Measurement and Simulation Methods for Assessing SRAM Reliability Against Random Telegraph Noise

Kiyoshi Takeuchi

1 Introduction

If we measure the drain-to-source current of a metal–insulator–semiconductor (MIS) field-effect transistor (FET) applying dc bias, excess low-frequency noise, called 1/*f* noise, is usually observed, whose power spectral density is approximately inversely proportional to frequency. A widely accepted explanation of 1/*f* noise today is that it is caused by superposition of multiple random telegraph noise (RTN) signals [1, 2]. If a trap site in the gate insulator of a MISFET captures or emits an electron, the conductance of the MISFET will change, depending on the charge state of the trap. An example of single-trap RTN waveform is shown in Fig. 1a. The signal can be characterized by three parameters: amplitude A, and time constants τ_0 and τ_1, where τ_0 and τ_1 are defined here as the mean time of stay in states 0 and 1, respectively. The power spectral density of the signal is given by [3]

$$S(f) = \frac{4\mathrm{A}^2}{(\tau_0 + \tau_1)\left\{1/\tau^2 + (2\pi f)^2\right\}}, \quad \frac{1}{\tau} \equiv \frac{1}{\tau_0} + \frac{1}{\tau_1}. \tag{1}$$

By adding many such signals, whose τ value is distributed uniformly per $\log\tau$ (i.e., the expected number of traps in the range of 1–10 Hz is the same as 10–100 Hz) for many orders, a 1/*f* power spectrum is generated. Therefore, when the channel length *L* and width *W* of a MISFET are large, in which many traps are expected to exist, 1/*f* noise will be observed. However, as MISFETs are scaled down, it becomes more and more likely that a limited number of traps exist in a FET. In such situations, RTN signals become apparent (Fig. 1).

K. Takeuchi (✉)
Institute of Industrial Science, The University of Tokyo, Tokyo, Japan
e-mail: takeuchi@nano.iis.u-tokyo.ac.jp

T. Grasser (ed.), *Noise in Nanoscale Semiconductor Devices*,
https://doi.org/10.1007/978-3-030-37500-3_8

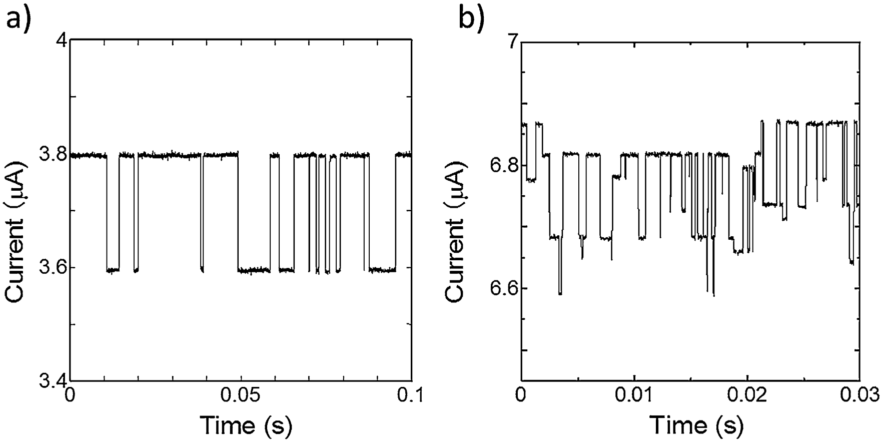


Fig. 1 Random telegraph noise waveforms for single trap (**a**) and multiple (three) traps (**b**) cases

While the expected number of traps per transistor will decrease by shrinking the channel area LW, on the contrary, the impact of RTN on transistor characteristics becomes more serious [4–6]. This is because the sensitivity of MISFET characteristics on a single charge carrier is increased in proportion to q/C_{OX}, where C_{OX} is the MISFET gate capacitance. As a result, recently, low-frequency noise has become a concern, not only for analog circuits but also for digital circuits using miniaturized transistors. RTN was first reported as a practical reliability problem for Flash memories [7–9], in which C_{OX} of the memory cell transistors is smaller than logic FETs, owing to the thicker gate dielectric and smaller LW used. Today, in state-of-the-art complementary metal–oxide–semiconductor (CMOS) integrated circuit technologies, L and W on the order of 10 nm are used, and hence C_{OX} has become extremely small, even for logic transistors. To realize reliable and error-free digital circuits using such advanced technologies, RTN must be taken into account [10–15].

To deal with RTN, a statistical approach is indispensable, since there is large inter-device "variability of noise." Depending on the number of traps, as well as the characteristics of each of the traps, one transistor exhibits statistically different noise waveforms from others. In such situations, effects of noise on MISFET circuits cannot be judged only from the averaged $1/f$ noise characteristics, which are obtained by measuring large transistors. Therefore, the author and coworkers proposed a set of methods for assessing "RTN reliability" of static random access memories (SRAMs) [16, 17], where "accelerated" SRAM measurements and fast Monte Carlo simulations are combined. SRAM cells are considered to be most vulnerable to RTN in logic integrated circuits, since the cell transistors are smaller than other logic transistors. In this chapter, these methods will be reviewed and given more thorough descriptions. Works in [6] are also briefly reviewed in the next section, which will serve as an additional introduction and motivation part of this chapter.

2 Individual FET Characterization

A simple and straightforward way of statistically characterizing RTN would be to measure noise waveforms for many individual transistors [5, 18–21]. For example, in [19], hundreds of n-channel and p-channel MISFETs were measured, applying dc bias on all the terminals. Then, the FETs exhibiting single-trap RTN signals were selected, and the parameters A, τ_0, and τ_1 for each RTN signal were extracted from the measured current vs. time waveforms, varying the dc bias conditions. By focusing on only such single-trap FETs, the determination of the trap parameters was simplified. This was possible, since single-trap RTN signals were found in a sufficiently large portion (20–30%) of all the FETs, thanks to the miniaturization. From such measurements, it was possible to discuss in detail the statistical distributions of the trap time constants, energy levels, and even vertical locations in the gate oxide. Correlation between the parameters was also easily examined. In [20, 21], transient measurements were also used, which is effective for covering a wider trap energy range than dc biasing. An advantage of this approach, i.e., simply measuring many individual transistors, is that it can be combined with almost any measurement and analysis methods of any sophistication. Detailed information on each trap can be obtained, e.g., by manipulating bias conditions or changing temperature. However, there is a disadvantage that single transistor measurements are usually time-consuming, and therefore it is difficult to measure a sufficient number of devices necessary for revealing detailed statistical distributions of the trap parameters in the low quantile range.

To alleviate this problem, addressable transistor arrays were used in [6]. Today, since variability in scaled down transistors is significant, characterization of variability in threshold voltage, drain current, etc. is indispensable, and therefore addressable transistor arrays are commonly used for this purpose [22–25]. By using addressable arrays, a large number of devices to be tested can be accommodated in a small area, by sharing area consuming pads by many devices. If appropriately designed, such arrays can be reused for characterizing RTN for a large number of FETs. The arrays used in [6] integrate 1024 identically designed MISFETs, where the gate and drain terminals of one selected FET out of the 1024 FETs, specified by an address value, are connected to external pads via FET switches (Fig. 2). Kelvin connections are used for the drain terminals (not shown) to reduce any voltage drop by parasitic resistance. Common source and body pads are also provided. Using this configuration, dc characterization including low-frequency noise measurements of each FET, one-by-one, using standard parametric testers is possible. However, if these 1024 FETs are measured sequentially, the measurement time per device will be essentially the same with single FET measurement cases, though the number of

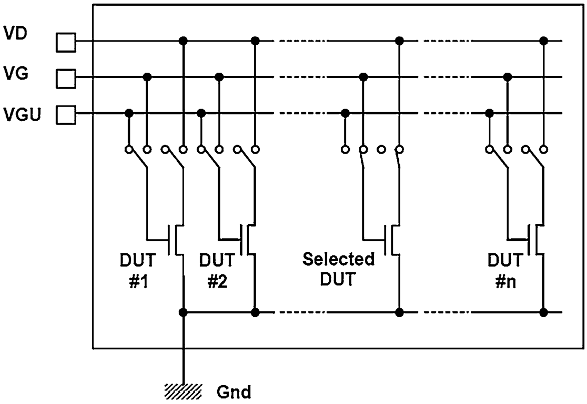


Fig. 2 Addressable transistor array configuration

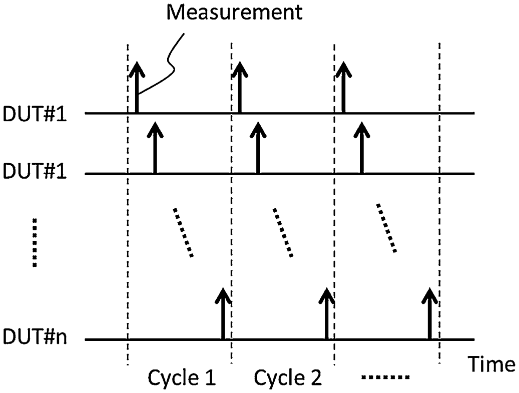


Fig. 3 Quasi-parallel measurement scheme

available FETs is increased. Therefore, to measure more FETs, while avoiding the unacceptable increase of measurement time, a quasi-parallel measurement scheme, as shown in Fig. 3, was adopted. First, only DUT #1 (DUT: device under test) is connected to the pads VD and VG, and its drain current is measured. Then, the connection is switched to DUT #2, and then to DUT #3, and so on, and finally to DUT #1024. After all the FETs are measured, the address goes back to #1, and the procedure is repeated for a desired number of times. Constant voltage is applied to all the pads VD, VG, and VGU throughout the measurements. By applying the same voltage on VG and VGU, transient trapping/de-trapping by DUT switching was avoided.

