{LT(c)}} \quad (8)$$

Since $LT(b) = LT(a) + PD(a,b)$ and $LT(c) = LT(a) + PD(a,c)$, Eq. (8) can be written as

$$\begin{aligned} &\rho_{LT(b),LT(c)} \\ &= \{E[(LT(a) + PD(a,b)) \cdot (LT(a) + PD(a,c))] \\ &-(\mu_{LT(a)} + \mu_{PD(a,b)}) \cdot (\mu_{LT(a)} + \mu_{PD(a,c)})\} \\ &/(\sigma_{LT(b)}\sigma_{LT(c)}) \\ &= \frac{E(LT(a)^2)-(\mu_{LT(a)})^2}{\sigma_{LT(b)}\sigma_{LT(c)}} = \frac{(\sigma_{LT(a)})^2}{\sigma_{LT(b)}\sigma_{LT(c)}} \end{aligned} \quad (9)$$

This indicates that the correlation of gate b and c is only determined by the variance of the leaving time of gate b, c and a. Actually gate a' is also a common gate of $IP(b)$ and $IP(c)$; however, $\rho_{LT(b),LT(c)}$ does not depend on gate a'. So the "last" common gate should be used to calculate the correlation coefficient of two gates. Obviously the "last" common gate must have largest delay among all the candidate common gates. Another essential condition is that gate a should be in both $CIP(b)$ and $CIP(c)$; otherwise as shown in Fig. 4b, the dotted line is the CIP of gate b, so $LT(a)$ has little impact on $LT(b)$, and $LT(b)$ is not correlated to $LT(c)$. The algorithm for calculating the correlation of two gates are shown in Algorithm 1.

Algorithm 1 Calculating the correlation of gate i and j.

```
1: S = CIP(i) ∩ CIP(j);
2: if S == ∅ then
3:     ρ_{LT(i),LT(j)} = 0;
4: else
5:     Find gate u ∈ S such that μ_{LT(u)} = MAX_{v∈S}{μ_{LT(v)}};
6:     ρ_{LT(i),LT(j)} = σ²_{LT(u)} / (σ_{LT(i)} σ_{LT(j)});
7: end if
```

D. Proposed Statistical Algorithm

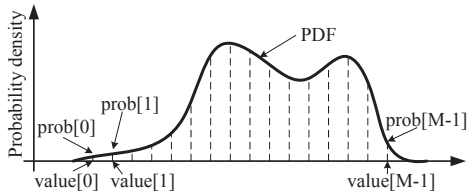


Fig. 5: Representing an arbitrary distribution by sampling the PDF.

The key idea of the proposed algorithm is that an arbitrary PDF that has no analytical form can be represented by sampling [24], as shown in Fig. 5. The number of sampled nodes is M, and larger M leads to better approximation. The sampled intervals can be uniform or non-uniform. Two vectors, $value$ and $prob$ with length M, are used to express the sampled value and the corresponding probability density, so the PDF becomes a discrete probability mass function (PMF). In this paper, $M = 100$ is used.

1) ADD operation: This is used to calculate the leaving time $LT(i) = AT(i) + D(i)$. Since $AT(i)$ and $D(i)$ are independent, their sum is the convolution of their PMFs. The convolution of two PMFs of sampled length M_1 and M_2 will have a length of $M_1 M_2$, which may exceeds M, so a grouping method is proposed to reconstruct the sum to be M-length. The pseudo code of ADD operation is shown in Algorithm 2.

Algorithm 2 Calculating $Z = X + Y$, X and Y are of sampled length M_X and M_Y.

```
1: Alloc two vectors, value and prob, both of length M_X M_Y;
2: //convolution
3: for i = 0 to M_Y − 1 do
4:     for j = 0 to M_X − 1 do
5:         value[i * M_X + j] = Y.value[i] + X.value[j];
6:         prob[i * M_X + j] = Y.prob[i] × X.prob[j];
7:     end for
8: end for
9: //grouping
10: min = MIN_{i<M_X M_Y} {value[i]}; max = MAX_{i<M_X M_Y} {value[i]};
11: step = (max − min)/M;
12: Clear Z;
13: for i = 0 to M_X M_Y − 1 do
14:     Z.value[(value[i] − min)/step] += prob[i] × value[i];
15:     Z.prob[(value[i] − min)/step] += prob[i];
16: end for
17: for i = 0 to M − 1 do
18:     Z.value[i] /= Z.prob[i];
19: end for
```

