A Cross-Layer Analysis of Soft Error, Aging and Process Variation in Near Threshold Computing

Anteneh Gebregiorgis, Saman Kiamehr, Fabian Oboril, Rajendra Bishnoi and Mehdi B. Tahoori
Chair of Dependable Nano Computing (CDNC), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
Email: {anteneh.gebregiorgis, kiamehr, fabian.oboril, rajendra.bishnoi, mehdi.tahoori}@kit.edu

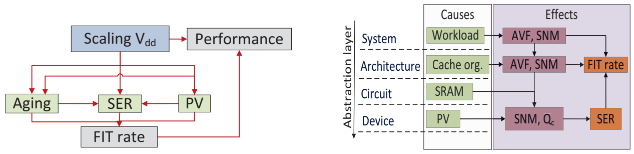
***Abstract*—Near Threshold Computing (NTC) is a promising approach to reduce the power consumption of modern VLSI designs. However, NTC designs suffer from functional failures and performance loss. Understanding the characteristics of the functional failures and variability effects is of decisive importance in order to mitigate them, and get the most out of NTC. This paper presents a cross-layer reliability analysis in the presence of soft errors, aging and process variation effects in the near threshold voltage domain. The objective is to quantify the reliability of different SRAM designs and to find a reliability-performance optimal cache organization for an NTC microprocessor. In this work, the Soft Error Rate (SER) and Signal Noise Margin (SNM) of 6T and 8T SRAM cells and their dependencies on aging and process variation are investigated by considering device, circuit and architecture level analysis. Their experimental results reveal that in NTC, process variation and aging-induced SNM degradation is 2.5X higher than in the super threshold domain while SER is 8X higher. The use of 8T instead of 6T SRAM cells can reduce the system-level SNM and SER by 14% and 22% respectively. Besides, we observe that we can find the right balance between performance and reliability by using an appropriate cache organization at NTC which is different from the super threshold.**

I. Introduction

Energy consumption is an important issue in today's computing systems [1]. One very promising way of minimizing the energy consumption is to reduce the supply voltage (V_{dd}) to a lower value. Scaling V_{dd} to near threshold voltage, commonly known as *Near Threshold Computing* (NTC), can gain up to 10X energy reduction at the expense of 10X performance degradation [2]. However, in addition to performance loss, NTC operation is facing several barriers such as an increase in functional failures and sensitivity to variation effects [2]. The key functional failures and resiliency challenges in NTC include aging, soft errors and process variation.

Aging is a major reliability challenge in semiconductor devices. Aging effects include failure mechanisms such as Bias Temperature Instability (BTI). BTI is a degradation phenomenon that leads to a gradual increase of the transistor threshold voltage ($|V_{th}|$) over a long period of time [3, 4]. In an SRAM cell, BTI degrades the Static Noise Margin (SNM) of the cell and makes it more susceptible to failures. In NTC designs aging has less impact on the SNM of SRAM cells due to the temperature reduction at lower supply voltage values. However, the sensitivity of SNM to aging increases significantly. Radiation-induced *soft errors* are another significant concern in nanoscale CMOS devices. They are the result of the interaction of high energy particles with sensitive regions of CMOS devices and can flip the data state of SRAM cells. In NTC, the susceptibility of SRAM cells to soft errors increases significantly due to the reduction of the critical charge (minimum amount of charge required to upset the stored value) of the cell and variation effects [5]. *Process Variation* (PV) due to mismatch of device parameters, such as transistor threshold voltage from their nominal design time values significantly affects the performance and reliability of NTC designs. In NTC, process variation mainly affects memory components as they are designed using smaller nodes for density (area) reason [6].

To utilize the energy efficiency of NTC designs, the performance loss and delay variations can be handled by guard-banding. However, the energy gain of NTC can be nullified by the usage of complex error correction hardware unless the functional failures in memory structures are addressed appropriately. To overcome the functional failures in memory structures employing Error Correction Code (ECC) mechanisms can be an effective solution in the super threshold domain. In NTC, however, since the functional failures and variability effects increase dramatically, ECC will not be an effective solution anymore. Besides, the interdependency of the failure mechanisms will make it worst. Hence, performing a combined analysis on the failure mechanisms (as shown in Figure 1(a)) across several layers (as in Figure 1(b)) is very important and it will help designers to select the most reliable components at each layer to tackle the reliability challenges of NTC.



(a) Reliability failures in NTC (b) Cross-layer reliability impacts

Fig. 1. Reliability failure mechanisms and their cross-layer impact on system FIT rate in NTC

To overcome the NTC barriers, most researches focus on addressing performance loss and variability, by design techniques such as multiple V_{dd} [7]. However, soft errors, process variation, aging and their interdependency are mainly studied in the super threshold voltage regime [8–10].

