Published in IET Circuits, Devices & Systems
Received on 22nd November 2012
Revised on 23rd January 2013
Accepted on 25th January 2013
doi: 10.1049/iet-cds.2012.0361

Special Issue: Design Methodologies for Nanoelectronic Digital and Analogue Circuits

ISSN 1751-858X

Evaluation and mitigation of performance degradation under random telegraph noise for digital circuits

Xiaoming Chen[1], Hong Luo[1], Yu Wang[1], Yu Cao[3], Yuan Xie[4], Yuchun Ma[2], Huazhong Yang[1]

[1]Department of Electronic Engineering, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, People's Republic of China
[2]Department of Computer Science, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, People's Republic of China
[3]Department of ECEE, Arizona State University, Tempe, Arizona 85287-5706, USA
[4]Department of CSE, Pennsylvania State University, Pennsylvania 16802, USA
E-mail: chenxm05@mails.tsinghua.edu.cn

Abstract: Random telegraph noise (RTN) has become an important reliability issue in nanoscale circuits recently. This study proposes a simulation framework to evaluate the temporal performance of digital circuits under the impact of RTN at 16 nm technology node. Two fast algorithms with linear time complexity are proposed: statistical critical path analysis and normal distribution-based analysis. The simulation results reveal that the circuit delay degradation and variation induced by RTN are both >20% and the maximum degradation and variation can be >30%. The effect of power supply tuning and gate sizing techniques on mitigating RTN is also investigated.

1 Introduction

In recent years, as the channel length of MOSFETs continues to shrink into nanoscale, a variety of reliability mechanisms, such as negative bias temperature instability [1, 2], time-dependent dielectric breakdown [3] and random telegraph noise (RTN) [4], are becoming key challenges for circuit designers. During the working life of devices, these physical phenomena will degrade the electrical parameters such as the drain current (I_d) and the threshold voltage (V_{th}), leading to degradation of the circuit operation speed and logic failure. This paper addresses RTN since it is an emerging research topic.

RTN can cause electrical parameters (such as V_{th} and I_d) to exhibit random fluctuations as a function of time [5]. Recent studies have shown that the RTN-induced fluctuation becomes quite large and can be more significant than the random dopant fluctuation at 22 nm technology node [6]. For example, the drain current fluctuation induced by RTN has been already identified as a large obstacle in both sub-V_{th} and super-V_{th} operation of digital circuits [7]. The variation of I_d caused by RTN can be up to 40% for 30×30 nm devices [8].

The physics of RTN has been widely investigated [7–10] and the RTN effect on SRAM and flash memories has been also studied [11–16]. Although some models which can be integrated into HSPICE analysis have been proposed [17–19], the impact of RTN on the temporal performance of digital circuits has been rarely studied [20]. Therefore our contributions in this paper distinguish itself in the following aspects:

- This paper proposes a simulation framework to evaluate the impact of RTN on the temporal performance of digital circuits. Two fast simulation methods are proposed: statistical critical path analysis (SCPA) and normal distribution-based analysis (NDA). The computational complexity of the two methods are both $O(N)$.
- The impact of RTN on circuit delay degradation and variation is investigated. The experimental results show that RTN degrades the circuit delay and increases the delay variation. The average delay degradation and variation are both >20% at 16 nm technology node. The results also demonstrate that the performance degradation and variation will grow rapidly with supply voltage scaling down.
- The effect of power supply tuning and gate sizing techniques on mitigating RTN is investigated. The simulation results show that gate sizing is better than power supply tuning.

The rest of the paper is organised as follows. Section 2 reviews some previous work on RTN. Section 3 introduces the RTN model used in this paper. Section 4 proposes the RTN simulation framework and the evaluation methods. The simulation results are presented in Section 5. The impact of design techniques on RTN mitigation is investigated in Section 6. Finally, Section 7 concludes the paper.

2 Related work

Over the last decade, studies on RTN mainly focused on the physics of RTN. It was suggested that RTN was originated

from the capture and emission of the channel carriers by interface traps [9]. A systematic study of the channel length, width and gate overdrive dependencies of RTN effects was carried out in [7]. A new method to characterise the oxide traps considering the energy band structure of high-*k*/metal gate MOSFETs was proposed in [10]. In [21], a method to determine whether an oxide trap leading to RTN was located in the high-*k* layer or the interface layer was proposed.

The RTN effect in SRAM and flash memories has been investigated recently. For example, the RTN effect in deca-nanometer flash memories was investigated in [11] and the statistical distribution of V_{th} was also analysed. The read/write margins of scaled-down SRAM with/without RTN were simulated in [12]. In [14], the impact of RTN on V_{min} in scaled SRAM was analysed. It was reported that RTN-induced V_{min} degradation could be up to 50 mV in 45 nm SRAM [13]. An accurate computational method for trap-level, non-stationary analysis of RTN in SRAMs was presented in [15] and a technique for predicting the impact of RTN on SRAMs/DRAMs in the presence of variability was further proposed in [16]. However, the continuous-time simulation approach used in [16] was too complex and not suitable for circuit-level performance evaluation.

It is believed that RTN can be also a serious issue in digital circuits. A Shockley–Read–Hall-based model to explain the RTN behaviour was proposed in [17]. A methodology to include RTN in circuit analysis was proposed in [18] and the transient analysis was applied on the four-quadrant Chible multiplier circuit. A two-stage L-shaped circuit to generate RTN signal which was fully compatible with SPICE was proposed in [19]. In [20], a time-domain delay model was used to simulate and measure the fluctuation of RTN. However, this approach could be only applied to simple circuits such as SRAM cells and ring oscillators because of the extraordinary computational complexity. Hence in this paper, the delay characterisation of digital circuits is investigated and two fast algorithms are performed on circuit-level analysis for RTN. Design techniques for mitigating RTN are further studied, enabling time-domain analysis in nanoscale digital circuit design.

3 Modelling random telegraph noise

This section first presents the physics of RTN and then the RTN-induced ΔV_{th} model for digital circuits is introduced.

3.1 Physics of RTN

The RTN effect is originated from the capture/emission of charge carriers by the oxide traps, which will induce correlated fluctuations of channel carrier number and mobility [9]. As shown in Fig. 1*a*, a carrier (the solid circle) is occasionally captured by a trap (the hollow circle) in the oxide and the carrier 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* [22]. 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 the two states is inherently random and the activity of a single trap can be modelled as a two-state time-inhomogeneous Markov chain [15].

In the time domain, because of 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 and the time spent in this state is the emission time τ_e. At the other side, the low level of I_d corresponds to the high level of RTN, at which the carrier is captured by the trap and the trap is filled and the time spent in this state is the capture time τ_c [9]. Both the capture time τ_c and emission time τ_e are time-varying and they depend on the position of the traps, the trap energy level and the gate overdrive voltage $V_{gs} - V_{th}$ [9, 15]. The typical values of τ_c and τ_e are about 1–1000 ms [9].

