Reliability Analysis and Mitigation of Near-Threshold Voltage (NTC) Caches

Anteneh Gebregiorgis, Rajendra Bishnoi, and Mehdi B. Tahoori

1 Introduction

SRAM based memory elements have been the prominent limiting factor in the near-threshold voltage domain as the supply voltage of SRAM cells does not easily downscale, as it is done for combinational logic. The supply voltage downscaling limitation is due to the significant increase in the failure rate of SRAM cells operating at lower supply voltage values, which in turn severely affects the yield. Various state-of-the-art solutions have been proposed to address this issue. These solutions include variation tolerant SRAM cell design [3, 13, 29] and heterogeneous cache design [31], improve the robustness of cache memories. However, the improvement comes at the cost of increased area and power overheads. Moreover, these approaches mostly ignore the impact of runtime failure mechanisms, such as aging and soft error, on the reliability of memory components. Therefore, design-time reliability failure analysis and mitigation schemes are crucial for the reliable operation of near-threshold caches.

Analyzing failures based on a particular reliability failure mechanism is insufficient for estimating the system-level reliability, as the interdependence among different failure mechanisms has a considerable impact on the overall system reliability. Moreover, the running workload affects the aging and SER of memory components as it determines the SP and AVF of the memory elements. Therefore, performing a combined analysis on the reliability failure mechanisms across different layers of abstraction (as shown in Fig. 1) is crucial, and it helps designers to choose the most reliable components at each abstraction layer, and tackle the reliability challenges of NTC operation.

A. Gebregiorgis · R. Bishnoi · M. B. Tahoori (✉)
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
e-mail: anteneh.gebregiorgis@kit.edu; rajendra.bishnoi@kit.edu; mehdi.tahoori@kit.edu

© The Author(s) 2021

J. Henkel, N. Dutt (eds.), *Dependable Embedded Systems*, Embedded Systems,
https://doi.org/10.1007/978-3-030-52017-5_13

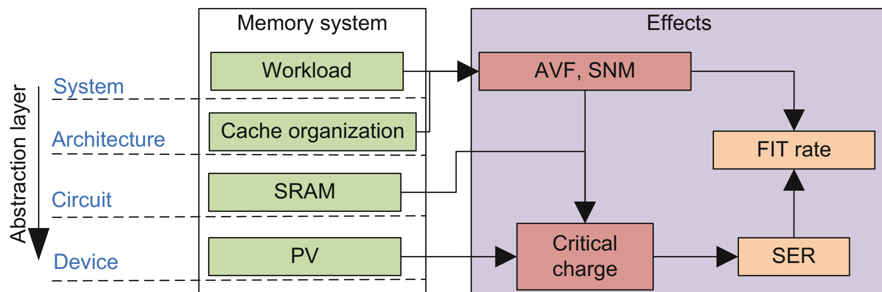

Fig. 1 Cross-layer impact of memory system and workload application on system-level reliability (Failure In-Time (FIT rate)) of NTC memory components, and their interdependence

For this purpose, a comprehensive cross-layer reliability analysis framework addressing the combined effect of aging, process variation, and soft error on the reliability of NTC cache designs is presented in this chapter. Moreover, the chapter presents the advantages and limitations of two different NTC SRAM cell designs (namely, 6T and 8T cells) in terms of reliability (SER and SNM) improvement, area, and energy overheads. The framework presented in this chapter helps to explore the cross-layer impact of different reliability failure mechanisms, and it is useful to study the combined effect of workload and cache organization on the SER and SNM of cache memories. The framework is also helpful to understand how the reliability issues change from super-threshold to the near-threshold voltage domain. Furthermore, it is important for architectural-level design space exploration to find the best cache organization for better reliability and performance trade-offs of NTC caches. Based on the comprehensive analysis using the framework, a memory failure mitigation scheme is developed to improve the energy efficiency of NTC caches.

2 Functional Failure and Reliability Issues of NTC Memory Components

The increase in sensitivity to process variation of NTC circuits affects not only the performance but also functionality. Notably, the mismatch in device strength due to process variation affects the state of positive feedback loop based storage elements (SRAM cells) [3, 10, 14]. The mismatch in the transistors makes SRAM cells to incline for one state over the other, a characteristic that leads to hard functional failure or soft timing failure [17, 20]. The variation-induced functional failure rate of SRAM cells is more pronounced in the nanoscale era as highly miniaturized devices are used to satisfy the density requirements [1]. SRAM cells mainly suffer from three main unreliability sources: (1) aging effects, (2) radiation-induced soft error, and (3) variation-induced functional failures [19]. The SRAM cell susceptibility to these issues increases with supply voltage downscaling.

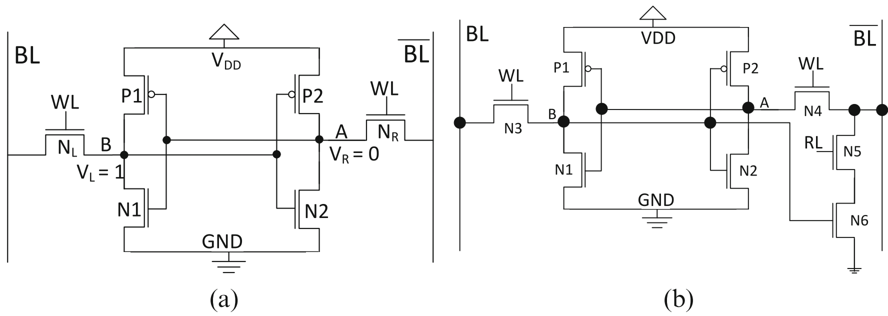

Fig. 2 Schematic diagram of 6T and 8T SRAM cell, where WL = word-line, BL = bit-line and RL = read-line. (**a**) 6T cell design. (**b**) 8T cell design

2.1 Aging Effects in SRAM Cells

Accelerated transistor aging is one of the main reliability concerns in CMOS devices. Among various mechanisms, Bias Temperature Instability (BTI) is the primary aging mechanism in nanoscale devices [18]. BTI gradually increases the threshold voltage of a transistor over a long period, which in turn increases the gate delay [18]. BTI-induced threshold voltage shift is a strong function of temperature as it has an exponential dependency. Hence, BTI-induced aging rate is higher at high operating voltage and temperature values. In SRAM cells, BTI reduces the Static Noise Margin (SNM)[1] of an SRAM cell, and makes it more susceptible to failures. BTI-induced SNM degradation is higher when the cell stores the same value for a longer period (e.g., storing "0" at node "A" of the SRAM cell shown in Fig. 2a). Hence, the effect of BTI on an SRAM cell is a strong function of the cell's Signal Probability (SP).[2]

2.2 Process Variation in SRAM Cells

Variation in transistor parameters such as channel length, channel width, and threshold voltage results in a mismatch in the strength of the transistors in an SRAM cell, and in extreme cases it makes the cell to fail [15]. The variation-induced memory failure rate increases significantly with supply voltage downscaling, for instance, SRAM cells operating at NTC (0.5 V) have 5× higher failure rate than the cells operating at a nominal voltage [15]. Process variation affects several aspects of SRAM cells, and the main variation-induced SRAM cell failures are:

[1]SNM is the minimum amount of DC noise that leads to a loss of the stored value.

[2]Probability of storing logic "1" in the SRAM cell.

Read Failure Read failure/disturb is a phenomenon where the stored value is distorted during read operation. For example, when reading the value of the cell shown in Fig. 2a, (V_L = "1" and V_R = "0"), due to the voltage difference between the access transistor N_R and pull-down transistor N_2, the voltage at node V_R increases [21, 39]. If this voltage is higher than the trip voltage (V_{trip}) of the left inverter, then the stored value of the cell is changed. Hence, the condition for read failure is expressed as [33]:

$$\text{read failure} = \begin{cases} 1, & \text{if } V_R > V_{trip} \\ 0, & \text{otherwise} \end{cases}$$

where $V_{trip} = V_{P_1} - V_{N_1}$ (here V_{P_1} and V_{N_1} indicate the voltages of the PMOS and NMOS transistors of the left inverter shown in Fig. 2a where P_1 and N_1 are the corresponding PMOS and NMOS transistors of the inverter).

Write Failure Write failure occurs when the cell is not able to write/change its state with the applied write voltage. For example, during a write operation (e.g., writing "0" to the SRAM cell shown in Fig. 2a), the node V_L is discharged through the bit-line BL. Write failure occurs when the node V_L is not reduced to be lower than V_{trip} of the right inverter (V_R) [21, 33]. In the standard 6T SRAM cell, write failure is a challenging issue as the cell cannot be optimized without reducing its read margin [21, 33, 39]. However, this is improved with the help of read/write assist circuitries or differential read/write access as it is done in the 7T, 8T, and 10T SRAM cell designs [3, 8, 10]. In order to illustrate the write failure issue, the write margin behaviors of 6T and 8T NTC SRAM cells are studied and compared in Fig. 3. As shown in the figure, the 6T SRAM cell has a smaller write margin as it

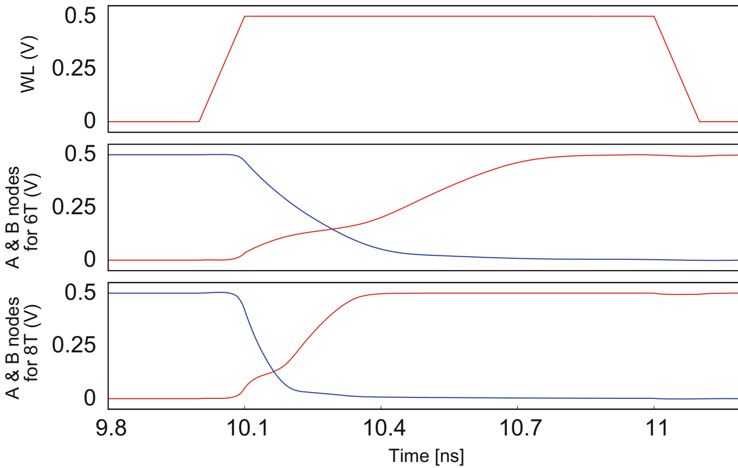

Fig. 3 Write margin (in terms of write latency) comparison of 6T and 8T SRAM cell operating in near-threshold voltage domain (0.5 V)

has longer write latency. On the other hand, the short write latency of the 8T design enables it to have a relatively larger write margin. The improvement in the write margin is because the 8T cell is optimized to improve the write operation without affecting its read operation, as the write and read operations are decoupled.

