Parametric and Functional Degradation Analysis of Complete 14-nm FinFET SRAM

Kraak, Daniël; Taouil, Mottagiallah; Agbo, Innocent; Hamdioui, Said; Weckx, Pieter; Cosemans, Stefan; Catthoor, Francky

DOI
10.1109/TVLSI.2019.2902881

Publication date
2019

Document Version
Final published version

Published in
IEEE Transactions on Very Large Scale Integration (VLSI) Systems

Parametric and Functional Degradation Analysis of Complete 14-nm FinFET SRAM

Daniel Kraak, *Student Member, IEEE*, Mottaqiallah Taouil, *Member, IEEE*, Innocent Agbo, *Member, IEEE*, Said Hamdioui, *Senior Member, IEEE*, Pieter Weckx, *Member, IEEE*, Stefan Cosemans, *Senior Member, IEEE*, and Francky Catthoor, *Fellow, IEEE*

***Abstract*— Designers typically add design margins to compensate for chip aging. However, this leads to yield loss (in case of overestimation) or low reliability (in case of underestimation). This paper analyzes the impact of aging on a complete high-performance industrial 14-nm FinFET SRAM. It investigates the impact on the memory's parametric (i.e., its delay) and functional (i.e., correct functionality) metrics. Moreover, it examines which components are the main contributors to the degradation of the memory's reliability and how it is impacted by workload and environmental conditions, i.e., temperature and voltage fluctuations. This paper not only investigates the impact of the memory's components individually, which is typically the case in prior work, but it also studies the contribution of components' interaction to the overall memory aging. The results show that the timing circuit, address decoder, and the output latches and buffers are the main contributors to the memory's parametric degradation, while the cell, sense amplifier, and address decoder are the main contributors to its functional degradation. Moreover, the results show that it is crucial to consider the impact of the interaction of components on the aging; individual analysis leads to overly pessimistic results and even wrong conclusions in certain cases.**

***Index Terms*— Aging, bias temperature instability (BTI), FinFET, reliability, SRAM.**

I. Introduction

THE aggressive downscaling of CMOS technology has significantly improved the performance and functionality of integrated circuits (ICs) over the past decades. However, this downscaling has resulted in reliability challenges due to the increased impact of time-zero and time-dependent variabilities [1], [2]. Time-zero variation (or process variation) is caused by imperfections during the production. As a result, circuits have deviating characteristics from the intended ones. Time-dependent variations are variations that occur during the lifetime of the ICs. They include environmental variations, such as supply voltage and temperature fluctuations, and aging variations due to, for instance, bias temperature instability (BTI) [3], [4]. In order to achieve a high quality product in terms of low failure rate and long lifetime at optimal design, it is crucial to estimate the impact of these variabilities for critical electronic components. In this paper, we focus on estimating the degradation of SRAM. Embedded SRAMs are optimized in terms of area, power, and performance as they often dominate the total area of SoCs and microprocessors. Therefore, optimizing the resources (and hence the area, power, performance) to enable the proper level of reliability is crucial.

Previous studies on SRAM reliability have mainly focused on the impact of aging on the memory cell array [5]–[10]; only limited work exists on the degradation of the memory's peripheral circuitry; examples are the sense amplifier (SA) [11], [12] and timing circuit [13]. The published works evaluate the circuits under investigation using appropriate metrics; for instance, noise margins for the memory cell array, sensing delay for the SA, and so on. Nevertheless, estimating and understanding the aging of each individual component do not mean that the impact on the overall memory reliability is straightforward to identify; first of all, because it is difficult to estimate the effects of the degradation of individual components on the operation of the memory as a whole; second, the degradation of a component will be affected by the degradation of other components. Hence, it is crucial to consider the aging of the whole memory system. Only two publications [14], [15] tried to address this limitation by investigating the impact of aging while considering the interaction between multiple components or the whole memory. Agbo *et al.* [14] investigated the impact of aging on the interaction between the memory cell and SA. However, the analysis does not include other important circuits such as the timing circuit and address decoder. On the other hand, Kinseher *et al.* [15] investigated the effects of aging on a complete 28-nm SRAM circuit. However, they did not examine how the interaction between components contributes to their mutual degradation and they did not explore the stochastic nature of aging. In our previous work, we investigated the impact of aging on a high performance, state-of-the-art 14-nm FinFET complete

Manuscript received October 8, 2018; revised January 10, 2019; accepted February 8, 2019. Date of publication April 1, 2019; date of current version May 22, 2019. This work was supported in part by the TRACE Project (CATRENE) under Grant 16ES0488K-16ES0502 and Grant 16ES0737 and in part by the PRYSTINE Project (ECSEL) under Grant 783190. *(Corresponding author: Daniel Kraak.)*

D. Kraak, M. Taouil, I. Agbo, and S. Hamdioui are with the Department of Computer Engineering, Delft University of Technology, 2628CD Delft, The Netherlands (e-mail: d.h.p.kraak@tudelft.nl; m.taouil@tudelft.nl; i.o.agbo@tudelft.nl; s.hamdioui@tudelft.nl).

P. Weckx and S. Cosemans are with imec, 3001 Leuven, Belgium (e-mail: pieter.weckx@imec.be; stefan.cosemans@imec.be).

F. Catthoor is with imec, 3001 Leuven, Belgium, and also with ESAT, Katholieke Universiteit Leuven, 3001 Leuven, Belgium (e-mail: francky.catthoor@imec.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2019.2902881

SRAM using a calibrated atomistic BTI model [16], [17]. We examined the degradation of one overall memory metric (i.e., read access time) and several individual component metrics (e.g., SA's sensing delay). In addition, we investigated how the aging of components affects each other. However, the work did not analyze which components are the main contributors to the overall degradation of the memory. Furthermore, the work only investigated the degradation of metrics related to the read operation of the memory.

In this paper, we analyze how aging impacts the parametric as well as the functional reliability of the memory; we consider the impact on all relevant memory metrics related to both read and write operations. Moreover, we investigate which components are the main contributors to the memory's degradation. In short, the contributions of this paper are as follows.

1) The impact of aging on the memory's parametric and functional metrics is studied.
2) The impact of component, workload, temperature, and voltage fluctuations on the memory's parametric and functional degradation is analyzed.
3) All the above was done by using a full 14-nm FinFET high-performance SRAM design based on realistic "industrial-strength" circuit design and a calibrated aging model for accurate simulation results.

The outline of this paper is as follows. Section II presents the simulation setup. Section III discusses the used methodology and the performed experiments. Sections IV–VI evaluate the experiments related to the component sensitivity, workload sensitivity, and the environmental conditions sensitivity, respectively. Section VII provides the discussion. Finally, Section VIII concludes this paper.

II. Simulation Setup

This section first discusses the memory model used for this paper. Subsequently, it discusses the evaluated metrics and the used aging model.

A. Memory Model

The memory model used as a case study for this paper is a high-performance SRAM from imec, implemented in FinFET 14-nm technology. The memory has a size of 1 kB and a logical word length of 32 bits. It is able to run at a frequency of 2 GHz under worst case conditions. Such a fast and compact memory is typically deployed in L1 caches. Fig. 1 shows a block diagram of the memory with the relevant components and signals. The functionality and specifications of these components are as follows.

1) *Input flip-flops*: The input flip-flops serve as the synchronous interface to the memory. The address signal holds the memory address for the read/write operation. The enable signal specifies whether the memory should perform an operation or remain idle. The r/w signal selects the type of operation (read or write).
2) *Cell array*: The cell array consists of 128 rows by 64 columns. It is implemented using high-performance cells with a 1:2:2 ratio (1 fin for the pull-up transistor, 2 fins for the pass gate as well as for the pull-down).

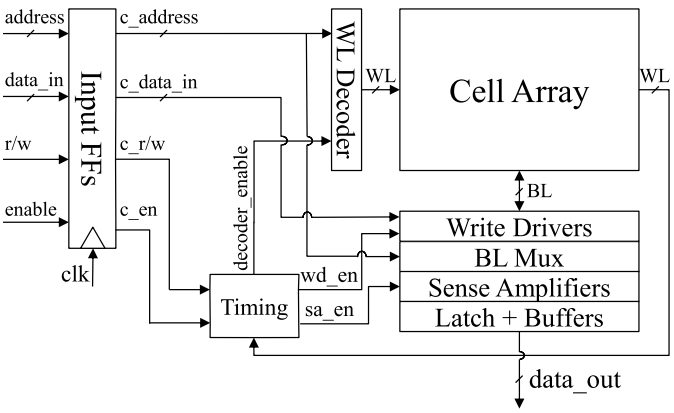


Fig. 1. Block diagram of memory.

3) *WL decoder*: The wordline (WL) decoder selects the appropriate row in the cell array based on the input address.
4) *BL Mux*: The bitline multiplexer (BL Mux) selects the appropriate columns in the cell array based on the input address.
5) *Timing*: The timing circuit is the main control circuit of the memory. It provides timed control signals to the other components during the operation (e.g., enabling the WL decoder). It uses a WL detect scheme, i.e., it is able to detect the activation of one of the WLs, as illustrated by the feedback loop in the figure.
6) *Sense amplifiers*: The SAs amplify the voltage difference on the BLs to full-swing read values. They are implemented using latch-type SAs [18]. Each of the two critical pull-down transistors is implemented using 16 fins in order to get a sufficiently low offset voltage.
7) *Output latches*: The output latches hold the read value. They are controlled by the outputs of the SAs and are implemented using SR latches.
8) *Output buffers*: the output buffers convert the weak outputs of the output latches into strong output signals. They are implemented using two cascaded inverters.
9) *Write drivers*: The write drivers write the values of the selected memory cells by driving one of the complementary BLs to a low value while keeping the other one high.