One significant advantage of parallel measurement is its ability to perform long time measurements for a large number of FETs, and hence information of slow traps can be obtained. Figure 4 shows examples of slow RTN signals found by the quasi-parallel scheme. It can be confirmed that traps with time constants on the order of hour actually exist. As will be discussed later, such slow traps are problematic in that their existence is not easily detected by short-time screening tests applicable to production. Drawback of the quasi-parallel scheme, compared with true parallel

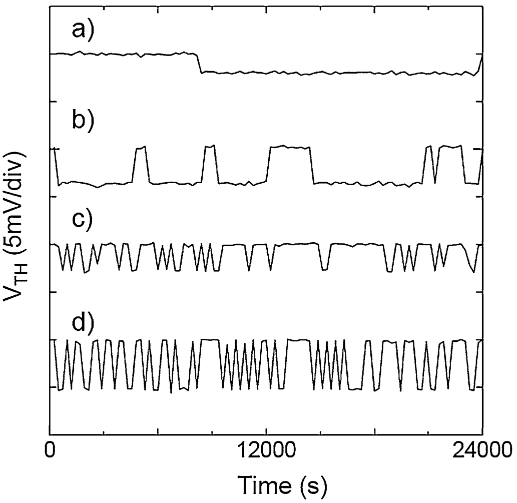


Fig. 4 Single-trap waveform examples obtained by quasi-parallel measurements. Waveforms (**a**) to (**d**) were taken from four different FETs in the same array. © 2009 the Japan Society of Applied Physics (JSAP). Reprinted, with permission, from [6]

measurements, is that the sampling interval becomes very long (at least 1024 times longer in this particular example). Because of the sparse sampling, obtained current vs. time data do not fully track true current vs. time RTN waveforms, unless the RTN time constants are much larger than the long sampling interval. Even if the measured current values for two consecutive samples are equal, it does not guarantee that the current was constant between the sampling events; it is possible that the current has travelled to a different value, and came back to the original one. In spite of this, if the measurements are repeated for many cycles, fast RTN signals whose time constants are much shorter than the interval can be detected, on a condition that the signal satisfies the following conditions: (1) Both f_0 and f_1 are sufficiently large compared to 1/N, where N is the total number of current sampling per FET,

$$f_0 \equiv \frac{\tau_0}{\tau_0 + \tau_1}, \; f_1 \equiv 1 - f_0 = \frac{\tau_1}{\tau_0 + \tau_1}, \tag{2}$$

and (2) both τ_0 and τ_1 are long enough compared to the time resolution of the source measure units (SMUs). From the data thus obtained, statistical distributions of trap number and amplitude could be efficiently determined, without spending the cost of true parallel measurements using 1024 sets of SMUs.

The measured source-to-drain current (I_{DS}) was first translated into an effective threshold voltage (V_{TH}) defined as

$$V_{TH} \equiv V_{GS} - I_{DS}/g_m, \tag{3}$$

where V_{GS} is the constant gate-to-source voltage, and g_m is the transconductance. The g_m value used for each FET was determined by individually measuring current–voltage characteristics of the same FET, to reduce the effects of g_m variability. Then, from the N sampled V_{TH} values per FET, the number of traps in the FET

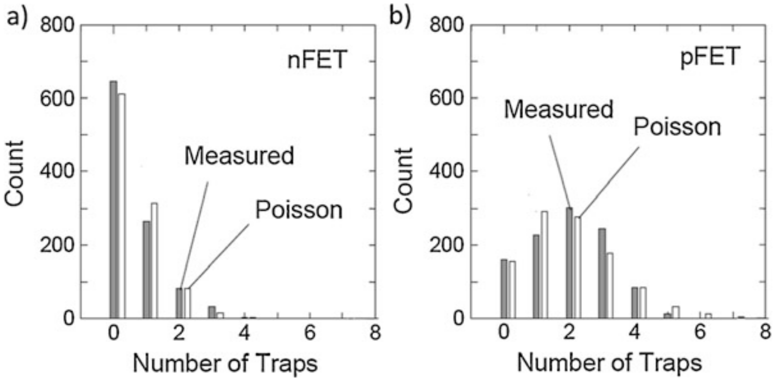


Fig. 5 Trap number distributions obtained by using addressable transistor array for nFETs (**a**) and pFETs (**b**). Poisson distributions whose mean values are equal to the measured values (0.52 for the nFETs, 1.90 for the pFETs) are also shown for comparison. © 2009 JSAP. Reprinted, with permission, from [6]

was determined using a simple algorithm: (1) Sort the N V_{TH} values for each FET, (2) find discontinuity of sorted V_{TH} values which is larger than a certain value (e.g., 0.5 mV), and (3) determine the number of traps n as

$$n = \mathrm{ceil}\left(\log_2 (m+1)\right), \tag{4}$$

where m is the number of discontinuities. Figure 5 shows the number distributions of detected traps for both n-channel and p-channel FETs, which were SRAM cell transistors fabricated by a conventional poly-Si/SiON gate stack technology. More traps were found in pFETs than nFETs in this example. Poisson distributions, whose expected values are set equal to the measured sample means, are also shown for comparison. Figure 5 shows that it is reasonable to assume Poisson distributions for the trap number in a FET. The slight disagreement could be attributed to the limited number of FETs measured and incompleteness of the simple number determination algorithm.

Next, RTN signal amplitudes were determined. To do this, similar to [19], only those FETs containing one trap were selected and used, to unambiguously determine the amplitude associated with a single trap. Figure 6 shows cumulative distributions of the amplitude of effective threshold voltage shift ΔV_{TH} (normalized by channel area LW) thus determined. It was found that the distributions can be approximated by exponential distributions, shown by the straight lines, which were also found in Flash memory cells [9]. For plotting the data in Fig. 6, care was taken to account for the fact that traps with too small amplitudes cannot be detected. Note that, in Fig. 6, there is a very small gap between $\Delta V_{\mathrm{TH}} = 0$ and the lowest ΔV_{TH} plot. Since the number of the traps below this detection limit is unknown, that number was used as a fitting parameter. Denoting the number of detected traps n_1, and the number of undetected ones n_2, accounting for n_2 results in shifting all the plots

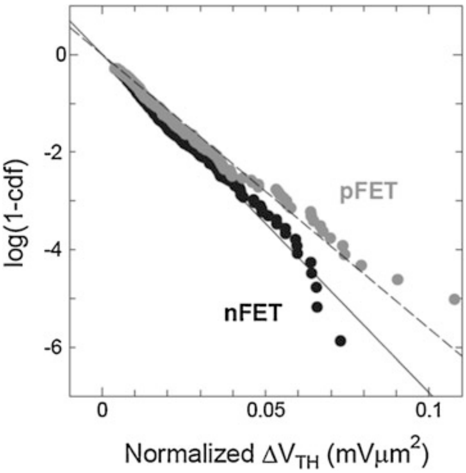


Fig. 6 Cumulative distributions of RTN amplitude obtained by using an addressable transistor array for n-channel FETs (black marks) and p-channel FETs (gray marks). Amplitude is normalized by transistor channel area. “Log” here is the natural logarithm. © 2009 JSAP. Reprinted, with permission, from [6]

(quantile values) in Fig. 6 downward by a constant ratio $n_1/(n_1 + n_2)$. The fitting parameter n_2 was determined, such that each straight regression line crosses the vertical axis $\Delta V_{\mathrm{TH}} = 0$ at cdf $= 0$, where cdf is cumulative distribution function of ΔV_{TH}. In this way, two fitting parameters were used for determining the straight lines in Fig. 6, though exponential distribution itself contains only one parameter. Similar fit could be obtained by using naturally two parameter distributions, such as lognormal distributions [5]. However, assuming nonzero n_2 seems to be more adequate, since it can decouple the effect of the measurement limitation.

So far, methods for collecting statistical information of RTN by measuring individual transistors have been discussed. It was shown that the measurement efficiency can be improved by using addressable transistor arrays. However, a question arises. Is it practically possible to fully understand or describe RTN phenomena by straightforwardly accumulating such statistical information (number, amplitude, time constants, as well as their voltage and temperature dependence) of the traps? To obtain such information, the problem of limited measurement window must be considered. If we perform current sampling measurements with an interval T and sampling count N, only those traps whose time constants are sufficiently larger than T, and sufficiently smaller than NT, can be detected. On the other hand, to fully predict product reliability, it would be necessary to know about those traps, whose time constants are between around the operation clock cycle (e.g., 1 ns) and product lifetime (e.g., 10 years). To straightforwardly achieve this, measurements of many devices (e.g., 1000) must be continued for more than 10 years, which is practically impossible. Even if all such information is available, there is another problem. How can we predict product failure probability? It would be possible to simulate product failures using Monte Carlo circuit simulations [26–28], using the perfect RTN parameter sets. However, it is not realistic to repeat transient simulations of 10-year duration for a large number of devices. Therefore, some additional methods must be provided that can link individual trap characterization and product reliability.