In the frequency domain, the power spectral density of the drain current I_d shows a Lorentzian shaped spectrum with the slope of $1/f^2$, as shown in Fig. 2*b* [10]. The cut-off frequency is

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

The time constant τ_{cut} is defined as [19]

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

3.2 RTN-induced V_{th} fluctuation in digital circuits

To model the RTN effect in digital circuits, the equivalent circuit is used [14], 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 the threshold voltage. The right device shows the low-current state induced by RTN, which is modelled by a shift in the threshold voltage ΔV_{th} and the shift is given by Ye *et al.* [19]

$$\Delta V_{th} = \frac{nq}{C_{ox}WL} \tag{3}$$

where n is the number of oxide traps, q is the elementary

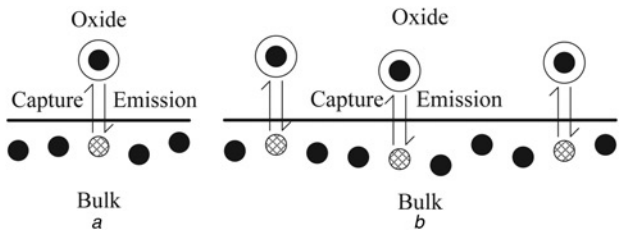

Fig. 1 *Capture/emission process of RTN*

a Single trap
b Multiple traps

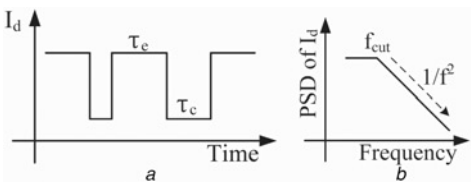

Fig. 2 *Drain current I_d caused by RTN*

a Time domain
b Frequency domain

doi: 10.1049/iet-cds.2012.0361

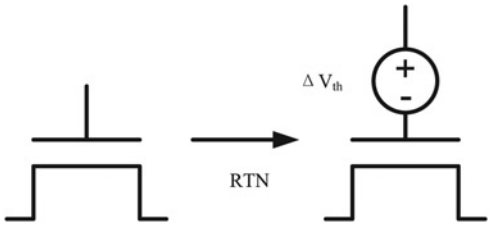

Fig. 3 *Equivalent circuit of RTN effect*

charge, C_{ax} is the unit area capacitance, whereas W and L are the channel width and channel length, respectively.

Since the magnitude of single-trap-induced RTN sharply goes up as device shrinks [19], this paper targets at the single-trap-induced RTN fluctuation as shown in Fig. 1*a*. Equation (3) indicates that RTN depends on the area of the device and experiments show that the gate overdrive voltage can also affect the RTN amplitude, and hence the V_{gs} dependence of ΔV_{th} is an approximate quadratic function [20]

$$\Delta V_{th} = \frac{\lambda\left(V_{gs} - V_{th}\right)^2}{WL} \tag{4}$$

where λ is a constant that can be fitted by experimental data. It is shown that ΔV_{th} can be >70 mV for the smallest devices at 22 nm technology node [6, 23] shows that the RTN amplitude increases superlinearly with the scaling down of the device's size. Hence, ΔV_{th} is expected to be as much as 130 mV at 16 nm technology node.

4 RTN evaluation in digital circuits

As described in Section 2, the capture time τ_c and emission time τ_e are both at millisecond-order [9], whereas the clock cycle of a digital circuit is at nanosecond-order. The operation of a digital circuit is much faster than the transition between high- and low-current states, thus during the operation time $[t,\ t+\Delta t)$ of the digital circuit, all the traps are considered to keep their filled/empty states. Therefore the 'sampling' method can be used 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 a MOSFET at time t can be described by a random variable $S(t)$, which has two discrete values: 0 corresponding to empty state and 1 corresponding to filled state. The probability distribution of $S(t)$ is determined by the capture time and emission time, which is given by

$$\begin{cases} P(S(t)=0) = \dfrac{\bar{\tau}_e}{\bar{\tau}_e + \bar{\tau}_c} = \dfrac{1}{1+r} \\ P(S(t)=1) = \dfrac{\bar{\tau}_c}{\bar{\tau}_e + \bar{\tau}_c} = \dfrac{r}{1+r} \end{cases} \tag{5}$$

where $r = \left(\bar{\tau}_c/\bar{\tau}_e\right)$, which is a constant only depending on the trap energy level and Fermi level and its typical value is from 0.1 to 10 [19].

Thus, when the circuit is 'sampled' at time t, the threshold voltage of a given MOSFET is

$$V_{th}(t) = V_{th\,0} + S(t)\Delta V_{th} \tag{6}$$

where V_{th0} is the initial threshold voltage.

Since all the traps in the device are independent, all S's are independent. Therefore by the 'sampling' method, Monte-Carlo (MC) simulations can be adopted to evaluate the circuit performance under RTN. One MC simulation can be considered as one 'sample' at some time node of the given circuit and the value of S can be randomly set to 0 or 1 according to the value of r. Then, traditional static timing analysis (STA) tools can be used for subsequent simulations. However, the MC simulations are time-consuming. Thus, new faster simulation algorithms will be proposed in the following sections.

4.1 RTN evaluation framework

The proposed framework for RTN evaluation is shown in Fig. 5. First, HSPICE is used to create a gate library based on the 16 nm predictive technology model (PTM) [24]. The gate library includes delay, area and oxide capacitance of each gate type (i.e. NAND2X1, NAND2X4, OR2X1 etc). Then, a private STA tool written in C++ is used to calculate the delay of all the paths in the circuit and find the critical paths. An RTN ΔV_{th} calculator is used to calculate ΔV_{th} of all the gates according to (4). Finally, the delay distribution of the circuit is calculated by a delay distribution calculator. In the next two sections, we will introduce two algorithms to perform the distribution