2) MAX operation: This is used to calculate the arrival time $AT(i) = MAX\{LT(j_1), LT(j_2), \cdots, LT(j_K)\}$, where gate $j_1, j_2, \cdots, j_K$ are the fan-ins of gate i. Consider a simple case $W = MAX(X, Y)$, X and Y are with correlation coefficient ρ_{XY}. The core operation is to calculate the integral

$$P(W \le w) = F_{XY}(w, w; \rho_{XY}) = \int_{-\infty}^{w}\int_{-\infty}^{w} f_{XY}(u, v; \rho_{XY})dudv \quad (10)$$

where f_{XY} is the joint PDF of X and Y and its analytical form is nonexistent. The integral is difficult to calculate. In this paper, we use an approximate method to fast obtain the result, based on the standard bivariate normal distribution.

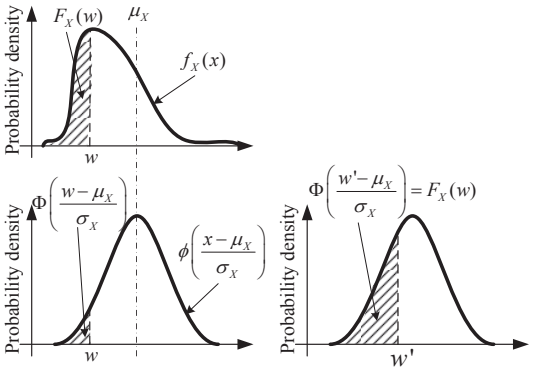


Fig. 6: Using normal distribution to calculate the approximate integral.

If the CDF of the corresponding bivariate normal distribution ($\Phi_2(\frac{w-\mu_X}{\sigma_X}, \frac{w-\mu_Y}{\sigma_Y}; \rho_{XY})$) directly replaces $F_{XY}(w, w; \rho_{XY})$, it will generate large error when $f_{XY}(u, v; \rho_{XY})$ is far away from the PDF of the bivariate normal distribution ($\phi_2(\frac{u-\mu_X}{\sigma_X}, \frac{v-\mu_Y}{\sigma_Y}; \rho_{XY})$). A univariate example is shown in Fig. 6, in which $\Phi\left(\frac{w-\mu_X}{\sigma_X}\right) << F_X(w)$. The reason of the error is that both F_X and Φ use the same integral end point w. If the integral end point of Φ is changed to $w' = \Phi^{-1}(F_X(w))\sigma_X + \mu_X$, we get $\Phi\left(\frac{w'-\mu_X}{\sigma_X}\right) = F_X(w)$. So the CDF of normal distribution can be used to replace the integral. Inspired by this result, in the bivariate case, the same method is used, which is given by

$$w_1 = \Phi^{-1}(F_X(w)), w_2 = \Phi^{-1}(F_Y(w))$$
$$F_{XY}(w, w; \rho_{XY}) \approx \Phi_2(w_1, w_2; \rho_{XY}) \quad (11)$$

Φ^{-1} is calculated by a lookup table. Though $\Phi_2(w_1, w_2; \rho)$ is still difficult to calculate, a fast and simple approximation is proposed in [25]. The pseudo code of MAX operation of X and Y is shown in Algorithm 3, in which X and Y are described by sampled PMFs.

We have validated this method using Monte-Carlo, and the accuracy is larger than 95% for some most common distributions. This accuracy is sufficient for RTN evaluation.

E. Complexity Analysis

The computational complexity of the proposed statistical algorithm is no more than $MAX\{O(M^2n), O(KMn)\}$, where n is the number of gates in a given circuit, and K is the maximum number of gate fan-ins. Since M and K are both constant, the complexity is $O(n)$, which is linear to the size of circuit.

