Impact and Mitigation of Sense Amplifier Aging Degradation Using Realistic Workloads

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

DOI
10.1109/TVLSI.2017.2746798

Publication date
2017

Document Version
Final published version

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

Impact and Mitigation of Sense Amplifier Aging Degradation Using Realistic Workloads

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

***Abstract*— Designers typically add design margins to compensate for time-zero variability (due to process variation) and time-dependent (due to, e.g., bias temperature instability) variability. These variabilities become worse with scaling, which leads to larger design margin requirements. As an alternative, mitigation schemes can be applied to counteract the variability. This paper investigates the impact of aging on the *offset voltage* of the memory's sense amplifier (SA). For the analysis, the degradation of the SAs in the L1 data and instruction caches of an ARM processor is quantified while using realistic workloads extracted from the SPEC CPU2006 Benchmark suite. Furthermore, the effect of our mitigation scheme, i.e., an online control circuit that balances the SA workload, is analyzed. The simulation results show that the mitigation scheme reduces the offset voltage degradation due to aging with up to 40% for the benchmarks, depending on the stress conditions (temperature, voltage, and workload).**

***Index Terms*— Aging, memory, mitigation, offset voltage, sense amplifier (SA), static random-access memory, time-dependent variability, time-zero variability.**

I. Introduction

The downscaling of CMOS technology over the past decades has significantly improved the performance of integrated circuits. However, this downscaling poses major challenges with respect to the device lifetime and reliability [1]. These challenges are caused by increasing variability coming from two sources: manufacturing and operational usage. Due to imperfections in the manufacturing process, devices will suffer from process variations and end up with different characteristics from the intended ones; these process variations are referred to as *time-zero* variability. Variations that occur during the lifetime include *environmental variations* (such as supply voltage fluctations and temperature variations) and *aging variations* due to, for instance, *bias temperature instability* (BTI) [1]; these variations are referred to as *time-dependent variability*. The impact of both time-zero variability and time-dependent variability becomes more severe with CMOS scaling [1].

If countermeasures are not taken against the increasing variability, failure rates of devices will increase. Traditionally, designers use guardbanding, which means that extra margins are added to the circuit, to guarantee that the circuit will function correctly during its lifetime. This method negatively impacts the performance of the design, as it affects its area, power consumption, speed, and/or yield. This is especially the case when the workload dependence is not properly incorporated and only worst case workloads are used. In contrast, we use an analysis flow, which properly incorporates the actual workloads. Moreover, instead of worst case margins, more cost-effective mitigation schemes can be employed to counteract the variability. These mitigation schemes can even adapt the workload by employing online control circuits. This paper focuses on the mitigation of the impact of aging on the *offset voltage* of the memory's sense amplifier (SA). The SA is very important for high-performance memories, as it forms an integral and critical part of the read path delay. The SA behavior influences the memory delay in two ways. First, a larger SA offset voltage requires a larger bitline swing, which means more time must be allocated for the bitline discharge; failing to provision for sufficient swing results in failures in the field. Second, the *sensing delay*, which is the delay from SA trigger to SA output, is on the cricital path. Therefore, understanding the impact of workload-dependent aging on the memory SA offset voltage and providing appropriate mitigation schemes are an important part of designing a robust and reliable memory system.

A lot of work has been published on the impact of aging and their countermeasures for static random-access memory cell arrays. However, very limited work has been done on the characterization and workload-dependent mitigation of aging in memory peripheral circuits such as SAs. In [2], a tunable SA is presented to compensate for within-die variations. In [3], the offset voltage is monitored using an on-chip circuit to estimate the yield. In [4], an accurate method to estimate the impact of both time-zero variability and time-dependent variability on the SA offset voltage is proposed; it considers the SA offset voltage dependence on temperature, voltage, and *workload*, but the mitigation is not the focus here. Furthermore, the workloads are artificial,

Manuscript received April 26, 2017; revised July 14, 2017; accepted August 9, 2017. Date of publication September 28, 2017; date of current version November 22, 2017. *(Corresponding author: Daniël 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 F. Catthoor are with imec, 3001 Leuven, Belgium, and also with ESAT, Katholieke Universiteit Leuven, 3001 Leuven, Belgium (e-mail: pieter.weckx@imec.be; francky.catthoor@imec.be).

S. Cosemans is with SureCore Ltd., 3001 Leuven, Belgium (e-mail: stefan.cosemans.ext@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.2017.2746798

as they are not extracted from applications. Prior work mainly focused on mitigating the SA offset voltage due to time-zero variability. Run-time mitigation schemes for workload-driven time-dependent variability have not been researched.

In our previous work, we proposed the *input switching sense amplifier (ISSA)* [5]. The ISSA switches its inputs periodically in order to create an online control-based balanced workload. Thanks to the balanced workload, the impact of time-dependent variability on the SA offset voltage is minimized. However, artificial workloads were used and the practical effectiveness of the ISSA could not be evaluated. In this paper, we use workloads extracted from applications. The workloads are obtained from the SPEC CPU2006 benchmark suite [6] and are simulated on an ARM processor using gem5 simulator [7]. Using these workloads, the degradation and effect of mitigation are evaluated for the L1 instruction and data cache. The main contributions of this paper are as follows.

1) Analysis of the workload-dependent offset voltage degradation due to BTI with and without mitigation for the ARM Cortex-A9 L1 data and instruction caches using real workloads extracted from applications.
2) Analysis of the offset voltage degradation for the whole L1 cache, including valid, dirty, tag, and data bits.
3) Investigation of impact of supply voltage and temperature on the offset voltage degradation due to BTI.

The rest of this paper is organized as follows. Section II provides the background and discusses the BTI model, the high-performance standard-latch type SA, and the offset voltage specification of the SA. Section III presents the proposed methodology. Section IV quantifies the offset voltage degradation due to aging. Section V analyzes the impact of the mitigation scheme. Finally, Section VI concludes this paper.

II. Background

This section first discusses the BTI aging model used in this paper. Thereafter, the standard latch-type SA and, finally, the method to determine the offset voltage specification are explained.