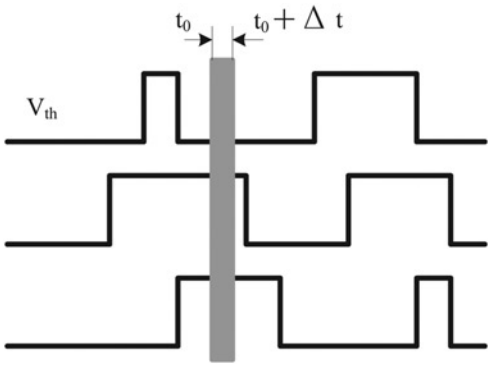

Fig. 4 *Sampling the high and low states of devices induced by RTN*

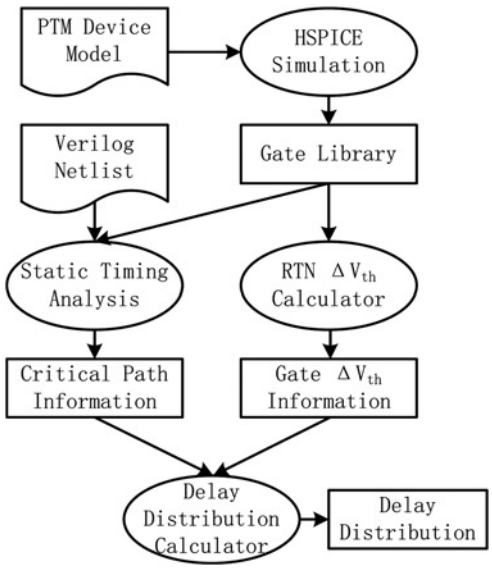

Fig. 5 *RTN evaluation framework*

calculation step. The first method is called SCPA method and the second is called NDA method.

4.2 Statistical critical path analysis

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

$$d_{\mathrm{c}} = \max_i \left\{ d_{\mathrm{cp},i} \right\} \tag{7}$$

where d_{c} is the maximum circuit delay and $d_{\mathrm{cp},i}$ is the delay of the ith critical path. The delay of a critical path is

$$d_{\mathrm{cp}} = \sum_j d_j \tag{8}$$

where d_j is the delay of the jth gate in the path. The propagation delay of a logic gate j is

$$d_j = \frac{K_j C_{L,j} V_{\mathrm{dd}}}{A_j \left(V_{\mathrm{dd}} - V_{\mathrm{th},j} \right)^{\alpha}} \tag{9}$$

where K_j is a coefficient related with device physical parameters, A_j is the equivalent area of the gate, $C_{L,j}$ is the load capacitance and α is the velocity saturation index. Combined with (6), the RTN-induced delay shift of gate j is

$$\Delta d_j \simeq \frac{\alpha S \Delta V_{\mathrm{th},j}}{V_{\mathrm{dd}} - V_{\mathrm{th0}}} \times d_j \tag{10}$$

Hence Δd_j is also a random variable and has a similar probability distribution as S, which is given by

$$\begin{cases} P(\Delta d_j = 0) = \dfrac{1}{1+r} \\ P\left(\Delta d_j = \dfrac{\alpha \Delta V_{\mathrm{th},j}}{V_{\mathrm{dd}} - V_{\mathrm{th0}}} \times d_j \right) = \dfrac{r}{1+r} \end{cases} \tag{11}$$

For simplicity, let

$$p = \frac{1}{1+r}, \quad q = \frac{r}{1+r} (p + q = 1), \quad \text{and}$$
$$t_j = \frac{\alpha \Delta V_{\mathrm{th},j}}{V_{\mathrm{dd}} - \mathrm{V}_{\mathrm{th0}}} \times d_j$$

The delay shift of a critical path is also a random variable

$$\Delta d_{\mathrm{cp}} = \sum_j \Delta d_j \tag{12}$$

where Δd_{cp} varies from 0 to $\Sigma_j t_j$. The probability distribution of Δd_{cp} can be calculated by convoluting all the probability distributions of Δd_j's in the path (i.e. first the convolution of d_1 and d_2 is calculated, then d_3 is added and finally all d_j's are summed up), since they are independent.

Finally, the delay shift of the circuit caused by RTN is the maximum distribution of all the critical paths

$$\Delta d_{\mathrm{c}} = \max_i \left\{ \Delta d_{\mathrm{cp},i} \right\} \tag{13}$$

The cumulative distribution function (CDF) of Δd_c is the product of all the CDF's of $\Delta d_{\mathrm{cp},i}$.

For a given critical path, since each Δd_j has two discrete values: 0 and t_j, Δd_{cp} will have 2^N discrete values, where N is the number of gates in the path. This indicates that it is impractical to directly calculate the distribution of (12), since the time and space complexity are both $O(2^N)$.

To reduce the complexity, we use a grouping method to construct the approximate distribution of the partial sum $\phi_L = \sum_{j=1}^{L<N} \Delta d_j$. First, a new random variable Φ is constructed, whose distribution is defined by

$$P(m\delta < \Phi \le (m+1)\delta) = \sum_{m\delta < x \le (m+1)\delta} p_L(x) \tag{14}$$

where $m = 0 \ \ldots \ M-1$, $\delta = (1/M) \sum_{j=1}^{L} t_j$ and $p_L(x)$ is the probability mass function (PMF) of ϕ_L. Here, M is a user-defined parameter and larger M leads to better approximation. Second, the probability distribution of Φ is denoted by the probability of M discrete values, which is given by

$$p_{\Phi}((m+0.5)\delta) = P(m\delta < \Phi \le (m+1)\delta) \tag{15}$$

This method redistributes 2^L discrete values into M discrete values. In this paper, $M = 64$ is adopted.

Obviously, by using the grouping method, the computational complexity reduces to $O(2MN)$. Since M is a constant, the computational complexity is $O(N)$. This algorithm is described in Algorithm 1 (see Fig. 6).

4.3 Normal distribution-based analysis

This section presents another alternative method to calculate the delay distribution of the circuit, called NDA, which is based on the following theorem.

Theorem: For a given critical path that has N gates, the delay shift of each gate caused by RTN is described by (11), then

$$\lim_{N \to \infty} \Delta d_{\mathrm{cp}} = N \left(\sum_{j=1}^{N} E\left(\Delta d_j \right), \sum_{j=1}^{N} D\left(\Delta d_j \right) \right) \tag{16}$$

where $N(\cdot,\cdot)$ denotes the normal distribution, $E(\cdot)$ and $D(\cdot)$ are the expectation and variance, respectively.

