Lehrstuhl für Technische Elektronik
der Technischen Universität München

Parametric Reliability of 6T-SRAM Core Cell Arrays

Stefan Drapatz

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. Georg Sigl

Prüfer der Dissertation:

1. Univ.-Prof. Dr. rer. nat. Doris Schmitt-Landsiedel
2. apl. Prof. Dr.-Ing. habil. Walter Stechele

Die Dissertation wurde am 14.10.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 7.02.2012 angenommen.

Summary

The increasing integration density of microelectronic circuits in combination with non-constantly scaled supply voltages results in higher electric fields in MOS transistors. This is one central source of several aging mechanisms, some of them shifting the parameters of MOS transistors during lifetime. These parametric degradation effects can be separated in two groups called 'Bias Temperature Instability' (BTI) and 'Hot Carrier Injection' (HCI). This work focuses on the impact of these degradation mechansisms on 6-Transistor Static Random Access Memory (SRAM) arrays in 65 nm low power CMOS technology.

First, some basic information is provided about SRAM cell functionality, key performance metrics, reliability and the four parametric degradation mechanisms covered in this work. Then, the sensitivity of the SRAM core cell to each degradation mechanism is simulated. Together with the effective device degradation under normal SRAM operations in real life, this results in the information about the impact of each mechanism. BTI for pMOS transistors, called Negative BTI (NBTI), could be identified as the main problem in actual 65 nm low power technology with conventional SiO_2 gate dielectrics.

NBTI shows strong variation- and recovery-effects, which both are not fully understood, although this degradation mechanism has been known for approx. 30 years. This is why there are no sufficient simulation models so far, thus, measurements have to be performed to do the step from single cell simulation to SRAM array conclusions.
Consequently, a major focus of this work is to develop unconventional new measurement techniques. Contrary to state-of-the-art methods they are faster, do not need dedicated test chips which do not represent mass product design, do not need highly accurate V-I measurements and therefore can be used in-field in products with the only precondition of dual-V_{DD} power routing.
By using these new techniques, the impact of the worst degradation mechanism NBTI was examined directly on large-scale SRAM arrays. Especially the fast-recovering component of NBTI was directly measured on SRAM array stability for the first time. Thus, it could be shown which use-cases are critical to provide long lifetimes, which is the first step to fight the impact of parametric degradation mechanisms.
Finally, a comparison of known countermeasure techniques was performed in order to choose the most promising methods.

Contents

Part I

Introduction and Background

Chapter 1

Introduction

Static Random Access Memory (SRAM) nowadays is a dominant part of Systems-on-Chip (SoC). Up to about half of the die area and 2/3 of the transistor count of a modern microprocessor consists of SRAM cells. Fig. 1.1 shows the die photo of an Intel Penryn processor manufactured in 45 nm technology [www.intel.com]; the SRAM area can be identified on the left half of the die with its characteristic homogeneous layout style. 6 MB of SRAM Cache memory equals approx. 300 million transistors, which is 73% of the complete number of 410 million transistors.

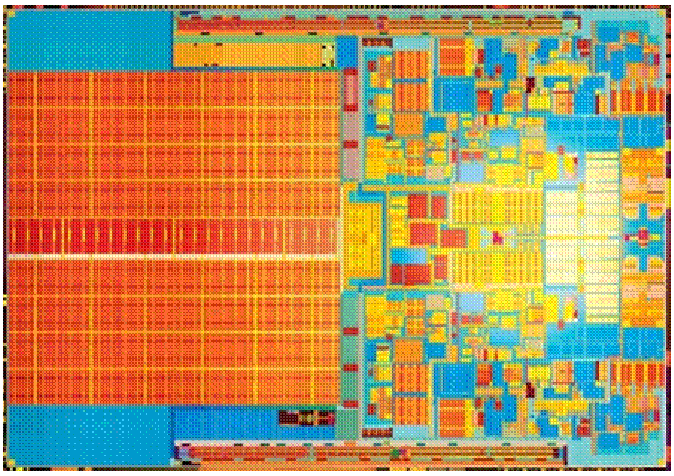
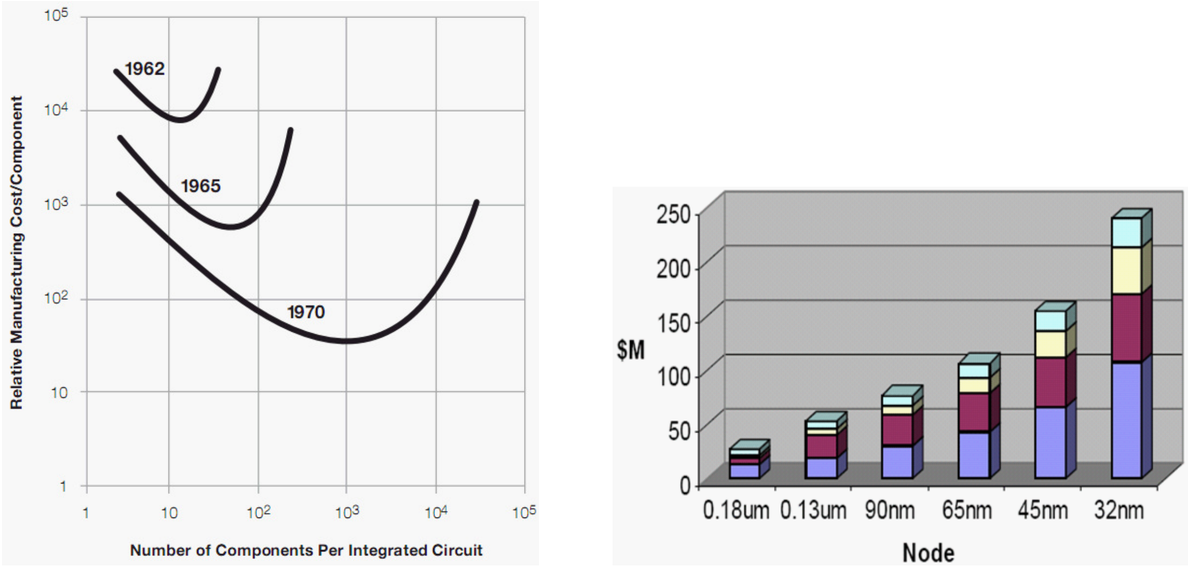


Fig. 1.1: Intel Penryn Processor: about half of the die area and 2/3 of transistor count consist of SRAM, identifiable on the left half of the die [www.intel.com].

Systems on Chip will require more and more memory in the future. 90% of the die area are projected to be memory in the next 10 years [1]. Die area directly translates to cost. To get maximum memory capacity on smallest possible area, the two obvious main approaches are: 1. minimize transistor sizes, 2. densify transistor packaging. This is why SRAM has the smallest transistors and the highest transistor density of the whole chip. This work focuses on the behavior of SRAM cells made of minimum size transistors with special tight design rules with respect to parametric reliability issues.

1.1 Technology Scaling: Benefits and Challenges

“The number of transistors incorporated in a chip will approximately double every 24 months“ (Gordon Moore, Co-founder of Fairchild and Intel) [www.intel.com].
This quote, better known as ”Moore’s Law“ from 1965, has just celebrated its 45th anniversary and is still valid. It is motivated from the falling cost per transistor on a chip when the integration density is ramped up (Fig. 1.2(a)). Increasing the number of transistors is feasible only by decreasing the size of each transistor. All 18 to 24 months, the length of one transistor is divided by $\sqrt{2}$ and therefore the transistor count is doubled. While in 1970, the minimum transistor length was about 10 µm, in 2011 product development has reached 32 nm or even less.



(a) Moore’s Law of higher integration over time is motivated from decreasing cost per transistor [2]

(b) Cost for each new technology generation is rising drastically in the last decade [I.B.S. Inc.]

Fig. 1.2: Moore’s Law and cost of new technology development

This ’law’ could only last that long because, contrary to power semiconductors, the optimum transistor in information technology is a small device. Smaller transistors not only need less area and create less cost, but they are faster and have less energy consumption. This is why scaling in the last 4 decades had almost nothing but advantages, which is an unlike behavior for all kinds of engineering. Only in the last 10 years, the disadvantages are increasing.
First of all, the cost of development for each new technology generation is increasing dramatically (Fig. 1.2(b)). It is getting more and more difficult to produce tiny structures in the deca-nanometer regime, e.g. for lithography to create 32 nm structures with the actual 193 nm wavelength. Immersion layers are state of the art before the next step to Extreme Ultra-Violet (EUV) with 13 nm wavelength can be done. Actually, this technology has too low throughput to replace the well-known state-of-the-art lithography. This is why conventional lithography in combination with highly regular layout pattern

is used.

Furthermore, the produced devices show more and more non-ideal behavior: short channel effects are increasing, so the small-signal output resistance r_{out} is reduced because of the slope in the output characteristics. Oxides are getting thinner to avoid decreasing transconductance g_m. But this increases subthreshold leakage, also because V_{DD} cannot be reduced as much as it should be in order to keep enough overdrive voltage ('non-constant voltage scaling'). Also gate leakage increases: the thickness of gate oxide is in the range of some atomic layers now, which causes direct tunneling, and static power dissipation is going up. Another topic is variability: the small number of dopants in the channel results in high V_{th} distribution.

The two worst effects on circuit perspective in the last 10 years were leakage and V_{th} variability, and many approaches have been done to fight against both effects [3]: 1. high-κ gate oxide materials and metal gate electrodes. 2. Bulk CMOS is replaced by Silicon-On-Insulator (SOI). 3. MuGFETs: multi-gate FETs, FINFETs.

But now, degradation effects are adding: With decreasing gate oxides, but almost constant V_{DD}, electric fields in the gate oxide increase. This is the motor for parametric degradation.

1.2 Scaling of SRAM: Motivation for this work

Volatile memory like SRAM has been one of the major driving forces for scaling in the last decades. This is because scaling has the greatest area effect on these high-density transistor structures, much more effect than for ordinary logic [4]. Fig. 1.3 shows the cell area shrink from 1995 till today including a view into the future until 2025 [1]. While in

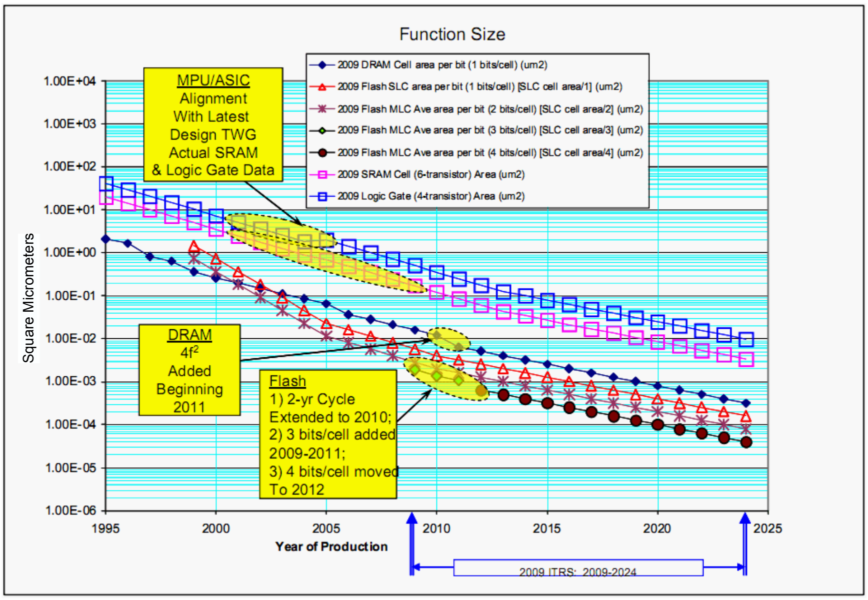


Fig. 1.3: ITRS roadmap for volatile memory incl. SRAM from 1995 until 2025 [1]

1995, a 6T-SRAM cell had an area of 20 µm², 2010 this got reduced to 0.15 µm², which

is about 1%. But how do the already discussed challenges of further shrinking translate to SRAM? Shrinking the node size from 250 nm down to 50 nm divides the cell stability by a factor of 4 [4]. This is the first reason why technology scaling challenges the SRAM cells. ITRS sees the difficult challenge in SRAM scaling in 'maintaining adequate noise margins and control key instabilities and soft-error rate' [1].
Summing up, leakage, variability and stronger electric fields are the three major challenges in the ongoing transistor scaling. Since SRAM must provide as much memory as possible on minimum space, SRAM suffers most from the drawbacks of scaling:

1. Up to hundreds of millions of transistors are a huge multiplier for single transistor leakage. This is why SRAM is one of the worst leakage current sources in a SoC.

2. Variability is increasing with decreasing transistor size, following Pelgroms law [5]. This is why variability on SRAM with its minimum-size transistors is much worse than for digital (and of course analog) transistors.

3. Degradation due to high electric fields can affect the SRAM behavior. Degradation and variability are always combined, i.e. degradation never appears without variability. This will be in the focus of this work.

Yield and Reliability are going to be more and more critical from technical and economical point of view, this is why degradation has to be investigated on SRAM cells.

1.3 Outline and Contributions of this work

Chapter 2 is about the functionality of 6-transistor SRAM cells as well as performance indicator metrics. Several metrics for the same performance exist, the advantages and disadvantages are discussed. **Chapter 3** discusses the basics of yield, quality and reliability. Four actually known parametrical degradation mechanisms on transistor level are introduced. In **Chapter 4** the impact of the four parametric degradation mechanisms on the SRAM cell is simulated. It is examined in which modes the SRAM circuit fulfills which degradation conditions and how it reacts to degradations. Together, this can state how strong each degradation takes effect on the circuit. Additionally, a combination of degradation mechanisms is considered. **Chapter 5** performs the step from the single SRAM cell to the SRAM array. It describes the newly developed method for fast analyzing stability of SRAM arrays. The impact of NBTI can be measured with this technique. **Chapter 6** examines the recovering NBTI component and its measurement on SRAM arrays. This has never been done on SRAM cells before, because all existing measurement approaches were too slow. But the new technique developed within this work is able to measure this component. **Chapter 7** is dealing with countermeasures to aged SRAM core cells. After only the impact of degradation on the memory cell was discussed so far, the focus now is on countermeasures. What can be done to achieve memory cells that are working correctly many years after production and usage? **Chapter 8** is the conclusion followed by an outlook.

Chapter 2

SRAM Fundamentals

In computer memory hierarchy, the fast and small-capacity memory types are on top, while the slow and huge-capacity memories are at the bottom (Fig. 2.1). While disc drives store Terabytes of data and have access time in the 10 ms region, main memory of DRAM type typically stores some Gigabytes but has a factor of 10^5-10^6 faster access time of approx. $10 - 100$ ns. Most of these faster techniques are based on charging or discharging of capacitors, which takes some time for transportation of charge. Often they represent dynamic memories, which must be refreshed in fractions of a second to enable long storage time [6]. To further improve access time by a factor of 10 or more in order to get to the top of the memory hierarchy pyramid, the principle of positive feedback is used. No charge must be stored, positive feedback is a technique that brings a circuit to its extreme values and therefore realizes bistable systems. In case of memory those are the two binary states '1' and '0'; systems using this technique are called 'Flip-Flops'. SRAM, latches and registers are based on that principle: their killer feature is having extremely fast read and write access. Latches, which are level-sensitive and typically used to build sequential logic circuits, are often based on cross-coupled NAND or NOR gates, compare Fig. 2.2(a). Its advantage is the simple usage and asynchronous data interface, it can

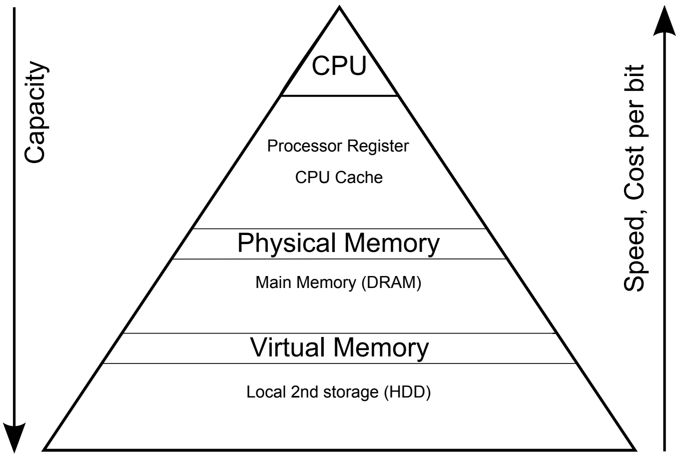
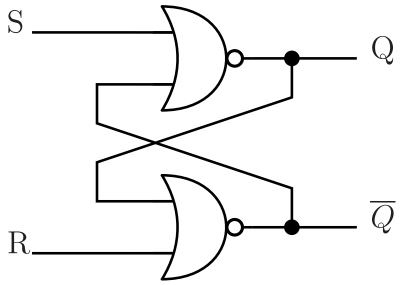


Fig. 2.1: Memory hierarchy of a Personal Computer (PC)

be easily read and written with 3 signals R, S and Q, compare Fig. 2.2(b). Q always keeps the stored information, and setting R or S to 1, while keeping the other signal at 0 resets or sets the latch. To build edge-triggered registers, the level-sensitive latch must be transformed to a synchronous circuit, which is adding some more transistors. So the disadvantage is the big area consumption: typically 10 to 30 transistors are required to store only one binary digit (bit).



(a) SR FlipFlop is a NOR-based latch with the typical cross-coupling for positive feedback.

S	R	Action
0	0	No change
0	1	Q=1
1	0	Q=0
1	1	forbidden

(b) Table for reading and writing the latch with 3 signals.

Fig. 2.2: Simplest latch: asynchronous SR Flip-Flop which can be used to build sequential logic circuits [6].