Hold Failure Hold failure commonly known as metastability issue is a reliability issue that occurs when the SRAM cell is not able to store the value for a longer period [20, 33]. This problem happens during a standby mode if the voltage at nodes V_L or V_R is smaller (smaller SNM value), then the stored value is easily destroyed by a noise voltage due to various sources such as particle strike and leakage current [20, 33].

2.3 Soft Error Rate in SRAM Cells

In SRAM cells, soft error is a transient phenomenon that occurs when charged particles penetrate the cell's cross junction creating an aberrant charge that changes the state of the cell [27]. The primary source of soft errors is related to cosmic ray events such as neutrons and alpha particles. Atmospheric neutrons are one of the higher flux components, and their reaction has a high energy transfer. Thus, neutrons are the most likely cosmic radiations to cause soft errors [16, 19]. Neutrons do not generate electron-hole pairs directly. However, their interaction with the Si-atoms generates secondary particles. These secondary particles produce charges/electron-hole pairs [16]. If the generated charges are larger than the *critical charge*[3] of an SRAM cell, then the internal value of the cell is inverted, this phenomenon is commonly referred to as soft error.

Radiation-induced Soft Error Rate (SER) of an SRAM cell increases significantly with decrease in the supply voltage. Previous experiments have shown that the radiation-induced SER increases by 50% for just 20% decrease in the supply voltage [40]. Moreover, the SER of NTC designs is affected by variation and aging-induced SNM degradation.

2.4 Interdependence and Combined Effects

Analyzing failures based on a particular reliability failure mechanism is insufficient for estimating the system-level reliability as the interdependence among different failure mechanisms (such as aging, soft error, and process variation) has a considerable impact on the overall system reliability [4, 19, 20]. Figure 4 shows how the interdependence between different reliability mechanisms (aging, SER, and process

[3]Minimum amount of charge required to upset the stored value, of an SRAM cell.

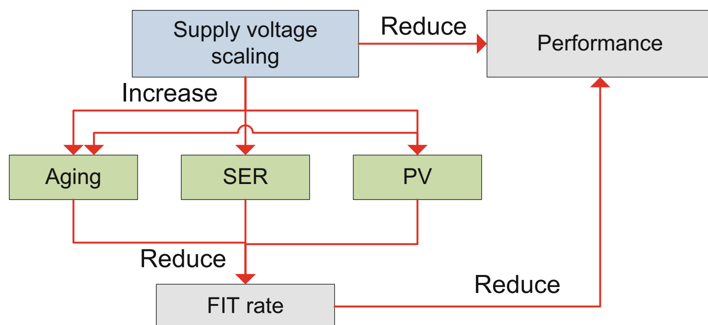

Fig. 4 Interdependence of reliability failure mechanisms and their impact on the system Failure In-Time (FIT) rate in NTC

variation) affects the overall system reliability of memory components in terms of Failure In-Time (FIT rate). As shown in the figure, variation-induced threshold voltage shift increases both aging and SER by reducing the SNM and critical charge of the cell. Similarly, aging-induced SNM degradation increases the sensitivity of SRAM cell to soft errors. The problem is more pronounced when the SRAM cell is operating at NTC domain due to the wide variation extent and higher sensitivity to aging effects [19]. It has been observed that aging has $\approx$5% SNM and critical charge degradation at NTC while process variation-induced SNM degradation reaches as high as 60% [19]. In the super-threshold voltage domain (1.0 V), however, the aging effect increases by $3\times$ to be 15% while variation effect is reduced significantly.

Moreover, the running workload affects the aging rate and SER of memory components, as it determines the signal probability and the Architectural Vulnerability Factor (AVF)[4] of the memory elements [19]. Therefore, to overcome these reliability challenges and improve the overall system reliability, combined analysis of the reliability failure mechanisms at different levels of abstraction is imperative. Besides, the cross-layer analysis should consider the impact of workload on signal probability as well as architectural vulnerability factor of memory components, and their circuit-level consequences on critical charge and SNM degradation.

2.5 Technology Scaling Effects on SRAM Reliability

Reliability has been an essential issue with the miniaturization of CMOS technology, as different design-time and runtime failures are among the limiting factors of technology scaling [24]. At smaller technology nodes, process variation increases the permanent and transient failures of memory components significantly [11, 15].

[4]AVF is the probability that an error in memory structure propagates to the data path. AVF = vulnerable period/total program execution period.

The authors in [15] show that SRAM cell failure rate increases by more than 2× with downscaling from 90 to 65 nm technology node. Similarly, the authors in [26] demonstrated that technology downscaling increases the radiation-induced soft error rate of SRAM cells significantly.

3 Cross-Layer Reliability Analysis Framework for NTC Caches

The comprehensive cross-layer reliability estimation framework that abstracts the impact of workload, cache organization, and reliability failure mechanisms at different levels of abstraction is illustrated in Fig. 5. The reliability analysis and simulation conducted in this work use the symmetric six-transistor (6T) and 8T SRAM cells shown in Fig. 2a and b. In this work, the device-level critical charge characterization is modeled according to the analytical model presented in [27].

This section presents the cross-layer reliability estimation framework in a top-down manner. The system-level *Failure In-Time* (FIT) rate and SNM extraction are described in Sect. 3.1 followed by the cross-layer SNM and SER estimation in Sect. 3.2.

3.1 System FIT Rate Extraction

The system-level FIT rate of a cache memory is the sum of the FIT rate of each row (cache line). The row FIT rate is calculated as the product of the row-wise SER (extracted based on the circuit-level SER information) and its

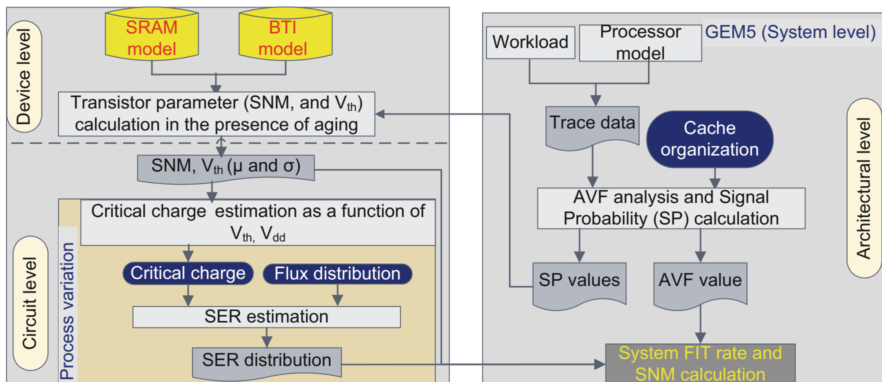

Fig. 5 Holistic cross-layer reliability estimation framework to analyze the impact of aging and process variation effects on soft error rate

Architectural Vulnerability Factor (AVF). Cache AVF is a metrics used to determine the probability that an error in a cache memory propagates to the datapath, and results in a visible error in a program's final output [38]. Equation (1) shows the system-level FIT rate calculation of cache memories.

$$\mathrm{FIT}_{\mathrm{system}} = \sum_{i=0}^{N-1} \mathrm{AVF}_i \times \mathrm{SER}_i \tag{1}$$

where N is the total number of rows in the cache.

3.1.1 Architecture-Level AVF Analysis

One step of determining the failure rate of memory (cache) due to soft errors is to determine the AVF value of the memory. AVF of a memory array is measured by the ratio of vulnerable periods, time interval in which the memory content is exposed to particle strike, to the total program execution period, and the probability of the erroneous value being propagated [38]. Hence, the vulnerability factor of a memory array is computed based on the liveness analysis commonly known as Architectural Correct Execution (ACE) analysis which is the ratio of ACE (vulnerable) cycles to the total number of operational cycles [42]. Therefore, the AVF value of a memory array with M cells is computed as shown in Eq. (2).

$$\mathrm{AVF}_{\mathrm{array}} = \frac{\sum_{i=0}^{M-1} \mathrm{ACE}_i}{T \times M} \tag{2}$$

where T is the total number of cycles.

3.1.2 Architecture-Level SNM Analysis

Aging-induced SNM degradation of an SRAM cell strongly depends on the Signal Probability (SP) of the cell. Thus, BTI-induced SNM degradation is minimized when the signal probability of the cell is balanced (close to 0.5) [18]. In order to determine the aging-induced SNM degradation, the worst-case SP of the memory row is obtained as the maximum SP distance from 0.5 ($D = |\mathrm{SP} - 0.5|$) as shown in Eq. (3). Then, the worst-case SP is used by the SNM estimation tool given in Fig. 5 to determine the corresponding aging-induced SNM degradation.

$$\mathrm{SP}_{\text{worst-case}} = \mathrm{MAX}_{i=1}^{Z} D_i \tag{3}$$

where $D_i = |\mathrm{SP}_i - 0.5|$ and Z is the total number of cells in the memory row.

In order to extract the AVF and SNM of a cache unit, first, it is necessary to extract the trace of the data stored in the cache, read-write accesses, and the duration (number of cycles) of the running workload. Once the information is available, the

reliability analysis tool uses it along with the cache organization to determine the AVF and SP of the cache memory according to Eqs. (2) and (3), and generates the SNM LUT for different signal probability values.

The cache organization (size and associativity) has significant impact on the SER and SNM of the cache, as it determines the hit ratio and the duration data is stored in a cache entry. Hence, different cache size and associativity combinations result in different SER and SNM values for the same workload application. Additionally, SER and SNM are highly dependent on the running workload. In order to explore the impact of cache organization and workload, various organizations and workload applications are investigated.

3.2 Cross-Layer SNM and SER Estimation

3.2.1 SNM Degradation Estimation

Device-Level Aging Analysis

BTI-induced aging degrades the carrier mobility of CMOS transistors, and leads to transistor threshold voltage (V_{th}) shift. In an SRAM cell, the V_{th} shift reduces the noise tolerance margin of the cell, and makes it more susceptible to failures. In the reliability analysis framework, the BTI-induced threshold voltage shift of the transistors in an SRAM cell is evaluated at device-level using a Reaction-Diffusion (RD) model [28]. Then, the device-level V_{th} shift results are used to estimate the corresponding SNM degradation of an SRAM cell at the circuit-level.

Circuit-Level SNM Estimation

The SNM of an SRAM cell is extracted by conducting a circuit-level SPICE simulation. The SPICE simulation uses device-level aging and architecture-level SP results to determine the SNM of the SRAM cell. Finally, the SNM degradation of a particular SP value is obtained according to Eq. (4).

$$DEG_{SP} = \frac{SNM_{SP} - SNM_{fresh}}{SNM_{fresh}} \times 100\% \tag{4}$$

where SNM_{SP} is the SNM of the SRAM cell for a particular signal probability value and SNM_{fresh} is the SNM of a fresh (new) SRAM cell.