Algorithm 3 Calculating $W = MAX(X, Y)$, X and Y are of sampled length M_X and M_Y.

```
min = MAX{MIN_{i<M_X}{X.value[i]}, MIN_{i<M_Y}{Y.value[i]}};
max = MAX{MAX_{i<M_X}{X.value[i]}, MAX_{i<M_Y}{Y.value[i]}};
step = (max − min)/M;
for i = 0 to M − 1 do
    w = min + step ∗ (i + 1);
    p1 = F_X(w); p2 = F_Y(w);
    W.prob[i] = Φ2(Φ^{-1}(p1), Φ^{-1}(p2); ρ_XY);
    W.value[i] = w;
end for
for i = M − 1 to 1 do
    W.prob[i] −= W.prob[i − 1];
end for
```

IV. Experimental Results

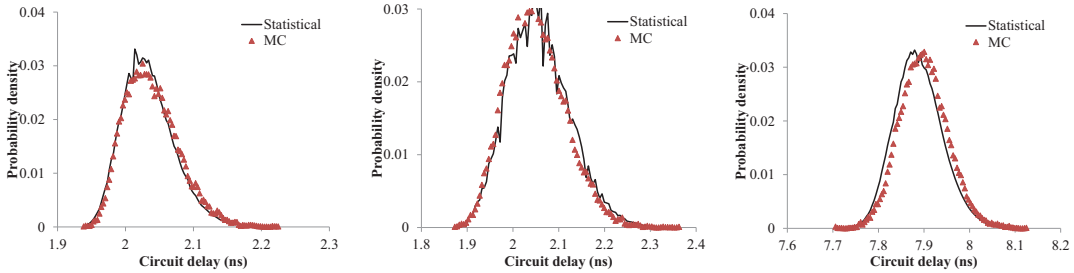
A. Experiment Setup

The experiments are implemented on a PC with an Intel Q9550 CPU. The STA tool, the logic simulator, and the statistical algorithm in Fig. 3 are written in C++. Twenty-four benchmarks including ISCAS85 and some ALU circuits are used to evaluate the proposed algorithm for RTN simulation. The 16nm high-performance PTM [23] is used, with nominal $V_{dd} = 0.9V$ and $|V_{th0}| = 0.4V$. Some key parameters are listed: $\alpha = 1.5$ (in Eq. (4)), single trap induced $|\Delta V_{th}| = 30$mV for smallest devices according to Eq. (2).

B. Results of RTN Evaluation

1) RTN-induced circuit delay degradation: In this experiment, the average number of detectable traps in each device (λ) is set to 2, $\tau_c^{(on)} = 0.01$s and $\tau_e^{(off)} = 0.1$s, $m_c = m_e = 12$. The temporal distribution of circuit delay of four circuits obtained by the proposed algorithm and Monte-Carlo (MC) simulation are shown in Fig. 7. MC is implemented by 10000 times. The proposed algorithm can generally obtain accurate results compared with MC. We find that the RTN-induced temporal distribution of circuit delay has a long tail, as shown in Fig. 8. As mentioned above, RTN-induced circuit delay varies in time randomly, so if chip operates long enough circuit delay can always get its worst value. Consequently, for a reliable design, we should ensure that the maximum possible delay does not violates the design specification, leading to large design redundancy. However, as can be seen from Fig. 8, though the tail is very long, the large delay values can hardly appear during the whole lifetime. Ensuring that circuit functions correctly with $1 - 10^{-10}$ probability or $1 - 10^{-50}$ probability have no difference in essence. Therefor, it is not practical to consider the maximum possible delay which will lead to large design overheads, but we should only ensure that circuit functions correctly with a certain probability P during the whole lifetime, such that timing violation can hardly happen. Consider a 10-year circuit lifetime and the typical values of τ_e and τ_c, we choose $P = 1 - 10^{-9}$. So our algorithm predicts that circuit functions correctly with $1 - 10^{-9}$ probability. The corresponding delay value is called "RTN-induced maximum delay".