In case the memory performs a read operation, the timing circuit is activated by the enable signal and the rising edge of the clock. Meanwhile, the selected address is applied to the WL decoder and BL Mux to select the appropriate rows and columns in the cell array. Next, the timing circuit activates the WL decoder using the decoder_enable signal. Note that the timing circuit allocates sufficient time before activating the decoder to ensure all logical paths in the WL decoder have finished propagating when a new address is selected. Once the timing circuit detects that the WL has been activated using its WL detect scheme, it keeps the WL activated for a certain amount of time in order to give the selected memory cells sufficient time to discharge their BLs. Finally, the timing circuit activates the SA_enable signal to

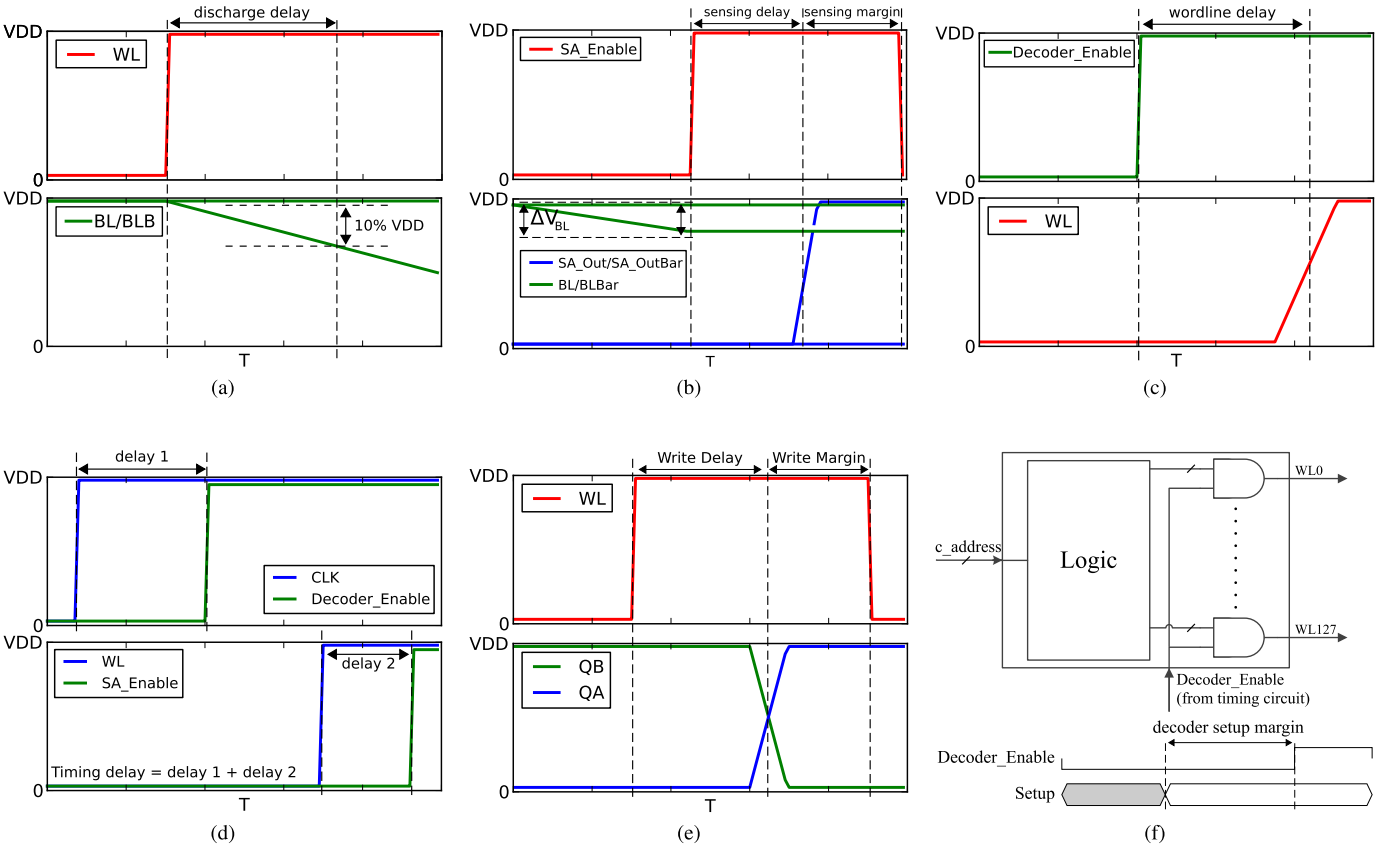


Fig. 2. Definitions of the used metrics. (a) Metric discharge delay. (b) Metrics sensing delay, sensing margin, and BL swing (ΔV_{BL}). (c) Metric WL delay. (d) Metric timing delay. (e) Metrics write delay and write margin. (f) Metric decoder setup margin.

enable the SAs, which then amplify the voltage difference on the BLs to full-swing read values.

When, instead, the memory performs a write operation, the timing circuit activates the WL decoder using the decoder_enable signal. Subsequently, the timing circuit activates the write drivers using the WD_enable signal. The timing circuit detects the WL activation using its WL detect scheme and keeps the WL activated long enough to ensure that the write drivers have sufficient time to write to the cells. Finally, the timing circuit deactivates the WL and write drivers.

B. Metrics

The analyzed metrics are divided into parametric and functional metrics [19]. The parametric metrics evaluate the performance of memory characteristics that do not directly impact the correct functionality of the memory. Examples include delay and power consumption. In this paper, we focus on the delay aspect only. The functional metrics evaluate whether the memory performs its function without any failures. The requirement for this is that the memory generates the correct output. An example of a failure is the flipping of an unstable memory cell.

1) Parametric Metrics: The parametric metrics are divided into overall metrics and individual metrics. The overall metrics evaluate the performance of the whole system, while the individual metrics evaluate the performance of individual components.

a) Overall: For the parametric metrics, we use the memory's read and write access times. The read access time is the delay between the start and end of the read operation. It is measured as the time between the rising edge of the clock and the data appearing at the memory's output. Similarly, the write access time is measured as the time between the rising edge of the clock and the transition of the cell value (i.e., when the voltages of the cell's inner nodes cross).

b) Memory cell: We use the discharge delay to evaluate the memory cell's parametric performance. It is illustrated in Fig. 2(a). In the figure, WL presents the WL signal and BL/BLBar the complementary BLs. The discharge delay is defined as the time between the activation of the WL and one of the BLs being discharged by 10% with respect to the supply voltage value.

c) SA: We use the metric sensing delay to evaluate the SA's parametric performance. Fig. 2(b) illustrates the sensing delay. In the figure, SAenable presents the enable signal of the SA, BL/BLBar the differential BL pair connected to the accessed cell, and SA_Out/SA_OutBar the differential output signal of the SA. The sensing delay is defined as the time between the enabling of the SA and its outputs being ready.

d) Address decoder: We use the metric WL delay to evaluate the decoder's parametric performance. Fig. 2(c) illustrates the WL delay. In the figure, Decoder_Enable presents the enable signal of the decoder coming from the timing circuit and WL presents the WL signal. The WL delay is defined as

the time between the enabling of the decoder and the WL signal being high.

e) Timing: We use the metric timing delay to evaluate the timing circuit's parametric performance, as shown in Fig. 2(d). In the figure, CLK represents the input clock signal, Decoder_Enable the WL decoder activation signal, WL the WL activation signal, and SAenable the SA activation signal. The timing delay is the cumulative delay across the paths of the timing circuit. It is defined as the sum of the delay between the rising clock and the enabling of the decoder (denoted by delay 1 in the figure) and the delay between the WL activation and the enabling of the SA (denoted by delay 2).

f) Write driver: We use the metric write delay to evaluate the write driver's parametric performance, as shown in Fig. 2(e). In the figure, WL shows the WL activation signal and QA/QB show the internal nodes of the memory cell during a write operation. The metric write delay is defined as the time between the activation of the WL and the intersection of QA and QB (i.e., the moment the cell's value changes).

2) Functional Metrics: The functional metrics evaluate whether the memory performs its functionality without any failures. Most components have one or more functional metrics. These components include the memory cell, SA, address decoder, and the write driver. We will discuss each component and its metric(s) next.

a) Memory cell: We use the metrics BL swing, hold static noise margin (SNM), and read SNM to evaluate the memory cell's functional performance. The BL swing metric is also illustrated in Fig. 2(b). The metric BL swing is defined as the differential voltage on the BLs generated by the cell at the moment, the SA is enabled. It is shown as ΔV_{BL} in the figure. We measure this at the inputs of the SA.

The hold SNM and read SNM evaluate the memory cell's stability. The hold SNM is defined as the maximum noise that can be tolerated on the internal nodes of the cell in its idle state (i.e., the pass transistors are disabled by the WL) before it flips. The read SNM is a measure of the maximum noise that can be tolerated when the cell is being read (i.e., when the pass transistors are enabled by the WL). We measure both the hold and read SNM using the simulation-based maximum square method from [20].

Finally, it is worth noting that the write-ability of the cell is not investigated. This is due to the fact that the write-ability improves with aging [21].

b) SA: We use the metrics sensing margin and offset voltage to evaluate the SA's functional performance. The sensing margin metric is also illustrated in Fig. 2(b). The sensing margin is defined as the time between the SA's output being ready and the deactivation of the SA.

