Analyzing the Effects of Peripheral Circuit Aging of Embedded SRAM Architectures

Josef Kinseher*, Leonhard Heiß*, Ilia Polian†
*Intel Deutschland, 85579 Neubiberg, Germany
†University of Passau, 94032 Passau, Germany
Email: {josef.kinseherleonhard.heiss}@intel.com, ilia.polian@uni-passau.de

Abstract—Modern System-on-Chips rely heavily on the performance of their embedded memories which are also most susceptible to the increasing reliability challenges of today’s nanoscale technology nodes. However, in contrast to memory core-cells, the effects of transistor aging inside the peripheral logic of SRAM architectures have received little attention. This study works out how BTI and HCI induced wear-out of the peripheral SRAM circuitry impacts various performance metrics of an industrially used memory library. We show that the degradation of the peripheral logic is the dominant driver for access speed loss while it tends to slightly lower memory read margin and lead to minor improvements of write margin. We furthermore show that in terms of access time margin the degradation of SRAM control circuitry counteracts aging effects inside core-cells and sense amplifiers. Surprisingly, wear-out of peripheral circuitry can even improve access time margin in case when the relative magnitude of PBTI is much lower compared with NBTI. Based on the example of an embedded memory library, this study further underlines the importance to analyze aging mechanisms at system level rather than for its individual interacting sub-circuits.

I. INTRODUCTION

With the continuous trend of technology shrinking and higher on-chip integration densities while simultaneously reaching the limits of voltage scaling circuit reliability has become a major concern [1]. Larger internal electrical fields as well as elevated on-chip temperatures accelerate transistor aging effects which in turn degrade circuit performance over lifetime. Among all chip components, Static Random Access Memories (SRAMs) have highest integration densities, smallest feature sizes and pose performance bottlenecks in high-performance chip architectures. Embedded SRAMs are therefore especially sensitive to these time-dependent degradation mechanisms which exacerbate design margins and ultimately lead to functional failures. Thus, in order to avoid overly pessimistic margins without compromising SRAM reliability accurate predictions of aging effects on memory failure mechanisms are essential.

The degradation of SRAM read and write margin is mainly driven by core-cell transistor strengths, and has already been thoroughly analyzed [2], [3]. Memory access time margin, however, relies not only on the core-cell but to a large extent on the SRAM peripheral circuitry which is affected by aging as well. For example, accurate statements on SRAM access time margins require to consider not only the sense amplifier but also the core-cell and control circuitry driving its inputs. Previous studies have already investigated the aging behavior of various SRAM peripheral sub-components such as sense amplifier [4], write driver [5], timing control logic [6] as well as of some internal memory paths [7], [8]. However, results from studies on completely or partially isolated building blocks cannot be extrapolated to complex SRAM libraries and the individual aging rates can hardly be interrelated to identify the most critical sub-components. Conclusions on overall SRAM failure mechanisms require a simultaneous analysis of all of these components. It is therefore still an open question whether the concurrent wear-out of the SRAM peripheral logic tends to either aggravate or mitigate core-cell aging effects.

For the first time, this work analyzes the interaction of wearout in all essential components of an industrial SRAM library including varying workloads, temperature and voltage stresses. It particularly works out how long-term degradation of SRAM peripheral logic influences overall memory performance metrics and determines for each functional failure mode the most critical building blocks. Most importantly, we show at the example of memory access time margin that, even though aging degrades the performance of each individual memory sub-component, their interaction compensates for these effects at system level. This behavior can even lead to improvements of overall memory access time margin.

The rest of the paper is structured as follows: The next section describes the studied aging phenomena and presents the simulation setup. Section III analyzes the access speed degradation of embedded memories. Section IV studies how peripheral logic impacts access margins while Section V analyzes its influence on read and write margins. Section VI concludes the paper.