Various proposals, which will be useful for achieving this goal, are already made. Transient measurements [29, 30] originally used in the field of bias temperature instability are known to be very effective for enlarging the measurement windows. New methods for long-term reliability simulations are also reported [31, 32]. The methods in [16, 17] were also intended for solving these problems.

3 Accelerated SRAM Test

As for the first problem mentioned above, the same also applies to other phenomena concerning long-term reliability, such as time-dependent dielectric breakdown (TDDB) and bias temperature instability (BTI). To determine TDDB lifetime, for example, a straightforward way would be to measure several devices for more than 10 years, under the normal operation conditions. However, practically this is not possible. Therefore, a common practice is to predict lifetime by combining accelerated tests and lifetime extrapolation [33, 34]. For example, time to dielectric breakdown (TBD) is measured by applying higher than normal voltage, to make TBD short enough to measure. By obtaining TBD values for several different such accelerated voltage conditions, TBD under normal operation voltage is estimated by extrapolation. A similar method would also be required for RTN. As for the second problem, it is considered that directly measuring circuit failures caused by RTN would be a good solution. As already mentioned, a circuit that will be most easily affected by RTN (or any random variability) in logic integrated circuits is static random access memory (SRAM). The reasons are that an SRAM cell uses smaller transistors than logic circuits, and that it is essentially an analog circuit relying on a subtle balance between the transistors constituting the memory cells. Based on these considerations, it was proposed to directly measure SRAM failures caused by RTN, by applying accelerated bias conditions [16, 17]. The work will be reviewed in the following.

Figure 7 shows a typical SRAM cell, consisting of six transistors. Transistors p_1 and d_1 form a first CMOS inverter, and p_2 and d_2 form a second. The two CMOS inverters are cross-coupled, and constitute a bistable latch. Transistors a_1 and a_2 serve as pass gates to connect the internal nodes n_1 and n_2 to the bit lines BL and BL$'$. Usually, an SRAM cell is disconnected from the bit lines, by turning off the pass gates. In this situation (retention state), the cell is very stable, owing to the near-ideal CMOS inverter transfer curves (Fig. 8a). The high node (either n_1 or n_2) voltage (V_1 or V_2) is close to the power supply voltage V_{CC}, and the low (the other) node voltage is close to zero. However, when the cell content needs to be read out, the pass gates are turned on by raising the word line (WL) voltage, while setting the voltage of the bit lines equal to V_{CC}. This deforms the inverter characteristics as shown in Fig. 8b. The low node voltage is pulled up by the bit line voltage. In this situation (read disturb state), the cell becomes less stable. SRAM cell failure due to RTN will occur, if any, almost certainly during this read disturb state. A static noise margin (SNM) [35, 36] is defined as the edge length of a square that nests

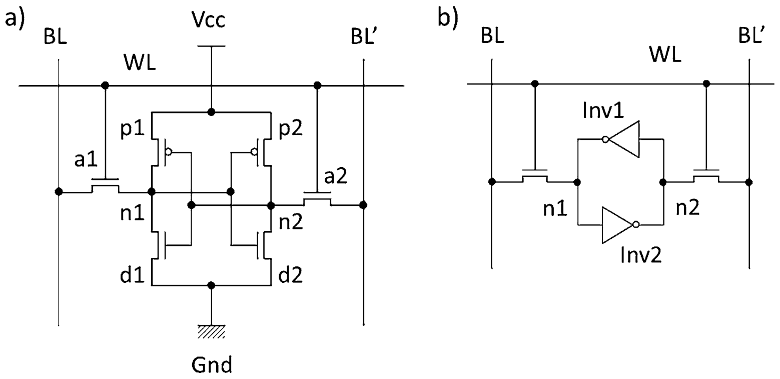


Fig. 7 Six-transistor CMOS SRAM cell (**a**) and its equivalent circuit (**b**)

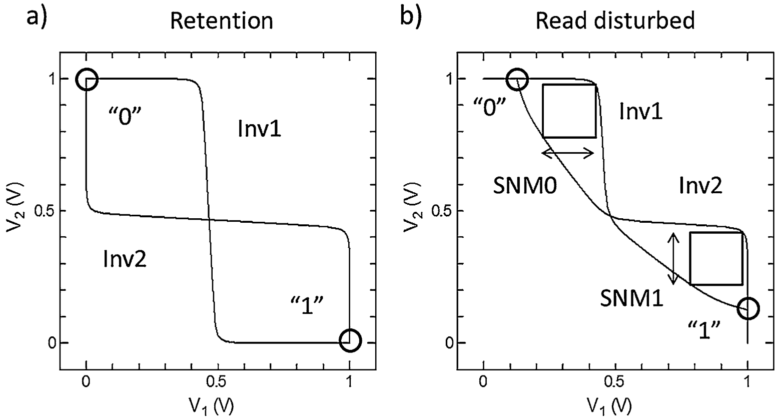


Fig. 8 Butterfly curves for retention (**a**) and read disturbed (**b**) states. V_1 and V_2 are the voltage at nodes n_1 and n_2, respectively. Circles show stable crossing points of two inverter transfer curves, which correspond to memory states "0" and "1," respectively. If SNM1 < 0, crossing point "1" disappears, and memory "1" is lost upon reading

in the "butterfly curves" in this read disturb state (see Fig. 8b). This definition can be modified to allow negative SNM, by defining SNM as the maximum distance between the two transfer curves (divided by the square root of two for compatibility with the original definition). A failure occurs if SNM becomes negative.

Figure 9 shows the concept of SRAM-accelerated test. In recent scaled SRAMs, there is large transistor variability. Usually, the threshold voltage of SRAM cell transistors is normally distributed, with a standard deviation of around a few tens of millivolts. As a result, SNM is also nearly normally distributed, because the nonlinearity between SNM and transistor threshold voltage is weak. To avoid yield loss due to the variability, SRAM cells are designed such that the mean SNM is larger than at least around six times the standard deviation σ (six sigma) of SNM. In this situation, SRAM cells whose SNM is small enough to be diminished by RTN is

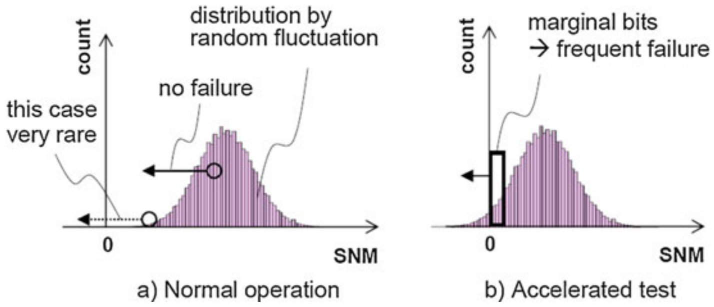
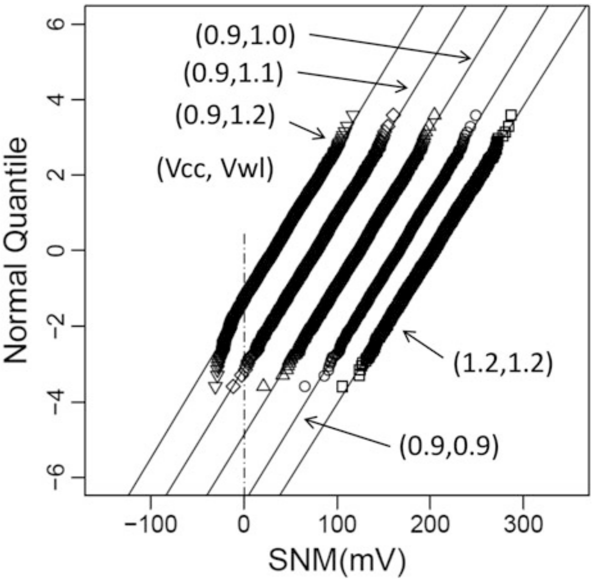


Fig. 9 Concept of accelerated SRAM test. Histogram of SNM in normal operation (**a**) is shifted to (**b**) by adjusting bias conditions. © 2010 IEEE. Reprinted, with permission, from [16]

Fig. 10 Quantile–quantile plots of SNM distributions for various combinations of cell power supply voltage (V_{CC}) and word line voltage (V_{WL}), obtained by Monte Carlo circuit simulations. Sixty-five nanometer technology transistor models were used. A straight line corresponds to a normal distribution. SNM distributions can be shifted keeping the same shape



extremely rare, since the amplitude of SNM change due to RTN (σ is on the order of mV) is small compared to the SNM variability range (σ is a few tens of mV). Hence, in normal operation of properly designed SRAMs, RTN failure can hardly be detected (Fig. 9a). Therefore, to observe many SRAM failure events by RTN experimentally, it was proposed to intentionally reduce SNM by applying a special bias to the cell (Fig. 9b). In this situation, there will be a large number of cells whose SNM is negative due to variability, and always fail. In addition, there are cells whose SNM is so small that RTN can easily cause their failure. As a result, SRAM cell failure events due to RTN can be frequently observed. Fortunately, almost ideal parallel shift of the SNM distribution as schematically shown in Fig. 9 is possible, by simply applying a voltage higher than the nominal value to the word lines, and/or setting V_{CC} to a lower than nominal value, as demonstrated in Fig. 10.