Besides the sensing margin, the offset voltage is also an important functional metric of the SA. The offset voltage is defined as the differential input voltage that results in a differential output voltage equal to zero. In order, for the SA, to be able to generate the correct read value, the generated BL swing [ΔV_{BL} in Fig. 2(b)] should be higher than the offset voltage. In the ideal case, the offset voltage of the SA equals zero. However, this is never the case as the transistors of the SA are unbalanced due to process variation and aging [18].

c) Address decoder: We use the metric decoder setup margin to evaluate the functional performance of the address decoder. Fig. 2(f) illustrates the decoder setup margin metric. The top part of the figure shows a simplified diagram of the WL decoder. As can be seen, the WL decoder is controlled by the timing circuit. AND gates in the final stage of the decoder ensure that a WL can only be activated when the decoder is enabled. Before the decoder is enabled, all setup signals coming from the logic block must be ready. Otherwise, the wrong WL might be activated. The decoder setup margin is defined as the minimum time between the WL decoder setup signals being ready and the activation of the decoder enable, as illustrated in the bottom part of Fig. 2(f).

d) Write driver: We use the metric write margin to evaluate the functional performance of the write driver. Fig. 2(e) also shows the write margin metric. The metric write margin is defined as the time between the intersection of QA and QB (i.e., the moment the cell's value changes) and the deactivation of the WL.

C. Aging Model

Several aging mechanisms exist, such as BTI, hot carrier injection (HCI), and time-dependent dielectric breakdown [22]. BTI is considered to be the most important device failure mechanism for SRAMs in scaled CMOS technologies [22], [23]. Therefore, this paper uses BTI to model aging in SRAM. BTI takes place inside the MOS transistors and causes an increment in the absolute threshold voltage (V_{th}). This happens under negative gate stress for pMOS transistors, referred to as negative BTI (NBTI). For nMOS transistors, this happens under positive gate stress, which is referred to as positive BTI (PBTI).

This paper uses the atomistic model presented in [24] to model BTI. It is based on the capture of traps during stress periods and emission during relaxation periods. Each occupied trap contributes to the total threshold voltage shift ΔV_{th}. The probabilities of the capture P_C and emission P_E of traps are defined by [25] as follows:

$$P_C(t_{\text{STRESS}}) = \frac{\tau_e}{\tau_c+\tau_e}\left\{1-\exp\left[-\left(\frac{1}{\tau_e}+\frac{1}{\tau_c}\right)t_{\text{STRESS}}\right]\right\} \tag{1}$$

$$P_E(t_{\text{RELAX}}) = \frac{\tau_c}{\tau_c+\tau_e}\left\{1-\exp\left[-\left(\frac{1}{\tau_e}+\frac{1}{\tau_c}\right)t_{\text{RELAX}}\right]\right\} \tag{2}$$

where τ_c is the mean capture time constant, τ_e is the mean emission time constant, t_{STRESS} is the stress period, and t_{RELAX} is the relaxation period. The model also incorporates the impact of voltage and temperature [26].

III. Methodology and Performed Experiments

This section first discusses the used methodology and, subsequently, the performed experiments.

A. Methodology

Fig. 3 shows the methodology that is used to investigate the impact of aging on the memory. The methodology is based on

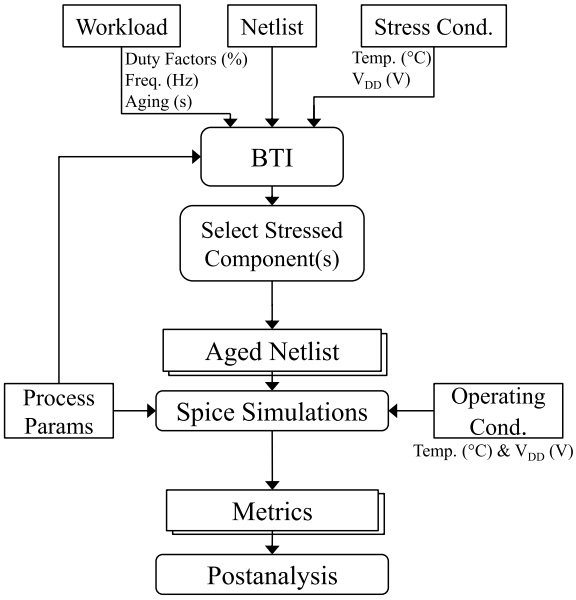


Fig. 3. Proposed methodology.

SPICE-level Monte Carlo simulations during which process variation and BTI are injected into the memory netlist. During each Monte Carlo run, the metrics described in the previous section are measured. Finally, postanalysis is performed on the measurement data, during which the average and spread are determined for each metric. To model BTI, the atomistic model is used; it was discussed in the previous section. The model has been calibrated through measurements on real devices [24]. The inputs to the BTI model are the netlist, workload, stress conditions, and process technology parameters. The workload is characterized by the duty factors and the frequency of the signals applied to the gates of the transistors and the aging time. The stress conditions consist of the junction temperature and supply voltage that are used for the aging of the circuit. The BTI model generates an aged netlist for each Monte Carlo iteration. The aged netlists are then simulated using Spectre at the supplied operating conditions. It is worth to note that these operating conditions may differ from the stress conditions. The used transistor model for the netlist is the PTM 14-nm FinFET LSTP library [27], which we calibrated with commercial 14-nm libraries to match their delay and power characteristics.

A key feature of the methodology is that it allows for the injection of aging/BTI into individual components of the memory. This allows to investigate the individual impact of aging of components and how it affects the interaction with other components. This paper distinguishes between the following components: the memory cells, the SAs, the output latches and buffers, the address decoder (includes both WL and BL decoders), the timing circuit, and the write drivers. Besides this individual aging analysis, we also analyze the combined aging of all components in order to examine the final impact of aging. Finally, it is worth to note that the proposed methodology is generic; not only can it be used on different memory designs but also on different circuits, such as sequential logic.

B. Performed Experiments

We analyze the impact of aging on the memory's parametric and functional metrics using the methodology from Fig. 3. We use four different workloads for the aging. These can be described using the following March algorithms.

1) Low activity Balanced (LB): $\Uparrow$(w0, r0, i^8, w1, r1, i^8).
2) Low activity Unbalanced (LU): $\Uparrow$(w0, r0, i^8).
3) High activity Balanced (HB): $\Uparrow$(w0, r0, w0, r0, i, w1, r1, w1, r1, i).
4) High activity Unbalanced (HU): $\Uparrow$(w0, r0, w0, r0, i).

Workloads LB and LU (HB and HU) assume that the memory performs operations 20% (80%) of the time and is idle in the remaining time. Their difference is that LB (HB) assumes a balanced workload for the memory cells and read/write circuitry, while LU (HU) assumes an unbalanced workload. What the workloads have, in common, is that they all iterate over the 256 memory addresses. As a result, the workloads stress each address an equal amount of time.

We perform the following experiments.

1) *Component sensitivity*: We investigate which components are responsible for the degradation of the memory. We simulate a stress time of 3 years at 85 °C and nominal supply voltage (0.8 V) for workload HU.
2) *Workload sensitivity*: We investigate the impact of workload on the degradation. We simulate a stress time of 3 years at 85 °C and nominal supply voltage (0.8 V) for four different workloads: LB, LU, HB, and HU.
3) *Environmental conditions sensitivity*: We investigate the impact of temperature and voltage fluctuations on the degradation. For the temperature-related experiments, we simulate a stress time of 3 years at nominal supply voltage for workload HU at the following temperatures: 25 °C, 85 °C, and 125 °C. For the voltage-related experiments, we first simulate that the memory is stressed for 3 years at 85 °C and nominal supply voltage for workload HU. Subsequently, we investigate the impact of voltage fluctuations on the aged memory netlist; we simulate the memory's operation at 85 °C for the following supply voltages: -10% V_{DD}, nominal V_{DD}, and $+10\%$ V_{DD}.

We perform 2000 Monte Carlo simulations for each experiment, in which we simulate process variation and stochastic aging. During each Monte Carlo simulation, the parametric and functional metrics are measured. Based on these Monte Carlo simulations, we then determine the degradation of the 6σ corner. It is obtained by dividing the value of the 6σ corner after aging by the 6σ corner at time-zero. Note that in case a deteratioration of the metric corresponds to a decrease of its value, the inverse ratio of the 6σ corners is taken (i.e., for the cell's noise margins and BL swing, the decoder's setup margin, and the write driver's write margin). The 6σ corner is calculated for each metric based on its mean μ and spread σ using the following equation:

$$\text{corner}_{6\sigma} = \mu \pm 6\sigma. \tag{3}$$

Here, addition is used, in case a deterioration of the metric corresponds to an increase of its value, while subtraction is

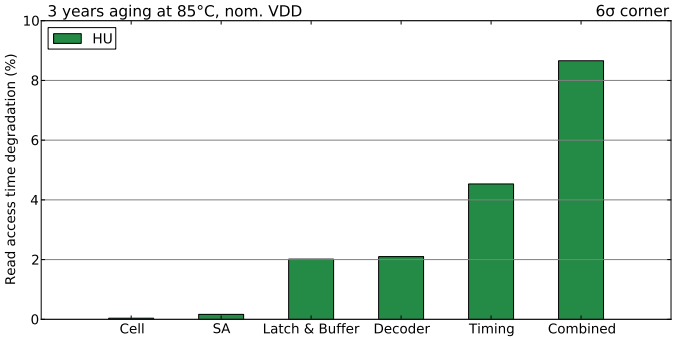


Fig. 4. Component sensitivity of read access time degradation.

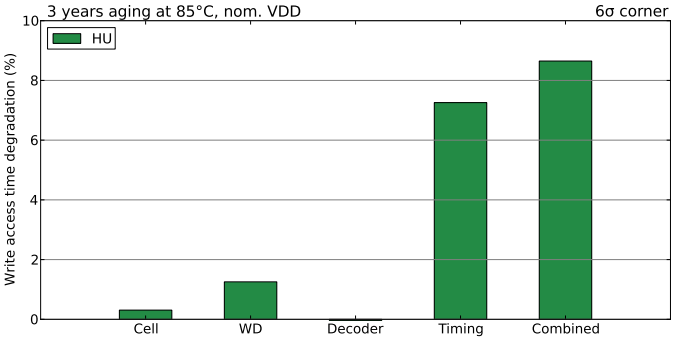


Fig. 5. Component sensitivity of write access time degradation.

used, in case a deterioration corresponds to a decrease of its value.