In this paper, we present a *cross-layer reliability analysis framework* addressing the impact of aging, process variation and soft errors on the reliability of NTC memory designs. Additionally, to explore the cross layer impact we study the combined effect of workload and cache organization on the Soft Error Rate (SER) and SNM of a memory array. This framework helps to understand how these issues change from super to the near threshold voltage domain. Additionally, it can be used for design space exploration to find the best cache organization for reliability-performance tradeoffs in NTC. Experimental results show that in NTC, the 8T SRAM design is *highly affected by process variation, yet more stable and reliable than its 6T counterpart*. Hence, by using 8T instead of 6T SRAM cells the system-level SNM and SER rate of NTC caches can be improved by 14% and 22% respectively. Moreover, we observe that for a certain FIT rate requirement the size of the cache in an 8T design can be double of the 6T cache size to obtain better performance without violating the system FIT rate requirement.

The rest of this paper is organized as follows. The background of reliability failures in NTC is presented in Section II. Section III presents the proposed cross-layer reliability analysis framework followed by the experimental results in Section IV. Finally, the paper is concluded in Section V.

II. Preliminaries and Related Work

Even though NTC is a promising way to provide better tradeoff for performance and energy efficiency, its applicability is limited by functional failures and variation effects. As shown in Figure 2, the energy consumption reduces exponentially by reducing the supply voltage while the memory failure rate increases dramatically.

978-3-9815370-6-2/DATE16/©2016 EDAA

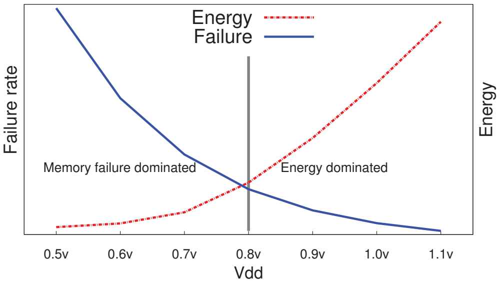


Fig. 2. Energy gain versus memory failure rate for various vdd ranges

A. Aging, Process Variation and Soft Error in NTC

1) ***Aging effects****:* Accelerated transistor aging is one of the main reliability concerns in nanoscale devices. Among various mechanisms, BTI is the primary aging mechanism in nanoscale designs [11]. BTI gradually increases the threshold voltage of a transistor over a long period of time, which in turn increases the gate delay [3]. BTI reduces the SNM of an SRAM cell and makes it more susceptible to failures. SNM is the minimum amount of DC noise that leads to a loss of the stored value. BTI-induced SNM degradation is higher when the cell stores the same value for a long period of time (e.g., storing '1' at node 'A' of the SRAM cells shown in Figures 3(a) and 3(b)). Hence, the effect of BTI on an SRAM cell is a strong function of the cell Signal Probability (SP)[1].

2) ***Process variation effects****:* Process variation is a manufacturing mismatch of transistor parameters, such as threshold voltage from their designed values. Random Dopant Fluctuation (RDF) and Line Edge Roughness (LER) are the main sources of variation in nanoscale devices. The effect of process variation is more pronounced in storage elements at lower voltage levels [6]. Hence, in NTC designs, SRAM cells are highly affected by variation effects.

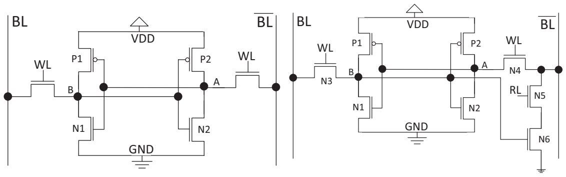
3) ***Radiation-induced soft errors****:* The primary source of soft errors is related to cosmic ray events. Atmospheric neutrons are one of the higher flux components and their reaction have a high energy transfer. Thus, neutrons are the most likely cosmic radiations to cause soft errors. Neutrons do not generate electron-hole pairs directly. However, their interaction with the Si-atoms generate secondary particles. These secondary particles produce electron-hole pairs which can result in a soft error.

Radiation-induced SER of an SRAM cell is increasing significantly with decrease in the supply voltage. Previous experiments have shown that the radiation-induced soft error rate can increase by 50% for a 20% decrease in supply voltage [12]. Moreover, the soft error rate of NTC designs is affected by aging-induced SNM degradation and process variation effects.

B. Related Works

With the increase in reliability challenges, various researchers have focused on developing mitigation schemes to address BTI, soft errors and process variation. In the super threshold voltage regime, several works are available in the literature to address these reliability issues independently such as [3, 4, 13], to only name a few. In NTC, however, most of the studies address performance loss and variability by design techniques such as multiple V_{dd} [7].

Regarding the interdependence of the reliability failure mechanisms, however, there is quite limited attempt to address the interdependence of failure mechanisms in the super threshold regime. For example, studies in [8, 10] discuss the impact of aging and process variation effects in soft error rate of an SRAM cell. In NTC dealing with failure mechanisms independently can underestimate the overall reliability impact by more than 2X. Hence, analyzing the combined effect of the failure mechanisms across several abstraction levels is of decisive importance for reliable NTC operation. This is the area where our cross-layer reliability analysis framework can play a big role to get the utmost NTC benefits during early design phases.