The results of all the benchmarks are shown in Table I. We show the circuit intrinsic delay and degradation ratio of the RTN-induced maximum delay. To ensure $1-10^{-9}$ correctness during the whole lifetime, circuit should be protected from the degraded delay rather than the intrinsic delay. The average degradation ratio is 20%. Our statistical algorithm is on



(a) booth9x9 (small size) (b) c7552 (medium size) (c) pmult32x32 (big size)

Fig. 7: Comparison with MC on circuit delay distribution.

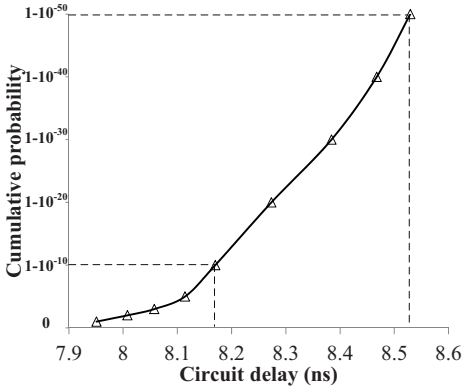


Fig. 8: Long tail of the RTN-induced temporal distribution of circuit delay, for pmult32x32.

average 41× faster than MC (10000 times). The simulation time is at millisecond magnitude, which is also expected to be much faster than SPICE-based approaches. We also compare the accuracy between MC and our method in the error of the mean value of delay distribution. The average error is only 0.53%.

2) Impact of gate voltage: Several publications show that RTN strongly depends on gate voltage at device level. It is shown that the gate voltage V_{gs} has large impact on $\tau_e^{(on)}$ and $\tau_c^{(on)}$, and the dependence is approximately exponential [20], [26]. So $\tau_e^{(on)}$ and $\tau_c^{(on)}$ can be described as

$$\tau_c^{(on)} = \gamma_c e^{-\theta_c V_{gs}}, \tau_e^{(on)} = \gamma_e e^{\theta_e V_{gs}} \quad (12)$$

Based on the data plotted in [26], we choose $\gamma_c = 32$, $\gamma_e = 3.2 \times 10^{-8}$, and $\theta_c = \theta_e = 11.5$. $m_c = m_e = 12$. λ also depends on V_{gs}, and the dependence is approximately linear [20]. When V_{gs} increases, λ also increases. The data plotted in [20] show that $\frac{d\lambda}{dV_{gs}} \approx 1.25$ and $\lambda \approx 2$ under nominal condition, so we choose

$$\lambda = 1.25 \times (V_{gs} - 0.9) + 2 \quad (13)$$

Take pmult32x32 as an example, the impact of V_{gs} on RTN is shown in Fig. 9 and Fig. 10. We get two important observations from the results.

- RTN-induced maximum delay degradation ($d_{max} - d_0$) has two different trends under different V_{gs} (Fig. 10). When $V_{gs} < 1$V, with V_{gs} increasing, τ_e increases, τ_c decreases, and the average number of filled traps $r\lambda = \frac{\lambda\tau_e}{\tau_e+\tau_c}$ also increases, so the maximum delay degradation increases. On the other hand, when $V_{gs} > 1$V, since $\tau_e >> \tau_c$, the average number of filled traps is almost constant ($\frac{\lambda\tau_e}{\tau_e+\tau_c} \approx \lambda$), which means all the detectable traps are almost in

TABLE I: Circuit delay degradation caused by RTN.