SRAM on the other hand needs a complex periphery to read or write a distinct cell in a huge array of core cells, compare Fig. 2.3 [7]. Reading and writing are complex procedures, which will be described in section 2.1. Due to the periphery overhead, SRAM cells do not make sense as single latch cells. So one SRAM cell never comes alone, the typical SRAM array size is at least some thousand to some millions of cells, which makes it a kBit or MBit array. Therefore SRAM is also a great test vehicle for variability examinations.

So the key performance of SRAM compared to all other memory types is speed: SRAM has about 1 ns read and less than 1 ns write access time. It is typically used inside a microcontroller for cache memory, which is divided in several, normally up to 3, speed or hierarchy levels. Level 1 cache is clocked with CPU frequency, which is some GHz. Therefore, a memory type with less than 1 ns access time is needed. This level nowadays normally has a size of 4 to 64 kB.

Level 2 is much bigger, about 64 kB to 12 MB. Sometimes it is not on the CPU itself, and it is always clocked slower, e.g. with some hundreds of megahertz.

For the advantage of high speed, one SRAM cell needs about 10 to 15 times more area than a DRAM cell [1], which directly translates to cost. One SRAM cell in 65 nm is about 0.5 -0.7 µm^2, the core cells examined in this work have a size between 0.6 -0.7 µm^2. To keep this area as small as possible, they are built with especially tight design rules. This is possible because of their regular layout. They allow to place more minimum size transistors than for conventional logic, so SRAMs do have special status in semiconductor manufacturing.

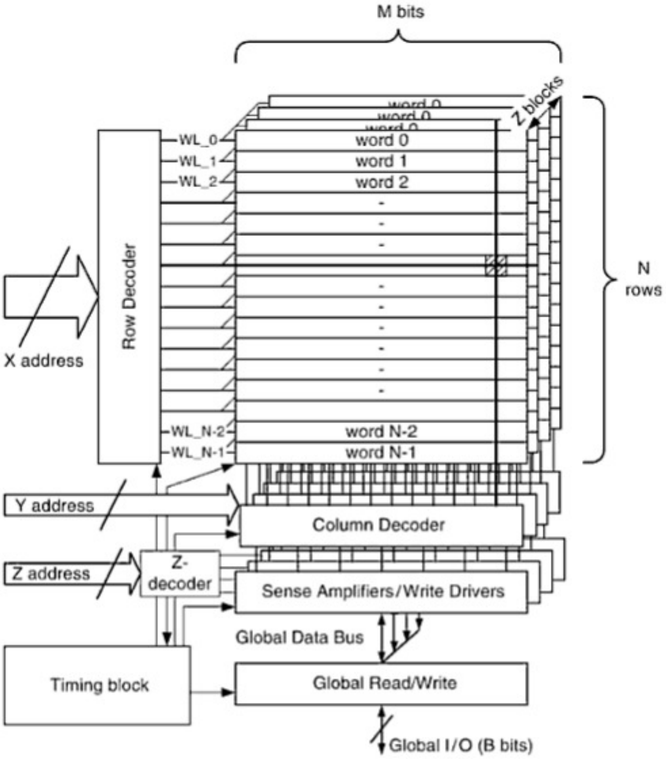


Fig. 2.3: SRAM block diagram showing the core cell array and the periphery containing row/column decoder and sense amplifier taken from [4]

2.1 Functionality of the 6T-SRAM Core Cell

Although many SRAM core cell designs exist with various number of transistors [8], the standard choice is the cell with 6 transistors, called the 6T cell. It is often used because at least until 65 nm technology, this type of cell provides the optimum trade-off between stability and area consumption [9]. The core cell consists of 2 cross-coupled inverters implementing the positive feedback and therefore the two memory nodes which keep the information (original S and inverted $\overline{S}$). 2 additional access- (or pass gate-) transistors are required to read and write the cell in a differential way, resulting in a more robust circuit. Fig. 2.4 shows the schematics of the 6T-SRAM core cell. The transistors in this work are named pullup (PL1, PL2), pulldown (PD1, PD2) and pass-gate or access (PG1, PG2). The pullups are p-type, while the pulldowns and pass-gates are n-type.
The 3 standard procedures hold, read and write work as follows [10].

- 'Hold': the access transistors are disabled (WL=0), the information is stored on the feedback-coupled inverter-pair.

- 'Read': both bitlines BL and $\overline{BL}$ (or BLB) are precharged to V_{DD}, then the access transistors are enabled (WL=1). The '0' memory node provides a conducting pull-down to ground and discharges the bitline via the opened access transistor on this side. A sense amplifier detects the sloping voltage on one of both bitlines and concludes this side to be the '0' memory node. The sense-amplifier serves to speed-up

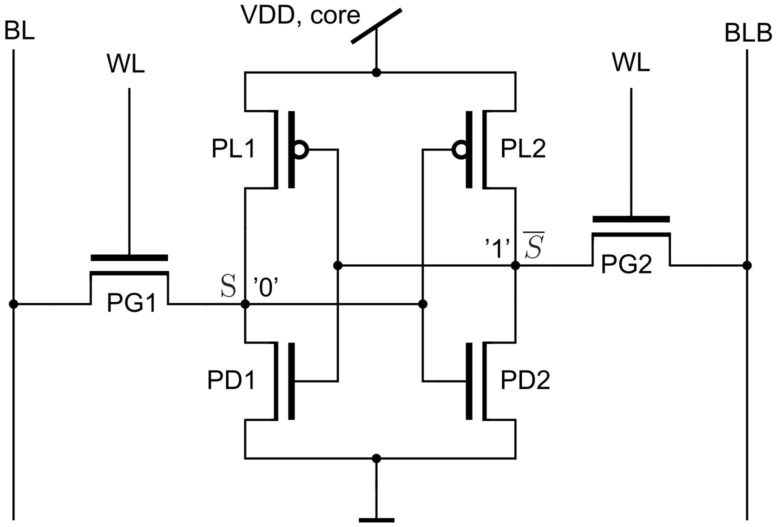


Fig. 2.4: The 6T-SRAM cell consists of 2 feedback-coupled inverters that only allow the 2 stable states '1' and '0' on the memory nodes S and $\overline{S}$ plus 2 access transistors. This SRAM circuit is in memory state S='0'. The inverted information is kept on memory node $\overline{S}$.

the read process, because the bitline then does not need to be discharged completely down to 0 V.

- 'Write': starting from 'Read' case (BL=BLB=1, WL=1), the bitline on the desired '0' memory node side is tied to ground, while the other bitline is kept at V_{DD}. If the cell is not in this state already, the voltage on the desired '0' node will drop below the switching level of the opposite inverter and flip the cell.

2.2 SRAM Performances and Figures of Merit

SRAM is a volatile memory and its task from the user's point of view is simple: as long as the cell is connected to supply voltage, it must keep data (hold) and enable to read and write data. Ideally, this must be done very fast, on minimum die area, with almost no leakage and great yield.

Each core cell has different qualities and strengths, depending on its design. All these different qualities must be measurable to be able to compare various core cells. While some qualities are unambiguos (e.g. area in µm^2), some others need Figures of Merits (FoMs) when they are not directly measurable (e.g. cell stability). Table 2.1 provides the list of all core cell performances and Figures of Merit if necessary.

It is important to note that, as in probably every technical system, not all performances can be improved together; some of them are oppositional requests. The most important fact for SRAM is that reading and writing are conflicting challenges. Generally, SRAM stability is limited by the switching levels of the 2 inverters. When the '0' memory node voltage surmounts the switching level of the '1' side-inverter, the cell flips. This must be avoided in read case, but must be achieved in write case. This means that a cell is either

Section	Performance	Meaning	Figures of Merit
2.2.1	Read stability	How easy cell flips during read access	SNM(read) [V] NCurve [V,A] RM [V]
2.2.2	Hold stability	How easy cell flips during hold	SNM(hold) [V] NCurve [V,A] $V_{min,ret}$ [V]
2.2.3	Write-ability	How easy cell flips during write access	Write Level [V] NCurve [V,A]
2.2.4	Speed	How fast the cell can be read	I(read) [A]
2.2.5	V_{min}	Lowest V_{DD} to provide full functionality	n/a [V]
2.2.6	Area	Size of one cell on the die	n/a [µm^2]
2.2.7	Leakage	Static current during hold state	n/a [A]
2.2.8	Yield	Fraction of functional cells	n/a [1]

Table 2.1: 8 performance parameters of an SRAM cell including 4 Figures of Merit

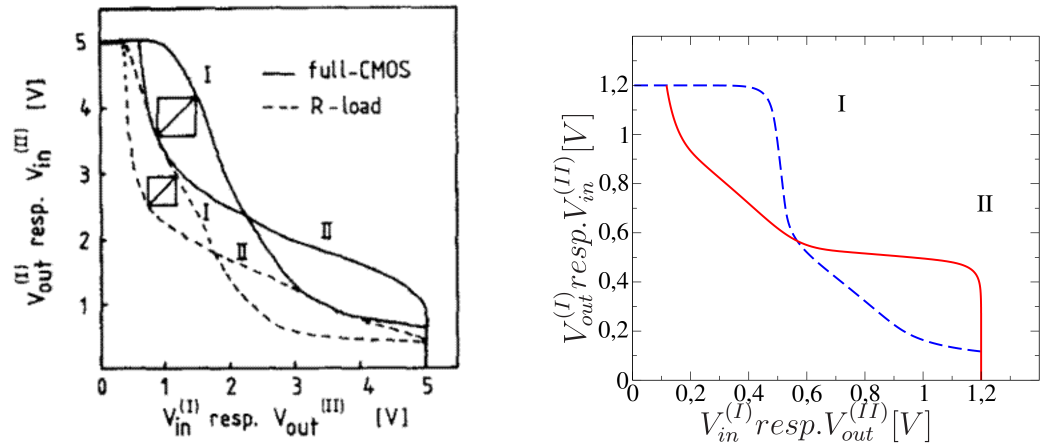
very stable and has good reading quality, but then does not flip easily during write access and therefore has poor writing quality. A cell must always be a trade-off between these 2 qualities. If this trade-off is met best, the cell is called 'centered'.
The next sections are about the 8 performance metrics of an SRAM cell.

2.2.1 Read Stability

These metrics express how easily a cell flips under read conditions, which is sometimes also called 'Access Disturb Margin' (ADM).

Static Noise Margin for read case: SNM(read)

Static Noise Margin (SNM) is the most prominent stability metrics and was introduced more than 40 years ago [11]. Seevinck et al. then wrote the maybe most cited paper in SRAM design, some analytical work and a simulation method about the stability of SRAM cells [12]. Nowadays, this FoM to measure read stability is universally accepted and automated using DC circuit simulators. With this metric, the cell is actually not being flipped, but it is estimated how much voltage room for static noise, i.e. DC voltage, is left until the cell will flip. Therefore, both inverters are first scanned independently to get the transfer curves, also known as voltage transfer characteristics VTC (compare Appendix A). Then, curves are plotted into one diagram, while one transfer curve is mirrored. This results in 2 overlaying curves, better known as 'butterfly curve', compare Fig. 2.5(a) from the original publication and Fig. 2.5(b) from simulation. The eye opening of the biggest inbuilt square is a measure for how much static noise is needed to reach the trigger level of the opposite inverter under read conditions. For this DC value, the curve still represents a 'butterfly', i.e. 2 stable regions and 1 metastable point in the middle. The higher the SNM value, the more stable the cell.



(a) SNM original plot from Seevinck

(b) Simulated butterfly curve for read case

Fig. 2.5: Butterfly curves for read case: original plot from [12] and simulated in 65 nm technology

SNM is perfectly suited for simulation. This is done with a trick: a 45 degrees rotated coordinate system is introduced to optimize calculating the inbuilt square [12]. Simulations in a 65 nm technology showed the results in the middle column of Table 2.2. It can be seen that elevated temperature decreases cell stability. This is due to the decrease in threshold voltages.

Temperature	SNM(read)	SNM(hold)
-40 °C	254 mV	487 mV
25 °C	243 mV	468 mV
125 °C	220 mV	436 mV

Table 2.2: Simulated SNM values for read and hold case in a 65 nm technology. Stability is reduced by 10% in read case and by 7% in hold case when the temperature is increased from 25 °C to 125 °C. Furthermore, read case is generally much more critical than hold case at nominal supply voltage.

Unfortunately, SNM is not very well suited for measurement, as access to the memory nodes is needed, which is not provided by conventional SRAM products. Dedicated test structures ('fly cells') are needed. Furthermore, calculating the eye opening is much too complex for extremely fast inline-testing.

Read N-Curve

This FoM came up recently [13]; its general idea is to flip the cell with an external force, which can be measured. This external force is a voltage source which is connected to the '0' memory node and then ramped up to V_{DD}, compare Appendix A. Not only the voltage (and therefore the voltage transfer characteristic) of this setup is monitored, but also the current that the sweeping source provides. This results in the so-called N-curve shown in the lower plot of Fig. 2.6, since it has the shape of an 'N'. Both SNM(read)

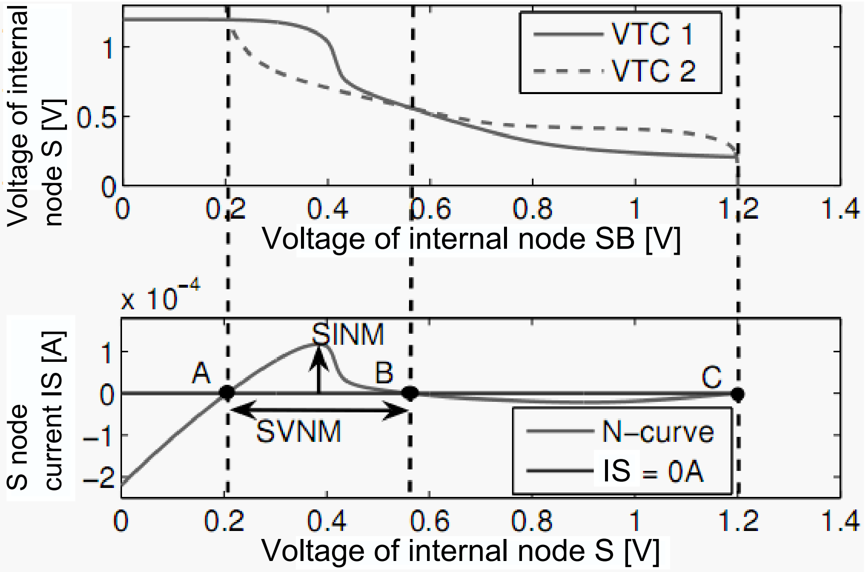


Fig. 2.6: Comparison of N-Curve (lower plot) with SNM (upper plot), taken and adapted from [3]. The voltage information in both stability analysis methods is almost the same. N-Curve provides some additional information with the measured current.

and Read N-Curve show correlated results, at least for small currents [14], depicted in Fig. 2.6: the 3 voltages where the current equals 0 are the 3 crossing points in SNM analysis: two stable and one meta-stable point. N-Curve provides additional information with the measured current [3] [15].

In contrast to SNM, the big advantage of Read N-Curve [13] is the ability to measure this metric very fast with automatic inline testers. But unfortunately, this method again requires access to the memory nodes and is therefore not capable for product measurement, compare Static Noise Margin.

Read Margin RM

For Read Margin, only the core voltage is reduced to zero, while the periphery is left on nominal V_{DD}, compare Appendix B. Lowering the core voltage decreases cell stability, and at some voltage, the bitline current drops, compare Fig. 2.7. The difference between the nominal core voltage and the reduced voltage where the bitline current drops is the Read Margin. Static Noise Margin and Read Margin are well correlated [17].

It is important to note that the bitline current only drops if the cell is not kept on its preferred side. Fig. 2.8 provides the two cases of a cell on its non-preferred side (Fig. 2.8(a): bitline current drops) and on its preferred side (Fig. 2.8(b): bitline current does not drop). An ideal SRAM cell (like in simulation) does not have a preferred side, but the unavoidable manufacturing variations in real life always result in a preferred side. Therefore, while SNM can be simply calculated with a DC simulation, RM is only accessible via Monte Carlo simulations. Only when the variability is known for the simulated technology, then the RM stability can be MC simulated.

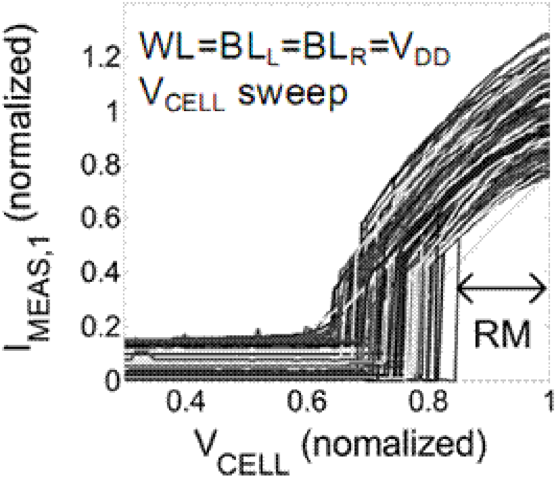
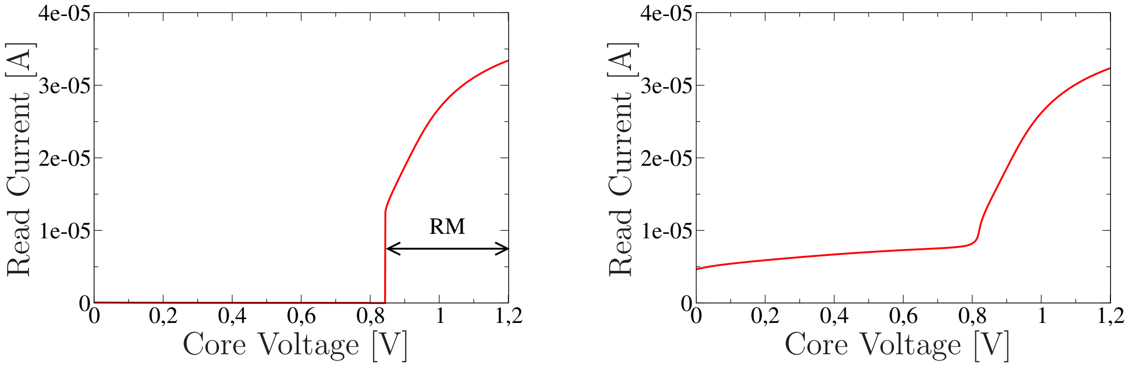


Fig. 2.7: Read Margin: the difference between nominal supply voltage and the reduced supply voltage where the bitline current drops. Original plot taken from [16]



(a) For about 50% of the cells, at some $V_{DD,core}$ the read current on the '0' memory side drops to zero. The difference between V_{DD} and $V_{DD,core}$ is the Read Margin.