A. Bias Temperature Instability

Several aging mechanisms exist, e.g., BTI [8], hot carrier injection [9], and time-dependent dielectric breakdown [10]. BTI is considered to be the most important of them [11] and, therefore, it is the focus of this paper. BTI is a failure mechanism that takes place inside the MOS transistors and causes an increment in the threshold voltage (V_{th}). This V_{th} increase happens under *negative gate stress* for pMOS transistors, which is referred to as negative BTI. For nMOS transistors, it happens under *positive gate stress*, which is referred to as positive BTI.

To model BTI, this paper uses the atomistic model presented in [12]; it is based on the capture and emission of traps during stress and relaxation phases of BTI. Each trap contributes to the threshold voltage shift ΔV_{th}. The ΔV_{th} of the transistor is the accumulated result of all gate oxide defect traps. The probabilities of the defects capture P_C and emission P_E are defined as follows [13]:

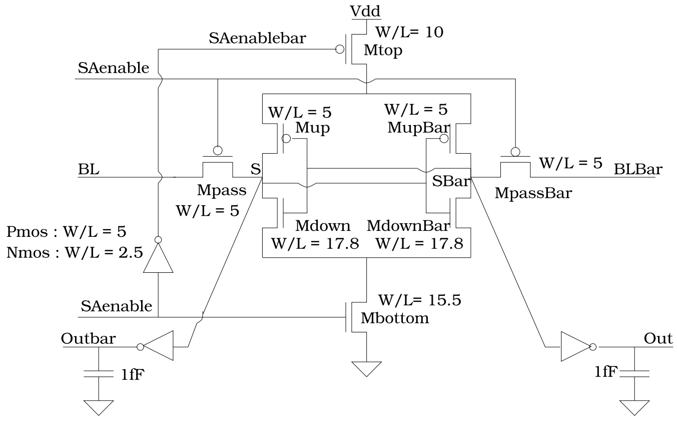


Fig. 1. Standard latch-type SA.

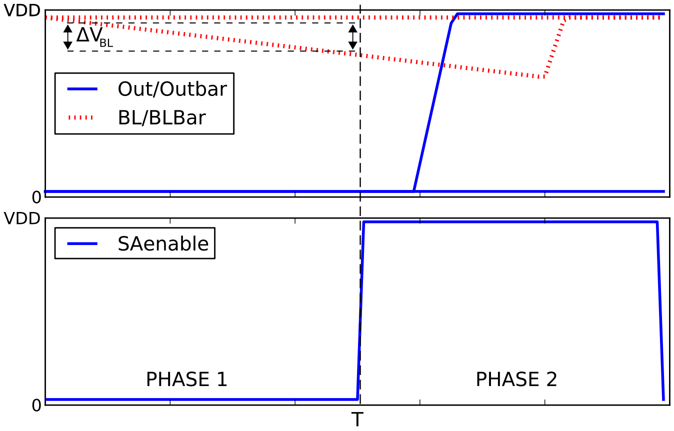


Fig. 2. Timing diagram of SA operation.

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

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

Here, τ_c and τ_e are the mean capture and emission time constants, t_{STRESS} is the stress period, and t_{RELAX} is the relaxation period. The impact of voltage and temperature is also included in the model [12], [14].

B. Sense Amplifier

The standard latch-type SA, shown in Fig. 1, will be used as a case study in this paper. In principle, other types of SAs could also be used as well, such as the look-ahead type SA [15] or the double-tail latch-type SA [16]. The SA amplifies a small voltage difference between bitlines BL and BLBar during read operations. The operation of the SA of Fig. 1 consists of two phases; the timing diagram is shown in Fig. 2. In the first phase, when SAenable is low, the voltage swing on BL and BLBar is passed to the internal nodes S and SBar. In the second phase, when SAenable is high, the pass transistors disconnect the internal nodes from BL and BLBar

Algorithm 1 Binary Search for Offset Voltage

```
Input: Number of $steps$ and search range ($V\_min$, $V\_max$)
Output: Offset Voltage
$V\_try$ = ($V\_min$ + $V\_max$) / 2
for $i = 1 \rightarrow steps$ do
    apply $V\_try$ on BL/BLBar
    if Out is low then
        $V\_min$ = $V\_try$
    else
        $V\_max$ = $V\_try$
    end if
    $V\_try$ = ($V\_min$ + $V\_max$) / 2
end for
return $V\_try$
```

and the amplification of the voltage difference between S and SBar takes place by the cross-coupled inverters. The cross-coupled inverters get current through Mtop and Mbottom and produce the result of the read on outputs Out and Outbar. The strong pull-down transistors, Mdown and MdownBar, of the cross-coupled inverters ensure a low amplification time. Therefore, this type of SA is often used in high-performance memories that require fast response times, such as L1 caches.

C. Offset Voltage Specification

An important metric of the SA is its input offset voltage. The offset voltage is defined as the differential input voltage that results in a differential output voltage of zero. Due to process variation and aging [4], the transistors of the SA are never matched perfectly. As a result, each fabricated SA has a different offset voltage. The offset voltage is an important metric, since it determines the minimum bitline swing ΔV_{BL} required to perform a successful read operation (see Fig. 2). An SA with a lower offset voltage requires a lower bitline voltage swing. Therefore, the memory is faster, since less time is required to generate enough bitline voltage swing.

To determine the offset voltage specification of the SA, the distribution of the offset voltages of the mismatched SAs must be analyzed; we use the same method as applied in [4]. It determines the offset voltage specification through Monte Carlo simulations. The Monte Carlo simulations consider both the impact of time-zero variability (i.e., local process variation) and time-dependent variability (i.e., variation due to aging, temperature, and voltage). During each simulation, the offset voltage of the specific sample is measured. Subsequently, the average and standard deviation of the measured offset voltages are calculated, in order to find its fitted normal distribution. Using the fitted normal distribution, it is then possible to calculate the offset voltage specification of the SA for a given target failure rate. In this paper, a target failure rate of $f_r = 10^{-9}$ is assumed, thus targeting applications with a *high reliability* requirement.