II. PRELIMINARIES

For this study, we consider Hot Carrier Injection (HCI) and Bias Temperature Instability (BTI) aging effects causing long-term transistor parameter shift during IC operation. Understanding and predicting their influence on all memory components is essential for the design of both high-performance and reliable memory libraries. We use the RelXpert simulator from Cadence with its proprietary AgeMOS model for HCI- and BTI-induced degradation to compute an aged transistor netlist from a memory circuit. The AgeMOS version used in this study features a temporal deterministic aging model and the simulation results need to be considered as average values. The AgeMOS model parameters have been extracted from silicon stress measurements on MOSFETs manufactured in 28nm high-k process technology.

A. Bias Temperature Instability (BTI)

BTI aging gradually deteriorates transistor current drive-strength by shifting the absolute threshold voltage value $|V_{TH}|$. 

978-3-9815370-8-6/17/$31.00 ©2017 IEEE 852
Its effects on PMOS devices under negative gate stress have been observed for a long time and are known as NBTI (Negative Bias Temperature Instability). With the introduction of high-k metal gate technology a similar wear-out mechanisms known as PBTI (Positive Bias Temperature Instability) became significant for NMOS devices under positive gate stress. In both cases the $V_{th}$ drift recovers to some extent after the stress is removed resulting in a differentiation between stress and recovery phase. The BTI phenomenon is commonly attributed to charge trapping in the gate dielectric or the interface between the gate dielectric and the substrate. The most prevalent models describe it as a Reaction-Diffusion process [9] or by an atomic-scale model based on the capture and emission of single traps during stress and relaxation phase [10]. Nevertheless, the exact physical mechanisms causing BTI are still not fully understood and a topic of active research.

B. Hot Carrier Injection (HCI)

The term ‘hot carriers’ refers to electrons or holes that gain high kinetic energy when traveling through the transistor channel. These carriers are accelerated by high lateral electrical fields across the channel and lead to device degradation if their energy is high enough to escape from the channel and to be injected into the gate/substrate interface. Thus, high transistor drain-source voltages in combination with short gate lengths increase overall device wear-out. HCI induced degradation takes place in both NMOS and PMOS devices and manifests itself in higher transistor threshold voltages. Moreover, as high drain-source voltages across the transistor channel imply an output change in the corresponding CMOS logic gate, HCI aging gets worse with higher gate switching activity as well as circuit clock frequency.

C. Underlying memory model

The underlying memory of this study is an industrial embedded single-port SRAM library for low power applications manufactured in 28nm high-k metal gate technology. The memory features several performance modes and was run at a nominal operation frequency of 500 MHz. Advanced SoCs easily contain a four-digit number of such memory instances in varying sizes and dimensions. Figure 1 depicts the main sub-components and their interaction during a read operation. These are as follows:

- The core-cell array contains the bit cells arranged in several sub-blocks to reduced bit line load. The examined 6T-memory cells of this study have transistor size ratios 1:1:1 if not otherwise mentioned.
- Row decoders select the word line of the accessed cell.
- Column decoders select the bit lines of the accessed cell and connect it to the corresponding read and write circuitry.
- The control unit generates internal memory signals to regulate the timing of the memory. This is usually based on a replica bit line based self-timing mechanism [11]. Furthermore, several control pins (CTRL) allow to adjust the internal speed and performance mode of the memory module.
- The R/W circuitry comprises bit line pre-charge logic as wells as latch-type voltage sense amplifiers and write drivers to read or write the memory cells.

Fig. 1: A simplified block diagram of a read operation. The signal paths for starting and ending the operation are highlighted in blue and red color, respectively.

One important property of virtually all embedded memories is their self-timing feature. That is, as outlined above, a rising clock edge initiates memory operations and the memory ends the operation itself using internal timing delay paths without considering the falling clock edge. This allows the memory to adapt itself to various process, temperature and voltage conditions while considering various workloads, voltages and temperature conditions. Details about the self-timing mechanisms inside the studied memory module are presented in [12]. Most importantly, the self-timing behavior causes signals which trigger the end of a memory operation to have much longer logic paths than those starting the operation. As we shall see later, this has fundamental effects on the aging of memory libraries.