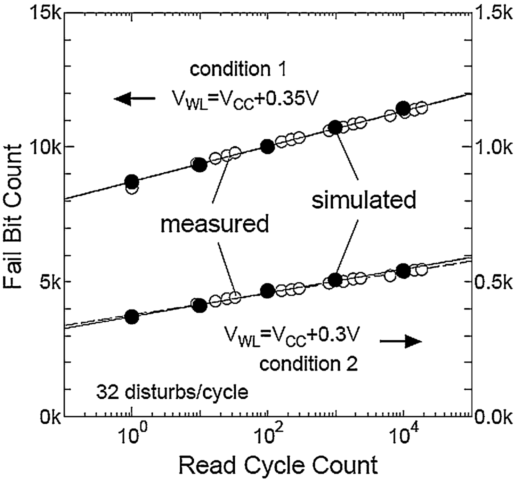


Fig. 11 Measured (open circles) and simulated (closed circles) fail bit count vs. read cycle count for two different bias (i.e., acceleration) conditions. Here, V_{WL} is constant. © 2011 JSAP. Reprinted, with permission, from [17]

For the accelerated SRAM measurements, a 512-k bit (32-k words × 16 bits) test SRAM cell array was used, which was capable of setting the high-state word line voltage (V_{WL}) and V_{CC} independently. A 40-nm bulk CMOS technology was used for the fabrication. First, zero was written to all the 512-k bits. Then, the SRAM array was repeatedly read out, one word (16 bits) at a time, by scanning the address starting from 0 to 32,767 ($= 32 \times 1024 - 1$). Note the similarity to the quasi-parallel measurements discussed in the previous section. If a cell operates normally, zero will be read out every time. However, if a failure occurs in a cell during a read disturb period, the bit state is flipped from 0 to 1. Once a failure occurs, the bit state will never return to 0, since state 1 should be more stable for that specific bit, and 1 is read out for all the readings after the failure. Figure 11 shows results of such an accelerated SRAM read test, where the number of 1 bits (i.e., fail bit count, FBC) vs. read cycle count is plotted. It can be seen that, even at the very first reading, many cells fail. This is because, as a result of the intentional margin reduction, there are a large number of cells whose SNM is negative owing to the variability. This initial FBC should stay constant, no matter how many times the reading process is applied, if there is no noise. However, the fact is that the FBC monotonically increased, as the number of read cycles increased. This shows that some noise source that causes bit failures certainly exists. It is suspected that the noise may be caused by some external source, such as that generated by the memory tester. Therefore, the same measurement for the same chip was repeated three times, to check the reproducibility. Figure 12 shows maps of those bits that failed at the first reading (gray marks), and those that did not fail at the first reading, but failed at the second or later readings (solid marks). Similarity of the locations of the solid marks between the three trials is apparent. The time-dependent failures tend to occur repeatedly at the same bits. Those failing bits seem to be randomly distributed over the chip area. Absence of any positional correlation suggests that the failures are not caused by any deficiency of the array circuitry. These results strongly suggest that

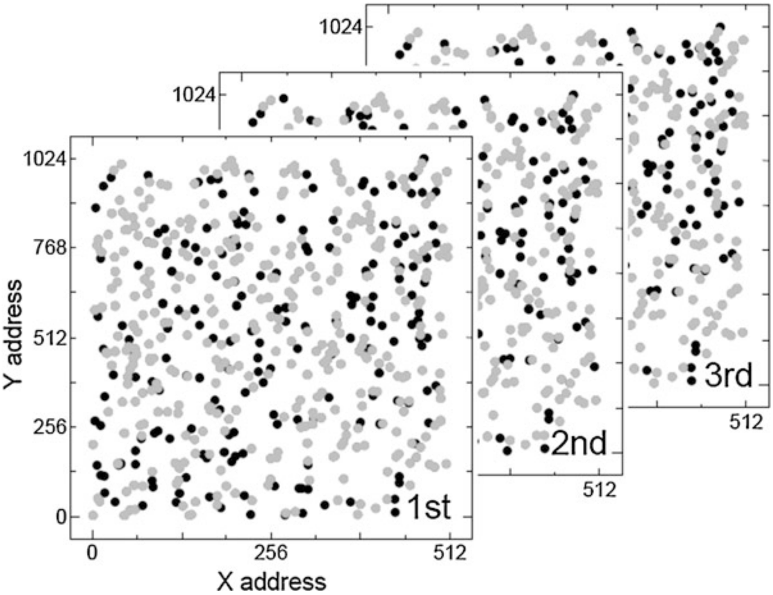


Fig. 12 Fail bit maps obtained by repeatedly measuring the same chip three times. Gray circles show always failing bits, and black ones show initially passed but finally failed bits. © 2010 IEEE. Reprinted, with permission, from [16]

the noise source exists in each SRAM cell, which should be RTN. Table 1 shows the statistics of the results. Among the bits that exhibited a time-dependent failure at least once, 60% consistently showed first-pass, last-fail behavior in all the three measurements. However, the repeatability is not perfect. This is natural considering that the failures caused by RTN will be statistical.

4 RTN Monte Carlo Simulation

By using SRAM arrays, direct observations of RTN failures seem to be possible. However, still, the link between the measurements and product reliability is missing. To solve this problem, realizing numerical simulations of the SRAM failures was considered, to understand the failure mechanism. Let us first discuss what is happening in a cell during the accelerated SRAM measurements. As already pointed out, a cell stays in the retention state for almost all the time, in which the cell is very stable, and no failure is expected to occur. However, trapping and de-trapping do occur in the cell transistors, and their threshold voltage (V_{TH}) will change over time. Since the applied bias does not change in the retention state, each FET is in a similar situation as during dc bias RTN measurements. This situation is occasionally interrupted by the read operations, which will cause some transient

Table 1 Statistics of bit failures for three repeated accelerated SRAM tests

Case	Bit count	Ratio (%)
Bias condition 1		
0–1	28	12.33
0–2	15	6.61
0–3	141	62.11
1–3	23	10.13
2–3	20	8.81
Subtotal	227	100.00
3–3	344	
Total	571	
Bias condition 2		
0–1	328	8.93
0–2	248	6.76
0–3	2252	61.35
1–3	461	12.56
2–3	381	10.38
1–2	1	0.03
Subtotal	3671	100.00
3–3	8086	
Total	11757	

Case *x*–*y* means that the bit failed *x* out of three times at the first reading, and *y* times at the last reading. 3–3 means the bit always failed, and 0–3 means the bit always passed at the first reading, but always failed at the last reading. © 2010 IEEE. Reprinted, with permission, from [16]

effects. However, since the duration of a read disturb state is very short (typically, on the order of 0.1–10 ns, for product SRAMs), it is assumed that the slow traps of interest for long-term reliability will be frozen during the reading operation, and the trap states of the cell established during the retention state will be sampled. Since SNM is a function of $V_{\mathrm{TH},i}$ ($i = 1,2,\ldots,6$), where $V_{\mathrm{TH},i}$ is the threshold voltage of the *i*th cell transistor, SNM changes over time depending on the trap states of all the cell transistors. The sampling result will be either 0 (SNM $\geq$ 0, pass) or 1 (SNM $<$ 0, fail).

Based on these considerations, a simple and fast Monte Carlo (MC) simulation method was proposed [16], which will be described below. The basic idea is that, given the parameters (amplitude and time constants) of all the traps in a cell, and assuming that the amplitudes are additive, it would be possible to estimate the worst-case amplitude that is likely to occur in N sampling events, by simple analytical considerations, even if N is a very large number (e.g., $N = 3 \times 10^{14}$ for 1-M samples/s × 10 years). First, let us introduce a linear approximation

$$\Delta \mathrm{SNM} = \sum_{i=1}^{6} a_i \Delta V_{\mathrm{TH},i}, \tag{5}$$

where ΔSNM is SNM deviation, and $\Delta V_{\mathrm{TH},i}$ is threshold voltage deviation of the ith transistor ($i = 1{,}2, \ldots, 6$) from the respective reference values. The set of coefficients a_i can be determined by iteratively searching for the most probable failure point (MPFP) using circuit simulations, applying SRAM cell variability design methods [37, 38]. We also assume that trap amplitudes are additive. That is,

$$\Delta V_{\mathrm{TH},i} = \sum_{j=1}^{n_i} \Delta V_{\mathrm{TH},i,j}, \tag{6}$$

where n_i is the number of traps in the ith transistor, $\Delta V_{\mathrm{TH},i,j}$ is the threshold voltage deviation caused by the jth trap in the ith transistor. By combining Eqs. (5) and (6), the SNM shift is now expressed as a simple linear combination of the V_{TH} shifts caused by all the traps. Here, $\Delta V_{\mathrm{TH},i,j}$ denotes the V_{TH} change from some reference value, and is not necessarily equal to the amplitude ($\equiv A_{i,j}$) of the trap (i, j), but can be either 0 (no trap state change) or $\pm A_{i,j}$ (low V_{TH} to high V_{TH} transition, and vice versa). For simulating the time-dependent FBC increase as in Fig. 11, it would be natural to select the initially sampled (read disturbed) state as the reference. If SNM + ΔSNM < 0 at the moment of any read disturb (including the first), the cell will fail. Note that SNM is also randomly distributed due to variability, and differs from cell to cell.