(a) 6T SRAM cell (b) 8T SRAM cell

Fig. 3. Schematic diagram of 6T and 8T SRAM cells

III. Holistic cross-layer Reliability Estimation Framework

Figure 4 illustrates our holistic cross-layer reliability estimation framework that abstracts the impact of workload, cache organization and different reliability failure mechanisms at different levels of abstraction. The reliability analysis and simulation conducted in this work use the symmetric six-transistor (6T) and 8T SRAM cells shown in Figure 3. In this work, the device-level critical charge characterization was modeled according to the model presented in [14].

In this section, we will discuss the proposed cross-layer reliability estimation framework in a top-down manner. The system-level *Failure In Time* (FIT) rate and SNM extraction is described in Section III-A followed by the cross-layer SNM and SER estimation in Section III-B.

A. 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 *Architectural Vulnerability Factor* (AVF). Equation (1) shows the system-level FIT rate calculation.

$$FIT_{system} = \sum_{i=0}^{N} AVF_i \times SER_i \tag{1}$$

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

Architecture level AVF and SNM analysis*:* One step of determining the failure rate of a memory (cache) due to soft errors is to determine the AVF value of the memory. The AVF of a memory is the probability that a fault (flip) in the memory cells will propagate down to the data path [15]. 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 cycles to the total number of operational cycles [16]. The computation of AVF value of a memory array with M cells is expressed in Equation (2).

$$AVF_{array} = \frac{\sum_{i=0}^{M} ACE_i}{T \times M} \tag{2}$$

Where T is the total number of cycles.

The SNM degradation of an SRAM cell strongly depends on the signal probability of the cell. The BTI-induced SNM degradation is minimized when the signal probability of the cell is close to 0.5. To determine the SNM degradation, the worst case SP of the memory row is obtained as the maximum SP distance from 0.5 (D = $|SP - 0.5|$) as shown in Equation (3):

$$SP_{worst-case} = MAX_{i=1}^{Z} D_i \tag{3}$$

[1] Probability of storing logic '1' in the SRAM cell

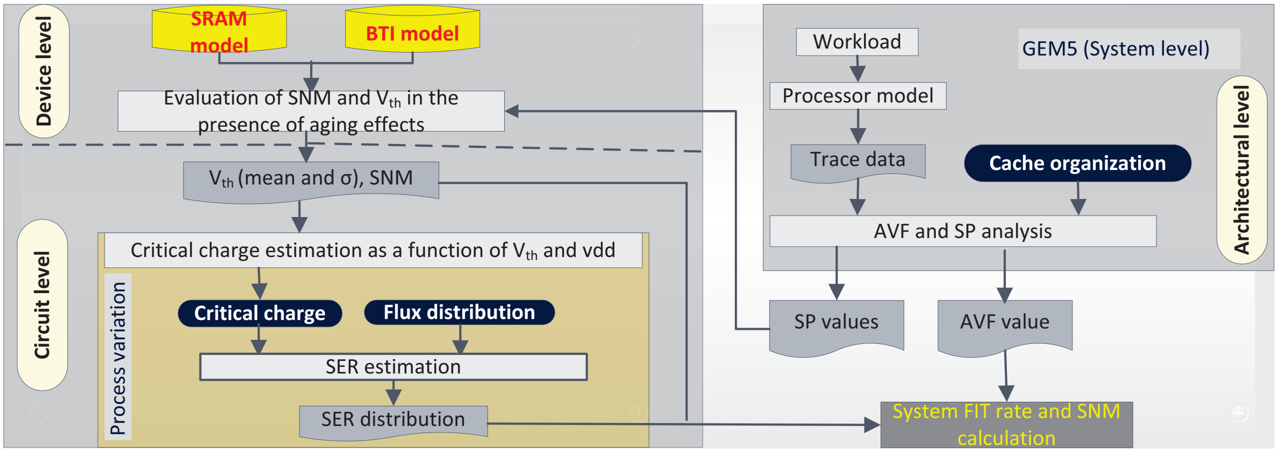


Fig. 4. Holistic cross-layer reliability estimation framework to analyze impact of aging and process variation effects on soft error rate

where D$_i$ = $|SP_i - 0.5|$ and Z is the total number of cells in the memory row.

To extract the AVF and SNM of a cache unit, first we extract the trace of the data stored in the cache, read-write accesses and the clock period of the running workload. Once all these information are available, our reliability analysis tool will use these information along with the cache organization to determine the AVF according to Equation (2) and SNM from the SP to SNM Look-Up Table(LUT).

The cache organization (size and associativity) has a big impact on the AVF and SNM of the cache, as it determines the hit ratio and the duration data can be stored in the cache entry. Hence, different cache size and associativity combinations can result in different AVF and SNM values for the same application. Additionally, AVF and SNM are highly dependent on the running workload. To explore this issue, various cache organizations and workloads are investigated in this work. From our study we identify and propose a reliability and performance optimal cache organization for the near and super threshold voltage regimes.