III. DEGRADATION OF SRAM ACCESS SPEED

Embedded memory blocks are often performance bottlenecks of critical circuit paths. Therefore, memory access speed, i.e. the time delay between applying a read operation and observing the data value at the memory output, is an important performance parameter. In this section, we analyze how BTI and HCI mechanisms impact SRAM access speed while considering various workloads, voltages and temperature conditions.

Workloads define switching activity and duty cycles of individual transistors and thus the extent to which they are affected by aging. In order to determine the most critical
workload profiles for memory access speed degradation, we apply various March stress patterns to the memory. They are run at a power-supply voltage ($V_{DD}$) of 1.1 V, 125°C environmental temperature and a simulated stress time of $10^8$ s. The March patterns were executed on a 2x2 section of the core-cell array. This ensures that all relevant paths inside the SRAM library are stressed while keeping simulation time within acceptable limits. Figure 2 depicts the applied test sequences and their impact on memory access speed. Each sequence consists of write (w) and read (r) operations which traverse all cells of the memory section in increasing address order (↑) before the next operation is performed. The notation $i^n$ denotes n idle clock cycles.

The figure shows that workloads W1-W3 which access the memory on average at one percent of all clock cycles (four memory accesses for every read and write operation together with 792 idle cycles) cause substantially less degradation than the corresponding extreme cases W4-W6 accessing the memory every single clock cycle. Moreover, comparing workloads W1 and W2 (W4 and W5) reveals that the data value stored inside a core cell has only minor impact on memory access speed. As the data value determines the wear-out of the internal cell transistors this suggests that overall access speed degradation is driven by the peripheral memory logic rather than the core-cell. Moreover, workload W3 (W6) which has the highest percentage of read operations only induces marginally higher degradation than the other workload profiles. This suggests that overall memory speed degradation is primarily driven by common logic paths inside the memory rather than circuitry which is specifically dedicated to either read or write operations.

Next, we further study how various voltage and temperature conditions influence access speed degradation using the worst-case workload profile W6. During that, the memory operation frequencies have been adapted to the respective performance modes of the memory. The results depicted in Figure 3 show that lower supply voltages and temperatures are efficient means to mitigate memory speed degradation. Besides that, the last two column groups of the chart reveal that the most severe performance impacts occur after switching from high- to low-$V_{DD}$ mode. This represents the most critical case as operating the memory at high-performance mode induces maximum $V_{th}$ drifts while lowering the supply voltage afterwards makes the transistors most sensitive to the induced $V_{th}$ drifts [13].

The results presented here indicate that overall access speed degradation is driven by the peripheral circuitry of the memory. Therefore, SRAM aging mitigation techniques such as bit flipping schemes [3] which focus on lessening core-cell wear-out are little useful to reduce access speed loss. As the most-severe path delays occur after switching from high-to low-$V_{DD}$ mode one potential mitigation scheme might be to postpone transitions into low-performance modes after the memory has experienced intense stress to better exploit BTI recovery effects. However, this can hardly be handled at circuit level and needs to be addressed by cross-layer approaches. In the next sections we study the SRAM periphery wear-out in more detail and examine how it affects the functional read, write and access time margins of memory architectures.

IV. IMPACT OF SRAM PERIPHERY DEGRADA TION ON ACCESS TIME MARGIN

Functional SRAM failure margins are naturally affected by transistor wear-out and drift over the operational lifetime of memories. While read and write margins are mainly determined by internal cell transistor strengths access time margin additionally depends largely on the performance of the peripheral memory components. In this section we study in detail how aging of a complete memory library affects its access time margins.