During each Monte Carlo simulation, the input offset voltage of the specific sample is determined using a binary search on its inputs. The binary search algorithm is provided in Algorithm 1 and is implemented in Verilog-A. It finds the offset voltage of the SA by repeatedly applying voltage $V_$try on the inputs of the SA, where $V_$try is the average of the search range ($V_$min and $V_$max). After applying $V_$try, either the lower or upper bound of the search range is updated to $V_$try based on the output value of the SA. Therefore, it eliminates half of the range, where the input offset voltage of the SA cannot lie. The algorithm repeats this process for a certain amount of steps. The algorithm is able to find the offset voltage in logarithmic time with a maximum error of $(V\_\max - V\_\min)/(2^{\mathrm{steps}})$.

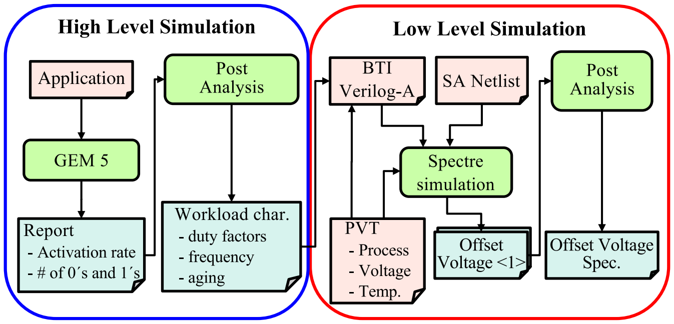


Fig. 3. Framework of the proposed methodology.

III. Methodology

Fig. 3 shows the framework of the proposed methodology. As can be seen, the framework consists of a *high-level* and a *low-level* simulation step. In the high-level simulation step, the workload of the SAs is characterized based on the read and write operations from/to the L1 data and instruction caches for application benchmarks. In the low-level simulation step, the workload characterization in combination with different VT conditions, is used to analyze the impact of aging/BTI on the offset voltage specification of the SAs. Note that the proposed methodology is generic and could also be applied to determine the impact on other metrics, designs (e.g., high-performance or lower power designs), and components (e.g., memory cells) with minor changes. The high- and low-level simulation steps are explained next.

A. High-Level Simulation

In the high-level simulation step, the workloads of the benchmark applications are characterized, which is subsequently used by the low-level aging simulation step. The L1 data and instruction cache accesses are simulated using the cycle-accurate gem5 simulator [7]; gem5 is configured to closely resemble the ARM Cortex-A9 architecture [17] by using the configuration proposed in [18]. The most important configuration parameters for gem5 can be found in Table I. A monitor is added to gem5 between the processor and memory to keep track of the write requests and read responses. Based on these requests and the internal cache architecture, an accurate behavior of the cache is simulated. For instance, when a read request is issued, the tag, valid, data, and dirty bits will be read from each set in the cache. By simulating this behavior, it is possible to characterize the workload of

TABLE I
gem5 CONFIGURATION

Processor	ARM Cortex-A9, single-core, out-of-order, @ 1GHz
L1 Data & Instruction Cache	32KB, 4-way set associative, 32B line size, 1 cycle latency
L2 Cache	512KB, 8-way set associative, 32B line size, 8 cycle latency

the SAs. This characterization is done by keeping track of the number of reads, i.e., the *activation rate*, and how often a zero or one is read per SA. This traffic is application-dependent. Subsequently, a report is generated by gem5 with this information for both the data and instruction caches. Finally, *post analysis* is performed on the report to translate this report into a workload for the SAs. The SA workload is characterized by the duty factor, frequency, and stress time of each transistor.

B. Low-Level Simulation

In the low-level simulation step, the method to calculate the offset voltage specification, discussed in Section II-C, is implemented. The BTI model is incorporated using a Verilog-A module, which calculates the threshold voltage shifts of the transistors based on the SA workload obtained in the previous step. The duty factors are calculated based on the duty factors of a single read operation (extracted from SPICE waveforms), the activation rate, and the amount of zeros and ones read. Furthermore, the BTI impact also depends on the process and the voltage and temperature at which the circuit is stressed. Monte Carlo simulations are performed using Spectre for the SA, while considering the effects of process, voltage, temperature, and BTI. During each Monte Carlo simulation, the parameter of interest is measured, which is the offset voltage in this paper; 800 Monte Carlo simulations are run for each experiment. Finally, *post analysis* is performed on the measured offset voltages to calculate the offset voltage specification for the target failure rate, for which we use $f_r = 10^{-9}$ in this paper.

IV. IMPACT OF AGING ON THE SA OFFSET VOLTAGE

This section evaluates the impact of aging on the SA offset voltage. It discusses the experimental setup followed by the results.

A. Experimental Setup

The methodology of Fig. 3 is used to determine the impact of aging on the SA offset voltage. In the high-level simulation step, several integer and floating point applications from the SPEC CPU2006 benchmark suite [6] are simulated. The motivation to use SPEC CPU2006 is that it is commonly used in academia and it is also used in similar works [19], [20]. The integer applications include *bzip2* (executes a compression algorithm), *gcc* (compiles code), *hmmer* (searches gene sequences), and *omnetpp* (simulates discrete events). The floating point applications include *cactusADM* (simulates general relativity), *GemsFDTD* (simulates computational electromagnetics), *soplex* (performs linear programming optimization), and *tonto* (simulates quantum chemistry). Each application is analyzed based on one billion instructions. From them, the workload characterization is generated for the low-level simulation. The low-level simulation step consists of the standard-latch type SA (Fig. 1) implemented with the 45-nm predictive technology model high-performance library [21]. Using this setup, the impact for three years of aging under nominal conditions is evaluated, i.e., $T = 25$ °C and $V_{\text{dd}} = 1$ V.