Aging and Process Variation-Induced SNM Degradation Analysis

BTI-induced SNM degradation of an SRAM cell depends not only on the cell signal probability but also on process parameters, such as channel length and

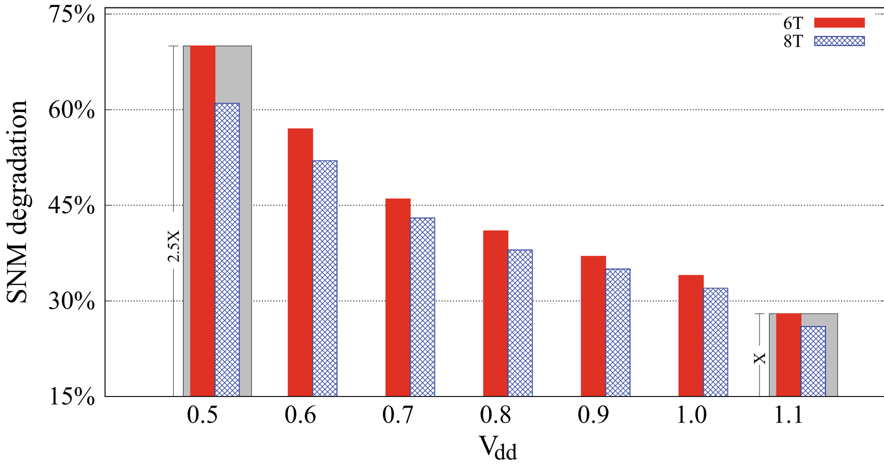

Fig. 6 SNM degradation in the presence of process variation and aging after 3 years of operation, aging+PV-induced SNM degradation at NTC is 2.5× higher than the super-threshold domain

oxide thickness, which are highly affected by manufacturing variabilities. Due to low operating temperature at NTC, aging has relatively less impact on the SNM degradation of near-threshold voltage SRAM cells. However, in combination with variation-induced threshold voltage shift, aging degrades the SNM of SRAM cells significantly.

Figure 6 shows the worst-case aging (SP = 0.0) and variation-induced SNM degradation of 6T and 8T SRAM cells after 3 years of operation for wide supply voltage range. The obtained SNM degradation confirms the analytical expectation as the SNM degradation in NTC is 2.5× higher than the degradation in the super-threshold voltage domain (as shown by the gray boxes). While the use of 8T instead of 6T SRAM cells in super-threshold voltage domain has limited improvement in SNM degradation (only 7.7%), it achieves more than 14% reduction in the SNM degradation in the near-threshold voltage domain.

3.2.2 SER Estimation

The SER of an SRAM cell depends on two main factors, the critical charge of the cell and the flux rate of the strike. To determine SRAM cell SER, first, the critical charge of an SRAM cell is obtained from a circuit-level model. Then, the SER value is calculated by combining the critical charge, flux distribution, and the area sensitive to strike.

Device-Level Critical Charge Characterization

The sensitivity of an SRAM cell to radiation-induced soft errors is determined by the critical charge ($Q_{critical}$) of the cell, as it determines the minimum amount of charge required to alter the state of the cell. The $Q_{critical}$ of an SRAM cell depends on several factors such as supply voltage, threshold voltage, and strength of the transistors of the SRAM cell [9]. The critical charge of an SRAM cell is computed using analytical models or circuit simulators. An analytical model developed in [27] is used to determine the $Q_{critical}$.

As shown in Fig. 5, the SPICE model of an SRAM cell along with the BTI model is employed to evaluate the impact of BTI on the threshold voltage (V_{th}) of the transistors of an SRAM cell. The BTI analysis uses the SP values of the memory array from higher (architecture-level) analysis to determine the BTI-induced V_{th} shift of the running workload. In this way, the aging effect of the workload is incorporated into the framework. Once the fresh and aged V_{th} values are available, the impact of process variation is incorporated as a normal distribution ($\mu \pm 3\sigma$) of the transistor threshold voltage where μ is the mean V_{th} value and the standard deviation (σ) which is obtained using an industrial standard, measurement based, model (the "Pelgrom model") given in Eq. (5) [30]. Finally, all these parameters are used by the model given in [27] to extract the $Q_{critical}$.

$$\sigma \Delta V_{th} = \frac{A_{VT}}{\sqrt{L \times W}} \tag{5}$$

where L and W are the length and width of transistors, and A_{VT} is process specific parameter (the "Pelgrom coefficient").

Circuit-Level SER Analysis

The circuit-level SER analysis is conducted using the SER extraction module of the framework given in Fig. 5. First, the critical charge of the SRAM cell is extracted using the device-level model [27]. Afterward, the critical charge along with the neutron-induced flux distribution is used to determine the SER of the cell using an experimentally verified empirical model given in Eq. (6) [23]. As shown in Eq. (6), the SER of an SRAM cell has an inverse exponential relation with its critical charge ($Q_{critical}$). Hence, the higher the $Q_{critical}$, the lower the SER will be.

$$\text{SER} \propto FAe^{\left(-\frac{Q_{critical}}{Q_s}\right)} \tag{6}$$

where F is the flux in particles/cm^2-s with energy higher than 1 MeV [6]; A is the area sensitive to a strike in cm^2, and Q_S is the charge collection efficiency.

The main observations from Eq. (6) are:

- The SER of an SRAM cell has an inverse exponential relation to its critical charge. Hence, a small decrease in the $Q_{critical}$ leads to an exponential increase in the cell SER.
- For the same atmospheric neutrons, a small drift in $Q_{critical}$ leads to a significant increase in the SER. Furthermore, transistor up-sizing increases the area which is sensitive to particle strike and hence, higher SER.

SER of 6T and 8T SRAM Cells

In the conventional 6T SRAM cell, the cell must maintain the stored value and it should be stable during read/write accesses. SRAM cell stability is a challenging task when the cell is operating in the near-threshold voltage domain, as the cell mainly suffers from read-disturb. To address this issue, either a read-write assist circuitry should be employed or the pull-down (NMOS) transistors of the SRAM cell should be strengthened by transistor up-sizing [35]. However, the up-sizing also increases the area of the cell that is sensitive to soft errors. Since the read-disturb of the 6T SRAM cell is worst when it operates at lower voltage values, transistor up-sizing cannot adequately mitigate the read-disturb issue which makes the 6T design less desirable for near-threshold voltage operation.

This issue is addressed by using alternative SRAM cell designs (such as 8T [32] and 10T [8] SRAM cells). For example, the read failure issue is solved in the 8T design by decoupling the read and write lines using two additional NMOS access transistors. The decoupling allows to downsize the pull-down NMOS transistors, and reduce the area sensitive to soft errors. Therefore, alternative SRAM designs (e.g., 8T) are recommended for NTC operation, which is verified by studying the reliability and energy efficiency improvement of the 8T SRAM design over the conventional 6T design. The transistor sizing specified in [32] is used for the design of the 6T and 8T SRAM cells used in this study.

Figure 7 shows the fresh and aged SER of the 6T and 8T SRAM designs for different supply voltage values. In the super-threshold voltage domain, (0.9–1.1 V) the 6T and 8T designs have negligible differences in their SER. In NTC, however, the 6T design has higher SER than the 8T design due to the effects of transistor up-sizing which increases the area sensitive to radiation. The combined effect of aging and process variation on 6T and 8T SRAM cells is shown in Fig. 8. Figure 8 shows variation effect has severe impact at NTC, as the SER of the 6T and 8T SRAM cell designs in the near-threshold voltage domain is 4× higher than their SER in the super-threshold voltage domain.

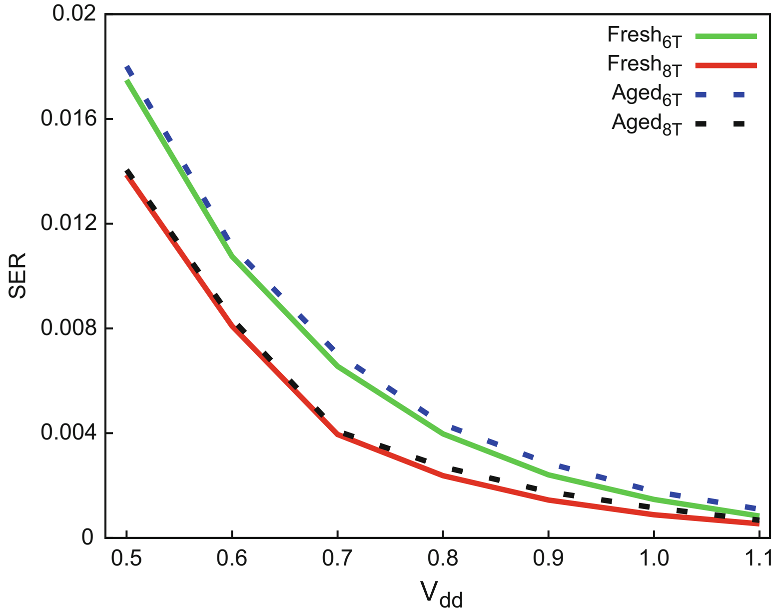

Fig. 7 SER rate of fresh and aged 6T and 8T SRAM cells for various V_{dd} values

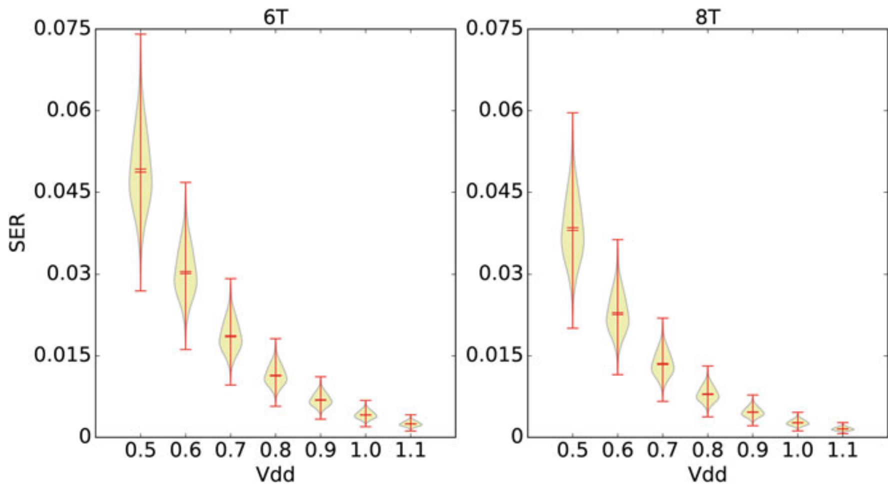

Fig. 8 SER of 6T and 8T SRAM cells in the presence of process variation and aging effects after 3 years of operation

3.3 Experimental Evaluation and Trade-Off Analysis

3.3.1 Experimental Setup

The reliability analysis is conducted using an ALPHA implementation of an embedded in-order core on the Gem5 architectural simulator [7]. Since cache memories are the main focus, various cache sizes (4–16 KB) and wide associativity range from simple directly mapped to 4-way set associative caches are assessed to perform a reliability and performance trade-off analysis. The evaluation is conducted using several workload applications from the SPEC2000 CPU benchmark suite [25]. The workload applications were executed for five million cycles by fast-forwarding to the memory intensive phases. The experimental setup used in this work is presented in Table 1.