Proof: Following (11), the expectation and variance of Δd_j are

$$\begin{cases} E\left(\Delta d_j \right) = q t_j \\ D\left(\Delta d_j \right) = p q t_j^2 \end{cases} \tag{17}$$

doi: 10.1049/iet-cds.2012.0361

Algorithm 1

$L = 1$;

$p_L(x) = PMF(\Delta d_1)$;

for $L = 2 \rightarrow N$ **do**

$p_L(x) = p_L(x) \otimes PMF(\Delta d_L)$;

if $2^L > M$ **then**

Construct Φ by the grouping method using Eq. (14) and (15);

end if

end for

return Distribution of Φ;

Fig. 6 *Algorithm for calculating critical path delay distribution*

Let $B_N^2 = \sum_{j=1}^{N} D\left(\Delta d_j\right) = pq \sum_{j=1}^{N} t_j^2$, for any positive constant $\delta > 0$, we have

$$
\begin{aligned}
f(N) &= \frac{1}{B_N^{2+\delta}} \sum_{j=1}^{N} E\left(\left|\Delta d_j - E\left(\Delta d_j\right)\right|^{2+\delta}\right) \\
&= \frac{\sum_{j=1}^{N}\left(p\left(qt_j\right)^{2+\delta} + q\left(pt_j\right)^{2+\delta}\right)}{\left(pq\sum_{j=1}^{N} t_j^2\right)^{1+(\delta/2)}} \\
&= \frac{\left(p^{1+\delta} + q^{1+\delta}\right)\sum_{j=1}^{N} t_j^{2+\delta}}{(pq)^{(\delta/2)}\left(\sum_{j=1}^{N} t_j^2\right)^{1+(\delta/2)}} \\
&= \gamma \frac{\sum_{j=1}^{N} t_j^{2+\delta}}{\left(\sum_{j=1}^{N} t_j^2\right)^{1+(\delta/2)}} \qquad (18)
\end{aligned}
$$

where $\gamma = (p^{1+\delta} + q^{1+\delta}/((pq)^{\delta/2}))$ is a positive constant.

In practice, all t_j's are limited in a range $[t_{\min}, t_{\max}]$ ($t_{\max}$ and $t_{\min}$ are constants, $t_{\max} > t_{\min} > 0$), hence we have

$$
\begin{aligned}
f(N) &\le \gamma \frac{N t_{\max}^{2+\delta}}{\left(N t_{\min}^2\right)^{1+(\delta/2)}} \\
&= \gamma \frac{1}{N^{(\delta/2)}}\left(\frac{t_{\max}}{t_{\min}}\right)^{2+\delta} \rightarrow 0 (N \rightarrow \infty) \qquad (19)
\end{aligned}
$$

This reveals that $\Delta d_{cp} = \sum_{j=1}^{N} \Delta d_j$ satisfies the condition of Lyapunov's central limit theorem (CLT) [25], hence Δd_{cp} is a normal distribution when N is infinite. The expectation and variance are $\sum_{j=1}^{N} E\left(\Delta d_j\right)$ and $\sum_{j=1}^{N} D\left(\Delta d_j\right)$, respectively. □

Based on the above theorem, we suppose that the delay of each critical path follows normal distribution since N is usually large enough to fit the CLT, then the distribution of circuit delay is the maximum distribution of several independent normal distributions, which can be calculated by Clark's formula [26] and the maximum distribution is still a normal distribution.

We believe that NDA is faster than SCPA, since the computation is much simpler. However, if N is small, NDA will get large error.

5 Experimental results

5.1 Experiment setup

The experiments are implemented on a PC with an Intel Q9550 CPU and 4 GB DRAM. 24 ISCAS85 and ALU circuits are used to evaluate the proposed algorithms. The device model is the 16 nm high-performance PTM model [24], with nominal $V_{dd} = 0.9$ V and $|V_{th0}| = 0.4$ V. Some key parameters are: $r = 1$ [in (5)], $\alpha = 1.5$ [in (9)], maximum $\Delta|V_{th}| = 120$ mV for the smallest devices and the load capacitance of each output pin is 1×10^{-17} F. HSPICE is used to build the gate library and other simulators in Fig. 5 are written in C++.

5.2 Comparison with MC

This section compares the results obtained from SCPA and NDA with MC simulation. Two examples (c3540 and log64) are shown in Figs. 7 and 8. The X-axis is the delay values and the Y-axis is the probability.

For c3540, the expectation of the circuit delay is 2.89 ns, which is obtained by MC; whereas SCPA and NDA both get 2.85 ns, the relative error is only 1.4%. In addition, SCPA, NDA and MC all get similar distributions for c3540.

For log64, SCPA and MC obtain similar distributions. However, the distribution shape obtained by NDA is significantly different from that obtained by MC or SCPA. The reason is that NDA assumes the circuit delay is a normal distribution, but the maximum length of the critical paths of log64 is only 11, which does not fit the CLT. Fortunately, for most circuits, the maximum length of the

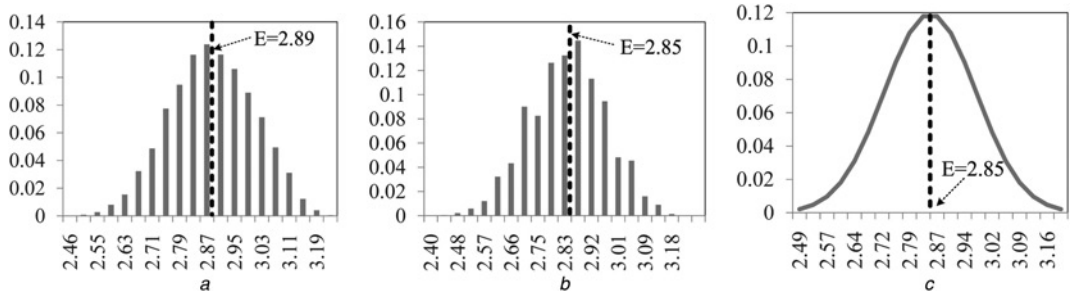

Fig. 7 *Delay distribution of c3540 caused by RTN*

a MC
b SCPA
c NDA

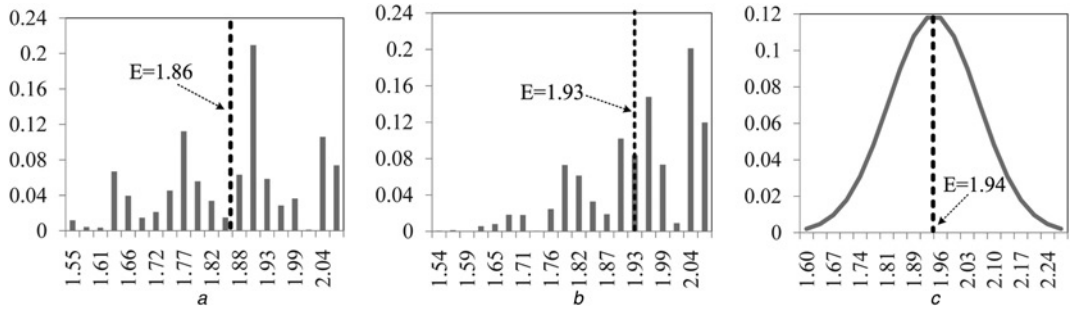

Fig. 8 *Delay distribution of log64 caused by RTN*

a MC
b SCPA
c NDA

critical paths are large enough to fit the CLT, and hence NDA is ineffective for only few circuits.