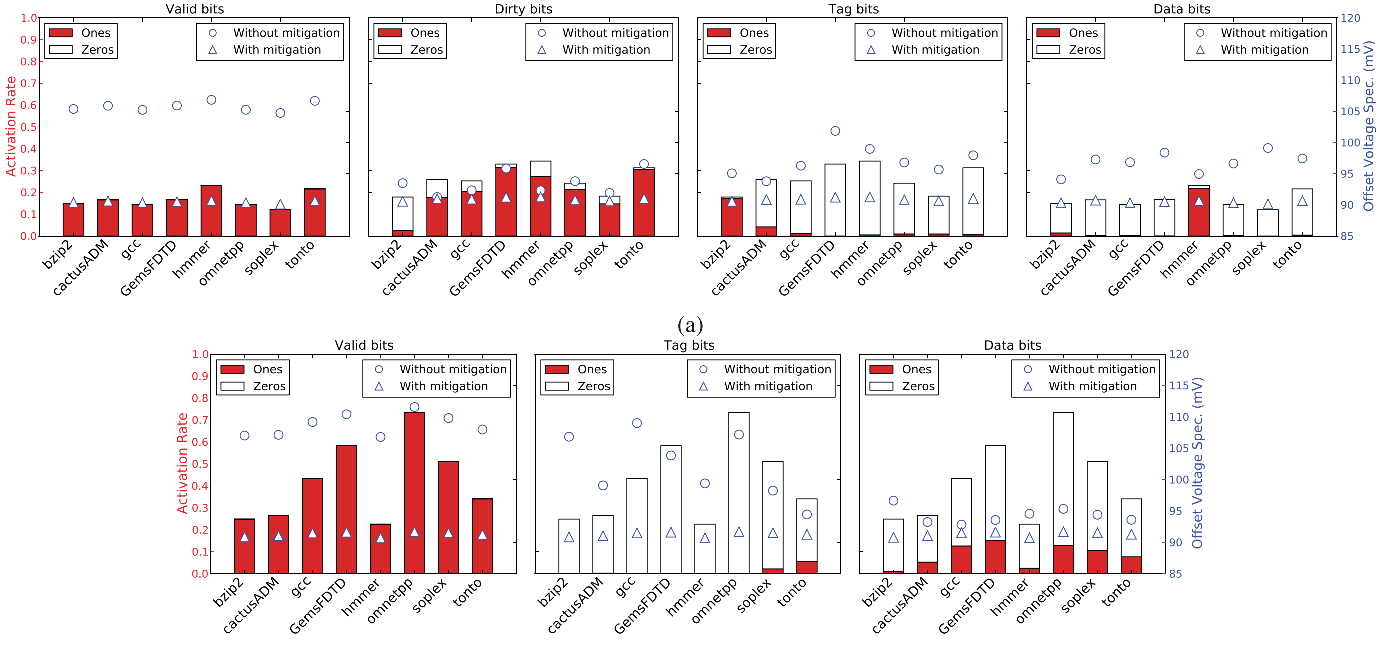
B. Results

Fig. 4(a) shows the offset voltage specification of the SAs for the data cache. Only the worst case SAs have been included per benchmark for the valid, dirty, tag, and data bits. For instance, for the data bits, only one out of every 256 SAs [4 (sets) $\times$ 64 (datawidth) $=$ 256 SAs] is shown. The heights of the bars represent the activation rate of each SA. The white and red colors within a bar represent the ratio of zeros and ones, respectively. Both are used to calculate the offset voltage specification, which is denoted by the circles. The triangles show the offset voltage specification when the mitigation scheme is used; they are further discussed in Section V.

The figure shows that the degradation for the valid bits is the highest. This happens due to very unbalanced SA workloads, i.e., mostly ones are read. Note that once the data starts to be filled in the cache, the valid bits remain high for the remainder of the execution part. The SA offset voltage is very susceptible to these unbalanced workloads. During a read one (zero) operation, the transistors that are responsible for generating a one (zero) at the output are stressed and become weaker. Therefore, the transistors of the SA will have an unbalanced degradation for unbalanced workloads. As a result, the input offset voltage increases. The valid bit for benchmark *hmmer* shows the highest degradation. In this case, the offset voltage specification increases up to 23% compared with the offset voltage specification at time-zero under nominal conditions, which is 87 mV. Note that the time-zero specification is obtained solely based on process variations, i.e., without aging. The degradation of the dirty, tag, and data bits is lower in general, due to more balanced workloads.

Additionally, it can be observed that the valid bits do not show a very strong dependence on the used benchmark, as comparable offset voltage degradations are observed. However, the dirty, tag, and data bits do show a strong dependence on the used benchmarks and larger differences are observed. This can be explained by the bigger unbalance differences in the SA workloads. For instance, when looking at the tag bits, *GemsFDTD* shows a high degradation due to a very unbalanced workload, while *cactusADM* shows a lower degradation due to a more balanced workload.

Similarly, Fig. 4(b) shows the results of the instruction cache. The dirty bits are omitted as there are no writes from the CPU to the instruction cache, and therefore, no data have to be written to a higher memory level (e.g., L2 cache). The results for the instruction cache show, similarly as for the data cache, that the degradation of the valid bits is the highest. This can be explained again by the very unbalanced workloads. The valid



(a)

(b)

Fig. 4. Worst case degradation of offset voltage specification for the L1 data and instruction caches. (a) Data cache. (b) Instruction cache.

bit for benchmark *omnetpp* shows the highest degradation. In this case, the offset voltage specification increases up to 28% compared with the offset voltage specification at time-zero under nominal conditions. Furthermore, on the contrary to the data cache, the tag bits show a significantly higher degradation than the data bits for the instruction cache. This can be explained by the fact that the program instructions use only a certain address range in the memory, while the data, due to a higher memory demand, are stored in a more distributed way. Finally, the data bits of the instruction cache show generally a lower degradation than the data bits of the data cache. The lower instruction cache degradation can be explained by the fact that there is more variation in the data bits due to the execution of different instructions, which use different instruction formats; hence, they lead to a more balanced workload. For the data cache, data representation standards, such as integer and floating point, cause certain bits to be biased toward certain values, such as sign bits and most significant bits (MSBs). As a result, the workload is very unbalanced for some SAs.