To evaluate the impact of aging on metrics related to read operations, we initialize the memory cells with zeroes and simulate a read from address "0." To evaluate the impact on metrics related to write operations, we initialize the memory cells with ones and simulate a write "0" operation to the cells from address "0." To evaluate the decoder setup margin, we simulate an address transition from address "255" to "0."

IV. Component Sensitivity Analysis

A. Parametric Metrics

Fig. 4 shows the degradation of the 6σ corner of the read access time for workload HU. The figure shows several aging scenarios along the x-axis. First, it shows cases where only individual components are aged. These components include the memory cell, SA, address decoder, and timing circuit. Finally, it shows the combined aging case.

The figure reveals that the impact on the read access time degradation is strongly component-dependent; the cell and SA have a marginal impact on the access time, while the output latch, address decoder, and, especially, the timing circuit have a high impact. The impact of the SA is marginal, since its delay forms only a small fraction of the total access time. Similarly, the impact of the cell is also marginal, since its aging causes a reduced BL swing, which mainly impacts the SA. The high impact of the output latch and buffer is caused by the fact that a relatively high output capacitance (i.e., 10 fF) was used during the simulation. Driving this capacitance takes a relatively large amount of time. Thus, the degradation of the output buffer has a high impact. The high impact of the address decoder is caused by the degradation of the gates that drive the respective WL. Driving the WL takes a significant portion of the total access time of the memory, as the WLs have a high parasitic capacitance. Therefore, the degradation of these gates results in a significant impact on the access time. The timing circuit has the highest impact on the read access time, as it has the longest paths. Hence, increasing cumulative delays along the timing paths result in its high impact.

Fig. 5 shows the degradation of the write access time for workload HU. Once again, the figure first shows cases where only individual components are aged and, finally, the combined aging case. Similar to the read access time, the timing circuit has the highest contribution to the degradation of the write access time. This is once again due to the fact that the timing circuit contains the longest paths. Furthermore, the cell and write driver both have a higher contribution to the write access time degradation than the cell and SA to the read access time degradation. This is due to the fact that the write access time is lower than the read access time. Hence, the delays of the cell and write driver take a bigger portion of the total write access time. Finally, the degradation of the address decoder does not have any impact on the write access time. The reason for this is that the WL is activated before the write drivers and, therefore, the delays on the address decoder are masked.

Fig. 6 shows the degradation of the individual delay metrics for each component. In addition, the figure shows the relative contribution of each component to the total access time on the right axis (at 85 °C, nom. process, nom. V_{DD}). Note that for the write driver, its delay and contribution are taken from a write operation, while for the other components, they are taken from a read operation. The figure reveals that the degradation of the delay is significantly higher for the output latch and buffer compared to the other components. The degradation of the output latch and buffer is $\sim$23%, while it is only between $\sim$4% and $\sim$9% for the other components. This can be explained by the fact that the transistors of the output latch and buffer are affected by a constant stress to hold the output value. In contrast, the other components are only active during a small portion of an operation. Hence, they are under significantly less stress.

Another observation that can be made is that even though certain components have a high degradation, they do not necessarily have the highest impact on the degradation of the total access time. This is revealed by comparing the degradation of the individual delay metrics in Fig. 6 with their impact on the access time in Fig. 4. For instance, the output latch and buffer have the highest individual degradation (Fig. 6). However, they do not have the highest contribution to the degradation of the access time (Fig. 4). Instead, components whose delay have the highest contribution to the access time, also have the highest degradation. This trend is revealed by comparing the contribution of each component to the total access time (orange bars in Fig. 6) with their contribution to the access time degradation (Fig. 4). For instance, the timing circuit has a high contribution to the access time and, therefore, also a

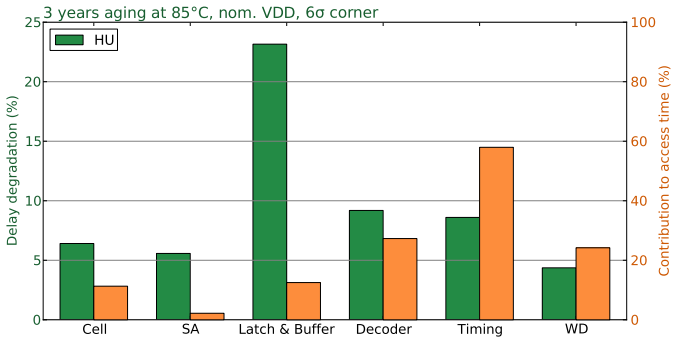


Fig. 6. Parametric degradation of individual components (green axis) and their contribution to the access time at time-zero (orange axis).

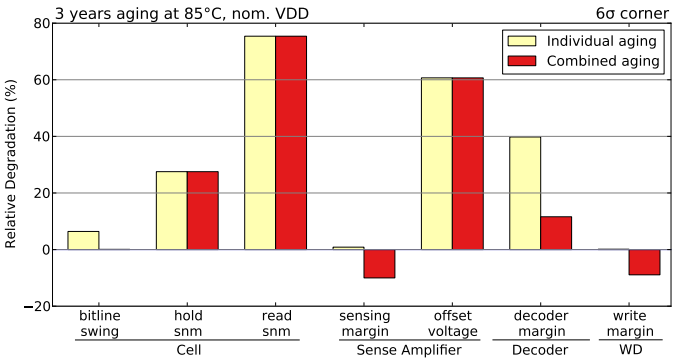


Fig. 7. Component sensitivity of functional degradation

high contribution to its degradation. The reason for this is that a slight degradation of a component with a high contribution will have a high overall impact. Hence, it is misleading to only analyze the degradation of individual components, because it ignores the above characteristics of the complete circuit.

B. Functional Metrics

Fig. 7 shows the degradation of the functional metrics for each component for workload HU. The figure depicts cases with individual aging and combined aging. In case of individual aging, only the component of the inspected metric is aged. In case of combined aging, the whole memory is aged; hence, it includes the impact of aging on the interaction between components. Note that a positive value for the degradation corresponds to a deterioration of the metric, while a negative value to an improvement.

The figure clearly shows that aging is strongly component-dependent. For example, the read SNM degrades by ∼75.4%, while the sensing margin improves by ∼10.0% (both for the combined aging case). Furthermore, we observe that it is misleading to consider a single metric per component. For instance, the SA's offset voltage degrades by ∼60.7%, while the sensing margin improves by ∼10.0%. In addition, the figure reveals that in a lot of cases, it is too pessimistic to only consider the aging of individual components. For example, the decoder margin degrades by ∼39.8% in case of individual aging. However, in case of combined aging, it actually degrades by only ∼11.6%. Thus, this demonstrates the importance of analyzing the impact of aging on the system as a whole, rather than on individual components. We will now discuss each component and its metric(s).

1) Memory Cell: In case of individual aging, we observe that the BL swing degrades. This happens due to the degradation of the pull down and pass transistors of the memory cell, which results in a slower discharge of the BLs. However, in case of combined aging, we observe that the BL swing only shows a marginal degradation. The reason for this is that aging in the timing circuit delays the activation of the SA. Hence, the cell gets more time for the BL discharge and its degradation is almost fully compensated.

The cell's hold and read SNM both show a significant degradation. This happens because aging causes a mismatch between the cross-coupled inverter pair of the cell. The read SNM shows the highest degradation. This is due to the fact that the read SNM strongly depends on the strength ratio between the pass and pull-down transistors [28]; stronger pass transistors improve the cell's capability of discharging the BL during a read at the cost of an increased disturbance at its inner nodes (i.e., a lower read SNM). The pass transistors show a low degradation, as they are only stressed when their respecitive WL is activated. In contrast, the pull-down transistors have a high degradation, as they are affected by a constant stress. Hence, the pass-transistors retain a higher portion of their original strength than the pull-down transistors, which negatively impacts the read SNM.

2) Sense Amplifier: in case of individual aging, we observe that the sensing margin shows a slight degradation. This is caused by the increase of the SA's sensing delay. In case of combined aging, however, we observe that the sensing margin improves. This can be explained by the aging in the timing circuit as it delays the deactivation of the SA. Due to longer logic paths in the timing circuit, the deactivation of the SA is delayed significantly more than the increase in delay of the SA. As a result, the sensing margin improves with aging.

The SA's offset voltage shows a significant degradation. Aging causes a bigger mismatch between the cross-coupled inverter pair of the SA. As a result, the offset voltage increases. At the same time, we observed that the BL swing does not increase enough due to aging to compensate for it. Therefore, sufficent margin for the BL discharge should be added to the design. Otherwise, the SA may generate incorrect read values.

3) Address Decoder: in case of individual aging, the setup margin shows a significant degradation due to increased path delays in the WL decoder. In case we consider combined aging, we observe that this degradation is partially compensated. This happens because the aging of the timing circuit delays the activation of the WL decoder. This compensation is only partial, because the path that activates the decoder is only stressed during an operation while parts of the address decoders are always stressed depending on which address is selected. Hence, the address decoder has a higher degradation.