The BTI-induced V_{th} shift is extracted by assuming 10% BTI-induced aging after 3 years of operation [37]. First, the 45 nm 6T and 8T SRAM cells are modeled using the PTM model. Afterward, the BTI-induced V_{th} shift LUT and the corresponding SNM degradation for various SP values (0.0–1.0) are obtained using a SPICE simulation. The impact of process variation is considered as a normal distribution of the transistor threshold voltage with a mean ($\mu = V_{th}$, 300 mV) and standard deviation (σ) obtained using the Pelgrom model given in Eq. (5).

To demonstrate the effect of soft error, neutron-induced soft errors are considered as they are the dominant soft error mechanisms at terrestrial altitudes. In order to ensure the proper functionality of both 6T and 8T SRAM cells in the near-threshold voltage domain, their transistors are sized according to the transistor sizing used to model and fabricate near-threshold 6T and 8T SRAM cells specified in [32]. It should be noted that L1 cache is used for illustration purpose only as most embedded

Table 1 Experimental setup, configuration, and evaluated benchmark applications

	Gem5	
Simulation environment	Near-threshold	Super-threshold
Core configuration		
Processor model	Embedded	Embedded
Architecture	Single in-order core	Single in-order core
ISA	ALPHA	ALPHA
Supply voltage	0.5 V	1.1 V
Frequency	100 MHz	1 GHz
Technology node	45 nm PTM	45 nm PTM
Cache configuration		
L1 Cache	Sizes = 4, 8, and 16 KB	Sizes = 4, 8, and 16 KB
	Associativity = 1, 2, and 4 way	Associativity = 1, 2, and 4 way
	Replacement policy = LRU	Replacement policy = LRU
	SRAM cells = 6T and 8T	SRAM cell = 6T
Benchmark	SPEC2000	SPEC2000

NTC processors have limited cache hierarchy. However, the framework is generic, and it is applicable to any cache levels such as L2 and L3.

3.3.2 Workload Effect Analysis

As discussed in Sect. 3.2, BTI-induced SNM degradation of SRAM cell highly depends on the cell's signal probability and the residency time of valid data which varies from one workload application to another. Similarly, the SER of memory components is dependent on the data residency period which is commonly measured using AVF. Hence, for SER analysis, the AVF of different workloads is obtained based on the workload application's data residency period. In order to show the effect of workload variation on SER and SNM degradation, the AVF and signal probabilities of the cache memory are extracted by running different workload applications from the SPEC2000 benchmark suite. Then, the corresponding SNM and SER of the cache memory are obtained using the SER and SNM models presented in Sect. 3.2.

3.3.3 Aging and Variation-Induced SNM Degradation

SNM degradation affects the metastability of SRAM cells. Metastability of SRAM cell determines the stability of the stored value, and it is highly dependent on the worst-case SNM degradation [18]. Therefore, for any workload application, the aging-induced SNM degradation should be evaluated based on the first cell to fail (worst-case SNM degradation).

The impact of workload on the SNM degradation of 6T and 8T based caches across wide supply voltage range is shown in Fig. 9a and b, respectively. For both cases, the SNM degradation increases significantly with supply voltage downscaling. Although the aging rate is slower at lower supply voltage values due to the lower temperature, the wide variation extent in NTC leads to higher aging sensitivity. Hence, in NTC the impact of process variation on SNM is more severe and leads to a significant increase in the aging sensitivity of SRAM cells.

3.3.4 Soft Error Rate Analysis

In order to analyze the impact of workload variation on the soft error rate of cache memories, the architectural vulnerability factor of each workload is extracted and combined with the circuit-level information. Figure 10 shows the contribution of the SPEC2000 workload applications on the SER of the 6T SRAM based cache. As shown in the figure, for all workload applications the SER increases significantly with supply voltage downscaling. For example, the SER of all workload applications increases by five orders of magnitude when the supply voltage is downscaled from the super-threshold voltage (1.1 V) to the near-threshold voltage domain (0.5 V).

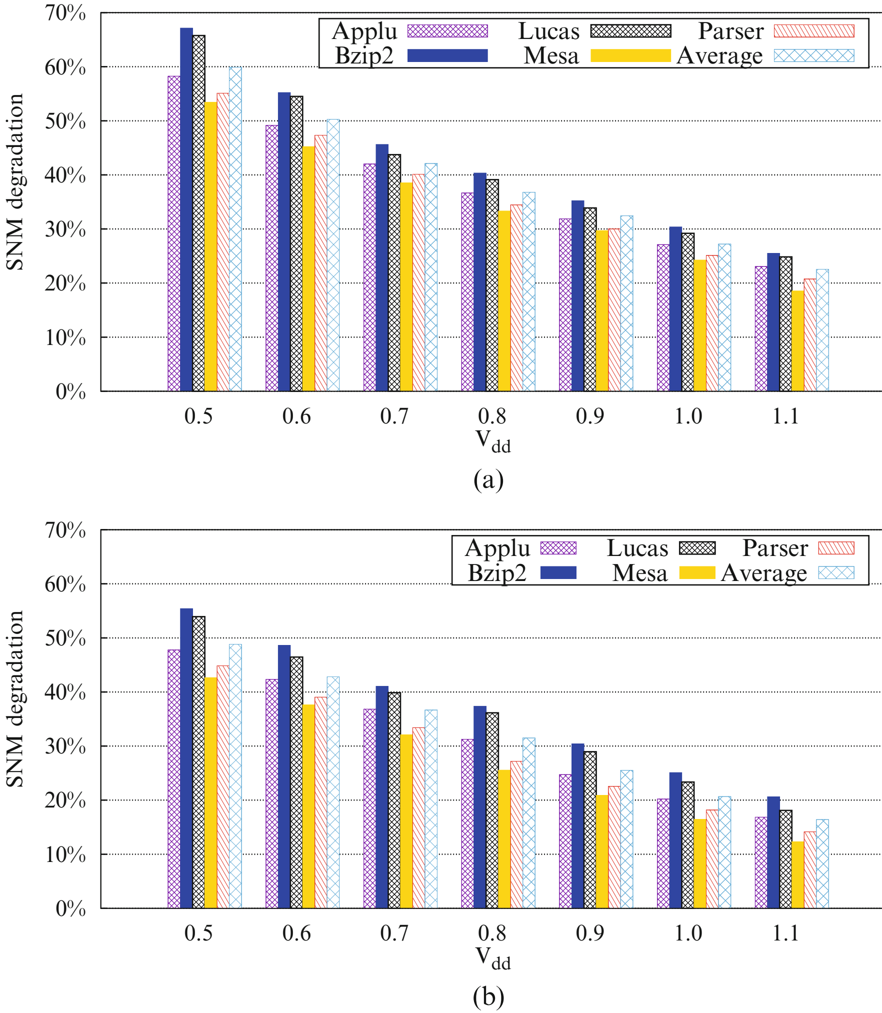

Fig. 9 Workload effects on aging-induced SNM degradation in the presence of process variation for 6T and 8T SRAM cell based cache after 3 years of operation (**a**) 6T SRAM based cache (**b**) 8T SRAM based cache

Additionally, the workload variation has a considerable impact on the soft error rate. For example, the SER of *Bzip2* is almost two orders of magnitude higher than the SER of *Mesa* and *Parser* workload applications. The workload variation impact is observed because *Bzip2* application has higher locality and hit rate which increases the data residency period when compared to the other workload applications. Although the higher hit rate of *Bzip2* leads to a better performance measured in Instructions Per Cycle (IPC), it has a significant impact on the soft error rate of the cache. Hence, it is essential to exploit the workload variation in

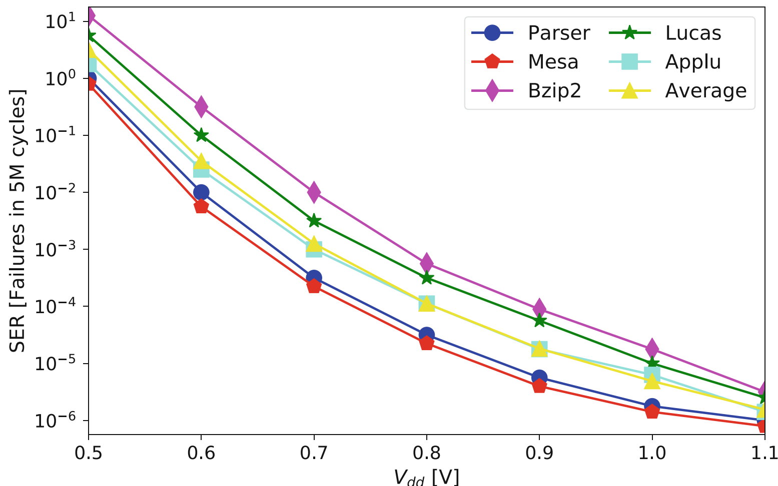

Fig. 10 Workload effect on SER rate of 6T SRAM cell based cache memory for wide supply voltage range

order to downscale the supply voltage of the cache memory in per-application bases for a given target error rate. For a given target FIT rate (e.g., 10^{-2}) the cache has to operate at 0.6 V for *Mesa* and *Parser* workload applications. However, for Bzip2 the cache has to operate at a higher voltage (0.7 V) for the same target error rate.

3.3.5 Cache Organization Impact on System FIT Rate

Cache organization has a significant impact on the performance of embedded processors [34]. Similarly, the organization has an impact on the reliability of cache units. In NTC, the reliability impact of cache organization is even more pronounced. Hence, a proper cache size and associativity selection should consider both performance and reliability as target metrics. The system failure probability (FIT rate and SNM) of a cache unit is highly dependent on the architectural vulnerability factor and the values stored in the cache as well as their residency time intervals, which is in turn is a strong function of the read-write accesses of the cache. Hence, these parameters are influenced by cache size and associativity.