A. Metrics

During memory read events the accessed cell discharges one of its corresponding pre-charged bit lines BL or BLB developing the differential voltage signal $\Delta V_{BL}$. Subsequently, the sense amplifier is activated and amplifies $\Delta V_{BL}$ to a full-swing differential signal which is then interpreted as a logic value by the data output circuit. An access time failure occurs if the weak differential bit line voltage cannot be translated to a full-swing signal within a given time period. This happens if $\Delta V_{BL}$ is not sufficiently high or the sense amplifier is too slow. Thus, the degradation of access time margin is governed
by the aging rate of the sense amplifier and the drift of $\Delta V_{BL}$. Accurately predicting memory access time margin requires therefore the analysis of all interacting memory components driving these parameters: the timing control circuitry, the sense amplifier as well as the memory core-cells. We use the following metrics to quantify their individual impact:

**Sensing delay:** The sensing delay defines the time period between the activation of the sense amplifier and the point when the weak differential input voltage has been converted to a full swing output signal.

**Sensing time margin:** The sensing time margin defines the time period between the generation of the full swing output signal and the deactivation point of the sense amplifier.

**Bit line voltage swing ($\Delta BL$):** The input voltage difference of the sense amplifier at its activation point. It needs to exceed the sense amplifier’s minimum offset voltage in order to prevent sensing faults.

These metrics are illustrated in Figure 4. The signals WL and SA_en denote the word line and sense amplifier activation signal, respectively, which are both driven by the SRAM control circuitry. SA_out and SA_outb are the full swing differential sense amplifier output signals while BL and BLB denote the bit line voltages of the accessed cell. All measurements are performed when the respective trigger or target voltage levels reach 50% of $V_{DD}$.

### B. Sense Amplifier Speed

As has been discussed above, the sensing delay quantifies the speed of the sense amplifier, whereas the sensing time margin describes the remaining time period between the successful sensing operation and the deactivation point of the sense amplifier. Thus, it allows to assess the severity of sense amplifier speed degradation.

In order to specifically work out the effects of aging on these metrics we apply the most critical stress setting from Figure 3, i.e. we stress the memory block using the high-performance mode and analyze how this affects its operation under lower voltages. We use workload profile W6 as it maximizes the sense amplifier degradation due to its high percentage of read operations of the same data value. Moreover, random process variations were injected into the core-cells and sense amplifier circuitry. Due to high simulation overhead, we had to restrict the analysis to 100 Monte Carlo samples, however, this small set already reveals a clear trend as depicted in Figure 5. It describes how the mean values of the sensing delay and the sensing time margin develop over various stress periods. The sensing delay increases only slightly over an operation time of $10^8$s while, at the same time, the sensing margin does not worsen but even improves to a much larger extent due to the simultaneous wear-out of the SRAM control logic. It turns out that aging slows down the internal gate delay chain deactivating the sense amplifier much more than it worsens its sensing speed. This can be explained by cumulative aging effects along that delay path which greatly exceed the degradation of the sense amplifier transistors. Since every read operation causes concurrent stress on both the sense amplifier as well as the timing logic controlling its activation pulse width both components are exposed to equivalent workloads. As a result, the aging of the SRAM control circuitry offsets the speed degradation of the sense amplifier in terms of access time margin.

### C. Sense Amplifier Offset Voltage

Another important parameter of the characterization of SRAM access time margin is the minimum offset voltage of the sense amplifier. If the bit line voltage swing $\Delta BL$ cannot exceed it, read operations will fail despite a sufficiently large sensing time margin as its cross-coupled inverters remain in their meta-stable state. Therefore, we conduct the following Monte Carlo experiment to analyze how transistor wear-out shifts the sense amplifiers minimum offset voltage: First, we subject a sense amplifier design to random process variations by injecting random $V_{th}$ fluctuations. Then, we add various PBTI and NBTI-induced $V_{th}$ drifts. Table I shows the shifts of the mean of these drifts for various stress periods. The sensing delay increases only slightly over an operation time of $10^8$s while, at the same time, the sensing margin does not worsen but even improves to a much larger extent due to the simultaneous wear-out of the SRAM control logic. It turns out that aging slows down the internal gate delay chain deactivating the sense amplifier much more than it worsens its sensing speed. This can be explained by cumulative aging effects along that delay path which greatly exceed the degradation of the sense amplifier transistors. Since every read operation causes concurrent stress on both the sense amplifier as well as the timing logic controlling its activation pulse width both components are exposed to equivalent workloads. As a result, the aging of the SRAM control circuitry offsets the speed degradation of the sense amplifier in terms of access time margin.