Finally, the results of the data and instruction caches reveal that an unbalanced workload has a higher impact on the offset voltage degradation than the activation rate. For example, when looking at the degradation of the data bits of the data cache, the worst case offset voltage is 99.1 mV (activation rate of 0.12, ∼0% ones, ∼100% zeros), while the best case is 94.1 mV (activation rate of 0.15, 9.6% ones, 90.4% zeros). Even though the best case has a higher activation rate, the worst case still has ∼5.3% higher offset voltage specification; hence, the impact of unbalanced workloads is higher. To better illustrate this, we investigated the offset voltage degradation using artificial workloads, with varying percentages of read ones for several activation rates, i.e., 25%, 50%, and 75%. The results for three years of aging under nominal conditions are shown in Fig. 5. The figure clearly reveals that the offset voltage specification is the highest for very unbalanced workloads, where there is a low or high percentage of read ones, and it is the lowest for balanced workloads. Furthermore, we observe that the offset voltage degradation already decreases significantly when the workload is slightly more balanced, e.g., 10% versus 0% read ones. Finally, it can be observed that the activation rate has a weak impact on the offset voltage degradation. It can be concluded from this that as a mitigation scheme, one should first aim at balancing the workloads, rather than lowering the activation rate, e.g., by adding redundant SAs and switching between them. In Section V, we analyze such a workload balancing scheme.

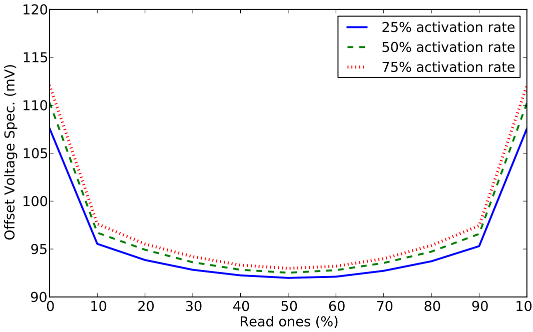


Fig. 5. Offset voltage degradation versus percentage of read ones.

V. Mitigation

This section first discusses the design of the mitigation scheme, followed by the performed experiments and the obtained results.

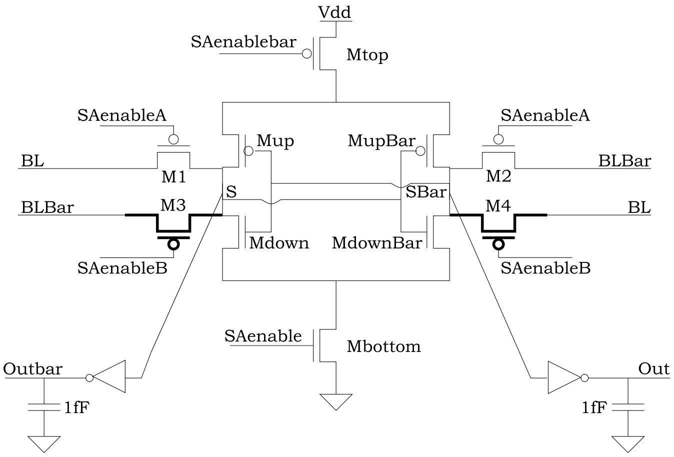


Fig. 6. Input Switching Sense Amplifier [5].

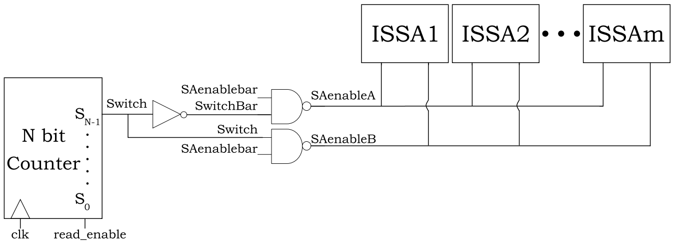


Fig. 7. Control logic for the ISSAs [5].

A. Design

In [4], we showed that the SA has a large offset voltage degradation when unbalanced workloads are applied. Therefore, we proposed the ISSA [5], where the SA switches its inputs periodically in order to create, at run-time, a balanced SA workload. As a result, the degradation of the offset voltage is mitigated.

Fig. 6 shows the structure of the ISSA. It uses a second pair of pass transistors, M3 and M4, to switch the SA inputs. Internal nodes S and SBar can be accessed now by M1 and M2 or M3 and M4. Pass transistors M1 and M2 are controlled by signal SAenableA and pass transistors M3 and M4 by signal SAenableB. When SAenableA is low/enabled and SAenableB is disabled/high, pass transistors M1 and M2 forward the voltage level on BL and BLBar to the internal nodes. Note that, in this case, the SA operates in the same way as that of the standard latch-type SA. When SAenableB is low/enabled and SAenableA is disabled/high, pass transistors M3 and M4 forward the voltage level on BL and BLBar to the internal nodes. However, in this case, the inputs BL and BLBar are switched. When the inputs of the SA are switched, the SA will effectively read the opposite value, i.e., a read one operation becomes a read zero operation and vice versa. Hence, by controlling this switching, it is possible to balance the amount of zeros and ones read by the internal nodes of the SA.

Fig. 7 shows the required control logic for the generation of signals SAEnableA and SAEnableB. Two NAND gates are used to generate SAenableA and SAenableB from the original SAenable(bar) and the Switch signal. When switch is low (high), only SAenableA (SAenableB) is active; SAenableB (SAenableA) is always high in this case and makes sure that the corresponding pass transistors M3 and M4 (M1 and M2) are switched OFF. The switch signal is generated by the MSB of an N-bit counter, which is only updated during reads and controlled by read_enable. Hence, the inputs of the SA switch each 2^{N-1} reads.

B. Performed Experiments

The ISSA is implemented in the L1 data and instruction caches of gem5. Using it, the following mitigation experiments are performed.

1) *Impact of Application:* The impact of several applications from the SPEC CPU2006 benchmark suite is evaluated at nominal temperature and V_{dd} for three years of aging while using a counter size of 1 bit.
2) *Impact of Counter Size:* The impact of the counter size is evaluated at nominal temperature and V_{dd} for three years of aging. The following counter sizes are used: 1, 2, 4, and 8 bit.
3) *Impact of Supply Voltage:* The impact of varying supply voltages is investigated. The following supply voltages are evaluated: $-10\%\ V_{\text{dd}}$, nom. $V_{\text{dd}} = 1.0$ V, $+10\%\ V_{\text{dd}}$.
4) *Impact of Temperature:* The impact of several temperatures is investigated. The following temperatures are evaluated: 25 °C, 75 °C, and 125 °C.