(b) For the other 50% of the cells, the current drop does not happen. The cell is alrady in its preferred state and therefore the current never reaches zero.

Fig. 2.8: Read Margin criterion only works for about 50% of all cases. Two Monte Carlo simulation plots with completely different result.

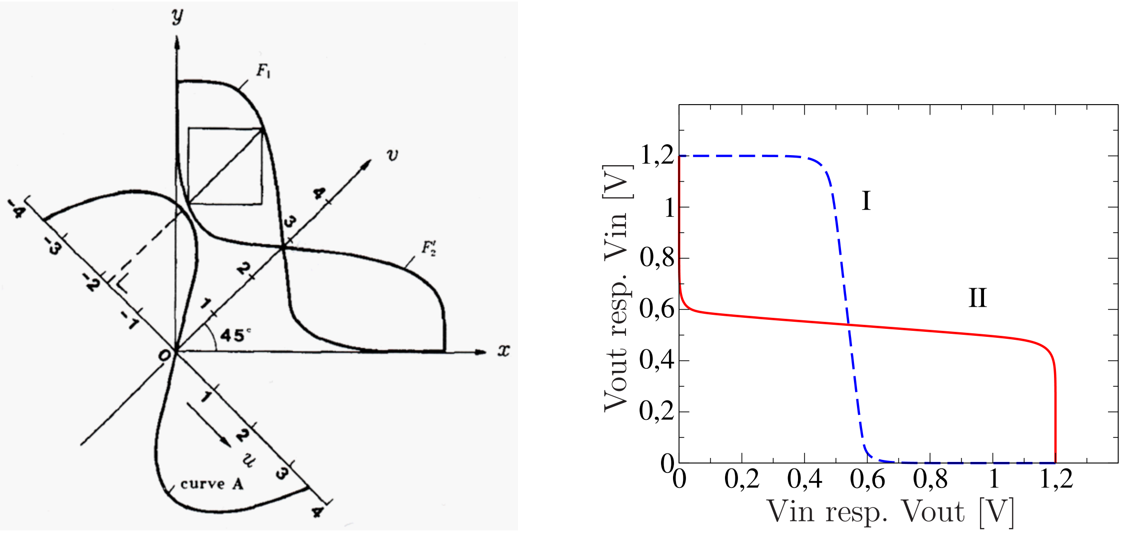
In contrast to SNM and N-Curve, RM is the only stability-FoM that can be measured with product-like core cells, as no access to the memory nodes is required. Unfortunately, V-I measurements are relatively slow and therefore, this technique is not well suitable for array characterization so far. This is the reason why RM is the better measurement metric.

2.2.2 Hold Stability = Data Retention Stability

These metrics express how easily a cell flips under hold conditions. In principle, the read stability FoMs could be used as well for hold stability characterization, with the difference of a disabled WL. But in practice, this is only valid for SNM and N-Curve: disabling the WL creates SNM(hold) and Hold N-Curve. Read Margin current measurement does not work for hold case, since the measured read current, which is analyzed for the current drop, does not exist in hold case.

Static Noise Margin for hold case: SNM(hold)

This is measured like SNM(read) (compare Appendix A), but the access transistor is disabled while the inverter characteristics are scanned. Since the access transistor is not conducting, the transfer characteristic has the typical inverter-shape, compare Figs. 2.9. Some simulated values for 65 nm technology are listed in the right column of Table 2.2 on page 12. The eye opening is much bigger than in read case, which means that in



(a) Original plot from Seevinck [12] including the 45 degrees rotated coordinate system

(b) Simulated butterfly curve for hold case. The inbuilt square is much bigger than for read case, because the stability-decreasing access transistor is switched off.

Fig. 2.9: Butterfly curves for hold case: original plot from [12] including the rotated coordinate system for DC circuit simulation and simulated in 65 nm technology

comparison to SNM(read), for nominal V_{DD} read case is much more critical than hold case.
To save leakage power, SRAM arrays are often operated with reduced supply voltage, compare sections 2.2.5 and 2.2.7. This is why hold stability must be considered as well. In systems with permanent nominal supply voltage, hold case is no problem.

Hold N-Curve

The same setup like for Read N-Curve can be used for Hold N-Curve (compare Appendix A), but with disabled wordline [17]. Again, the 3 voltages where the current equals 0 are the 3 crossing points in SNM analysis: two stable and one meta-stable point, compare section 2.2.1.

Minimum Retention Voltage $V_{min,ret}$

Another FoM to characterize hold stability is to lower V_{DD} when WL=0, i.e. during hold condition. The minimum voltage at which the cell still does not lose its memory state is the minimum retention voltage $V_{min,ret}$. Reading is not allowed with this drastically reduced voltage. Since the hold case is very stable compared to read case, the retention voltage can be reduced drastically until the cell loses its state.

2.2.3 Write-ability

Write-ability describes how easily the cell can be flipped under write conditions and is therefore the opposite to stability. This is the reason why it is not possible to optimize both qualities together.

Write Level or Write-Trip-Point

The commonly used FoM for writeability is the Write Level, also known as Write-Trip-Point. First, read conditions are applied to a cell (WL=BL=BLB=1), then the BL voltage is lowered on the '1' memory side, compare Appendix C. When the BL voltage is low enough so that the cell flips, then this is called the Write Level. The lower this BL voltage must be to flip the cell, the harder the cell is flippable. So the higher the Write Level value, the better the cell is writeable. Table 2.3 shows some simulated Write Level values. Increasing temperature increases writeability, again of course opposite to stability.

Write N-Curve

Similar to Read N-Curve, an external voltage source which is connected to the '0'-memory node is ramped up to V_{DD}, compare Appendix D. This time the cell is under write conditions, i.e. the '1'-bitline is high, the '0'-bitline is low. The current flowing into the memory node is a measure for the writeability of the cell [13]. The resulting curve is similar to an

Temperature	Write Level
-40 °C	376 mV
25 °C	406 mV
125 °C	440 mV

Table 2.3: Simulated WL values in a 65 nm technology. Write-ability is increased by 8% when the temperature is increased from 25 °C to 125 °C.

'N', compare Fig. 2.10. The so-called critical writeability current I_{crit} can be read from the graph, it is defined as the current valley in the right half of the plot. If this critical current is a positive value for all applied voltages (like in Fig. 2.10), then the write process was successful. A more positive current value represents a better writeable cell, a critical current <0 represents a write failure.

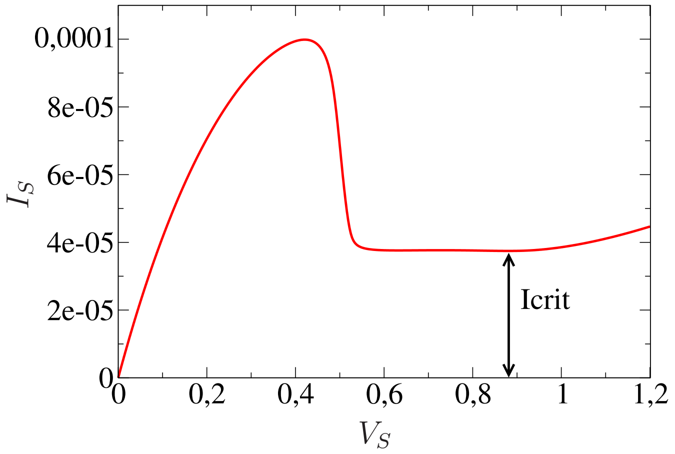


Fig. 2.10: Write N-Curve: the current valley close to the right end is the writeability current. Bigger values represent better writeable cells.

It is also possible to use the read stability N-Curve plot to analyze the circuit for writeability [18]. Then, the read N-Curve in Fig. 2.6 on page 13 is read from right to left because this represents lowering of the memory node voltage. The Write Trip Voltage in this approach is the voltage difference between the last 2 zero crossings, namely the points B and C.

2.2.4 Speed / Performance

One of the most important SRAM parameters is the speed of a cell, i.e. how fast the cell can be read. This parameter is sometimes called performance. A Figure of Merit is introduced that does not directly measure speed or a time, but the read current. This read current is discharging the parasitic capacitances of the long bitlines. It is called I_{read} and determines the time necessary to discharge the bitlines to a value where the sense amplifiers detect the zero node. The bigger this current, the faster the BLs can be discharged, and the faster the sense amplifier is able to detect the cell state. Since speed is the most important parameter and the reason to use SRAM instead of another volatile memory type, it is the main performance of the cell.

2.2.5 Minimum supply voltage V_{min}

In times of high leakage currents, lowering V_{DD} is the simplest way to save energy in low-power systems, which of course costs performance. But lowering the supply voltage is not only possible during retention mode, but also during normal operation of SRAM, i.e. including read- and write-accesses.

V_{min} is a very prominent parameter for SRAM circuits. It defines the minimum supply voltage that provides full functionality, namely hold, read and write. It is important to note that not only the core voltage is lowered, but also the complete periphery voltage, i.e. BL, BLB and WL voltages. The difference to $V_{min,ret}$ (which is a parameter of hold stability, compare section 2.2.2) is that also read and write must still work properly.

It is remarkable how deep V_{DD} can be lowered until the first cells start to flip during read condition. Fig. 2.11 shows the measured number of flipped cells for Read Margin condition (right), hold condition (WL=0, left) and V_{min} (middle). This plot does not include the writing procedure, but it shows that the supply voltage can be lowered from 1.2V to approx. 0.6V before the first cells start to flip. When only the core voltage is lowered (Read Margin case), cells start to flip at much higher voltages, because the access transistors are connected to nominal V_{DD}.

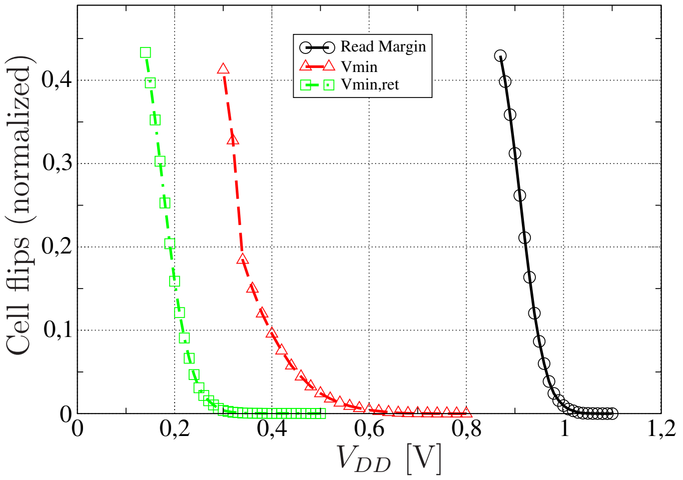


Fig. 2.11: Measured flips due to lowering core voltage in read state (RM criterium), supply voltage (V_{min}), and core voltage in hold state ($V_{min,ret}$)

V_{min} cannot be used as a metric for stability, but it is of big importance in mass production. Contrary to many other FoMs, this metric can be determined very easily without special test structures. Hold, read and write procedures are repeated with continuously lowered supply voltage until one procedure fails. This is the V_{min} value. While SNM and

RM represent the stability of the core cell alone, V_{min} represents the minimum voltage for which the complete system incl. periphery still works. So SNM and RM show Gaussian Distribution, while V_{min} does not: it includes failure of the periphery.

2.2.6 Area

No FoM is necessary to determine the area of the core cell. For a 65 nm low power CMOS technology, this is a value between 0.6 and 0.7 µm^2. It is very important to keep this area as small as possible, since millions of core cells represent one memory array and a big fraction of area on the die.

2.2.7 Leakage

With shorter channels and thinner gate oxides, channel- and gate-leakage are increasing. Leakage in one transistor is not a big issue, since it normally is in the pA region. But in SRAM arrays with millions of devices, the currents are adding up to a big fraction of complete SoC's power consumption. Measurements have shown that the 1 MBit array in 65 nm low power CMOS technoloy has approx. 25 µA of leakage current at room temperature and nominal V_{DD}. Of course this is highly temperature-dependent, so operation at 125 °C will multiply this leakage current. Since SRAM is volatile memory, it cannot be switched off, otherwise the data is lost. This is why often the supply voltage is reduced to minimize leakage, compare section 2.2.5. This can be done on two levels. First, only reducing the supply voltage about some 10 mV to ensure full functionality of hold, read and write. This is the V_{min} value. Second, drastically lower the supply voltage to about half of the nominal supply voltage to ensure only retention of data. This is also called retention mode and can only be done with disabled WL. The corresponding voltage is called minimum retention voltage $V_{min,ret}$, compare section 2.2.2.

2.2.8 Yield

Yield is the fraction of working cells of all produced cells. The goal of every production is to reach a yield of close to 1. A cell will fail either due to read or due to write problems. If it is not able to hold data, then it is totally misdesigned. The next chapter 3 will discuss the topic of yield more in detail.

2.3 Summary

SRAM cells must be able to hold, read and write data. Ideally very fast, on minimum area, with low leakage and high yield. The goal is to design a cell that provides the best trade-off between all these qualities; it is then called 'centered'. Since it is not possible to optimize all these qualities at the same time and every core cell design has its characteristic strengths and weaknesses, performance parameters are introduced to quantify every single quality of the cell. The main FoMs are for read stability, hold stability, write-ability and speed. FoMs do not provide an absolute value of each quality, but a scalar quantity, so different cells can be compared between each other. This gives the result that read stability is much more critical than hold stability during nominal V_{DD} usage. Furthermore, stability decreases with increasing temperature.

Comparing the 3 stability metrics SNM, N-Curve and RM, it must be stated that SNM and N-Curve are not very well suited for measurement purpose. They need dedicated test structures with access to the memory nodes ('fly cells') which is not feasible on product chips. Read Margin can be determined with nominal core cell design, but needs slow V-I measurements.

Chapter 3

Degradation and Reliability

Moore's Law results in an exponentially growing number of devices per chip. This directly leads to much more complex systems, and reliability could be expected to sink drastically. Interestingly, this is not the case: over the last years, reliability was even increasing in spite of growing product complexity and application stresses [19]. This was only possible because of distinct reliability engineering, which is examining the reasons for failure and analyzing the systems with statistical means. The basics to this approach, called the physics-of-failure-concept, as well as the degradation mechanisms will be introduced in this chapter.

3.1 Quality, Yield, Variations and Redundancy

'Quality' is the degree to which products or services satisfy or even exceed the requirements and expectations [20]. Simply spoken, this means the product is 'fit for use'.
In terms of SRAM, this means that the core cell is able to hold, read and write data.

'Yield' is the fraction of high-quality (i.e. fit-for-use) chips directly after production.

$$Yield = \frac{\text{number of high-quality SRAM cells}}{\text{all SRAM cells}} \tag{3.1.1}$$

The two main reasons for reduced yield are local defects (randomly distributed in wafer- and assembly lots) and variations of physical and electrical parameters. The 'random defects' are caused by e.g. irregularities in material structure, effects of introduced particles, damages due to handling of wafers and assembly, etc. Their effect on yield and reliability depends on their size in relation to the dimensions of the affected functional structure. They are addressed by modern automated production techniques in high-class clean rooms. The 'parameter variations', however, are of statistical nature and represent differences between ideally identical physical elements of today's deca-nanometer transistors. The reasons for this are very broad and refer to variations of e.g. doping fluctuations, dimensions due to deposition, lithography and etching processes and related functional electrical parameters (e.g. V_{th}, g_m, ...).

The variations are divided in global and local variations. While global variations, i.e. differences between two dies, can be kept in narrow margins, local variations increase with decreasing feature size. They obey Pelgrom's law [5], which says that the standard deviation of the threshold voltage is increasing with decreasing area of the device.

$$(\sigma = \frac{A}{\sqrt{WL}}). \tag{3.1.2}$$

So if the transistor area is divided by 2 (which is the case from one technology step to the next), the standard deviation will increase by a factor of $\sqrt{2}$. This is why variations, together with leakage, are the main challenge in semiconductor manufacturing today.

Generally, microelectronic circuits must be designed in a way that they have high yield after production. This is done by e.g. simulations in all corners (different combinations of PVT: process, voltage, temperature) and design techniques that make the circuit robust against variations.

Although this is also done for memory circuits, it still cannot be guaranteed that all cells in a Mega-Bit array are working correctly after production. This is mainly due to variations on the tiny devices and the huge device count, which results in a non-vanishing probability of single cells with up to 6σ variability. If 1σ of V_{th} variation is assumed to be around 40 mV in the 65 nm technology, this means V_{th} variations of up to 240 mV. This is approx. half of the absolute V_{th} value!

Consequently, redundancy is commonly used to fight SRAM yield problems. After a burn-in step to eliminate the weak cells, a functionality check of every single cell is performed, which can be done in a reasonable time frame. Every cell (or every wordline) that does not work is then re-routed with fuses in the periphery to a spare, additional cell in the array. Then, if not too many cells were failing and enough spare cells were available, the complete memory array is working properly. But this concept only works directly after production, the fusing step cannot be done in-field.