4) Write Driver: in case of individual aging, we observe that the write margin shows a marginal degradation. This degradation is caused by the increase of the write driver's write delay. Due to the fact that the write margin is relatively high at time-zero, the increased write delay only has a marginal impact. In case of combined aging, we observe that the write

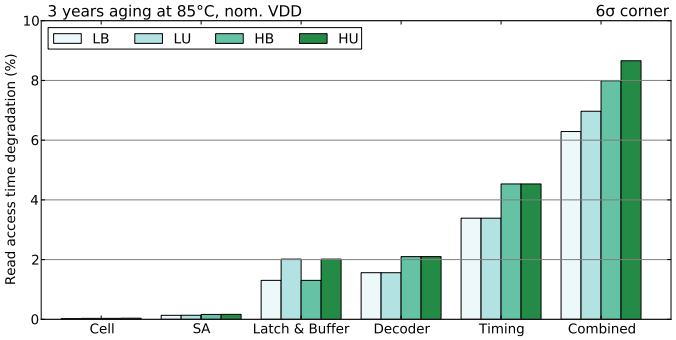


Fig. 8. Workload dependence of read access time degradation.

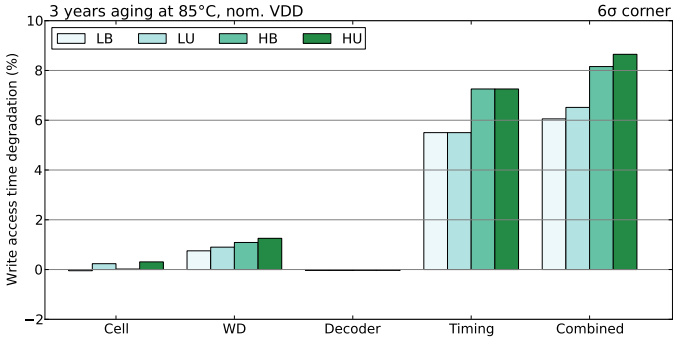


Fig. 9. Workload dependence of write access time degradation.

margin improves. The reason for this is that aging in the timing circuit and decoder delays the deactivation of the WL. Due to longer paths in the timing circuit and decoder, the increase in delay of the WL deactivation is significantly higher than the increase of the write delay. Therefore, the write margin improves with aging.

C. Parametric Versus Functional Metrics

In case we compare the component sensitivity of the memory's parametric degradation (Fig. 4) and functional degradation (Fig. 7), we observe that there is little correlation between the degradation of functional metrics of individual components and their impact on the overall access time. Therefore, both functional and parametric metrics should be considered during design. The main contributors to the parametric degradation are the output latch and buffer, the decoder, and the timing circuit. For the parametric degradation, the components that have the highest contribution to the access time also have the highest contribution to its degradation. The main contributing components to the memory's functional degradation are the cell (i.e., its hold and read SNM), the SA (i.e., its offset voltage), and the decoder. For the functional degradation, the most contributing components are components whose aging is not or not fully compensated by the degradation of other components.

V. Workload Sensitivity Analysis

A. Parametric Metrics

Fig. 8 shows the workload dependence of the degradation of the read access time. The figure first shows cases where only single components are aged and, finally, the combined aging case. Note that the same information as in Fig. 4 is presented for HU. The figure reveals that workloads that access the memory more often result in a higher degradation. For instance, the read access time increases with ∼7.0% for LU, while it increases with ∼8.7% for HU. Analyzing the contribution of individual components reveals that the main contributors to this behavior are the timing circuit and address decoder. This is due to the fact that workloads with higher access rates stress the paths of these components more often and, therefore, they cause a higher degradation.

Furthermore, comparing between workloads with balanced read/write values (LB and HB) and unbalanced values (LU and HU) reveals that unbalanced workloads cause a higher degradation. For instance, the read access time increases with ∼8.0% for HB, while it increases with ∼8.7% for HU. The contributors to this behavior are the cell, SA, and output latch and buffer. Due to the unbalanced workload, the transistors responsible for generating the favored read value degrade more. Hence, the delay to generate this value increases the most. It is worth to note that the output latch and buffer have the highest contribution to this behavior, as a high output capacitance was used during simulation. Therefore, minimizing the output capacitance of the memory would significantly reduce the impact of unbalanced workloads on the read acces time.

Fig. 9 shows the workload dependence of the write access time degradation. Analyzing the impact of activity reveals that workloads with a higher activity cause a higher degradation. The main contributors to this behavior are the timing circuit and write driver. This is due to the fact that high-activity workloads stress these components more often. Furthermore, unbalanced workloads result in a higher degradation. Mainly, the cell and write drivers are responsible for this behavior. Due to the unbalanced workload, the cell's transistors that are active when it stores a zero have a higher wearout. As a result, the cell becomes slower at flipping to a zero value. In case of the write driver, the circuitry responsible for writing a zero is stressed more by the unbalanced workload.

B. Functional Metrics

Fig. 10 shows the workload dependence of the degradation of the functional metrics for combined aging. From the figure, we observe that the impact of workload is strongly component-dependent; for some components, the degradation mainly depends on the balancing of the read/write values, while for others, it mainly depends on the activity of the workload. Also, in some cases, the degradation is strongly dependent on both the balancing of read/write values and the activity. We will now discuss the observed trends for each component and its metric(s) separately.

1) Memory Cell: The BL swing degradation is marginally dependent on the workload. This happens due to the fact that the delayed activation of the SA and the cell degradation even out, irrespective of the workload; the SA activation becomes more delayed for workloads with a higher activity and, thus,

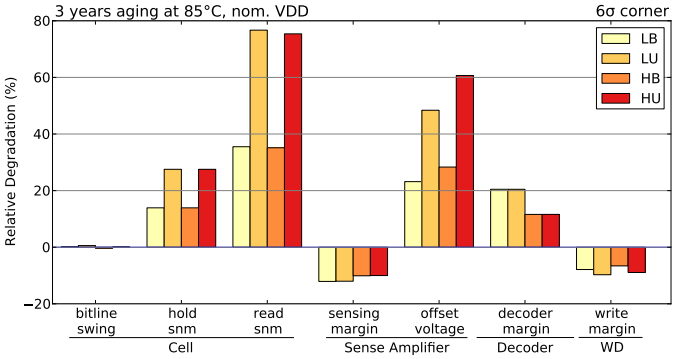


Fig. 10. Workload dependence of functional degradation.

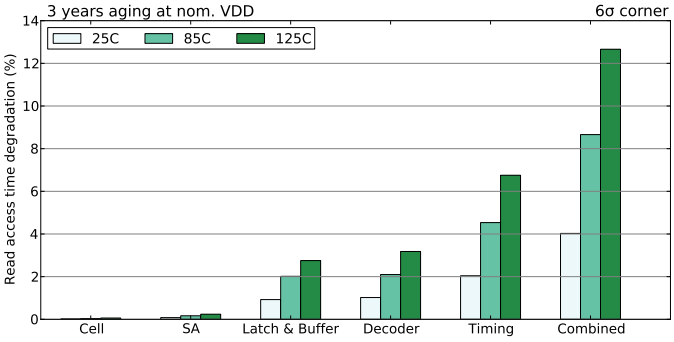


Fig. 11. Temperature sensitivity of read access time degradation. The memory is stressed at 25 °C, 85 °C, and 125 °C.

the cell gets more time for the BL discharge. However, workloads with a higher activity also activate the cell's pass transistors more often and, therefore, the cell becomes slower at discharging the BL.

We observe for the cell's hold and read SNM that they both strongly depend on the balancing of the read/write values. Their degradation is highest for unbalanced workloads (LU and HU). This is due to the fact that unbalanced workloads cause an unbalanced degradation of the transistors of the cross-coupled inverter pair. In addition, the degradation of the read SNM also slightly depends on the activity of the workload; workloads with a lower activity result in a higher degradation. This can be explained by the fact that these stress the cell's pass transistors less and, hence, they have a lower degradation. As a result, the pass transistors give a higher disturbance to the inner cell nodes during a read operation.

2) Sense Amplifier: Fig. 10 reveals that the sensing margin mainly depends on the activity of the workload. High-activity workloads give a lower improvement to the sensing margin. This is due to the fact that they stress the SA more and, thus, the sensing delay has a higher increase. Furthermore, they cause a lower degradation of the delay of the deactivation of the SA. In case the memory is not performing a read operation, the path responsible for the deactivation of the SA is active and, thus, stressed. Hence, workloads with a lower activity result in a higher stress for this path. The combination of the increased sensing delay and the reduced delay of the deactivation of the SA results in a lower relaxation of the sensing margin for high-activity workloads.

For the SA's offset voltage, we observe that it is especially sensitive to the balancing of the read values. The offset voltage is very susceptible to unbalanced workloads (LU and HU). Unbalanced workloads cause an unbalanced degradation of the transistors of the SA and, therefore, its offset voltage increases significantly. For example, balanced workload HB gives a degradation of ∼28.3%, while unbalanced workload HU gives a degradation of ∼60.7%. In addition, the degradation of the offset voltage is dependent on the activity of the workload. The high-activity workloads give a higher degradation as they stress the transistors of the SA more. For example, workload LU gives a degradation of ∼48.4%, while workload HU gives a degradation of ∼60.7%.

3) Address Decoder: The figure shows that the decoder setup margin only depends on the activity of the workload. A higher activity workload results in a lower degradation. This can be explained by the fact that higher activity workloads stress the timing circuit more. As a result, the activation of the decoder becomes more delayed and a bigger portion of the increased path delay of the decoder is compensated.

4) Write Driver: The figure reveals that the write margin mainly depends on the balancing of the write values. Unbalanced workloads further relax the write margin. This is due to the fact that the unbalanced workload cause an unbalanced degradation of the memory cell's transistors. As a result, the cell flips easier when a one is written.

C. Parametric Versus Functional Metrics

In case we compare the workload dependence of the memory's parametric degradation (Figs. 8 and 9) and functional degradation (Fig. 10), we observe that the parametric degradation is mainly impacted by the activity of the workload. Workloads with a higher activity stress the paths responsible for read/write operations more, and, therefore, cause a higher degradation. On the other hand, the memory's functional degradation mainly depends on the balancing of the read/write values of the workload. Both the cell and SA are very sensitive to unbalanced workloads. This is due to the fact that unbalanced workloads cause an unbalanced degradation of their cross-coupled inverter pairs.