Noting that an SRAM read failure is determined only by the trap states at the moment of a read disturb, a simple time domain MC simulation would be, in principle, possible. The probability that a trap is in state 0 at time zero, and is found in state 0 (denoted P_{00}) or 1 (denoted P_{01}) after a time t is given by [3]

$$P_{00} = f_0 + f_1 \exp\left(-\frac{t}{\tau}\right), \; P_{01} = f_1\left(1 - \exp\left(-\frac{t}{\tau}\right)\right). \tag{7}$$

Using this formula, it is possible to track the state of a trap over time by MC simulations. That is, starting from an initial state, the state of a trap at the next reading is probabilistically determined using Eq. (7). Then, collecting all the states of the traps in a cell, Eqs. (5) and (6) are calculated to judge if SNM of the cell is negative or not. This can be repeated for a desired number of times to simulate the behavior as in Fig. 11. This discretized time domain simulation, based on a Markov chain model, is much more efficient than industry standard general purpose transient circuit simulations. However, even using this method, simulations of a sufficiently large number of cells (e.g., 1-M cells) for a sufficiently large number of readings (e.g., $N \sim 10^6$ for 1 s, not to mention $N \sim 3 \times 10^{14}$ for 10 years) is still computationally too demanding. It should also be noted that the number of traps in a cell to be simulated is much larger than the measured numbers shown in Fig. 5, since traps with small probability of transition, which did not fall within the measurement window, must be taken into account. Therefore, in [16, 17], an even simpler method of MC simulation was adopted. That is, the largest amplitude of a cell that is likely to occur in N reading events with a constant interval T was directly calculated from the trap parameters. The procedure is as shown below. In the following, the state of a trap, which corresponds to the better SNM (good state), will be denoted state 0, and the other (bad state) will be denoted state 1.

(S1) Assign an SNM value to cell, according to random variability of transistors.
(S2) Assign number of traps n to each transistor.
(S3) Assign amplitude A and two time constants τ_0 and τ_1 to each trap.
(S4) Randomly select initial state of all the traps, according to the ratio of τ_0 and τ_1.
(S5) Find a combination of traps which maximizes $-\Delta$SNM, while the probability of finding all the traps simultaneously in state 1 after *N* read disturb events is high enough.
(S6) If SNM + ΔSNM<0, judge that the cell will fail.

By repeating this for many cells, and for several *N* values, FBC vs. *N* relationship can be simulated. Note that the simulation time using this method does not depend on *N*, and is much more simplified than [27, 28]. Therefore, simulations of 10 years operation can be easily performed.

It can be noticed that, by following this procedure, a cell failure is deterministic. That is, a cell is always assumed to fail, if *N* exceeds a certain value. However, in reality, a failure of the same cell may occur much earlier or later. This simplification (i.e., use of an expected lifetime for a given set of trap parameters) could be justified by the following reasons. Firstly, the simulation will be performed for a large number of cells. Because of the simplification, simulating 1-M bits is easily accomplished. As a result, the stochastic difference between real and simulated time-to-failure will be averaged, and its effect on FBC will be reduced. This is supported by the fact that almost the same FBC vs. cycle count results are obtained by measuring the same array for three times (Fig. 11). Second, consideration of stochastically different waveforms for only a single bit introduces additional dimension of variability. At an early stage of study, removal of such complication would be desirable.

For the MC simulations, normal distributions for the random variability of SNM, Poisson distributions for trap number, and exponential distributions for trap amplitude are assumed, taking into account the results shown in Sect. 2. As for the time constants, it is assumed that the occupancy ratio follows Fermi–Dirac type relationship [1, 2].

$$\frac{\tau_1}{\tau_0+\tau_1}=\frac{1}{1+\exp\left(\frac{E-E_\mathrm{F}}{k_\mathrm{B}T}\right)}, \tag{8}$$

where the trap energy *E* is uniformly and symmetrically distributed around $E_{\mathrm{F}.}$ It is also assumed that the effective time constant τ defined as

$$\frac{1}{\tau}\equiv\frac{1}{\tau_0}+\frac{1}{\tau_1} \tag{9}$$

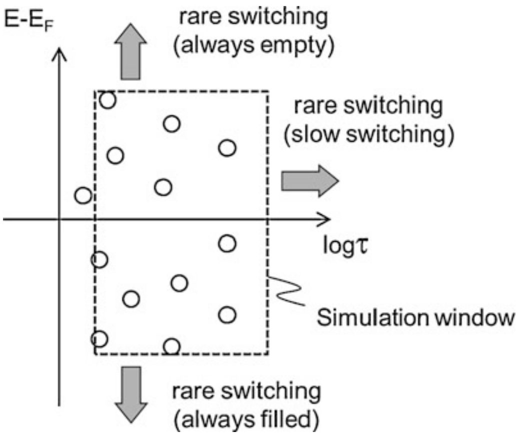


Fig. 13 Energy vs. effective time constant plane. Traps are assumed to be uniformly distributed in this plane. Traps in a sufficiently large rectangle are considered for simulation

is uniformly distributed per logτ, as already mentioned in connection with 1/*f* noise. Then, τ_0 and τ_1 are given as

$$\tau_0 = \tau \left\{1 + \exp\left(\frac{E - E_\mathrm{F}}{k_\mathrm{B}T}\right)\right\}, \ \tau_1 = \tau \left\{1 + \exp\left(\frac{E_\mathrm{F} - E}{k_\mathrm{B}T}\right)\right\}. \tag{10}$$

The uniform distributions were chosen here as crude initial assumption. The lower bound of τ was set equal to the duration of read disturb, whereas, the upper bound of τ was selected to be large enough compared with the time range to be simulated. The upper and lower bounds of E–E_F were selected to be sufficiently away from zero, so that the probability of switching outside these bounds is negligible. In other words, it was assumed that traps are uniformly distributed in a rectangle placed in the E vs. logτ plane (Fig. 13); the rectangle is selected to be large enough so that switching outside its top, bottom, and right edges is so rare and can be ignored. The mean number of traps λ and their mean amplitude Λ were selected by fitting to SRAM measurement results, using the values obtained by the addressable transistor array measurements as initial guess. Note that the number of traps to be fed to the simulator should not be equal to the actually measured number. Since the area of the rectangle in the E vs. logτ plane used for the simulation is much larger than that covered by measurement windows, the measured number should be multiplied by the ratio of the areas (simulated over measured). It was assumed that Λ, E, and τ, are independent, and that SNM variability and RTN are also independent, according to our measurement results, part of which are reported in [19].

FBC vs. cycle count simulated using the above described method is overlaid in Fig. 11. The MC simulation could reproduce the accelerated SRAM measurement results quite well, using reasonable trap parameters consistent with individual transistor measurements. For the comparison, the fact that a cell suffers from 32 read disturbs in a cycle was taken into account. This is because a word line is shared by 512 bits, while only 16 bits are read out at a time. Therefore, N is set equal to 32 times the cycle count. Because of this shared word line architecture of the array, which is customary for SRAMs, the first reading does not necessarily correspond to the first disturb, but may be any of the 1st to 32nd disturb. This will cause some overestimation of SNM variability, since any failure at the first reading is regarded as caused by SNM variability, not RTN. This inaccuracy is ignored here, expecting that its impact on long-term results will be small. Assuming that SNM is normally distributed, the mean SNM (μ) normalized by its standard deviation σ can be estimated from the initial fail bit ratio (FBR = FBC/512 k), using a relationship

$$1-\Phi\left(\frac{\mu}{\sigma}\right)=\mathrm{FBR},\ \Phi(x)\equiv\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x. \tag{11}$$

Since SNM of the memory cells in the array was not directly measurable, it was decided to assume that mean SNM estimated using Eq. (11) is the true mean SNM. This results in automatic alignment of the measured and simulated first FBC. The increase of FBC with the number of cycles depends on the trap parameters. It was found that good agreement between the measurement and simulation can be obtained without much effort of parameter fitting. The results for the two different bias conditions were reproduced by the same set of trap parameters without bias dependence, in spite of the 50-mV V_{CC} difference. It was also found that different sets of parameters can yield almost the same simulation results. This suggests that, for the modeling of RTN failures, a simplified noise model with reduced number of parameters, e.g., by assigning traps to only one or two transistors, could be used. In the following, to save simulation time, only the three most relevant transistors for read stability were taken into account, while keeping the parameters in the range reasonably consistent with individual FET measurements.

5 Reliability Extrapolation

With the aid of MC simulations, it is now possible to discuss, in more detail, what is happening during the accelerated SRAM measurements. Figure 14 shows the simulated probability density function (pdf) of worst RTN amplitude (i.e., $-\Delta$SNM determined in step (S5) of the simulation, or largest amplitude expected to be found in at least one of N read disturbs). As the number of disturbs N increases, it becomes more likely that a cell encounters a larger SNM amplitude, since the probability of simultaneous switching of more traps towards the bad direction increases. This is because slow traps with large τ or rarely switching traps with large $|E\text{–}E_{\mathrm{T}}|$ contributes to the amplitude. As a result, the mode (peak position) of the amplitude pdf increases with N, while its variance also increases. It was found that these simulated pdfs can be approximated by gamma distributions. This is expected, since exponential distributions are assumed for the single-trap amplitude, and that there is a close relationship between exponential and gamma distributions.