But so far, yield was only considered directly after production. This will now be extended to lifetime and is then called 'reliability'.

3.2 Reliability Basics

Reliability = quality during use = stability of its properties and characteristics at time of delivery during subsequent use under the intended conditions during the planned period of use = yield over lifetime [20].

This means that a completely working circuit directly after production must keep its properties over its lifetime so that after a couple of years, this circuit still is of high quality. Unfortunately, there are degrading mechanisms based on the usage (electric fields, temperature, flowing currents, ...). Some of them shift the parameters of the devices over time, these are called parametrical degradation mechanisms and are subject of this work.

Parametrical degradation mechanisms are adding to variations over lifetime [21].

$$\text{Variations during lifetime} = \text{variations after production} + \text{degradation during lifetime} \tag{3.2.3}$$

So it makes no sense to consider degradation effects isolated, they always appear together and on top of the unavoidable variations. This approach will be taken into account for SRAM circuits in this work.

The fundamental reliability aspects are now discussed with the help of an example of working semiconductor devices over time, taken from [22]. Table 3.1 shows the number of fails counted after each time segment. The cumulative number of fails is the complete number of non-working chips over time. Normalized to the sample size of 5000, it is called the probability of failure F(t), better known as the cumulative density function (CDF) of failure. The derivative of the CDF is the failure probability density, which is the number of fails per time segment normalized to the complete sample size. It is better known as the probability density function (PDF). The failure rate $\lambda(t)$ (or hazard rate h(t)) is the number of fails per time segment normalized to the number of working samples. Since the number of fails is low in this example, both characteristics coincide, compare Fig. 3.1. The shape of the failure rate curve $\lambda(t)$ is like the cross-section of a bathtub, it is therefore called the bathtub curve [23]. The characteristic shape of this most famous reliability

Time [years]	0.1	0.2	0.3	0.4	0.5	1	2	3	4	5	6	7	8	9	10
Fails count	15	1	0	0	0	1	2	2	2	2	2	2	5	15	25
$\sum$ fails	15	16	16	16	16	17	19	21	23	25	27	29	34	49	76

Table 3.1: Number of failures of a device over time with a sample size of 5000

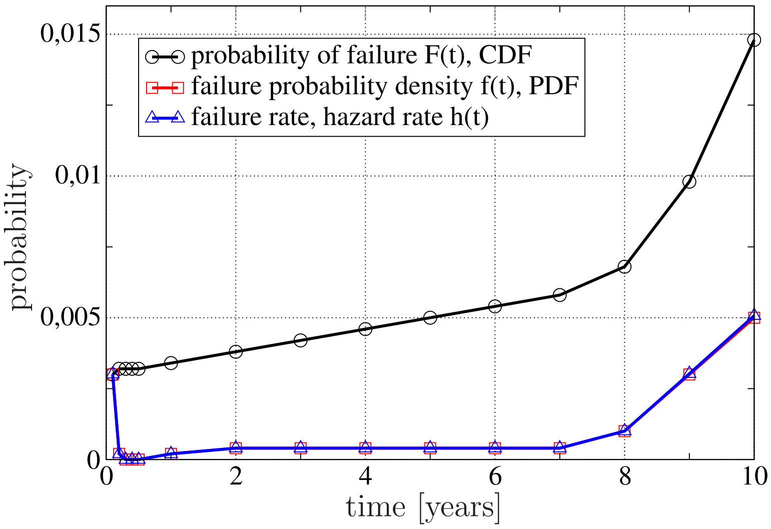


Fig. 3.1: Cumulative density function (CDF), Probability density function (PDF) and bathtub-like failure rate h(t). PDF and h(t) coincide in this example.

curve can be divided in 3 parts and results from the superposition of 3 fundamental failure rate curves, compare Fig. 3.2.

1. Infant mortality region with early failure rate
2. Useful operating life region with randomly distributed failure events vs. time, therefore almost constant failure rate
3. System wear-out region with wear-out failure rate (due to failure modes covered in this work)

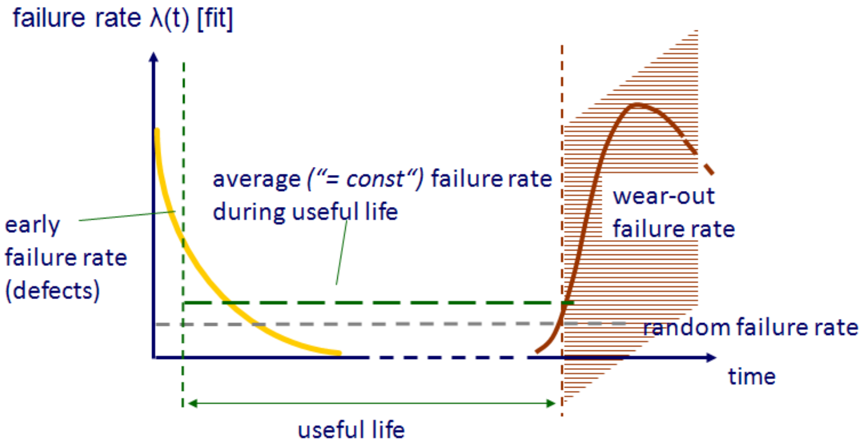


Fig. 3.2: The characteristic shape of the bathtub curve results from the superposition of 3 failure rates: early-, constant- and wear-out-failure rate. This work is about the wear-out failure rate. [20]

The early failures are because of products with production weaknesses, which still pass the functionality check but have some built-in defects. These normally do not live long, they fail during the first usage. In semiconductor industry, this is fighted by special screening or 'burn-in': usage with raised temperature and voltage stresses the device so much that most of the infant mortality candidates fail during production test in the company and not at the customer.

The random failure rate phase is the phase of useful life. Failures in this period are not built in the product, but caused externally by the conditions of use, like overvoltage, radiation etc. The dominant value in this region is the mean time between failure MTBF. It is not predictable for a single device, it can only be determined for a big amount of devices via statistics. But with a huge number of devices, these statistics work very precisely.

The random failure rate phase can be compared with radioactive decay: nobody knows about one specific atom when it will decay, but with the huge number of atoms in materials, the radioactive half-life can be determined very precisely. This half-life is an equivalent information to MTBF, but half-life is not a used term in semiconductor industry.

The wear-out region is the end-of-life region. Wear-out happens because of degrading mechanisms. These mechanisms do not start after some years, they start directly after production when the device is used. But after useful life, the degradation becomes so bad

that the quality of the device is no longer given, it is not fit-for-use anymore. The question that is most important: when does the wear-out failure rate reach a certain limit, which determines end of useful life?
This thesis is about the parametrical wear-out mechanisms of 6T SRAM cells. What different mechanisms exist, which of them have bad impact on the performance or quality and can it be quantified? And, in the end, are there any means against that?

3.3 Physics-of-Failure Concept

The physics-of-failure concept is an approach to design and development of reliable products to prevent failure based on the knowledge of root cause failure processes. It is based on understanding the effects of loads (stressors) on product materials and their influence on the life time with respect to the use conditions and time [20].
This concept dominantly works for the wear-out zone of the bathtub curve. The root cause failure processes in this work are parametrical degradation mechanisms that are discussed in the following section 3.4.
These parametrical degradation mechanisms did not occur so drastically in the last decade, they are the result of continuous scaling. While scaling down the geometry of a microelectronic circuit with the factor 1/k, the supply voltage must be scaled with the same factor, otherwise the stressors are increasing. Table 3.2 shows the scaling principles of microelectronic components. The right column repesents constant field scaling, the middle column non-constant field scaling at constant V_{DD}. In practice, a value between those two extremes is given. The factor k typically is approx. $\sqrt{2}$, e.g. between 90 nm and 65 nm technology.

supply voltage scaling	1	1/k
packing density	k^2	k^2
drain current per channel width	k^2	1
current density	k^3	k
oxide field strength	k	1
power dissipation density	k^3	1
power dissipation per gate	k	1/k
gate delay	$1/k^2$	1/k

Table 3.2: Principle of scaling: if the supply voltage was scaled with the same factor 1/k like the geometry (typically $1/\sqrt{2}$ from one technology node to the next), all values except current density would either improve or stay the same. Unfortunately, supply voltage scaling is limited by noise and threshold voltage; therefore the voltage scaling is somewhere between both columns. This means increasing stressors. [24]

The clear trend is, that operational stresses increase dependent on voltage scaling, but voltage scaling is limited by noise and threshold voltage. This is why the voltage cannot be scaled according to the geometry scaling factor ('non-constant field scaling'), and the result is stronger electrical fields and therefore stronger degradation mechanisms in every technology node. They are the reason why this work is now done for 65 nm technology node, when the influence is starting to have heavy impact on the circuit behavior and its reliability.

When the stresses become so huge that the next scaling step is not possible anymore, materials research has to supply a new material with different properties. This is the case for high-κ materials. With conventional SiO_2 gate dielectrics, the 65 nm technology node was one of the last to have reasonable performance in terms of leakage. The next miniaturisation step would have increased leakage so much that a new material with higher dielectric constant had to be found. It enables to increase the gate insulator thickness but keep or even improve the gate capacity, which is necessary for good transistor behavior. On the other hand, however, it is the source for a new parametrical degradation mechanism.

3.4 Parametrical Degradation Mechanisms

There are 4 different voltage and current scenarios for a metal oxide semiconductor field effect transistor (MOSFET) causing 3 types of parametrical degradation mechanisms:

1. Conducting transistor, but no electrical field over the channel and therefore no current → BTI (Bias Temperature Instability), left half of Fig. 3.3

2. Conducting transistor and electrical field over the channel, therefore current flowing → HCI (Hot Carrier Injection), right scenario in Fig. 3.3

3. Non-conducting transistor, but electrical field over the channel → off-state Stress or NCHCI (Non-conducting Hot Carrier Injection), 2nd from right in Fig. 3.3

4. Non-conducting transistor, no electrical field over the channel → no degradation mechanism (not considered further)

In Fig. 3.4 the corresponding regions in the output characteristic of a MOSFET are shown. In the following sections, these 4 parametric degradation effects will be discussed in detail.

BTI also occurs when the transistor is conducting, but the effect is reduced due to the lowering electric field along the channel from source to drain. As this mode is not of importance for SRAM cells, it will not be regarded here.

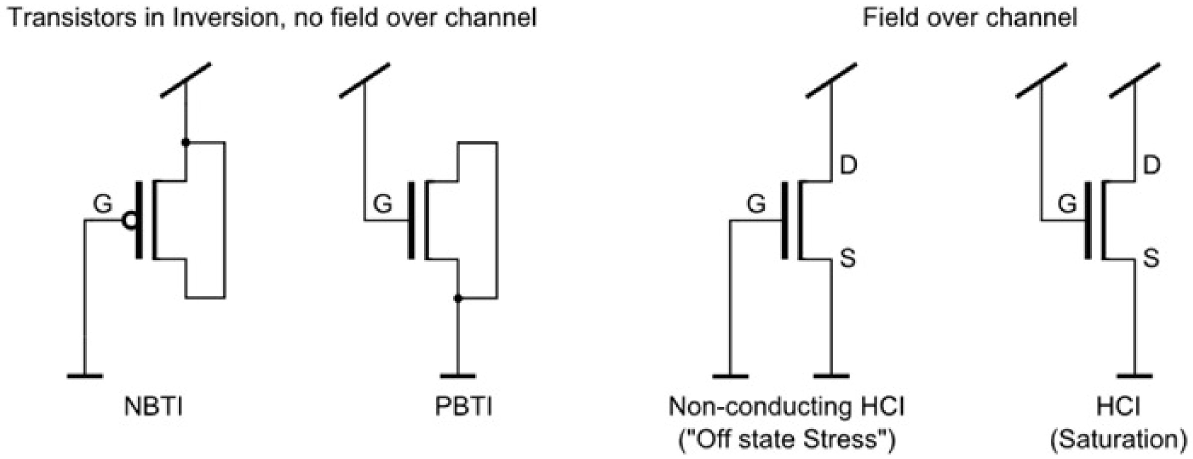


Fig. 3.3: The 4 parametrical degradation effects caused by the 3 different connection scenarios. NBTI is only active for pMOS and PBTI is only active for nMOS, while HCI and NCHCI are active for both polarities.

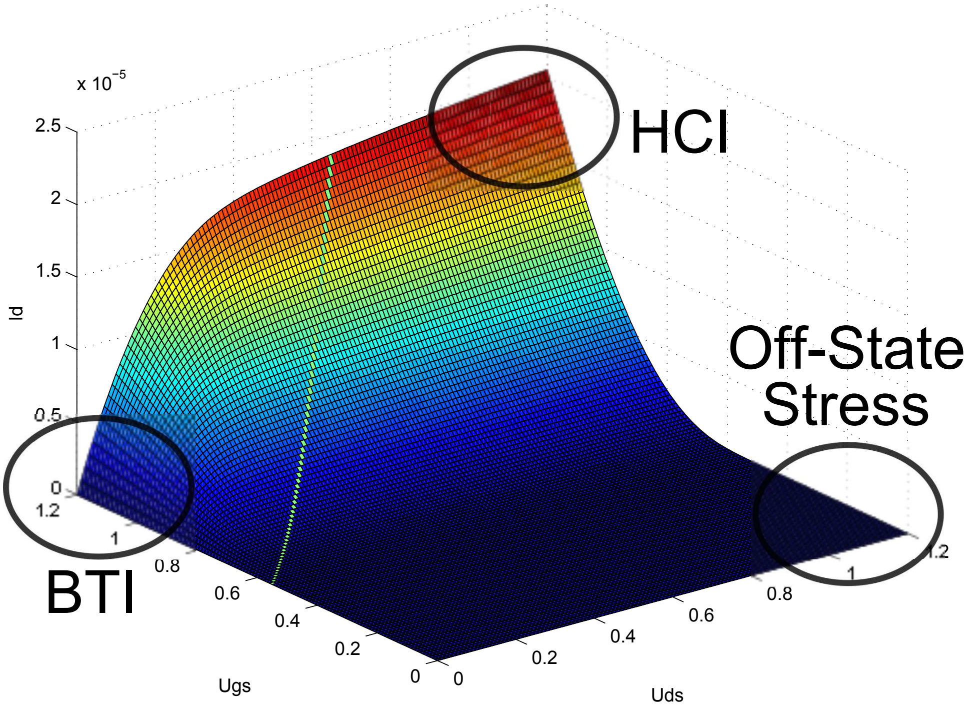


Fig. 3.4: 3D plot of the output characteristic of a MOSFET. The fine line is separating the triode region from the saturation region. In every section of the plot, there is one active degradation mechanism.

3.5 Negative Bias Temperature Instability (NBTI)

3.5.1 NBTI threshold voltage drift ΔV_{th}

Negative Bias Temperature Instability (NBTI) is a degradation effect that occurs on pMOS transistors in inversion without electric field over the channel [25] [26]. Fig. 3.5 on the left shows the stress conditions: gate is pulled to ground, while source and drain are both connected to V_{DD}. Many parameters will be shifted by NBTI, e.g. the transconductance g_m, the channel mobility μ_0 or the on and off current. But all these shifted parameters can be modeled with the shift of only one core parameter, the threshold voltage. Fig. 3.5 on the right shows the transfer characteristic of a stressed vs. a non-stressed device. The threshold voltage V_{th} got shifted to smaller values, i.e. a transistor with e.g. $V_{th} = -0.5V$ will degrade to e.g. $V_{th} = -0.55V$.

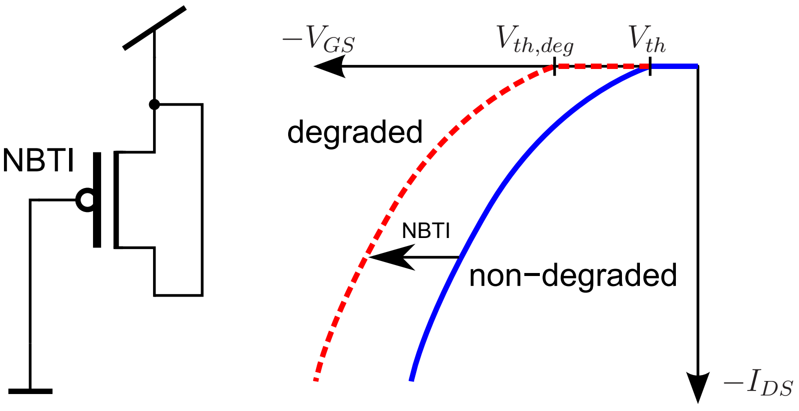


Fig. 3.5: Stress conditions (left) for NBTI and the impact of this degradation mechansims on the output charactersitc: shift of V_{th} (right)

The shift of this threshold voltage due to NBTI was modeled. This is the formula which gives an estimation of the V_{th} drift [27].

$$\Delta V_{th} = A \cdot (\frac{|V_{gs}|}{t_{inv}})^m \cdot exp(\frac{\Delta E}{kT}) \cdot L^\alpha \cdot W^\beta \cdot t^n \qquad (3.5.4)$$

The formula can be divided in four parts, which represent several factors:

1. Electrical field dependency with power law. The exponent m is getting bigger with smaller technology (90 nm: ≈ 2, 65 nm: ≈ 4, 32 nm: ≈ 5). t_{inv} is not the geometrical oxide thickness, but the effective electrical oxide thickness in inversion given by the location of the inversion layer, which is thicker. Here, the geometrical oxide thickness is 1.8 nm, but the t_{inv} value is 2.85 nm.

2. Exponential Arrhenius temperature dependency with Boltzmann constant k=$8.617 \cdot 10^{-5}$ eV/K and effective activation energy ΔE.