C. Results

1) Impact of Application: The triangles in Fig. 4(a) denote the offset voltage specification of the data cache when the mitigation scheme is applied. The figure shows that the offset voltage specification is significantly reduced in most cases. Especially, for unbalanced workloads, a high reduction is achieved, such as the valid bits. The highest achieved reduction in offset voltage specification is ∼15.1%, which is the case for the valid bits of benchmark *hmmer*. Another observation is that the offset voltage specification is independent of the used benchmark once the mitigation scheme is applied and, therefore, also the bit type (valid, dirty, tag, or data bits). This happens because workload balancing has a higher impact on the offset voltage than the activation rate.

Similarly, the triangles in Fig. 4(b) denote the offset voltage specification of the instruction cache when mitigation is applied. Similar trends to the data cache can be observed for the instruction cache; the offset voltage specification is reduced up to 17.8% and becomes also independent of the used benchmark and bit type.

2) Impact of Counter Size: Fig. 8 shows the offset voltage specification with and without mitigation for various counter sizes. As a case study, only the valid bit in the instruction cache from the benchmark *omnetpp* is shown, as it has the highest degradation among all benchmarks. The height of the bars represents the activation rate of the SA and the white and red colors of the bars the ratio of zeros and ones, respectively. The corresponding offset voltage specification is denoted by “x.” The x-axis contains the used counter sizes, which are denoted by C_n with n as the counter size (e.g., C_8 uses an 8-bit counter). Fig. 8 clearly shows that the impact of the counter size is negligible. The differences in offset voltage

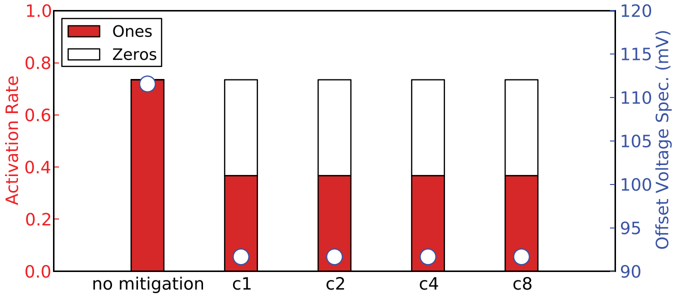


Fig. 8. Impact of counter size on offset voltage for three years of aging for worst case workload.

TABLE II

IMPACT OF COUNTER SIZE ON MEAN AND MAX UNBALANCE ERROR OF ALL BENCHMARKS

Counter Size	C_1	C_2	C_4	C_8
Mean	0.11%	0.05%	0.00%	0.00%
Max	3.60%	6.16%	0.04%	0.02%

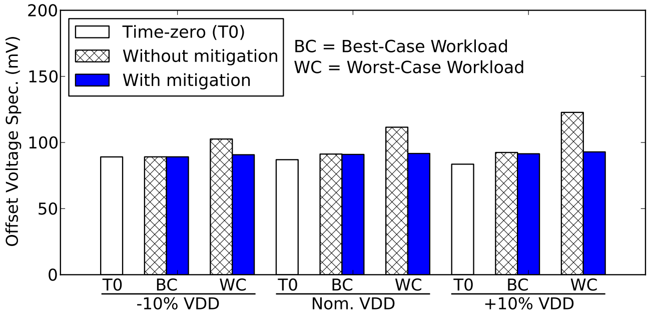


Fig. 9. Impact of supply voltage on offset voltage for three years of aging.

are minimal between all counter sizes, as the resulting SA workloads are all similarly balanced.

We also investigated the impact of the counter size for all the other benchmarks for all bit types. Table II contains the mean and maximum errors with respect to a perfectly balanced workload. The error is expressed as the absolute difference between the percentage of 1's (or 0's) and 50%. For example, a workload with 45% 0's and 55% 1's has an error of 5%. Table II shows that the workload is slightly better balanced for bigger counter sizes, as both the smaller mean and max errors are observed. However, these maximum values were only observed for one benchmark and only for the smaller counter sizes. The other benchmarks had a much lower maximum error. Furthermore, even the biggest error of 6.16% has only a very weak impact on the offset voltage degradation compared with a perfectly balanced workload (see Fig. 5). Therefore, we conclude that the impact of counter size is marginal.

3) Impact of Supply Voltage: Fig. 9 shows the supply voltage dependence of the offset voltage with and without mitigation at nominal temperature for three years of aging. As a case study, the best case (BC) workload and the worst case (WC) workload are shown. The BC workload is the dirty bit of data cache for *cactusADM* (activation rate of 0.26, 68% ones, 32% zeros) and the WC workload is the valid bit of the instruction cache for *omnetpp* (activation rate of 0.73, ∼100% ones, ∼0% zeros). The offset voltage specs at time-zero (T0) are also included.

The figure shows that at higher V_{dd}, the increase in offset voltage specification is significant without mitigation for the WC workload. For example, the offset voltage specification is ∼26% higher at +10% V_{dd} than at −10% V_{dd}. The mitigation scheme reduces the offset voltage specification for the WC workload at higher V_{dd} up to 22.5%. The figure also reveals that the offset voltage specification only shows a marginal improvement for the BC workload when the mitigation scheme is applied. This can be explained by the fact that the BC workload is already quite balanced and further balancing by the mitigation scheme does not give a big improvement. Furthermore, it can be observed that the impact of supply voltage on the offset voltage degradation is significantly smaller when the mitigation scheme is used.

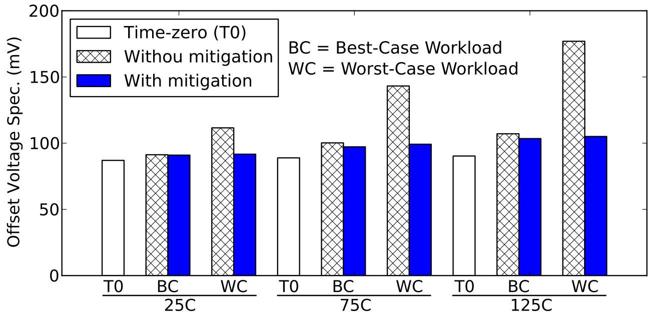


Fig. 10. Impact of temperature on offset voltage for three years of aging.