VI. Environmental Conditions Sensitivity Analysis

A. Parametric Metrics

Fig. 11 shows the impact of temperature on the degradation of the read access time for workload HU. Once again, the figure first shows cases where single components are aged and, finally, the impact of their combined aging. Note that the degradation is determined by comparing the access time after stress with respect to the access time at time-zero (i.e., no aging and only process variation) while considering the same operational and stress temperature. For example, in case the memory is stressed at 85 °C, its resulting access time is compared to the access time at time-zero at 85 °C. The figure reveals that a higher temperature results in a

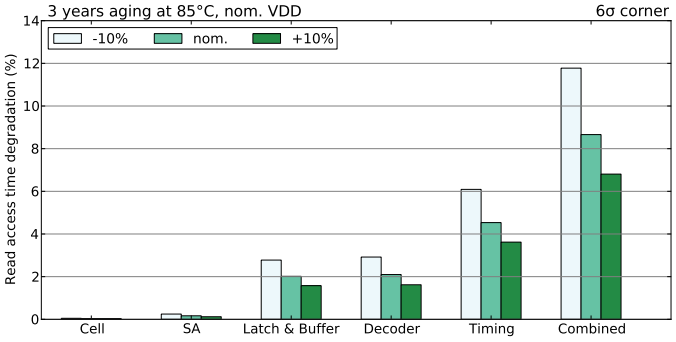


Fig. 12. Voltage sensitivity of read access time degradation. Nominal voltage is used during the stress period. The degradation is then measured at -10%, nominal, and $+10\%$ V_{DD}.

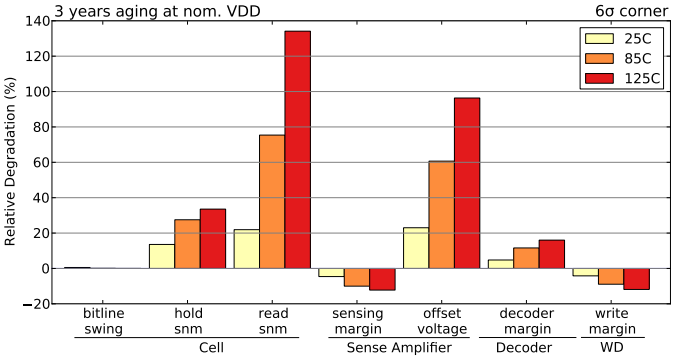


Fig. 13. Temperature sensitivity of functional degradation. The memory is stressed at 25 °C, 85 °C, and 125 °C.

higher degradation. For example, in case of combined aging, the access time degrades by $\sim$4.0% at 25 °C, while it degrades by $\sim$8.7% and $\sim$12.7% at 85 °C and 125 °C, respectively. The reason for this is that elevated temperatures accelerate the BTI mechanism [29].

Moreover, analyzing the impact of individual components reveals that for all temperatures, the same components are the main contributors to the degradation of the access time. The output latch and buffer, decoder, and timing circuit are the main contributors. Once again, this is due to the fact that these components have the highest contribution to the overall access time and, therefore, also the highest contribution to its degradation.

Fig. 12 shows the impact of the operational voltage on the degradation of the read access time for workload HU. Note that, in all cases, the memory is first stressed at nominal V_{DD}. After stress, a voltage fluctuation (i.e., -10%, nominal, and $+10\%$ V_{DD}) is then simulated. The degradation is determined by comparing the access time before stress with the access time after stress using the same operational voltage. For example, in case of a -10% voltage fluctuation, the access times before and after stress at supply voltages of 0.72 V are compared. The figure reveals that the impact of aging is highest during voltage drops. For instance, in case of combined aging, the access time degrades by $\sim$8.7% at nominal voltage, while it degrades by $\sim$11.8% at a -10% voltage fluctuation. The reason for this is that the threshold voltage shifts induced by BTI have a higher relative impact at lower supply voltages. Finally, it is worth noting that the main contributors to the degradation of the access time are once again the output latch and buffer, decoder, and timing circuit.

B. Functional Metrics

Fig. 13 shows the impact of the temperature on the degradation of the functional metrics for workload HU. The degradation is determined by comparing the metrics after stress with the metrics at time-zero while considering the same operational and stress temperature. The figure shows that a higher temperature results, in general, in a higher degradation and, in some cases, in a higher relaxation. For the cell's hold and read SNM, the SA's offset voltage, and the decoder margin, a higher temperature results in a higher degradation. In particular, the cell's read SNM shows a high sensitivity to temperature; its degradation increases by $\sim$7X at 125 °C compared to 25 °C. For the other components, their degradation increases between 2.5 times and 5 times at 125 °C compared to 25 °C. The accelerated aging at higher temperatures causes a bigger mismatch between the cross-coupled inverter pairs of the cell and SA. In case of the decoder margin, the paths of the WL decoder have a higher relative degradation at higher temperatures than the path that activates the decoder. This is due to the fact that the decoder's paths have a higher stress; they are also stressed when the memory is idle. Therefore, the decoder margin decreases at higher temperatures.

The SA's sensing margin and the write driver's write margin show a higher relaxation at higher temperatures. This is due to the fact that the paths that deactivate the SA and write driver are under a higher stress than the SA and write driver; these paths are always enabled during an idle cycle and also during a significant portion of a read/write operation. On the contrary, the SA and write driver are only active during a small portion of a read/write operation. Therefore, the deactivation paths degrade more, and an increase in the margins is observed.

Fig. 14 shows the impact of voltage fluctuations on the degradation of the functional metrics for workload HU. Note that the degradation is again determined by comparing the access time after stress with the access time at time-zero while considering the same voltage fluctuation for both. From the figure, we observe that voltage drops result in either a higher degradation or a higher relaxation. It is worth to note that, especially, the cell's read SNM and decoder margin are sensitive to voltage fluctuations; their degradation is 2 times higher at a -10% fluctuation compared to a $+10\%$ fluctuation. The degradation of the cell's hold SNM is $\sim$1.6 times higher at a -10% fluctuation compared to a $+10\%$ fluctuation. For the cell's hold and read SNM and the decoder margin, a lower voltage results in a higher degradation. During voltage drops, the induced threshold voltage shifts of the cell's transistors have a higher relative impact. Hence, voltage drops negatively impact the hold and read SNM. In case of the decoder margin, the aged paths of the WL decoder have a higher relative degradation than the path that activates the decoder during voltage drops. As mentioned, the decoder's paths have a higher stress. Hence, its transistors have higher induced

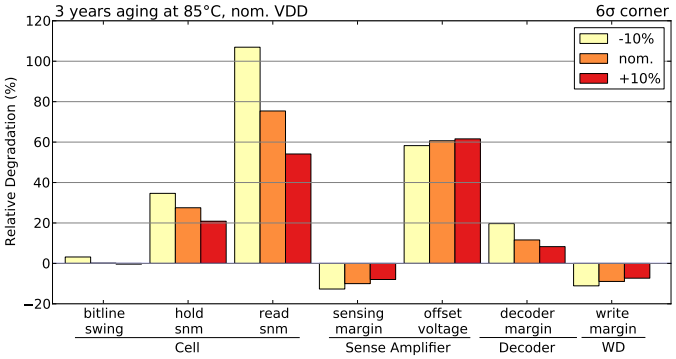


Fig. 14. Voltage sensitivity of functional degradation. Nominal voltage is used during the stress period. The degradation is then measured at −10%, nominal, and +10% V_{DD}.

threshold voltage shifts. As a result, they have a higher relative degradation during voltage drops.

The SA's sensing margin and the write driver's write margin show a higher relaxation during voltage drops. This is explained by the fact that the paths that deactivate the SA and write driver have a higher degradation than the SA and write driver. As already stated, the paths that activate the SA and write driver are under more stress than the SA and write driver themselves. Therefore, the transistors of these paths have higher induced threshold voltage shifts, which cause a relatively higher degradation during voltage drops.

C. Parametric Versus Functional Metrics

In case we compare the impact of temperature on the memory's parametric degradation (Fig. 11) and functional degradation (Fig. 13), we observe that the parametric degradation of all components shows a similar sensitivity to temperature. Each component shows a ∼3 times increase of its degradation at 125 °C compared to 25 °C. The functional metrics, however, show very different sensitivities to temperature. For instance, the cell's read SNM shows a ∼7 times increase of its degradation at 125 °C compared to 25 °C, while the SA's offset voltage only shows a ∼4 times increase. The same trend is observed for the sensitivity to voltage fluctuations; the parametric metrics (Fig. 12) show a similar sensitivity to voltage fluctuations for all components, while their functional metrics (Fig. 14) show very different sensitivities. The similar sensitivity to environmental conditions for the parametric metrics is explained by the fact that they are all delay metrics. Hence, it is likely that their response to environmental conditions is in the same order of magnitude. More variety is present for the functional metrics, however, for example, some metrics are related to delay (e.g., decoder margin), while others are related to the maximum voltage noise they can withstand (e.g., hold SNM). Hence, they are very different in nature and, therefore, their sensitivity to environmental conditions as well.

VII. Discussion

The reliability of embedded memories is extremely important for the overall system reliablity. Based on this paper, we make the following observations with respect to FinFET SRAM reliability.

A. Interaction

The reliability of memories is a global problem. It is crucial to consider the impact of aging on all components and, thus, on their interaction in order to obtain accurate predictions. Analyzing the individual degradation of components leads to too pessimistic results and even wrong conclusions. For instance, the decoder margin decreased with ∼39.8% when only considering the individual aging of the address decoder, while in case of considering the combined aging of all components, it degraded by only ∼11.6%.