3. Geometry dependency over power law (L=length, W=width of the device)

4. Time dependency over power law

Using this formula for typical scenarios, Fig. 3.6 shows the plot of ΔV_{th} over the gate source voltage V_{gs} for an example of 10^4 s stress time for two different temperatures. It is getting clear that the V_{th} drift is depending heavily on temperature and gate-source voltage.

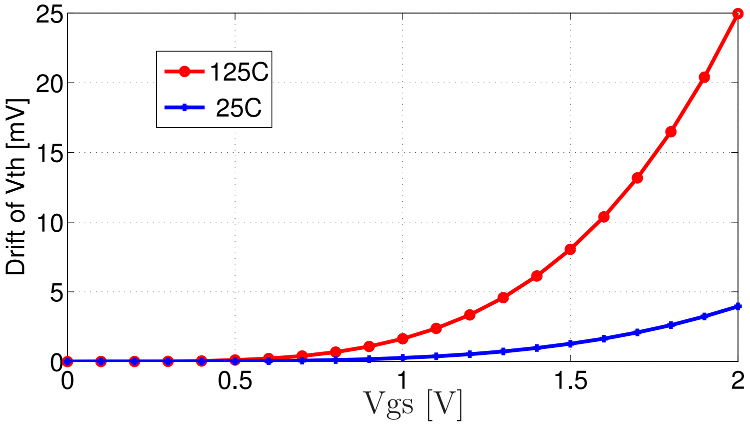


Fig. 3.6: ΔV_{th} over V_{gs}. Increasing temperature from room temperature to 125 °C increases the V_{th} drift by a factor of 6. Doubling V_{gs} increases the V_{th} drift by a factor of 15.

The dependency on V_{gs} is calculated from:

$$\frac{\Delta V_{t1}}{\Delta V_{t2}} = \frac{(\frac{V_{g1}}{t_{inv}})^m}{(\frac{V_{g2}}{t_{inv}})^m} = (\frac{V_{g1}}{V_{g2}})^m \tag{3.5.5}$$

So compared to nominal supply voltage of 1.2 V, for 1.0 V V_{gs} there is still 48% ΔV_{th}, at 0.8 V V_{gs} there is 20% and at 0.6 V V_{gs} a rest of 6%. In other words, doubling V_{gs} increases ΔV_{th} by about a factor of 15. But NBTI is also heavily depending on temperature. This shows the comparison of both curves in Fig. 3.6: one is for 25 °C, the other one for 125 °C. Increasing the temperature from room temperature to 125 °C increases the ΔV_{th} by about a factor of 6.

Fig. 3.7 shows ΔV_{th} shift over linear lifetime. This shows the typical log-like behaviour so that 50% of the final degradation after 10 years is reached after approx. 1 year.

To add several degradation scenarios resulting from several consecutive stress-steps, the single ΔV_{th} cannot simply be added, since they are not adding linearly. Instead, the following equation must be used:

$$\Delta_{tot} = (\Delta_1^{(1/n)} + \Delta_2^{(1/n)} + \Delta_3^{(1/n)} + ...)^n \tag{3.5.6}$$

The worst case for SRAM products is 10 years @ 125 °C and 1.32 V (110% nominal V_{DD}),

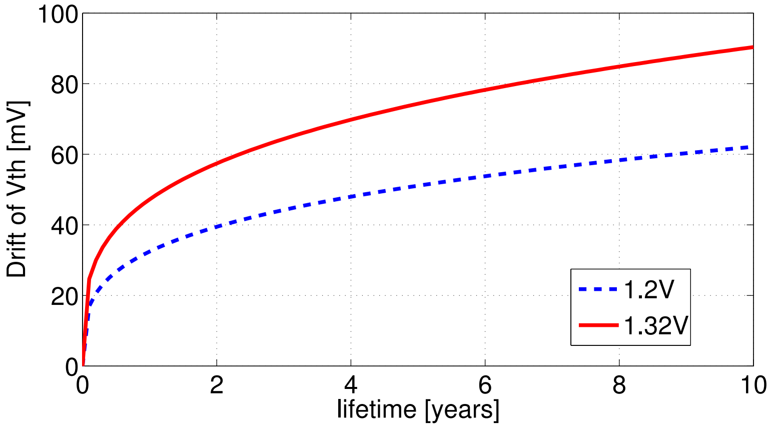
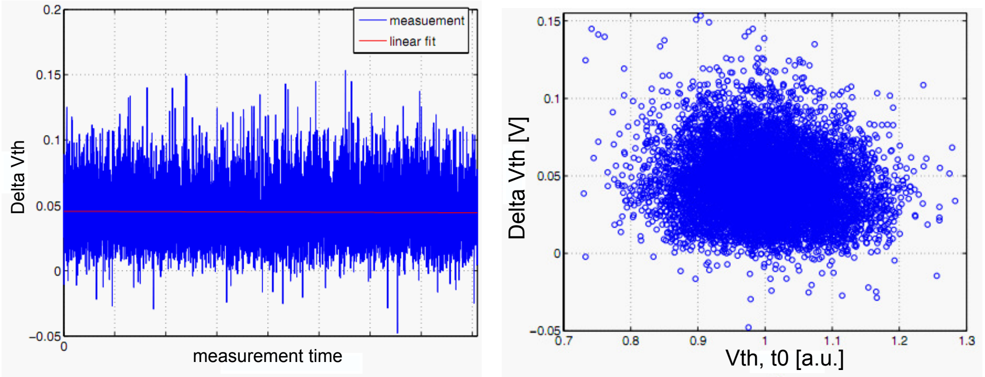


Fig. 3.7: ΔV_{th} over time. 50% of the degradation after 10 years is already reached after approx. 1 year.

resulting in almost 100 mV ΔV_{th}, compare Fig. 3.7. But the biggest problem of NBTI is not only the V_{th} shift, but two other major issues: variability and recovery [28], which are covered in the next two sections.

3.5.2 NBTI variability

The model provided so far only supports the mean value of threshold voltage drifts of big transistors (area in the µm^2 region). SRAM transistors have an area in the 5000 nm^2 region, which is about 2 orders of magnitude smaller. NBTI is a highly statistical process, as measurements on SRAM-sized pMOS transistors (W/L=90 /65 nm) show in Fig. 3.8(a) and Table 3.3 [29]. After 10^4 s stress with 2.4 V stress voltage at 125 °C, the threshold



(a) pMOS V_{th} shifts caused by NBTI after 10^4 s stress with 2.4 V and 500 s recovery [29]. Shift values between +40 mV and -150 mV show huge variability.

(b) High threshold voltage values before stress and high threshold voltage drift due to NBTI are not correlated

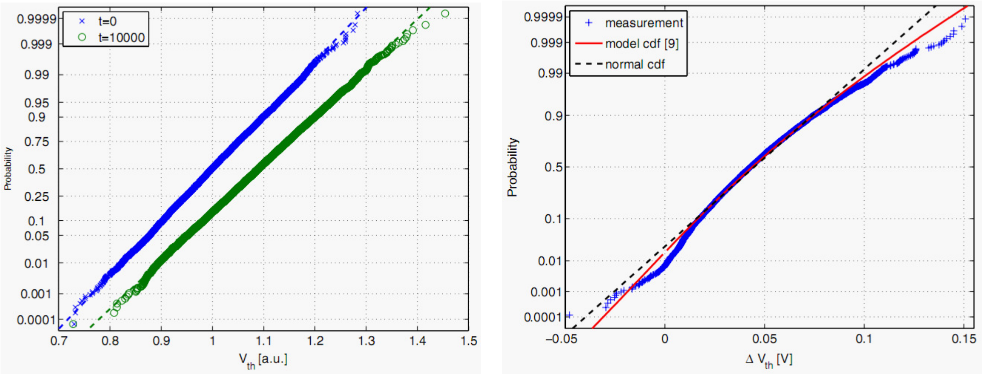
Fig. 3.8: Measurements on SRAM-sized pMOS transistors (W/L=90/65 nm) taken from [29]

voltage drift of each device varies between +40 mV and -150 mV. It is important to note that there are only a few mavericks that show V_{th} drift to more positive values. The scatter plot in 3.8(b) shows that the transistors with high threshold voltage values are not the devices with high threshold voltage drift. [29]. Both values are uncorrelated, which is good news for circuit design; this avoids single devices with extremely high threshold voltage after stress. While the threshold voltage is normal distributed over SRAM cells,

	σ/μ	μ	σ
V_{th} (t=0)	0.077	-	-
V_{th} (t=10^4s)	0.078	-	-
ΔV_{th}	0.51	44.9mV	22.8mV

Table 3.3: Ratio of mean value and standard deviation shows a highly statistical process [29]

NBTI creates a shift of the mean value of this distribution. Additionally, the standard deviation seems to increase slightly (Fig. 3.9(a) and Table 3.3). Another hint to a wider V_{th} distribution after NBTI stress is the ΔV_{th} distribution, which does not follow a Normal distribution exactly (Fig. 3.9(b)).



(a) NBTI causes a mean value right-shift of threshold voltage, while the standard deviation is slightly increased. V_{th} is normal distributed pre and post NBTI stress.

(b) The threshold voltage drift does not show perfect Normal distribution. Too many transistors show a huge V_{th} shift.

Fig. 3.9: Distributions of absolute threshold voltage and threshold voltage drift taken from [29]

3.5.3 NBTI recovery

NBTI not only has strong variability, but also shows recovering behavior after end of stress, compare Fig. 3.10. Measurements on huge single transistors in 90 nm technology (W=10 µm, L=0.12 µm) have shown that directly after termination of NBTI stress conditions, the V_{th} shift starts to drop [30]. In Fig. 3.11 the x axis represents the time after end of stress, while the y axis represents the V_{th} shift. This happens with extremely short

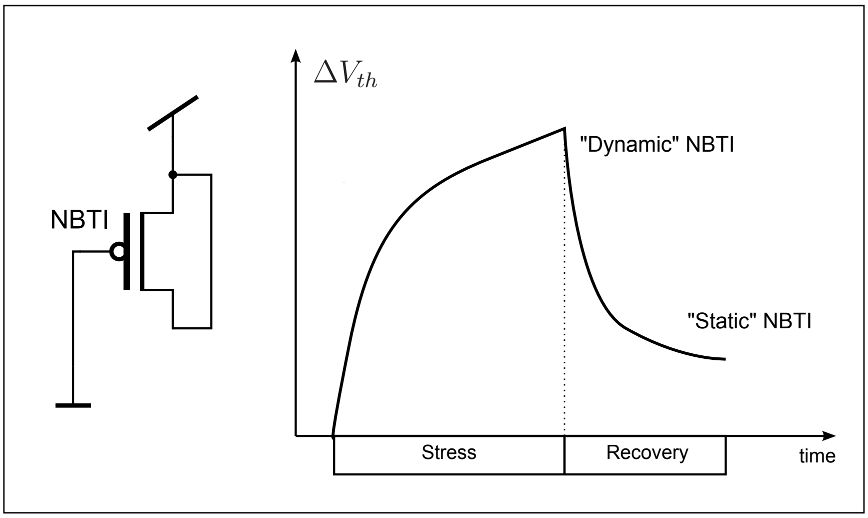


Fig. 3.10: Stress condition of NBTI and qualitative diagram of static and dynamic NBTI on single pMOS transistor: During stress, V_{th} shift grows quasi-logarithmically. Directly after end of stress, V_{th} rapidly decreases to a quasi-static value, which then only changes in very long timescales.

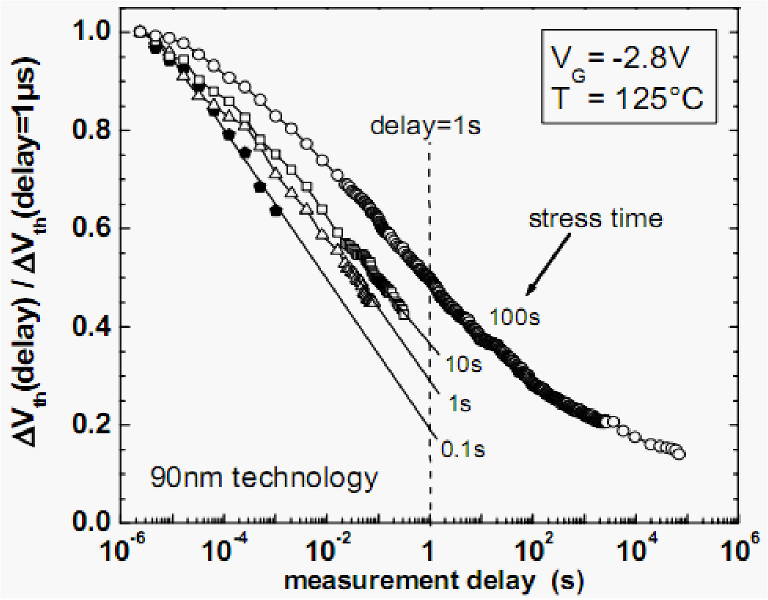


Fig. 3.11: Recovering behaviour of NBTI directly after end oft stress. At 1 s after end of stress, approx. 50% of the ΔV_{th} has disappeared. Original plot taken from [30]

time constants, so that 1 s after end of stress, 50% of the maximum ΔV_{th} @ 1 µs after end of stress has disappeared. This behaviour leads to 2 problems.

First, the NBTI ΔV_{th} models would have to refer to a specific time after stress, since this value is changing drastically. For the currently used models, this is not the case.

Second, all measurements would have to be done immediately (i.e. some ns after end of stress). This is not possible, and considering this effect in the measurements is a huge challenge.

3.5.4 NBTI physical background and modeling

NBTI appeared in 1967 [25], and in the 1970s, the first model describing this phenomenon was created: the reaction-diffusion model. This is based on the theory that H+ ions are trapped in the oxide by a diffusion process, which also explains the high temperature dependency [26]. Although this model was under discussion and was almost accepted for decades, it could not explain the recovery behavior precisely [31]. The model therefore got changed from 'reaction-diffusion' to 'capture-emission' in the last 2 years [32] [33]. Now it seems that diffusion is not involved in this process [31]. With this theory, degradation means trapping ('Capture') of charges with a broad range of time constants, similar to 1/f noise. Recovery on the other hand means detrapping ('Emission') of charges with comparably, but independent broad range of time constants, again similar to 1/f noise.

3.6 Positive Bias Temperature Instability (PBTI)

With scaling the transistor sizes and widths of dielectrics, leakage is becoming more and more dominant. Until 65 nm technology, leakage was in an acceptable region. But going below 65 nm requires the introduction of new gate dielectrics, which allows to increase gate thickness on one hand to decrease leakage, but at the same time increase or at least keep the dielectric properties. This can be done with new materials with a higher dielectric constant, so-called high-κ dielectrics.

Unfortunately, this is the source for another parametric degradation effect: the Positive Bias Temperature Instability (PBTI) for nMOS transistors. The behavior is very similar to NBTI: charges are trapped in the gate oxide [34]. But contrary to NBTI, no positive charges, but electrons are trapped. The result is threshold voltage drift to more positive values, so that the transistor is also getting weaker over time. Fig. 3.12 shows the stress conditions and the impact of this degradation mechanism. PBTI is even less understood than NBTI. Right now, PBTI is discussed not so often in literature, as even the much better known NBTI modeling is still not well understood.

First measurements in Fig. 3.13 have shown that the amount of threshold voltage drift is similar to NBTI. Also, PBTI shows the same variability and recovery like NBTI [36]. So the analytic model describing the mean V_{th} shift is similar to NBTI, but with a different set of parameters. But since 65 nm technology is a non-high-κ technology, no PBTI exists, so this is not discussed further.

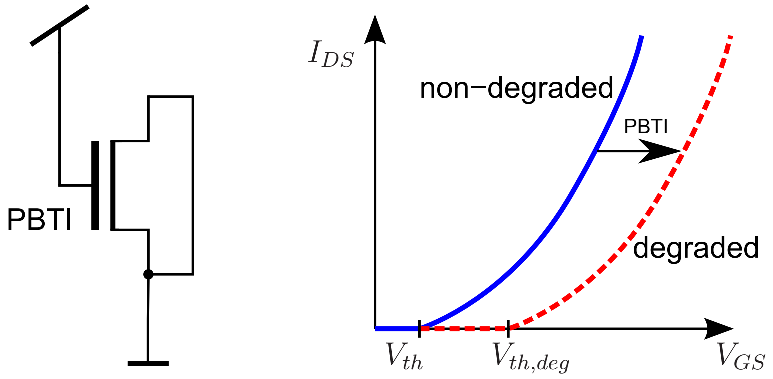


Fig. 3.12: Stress conditions (left) for PBTI and the impact of this degradation mechansim on the output characteristic: shift of V_{th} (right)

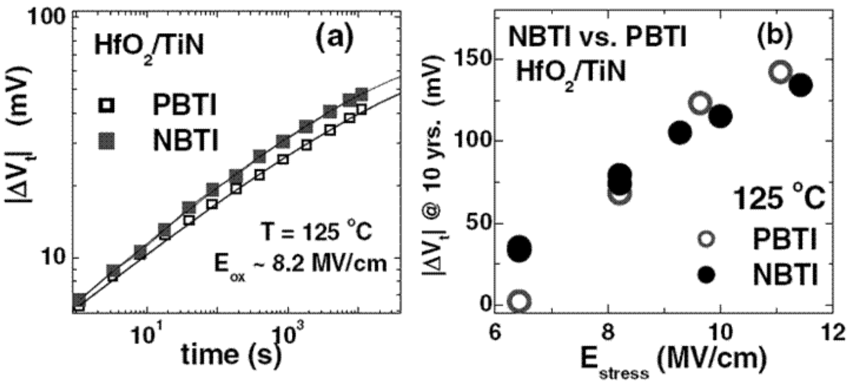


Fig. 3.13: Measurements show that the threshold voltage drift of NBTI and PBTI is about the same value under the same stress conditions. High-κ material is a stack of HfO_2/TiN. Original plot taken from [35]

3.7 Hot Carrier Injection (HCI) and Non-conducting HCI (NCHCI)

Hot Carrier Injection (HCI) is the best understood parametrical degradation mechanism of all covered in this work. It was very famous in the 1980s, when supply voltages where higher [37]. This effect got reduced in its impact in the last decades by technology development, esp. drain engineering like Lightly Doped Drain (LDD). So in 130 nm technologies and below, the HCI effect almost disappeared. But now, with smaller and smaller channel lengths, the effect might return.