![Fig. 4: The metrics used to characterize memory access time margin degradation.](image)

![Fig. 5: Peripheral circuit aging improves the sensing time margin which compensates for higher sensing delays.](image)

**Table I:** The shift of mean minimum offset voltage for various NBTI- and PBTI-induced $V_{th}$ drifts.

<table>
<thead>
<tr>
<th>Stress Period [s]</th>
<th>NBTI $V_{th}$ drift</th>
<th>PBTI $V_{th}$ drift</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 mV</td>
<td>10 mV</td>
<td>15 mV</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>-0.39 mV</td>
<td>-0.25 mV</td>
</tr>
<tr>
<td>Balanced</td>
<td>0.20 mV</td>
<td>0.63 mV</td>
</tr>
</tbody>
</table>

**Fig. 5:** Peripheral circuit aging improves the sensing time margin which compensates for higher sensing delays.
the distributions. The standard deviation $\sigma$ was found to be nearly constant.

The first column of the table denotes the workload to which the internal sense amplifier transistors are exposed during a read operation. In the balanced case, reading '0' and '1' is equally probable. This leads to a balanced transistor degradation which does not, as shown in the table, cause major degradation of minimum offset voltage. It also reveals that the sense amplifier's offset voltage is very susceptible to asynchronous PBTI stress of its pull-down transistors while the effects of NBTI inside the pull-up devices are minor. This behavior stems from the fact that both bit lines are close to $V_{DD}$ when the sense amplifier is activated which causes the NMOS device pair to immediately operate in saturation region while the PMOS devices are turned off. Hence, device mismatches inside the NMOS transistors have dominant influence for the sense amplifier's decision to exit its metastable region while those inside PMOS devices are less important [14]. The behavior observed in Table I is important for modeling the wear-out of sense amplifiers because it makes PBTI the commanding effect for the degradation of their offset voltage by further worsening initial transistor mismatches.

### D. Bit Line Voltage Swing

The voltage difference between the bit lines $\Delta V_{BL}$ is measured at the point when the sense amplifier activates; it depends on the degradation of the accessed core cell as well as the control logic which times the activation point of the sense amplifier. Triggering the sense amplifier too early causes read failures while activating it unnecessarily late leads to longer read cycles and higher power dissipation. Therefore, replica bit line based self-timing techniques are commonly used which tightly track the bit line discharge delay over various memory sizes and PVT conditions [11]. With aging of this timing control path the sense amplifier activation gets delayed which in turn gives the bit lines more time to discharge resulting in a higher differential voltage $\Delta V_{BL}$. However, at the same time PBTI effects on pull-down cell transistors counteract this trend and weaken the path over which bit lines discharge. Depending on SRAM workload and cell transistor duty factors core-cells and peripheral circuits degrade with different rates and therefore both need to be analyzed separately.

First, we exclusively subject the peripheral circuits of the SRAM architecture to aging stress in order to specifically work out how their wear-out influences the bit line swing metric. Moreover, as the memory cell type determines the discharge path of the bit lines we also include a low-power cell design having a stronger pull-down device with transistor ratios 1:1:2 into the analysis. The simulation results for both high-density as well as low-power cells when applying the worst case setup from Figure 3 are depicted in Figure 6. Most importantly, it shows that degradation of the SRAM periphery improves the differential bit line swing and therefore access time margin due to the reasons described above. Higher memory workloads even amplify this trend.

Next, we evaluate the extent to which aging of the internal core-cell transistors counteracts that behavior. We apply the same simulation setup whereas this time we only subject the core-cells to various PBTI-induced $V_{th}$ drifts degrading the strength of the NMOS pull-down transistor and therefore the bit line discharge path. Its impacts on the bit line voltage swing metric are also depicted in Figure 6. It shows that $\Delta V_{BL}$ of high-density cells is most susceptible to PBTI weakening their discharge path, however the improvements due to simultaneous aging of the periphery are also highest for this cell type. Moreover, comparing both figures reveals that aging of the SRAM peripheral circuits can offset strong PBTI effects inside the core-cell transistors.