B. Component Sensitivity

In this paper, we examined the impact of aging on the memory's parametric and functional metrics. The parametric metrics evaluated the delay of the individual components and the overall system (e.g., read access time). The functional metrics evaluated whether the memory is able to perform its intended functionality. Our study revealed that there is little correlation between the degradation of the functional metrics of individual components and their impact on overall parametric metrics, such as the read access time. Therefore, both functional and parametric metrics need to be considered during design.

Our case study revealed that the timing circuit has the highest impact on the memory's overall parametric degradation (read/write access time). This will apply to any self-timed memory design, like our design, due to the fact that the timing circuit has the longest paths and, thus, also the highest aging-induced delays. In addition, our study showed that after the timing circuit, the address decoder and output latch and buffer have the highest impact on the memory's parametric degradation. It can be expected that the address decoder will have a high impact on the parametric degradation for other memory designs as well, as the delay to drive the WLs increases in case the size of the memory array increases. On the contrary, the impact of the output latch and buffer is expected to decrease for bigger memories, as their delay will have a lower contribution to the overall delay.

Moreover, our case study revealed that the cell, SA, and address decoder are the responsible components for the memory's functional degradation. It can be expected that these components will be the most critical for other memory designs as well. In the case of the memory cell, this is due to the fact that its noise margins are not affected by other components and, thus, their degradation is also not compensated. Hence, sufficient margin should be added to the memory cells for any SRAM design in order to gaurantee a high reliability. In the case of SA, its offset voltage showed a severe degradation (up to 100%). The BL swing, however, showed only a marginal improvement and, thus, it is unable to compensate for the SA's aging. It is unlikely that other memory designs will have a significantly lower SA degradation or a significantly higher compensating behavior for the BL swing. In the case of the address decoder, its setup margin showed a degradation due

to the fact that the increased decoder delays are not fully compensated by aging of the timing circuit. This is due to the fact that the logic paths of the decoder have a higher stress than the paths of the timing circuit responsible for the activation of the decoder; the decoder's paths are always stressed depending on the selected address, while these timing circuit paths are only stressed during operations. This will also be the case for other memory designs and, thus, it is likely that the decoder has a higher degradation than the timing paths. Hence, sufficient margin needs to be added during design to the memory cell, SA, and address decoder to guarantee a reliable memory.

C. Impact of Workload

The impact of aging is strongly workload-dependent. This suggests that memory aging depends on the used application. For the overall parametric metrics (read/write access time), we observed that the activity of the workload (i.e., the amount of operations) was the main contributor to their degradation. Hence, memory-intensive applications result in a higher degradation of the memory's access time. For the functional metrics, we observed that the balancing of the read/write values is the main contributor to their degradation. The cross-coupled inverter pairs of the cell and SA are very sensitive to unbalanced workloads, as these create a mismatch between them. Therefore, it is very important to balance the read/write values by using techniques such as bit flipping [10], [30].

D. Impact of Temperature

A higher temperature results, in general, in a higher degradation. This is both the case for the parametric and functional metrics; although the increased parametric degradation relaxes the timing signals, it does not compensate fully for the increased functional degradation of most components. Hence, it is important to ensure proper cooling. Furthermore, techniques that reduce the power consumption and, therefore, heat generation may be applied, such as adaptive voltage and frequency scaling [31].

E. Impact of Voltage Fluctuations

The impact of aging is highest during voltage drops. In general, this results in the highest degradation of the analyzed metrics. Due to shrinking device dimensions, increasing switching frequencies, and increasing power consumption, cutting-edge technology is affected by higher voltage fluctuations [32]. Thus, a robust power network design is gaining an increasing importance as technology scales down further.

F. FinFET Versus Planar

Although FinFET has reduced process variation compared to planar technology [33], [34], comparing our results with other works suggests that it is less resilient to aging. For example, Kraak *et al.* [35] obtained for a planar 45-nm SA, an offset voltage degradation of up to 106.3% for the 6σ corner for a workload that assumes 80% read operations. In this paper, we observed for 14-nm FinFET that the 6σ corner of the offset voltage degrades with up to 96.4% for a workload that consists of 40% read operations. Hence, FinFET already shows a similar degradation for half the amount of operations. This suggests that FinFET has a higher degradation than planar technology. Degradation is, however, also dependent on other factors than the technology, such as the design type (e.g., high performance or low power) [36], making it difficult to draw fair conclusions. Therefore, more research on the reliability of FinFET versus planar should be performed, to help designers choose the most appropriate technology to meet performance and reliability targets for their designs.

VIII. Conclusion

This paper investigated the impact of aging on a high-performance 14-nm FinFET SRAM. The results showed that the timing circuit, followed by the address decoder, and output latches and buffers are the main contributors to the memory's parametric degradation. On the other hand, this paper showed that the cell, SA, and address decoder are the main contributors to the memory's functional degradation. Finally, we would like to underline the importance of the simulation framework that enabled this analysis. Such a framework is especially important for cutting-edge technologies, as it suffers from higher aging. Moreover, the framework could even be extended to allow for the reliability analysis of emerging technologies. Based on the outcomes of the framework, designers can take precautions to achieve the desired level of reliability depending on the targeted application.

References

[1] S. Borkar, "Microarchitecture and design challenges for gigascale integration," in *Proc. 37th Int. Symp. Microarchitecture*, Dec. 2004, p. 3.

[2] S. Hamdioui, D. Gizopoulos, G. Guido, M. Nicolaidis, A. Grasset, and P. Bonnot, "Reliability challenges of real-time systems in forthcoming technology nodes," in *Proc. Design Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2013, pp. 129–134.

[3] B. Kaczer *et al.*, "Atomistic approach to variability of bias-temperature instability in circuit simulations," in *Proc. Int. Rel. Phys. Symp.*, Apr. 2011, pp. XT.3.1–XT.3.5.

[4] S. Khan and S. Hamdioui, "Modeling and mitigating NBTI in nanoscale circuits," in *Proc. IEEE 17th Int. On-Line Test. Symp.*, Jul. 2011, pp. 1–6.

[5] S. V. Kumar, K. H. Kim, and S. S. Sapatnekar, "Impact of NBTI on SRAM read stability and design for reliability," in *Proc. 7th Int. Symp. Qual. Electron. Design (ISQED)*, Mar. 2006, p. 218.

[6] S. Khan *et al.*, "Bias temperature instability analysis of FinFET based SRAM cells," in *Proc. Design Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2014, pp. 1–6.

[7] B. Cheng, A. R. Brown, and A. Asenov, "Impact of NBTI/PBTI on SRAM stability degradation," *IEEE Electron Device Lett.*, vol. 32, no. 6, pp. 740–742, Jun. 2011.

[8] S. V. Kumar, K. H. Kim, and S. S. Sapatnekar, "Impact of NBTI on SRAM read stability and design for reliability," in *Proc. 7th Int. Symp. Qual. Electron. Design (ISQED)*, Mar. 2006, p. 218.

[9] A. Bansal, R. Rao, J.-J. Kim, S. Zafar, J. H. Stathis, and C.-T. Chuang, "Impacts of NBTI and PBTI on SRAM static/dynamic noise margins and cell failure probability," *Microelectron. Rel.*, vol. 49, no. 6, pp. 642–649, Jun. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0026271409000894

[10] A. Gebregiorgis, M. Ebrahimi, S. Kiamehr, F. Oboril, S. Hamdioui, and M. B. Tahoori, "Aging mitigation in memory arrays using self-controlled bit-flipping technique," in *Proc. 20th Asia South Pacific Design Autom. Conf.*, Jan. 2015, pp. 231–236.

[11] R. Menchaca and H. Mahmoodi, "Impact of transistor aging effects on sense amplifier reliability in nano-scale CMOS," in *Proc. 13th Int. Symp. Qual. Electron. Design (ISQED)*, Mar. 2012, pp. 342–346.

[12] I. Agbo *et al.*, "Integral impact of BTI, PVT variation, and workload on SRAM sense amplifier," *IEEE Trans. Very Large Scale Interact. (VLSI) Syst.*, vol. 25, no. 4, pp. 1444–1454, Apr. 2017.

[13] H. I. Yang, S. C. Yang, W. Hwang, and C. T. Chuang, "Impacts of NBTI/PBTI on timing control circuits and degradation tolerant design in nanoscale CMOS SRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 6, pp. 1239–1251, Jun. 2011.

[14] I. Agbo *et al.*, "Read path degradation analysis in SRAM," in *Proc. 21st IEEE Eur. Test Symp. (ETS)*, May 2016, pp. 1–2.

[15] J. Kinseher, L. HeiSS, and I. Polian, "Analyzing the effects of peripheral circuit aging of embedded SRAM architectures," in *Proc. Design Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2017, pp. 852–857.

[16] D. Kraak *et al.*, "Degradation analysis of high performance 14 nm FinFET SRAM," in *Proc. Design, Automat. Test Eur. Conf. Exhib. (DATE)*, Dresden, Germany, 2018, pp. 201–206. [Online]. Available: https://ieeexplore.ieee.org/document/8342003. doi: 10.23919/DATE.2018.8342003.

[17] D. Kraak *et al.*, "Device aging: A reliability and security concern," in *Proc. IEEE 23rd Eur. Test Symp. (ETS)*, May 2018, pp. 1–10.

[18] I. Agbo *et al.*, "Quantification of sense amplifier offset voltage degradation due to zero-and run-time variability," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI (ISVLSI)*, Jul. 2016, pp. 725–730.

[19] D. Rodopoulos *et al.*, "Classification framework for analysis and modeling of physically induced reliability violations," *ACM Comput. Surv.*, vol. 47, no. 3, p. 38:1–38:33, Feb. 2015. [Online]. Available: http://doi.acm.org/10.1145/2678276

[20] E. Seevinck, F. J. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, Oct. 1987.