When carriers travel through regions of high electric field, they can gain large kinetic energy. When the mean energy is getting larger than that associated with the lattice in thermal equilibrium, they are called 'Hot' because the carriers were historically assumed to be thermally distributed at an effective temperature higher than that of the lattice [23]. These high-energy carriers can be injected into the gate oxide or cause interfacial damage.

To get enough kinetic energy, high electric fields are needed. Since E=V/d, the voltage V_{ds} must be high and the distance L (the length of the device) must be small. In 1970, L was big, but so was V_{ds}. Nowadays, V_{ds} got reduced to about 1 V, but the length got down to about 65 nm. This is why this effect came back.

Two mechanisms of HCI must be distinguished: conducting and non-conducting Hot Carriers. Non-conducting HCI is also called 'off-state stress'. While in the conducting HCI the source of the channel carriers is the drain current in the pinch-off region (at SRAM size in the 10 µA range), in the NCHCI case this is the channel off-current (at SRAM size in the pA range). Since the on-current is a factor of 10^6 higher, it is clear that the effect of NCHCI is small compared to HCI.

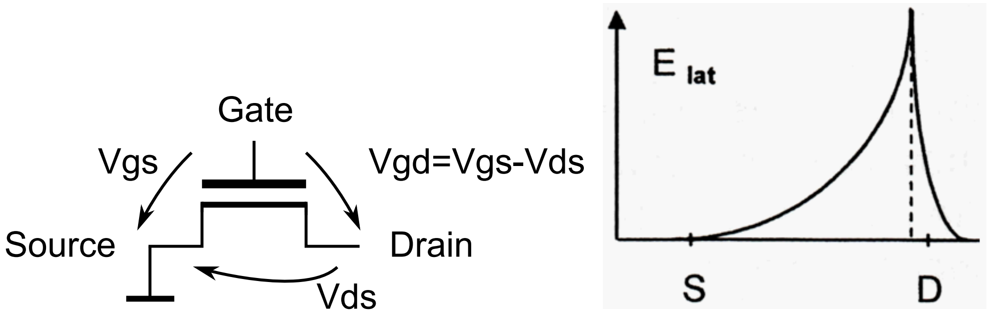
The hot carrier damage will modify the electrical characteristics of the MOSFET device (V_{th}, I_{on},etc.). These modifications can impact the functionality of SRAM cells.

3.7.1 Hot Carrier Injection (HCI)

When V_{gs} is relatively high compared with V_{ds}, the resistivity along the inverted channel is nearly constant and the potential varies linearly between source and drain (Fig. 3.14(a)). If, however, V_{gs} is comparable to or lower than V_{ds}, the inversion layer is much stronger on the source side than on the drain side, and the voltage drop due to the channel current is concentrated on the drain side, compare Fig. 3.14(b). Carriers traveling from the source to the drain can gain a considerable amount of energy in this drain-side high-field region [20]. The field can be so high that carriers gain a significant amount of energy between two scattering events.

So the device must be in saturation mode, which is $V_{gs} > V_{th}$ and $V_{ds} > V_{dsat}$. Only then the electrical field has a strong peak on the drain side of the device, which does not happen in triode region. High V_{ds} and short L lead to an extremely high lateral electric field on the drain side (compare Fig. 3.14(b)). While in analog circuits, minimum transistor lengths do not occur, it is the case for digitial (and especially SRAM) circuits

and therefore HCI is a possible candidate for a degradation mechanism.
Some of the hot carriers gain so much energy that they can surmount the energy barrier at the Si/SiO_2 interface and be injected into the oxide. If the hot-carrier injection is strong enough, the trapped charges or generated defects will permanently modify the electric field and hence the electrical characteristics of the FET. This especially results in a degradation of the on-current. HCI is mostly degrading transistors with full V_{ds} and V_{gs}. Then the transistor is in saturation with maximum overdrive voltage. In case V_{gs} and/or V_{ds} are reduced, the HCI degradation is lowered.



(a) Inversion needs voltages $> V_{th}$ over the gate. For $V_{gs} < V_{ds} - V_{th}$, the threshold voltage is not reached anymore on the drain side of the device. This results in pinch-off.

(b) The drain-sided pinch-off leads to high lateral electric fields.

Fig. 3.14: Voltages and high lateral electric field on drain side

Models have been developed to calculate the lowered current in the channel. This formula calculates the on-current degradation in percent [38]. Each device needs a distinct set of parameters A, m, V_0 and n.

$$\Delta I_D(\%) = A \cdot L^m \cdot exp(\frac{V_{ds}}{V_0}) \cdot t^n \tag{3.7.7}$$

Especially for nFETs, HCI is depending strongly on the applied drain-source voltage V_{ds}. This can be calculated from 3.7.7

$$\frac{\Delta I_{D1}}{\Delta I_{D2}} = \frac{exp(-\frac{V_0}{V_{DS1}})}{exp(-\frac{V_0}{V_{DS2}})} = exp(-V_0(\frac{1}{V_{DS1}} - \frac{1}{V_{DS2}})) \tag{3.7.8}$$

So for nFETs, below a drain-source voltage of 1.0 V, this effect can be neglected, compare Fig. 3.15.
For pFETs, the dependency on the drain-source voltage is smaller, comparable to NBTI voltage dependency, compare Fig. 3.16.

3.7.2 Non-conducting HCI

For the non-conducting case, V_{gs}=0 (or at least below V_{th}) and V_{ds} is high. This means that the device is switched off, but an electric field over the channel exists. This is the reason

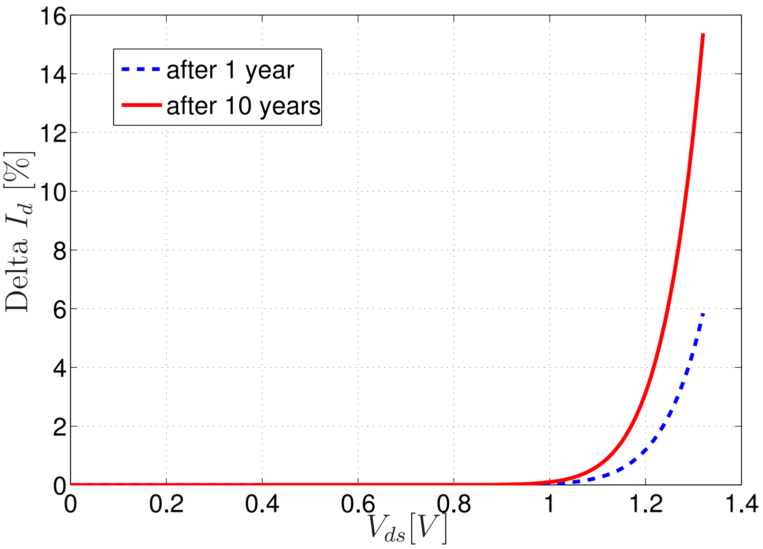


Fig. 3.15: HCI sensitivity of nFET on V_{ds}: strong degradation after 10 years with nominal or increased supply voltage. When V_{ds} is lowered to 1 V, HCI damage has disappeared.

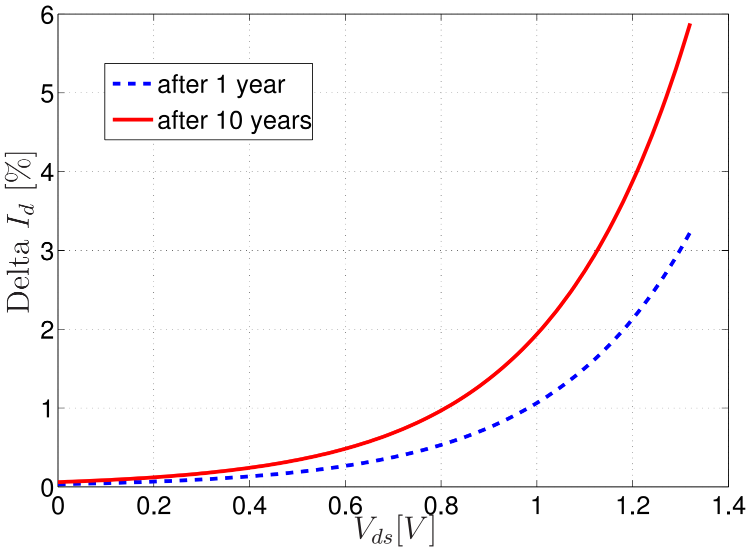


Fig. 3.16: HCI sensitivity of pFET on V_{ds}: less strong degradation compared to nFET, but still active with lowered supply voltage.

for subthreshold leakage currents in the channel, which causes the NCHCI phenomenon. NCHCI shows a strong dependency on channel length [39] [40]. Even small variations of channel length have high impact on the device lifetime [41].

The off-state degradation for nFETs is dominated by holes. The damage of the silicon dielectric is located at the drain region and results in increase of the threshold voltage [42]. This is the formula in 65 nm technology which gives an estimation of the NCHCI degradation in saturation current for nFET devices:

$$\frac{\Delta I_{on}}{I_{on}} = \frac{A}{100} exp(-\frac{V_0}{V_{DD}})(I_{off} t_{eq})^n \tag{3.7.9}$$

But this formula is only accurate under burn-in or similar conditions, otherwise it might not be accurate. While I_{off} represents the current density at burn-in conditions including all leakage currents, t_{eq} is the total time the device is turned off. The parameters A, V_0 and n are provided for each technology. Since the very important channel length L does not appear in the formula, it is clear that this is only a rough estimation. The degradation impact is so small that a precise formula cannot be provided.

For p-channel devices in this technology, no significant current shifts by NCHCI have been observed.

3.8 Accelerated Stress Measurements

All the described degradation mechanisms need years under normal working conditions to significantly shift the transistor parameters until they reach the wear-out zone. But waiting for some years is not possible. Therefore the degradation mechanisms have to be accelerated. While in automotive industry, acceleration is done by extreme hot and cold temperatures and salty environment that lets the car age in 19 weeks to a simulated age of 12 years, in microelectronics it is done by elevated electric field and temperature. The ratio between the times necessary to obtain the same degradation result at different levels of one stress as effect of the same failure mechanisms is called the acceleration factor. In this work, V_{DD} is increased up to 2.2 V at a max. (and still allowed) temperature of 125 °C. Then 1.5 years of normal life at 125 °C are already reached after 10^4 s, which is an acceleration factor of 5000. It is important not to increase the stressors (V, T) so much that another degradation mechanism is getting active. This is why in this work, the voltage was only elevated not higher than 2.2 V to keep the same degradation mechanism. It is important to keep in mind that elevated V_{DD} and temperature is not a simple way to accelerate the aging of a complex circuit [27]. Different transistors see different voltages, and raising V_{DD} does not accelerate all transistors and degradation effects by the same factor. So generally, raising V_{DD} over the specified limits is not a realistic aging acceleration. Of course, the circuit will fail sooner or later, but this does not generally allow conclusions when the wear-out zone is reached under normal working conditions. This is only the case if only one dominant degradation mode is active.

Another problem of acceleration is that the acceleration voltage must be reset to normal operating condition for measurements, which does not happen in real circuits. This especially evokes the problem of the BTI recovery effect. The stress voltage must be reduced to normal supply voltage, which cannot be done arbitrarily fast. Especially for BTI, where 50% of the recovery process has happened in the first second after end of stress, this is a real problem and will be subject of chapter 6.

3.9 Summary

Yield is the fraction of fit-for-use circuits directly after production, while reliability is yield during or after lifetime. This work uses the physics-of-failure concept to identify the worst degradation mechanisms that affect the wear-out failure rate for 6T-SRAM core cells. Non-constant field scaling causes growing electrical fields in every new technology node, which is the motor of the parametric degradation effects BTI and HCI.

BTI, especially the better researched NBTI, shows strong variability and recovery, which is not yet modeled sufficiently. NBTI is active on pMOS devices on current technologies, while PBTI is active on nMOS devices and plays a role only in combination with high-κ gate dielectrics.

HCI is active on small gate lengths, which is the case especially on SRAM circuits. HCI on nFETs shows strong degradation only with drain-source voltages >1 V. HCI on pFETs is generally the weaker of both degradation effects, however, it shows degradation also for smaller drain-source voltages.

Non-conducting Hot Carrier degradation is based on the leakage current of the devices. It is therefore generally much smaller than conventional HCI, which is caused by on-currents. For the current technology, it is almost not observable.

To reach sufficient degradation values in reasonable timeframes for measurement, all degradation mechanisms need to be accelerated. With raised voltage and temperature, the effect can be speeded up by factors of some thousand, so a stress corresponding to end-of-lifetime can be reached within hours. It is important not to increase the stressors too much to guarantee the same degradation mechanisms as in the real use case. Especially for BTI and its recovery behavior, the increased stress voltage is a huge problem for the experiments, because after only 1 s, degradation has decreased to about 50% of the maximum degradation value directly after end of stress.

Part II

Parametric Degradations on 6T-SRAM Core Cells

Chapter 4

Simulations of the four Parametric Degradation Mechanisms

Abstract

The basics of degradation mechanisms and the 6T-SRAM circuit were given in the first part. Now the impact of each of the four parametric degradation mechanisms on the performance of the 6T-SRAM cell is examined. This is done in two steps. First, voltage and current conditions for each of the 6 transistors are simulated in a 65 nm low power technology during normal SRAM operations (Section 4.1). This shows when which mechnism is active and allows an estimation on how big each transistor degradation will be. Second, the sensitivity of SRAM performance to each transistor degradation is examined (Sections 4.2 to 4.5). The product of transistor degradation and cell sensitivity will result in the impact of each degradation mechanism and allows a ranking of these four. It is then possible to focus on the worst effects.
The most important results are reported in

S. Drapatz, G. Georgakos, and D. Schmitt-Landsiedel: "Impact of negative and positive bias temperature stress on 6T-SRAM cells", Advances in Radio Science, vol. 7, pp. 191-196, 2009

4.1 Voltage and Current in each Transistor

Each degradation mechanism can be related to one region in the output characterics of a MOS transistor, compare Fig. 3.4 on page 27. Performing DC and transient analyses for the 3 use cases hold, read and write, the active degradation mechanisms can be determined for every single transistor.
First, 6 DC analyses are done for hold state 0 with WL=BL=BLB=0. Fig. 4.1 shows the constant voltages V_{GS} and V_{DS} for each transistor. Fig. 4.2 shows the qualitative location of the bias points of the 6 transistors in one output characteristics. During

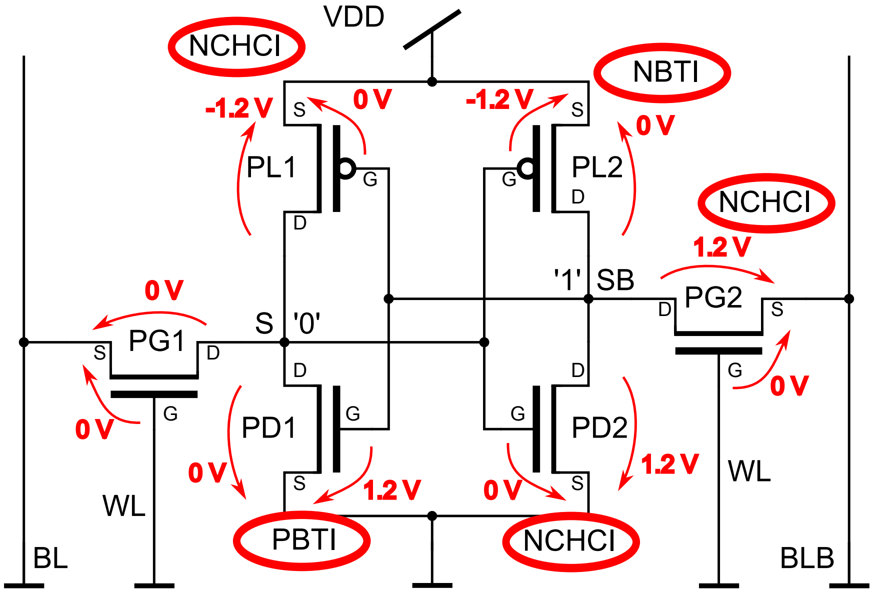


Fig. 4.1: Gate-Source voltages V_{gs} and Drain-Source voltages V_{ds} with the corresponding degradation mechanisms of a 6T-SRAM circuit in hold state 0.