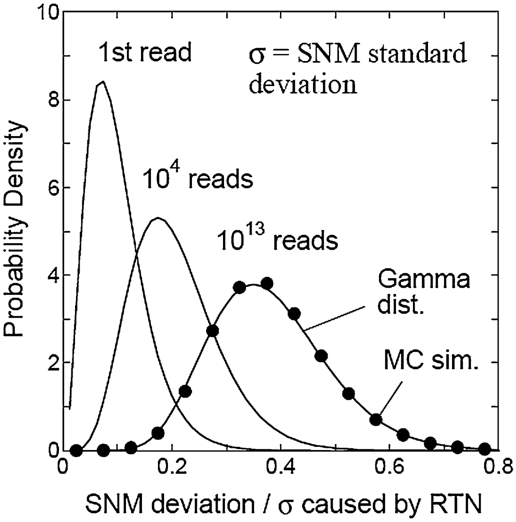


Fig. 14 Simulated worst RTN amplitude distributions, normalized by SNM standard deviation. © 2011 JSAP. Reprinted, with permission, from [17]

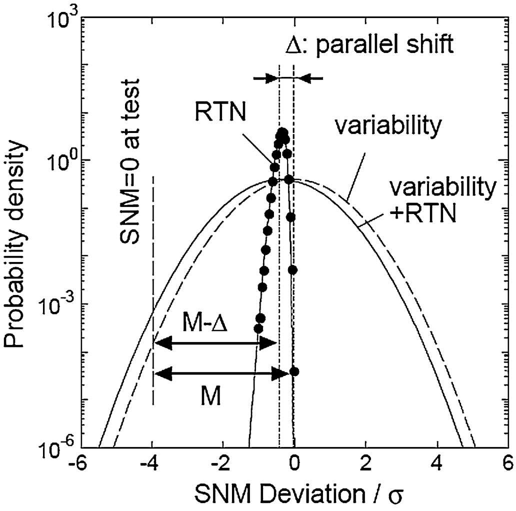


Fig. 15 Explanation of probability distribution shift of worst-case SNM, caused by coexistence of variability (variability without time dependence) and RTN (time-dependent variability). Original normal distribution of variability is shifted horizontally. © 2011 JSAP. Reprinted, with permission, from [17]

It should be noted that there is large SNM variability, which is not time-dependent. RTN amplitude calculated in Fig. 14 is still much smaller than the SNM variability. Let us now consider what happens when variability and RTN coexist. Assuming that SNM variability and RTN are independent, the pdf of the sum of the two can be calculated by convolution of the respective pdfs. Figure 15 shows an exemplary result of such convolution obtained numerically. It can be seen that the resulting summed amplitude distribution is almost equal to the original SNM normal distribution with a shifted mean. Although the RTN amplitude pdf has some width, since the distribution is much narrower than SNM variability, the original normal distribution is scarcely broadened.

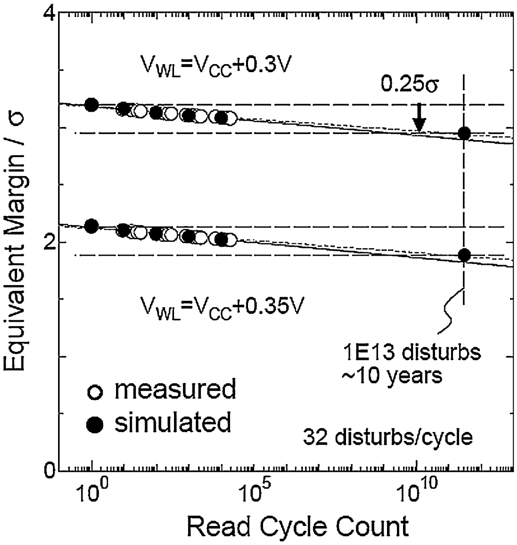


Fig. 16 Measured (open circles) and simulated (closed circles) effective margin vs. read cycle count. © 2011 JSAP. Reprinted, with permission, from [17]

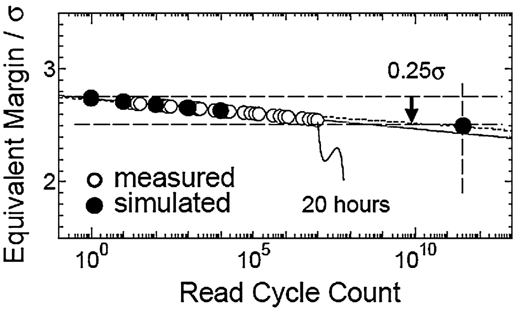


Fig. 17 Measured (open circles) and simulated (closed circles) effective margin vs. read cycle count for an extended number of cycles. © 2011 JSAP. Reprinted, with permission, from [17]

Considering that RTN effectively shifts the SNM distribution as the number of read disturbs N increases, we can define an effective SRAM margin M as a function of N. $M(N)$ normalized by the SNM standard deviation σ is calculated from the measured fail bit ratio (FBR) at the Nth read disturb using Eq. (11), where $M(N)$ is obtained as μ, which satisfies Eq. (11). By using this translation of FBR to $M(N)$, we can plot effective SRAM margin vs. read disturb counts, as shown in Fig. 16. It can be seen that the effective margin linearly decreases with $\log(N)$. The fact that the slope of the decrease does not depend on the initial FBC (i.e., the degree of acceleration) supports the assumption that the effective SNM distribution is shifted in parallel. It has to be pointed out here that the linear relationship is a result of assuming a uniform trap number distribution in the energy vs. $\log\tau$ plane. If the real distribution deviates from this assumption, a nonlinear dependence should be observed. To confirm the linearity for a larger N range, an accelerated measurement for extended number of cycles was performed. A linear relationship up to 10-M cycles (320-M disturbs) was confirmed (Fig. 17). These results suggest that effective

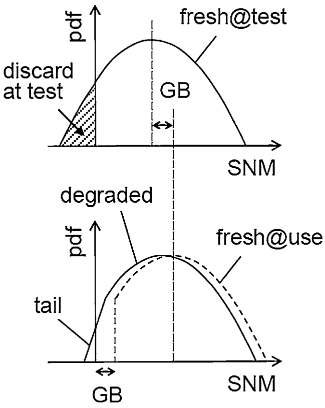


Fig. 18 Explanation of guard banding (GB). © 2011 JSAP. Reprinted, with permission, from [17]

margin loss due to RTN after 10 years operation can be empirically estimated by linearly extrapolating $M(N)$ vs. log(N). If a 1-M byte (= 8-M bits) SRAM in which each word line is shared by 512 bits, and all the word lines are evenly activated by a 500-MHz clock, a cell will be disturbed 30-k times per second (=500 MHz/8-M bits × 512 bits). Then, the margin loss ΔM after 10 years at $N = 1 \times 10^{13}$ (~30-k disturbs/s × 10 years) is estimated to be around 0.25σ, both by extrapolation and direct simulation. Though this value is much smaller than the usually assumed worst-case variability of 6σ, it is not negligible, since this degradation is time-dependent. While time-independent variability only degrades the product yield at shipment, the margin loss is relevant to reliability in the field.

Though the effective margin loss can now be estimated by the extrapolation technique, it has to be pointed out that the tolerance necessary for guaranteeing reliability would be somewhat larger than the ΔM value thus obtained, as discussed below. Consider a situation as shown in Fig. 18, where SRAMs, whose cells suffer from SNM variability, are tested before shipment, and a certain part of the products are discarded because of a failure caused by negative SNM. Since RTN should effectively shift the SNM variability distribution to the left during use, this testing condition should be stricter than the real use conditions. Then, the SNM distribution of an SRAM in use that passed the test would look like the dashed line in Fig. 18b. It can be assured that, if RTN is ignored, the worst SNM in the SRAM is greater than zero by at least a certain amount. This tolerance will be called a guard band (GB). If RTN simply shifts this truncated SNM distribution in parallel, similarly to Fig. 15, a GB width slightly larger than the extrapolated ΔM would suffice. However, this is too optimistic. Figure 19 shows a result of convoluting a truncated normal distribution (corresponding to variability) with a gamma distribution (corresponding to RTN). In addition to a parallel shift as in Fig. 15, a tail emerges at the left side, extending further than the parallel shift, owing to the nonzero width of the gamma distribution. The GB width should be determined such that the probability of failure during use does not exceed a certain acceptable limit (may be 1 ppm or 0.1%, depending on the applications). In doing so, existence of the tail portion

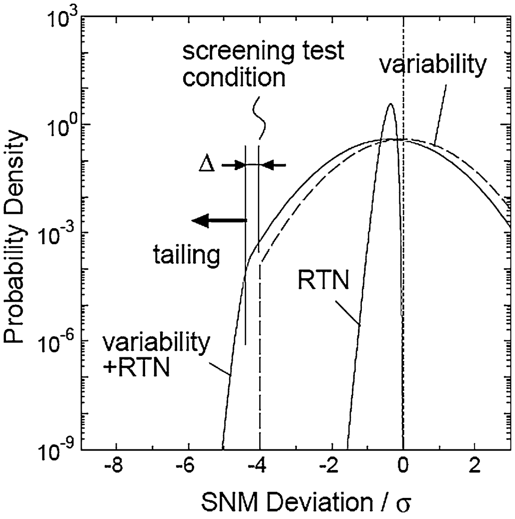


Fig. 19 Explanation of probability distribution change of worst-case SNM, caused by coexistence of variability and RTN, where the initial variability distribution is truncated by a screening test. In addition to a horizontal parallel shift, additional tailing caused by RTN amplitude variability emerges. The screening condition -4σ is too strict practically, and is chosen here for illustrative purpose. © 2011 JSAP. Reprinted, with permission, from [17]

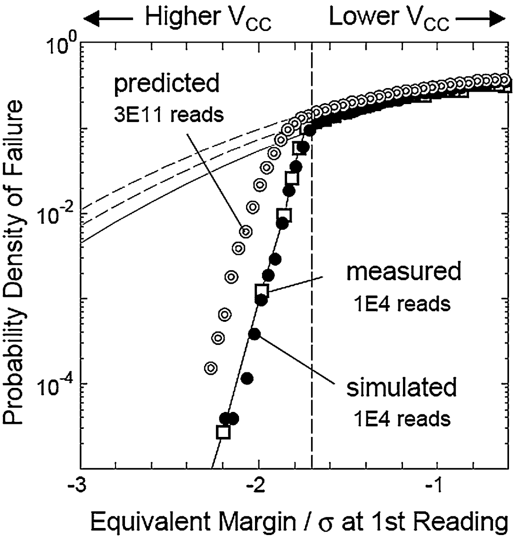


Fig. 20 Measured (open squares) and simulated (closed circles) distributions of worst-case SNM, corresponding to Fig. 19. Double circles show simulation results corresponding to 10-years operation ($=3 \times 10^{11}$ reads $= 1 \times 10^{13}$ disturbs). © 2011 JSAP. Reprinted, with permission, from [17]

should be taken into account. In the following, it will be shown that the shape of the distribution tail in Fig. 19 can be estimated, again using both the accelerated SRAM measurements and MC simulations, as shown in Fig. 20.