The performance and reliability impacts of different cache organizations in the near and super-threshold voltage domains are evaluated using the configurations described in Table 1. For near-threshold voltage (0.5 V) the processor core frequency is set to 100 MHz, and the cache latency is set to 1 cycle as gate delay is the dominant factor in the near-threshold voltage domain [12]. In the super-threshold voltage domain, however, the cache latency and interconnect delay have a significant impact on the overall delay. Thus, the cache hit latency is set to 2 cycles for 4 and 8 K cache sizes and 3 cycles for the 16 K cache size [41].

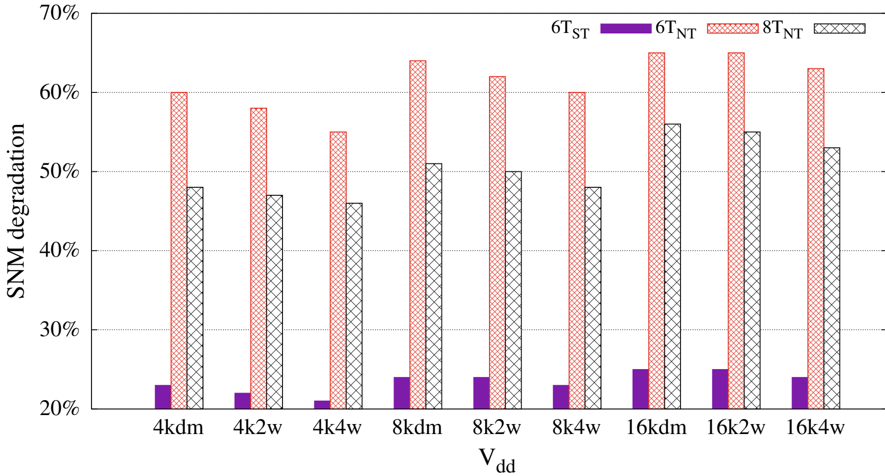

Fig. 11 Impact of cache organization on SNM degradation in near-threshold (NTC) and super-threshold (ST) in the presence of process variation and aging effect after 3 years of operation

3.3.6 Cache Organization and SNM Degradation

Since cache organization determines the data residency period, it has a direct impact on the SNM degradation. Figure 11 illustrates the impact of cache organization on the SNM degradation of near and super-threshold voltage 6T and 8T SRAM cell based memory arrays in the presence of process variation and aging effects after 3 years of operation. The figure shows smaller cache size with higher associativity (4 k-4 w) has less impact on SNM degradation as the data resides in the cache for a smaller duration.

3.3.7 Cache Organization and SER FIT Rate

The cache size and associativity also affect the ACE cycles of cache lines and their failure probabilities. The impact of the cache organization on the FIT rate and performance (IPC) varies along various supply voltage domains. In the super-threshold voltage, an increase in cache size and associativity improves the performance. However, from a FIT rate point of view, an increase in the cache size has a negative impact on FIT rate as it increases the FIT of the cache. Smaller cache sizes, however, have lower performance and better FIT rate. Figure 12 shows the design space of FIT rate and performance (IPC) impact of various cache organizations in the super-threshold voltage domain. In the figure, the FIT rate and performance optimal configuration is (8 k-4 w) as indicated by the blue italic font in Fig. 12.

In the near-threshold voltage domain, the performance is mainly dominated by the delay of the logic unit and the memory failure rate is significantly high. Therefore, it is essential to select a cache organization that gives better reliability

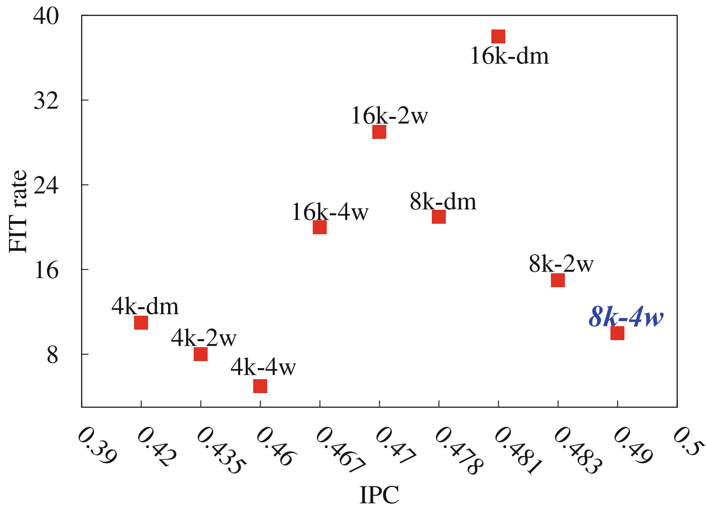

Fig. 12 FIT rate and performance design space of various cache configurations in the super-threshold voltage domain by considering average workload effect (the *blue italic font* indicates optimal configuration)

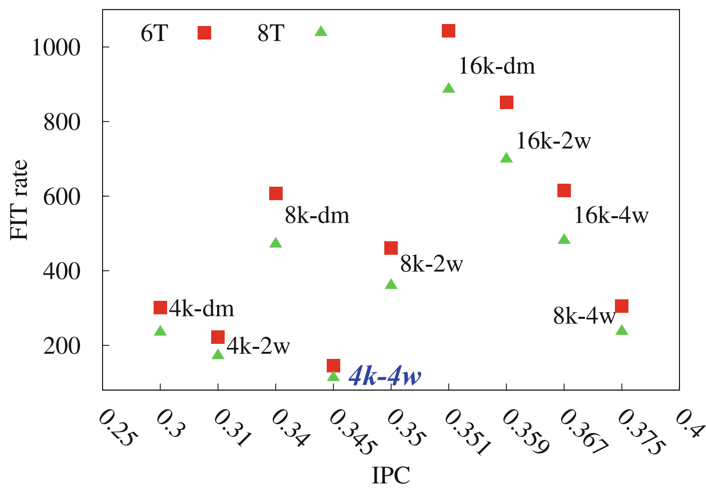

Fig. 13 FIT rate and performance design space of 6T and 8T designs for various cache configurations in the near-threshold voltage domain by considering average workload effect (the *blue italic font* indicates optimal configuration)

(FIT rate and SNM) than performance. Hence, in NTC a smaller cache size with higher associativity gives the best reliability and performance trade-off. Figure 13 shows the design space for the FIT rate and performance trade-off for 6T and 8T designs in NTC.

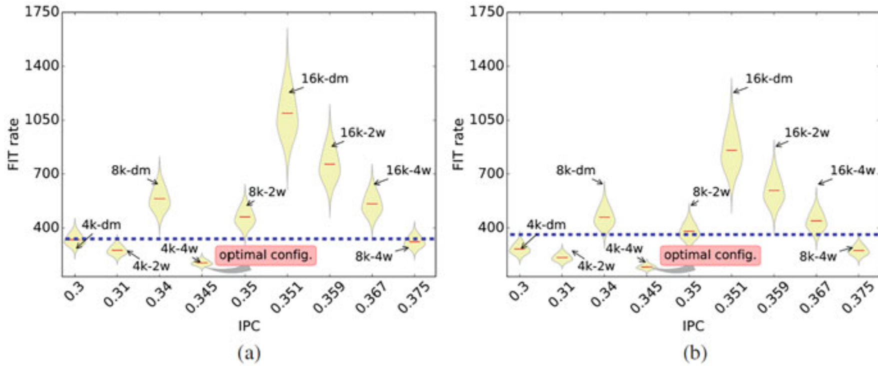

Fig. 14 FIT rate and performance trade-off analysis of near-threshold 6T and 8T caches for various cache configurations and average workload effect in the presence of process variation and aging effects. (**a**) Near-threshold 6T. (**b**) Near-threshold 8T

3.3.8 Reliability-Aware Optimal Cache Organization

The experimental results reported in Figs. 11, 12, 13, and 14 show an increase in the cache associativity improves the performance and reliability (both FIT rate and SNM). Hence, in the super-threshold voltage domain, medium cache size (e.g., 8 KB) with higher associativity has a better reliability and performance trade-off. In NTC, however, smaller cache sizes with higher associativity are preferable for two main reasons: (1) The performance is mainly dominated by the processor core, not by the cache units and hence, cache latency is not an important issue. (2) The soft error rate and SNM degradation are higher in NTC than in the super-threshold voltage domain. Hence, the cache size is reduced by half to obtain a better reliability and performance trade-off in NTC.

In the NTC domain, the selection of an optimal cache organization for the 6T SRAM cell based caches is different from the 8T based caches, depending on the FIT rate and performance requirement. For example, for a target tolerable FIT rate of 350 at NTC (as shown by the dotted line in Fig. 14a and b), only 4 KB 4-way associative cache organization is within the acceptable zone for the 6T-based cache. In the 8T-based cache, however, three additional cache organizations (4 K-dm, 4 k-2 w, and 8 k-4 w) are within the acceptable zone. Hence, the 8 k-4 w cache is used in the 8T-based cache to get ≈10% performance improvement without violating the reliability constraint.

To implement the suggested cache organizations for a specific supply voltage value (only near-threshold or super-threshold) is straightforward. For caches that are expected to operate in both super and near-threshold voltage domains, the reliability-performance optimum cache organization in the super-threshold voltage (e.g., 4-way 8 KB in this case) is preferable. Then, when switching to the near-threshold voltage domain, some portion of the cache is disabled (power gated) in order to maintain the reliability-performance trade-off at NTC.

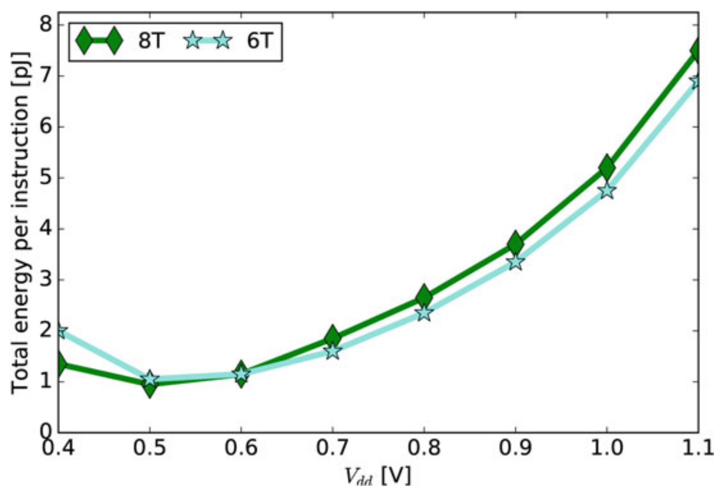

Fig. 15 Energy consumption profile of 6T and 8T based 4 K 4-way cache for wide supply voltage value ranges averaged over the selected workloads from SPEC2000 benchmarks

3.3.9 Overall Energy-Saving Analysis of 6T and 8T Caches

The energy-saving potential of supply voltage downscaling is evaluated by extracting the average energy consumption profile of the 4 K-Byte 4-way set associative cache (i.e., the reliability-performance optimal cache configuration) using 6T and 8T implementations. The energy consumption of the cache memory consists of three different components. These components are peripheries, row and column decoders, and bit-cell array energy consumptions. Since the energy consumption of the periphery and row/column decoder is independent of the bit-cell used, they are assumed to be uniform for both 6T and 8T based caches. Hence, the energy-saving comparison is done based only on the energy consumption of the bit-cell array.