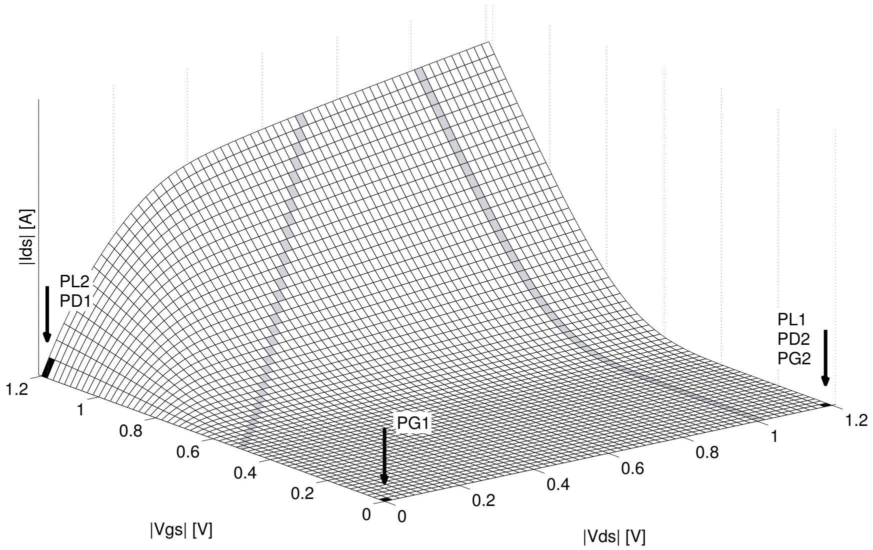


Fig. 4.2: Stable DC points of the 6 transistors in hold state 0. While PL2 and PD1 suffer BTI stress, PL1, PD2 and PG2 experience off-state stress. PG1 is not affected at all, compare Fig. 4.1.

'hold' of '0', PL2 suffers NBTI, while PD1 suffers PBTI. PL1, PD2 and PG2 suffer off-state stress. PG1 on all terminals is connected with 0 V, so no degradation mechanism is active. Of course, the conditions switch between the pairs of pullup-, pulldown- and access-transistors when the cell holds a '1'.
Next, 6 DC analyses are performed for read state with WL=BL=BLB=1. Fig. 4.3 shows the voltages V_{GS} and V_{DS} for each transistor. Figs. 4.4, 4.5 and 4.6 show the bias points of the 6 transistors in 3 output characteristics. The situation for the four pullup- and pulldown-transistors is close to 'hold' state, only the situation for the access-transistors changes. Node 'S' is raised to approx. 0.1 V, and a current, namely the read current, is visible through PG1 and PD1, which puts PG1 into the HCI area. PL2 keeps suffering NBTI, and PD2 is still in the PBTI area. PL1 and PD2 suffer off-state stress. PG2 is on all terminals connected to 1.2 V, so no degradation mechanism is active.

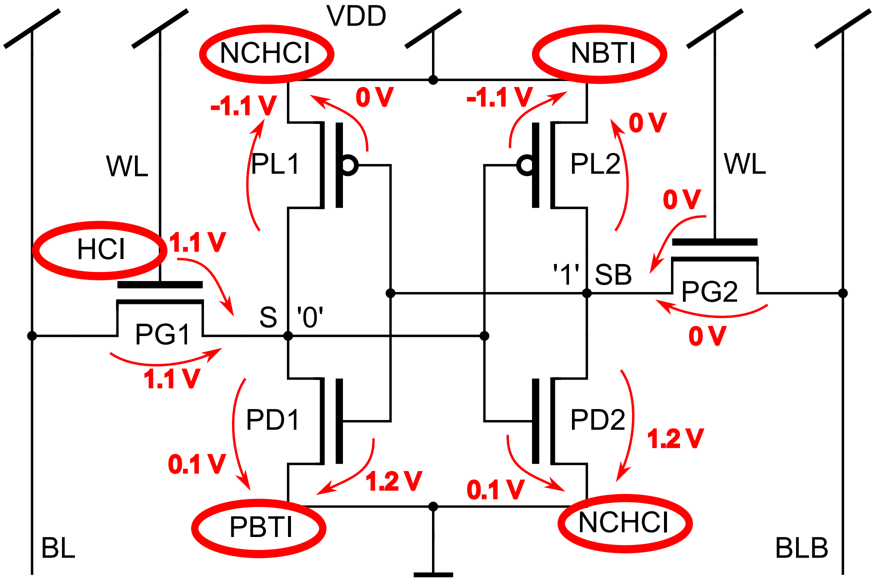


Fig. 4.3: Voltage conditions of a 6T-SRAM circuit in read state 0. Node S is raised to approx. 0.1 V and read current is flowing on the '0' memory side, resulting in HCI degradation for access transistor PG1.

DC analysis of read state alone does not capture a read cycle completely. Contrary to hold, read state only lasts until the sense amplifier kicks in and switches off the WL, which terminates the read state and goes back to hold state. So read state never gets to a steady state, this is why an additional transient analysis is necessary (Fig. 4.7). It shows that for typical capacitances of BL=100 fF and BL to BLB=1.6 fF, read current only flows for less than 1 ns. Then the BL on '0' side is discharged to approx. 1 V, and the information is read.
The last simulations are done for a write cycle. This is the only cycle where memory state changes, this is why the results of the 6 transient analyses are printed individually for each transistor (Figs. 4.8 to 4.13). First in read state 1, the BL is lowered to 0, which makes the cell switch to S=0. So in this simulation, the WL is opened first with BL=BLB=1 (read state), then BL or BLB side is switched to 0 (write the 0). It is also possible to set BL and BLB to 1/0 or 0/1 first, and then open the WL. But the difference in the

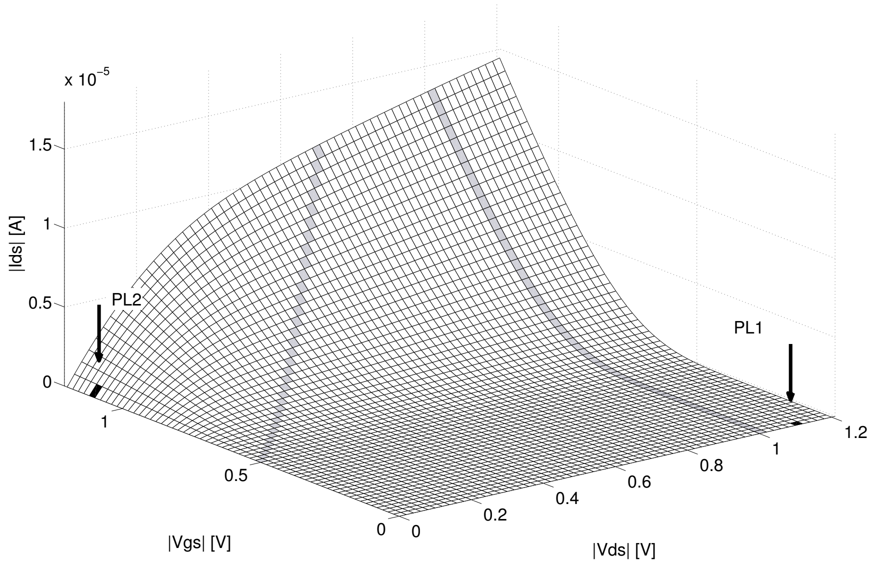


Fig. 4.4: Stable bias points of the two pMOS pullup transistors during read state 0. While PL2 suffers NBTI stress, PL1 experiences off-state stress.

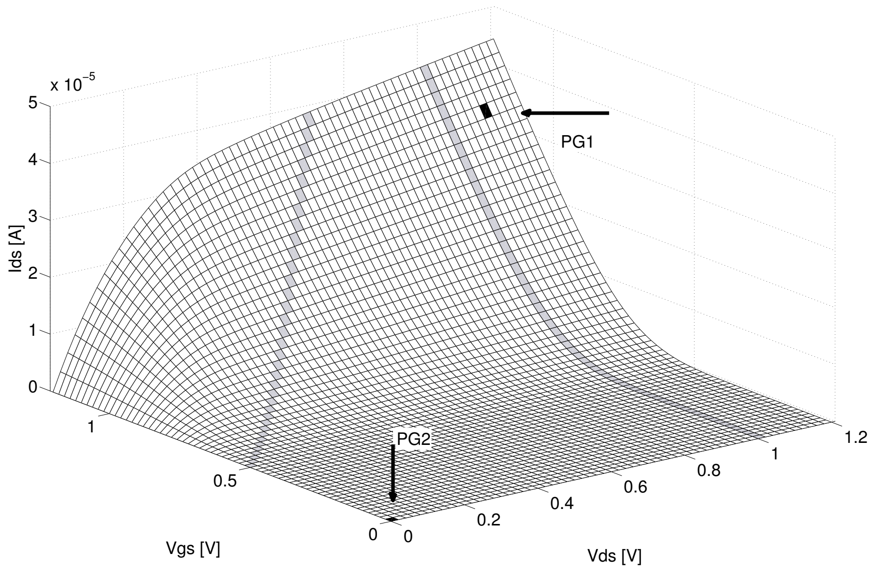


Fig. 4.5: Stable bias points of the two nMOS access transistors during read state 0. While PG2 does not see any electrical stress, PG1 experiences HCI degradation.

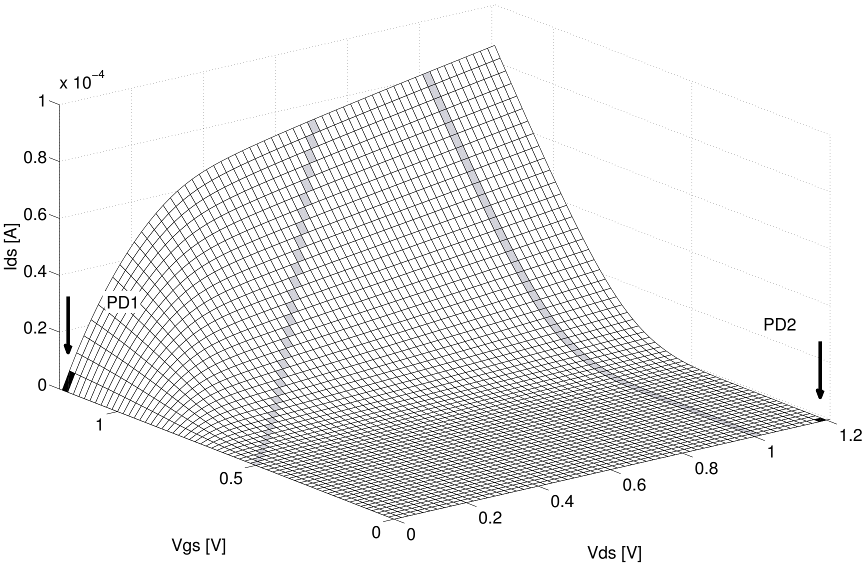


Fig. 4.6: Stable bias points of the two nMOS pulldown transistors during read state 0. While PD1 suffers PBTI stress, PD2 experiences off-state stress.

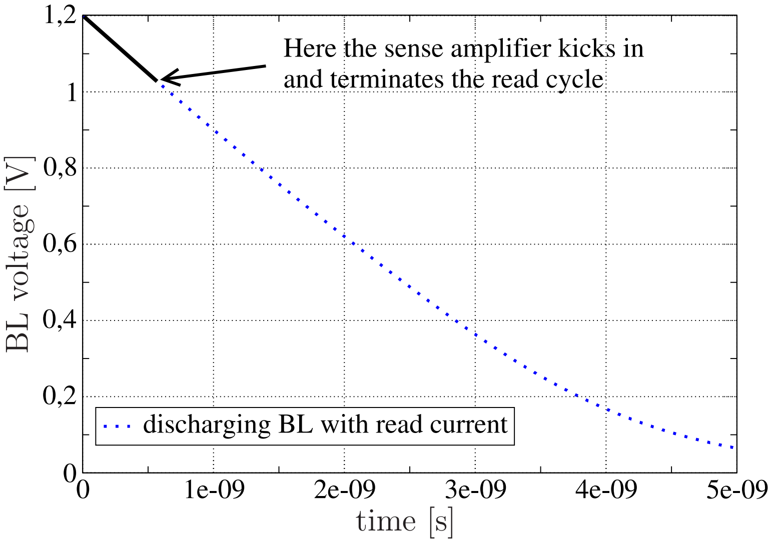


Fig. 4.7: Discharging BL capacity with read current takes about 0.5 ns to reduce voltage to 1 V. Then the sense amplifier kicks in and terminates the read cycle.

transient analyses is marginal.
The result of the switching is that the pullups and pulldowns are switched from BTI to NCHCI area, but without going deeply into HCI area. Only access transistors start in HCI area, then go to PBTI- and no-stress-area.

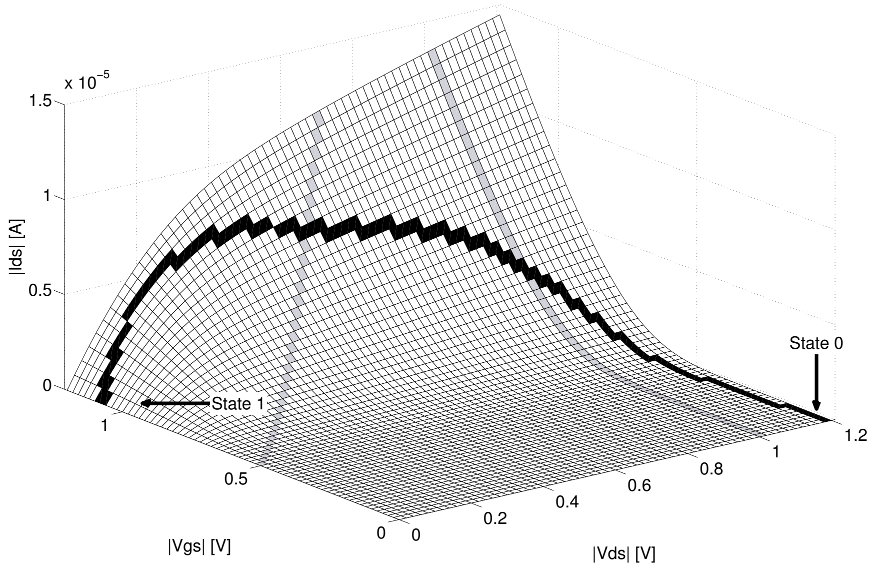


Fig. 4.8: Transient behavior of pullup PL1 during write cycle from 1 to 0. After NBTI stress, PL1 suffers off-state stress.

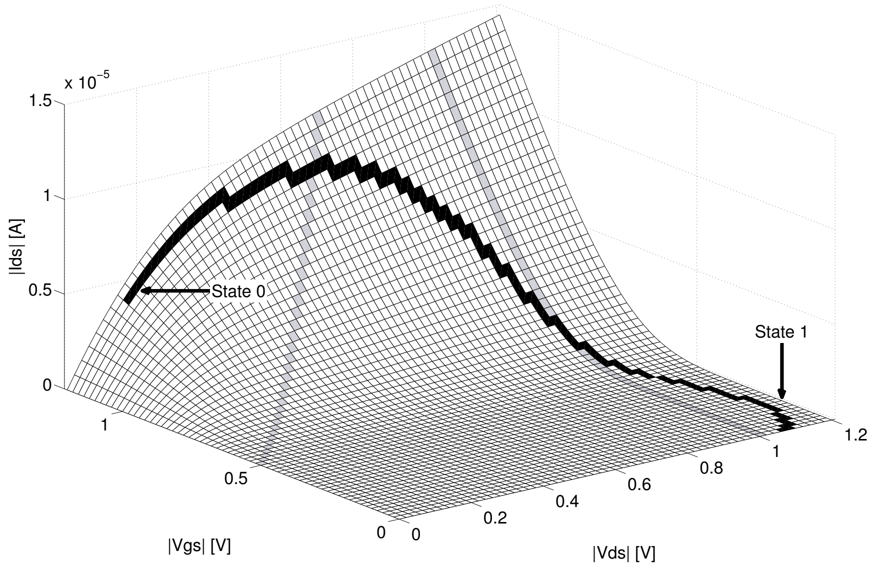


Fig. 4.9: Transient behavior of pullup PL2 during write cycle from 1 to 0. After off-state stress, PL2 suffers NBTI stress.

Again, additional to the voltage curves, the timing information is necessary. In Fig. 4.14 the transient analysis of a write cycle is plotted. Since the BL need not be discharged, this cycle is even faster than reading. The complete shift is done in approx. 0.5 ns.

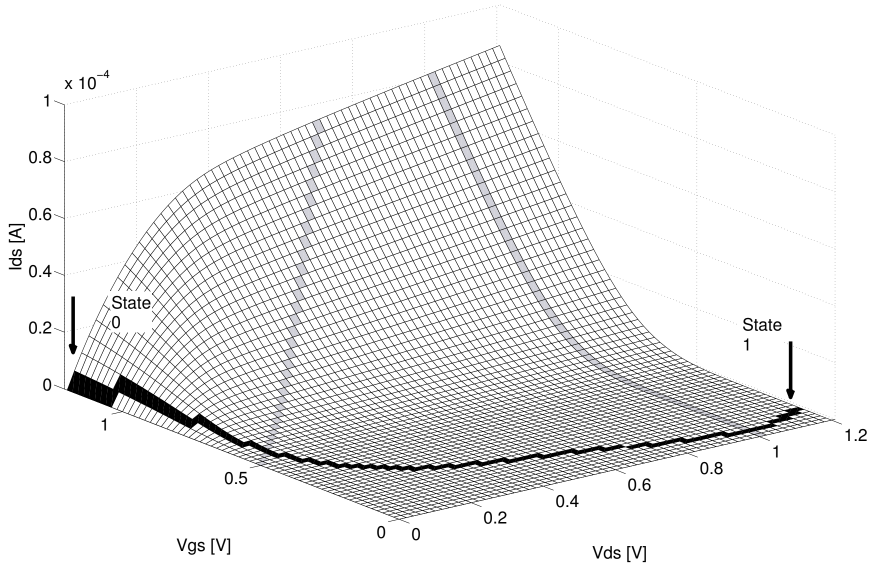


Fig. 4.10: Transient behavior of pulldown PD1 during write cycle from 1 to 0. After off-state stress, PD1 suffers PBTI stress.