benchmark	#gate	d_0(ns)	MC	Statistical algorithm			error(%)
			T(s)	T(s)	Δ(%)	speedup	
c432	169	2.81	0.256	0.008	16.6	31	0.21
c499	204	2.23	0.316	0.007	39.9	46	2.03
c880	383	1.13	0.564	0.018	13.8	32	0.43
c1355	548	1.91	0.782	0.031	21.5	25	0.49
c1908	911	2.77	1.364	0.039	27.0	35	0.94
c2670	1279	1.38	2.044	0.061	27.5	34	0.73
c3540	1699	2.14	2.492	0.058	32.5	43	2.01
c5315	2329	1.87	3.667	0.110	28.6	33	0.13
c6288	2447	6.36	3.411	0.091	12.0	37	0.37
c7552	3566	1.80	5.541	0.193	30.8	29	0.33
array4x4	69	0.84	0.101	0.004	18.2	24	0.00
array8x8	375	2.86	0.541	0.025	17.5	22	2.93
bkung16	81	1.00	0.124	0.002	9.1	57	0.27
bkung32	165	1.94	0.257	0.004	7.1	63	0.13
booth9x9	412	1.90	0.593	0.014	22.4	42	0.23
kogge16	81	1.00	0.124	0.002	9.1	58	0.26
kogge32	164	1.97	0.259	0.004	7.0	64	0.13
log16	140	0.54	0.208	0.006	19.7	32	0.22
log32	371	0.85	0.556	0.020	31.0	28	0.34
log64	862	1.52	1.318	0.038	31.5	35	0.13
pmult4x4	72	0.93	0.107	0.004	18.2	25	0.01
pmult8x8	356	1.93	0.515	0.011	17.1	47	0.05
pmult16x16	1672	3.89	2.486	0.058	11.6	43	0.12
pmult32x32	6814	7.44	22.442	0.235	9.5	96	0.15
average					**20.0**	**41**	**0.53**

d_0 = circuit delay without RTN (intrinsic delay)
T = simulation time
$\Delta = \frac{d_{max}-d_0}{d_0}$, d_{max} = RTN-induced maximum delay
error = error of the mean value of delay distribution

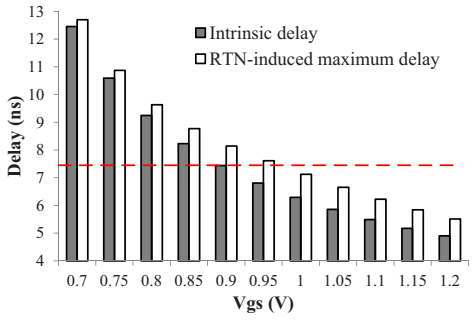


Fig. 9: Intrinsic delay (d_0) and RTN-induced maximum delay (d_{max}), for pmult32x32.

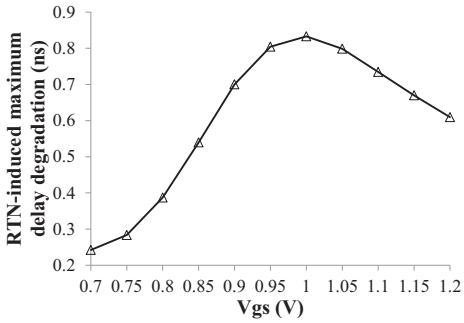


Fig. 10: RTN-induced maximum delay degradation ($d_{max} - d_0$), for pmult32x32.

filled state (although λ also depends on V_{gs}, the dependence is much weaker than that of the time constants); in the mean time, higher V_{gs} leads to lower intrinsic delay, so RTN-induced maximum delay degradation decreases. This phenomenon can be called "saturation" of traps.

- Although the RTN-induced maximum delay degradation ($d_{max} - d_0$) increases when $V_{gs} < 1$V, the maximum delay (d_{max}) still keeps decreasing with V_{gs} increasing. This indicates that a simple guard-banding approach can protect circuit from the effect of RTN. As shown in Fig. 9, if we choose the intrinsic delay when V_{gs} =0.9V (nominal voltage) as the design constraint, the RTN-induced maximum delay satisfies the constraint when $V_{gs} \geq 1$V.

V. Conclusions

This paper integrates RTN into timing analysis and analyzes the impact of multiple traps induced RTN on the temporal performance of digital circuits. We show the temporal distribution of circuit delay under RTN, and circuit should be protected from a degraded delay which is on average 20% larger than intrinsic delay to ensure $1-10^{-9}$ correctness during the whole lifetime. The impact of gate voltage on RTN is investigated. Our results show that a simple guard-banding approach can effectively protect circuit from the effect of RTN.

References

[1] J. Campbell, J. Qin, K. Cheung, L. Yu, J. Suehle, A. Oates, and K. Sheng, "The origins of random telegraph noise in highly scaled SiON nMOSFETs," in *IRW*, oct. 2008, pp. 1–16.

[2] 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. 3C6, pp. 293–300, 2003.