[21] K. Kang, S. P. Park, K. Roy, and M. A. Alam, "Estimation of statistical variation in temporal NBTI degradation and its impact on lifetime circuit performance," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2007, pp. 730–734.

[22] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, "Predictive modeling of the NBTI effect for reliable design," in *Proc. IEEE CICC*, Sep. 2006, pp. 189–192.

[23] R. Newhart, "Early reliability modeling for aging and variability in silicon system IBM view, Ermavss workshop," in *Proc. DATE*, 2016.

[24] P. Weckx *et al.*, "Defect-based compact modeling for RTN and BTI variability," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2017, p. CR–7.1–CR–7.6.

[25] M. Toledano-Luque *et al.*, "Response of a single trap to AC negative bias temperature stress," in *Proc. IRPS*, Apr. 2011, pp. 4A.2.1–4A.2.8.

[26] B. Kaczer *et al.*, "Origin of NBTI variability in deeply scaled pFETs," in *Proc. IEEE Int. Rel. Phys. Symp.*, May 2010, pp. 26–32.

[27] *Predictive Technology Model*. Accessed: Feb. 1, 2018. [Online]. Avalaible: http://ptm.asu.edu/

[28] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, "Read stability and write-ability analysis of SRAM cells for nanometer technologies," *IEEE J. Solid-State Circuits*, vol. 41, no. 11, pp. 2577–2588, Nov. 2006.

[29] T. Grasser *et al.*, "Analytic modeling of the bias temperature instability using capture/emission time maps," in *IEDM Tech. Dig.*, Dec. 2011, pp. 27.4.1–27.4.4.

[30] D. Kraak *et al.*, "Impact and mitigation of sense amplifier aging degradation using realistic workloads," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 12, pp. 3464–3472, Dec. 2017.

[31] S. Herbert and D. Marculescu, "Analysis of dynamic voltage/frequency scaling in chip-multiprocessors," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, Aug. 2007, pp. 38–43.

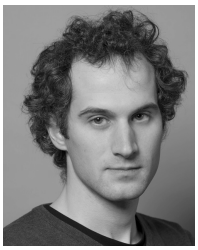
[32] S. K. Nithin, G. Shanmugam, and S. Chandrasekar, "Dynamic voltage (IR) drop analysis and design closure: Issues and challenges," in *Proc. 11th Int. Symp. Qual. Electron. Design (ISQED)*, Mar. 2010, pp. 611–617.

[33] M. D. Giles *et al.*, "High sigma measurement of random threshold voltage variation in 14nm Logic FinFET technology," in *Proc. Symp. VLSI Technol. (VLSI Technol.)*, Jun. 2015, pp. T150–T151.

[34] H. Kukner *et al.*, "BTI Reliability from Planar to FinFET nodes: Will the next node be more or less reliable?" Mar. 2014, Art no. 7564296.

[35] D. Kraak *et al.*, "Mitigation of sense amplifier degradation using input switching," in *Proc. Design Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2017, pp. 858–863.

[36] I. Agbo *et al.*, "Comparative BTI analysis for various sense amplifier designs," in *Proc. IEEE 19th Int. Symp. Design Diag. Electron. Circuits Syst. (DDECS)*, Apr. 2016, pp. 1–6.



Daniël Kraak (S'16) received the bachelor's degree in electrical engineering and the master's degree in computer engineering from Delft University of Technology, Delft, The Netherlands, in 2013 and 2016, respectively. He is currently working toward the Ph.D. degree at the Computer Engineering Laboratory, Dependable Nanocomputing Group, Delft University of Technology. His master thesis was focused on variability resilient design methods and was performed in cooperation with NXP Semiconductors.

His current research interests include robust memory design and mitigation for aging.

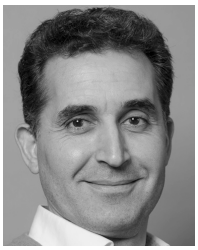
Mottaqiallah Taouil (S'10–M'15) received the M.Sc. and Ph.D. degrees (Hons.) in computer engineering from Delft University of Technology, Delft, The Netherlands, in 2009 and 2014, respectively.

He is currently an Assistant Professor at the Computer Engineering Laboratory, Dependable Nanocomputing Group, Delft University of Technology. His current research interests include reconfigurable computing, embedded systems, very large scale integration design and test, built-in-self-test, and 3-D stacked integrated circuits, architectures, design for testability, yield analysis, and memory test structures.

Innocent Agbo (S'10–M'19) received the M.Sc. and Ph.D. degrees in computer engineering from Delft University of Technology, Delft, The Netherlands, in 2010 and 2018, respectively.

He is currently a Postdoctoral Researcher at the Computer Engineering Laboratory, Dependable Nanocomputing Group, Delft University of Technology. His current research interests include the reliability analysis, design, monitoring and mitigation of memory systems.

Dr. Agbo was a recipient of the Best Paper Award at IEEE Computer Society Annual Symposium on VLSI (ISVLSI) 2016 and the Best Paper Award at International Conference on Design and Test of Integrated Systems in the nanoera DTIS 2015.



Said Hamdioui (M'99–SM'11) received the M.S.E.E. and Ph.D. degrees (Hons.) from Delft University of Technology, Delft, The Netherlands, in 1997 and 2001, respectively.

He was at Intel Corporation, Santa Clara, CA, USA, Philips Semiconductors Research and Development, Crolles, France, and Philips/NXP Semiconductors, Nijmegen, The Netherlands. He was as a Professor at Delft University of Technology. He is currently the Chair Professor on Dependable and Emerging Computer Technologies and the Head of the Computer Engineering Laboratory, Delft University of Technology. He has authored or coauthored one book and contributed to two others. He has coauthored over 180 conference and journal papers. He holds two patents. His current research interests include dependable CMOS nanocomputing, including reliability, testability, hardware security and emerging technologies and computing paradigms, including 3-D stacked ICs, memristors for logic and storage, in-memory-computing. He has consulted for many companies. He delivered dozens of keynote speeches, distinguished lectures, and invited presentations and tutorials at major international forums/conferences/schools and at leading semiconductor companies.

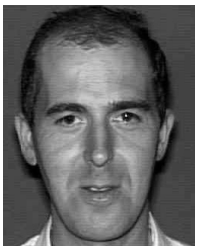
Dr. Hamdioui is a member of AENEAS/ENIAC Scientific Committee Council. He is an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He serves on the Editorial Board for IEEE DESIGN AND TEST, the *Elsevier Microelectronic Reliability Journal*, and the *Journal of Electronic Testing: Theory and Applications*.

Pieter Weckx (M'14) received the B.Sc. degree in electronic engineering, the M.Sc. degree in nanoscience and technology, and the Ph.D. degree in engineering from the Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium, in 2009, 2011, and 2016, respectively.

In 2015, he joined Interuniversity Microelectronics Center (imec), Leuven, where he is currently a Senior Researcher in circuit and device architecture for future scaled CMOS technologies. His current research interests include device variability and reliability and design and technology cooptimization of deeply scaled LOGIC and SRAM technologies.

Stefan Cosemans (S'04'–M'10–SM'18) received the Ph.D. degree from Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium, in 2009, with a focus on low-power SRAM.

From 2010 to 2014, he was at Interuniversity Microelectronics Center (imec), Leuven, where he was involved in circuit design for RRAM, STT-MRAM, and other novel memory devices. From 2015 to 2017, he was at sureCore Ltd., Leuven, where he led the development of single rail near-threshold SRAM compilers. In 2017, he joined imec Machine Learning Program. He is currently a Principal Member of Technical Staff at imec. His current research interests include circuits and memory devices for machine learning accelerators.



Francky Catthoor (F'05) received the Engineering degree and the Ph.D. degree in electrical engineering from the Katholieke Universiteit Leuven (KU Leuven), Leuven, Belgium, in 1982 and 1987, respectively.

From 1983 to 1987, he was a Researcher of VLSI design methodologies for digital signal processing, with Prof. H. De Man and Prof. J. Vandewalle as Ph.D. Thesis Advisors. Since 1987, he has been with the Systems Division, Interuniversity Microelectronics Center (imec), Leuven, where he has headed several research domains in high-level and system synthesis techniques and architectural methodologies. Since 1989, he has been an Assistant Professor with the Department of Electrical Engineering, KU Leuven, where he has been a Full Professor (part-time) since 2000. He is currently a fellow at imec. His current research interests include architecture design methods and system-level exploration for power and memory footprint within real-time constraints, oriented toward data storage management, global data transfer optimization, and concurrence exploitation. The major target application domains are real-time signal and data processing algorithms in image, video and end-user telecoms applications, and data-structure-dominated modules in telecoms networks. Platforms that contain both customized architectures (potentially on an underlying configurable technology) and (parallel) programmable instruction-set processors are targeted. In addition, deep-sub-micrometer technology issues are included.

Dr. Catthoor became a member of the steering board for the VLSI Technical Committee of the IEEE Circuits and Systems Society in 1997. Since 1999, he has served on the steering board for the IEEE TRANSACTION ON VLSI SYSTEMS. He was the Program Chair of the 1997 IEEE International Symposium on System Synthesis and the General Chair of the 1998 ISSS. He was also the Program Chair and Main Organizer of the 2001 IEEE Signal Processing Systems (SIPS) Conference. He was a recipient of the Young Scientist Award from the Marconi International Fellowship Council in 1986. He was an Associate Editor of the IEEE TRANSACTIONS ON VLSI SYSTEMS in 1995 and 1998 and the IEEE TRANSACTIONS ON MULTIMEDIA from 1999 to 2001. Since 2002, he has been an Associate Editor of the *ACM Transactions on Design Automation for Embedded Systems*, and since 1996, he has been an Editor for Kluwer's *Journal of VLSI Signal Processing*.