Table 1 shows the simulation time of MC, SCPA and NDA, together with the setup time, number of gates and the maximum length of the critical paths. Obviously, SCPA and NDA are both much faster than MC. It shows that on average, SCPA is about 1000× faster than MC and NDA is about 50× faster than SCPA. Hence SCPA and NDA can be both used for larger-scale circuits.

5.3 Circuit delay distribution analysis

Table 2 shows the delay distribution obtained by MC, SCPA and NDA. The average delay degradation is calculated by $\Delta d_{\text{avg}} = ((E(d_c) - d_0)/d_0)$, where $E(d_c)$ is the expectation of the circuit delay under RTN. For MC and SCPA, the delay variation is calculated by $\Delta d_{\text{var}} = ((d_{\max} - d_{\min})/(E(d_c)))$; whereas for NDA, $\Delta d_{\text{var}} = (6\sigma/(E(d_c)))$, where $\sigma = \sqrt{D(\Delta d_c)}$, $D(\Delta d_c)$ is the variance of circuit delay (shift) obtained by NDA.

According to Table 2, the average delay degradation and variation are both >20%. Meanwhile, the maximum delay degradation and variation can be >30%. The results demonstrate that RTN will be a very serious obstacle in circuit reliability in the deca-nanometer regime, which exhibits in the following two aspects:

- RTN can cause significant circuit performance degradation, leading to serious timing violation. The possible minimum delay as shown in Figs. 7 and 8 is still greater than d_0. Hence, the RTN effect must be considered in circuit design.
- The RTN-induced delay variation can lead to greater non-determinacy on circuit delay. Thus, statistical analysis should be considered in RTN evaluation.

Table 1 Comparison of simulation time, all the time values are shown in milliseconds

Benchmark	#gate	#len[a]	Setup[b], ms	MC[c], ms	SCPA, ms	NDA, ms
c432	169	21	2	228	0.14	0.008
c499	204	14	3	226	0.30	0.013
c880	383	22	4	498	0.28	0.010
c1355	548	27	8	594	1.16	0.016
c1908	911	46	14	1031	0.75	0.014
c2670	1279	30	18	1460	0.41	0.023
c3540	1699	38	25	1961	0.50	0.010
c5315	2329	43	51	2685	1.61	0.023
c6288	2447	125	54	2970	2.80	0.020
c7552	3566	37	93	4123	1.25	0.018
array4 × 4	69	20	1	73	0.09	0.013
array8 × 8	375	53	4	420	0.67	0.012
bkung16	81	31	1	86	0.30	0.014
bkung32	165	59	2	180	1.19	0.021
booth9 × 9	412	30	5	467	0.35	0.014
kogge16	81	31	1	86	0.29	0.014
kogge32	164	61	2	178	1.25	0.018
log16	140	8	2	159	0.05	0.008
log32	371	10	4	427	0.19	0.012
log64	862	11	14	1025	0.27	0.012
pmult4 × 4	72	15	1	78	0.06	0.013
pmult8 × 8	356	35	4	408	0.39	0.010
pmult16 × 16	1672	75	41	2085	1.43	0.017
pmult32 × 32	6814	165	382	7924	6.22	0.029

[a]'#len' means the maximum length of the critical paths
[b]'steup' means the setup time, including reading circuit netlist, building internal data structure, STA and gate ΔV_{th} calculation
[c]Simulation time of 10 000-time MC simulations

doi: 10.1049/iet-cds.2012.0361

Table 2 Circuit delay distribution caused by RTN

Benchmark	d_0^a, ns	MC		SCPA		NDA	
		Δd_{avg}, %	Δd_{var}, %	Δd_{avg}, %	Δd_{var}, %	Δd_{avg}, %	Δd_{var}, %
c432	2.81	19.2	22.9	19.9	22.5	20.1	19.1
c499	2.23	39.3	25.5	32.8	32.3	33.3	42.0
c880	1.13	20.3	26.9	22.6	26.0	22.9	19.7
c1355	1.91	28.5	20.0	24.6	26.1	24.7	23.2
c1908	2.77	22.0	27.6	25.0	27.8	26.1	29.3
c2670	1.38	30.8	32.3	32.0	33.1	32.4	26.0
c3540	2.14	34.5	28.1	32.5	34.7	32.8	26.1
c5315	1.87	33.6	25.2	31.7	34.3	32.1	29.3
c6288	6.36	21.3	7.9	15.5	9.3	19.2	6.4
c7552	1.80	31.1	29.1	31.9	33.6	32.3	30.3
array4 × 4	0.84	17.2	24.2	19.2	22.8	19.4	19.1
array8 × 8	2.86	21.7	13.9	16.1	16.4	19.9	12.2
bkung16	1.00	14.1	17.5	15.9	17.0	16.3	11.8
bkung32	1.94	13.8	12.8	13.3	12.2	15.4	8.0
booth9 × 9	1.90	18.1	20.1	19.2	21.1	19.8	20.0
kogge16	1.00	14.0	17.6	15.9	17.0	16.3	11.8
kogge32	1.97	13.8	12.4	13.3	12.2	15.4	7.9
log16	0.54	14.4	23.0	19.9	22.2	20.1	30.1
log32	0.85	21.2	27.2	26.7	27.7	26.8	38.5
log64	1.52	22.4	29.3	27.1	28.9	27.3	36.8
pmult4 × 4	0.93	16.0	27.4	19.3	22.8	19.6	19.0
pmult8 × 8	1.93	16.3	20.5	18.8	20.1	18.6	12.9
pmult16 × 16	3.89	16.7	13.5	11.7	11.2	17.6	9.2
pmult32 × 32	7.44	16.2	10.2	12.8	9.6	17.0	6.9
average		21.4	21.4	20.5	22.5	22.7	20.6

[a] d_0 is the circuit intrinsic delay, without the RTN effect

5.4 Power supply scaling analysis

Equation (4) shows that the circuit delay degradation can be affected by the power supply voltage (V_{dd}) and scaling down of V_{dd} decreases the RTN effect. The performance degradation and variation under different V_{dd} for c1355 and c3540 are shown in Fig. 9, which are obtained by NDA. The results show that with V_{dd} scaling down, both the temporal performance degradation and variation decrease. However, when V_{dd} decreases, the intrinsic delay increases.

6 RTN mitigation in digital circuits

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

6.1 Power supply tuning

This section investigates the impact of V_{dd} tuning on the maximum circuit delay under RTN. Although increasing V_{dd} increases the delay degradation and variation (Fig. 9), the circuit intrinsic delay is reduced and the maximum circuit delay under RTN still decreases, as shown in Fig. 10. However, if the intrinsic delay at $V_{dd} = 09$ V is chosen as the design specification (i.e. $d_0(V_{dd} = 0.9$ V)), the maximum circuit delay at $V_{dd} = 1.1$ V can not satisfy the design specification. In addition, the dynamic power increases by 49.4% when $V_{dd} = 1.1$ V.