The measured plots in Fig. 20 were obtained as follows. The accelerated SRAM measurements of 10-k cycles were repeated for the same array, by changing the cell power supply voltage V_{CC} with a small interval from 0.88 V to 0.98 V, while keeping the word line voltage V_{WL} constant at a higher than nominal value (1.3 V). At these bias conditions, the mean SNM normalized by σ, or margin M (Fig. 15), is moved to around 1–2, depending on the V_{CC} value. Lower V_{CC} corresponds to lower M. For each condition, the address values of all the fail bits at the first and the last readings were recorded. Let us denote fail bit counts for the first and last cycles FBC_1 and FBC_2, respectively ($\mathrm{FBC}_1 < \mathrm{FBC}_2$). By translating FBC_1 into a SNM deviation using Eq. (11), M for a certain V_{CC} value can be estimated. That is, the μ/σ that satisfies Eq. (11) is an estimate of M for each V_{CC}. Since FBC_1 was measured for several V_{CC} values, it was also possible to approximately determine the pdf of FBC_1 (pdf_1) by calculating (ΔFBC1/512 k)/ΔM, where Δx means the difference of x between two adjacent V_{CC} conditions and 512 k is the total number of bits. By definition, pdf_1 vs. $-M$ plots should fall on a standard normal distribution $N(0,1)$. Similarly, pdf_2 can be obtained from FBC_2, which will fall on a shifted normal distribution $N(-\Delta M,1)$. Note that the negative sign accompanying M comes from the fact that in the evaluation procedure here, the shape of the pdf is obtained by changing M and counting bits whose margin deviation is smaller than $-M$ (i.e., SNM < 0). By increasing M, the pdf information at a more negative deviation $-M$ is accessed. To obtain the tail distribution, the measured data were further manipulated. First, one of the V_{CC} condition ($V_{\mathrm{CC}} = 0.92$ V, $M = 1.7\sigma$ denoted M_{REF}) was selected as a reference and a special fail bit count $\mathrm{FBC}_2{}'$ was defined. $\mathrm{FBC}_2{}'$ is the number of those bits that failed after 10-k cycles at the respective V_{CC} (or M) condition, but passed at the reference condition at the first reading. Since the entire fail bit maps were recorded, determination of $\mathrm{FBC}_2{}'$ is straightforward. Then, similar to other cases, $\mathrm{FBC}_2{}'$ was converted to $\mathrm{pdf}_2{}'$. The open marks in Fig. 20 show $\mathrm{pdf}_2{}'$ vs. $-M$ thus obtained. The vertical line shows the reference condition. This mimics a shipment test discussed earlier, though the condition is unrealistically strict such that many failure bits can be measured (i.e., this is an accelerated condition). On the left-hand side of the line, an exponentially decaying tail was obtained, which should correspond to the tail in Fig. 19. These tail bits passed a stricter test ($M = M_{\mathrm{REF}} = 1.7\sigma$) at the first reading, but failed under a looser condition ($M > 1.7\sigma$) after 10-k cycles (= 320-k disturbs).

In Fig. 20, MC simulation results are also shown (solid circles). It is first mentioned that the procedures for obtaining the measured and simulated plots in Fig. 20 would look quite different. This is because while the measured pdf is determined by scanning it at SNM = 0 by shifting the mean of the pdf by sweeping V_{CC}, the simulated pdf is determined directly. The simulations were performed by modifying the procedure in Sect. 4, by replacing steps (S5) and (S6) with the following.

(S5′) Find ΔSNM for $N = N_1$ ($\equiv \Delta\mathrm{SNM}_1$), using the procedure as in (S5).
(S6′) Find ΔSNM for $N = N_2$ ($\equiv \Delta\mathrm{SNM}_2$), using the procedure as in (S5).
(S7′) If $(\mathrm{SNM} + \Delta\mathrm{SNM}_1 - \mu)/\sigma < -M_{\mathrm{REF}}$, mark the cell as rejected at test.
(S8′) Record the values of $\mathrm{SNM} + \Delta\mathrm{SNM}_1$ and $\mathrm{SNM} + \Delta\mathrm{SNM}_2$.

The simulated solid marks in Fig. 20 were obtained by creating a histogram of $\mathrm{SNM} + \Delta\mathrm{SNM}_2$ for a large number of bits (i.e., MC simulation runs), and normalizing it both horizontally and vertically. The screening operation at $M = M_{\mathrm{REF}}$ was taken into account by assigning an out-of-range value to $\mathrm{SNM} + \Delta\mathrm{SNM}_2$ of the rejected bits. For the horizontal normalization, an SNM + ΔSNM_2 value must be translated into a normalized SNM deviation. To do this, a seemingly natural choice is to use the transformation

$$f(x) = \frac{x - \mu}{\sigma}, \tag{12}$$

where μ and σ are the mean and standard deviation of SNM determined in step (S1) of the MC simulations. However, in real measurements, it is not possible to perfectly separate the effects of variability and RTN. Some randomness induced by RTN is always mixed up with variability. This is also the case for the measured plots in Fig. 20, where the normalized deviation was estimated from measured FBC_1 data, which is certainly affected by RTN. Therefore, for the horizontal normalization, μ and σ in Eq. (12) were replaced by those of SNM + ΔSNM_1 to make the results consistent with the measurements. In addition, N_1 in step (S5$'$) was set to 16 (i.e., average number of disturbs before the first reading in the SRAM measurements). By properly taking into account the effects of RTN on the first reading in this way, the good agreement between the measurements and simulations was obtained, using the same set of parameters that reproduce FBC vs. cycle count measurements. Without such measures, the simulated extension of the tail portion will be overestimated. It is also pointed out that, if we perform multiple tests, or increase N_1, the extension of the tail can be reduced, because more vulnerable bits can be screened. Once the MC simulation is calibrated in this way, the tail distribution after 10 years operation can be estimated using the simulation, as shown by the dotted circle plots in Fig. 20.

6 Conclusion and Remarks

A systematic set of methods for assuring SRAM reliability against RTN proposed earlier [16, 17] has been reviewed, supplementing additional details. It was pointed out that there is similarity between RTN reliability and other long-term reliability issues (e.g., TDDB, BTI), owing to the existence of extremely rarely switching traps, and that something similar to "lifetime extrapolation" is necessary. Another important point is that there is large inter-device variability of RTN waveforms and statistical evaluation of a large number of devices is mandatory. It was argued that these requirements could be met by using accelerated SRAM array measurements, combined with an extremely simplified and fast Monte Carlo simulation. A procedure for evaluating SRAM read failure probability in a product lifetime (e.g., 10 years) was described.

It is considered that, regarding the methods presented, there are still problems that remain to be addressed. Following are some of them that seem to be important. (1) The methods should be extended to cover SRAM write stability, which is equally important as read stability. (2) Validity of the MC simulation algorithm should be proved mathematically. The agreement with the measurements is still empirical. (3) The trap parameter distributions assumed, which strongly affect extrapolation results, seems to be too simplified. More trustworthy parameter settings based on experimental and/or theoretical studies are desired.

Acknowledgments The author would like to express sincere gratitude to Toshiharu Nagumo, Koichi Takeda, Shinobu Asayama, Shinji Yokogawa, Kiyotaka Imai, Yoshihiro Hayashi, and Takashi Hase. The work could not be accomplished without their collaboration and/or support.