B. Cross-Layer SNM and SER Estimation

*1) **SNM estimation**:* The SNM of an SRAM cell is extracted by conducting a circuit level SPICE simulation. The SPICE simulation uses the BTI model, device level parameters and architecture level signal probability values to determine the SNM of the SRAM cell after 3 years of operation. Finally, the SNM degradation of a particular SP value is obtained according to Equation (4).

$$DEG_{SP} = \frac{SNM_{SP} - SNMfresh}{SNM_{fresh}} \times 100\% \qquad (4)$$

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

*2) **SER estimation**:* The SER of an SRAM cell depends on two main factors, the critical charge of the cell and the flux rate.

***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 flip 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 cross-coupled inverters [17]. Q$_{critical}$ can be computed using analytical models or circuit simulators. In this work we use the analytical model developed in [14] to determine Q$_{critical}$ of an SRAM cell.

As shown in Figure 4, 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 signal probability (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 workloads is incorporated in our analysis framework. Once the fresh and aged V$_{th}$ values are available, process variation (i.e., the main NTC challenge) is incorporated as a normal distribution ($\mu \pm 3\sigma$) of the transistor threshold voltage where $\mu = V_{th}$ and σ is obtained using Equation (5) [18]. All these information will be used by the critical charge model to extract the corresponding Q$_{critical}$ values.

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

where, L and W are the length and width and A$_{VT}$ is process specific parameter (the ”pelgrom coefficient”), dependent on $\sqrt{L \times W}$.

***Circuit-level SER analysis**:* The circuit-level SER analysis is conducted using the SER extraction tool depicted on the left side of Figure 4. First, the critical charge of the SRAM cell is extracted using a device-level model. Afterwards, the critical charge along with the neutron-induced flux distribution is used to determine the SER of the cell according to Equation (6) [19]. As shown in Equation (6), the SER of an SRAM cell has an exponential relation with the Q$_{critical}$. Hence, the higher the Q$_{critical}$ the lower the SER will be.

$$SER \propto FAe^{(-\frac{Q_{critical}}{Q_s})} \qquad (6)$$

where F is the neutron flux in particles/cm^2-s with energy greater than 1MeV [20]; A is the area sensitive to a strike, in cm^2; Q$_{critical}$ is the critical charge and Q$_S$ is the charge collection efficiency.

From Equation (6), we can make the following observations:

- The SER of an SRAM cell has exponential relation to its critical charge. Hence, a small decrease in Q$_{critical}$ lead to an exponential increase in SER of the cell.
- Since atmospheric neutrons are an external factors, a small drift in Q$_{critical}$ can leads to a significant increase in the amount of flux. Furthermore, transistor up-sizing will increase the area which is sensitive to soft errors.

IV. Experimental Results

In this section, the experimental setup used in our holistic cross-layer reliability analysis is presented first. Afterwards, the obtained reliability and performance results are discussed.

A. Experimental Setup

For evaluation purpose, we use an ALPHA implementation of an embedded in-order core on the Gem5 [21] performance simulator.

TABLE I. EXPERIMENTAL SETUP

Configuration	Near threshold	Super threshold
Processor model	Embedded	Embedded
Architecture	Single in-order core	Single in-order core
Supply voltage	0.5V	1.1V
Frequency	100MHz	1GHz
Technology node	45nm PTM	45nm PTM
Cache size	4, 8 and 16 KByte	4, 8 and 16 KByte
Associativity	Direct mapped - 4-way	Direct mapped - 4-way
Benchmark	SPEC2000	SPEC2000

Various cache sizes (4KByte-16KByte) and wide associativity range from simple directly mapped to 4-way set associative caches are assessed to perform a reliability-performance tradeoff. Four workloads (*applu, bzip2, lucas and parser*) from the SPEC2000 benchmark suite [22] were executed for 5 million cycles on the simulated processor model to analyze the impact of workload. The experimental setup used in this work is presented in Table I.

To extract the BTI-induced V_{th} shift, we assume 10% BTI-induced aging after 3 years of operation. First, 45nm 6T and 8T SRAM cells are modeled using PTM model. Afterwards, we conduct SPICE simulation to extract the SNM and V_{th} shift Look-Up Table (LUT) for various SP values (0.0-1.0). In this work the impact of process variation is considered as a normal distribution of the transistor threshold voltage with a mean ($\mu = V_{th}$, 0.3V) and standard deviation (σ) obtained using the model given in Equation (5).

To demonstrate the effect of soft error, neutron induced soft errors are considered in this work as they are the dominant soft error mechanisms at terrestrial altitudes. Our results are according to the setup given in Table I and transistor sizing specified in [23].

B. SNM and SER Analysis of 6T and 8T SRAM Cells

*1) **Impact of aging and process variation on SNM**:* BTI-induced SNM degradation of an SRAM cell depends not only on SP but also on process parameters such as threshold voltage, which are highly affected by manufacturing variabilities. Though aging has less impact on SNM degradation in near threshold designs, as the temperature is low. In combination with process variation induced v_{th} shift, however, it can degrade the SNM of an SRAM cell significantly.