To simultaneously reduce the RTN-induced maximum delay and the dynamic power overhead by V_{dd} tuning, the

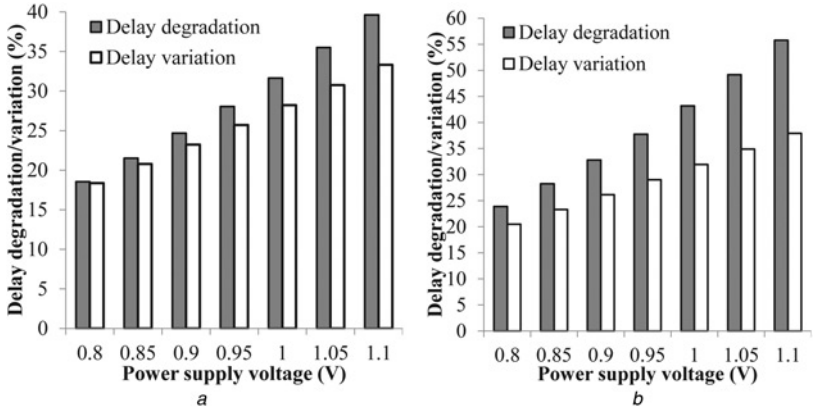

Fig. 9 *Percentage of delay degradation and variation with different V_{dd}, obtained by NDA*

a c1355
b c3540

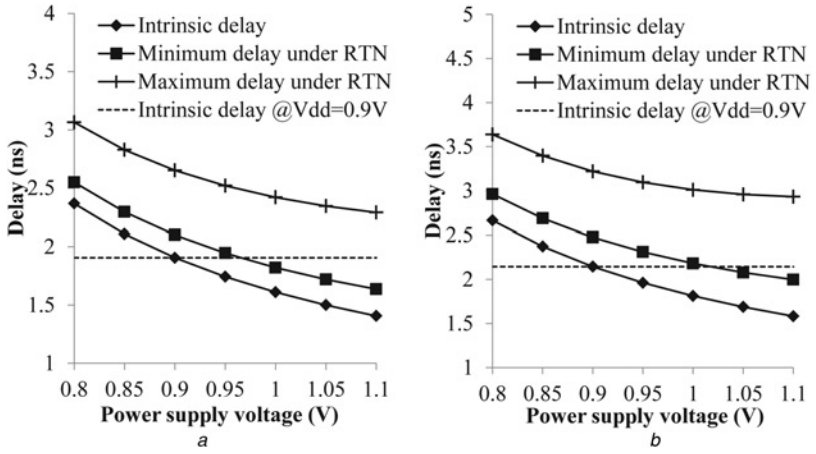

Fig. 10 *Delay degradation and variation using V_{dd} tuning, obtained by NDA*
a c1355
b c3540

dual V_{dd} assignment technique can be adopted. In this method, only the gates along the critical paths are tuned to high V_{dd}. The simulation results are shown in Table 3, obtained by MC. In this table, 'full' means that all the gates are tuned to high V_{dd} and 'critical' means the dual V_{dd} method.

$$\Delta d_{max} = \frac{d_{max} - d_0(V_{dd} = 0.9\,\text{V})}{d_0(V_{dd} = 0.9\,\text{V})},$$

$$\Delta d_{var} = \frac{d_{max} - d_{min}}{E(d_c)}, \quad \text{and} \quad \Delta P$$

is the dynamic power overhead. In this experiment, high V_{dd} is 1.1 V. The results reveal that average $\Delta d_{max} = 33.6\%$ when $V_{dd} = 0.9$ V (nominal design). By using the 'full' tuning method, the maximum delay is 12.8% larger than the design specification and the delay variation is increased to 27.2%. By using the dual V_{dd} method, the maximum delay is 13.9% larger than the design specification and the power overhead is 20.3%, less than a half of that in the 'full' tuning method. This reveals that the V_{dd} tuning method can reduce the RTN-induced maximum delay compared with the nominal design. However, the efficiency is very limited and the power overhead is large. Actually the effect of V_{dd} tuning completely comes from the reduction of the intrinsic delay.

6.2 Gate sizing and replacement

Equation (4) indicates that RTN strongly depends on the area of the device. Thus, this section investigates the effect of the

Table 3 Results of V_{dd} tuning, obtained by MC

Benchmark	Nominal[a]	'full'			'critical'		
	Δd_{max}, %	Δd_{max}, %	Δd_{var}, %	ΔP, %	Δd_{max}, %	Δd_{var}, %	ΔP, %
c432	31.7	9.7	28.6	49.4	10.1	29.6	23.3
c499	45.2	28.5	29.1	49.4	28.5	27.3	35.9
c880	36.0	15.1	34.3	49.4	14.9	32.2	12.1
c1355	38.3	18.2	25.9	49.4	18.4	22.9	20.5
c1908	38.2	17.3	36.4	49.4	17.7	36.5	25.2
c2670	52.2	39.4	45.4	49.4	39.4	41.3	17
c3540	52.3	35.8	29.8	49.4	35.5	34.6	20
c5315	50.0	35	27.2	49.4	35.9	28.5	15.4
c6288	25.8	4.6	9.8	49.4	4.4	9	34.1
c7552	49.2	34.5	35.2	49.4	34.1	36.4	18.2
array4 × 4	31.3	6.6	29.1	49.4	7.7	33.9	30.7
array8 × 8	29.8	8.7	17.3	49.4	9.7	16.8	28.1
bkung16	24.3	0	22.3	49.4	−0.5	22.3	26.3
bkung32	21.2	−2.2	17.7	49.4	−3.3	16.8	23
booth9 × 9	29.6	8.9	26.2	49.4	11.5	18.8	9.6
kogge16	23.9	0	22.3	49.4	1.3	23.6	26.3
kogge32	21.2	−2.7	16.7	49.4	−2.9	16.1	23.8
log16	26.6	3.9	32.5	49.4	4.1	31.1	21.2
log32	34.9	14.9	36.2	49.4	22.3	32.1	17.3
log64	36.7	17	40.4	49.4	24.9	36.4	18.1
pmult4 × 4	31.8	8.8	35	49.4	10	35.6	26.9
pmult8 × 8	28.7	4.5	25.8	49.4	5.9	19.8	9
pmult16 × 16	24.7	0.1	15.2	49.4	3.3	14.1	4.2
pmult32 × 32	21.8	-0.5	14.9	49.4	1.6	11.2	1.5
average	33.6	12.8	27.2	49.4	13.9	26.1	20.3

[a]'nominal' means $V_{dd} = 0.9$ V

doi: 10.1049/iet-cds.2012.0361

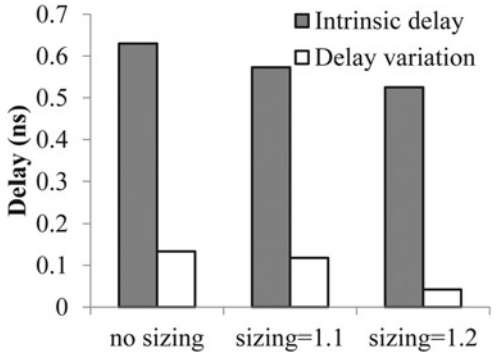

Fig. 11 *Gate sizing for an AND2X1 gate*

gate sizing and replacement technique on mitigating the RTN effect.