### E. Evaluation

The discussed results show that SRAM control logic wear-out compensates for aging effects inside SRAM core-cells and sense amplifiers. It offsets slower sense amplifiers by widening their activation window. Furthermore, it increases the differential bit line voltage swing which counteracts degrading sense amplifier offset voltages and weaker core-cell discharge paths. As every memory operation induces stress on the control circuitry, higher workloads even amplify this trend. Most importantly, the wear-out of SRAM control logic is susceptible to both NBTI as well as PBTI effects while degradation of sense amplifier offset voltages and internal core-cell discharge
paths is solely dominated by PBTI. Thus, this behavior can cause aging to even improve overall access time margins provided that the relative magnitude of PBTI is sufficiently low compared with NBTI.

V. IMPACT OF SRAM PERIPHERY DEGRADATION ON READ AND WRITE MARGIN

In this section, we briefly analyze how aging of the SRAM periphery impacts memory read and write margins. We only work out the trend direction as these margins are primarily determined by the internal core-cell transistor strengths and worsen with higher variability and lower supply voltages.

A. Read margin

During a read event the word line activates and exposes the core-cell to pre-charged bit lines which causes the voltage level of the cell node storing 0 to increase to a positive value. If it exceeds the trip point of the opposite inverter the cell flips and a read or stability fault occurs. SRAM read margin is subjected to both NBTI as well as PBTI whose effects tend to magnify each other and cause severe loss of internal core-cell stability [15]. The simultaneous wear-out of the SRAM peripheral circuitry leads to delays in memory operation and therefore also shifts the word line activation as depicted in Figure 7. Most importantly, its deactivation is shifted by a much greater extent, as the respective triggering mechanism involves a longer timing path which in turn also experiences higher cumulative aging delays. Due to this fundamental behavior SRAM aging enlarges the word line window length and therefore also the time period the core-cell is exposed to the bit lines. In terms of read margin, this causes higher cell disturbance as more current traverses through the access transistors into the cell. Thus, SRAM periphery wear-out worsens read margins, however, using statistical simulations these effects were found to be minor in comparison to the degradation effects of the internal core-cell transistors.

B. Write margin

Generally, the write-ability of SRAM core-cells improves with aging as NBTI effects reduce PMOS device strengths and thereby the contention between pull-up and access-transistor during the discharge of the storage node holding '1' [16]. In terms of write operations, SRAM periphery degradation mainly affects the write-driver [5] which drives one of the bit lines to zero before the word line is activated as well as the timing control circuitry. Similar to the improvements in sensing time margin, delays of the write driver transistors can be neglected in comparison to much longer delays along the timing paths of the SRAM control circuitry. This leads to longer word line windows giving the memory more time to discharge the storage node holding '1'. Thus, the wear-out of SRAM peripheral circuits does not degrade write margins but tends to slightly improve it in addition to the dominating aging effects of core-cell transistor strengths.

VI. CONCLUSION

This paper analyzed how aging of the SRAM periphery affects important memory performance metrics based on transistor aging models extracted from 28nm technology. It showed that SRAM peripheral logic wear-out is the main driver for access speed degradation while the impact of core-cells on that metric can be neglected. The worst-case speed loss occurs when memories experience high stress before being operated at lower VDD modes. It furthermore points out that wear-out of the SRAM control logic compensates for aging effects inside sense amplifiers and core-cells in terms of access time margin. Higher memory workloads even amplify this behavior. Moreover, SRAM peripheral circuit aging tends to slightly degrade SRAM read margin and further improve write margin. We also highlighted that the described effects are caused by the fundamental fact that signals which are triggered later within a memory operation cycle have longer timing paths and therefore experience stronger aging-induced delays. Thus, the results presented here can be generalized to all self-timed memory architectures.

REFERENCES