4) Impact of Temperature: Fig. 10 shows the temperature dependence of the offset voltage with and without mitigation at nominal V_{dd} for three years of aging. Once again, the results for the BC and WC workloads are shown, which are the same ones used in the supply voltage experiments. The offset voltage specs at time-zero are also included. Fig. 10 reveals that the impact of temperature on the offset voltage is much higher than the impact of the supply voltage. The offset voltage specification is ∼64% higher at 125 °C without mitigation than at 25 °C. Furthermore, the offset voltage specification increases up to 100.6% at 125 °C compared with time-zero. The mitigation scheme reduces the offset voltage specification up to 40% at 125 °C for the WC workload. Hence, the offset voltage specification is significantly better with the mitigation scheme. Similar to the supply voltage experiments, the offset voltage specification only shows a marginal improvement for the BC workload using the mitigation scheme. Furthermore, it can be observed that the impact of temperature on the offset voltage degradation is significantly smaller when the mitigation scheme is used.

D. Discussion

This paper investigated the impact and effect of mitigation on the degradation of the offset voltage of the SAs in the L1 data and instruction caches. The results showed that the mitigation scheme reduces the degradation of the offset voltage specification up to 40%. As a result, the bitline swing requirement can be reduced. This leads to a more efficient memory, since less energy is wasted on the bitline discharge. Furthermore, the memory can run at a higher speed, since less time needs to be allocated for the bitline discharge.

The area overhead of the ISSA scheme is negligible. The scheme consists of an *n*-bit counter (to generate *switch*),

TABLE III
AREA OVERHEAD OF ISSA

Component	Transistors per component	Total transistors
Inverter	2	2
NAND	4	8
Frequency Divider	-	70 (n=4)
Sense Amplifier	2	780
Total	-	860

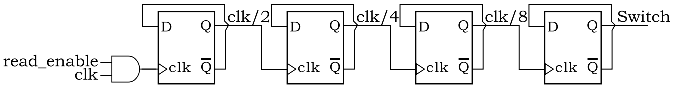


Fig. 11. Frequency divider.

an inverter (to generate *SwitchBar*), two NAND gates (to generate *SAenableA* and *SAenableB*), and two additional pass transistors, per SA (to select between its inputs). Table III contains the overhead of each component and the total overhead in terms of transistor count. The inspected L1 cache has a total of 340 SAs; 256 for the data bits, 76 for the tag bits, 4 for the dirty bits, and 4 for the valid bits. Hence, this corresponds to an overhead of 780 transistors. Instead of a counter, a frequency divider is used to create an equivalent signal as the highest MSB of a 4-bit counter. The circuit is shown in Fig. 11; the 4-bit counter showed a lower maximum error than the 1-bit and 2-bit counters, while the 8-bit counter gives a negligible improvement (see Table II). The frequency divider can be implemented with four flip-flops and an AND gate. This results in a total of 70 transistors when the flip-flops are implemented with 16 transistors [22]. The investigated L1 cache of the ARM Cortex-A9 consists of a total of 283 648 memory bits (when including the data, tag, dirty, and valid bits). This results in a total of 1 701 888 transistors when a 6T cell is used. Therefore, when the transistor count of the memory array and the mitigation scheme are compared, a totally negligible overhead of 0.051% is obtained. Note that the peripheral circuitry is not even included yet in this analysis. Therefore, the actual area overhead is even lower.

In terms of power overhead, the main added power comes from the counter/frequency divider, which is activated each read cycle. Note that the power consumption of the two extra NAND gates to generate SAenableA and SAenableB can be neglected; only one of the two gates is active at a time and in the design without the mitigation scheme, a similar gate is needed to drive the SAs. Compared with the leakage power of the memory matrix and the power consumption to drive the wordlines and discharge the bitlines of the memory, it can be expected that the power consumption of the counter is negligible.

Finally, the delay overhead of the scheme was investigated in [5]. The results showed that the scheme improves the sensing delay by up to 10% thanks to the balanced workload. It must be noted that when the inputs are switched, the resulting read value should still be inverted. Additional circuitry is needed for this (e.g., two transmission gates to select between the $data$ and $\overline{data}$ values from the output latch). However, the delay impact of this is marginal.

Therefore, the mitigation scheme provides a large gain at negligible overheads.

VI. CONCLUSION

In this paper, we investigated the impact of aging on the offset voltage of the SAs in the L1 data and instruction caches of an ARM processor. We proposed a mitigation scheme to minimize the impact of aging. The results showed that the mitigation scheme reduces the offset voltage degradation up to 40%. This means that the bitline swing of the memory can be reduced during read operations, which leads to a faster and more efficient memory. Therefore, this paper clearly shows that run-time mitigation schemes are a good alternative or supplement to the traditional guardbanded designs. They provide a more optimal design, without compromising on the lifetime and reliability. The mitigation schemes are extremely important for the deployment of cutting edge technology as they suffer from a reduced lifetime.

REFERENCES

[1] J. W. McPherson, “Reliability trends with advanced CMOS scaling and the implications for design,” in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2007, pp. 405–412.

[2] M. H. Abu-Rahma *et al.*, “Characterization of SRAM sense amplifier input offset for yield prediction in 28 nm CMOS,” in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2011, pp. 1–4.

[3] J. Vollrath, “Signal margin analysis for DRAM sense amplifiers,” in *Proc. 1st IEEE Int. Workshop Electron. Design, Test Appl.*, Jan. 2002, pp. 123–127.

[4] 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.

[5] D. Kraak *et al.*, “Mitigation of sense amplifier degradation using input switching,” in *Proc. Design Test Conf. Eur. (DATE)*, Lausanne, Switzerland, Mar. 2017, pp. 858–863.

[6] J. L. Henning, “SPEC CPU2006 benchmark descriptions,” *ACM SIGARCH Comput. Archit. News*, vol. 34, no. 4, pp. 1–17, Sep. 2006.

[7] N. Binkert *et al.*, “The gem5 simulator,” *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, 2011.