Figure 5 shows the worst case aging (SP=0.0) after 3 years and process variation induced SNM degradation of 6T and 8T SRAM cells, confirming our expectation about the significant SNM degradation in NTC which is 2.5X higher than the degradation in the super threshold domain (as shown using gray boxes). While the use of 8T instead of 6T SRAM cells in super threshold regions has limited difference in SNM degradation (only 7.7%), this difference is significant (14.7%) in NTC.

*2) **SER of 6T and 8T SRAM cells**:* In the conventional 6T SRAM cell (Figure 3(a)), the cell must be stable during read access. For this purpose, either read-write assist circuitries should be employed or the pull-down (NMOS) transistors of the cross-coupled inverters should be strengthened by transistor sizing [1]. However, the sizing will also increase the area of the cell sensitive to soft errors. In low voltage designs, the 6T SRAM cell is more sensitive to read-disturb such that transistor sizing cannot handle it. This makes the 6T design less preferable for near threshold designs.

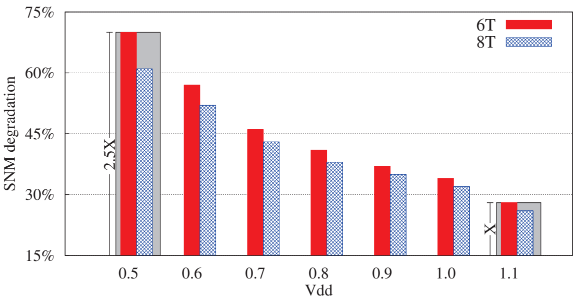


Fig. 5. SNM degradation in the presence of process variation and aging after 3 years of operation

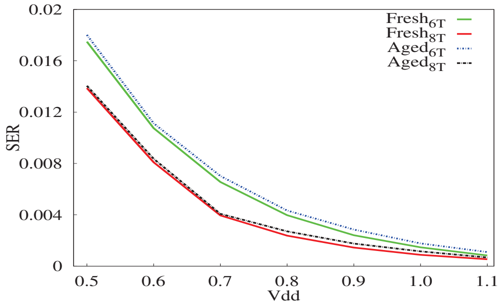


Fig. 6. SER rate of fresh and aged 6T and 8T SRAM cells for various vdds

To address this issue, alternative SRAM designs (such as 8T [23] and 10T [24] SRAM cells) are recommended for NTC. For this regard, the 8T SRAM cell design (see Figure 3(b)) is studied in this work. The read-disturb issue is solved in the 8T design by decoupling the read and write lines using two additional NMOS access transistors. Hence, the size of the pull-down transistors is decreased to reduce the area sensitive to soft errors. The transistor sizing specified in [23] is used for the design of the 6T and 8T SRAM cells.

The critical charge of the 6T and 8T designs is almost the same. However, due to the increase in the size of the cross-coupled inverters for read stability, the 6T SRAM cell is more vulnerable to soft errors. Hence, the SER of the 6T cell is higher than the 8T cell in NTC. Figure 6 shows the fresh and aged soft error rate of a 6T and 8T SRAM designs for different supply voltage levels. At super threshold voltage (0.9V-1.1V) the 6T and 8T designs have negligible SER difference. In NTC, however, the 6T design has higher SER than the 8T design due to the effects of transistor 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 Figure 7. From Figure 7, we observe that in near threshold voltages the SER of the 6T and 8T designs is 4X higher than the SER in the super threshold voltage which shows the severity of sensitivity to variation effects in NTC designs. Due to its smaller size the 8T design is highly affected by process variation. However, it is more reliable than the 6T design.

The circuit level SER of the 6T and 8T designs is used along with the architectural level AVF to determine the system level FIT rate of the cache. Figure 8 shows the system level FIT rate of different cache organizations in NTC by considering average aging impact of the four workloads.

*3) **Reliability improvement and area overhead analysis**:* In a near threshold SRAM design, the 8T design improves the soft error rate in the presence of aging and variation effects by up to 25%. Similarly,

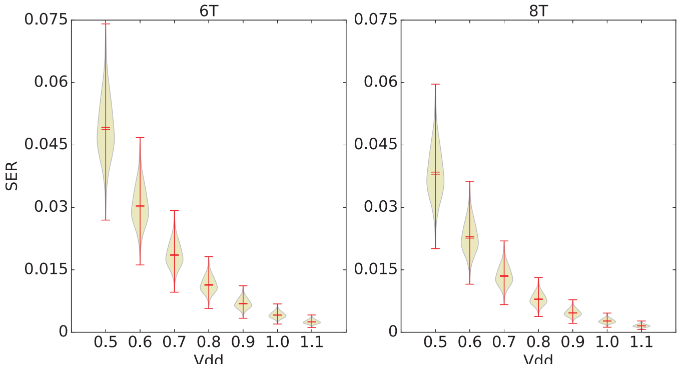


Fig. 7. SER of 6T and 8T SRAM cells in the presence of process variation and aging effects after 3 years of operation