References

1. K.S. Ralls, W.J. Skocpol, L.D. Jackel, R.E. Howard, L.A. Fetter, R.W. Epworth, D.M. Tennant, Discrete resistance switching in submicrometer silicon inversion layers: Individual interface traps and low-frequency (1/f?) noise. Phys. Rev. Lett. **52**(3), 228–231 (1984)
2. M.J. Kirton, M.J. Uren, Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency (1/f) noise. Adv. Phys. **38**(4), 367–468 (1989)
3. S. Machlup, Noise in semiconductors: Spectrum of a two-parameter random signal. J. Appl. Phys. **25**(3), 341–343 (1954)
4. M. Tsai, T. Ma, The impact of device scaling on the current fluctuations in MOSFET's. IEEE Trans. Electron Devices **41**(11), 2061–2068 (1994)
5. N. Tega, H. Miki, F. Pagette, D.J. Frank, A. Ray, M.J. Rooks, W. Haensch, 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 (VLSIT)*, 2009, pp. 50–51
6. K. Takeuchi, T. Nagumo, S. Yokogawa, K. Imai, Y. Hayashi, Single-charge-based modeling of transistor characteristics fluctuations based on statistical measurement of RTN amplitude, in *Symposium on VLSI Technology (VLSIT)*, 2009, pp. 54–55
7. K. Fukuda, Y. Shimizu, K. Amemiya, M. Kamoshida, C. Hu, Random telegraph noise in flash memories—model and technology scaling, in *IEEE International Electron Devices Meeting (IEDM)*, 2007, pp. 169–172
8. H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, O. Tsuchiya, Random telegraph signal in flash memory: Its impact on scaling of multilevel flash memory beyond the 90-nm node. IEEE J. Solid State Circuits **42**(6), 1362–1369 (2007)
9. C.M. Compagnoni, R. Gusmeroli, A.S. Spinelli, A.L. Lacaita, M. Bonanomi, A. Visconti, Statistical model for random telegraph noise in flash memories. IEEE Trans. Electron Devices **55**(1), 388–395 (2008)
10. N. Tega, H. Miki, M. Yamaoka, H. Kume, T. Mine, T. Ishida, Y. Mori, R. Yamada, 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
11. S.O. Toh, Y. Tsukamoto, Z. Guo, L. Jones, T.K. Liu, B. Nikolić, Impact of random telegraph signals on Vmin in 45nm SRAM, in *IEEE International Electron Devices Meeting (IEDM)*, 2009, pp. 767–770
12. S.O. Toh, T.K. Liu, B. Nikolić, Impact of random telegraph signaling noise on SRAM stability, in *Symposium on VLSI Technology (VLSIT)*, 2011, pp. 204–205
13. M. Yamaoka, H. Miki, A. Bansal, S. Wu, D.J. Frank, E. Leobandung, K. Torii, Evaluation methodology for random telegraph noise effects in SRAM arrays, in *IEEE International Electron Devices Meeting (IEDM)*, 2011, pp. 745–758

14. M. Fan, V.P. Hu, Y. Chen, P. Su, C. Chuang, Analysis of single-trap-induced random telegraph noise on FinFET devices, 6T SRAM cell, and logic circuits. IEEE Trans. Electron Devices **59**(8), 2227–2234 (2012)
15. T. Matsumoto, K. Kobayashi, H. Onodera, Impact of random telegraph noise on CMOS logic circuit reliability, in *IEEE Custom Integrated Circuits Conference (CICC)*, 2014, pp. 1–8
16. K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, Y. Hayashi, Direct observation of RTN-induced SRAM failure by accelerated testing and its application to product reliability assessment, in *Symposium on VLSI Technology (VLSIT)*, 2010, pp. 189–190
17. K. Takeuchi, T. Nagumo, T. Hase, Comprehensive SRAM design methodology for RTN reliability, in *Symposium on VLSI Circuits (VLSIC)*, 2011, pp. 130–131
18. T. Nagumo, K. Takeuchi, S. Yokogawa, K. Imai, Y. Hayashi, New analysis methods for comprehensive understanding of random telegraph noise, in *IEEE International Electron Devices Meeting (IEDM)*, 2009, pp. 759–762
19. T. Nagumo, K. Takeuchi, T. Hase, Y. Hayashi, Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps, in *IEEE International Electron Devices Meeting (IEDM)*, 2010, pp. 628–631
20. H. Miki, N. Tega, Z. Ren, C.P. D'Emic, Y. Zhu, D.J. Frank, M.A. Guillorn, D. Park, W. Haensch, K. Torii, Hysteretic drain-current behavior due to random telegraph noise in scaled-down FETs with high-κ/metal-gate stacks, in *IEEE Electron Devices Meeting (IEDM)*, 2010, pp. 620–623
21. H. Miki, N. Tega, M. Yamaoka, D.J. Frank, A. Bansal, M. Kobayashi, K. Cheng, C.P. D'Emic, Z. Ren, S. Wu, J.-B. Yau, Y. Zhu, M.A. Guillorn, D.-G. Park, W. Haensch, E. Leobandung, K. Torii, Statistical measurement of random telegraph noise and its impact in scaled-down high-κ/metal-gate MOSFETs, in *IEEE International Electron Devices Meeting (IEDM)*, 2012, pp. 450–453
22. N. Izumi, H. Ozaki, Y. Nakagawa, N. Kasai, T. Arikado, Evaluation of transistor property variations within chips on 300-mm wafers using a new MOSFET array test structure. IEEE Trans. Semicond. Manuf. **17**(3), 248–254 (2004)
23. K. Agarwal, F. Liu, C. McDowell, S. Nassif, K. Nowka, M. Palmer, D. Acharyya, J. Plusquellic, A test structure for characterizing local device mismatches, in *Symposium on VLSI Circuits (VLSIC)*, 2006, pp. 67–68
24. K.Y. Doong, T.J. Bordelon, L. Hung, C. Liao, S. Lin, S.P. Ho, S. Hsieh, K.L. Young, Field-configurable test structure array (FC-TSA): Enabling design for monitor, model, and manufacturability. IEEE Trans. Semicond. Manuf. **21**(2), 169–179 (2008)
25. T. Tsunomura, A. Nishida, T. Hiramoto, Verification of threshold voltage variation of scaled transistors with ultralarge-scale device matrix array test element group. Jpn. J. Appl. Phys. **48**(12R), 124505 (2009)
26. M. Tanizawa, S. Ohbayashi, T. Okagaki, K. Sonoda, K. Eikyu, Y. Hirano, K. Ishikawa, O. Tsuchiya, Y. Inoue, Application of a statistical compact model for random telegraph noise to scaled-SRAM Vmin analysis, in *Symposium on VLSI Technology (VLSIT)*, 2010, pp. 95–96
27. K.V. Aadithya, A. Demir, S. Venugopalan, J. Roychowdhury, SAMURAI: An accurate method for modelling and simulating non-stationary random telegraph noise in SRAMs, in *Design, Automation & Test in Europe (DATE)*, 2011, pp. 1–6
28. K. Aadithya, S. Venogopalan, A. Demir, J. Roychowdhury, MUSTARD: A coupled, stochastic/deterministic, discrete/continuous technique for predicting the impact of random telegraph noise on SRAMs and DRAMs, in *ACM Design Automation Conference (DAC)*, 2011, pp. 292–297
29. H. Reisinger, T. Grasser, W. Gustin, C. Schlünder, The statistical analysis of individual defects constituting NBTI and its implications for modeling DC- and AC-stress, in *IEEE International Reliability Physics Symposium (IRPS)*, 2010, pp. 7–15
30. T. Grasser, K. Rott, H. Reisinger, P.J. Wagner, W. Goes, F. Schanovsky, M. Waltl, M. Toledano-Luque, B. Kaczer, Advanced characterization of oxide traps: The dynamic time-dependent defect spectroscopy, in *IEEE International Reliability Physics Symposium (IRPS)*, 2013, pp. 2D. 2.1–2D. 2.7

31. P. Weckx, B. Kaczer, M. Toledano-Luque, T. Grasser, P.J. Roussel, H. Kukner, P. Raghavan, F. Catthoor, G. Groeseneken, Defect-based methodology for workload-dependent circuit lifetime projections-application to SRAM, in *IEEE International Reliability Physics Symposium (IRPS)*, 2013, pp. 3A. 4.1–3A. 4.7
32. K. Giering, C. Sohrmann, G. Rzepa, L. Heiß, T. Grasser, R. Jancke, NBTI modeling in analog circuits and its application to long-term aging simulations, in *IEEE International Integrated Reliability Workshop (IIRW)*, 2014, pp. 29–34
33. E.Y. Wu, J. Suñé, Power-law voltage acceleration: A key element for ultra-thin gate oxide reliability. Microelectron. Reliab. **45**(12), 1809–1834 (2005)
34. Z. Ji, L. Lin, J.F. Zhang, B. Kaczer, G. Groeseneken, NBTI lifetime prediction and kinetics at operation bias based on ultrafast pulse measurement. IEEE Trans. Electron Devices **57**(1), 228–237 (2010)
35. E. Seevinck, F.J. List, J. Lohstroh, Static-noise margin analysis of MOS SRAM cells. IEEE J. Solid State Circuits **22**(5), 748–754 (1987)
36. A.J. Bhavnagarwala, X. Tang, J.D. Meindl, The impact of intrinsic device fluctuations on CMOS SRAM cell stability. IEEE J. Solid State Circuits **36**(4), 658–665 (2001)
37. Y. Tsukamoto, K. Nii, S. Imaoka, Y. Oda, S. Ohbayashi, T. Yoshizawa, H. Makino, K. Ishibashi, H. Shinohara, Worst-case analysis to obtain stable read/write DC margin of high density 6T-SRAM-array with local Vth variability, in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2005, pp. 398–405
38. D. Khalil, M. Khellah, N. Kim, Y. Ismail, T. Karnik, V.K. De, Accurate estimation of SRAM dynamic stability. IEEE Trans. Very Large Scale Integr. Syst. **16**(12), 1639–1647 (2008)