Assuming that the area of a gate j in (9) becomes ρA_j ($\rho > 1$ is the sizing coefficient), according to (4), the RTN-induced delay of this gate becomes

$$d_j = \frac{K_j C_{L,j} V_{dd}}{\rho A_j \left(V_{dd} - V_{th0} - \left(S\Delta V_{th,j}/\rho\right)\right)^{\alpha}} \tag{20}$$

Thus, the delay will degrade by

$$\Delta d_j \simeq \left(\frac{1}{\rho} - 1 + \frac{\alpha S \Delta V_{th,j}}{\rho^2 \left(V_{dd} - V_{th0}\right)}\right) \times d_j \tag{21}$$

Compared with (10), sizing can mitigate the RTN-induced delay degradation. Meanwhile, the term $(1/\rho^2)$ indicates that the delay variation can be also reduced.

The gate sizing technology on an 'AND2X1' gate is shown in Fig. 11. The intrinsic delay is 0.63 ns when driving an 1 fF load capacitance. The delay varies from 0.63 to 0.763 ns

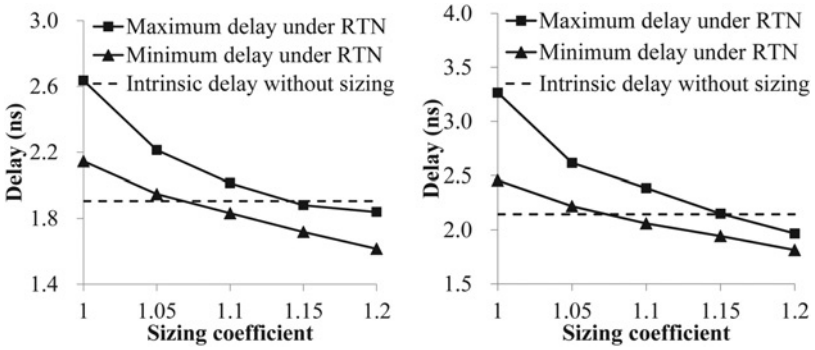

Fig. 12 *Gate sizing for c1355 and c3540, obtained by MC*

a c1355
b c3540

Table 4 Results of gate sizing, obtained by MC

Benchmark	'full'			'critical'		
	Δd_{max}, %	Δd_{var}, %	ΔA, %	Δd_{max}, %	Δd_{var}, %	ΔA, %
c432	−7.5	4.7	15.0	−7.4	4.6	8.0
c499	−1.2	9.4	15.0	−1.2	9.2	9.0
c880	−5.7	7.2	15.0	−5.4	7.9	3.2
c1355	−4.2	6.3	15.0	−1.3	9.3	6.2
c1908	−4.7	8.4	15.0	−4.4	8.6	7.4
c2670	1.9	13.7	15.0	1.0	12.3	3.7
c3540	1.0	11.2	15.0	0.7	11.1	6.6
c5315	0.4	10.4	15.0	−0.1	9.8	5.4
c6288	−7.6	2.2	15.0	−2.2	6.8	10.0
c7552	−0.2	11.9	15.0	0.3	12.1	5.2
array4 × 4	−7.4	5.8	15.0	−7.4	5.8	8.9
array8 × 8	−7.2	3.6	15.0	3.6	13.5	9.3
bkung16	−9.3	3.5	15.0	−9.6	3.1	4.0
bkung32	−10.3	1.9	15.0	−10.2	2.2	3.2
booth9 × 9	−7.7	5.2	15.0	8.3	17.3	3.1
kogge16	−9.3	3.5	15.0	−9.6	3.1	4.0
kogge32	−10.2	2.1	15.0	−10.1	2.2	3.3
log16	−9.6	3.9	15.0	−9.2	4.2	9.1
log32	−5.9	7.7	15.0	6.3	14.8	7.0
log64	−5.0	8.8	15.0	7.6	16.0	7.6
pmult4 × 4	−7.2	6.4	15.0	−7.1	6.2	7.6
pmult8 × 8	−8.0	4.9	15.0	1.7	13.8	2.1
pmult16 × 16	−8.6	3.3	15.0	−0.7	9.6	1.2
pmult32 × 32	−9.2	2.2	15.0	−3.0	5.3	0.4
average	−6.0	6.2	15.0	−2.5	8.7	5.7

without sizing. When $\rho = 1.1$, the delay varies from 0.573 to 0.691 ns; whereas for $\rho = 1.2$, the delay varies from 0.525 to 0.567 ns.

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 gates can be replaced by the corresponding larger gates in the library. Two replacement strategies are applied: 'full' replacement (replace all the gates) or 'critical' replacement (only replace the gates along the critical paths).

Fig. 12 shows the sizing results for c1355 and c3540, using the 'critical' replacement method. The intrinsic delay is still chosen as the design specification. It indicates that when $\rho = 1.15$, the maximum delay under RTN is almost below the specification line. Hence, $\rho = 1.15$ is chosen for the subsequent experiments.

The results of gate sizing for all the benchmarks are shown in Table 4, where ΔA is the area overhead. The results reveal that by using the 'full' replacement method, the maximum delay is on average 6% smaller than the design specification and the delay variation is 6.2%, which is much smaller than the results without sizing. By using the 'critical' replacement method, the maximum delay still satisfies the design specification and the area overhead is only on average 5.7%. Compared with V_{dd} tuning, gate sizing is much better: the efficiency is higher and the overhead is smaller.

7 Conclusions

This paper proposes a simulation framework to evaluate the RTN-induced temporal performance degradation and variation of digital circuits. Two fast evaluation methods with linear time complexity are proposed. The experimental results show that the average degradation and variation at 16 nm can be both >20%. Two design techniques, power supply tuning and gate sizing, are applied to mitigate the RTN effect and the simulation results show that gate sizing is better than power supply tuning.

The RTN-induced fluctuations are independent in all the devices, which causes very random performance distribution. Enough performance margin should be reserved to compensate the impact of RTN and design techniques, such as power supply tuning and gate sizing, should be investigated to mitigate the RTN effect. In addition, more efficient circuit-level and architectural-level techniques with less overheads should be investigated in future work.