Figure 15 compares the total energy consumption of the 6T and 8T based cache memories for a wide supply voltage range. As shown in the figure, the 8T based cache has slightly higher energy consumption in the super-threshold voltage domain (0.7–1.1 V) than the 6T based cache. The slightly higher energy consumption is because of the additional transistors used for read/write decoupling. However, due to the increase in the failure rate in the near-threshold domain, the 6T based cache consumes more energy than the corresponding 8T based implementation. The energy cost of the higher failure rate is considered as an increase in the read/write latency of the cache. This shows addressing the failures of the 6T cache in NTC results in additional energy cost which makes it less attractive for operating at lower supply voltage values (e.g., below 0.6 V).

3.3.10 Reliability Improvement and Area Overhead Analysis of 8T Based Caches

In a near-threshold voltage SRAM design, the 8T cell improves the soft error rate in the presence of aging and variation effects by up to 25%. Similarly, the SNM is improved by ≈15% using 8T SRAM cells in NTC caches. However, it is expected that the 8T SRAM design has 30% area overhead than the 6T design due to the two additional access transistors. In practice, however, the overhead is much less. Since the 6T SRAM has to be up-sized to increase its read stability, the up-sizing increases the cell area of the 6T design to the extent of being larger than the area of 8T design, as experimentally demonstrated in [32].

4 Voltage Scalable Memory Failure Mitigation Scheme

As shown in the analysis presented in Sect. 3, process variation has a significant impact on the failure rate of memory components operating in the near-threshold voltage domain. Hence, addressing variation-induced memory failures plays an essential role in harnessing NTC benefits. One way of mitigating variation-induced memory failures is by determining the voltage downscaling potential of cache memories without surpassing the tolerable/correctable error margins. For this purpose, the operating voltage of caches should be gracefully reduced so that the number of failing bits due to permanent and transient failures remains tolerable.

This section presents a BIST based voltage scalable mitigation technique to determine an error-free supply voltage downscaling potential of caches at runtime. In order to reduce the runtime configuration complexities, the cache organizations such as size, associativity, and block size are determined during design time. In this work, the block size is considered as the smallest unit used to transfer data to and from the cache. Then, a BIST based runtime cache operating voltage downscaling analysis is performed for a given cache organization. To illustrate the impact of block size selection, the voltage downscaling potential of two block sizes is studied.

4.1 Motivation and Idea

Due to the wide variation extent in NTC, different memory cells have different SNM values; as a result, their minimum operating voltages for a proper functionality vary significantly. The cells with smaller SNM values need to operate at a higher supply voltage than the cells with larger SNM values. Therefore, the supply voltage of some cells (cells with smaller SNM value) should be scaled down more conservatively than the cells with larger SNM in order to maintain the overall reliability. This idea is exploited in order to minimize the effect of process variation and determine error tolerant/error-free voltage downscaling potential of near-threshold caches. Since

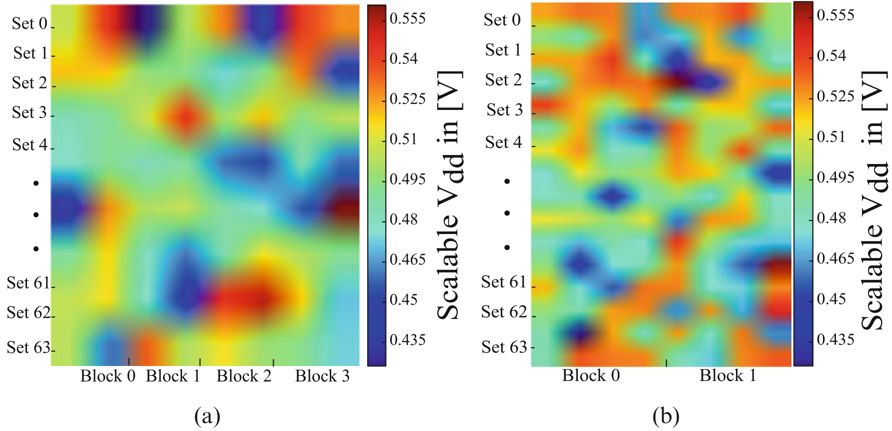

Fig. 16 Error-free minimum operating voltage distribution of 8 MB cache, Set size = 128 Byte (**a**) block size = 32 Bytes (4 blocks per set) and (**b**) block size = 64 Bytes (two blocks per set), the cache is modeled as 45 nm node in CACTI

cache memories are divided into several blocks, block size selection has a significant impact on the supply voltage downscaling potential of cache memories. Hence, one needs to analyze the impact of process variation and supply voltage downscaling potential of cache memories in a per block bases.

Cache block size has a substantial impact on the miss rate and miss penalty of caches at the same time. In order to reduce the cache miss rate and its associated penalty, a larger block size is preferable as it improves locality and reduces the miss rate. From a reliability point of view, however, larger block sizes have wide variation extent, and as a result more failing cells in NTC, which makes the entire block fail. These failures force the cache memory to operate at a much higher voltage (i.e., more conservative scaling) leading to a significant reduction in the energy efficiency. However, this is addressed by decreasing the cache block size in order to reduce cache operating voltage as the variation extent is minimal in comparison to larger block sizes.

To exploit this fact, the impact of block size selection on the supply voltage downscaling potential of a near-threshold voltage 8 KB cache is evaluated as shown in Fig. 16. The cache is modeled in CACTI [36] with 128 Byte set size and two different block sizes, and the impact of process variation is modeled using the threshold voltage variation model given in Eq. (5). As shown in the figure, the smaller block size (Fig. 16a) has narrow variation extent, and hence, it has more supply voltage downscaling potential than its larger block size counterpart (Fig. 16b) at design time. During operation time, the supply voltage downscaling potential of the larger block size cache is reduced further due to various runtime factors such as aging-induced SNM degradation and SER. Moreover, smaller block sizes have lower multiple bit failure rates, and hence, simpler ECC schemes are adopted at a minimum cost [2]. Table 2 shows the ECC overhead comparison for 64 and 32 Byte

Table 2 ECC overhead analysis of different block sizes and correction capabilities

ECC schemes	Block size = 64 Byte			Block size = 32 Byte		
	Area overhead	Storage overhead	Latency overhead	Area overhead	Storage overhead	Latency overhead
SECDED	13k gates	11 bits	2 cycle	≈4k gates	10 bits	1 cycle
DECTED	>50k gates	21 bits	4 cycles	≈10k gates	19 bits	2 cycle
4EC5ED	≈60k gates	41 bits	15 cycles	≈50k gates	37 bits	9 cycle

block sizes according to [2]. The table shows dividing the cache into smaller blocks has an advantage in terms of ECC overhead. Therefore, appropriate cache block size selection should consider both performance and reliability effects at the same time in order to achieve maximum performance while operating within the tolerable reliability margin. Once the cache block size is determined, the cache supply voltage should be tuned at runtime to incorporate the runtime reliability effects such as aging. For this purpose, a BIST based supply voltage tuning is used, and its concept is discussed in the following subsection.

4.2 Built-In Self-Test (BIST) Based Runtime Operating Voltage Adjustment

Built-In Self-Test (BIST) is a widely used technique to test VLSI system on chip [22]. Since memory components occupy majority of the chip area, BIST plays a significant role in testing large and complex memory arrays easily [5, 22]. In order to determine the runtime supply voltage downscaling potential of caches, it is essential to assume a cache memory is equipped with BIST infrastructure to test the entire memory.

In a conventional BIST, the BIST controller generates the test addresses and test patterns (finite number of read/write operations). Then, the test is performed, and the test result is compared with the expected response to determine the failing cells [5]. In this case, however, since the BIST module has to determine the minimum scalable voltage of each block, the test controller has to be modified in order to iteratively test and generate the minimum scalable voltages of each block. The goal is first to determine the error-free minimum scalable voltage of each cache block with/without error correction hardware. Then, the cache operating voltage is determined based on the block with higher operational voltage as shown in Eq. (7), such that the runtime memory failure is minimized.

$$\mathrm{V}_{dd}^{\mathrm{cache}} = \max_{0 \leq i \leq N-1} \mathrm{V}_{dd}^{B_i} \tag{7}$$

where N is the total number of cache blocks, and $\mathrm{V}_{dd}^{B_i}$ is the runtime minimum scalable voltage of block B_i obtained using the iterative BIST.

Algorithm 1 Runtime cache operating voltage adjustment

1: **function** CACHE-V_{dd}-SCALING (C_s, V_{dd}, B_s, F_m){C_s=cache size, V_{dd}= operating voltage, B_s=block size, F_m=tolerable margin of failing bits}
2: $B_t \leftarrow \frac{C_s}{B_s}$;{ B_t=total number of cache blocks}
3: **for** block i←1 to B_t **do**
4: $F_c \leftarrow 0$; {F_c=failing cells counter}
5: $V_{dd}^{new} \leftarrow V_{dd}$; { V_{dd}^{new}=voltage used to perform BIST}
6: **while** $F_c \leq F_m$ **do**
7: Perform BIST using V_{dd}^{new};
8: $F_c \leftarrow$ failing cells;{total number of failing cells per block}
9: $V_{dd}^{new} \leftarrow V_{dd}^{new} - \Delta V_{dd}$;{reduce operating voltage by ΔV_{dd}}
10: **end while**
11: **end for**
12: $V_{dd}^{cache} \leftarrow \max_{1 \leq i \leq B_t} V_{dd_i}^{new}$;{ $V_{dd_i}^{new}$ new operating voltage of $block_i$ }

Algorithm 1 presents the iterative BIST technique used to determine the minimum scalable voltage of cache memory by considering permanent and runtime memory failures. The algorithm takes cache size (C_s), operating voltage (V_{dd}), block size (B_s), and tolerance margin (F_m) as its input. Then, the number of cache blocks is determined by dividing the cache size by the block size (Step 2). Afterward, the minimum scalable voltage of each block is obtained by gradually reducing the operating voltage, and conducting block-level BIST to determine the total number of failing bits at each operating voltage level (Steps 3–10). It should be noted that, the supply voltage is reduced as long as the number of failing bits per block is within the tolerable/correction capability of the adopted error correction scheme. For example, a cache memory equipped with a Single Error Correction Double Error Detection (SECDED) infrastructure tolerates two failing bits per block (hence $F_m = 2$) as SECDED corrects only one bit and detects two erroneous bits at a time. Hence, whenever two failing bits are detected the error-free version is loaded from the lower-level memory which makes SECDED sufficient solution for tolerating two failing bits per block. Finally, the algorithm determines the operating voltage of the cache based on the block with the highest voltage as shown in Step 12.