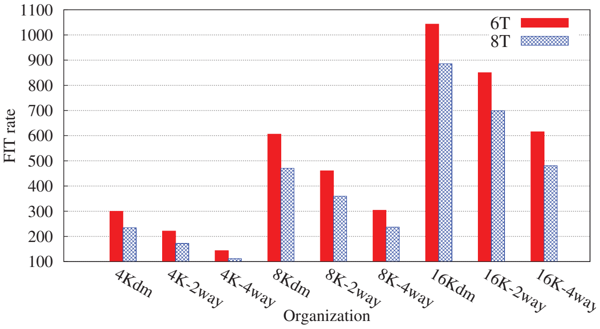


Fig. 8. System level FIT rate comparison of aged 6T and 8T SRAM in NTC

the SNM can be 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, this overhead is much less. Since the 6T SRAM has to be up-sized to increase its read stability in NTC, this will increase the cell area of the 6T design to be larger than the area of 8T design as experimentally demonstrated in [23].

C. Cache Organization Impact on System FIT Rate

Cache organization has a big impact on the performance of embedded processors [25]. Similarly, the organization has an impact on the reliability of the cache units. In NTC, the reliability impact of the 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 highly depends on the architectural vulnerability factor and the values stored in the cache as well as their residency time intervals which is in turn highly dependent on the read-write accesses of the cache. Hence, these parameters are influenced by cache size and associativity.

To evaluate the performance and reliability impact of different cache organizations in the near and super threshold voltage regimes we use the configurations described in Table I. For near threshold (0.5V) the processor core frequency is set to 100MHz (10X slower than the super threshold regime [2]). Since gate delay is the dominant factor in near threshold domain, the cache latency is set to 1 cycle [26]. In the super threshold domain, however, the cache latency and interconnect delay has significant impact on the overall delay. Thus, the cache hit latency is set to 2 cycle for 4K and 8K caches and 3 cycles for 16K cache [27].

*1) **Cache organization and SNM degradation**:* Since cache organization determines the residency time of a data, it has a direct impact on the SNM degradation. Figure 9 illustrates the impact of cache organization on SNM degradation of near and super threshold 6T and 8T memory arrays in the presence of process variation and aging after 3 years of operation. From the figure we can observe that smaller cache size with higher associativity (4k-4w) has less impact on SNM degradation as the data residency time is smaller.

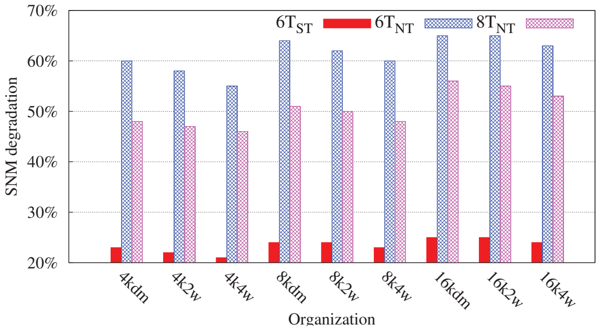


Fig. 9. Impact of cache organization on SNM degradation in near threshold (NT) and super threshold (ST) in the presence of process variation and aging effect after 3 years of operation

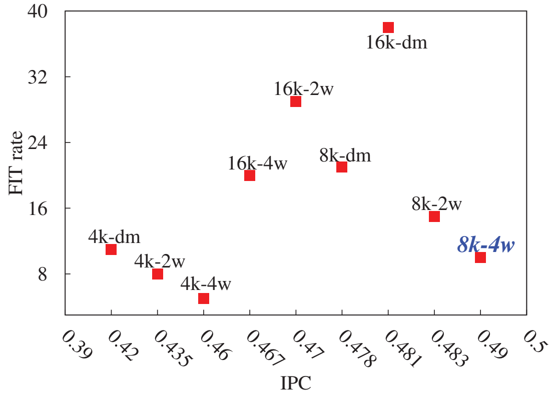


Fig. 10. FIT rate-performance design space of various cache configurations in super threshold voltage (the *blue italic font* indicates optimal configuration)

*2) **Cache organization and SER FIT rate**:* The cache size and associativity determines the read-write accesses of the cache lines. which in turn affects the ACE cycles of a cache line and its failure probability. The impact of the cache organization on FIT rate and performance (in terms of committed Instructions Per Cycle (IPC)) varies along various supply voltage domains. At the super threshold, an increase in cache size and associativity improves the performance. However, from FIT rate point of view, an increase in cache size has a negative impact due to the increase in the AVF of the cache. Smaller cache sizes, however, have lower performance and better FIT rate. Figure 10 shows the design space of FIT rate and performance (in terms of IPC) impact of various cache organizations in the super threshold voltage. In the figure the FIT rate-performance optimal configuration is (8k-4w) as indicated by bold italic font.