8 Acknowledgments

This work was supported by the National Science and Technology Major Project (grant no. 2011ZX01035-001-001-002), National Natural Science Foundation of China (grant numbers 61028006, 61076035 and 61261160501) and Tsinghua University Initiative Scientific Research Programme.

9 References

1 Wang, Y., Luo, H., He, K., Luo, R., Yang, H., Xie, Y.: 'Temperature-aware NBTI modeling and the impact of standby leakage reduction techniques on circuit performance degradation', *IEEE Trans. Dependable Secur. Comput.*, 2011, **8**, (5), pp. 756–769

2 Chen, X., Wang, Y., Cao, Y., Ma, Y., Yang, H.: 'Variation-aware supply voltage assignment for simultaneous power and aging optimization', *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, 2012, **20**, (11), pp. 2143–2147

3 Luo, H., Chen, X., Velamala, J., *et al.*: 'Circuit-level delay modeling considering both TDDB and NBTI'. Int. Symp. Quality Electronic Design (ISQED), March 2011, pp. 1–8

4 Luo, H., Wang, Y., Cao, Y., Xie, Y., Ma, Y., Yang, H.: 'Temporal performance degradation under RTN: evaluation and mitigation for nanoscale circuits'. IEEE Computer Society Annual Symp. VLSI (ISVLSI), August 2012, pp. 183–188

5 Tega, N., Miki, H., Ren, Z., *et al.*: 'Reduction of random telegraph noise in high-*k*/metal-gate stacks for 22 nm generation FETs'. IEEE Int. Electron Devices Meeting (IEDM), December 2009, pp. 1–4

6 Tega, N., Miki, H., Pagette, F., *et al.*: 'Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm'. Symp. VLSI Technology, June 2009, pp. 50–51

7 Campbell, J.P., Yu, L.C., Cheung, K.P., *et al.*: 'Large random telegraph noise in sub-threshold operation of nano-scale nMOSFETs'. IEEE Int. Conf. IC Design and Technology (ICICDT), May 2009, pp. 17–20

8 Lee, A., Brown, A.R., Asenov, A., Roy, S.: 'Random telegraph signal noise simulation of decanano MOSFETs subject to atomic scale structure variation', *Superlattices Microstruct.*, 2003, **34**, (3–6), pp. 293–300

9 Campbell, J.P., Qin, J., Cheung, K.P., *et al.*: 'The origins of random telegraph noise in highly scaled SiON nMOSFETs'. IEEE Int. Integrated Reliability Workshop (IRW), October 2008, pp. 1–16

10 Campbell, J.P., Qin, J., Cheung, K.P., *et al.*: 'Random telegraph noise in highly scaled nMOSFETs'. IEEE Int. Reliability Physics Symp. (IRPS), April 2009, pp. 382–388

11 Ghetti, A., Compagnoni, C.M., Spinelli, A.S., Visconti, A.: 'Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer flash memories', *IEEE Trans. Electron Devices*, 2009, **56**, (8), pp. 1746–1752

12 Tega, N., Miki, H., Yamaoka, M., *et al.*: 'Impact of threshold voltage fluctuation due to random telegraph noise on scaled-down SRAM'. IEEE Int. Reliability Physics Symp. (IRPS), May 2008, pp. 541–546

13 Toh, S.O., Tsukamoto, Y., Guo, Z., Jones, L., Liu, T.K., Nikolic, B.: 'Impact of random telegraph signals on Vmin in 45 nm SRAM'. IEEE Int. Electron Devices Meeting (IEDM), December 2009, pp. 1–4

14 Tanizawa, M., Ohbayashi, S., Okagaki, T., *et al.*: 'Application of a statistical compact model for random telegraph noise to scaled-SRAM Vmin analysis'. Symp. VLSI Technology (VLSIT), June 2010, pp. 95–96

15 Aadithya, K.V., Demir, A., Venugopalan, S., Roychowdhury, J.: 'SAMURAI: an accurate method for modelling and simulating non-stationary random telegraph noise in SRAMs'. Design, Automation Test in Europe Conf. Exhibition (DATE), March 2011, pp. 1–6

16 Aadithya, K.V., Venogopalan, S., Demir, A., Roychowdhury, J.: 'MUSTARD: a coupled, stochastic/deterministic, discrete/continuous technique for predicting the impact of random telegraph noise on SRAMs and DRAMs'. ACM/EDAC/IEEE Design Automation Conf. (DAC), June 2011, pp. 292–297

17 Leyris, C., Pilorget, S., Marin, M., Minondo, M., Jaouen, H.: 'Random telegraph signal noise SPICE modeling for circuit simulators'. European Solid State Device Research Conf. (ESSDERC), September 2007, pp. 187–190

18 Tang, T.B., Murray, A.F.: 'Integrating RTS noise into circuit analysis'. IEEE Int. Symp. Circuits and Systems (ISCAS), May 2009, pp. 585–588

19 Ye, Y., Wang, C.-C., Cao, Y.: 'Simulation of random telegraph noise with 2-stage equivalent circuit'. IEEE/ACM Int. Conf. Computer-Aided Design (ICCAD), November 2010, pp. 709–713

20 Ito, K., Matsumoto, T., Nishizawa, S., Sunagawa, H., Kobayashi, K., Onodera, H.: 'Modeling of random telegraph noise under circuit operation – simulation and measurement of RTN-induced delay fluctuation'. Int. Symp. Quality Electronic Design (ISQED), March 2011, pp. 1–6

21 Lee, S., Cho, H.-J., Son, Y., Lee, D.S., Shin, H.: 'Characterization of oxide traps leading to RTN in high-k and metal gate MOSFETs'. IEEE Int. Electron Devices Meeting (IEDM), December 2009, pp. 1–4

22 Nagumo, T., Takeuchi, K., Hase, T., Hayashi, Y.: 'Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps'. IEEE Int. Electron Devices Meeting (IEDM), December 2010, pp. 28.3.1–28.3.4

23 Ghetti, A., Compagnoni, C.M., Biancardi, F., *et al.*: 'Scaling trends for random telegraph noise in deca-nanometer flash memories'. IEEE Int. Electron Devices Meeting (IEDM), December 2008, pp. 1–4

24 http://ptm.asu.edu/ (accessed November 2012)

25 Billingsley, P.: 'Probability and measure' (Wiley Press, 1979, 2nd edn., 1986, 3rd edn., 1995)

26 Clark, C.E.: 'The greatest of a finite set of random variables', *Oper. Res.*, 1961, **9**, (2), pp. 145–162

doi: 10.1049/iet-cds.2012.0361