[3] 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 *VLSIT*, june 2009, pp. 50–51.

[4] A. Ghetti, C. Compagnoni, F. Biancardi, A. Lacaita, S. Beltrami, L. Chiavarone, A. Spinelli, and A. Visconti, "Scaling trends for random telegraph noise in deca-nanometer flash memories," in *IEDM*, dec. 2008, pp. 1–4.

[5] T. Matsumoto, K. Kobayashi, and H. Onodera, "Impact of random telegraph noise on cmos logic delay uncertainty under low voltage operation," in *IEDM*, 2012, pp. 25.6.1–25.6.4.

[6] M. J. Uren, D. J. Day, and M. J. Kirton, "1/f and random telegraph noise in silicon metal-oxide-semiconductor field-effect transistors," *Applied Physics Letters*, vol. 47, no. 11, pp. 1195–1197, dec 1985.

[7] K. Hung, P. Ko, C. Hu, and Y. Cheng, "A unified model for the flicker noise in metal-oxide-semiconductor field-effect transistors," *TED*, vol. 37, no. 3, pp. 654–665, mar 1990.

[8] T. Grasser, H. Reisinger, W. Goes, T. Aichinger, P. Hehenberger, P.-J. Wagner, M. Nelhiebel, J. Franco, and B. Kaczer, "Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise," in *IEDM*, dec. 2009, pp. 1–4.

[9] S.-M. Joe, J.-H. Yi, S.-K. Park, H. Shin, B.-G. Park, Y. J. Park, and J.-H. Lee, "Threshold voltage fluctuation by random telegraph noise in floating gate nand flash memory string," *TED*, vol. 58, no. 1, pp. 67–73, jan. 2011.

[10] 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 *IRPS*, 27 2008-may 1 2008, pp. 541–546.

[11] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, "Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate flash memory," in *IEDM*, dec. 2006, pp. 1–4.

[12] 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 *VLSIT*, june 2010, pp. 95–96.

[13] 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*, dec. 2009, pp. 1–4.

[14] 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*, march 2011, pp. 1–6.

[15] 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*, june 2011, pp. 292–297.

[16] 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.

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

[18] T. B. Tang, A. Murray, and S. Roy, "Methodology of statistical RTS noise analysis with charge-carrier trapping models," *TCAS-I*, vol. 57, no. 5, pp. 1062–1070, may 2010.

[19] H. Luo, Y. Wang, Y. Cao, Y. Xie, Y. Ma, and H. Yang, "Temporal performance degradation under RTN: Evaluation and mitigation for nanoscale circuits," in *ISVLSI*, aug. 2012, pp. 183–188.

[20] T. Nagumo, K. Takeuchi, S. Yokogawa, K. Imai, and Y. Hayashi, "New analysis methods for comprehensive understanding of random telegraph noise," in *IEDM*, dec. 2009, pp. 1–4.

[21] A. Van Der Wel, E. Klumperink, L. K. J. Vandamme, and B. Nauta, "Modeling random telegraph noise under switched bias conditions using cyclostationary rts noise," *TED*, vol. 50, no. 5, pp. 1378–1384, 2003.

[22] C. Amin, N. Menezes, K. Killpack, F. Dartu, U. Choudhury, N. Hakim, and Y. Ismail, "Statistical static timing analysis: how simple can we get?" in *Design Automation Conference, 2005. Proceedings. 42nd*, 2005, pp. 652–657.

[23] Nanoscale Integration and Modeling (NIMO) Group, ASU, "Predictive Technology Model (PTM)." [Online]. Available: http://ptm.asu.edu/

[24] J.-J. Liou, K.-T. Cheng, S. Kundu, and A. Krstic, "Fast statistical timing analysis by probabilistic event propagation," in *DAC*, 2001, pp. 661–666.

[25] W.-J. Tsay and P.-H. Ke, "A simple approximation for bivariate normal integral based on error function and its application on probit model with binary endogenous regressor," IEAS Working Paper : academic research 09-A011, Nov. 2009.

[26] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps," in *IEDM*, dec. 2010, pp. 28.3.1–28.3.4.