In the near threshold regime the performance is mainly dominated by the delay of the logic unit and failure rate is significantly high, it is important to select a cache organization that gives better reliability (FIT rate and SNM) than performance. Hence, in NTC a smaller cache size with a higher associativity gives the best reliability-performance tradeoff. Figure 11 shows the design space for the FIT rate-performance tradeoff for 6T and 8T designs in NTC. The reliability impact of the cache organization is more pronounced in the presence of aging and process variation effects. Figure 12(a), shows the FIT rate-performance plot in the presence of aging and process variation effects in the super threshold regime. Figures 12(b) and 12(c) show the FIT rate-performance plot of 6T and 8T designs in NTC. From the figures we can observe that the 8T design has better FIT rate for the same performance.

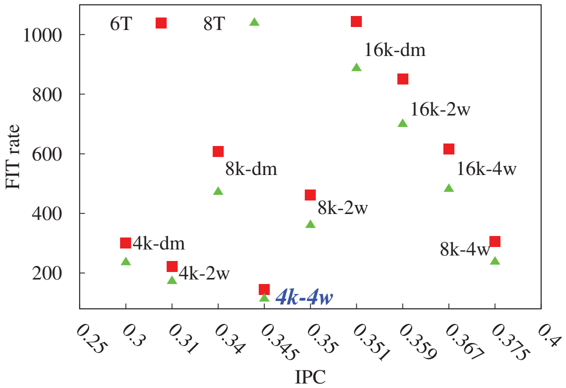
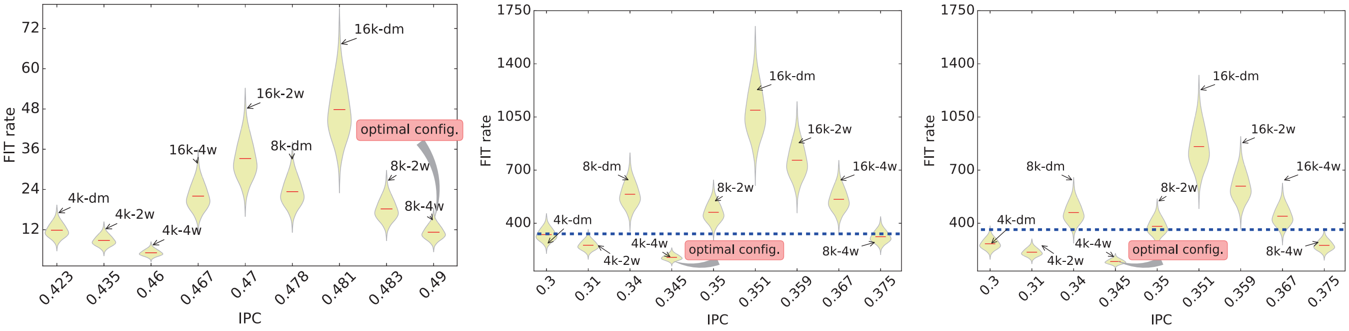


Fig. 11. FIT rate-performance design space of 6T and 8T designs for various cache configurations in near threshold voltage (the *blue italic font* indicates optimal configuration)



(a) Super threshold (6T SRAM)

(b) Near threshold 6T

(c) Near threshold 8T

Fig. 12. FIT rate-performance tradeoff for various cache configurations in the presence of process variation and aging effects

*3) **Supply voltage aware optimal cache organization**:* From our experimental results reported in Figures 9, 10, 11 and 12 we observe that an increase in the cache associativity improves the performance and reliability (both FIT rate and SNM). Hence, in the super threshold domain, medium cache size (e.g., 8 KByte) with higher associativity gives better reliability-performance tradeoff. In the NTC domain, however, smaller cache sizes with a higher associativity are preferable due to two main reasons: 1) The performance is mainly dominated by the processor core not by the cache units. 2) The soft error rate and SNM degradation are higher in NTC than in the super threshold regime. Hence, the cache size can be reduced by half to obtain a better reliability-performance tradeoff in NTC.

In the NTC domain, the selection of an optimal cache organization can be different for 6T SRAM cell based caches from the 8T caches, depending on the FIT rate and performance requirement. For example, if we assume that a FIT rate of 350 is tolerable in NTC (as shown by the dotted line in Figures 12(b) and 12(c)), then only 4 KByte 4-way associative cache is within the acceptable zone for 6T design. In the 8T design, however, three additional cache organizations (4K-dm, 4k-2w and 8k-4w) are within the acceptable zone. Hence, the 8k-4w cache can be used in the 8T design to get ≈10% performance improvement without violating the reliability constraint.

To implement the suggested cache organizations for a specific supply voltage regime (only near threshold or super threshold) is straightforward. In a design that is expected to operate in both super and near threshold regions, the cache can be designed according to the super threshold voltage (e.g., 4-way 8 KByte in our case). Then when switching to a near threshold region, half of the cache can be disabled (power gated) or can be used for error checking [28].