The overall flow of the cache access control logic along with the BIST infrastructure as well as mapping logic is presented in Fig. 17. The cache controller first decodes the address and identifies the requested block. Then, it determines if the requested block is functional or failing block for the specified operating voltage. If the requested block is functional, then a conventional block access is performed. In case the requested block is a failing one, the error tolerant block mapping scheme is employed to redirect the access request.

Since this approach considers the effect of permanent and transient failure mechanisms, it is orthogonal with different dynamic cache mitigation schemes such as block disabling [1, 43] and strong ECC schemes [2]. For energy-critical systems, block disabling technique is applied in combination with this approach to downscale the cache operating voltage aggressively by disabling the failing blocks at lower operating voltages at the cost of performance reduction (increase in miss rate).

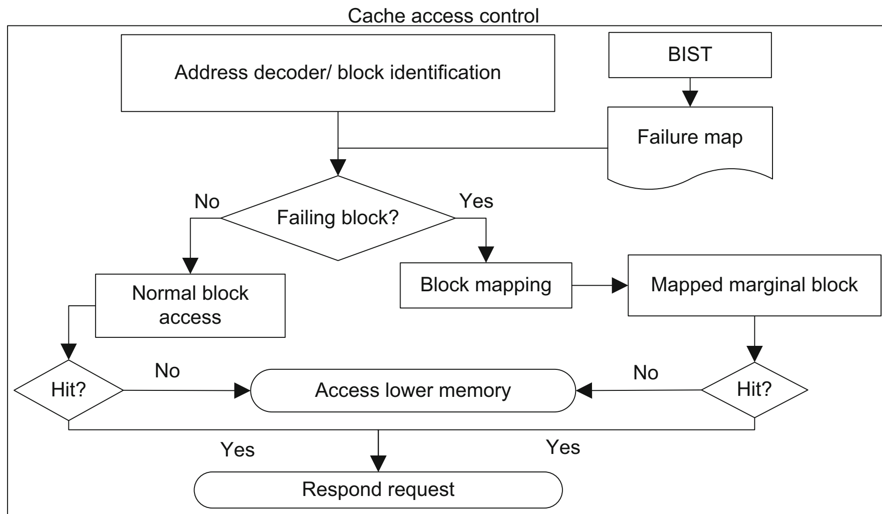

Fig. 17 Cache access control flowchart equipped with BIST and block mapping logic

4.3 Error Tolerant Block Mapping

Once the minimum scalable voltages of the cache blocks are determined, the next task is to disable the failing blocks, and map their read/write accesses to the corresponding non-failing blocks in order to ensure reliable cache operation. Additionally, in order to reduce the vulnerability to runtime failures (such as noise and soft errors), the non-failing blocks are stored in a stack frame sorted by their minimum scalable voltage values. Since the marginal blocks (blocks with less voltage downscaling potential) are more sensitive to runtime failures, they are stored at the top of the stack. Then, access to a disabled block is mapped to the marginal blocks in the stack. The mapping enables to reduce soft error vulnerability of the marginal blocks by reducing their data residency period. Since a stack is a linear data structure in which the insertion and deletion operations are performed at only one end commonly known as "top," the marginal blocks need to be at the top (upper half) of the stack to ensure their fast replacement.

The mapping process is illustrated in Fig. 18 by using an illustrative example. As shown in the figure, the cache blocks are divided into three categories: (1) red blocks are failing blocks. (2) yellow blocks are marginal blocks (non-failing but with limited supply voltage downscaling potential). (3) blue blocks are robust blocks (i.e., non-failing with higher supply voltage downscaling potential). Hence, the marginal blocks are stored at the top of the stack frame. Then, when a disabled (failing) block is requested (e.g., B5) its access request is mapped to a marginal block at the top of the stack frame (e.g., B4), and the stack pointer is updated to point to the next element in the stack. This process continues until all the disabled blocks are mapped. It should be noted that once a block is mapped, it is removed from the mapping stack

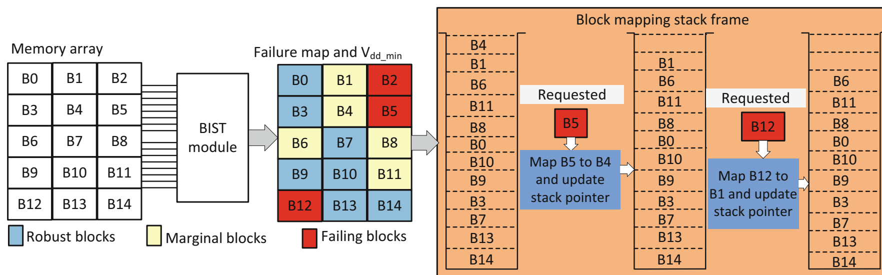

Fig. 18 Error tolerant cache block mapping scheme (mapping failing blocks to marginal blocks)

Table 3 Minimum scalable voltage analysis for different ECC schemes

	Minimum scalable voltage in [V]		
ECC-Scheme	Block size = 16Byte	Block size = 32Byte	Block size = 64Byte
No-ECC	0.50	0.53	0.54
Parity	0.47	0.51	0.53
SECDED	0.43	0.48	0.50

when updating the stack pointer. For example, when block B5 is mapped to block B4, then, block B4 is removed from the stack as shown by the empty slot in Fig. 18.

4.4 Evaluation of Voltage Scalable Mitigation Scheme

4.4.1 Variation-Aware Voltage Scaling Analysis

The supply voltage scalability of three different block sizes (16, 32, and 64 Byte) with different error correction schemes is compared in order to analyze the impact of block size selection on the supply voltage downscaling potential of cache memories with and without error correction schemes. The error-free (correctable error) minimum voltage of three block sizes is studied for 8 KB cache memory without ECC, parity, and Single Error Correction Double Error Detection (SECDED) configurations. Table 3 shows the supply voltage downscaling potential of the studied block sizes. For all ECC schemes (given in Table 3), the cache operating voltage has to be downscaled more conservatively when the block size is larger (64 Bytes). However, larger block sizes help to reduce the cache miss rate that results in a better cache performance. Therefore, for an aggressive supply voltage downscaling, the block size should be selected as small as possible by making performance and energy-saving trade-off analysis.

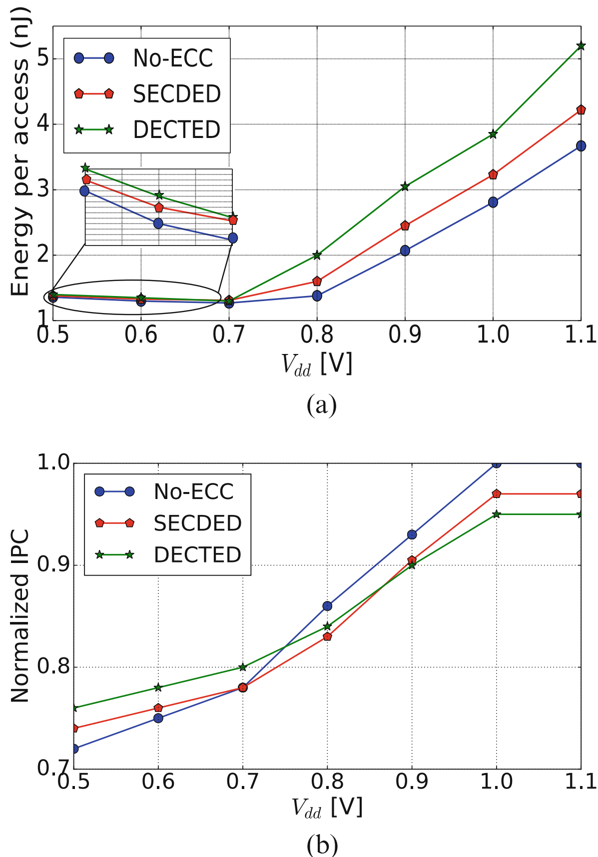

Fig. 19 Comparison of voltage downscaling in the presence of block disabling and ECC induce overheads for gzip, parser, and mcf applications from SPEC2000 benchmark (**a**) energy consumption comparison (**b**) Performance comparison in terms of IPC

4.4.2 Energy and Performance Evaluation of Voltage Scalable Cache Different ECC Schemes

The average energy reduction and performance comparison of voltage scaled cache memory with and without ECC are given in Fig. 19a and b by running selected workloads (*gzip*, *parser*, and *mcf*) from the SPEC2000 benchmark. The energy results in Fig. 19a are extracted from CACTI by considering block disabling, and ECC induced delay and energy overheads. As shown in the figure, supply voltage downscaling improves the energy efficiency significantly. However, the overheads of this scheme, namely ECC energy overhead, block disabling induced cash miss rate, and ECC encoding/decoding delay overhead outweigh the energy gain of supply voltage downscaling when the cache operating voltage is below 0.7 V. Therefore, the energy per access of Double Error Correction Triple Error Detection (DECTED) is higher than SECDED when the supply voltage is scaled down to 0.7 V or below. Similarly, Fig. 19b shows the cache performance (IPC) is reduced significantly with the supply voltage downscaling as more blocks are disabled for reliable operation.

5 Conclusion

Embedded microprocessors, particularly for battery-powered mobile applications, and energy-harvested Internet of Things (IoT) are expected to meet stringent energy budgets. In this regard, operating in the near-threshold voltage domain provides better performance and energy efficiency trade-offs. However, NTC faces various challenges among which increase in functional failure rate of memory components is the dominant issue. This chapter analyzed the combined effect of aging, process variation, and soft error on the reliability of cache memories in super and near-threshold voltage domains. It is observed that the combined effect of process variation and aging has a massive impact on the soft error rate and SNM degradation of NTC memories. Experimental results show process variation and aging-induced SNM degradation is $2.5\times$ higher in NTC than in the super-threshold voltage domain while SER is $8\times$ higher. The use of 8T instead of 6T SRAM cells reduces the system-level SNM and SER by 14% and 22%, respectively. Additionally, workload and cache organization have a significant impact on the FIT rate and SNM degradation of memory components. This chapter demonstrated that the reliability and performance optimal cache organization changes when going from the super-threshold voltage to the near-threshold voltage domain.