[8] B. Kaczer *et al.*, “Atomistic approach to variability of bias-temperature instability in circuit simulations,” in *Proc. IRPS*, Apr. 2011, pp. 915–919.

[9] F. Cacho *et al.*, “Hot carrier injection degradation induced dispersion: Model and circuit-level measurement,” in *Proc. IEEE Int. Integr. Rel. Workshop Final Rep.*, Oct. 2011, pp. 137–141.

[10] K. B. Yeap, F. Chen, H. W. Yao, T. Shen, S. F. Yap, and P. Justison, “A realistic method for time-dependent dielectric breakdown reliability analysis for advanced technology node,” *IEEE Trans. Electron Devices*, vol. 63, no. 2, pp. 755–759, Feb. 2016.

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

[12] B. Kaczer *et al.*, “Origin of NBTI variability in deeply scaled pFETs,” in *Proc. IRPS*, May 2010, pp. 26–32.

[13] M. Toledano-Luque *et al.*, “Response of a single trap to AC negative bias temperature stress,” in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2011, pp. 4A.2.1–4A.2.8.

[14] 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.

[15] T. Asano *et al.*, “Low-power design approach of 11FO4 256-Kbyte embedded SRAM for the synergistic processor element of a cell processor,” *IEEE Micro*, vol. 25, no. 5, pp. 30–38, Sep. 2005.

[16] D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, “A double-tail latch-type voltage sense amplifier with 18ps setup+hold time,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 314–605.

[17] “Cortex-A9 technical reference manual (ID050110),” ARM Ltd., Cambridge, U.K., Tech. Rep., Apr. 2010.

[18] F. A. Endo, D. Couroussé, and H.-P. Charles, “Micro-architectural simulation of in-order and out-of-order ARM microprocessors with gem5,” in *Proc. Int. Conf. Embedded Comput. Syst., Archit., Modeling, Simulation (SAMOS XIV)*, Jul. 2014, pp. 266–273.

[19] N. Rohbani, M. Ebrahimi, S.-G. Miremadi, and M. B. Tahoori, "Bias temperature instability mitigation via adaptive cache size management," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 3, pp. 1012–1022, Mar. 2017.

[20] A. Valero, N. Miralaei, S. Petit, J. Sahuquillo, and T. M. Jones, "On microarchitectural mechanisms for cache wearout reduction," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 3, pp. 857–871, Mar. 2017.

[21] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2816–2823, Nov. 2006.

[22] U. Ko and P. T. Balsara, "High performance, energy efficient master-slave flip-flop circuits," in *IEEE Symp. Low Power Electron. Dig. Tech. Papers*, Oct. 1995, pp. 16–17.

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

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 the Delft University of Technology, Delft, The Netherlands.

He is currently a Post-Doctoral Researcher with the Dependable Nano-Computing 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) received the M.Sc. degree in computer engineering from the Delft University of Technology, Delft, The Netherlands, in 2010, where he is currently pursuing the Ph.D. degree.

His current research interests include the reliability analysis, design, monitoring, and mitigation of memory systems.

Said Hamdioui (M'99–SM'11) was with Intel Corporation, Mountain View, CA, USA, Philips Semiconductors Research and Development, Crolles, France, and Philips/NXP Semiconductors, Nijmegen, The Netherlands. He is currently a Chair Professor on dependable and emerging computer technologies with the Computer Engineering Laboratory, Delft University of Technology, Delft, The Netherlands. He has authored or co-authored one book and co-authored over 170 conference and journal papers. He holds one patent. His research focuses on two domains: dependable CMOS nanocomputing, including reliability, testability, and hardware security, and emerging technologies and computing paradigms, including 3-D stacked ICs, memristors for logic and storage, and in-memorycomputing.

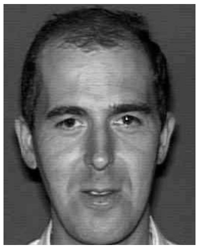
Dr. Hamdioui is also a member of the AENEAS/ENIAC Scientific Committee Council (AENEAS =Association for European NanoElectronics Activities). He delivered dozens of keynote speeches, distinguished lectures, and invited presentations and tutorial at major international forums/conferences/schools and at leading semiconductor companies. He is an Associate Editor of the IEEE TRANSACTIONS ON VLSI SYSTEMS. He serves on the Editorial Board of the *IEEE Design & Test* and the *Journal of Electronic Testing: Theory and Applications*.

Pieter Weckx (M'14) received the M.Sc. degree in nanoscience and nanotechnology from Katholieke Universiteit Leuven, Leuven, Belgium, in 2011, where he is currently pursuing the Ph.D. degree in electronics and electrical engineering.

His current research interests include the modeling and simulation of time-dependent variability problems in nanoscaled electronic devices, statistical circuit simulations, and stochastic/deterministic clustering of circuit degradation behavior.

Stefan Cosemans (S'04–M'10) received the M.S. degree in electrical engineering and the Ph.D. degree, with a focus on the variability-aware design of embedded memories, from Katholieke Universiteit Leuven, Leuven, Belgium, in 2004 and 2009, respectively.

He is currently a Principal Design Engineer with SureCore Ltd., Leuven, where he is involved in developing low power, low voltage-capable SRAM IP.



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

From 1983 to 1987, he was a Researcher in 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 headed several research domains in the area of high-level and system synthesis techniques and architectural methodologies with the Systems Division, imec, Leuven. Since 1989, he has been an Assistant Professor with the Department of Electrical Engineering, KU Leuven, where he has been a part-time Full Professor since 2000. He is currently a fellow with imec. His current research activities belong to the field of 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-submicrometer 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. He was a recipient of the Young Scientist Award from the Marconi International Fellowship Council in 1986. He was the Program Chair of the 1997 IEEE International Symposium on System Synthesis (ISSS) and the General Chair of the 1998 ISSS. He was also the Program Chair and the Main Organizer of the 2001 IEEE Signal Processing Systems Conference. Since 1999, he has served on the Steering Board for the IEEE TRANSACTIONS ON VLSI SYSTEMS. He was an Associate Editor of the IEEE TRANSACTIONS ON VLSI SYSTEMS from 1995 to 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 of the *Journal of VLSI Signal Processing* (Kluwer).