V. Conclusions

In this paper, we studied the combined impact of aging, process variation and soft error on the reliability of cache memories in super and near threshold voltage regimes. We observe that, the combined effect of process variation and aging has huge impact on the soft error rate and SNM degradation in NTC. Experimental results show that, process variation and aging-induced SNM degradation is 2.5X higher in NTC than in the super threshold domain while SER is 8X higher. The use of 8T instead of 6T SRAM cells can reduce 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 a memory structure. We show that the cache organization changes when going from the super to near threshold voltage domain. Hence, smaller caches size with higher associativity give better reliability-performance tradeoff in the NTC.

Acknowledgments

This work was partly supported by the German Research Foundation (DFG) as part of the national focal program Dependable Embedded Systems (SPP-1500, http://spp1500.ira.uka.de)

References

[1] M. Seok *et al.*, “Cas-fest 2010: Mitigating variability in near-threshold computing,” *Emerging and Selected Topics in CS, IEEE Journal on*, 2011.
[2] R. G. Dreslinski *et al.*, “Near-threshold computing: Reclaiming moore’s law through energy efficient integrated circuits,” *Proceedings of the IEEE*, 2010.
[3] A. Gebregiorgis *et al.*, “Aging mitigation in memory arrays using self-controlled bit-flipping technique,” in *ASP-DAC*, 2015.
[4] V. Huard *et al.*, “From bti variability to product failure rate: A technology scaling perspective,” in *IRPS*, 2015.
[5] D. Heidel *et al.*, “Single-event-upset critical charge measurements and modeling of 65 nm silicon-on-insulator latches and memory cells,” *IEEE, nuclear science*, 2006.
[6] Q. Ding *et al.*, “Impact of process variation on soft error vulnerability for nanometer vlsi circuits,” in *ASIC, 2005. ASICON 2005. 6th International Conference On*, 2005.
[7] U. R. Karpuzcu *et al.*, “Coping with parametric variation at near-threshold voltages,” *Micro, IEEE*, 2013.
[8] M. Bagatin *et al.*, “Impact of nbti aging on the single-event upset of sram cells. nuclear science,” *IEEE Transactions on*, 2010.
[9] H. Amrouch *et al.*, “Towards interdependencies of aging mechanisms,” in *ICCAD*, 2014.
[10] E. H. Cannon *et al.*, “The impact of aging effects and manufacturing variation on sram soft-error rate,” *Device and Materials Reliability*, pp. 145–152, 2008.
[11] K. K. Kim *et al.*, “On-chip aging sensor circuits for reliable nanometer mosfet digital circuits,” *Circuits and Systems I*, 2010.
[12] J. Tonfat *et al.*, “Analyzing the influence of voltage scaling for soft errors in sram-based fpgas,” in *RADECS*, 2013.
[13] M. Ebrahimi *et al.*, “Comprehensive analysis of alpha and neutron particle-induced soft errors in an embedded processor at nanoscales,” in *DATE*, 2014.
[14] S. M. Jahinuzzaman *et al.*, “An analytical model for soft error critical charge of nanometric srams,” *VLSI*, 2009.
[15] V. Sridharan *et al.*, “Using hardware vulnerability factors to enhance avf analysis,” in *ACM SIGARCH Computer Architecture News*, 2010.
[16] M. Wilkening *et al.*, “Calculating architectural vulnerability factors for spatial multi-bit transient faults,” in *International Symposium on Microarchitecture*, 2014.
[17] J. M. Cazeaux *et al.*, “On transistor level gate sizing for increased robustness to transient faults,” in *IOLTS. 11th IEEE International*, 2005.
[18] K. J. Kuhn *et al.*, “Process technology variation,” *Electron Devices*, 2011.
[19] P. Hazucha *et al.*, “Impact of cmos technology scaling on the atmospheric neutron soft error rate,” *Nuclear Science, IEEE Transactions on*, 2000.
[20] J.-L. Autran *et al.*, *Soft-error rate of advanced SRAM memories: Modeling and monte carlo simulation*. INTECH Open Access Publisher, 2012.
[21] N. Binkert *et al.*, “The gem5 simulator,” *ACM Computer Architecture News*, 2011.
[22] J. L. Henning, “Spec cpu2000: Measuring cpu performance in the new millennium,” *Computer*, 2000.
[23] Y. Morita *et al.*, “An area-conscious low-voltage-oriented 8t-sram design under dvs environment,” in *VLSI Circuits*, 2007.
[24] K. Tae-Hyoung *et al.*, “A high-density subthreshold sram with data-independent bitline leakage and virtual ground replica scheme,” in *Solid-State Circuits*, 2007.
[25] O. Olorode *et al.*, “Improving performance in sub-block caches with optimized replacement policies,” *ACM Journal on JETC*, 2015.
[26] H. Chen *et al.*, “Opportunistic turbo execution in ntc: exploiting the paradigm shift in performance bottlenecks,” in *DAC*, 2015.
[27] “Understanding cpu caching and performance,” http://http://arstechnica.com/gadgets/2002/07/caching/2/, accessed: 2015-09-28.
[28] A. BanaiyanMofrad *et al.*, “Protecting caches against multi-bit errors using embedded erasure coding,” in *ETS*, 2015.