References

1. Agarwal, A., Paul, B.C., Mahmoodi, H., Datta, A., Roy, K.: A process-tolerant cache architecture for improved yield in nanoscale technologies. IEEE Trans. Very Large Scale Integr. Syst. **13**, 27–38 (2005)
2. Alameldeen, A.R., Wagner, I., Chishti, Z., Wu, W., Wilkerson, C., Lu, S.L.: Energy-efficient cache design using variable-strength error-correcting codes. In: ACM SIGARCH Computer Architecture News (2011)
3. Aly, R.E., Faisal, M.I., Bayoumi, M.A.: Novel 7T sram cell for low power cache design. In: Proceedings of the 2005 IEEE International SOC Conference (2005)
4. Amrouch, H., van Santen, V.M., Ebi, T., Wenzel, V., Henkel, J.: Towards interdependencies of aging mechanisms. In: Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (2014)
5. Au, A., Pogiel, A., Rajski, J., Sydow, P., Tyszer, J., Zawada, J.: Quality assurance in memory built-in self-test tools. In: 17th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (2014)
6. Autran, J.L., Serre, S., Munteanu, D., Martinie, S., Semikh, S., Sauze, S., Uznanski, S., Gasiot, G., Roche, P.: Real-time soft-error testing of 40 nm SRAMS. In: Proceedings of IEEE International Reliability Physics Symposium (IRPS) (2012)
7. Binkert, N., Beckmann, B., Black, G., Reinhardt, S.K., Saidi, A., Basu, A., Hestness, J., Hower, D.R., Krishna, T., Sardashti, S., et al.: The gem5 simulator. ACM SIGARCH Comput. Archit. News **39**, 1–7 (2011)
8. Calhoun, B.H., Chandrakasan, A.: A 256kb sub-threshold SRAM in 65nm CMOS. In: 2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers (2006)
9. Cazeaux, J.M., Rossi, D., Omana, M., Metra, C., Chatterjee, A.: On transistor level gate sizing for increased robustness to transient faults. In: 11th IEEE International On-Line Testing Symposium, IOLTS (2005)

10. Chang, L., Montoye, R.K., Nakamura, Y., Batson, K.A., Eickemeyer, R.J., Dennard, R.H., Haensch, W., Jamsek, D.: An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches. IEEE J. Solid-State Circ. **43**, 956–963 (2008)
11. Chatterjee, I., Narasimham, B., Mahatme, N., Bhuva, B., Reed, R., Schrimpf, R., Wang, J., Vedula, N., Bartz, B., Monzel, C.: Impact of technology scaling on SRAM soft error rates. IEEE Trans. Nuclear Sci. **61**, 3512–3518 (2014)
12. Chen, H., Manzi, D., Roy, S., Chakraborty, K.: Opportunistic turbo execution in NTC: exploiting the paradigm shift in performance bottlenecks. In: Proceedings of the 52nd Annual Design Automation Conference (2015)
13. Chen, Y.H., Chan, W.M., Wu, W.C., Liao, H.J., Pan, K.H., Liaw, J.J., Chung, T.H., Li, Q., Lin, C.Y., Chiang, M.C., et al.: A 16 nm 128 Mb SRAM in high-κ metal-gate FinFET technology with write-assist circuitry for low-VMIN applications. IEEE J. Solid-State Circuits **50**, 170–177 (2015)
14. Dreslinski, R., Wieckowski, M., Blaauw, D.S., Mudge, T.: Near threshold computing: Overcoming performance degradation from aggressive voltage scaling. In: Proceedings of the Workshop Energy-Efficient Design (2009)
15. Dreslinski, R.G., Wieckowski, M., Blaauw, D., Sylvester, D., Mudge, T.: Near-threshold computing: reclaiming Moore's law through energy efficient integrated circuits. Proc. IEEE **98**, 253–266 (2010)
16. Ebrahimi, M., Evans, A., Tahoori, M.B., Seyyedi, R., Costenaro, E., Alexandrescu, D.: Comprehensive analysis of alpha and neutron particle-induced soft errors in an embedded processor at nanoscales. In: Proceedings of the conference on Design, Automation & Test in Europe (2014)
17. Gebregiorgis, A., Tahoori, M.B.: Reliability and performance challenges of ultra-low voltage caches: A trade-off analysis. In: IEEE 24th International Symposium on On-Line Testing and Robust System Design (IOLTS) (2018)
18. Gebregiorgis, A., Ebrahimi, M., Kiamehr, S., Oboril, F., Hamdioui, S., Tahoori, M.B.: Aging mitigation in memory arrays using self-controlled bit-flipping technique. In: 20th Asia and South Pacific Design Automation Conference (ASP-DAC) (2015)
19. Gebregiorgis, A., Kiamehr, S., Oboril, F., Bishnoi, R., Tahoori, M.B.: A cross-layer analysis of soft error, aging and process variation in near threshold computing. In: Design, Automation & Test in Europe Conference & Exhibition (DATE) (2016)
20. Gebregiorgis, A., Bishnoi, R., Tahoori, M.B.: A comprehensive reliability analysis framework for NTC caches: a system to device approach. IEEE Trans. Comput. Aided Design Integr. Circuits Syst. **38**, 439–452 (2018)
21. Grossar, E., Stucchi, M., Maex, K., Dehraene, W.: Read stability and write-ability analysis of SRAM cells for nanometer technologies. IEEE J. Solid-State Circuits **41**, 2577–2588 (2006)
22. Hamdioui, S., Al-Ars, Z., Gaydadjiev, G.N., Van de Goor, A.: Generic march element based memory built-in self test. US Patent 8,910,001, 2014
23. Hazucha, P., Svensson, C.: Impact of CMOS technology scaling on the atmospheric neutron soft error rate. IEEE Trans. Nuclear Sci. **47**, 2586–2594 (2000)
24. Henkel, J., Bauer, L., Dutt, N., Gupta, P., Nassif, S., Shafique, M., Tahoori, M., Wehn, N.: Reliable on-chip systems in the nano-era: Lessons learnt and future trends. In: Proceedings of the 50th Annual Design Automation Conference (2013)
25. Henning, J.L.: SPEC CPU2000: Measuring CPU performance in the new millennium. Computer **33**, 28–35 (2000)

26. Hubert, G., Artola, L., Regis, D.: Impact of scaling on the soft error sensitivity of bulk, FDSOI and FinFET technologies due to atmospheric radiation. Integration **50**, 39–47 (2015)
27. Jahinuzzaman, S.M., Sharifkhani, M., Sachdev, M.: An analytical model for soft error critical charge of nanometric SRAMs. IEEE Trans. Very Large Scale Integr. Syst. **17**, 1187–1195 (2009)
28. Jeppson, K.O., Svensson, C.M.: Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices. J. Appl. Phys. **48**, 2004–2014 (1977)
29. Jiang, C., Zhang, D., Zhang, S., Wang, H., Zhuang, Z., Yang, F.: A yield-driven near-threshold 8-T SRAM design with transient negative bit-line scheme. In: 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC) (2017)
30. Kuhn, K.J., Giles, M.D., Becher, D., Kolar, P., Kornfeld, A., Kotlyar, R., Ma, S.T., Maheshwari, A., Mudanai, S.: Process technology variation. IEEE Trans. Electr. Devices **58**, 2197–2208 (2011)
31. Maric, B., Abella, J., Valero, M.: Adam: An efficient data management mechanism for hybrid high and ultra-low voltage operation caches. In: Proceedings of the Great Lakes Symposium on VLSI (2012)
32. Morita, Y., Fujiwara, H., Noguchi, H., Iguchi, Y., Nii, K., Kawaguchi, H., Yoshimoto, M.: An area-conscious low-voltage-oriented 8T-SRAM design under DVS environment. In: IEEE Symposium on VLSI Circuits (2007)
33. Mukhopadhyay, S., Mahmoodi, H., Roy, K.: Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. IEEE Trans. Comput. Aided Design Integr. Circuits Syst. **24**, 1859–1880 (2005)
34. Olorode, O., Nourani, M.: Improving performance in sub-block caches with optimized replacement policies. ACM J. Emerg. Technolo. Comput. Syst. **11**, 1–22 (2015)
35. Seok, M., Chen, G., Hanson, S., Wieckowski, M., Blaauw, D., Sylvester, D.: CAS-FEST 2010: Mitigating variability in near-threshold computing. IEEE J. Emer. Sel. Topics Circuits Syst. **1**, 42–49 (2011)
36. Shivakumar, P., Jouppi, N.P.: Cacti 3.0: An integrated cache timing, power, and area model. Technical Report, Compaq Computer Corporation (2001)
37. Siddiqua, T., Gurumurthi, S., Stan, M.R.: Modeling and analyzing NBTI in the presence of process variation. In: 12th International Symposium on Quality Electronic Design (ISQED) (2011)
38. Sridharan, V., Kaeli, D.R.: Using hardware vulnerability factors to enhance AVF analysis. ACM SIGARCH Comput. Archit. News **38**, 461–472 (2010)
39. Takeda, K., Hagihara, Y., Aimoto, Y., Nomura, M., Nakazawa, Y., Ishii, T., Kobatake, H.: A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications. IEEE J. Solid-State Circuits **41**, 113–121 (2006)
40. Tonfat, J., Azambuja, J.R., Nazar, G., Rech, P., Frost, C., Kastensmidt, F.L., Carro, L., Reis, R., Benfica, J., Vargas, F., et al.: Analyzing the influence of voltage scaling for soft errors in SRAM-based FPGAs. In: 14th European Conference on Radiation and Its Effects on Components and Systems (RADECS) (2013)
41. Understanding CPU caching and performance (2015). http://http://arstechnica.com/gadgets/2002/07/caching/2/

42. Wilkening, M., Sridharan, V., Li, S., Previlon, F., Gurumurthi, S., Kaeli, D.R.: Calculating architectural vulnerability factors for spatial multi-bit transient faults. In: Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (2014)
43. Wilkerson, C., Gao, H., Alameldeen, A.R., Chishti, Z., Khellah, M., Lu, S.L.: Trading off cache capacity for reliability to enable low voltage operation. In: ACM SIGARCH Computer Architecture News (2008)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.