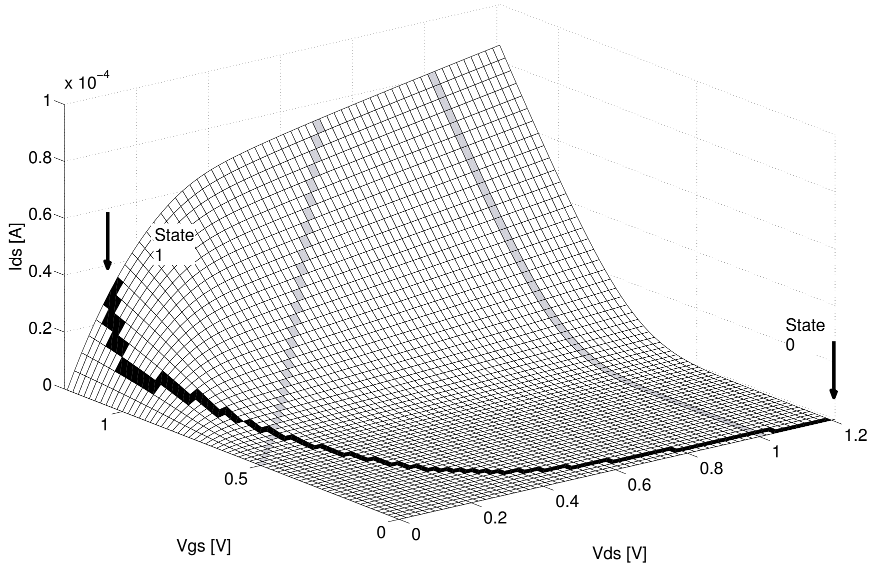


Fig. 4.11: Transient behavior of pulldown PD2 during write cycle from 1 to 0. After PBTI stress, PD2 suffers off-state stress.

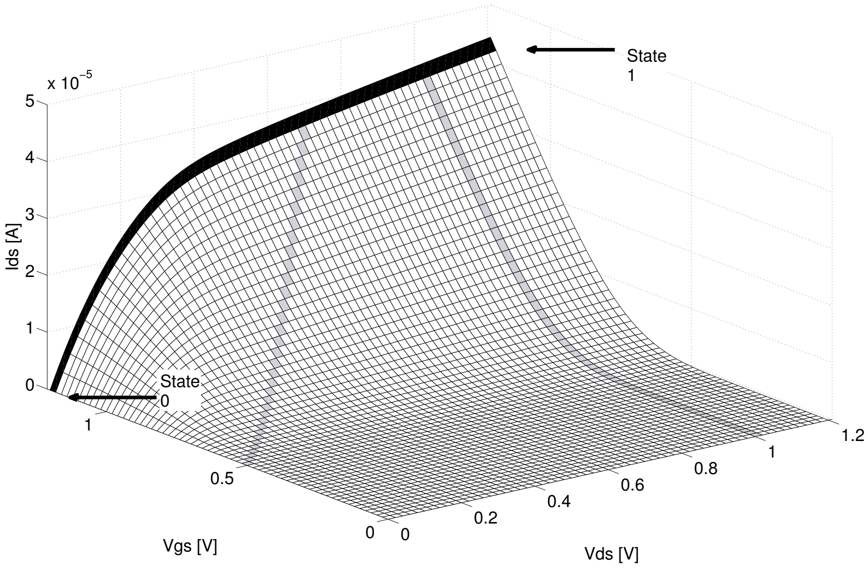


Fig. 4.12: Transient behavior of access PG1 during write cycle from 1 to 0. After HCI stress, PG1 experiences PBTI stress.

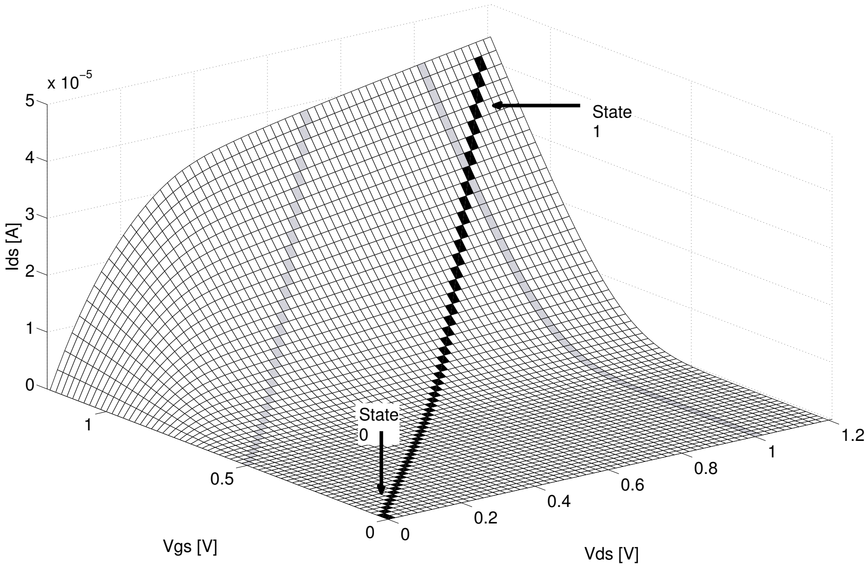


Fig. 4.13: Transient behavior of access PG2 during write cycle from 1 to 0. After HCI stress, PG2 does not see any more stress.

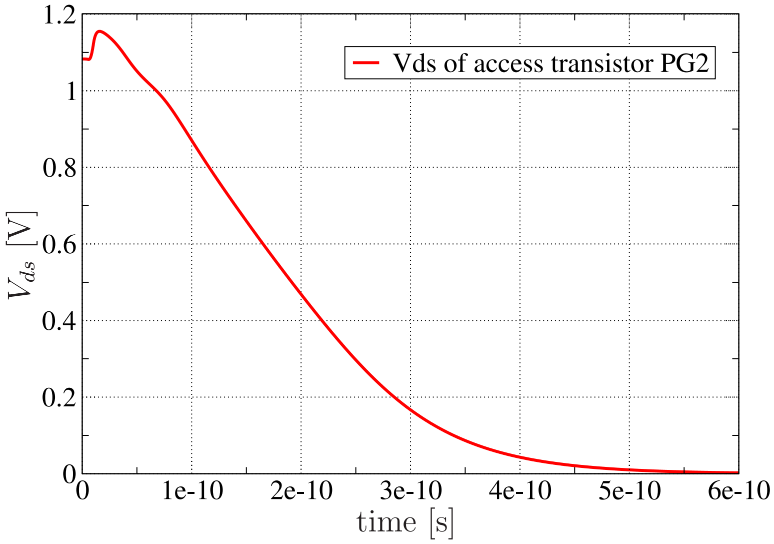


Fig. 4.14: Timing behavior of a write cycle: done in 0.5 ns.

So the result of all simulations is that one pullup and the opposite pulldown are suffering BTI during hold and read. HCI is only active during the very short read and the even shorter write cycle on one access transistor. Tables 4.1 and 4.2 show an overview of all transistors in all possible situations, describing their degradation mechanism. Table 4.3 shows the core results of these simulations.

After simulation of the voltages and currents to determine the different degradation mechanisms, the sensitivities to each degradation mechanism are examined individually in the next sections.

	Hold 0	Read 0	Write 1 to 0	Write 0 to 0
Pullup PL1	NCHCI	NCHCI	first NBTI then NCHCI	NCHCI
Pullup PL2	NBTI	NBTI	first NCHCI then NBTI	NBTI
Pulldown PD1	PBTI	PBTI	first NCHI then PBTI	PBTI
Pulldown PD2	NCHCI	NCHCI	first PBTI then NCHCI	NCHCI
Access PG1	n/a	HCI	first HCI then PBTI	PBTI
Access PG2	NCHCI	n/a	first HCI then n/a	n/a

Table 4.1: Overview of the four degradation mechanisms NBTI, PBTI, HCI and NCHCI on all possible SRAM modes. This table focuses on the '0' memory state. It shows which degradation mechanism is activated during which operation mode. Write 0 to 0 means to write a 0 to the cell which is already in 0 state.

	Hold 1	Read 1	Write 0 to 1	Write 1 to 1
Pullup PL1	NBTI	NBTI	first NCHCI then NBTI	NBTI
Pullup PL2	NCHCI	NCHCI	first NBTI then NCHCI	NCHCI
Pulldown PD1	NCHCI	NCHCI	first PBTI then NCHCI	NCHCI
Pulldown PD2	PBTI	PBTI	first NCHCI then PBTI	PBTI
Access PG1	NCHCI	n/a	first HCI then n/a	n/a
Access PG2	n/a	HCI	first HCI then PBTI	PBTI

Table 4.2: Overview of the four degradation mechanisms NBTI, PBTI, HCI and NCHCI on all possible SRAM modes. This table focuses on the '1' memory state. The pairs of pullups, pulldowns and access-transistors have switched their degradation mechanisms, compare Table 4.1. Write 1 to 1 means to write a 1 to the cell which is already in 1 state.

State	Most important effect
Hold	NBTI on '1'-side pullup PBTI on opposite pulldown NCHCI on the rest
Read	like 'Hold', but with HCI on '0'-side access transistor
Write	like 'Hold', but switching from one SRAM side to the other

Table 4.3: Summary of the most important degradation effects during the 3 states Hold, Read and Write

4.2 Negative Bias Temperature Instability

The voltage and current characteristics show that NBTI is active during all cycles hold, read and write. Especially the 'hold' cycle is important, because reading and writing only takes very short time in the ns-range. Hold, on the other hand, can be done for years! This section will explore the sensitivity of SRAM performance to NBTI-degraded pullup transistors. The pulldown- and access-transistors are all n-type and therefore do not show NBTI degradation in the SRAM modes of operation.

4.2.1 NBTI sensitivity

To simulate the NBTI degradation, the equivalent circuit in Fig. 4.15 (based on the findings of section 3.5.1 on page 28) is used. Fig. 4.16 shows the sensitivity simulation setup: Artificial threshold voltage drift ΔV_{th} is applied to one of both pMOS pullups. Then the 4 performance metrics SNM(read), SNM(hold), I(read) and WriteLevel (compare Section 2.2 starting on page 10) are simulated. This is done for max. $\Delta V_{th} = 100$ mV , because this is worst case for products: 10 years @ 125 °C and 110%V_{DD}=1.32 V.

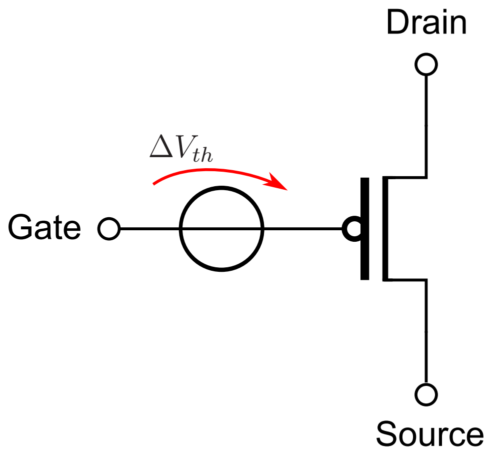


Fig. 4.15: Modeling NBTI on degraded pMOS transistors: shift in threshold voltage, $\Delta V_{th} > 0$

Fig. 4.17 shows the normalized impact on hold and read stability. 2 curves are plotted for each stability, because the result depends on whether the artificial degradation was applied to the '0' or '1' side of the memory cell, i.e. if PL1 or PL2 got degraded. Generally, NBTI degradation on the '1' side makes the cell less stable. This is because the conducting pullup has to be turned off to switch the cell, which is easier for a degraded transistor. In contrast, NBTI degradation on the '0' side of the cell means that the non-conducting pullup has to be turned on to switch the cell, which is harder for a degraded transistor. This is why NBTI degradation on the '0' side does not affect (or even improves) the stability of the cell. But since the cell must keep working for both directions, we focus on the bad case of losing stability. A V_{th} shift of -100 mV on one pMOS pullup due to NBTI result in approx. -7% hold stability and -10% read stability. So read stability shows more

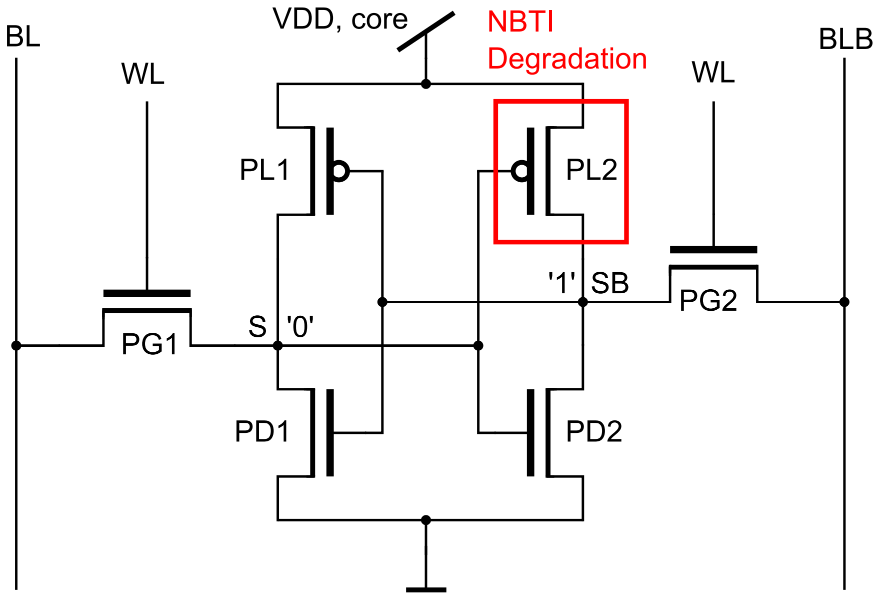


Fig. 4.16: Simulation setup of NBTI sensitivity: one pullup is artificially NBTI degraded by applying a threshold voltage shift like in Fig. 4.15. The other pullup is kept with nominal threshold voltage.

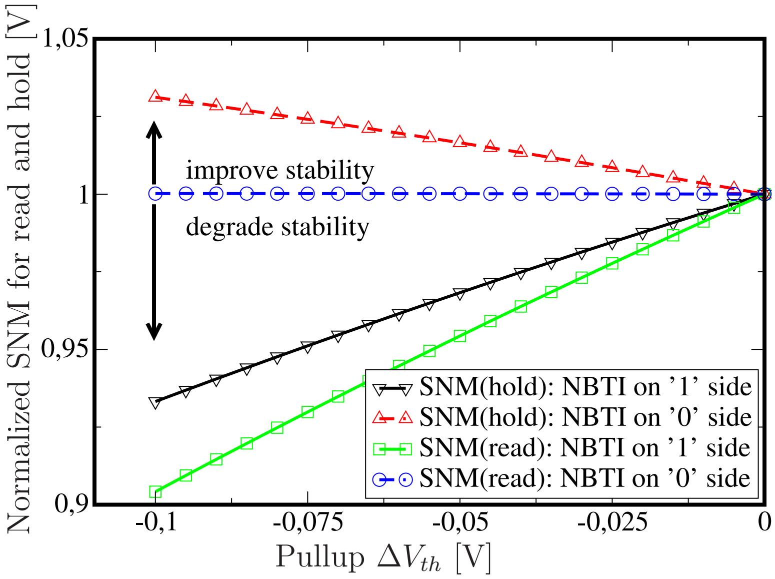


Fig. 4.17: Simulation of hold- and read-stability SNM: if drift is applied on the ’1’ memory side, SRAM is losing stability. 100 mV V_{th} drift result in approx. 10% read stability loss.

sensitivity to NBTI, but this is only true for nominal operation voltage.
Fig. 4.18 shows the impact on I_{read} and therefore the speed of the cell. NBTI degradation does not have any impact on read current. This is because read current only flows through access- and pulldown-transistors, the pullup does not matter. So the maximum speed of reading a cell is not influenced by NBTI.

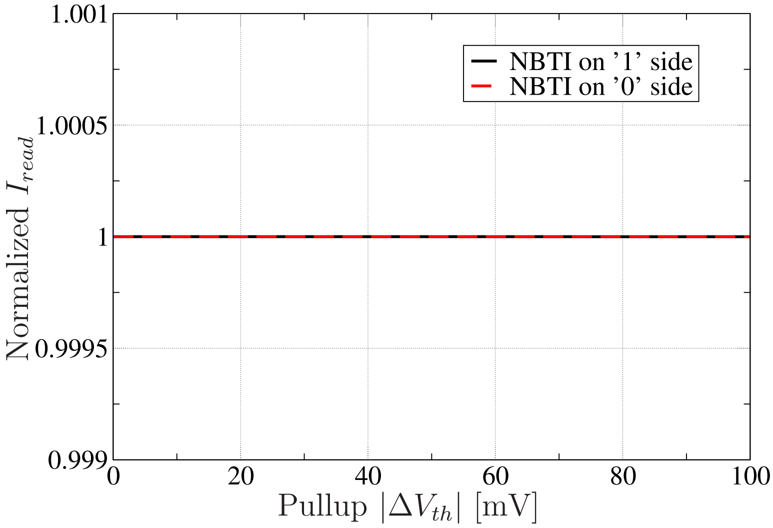


Fig. 4.18: Simulation of read current: NBTI does not affect this value.

Fig. 4.19 shows the impact on Write Level and therefore the writeability of the cell. Since writeability and stability behave oppositely, the already observed loss of stability makes the cell easier to write. This again depends on whether the degradation happened on the '0' or on the '1' side of the cell. 100 mV degradation on the '1' side make the cell approx. 11% better writeable, which represents less stability. Degradation on the '0' side does not impact the cell, like it did not impact read stability before.

4.2.2 NBTI conclusion

Altogether, the SRAM cell is quite sensitive to NBTI pullup degradation. During long hold times, the pullup on the '1' memory side degrades and reduces the stability of the cell, compare Table 4.4.

SNM(read)	SNM(hold)	I(read)	Write Level
-10%	-7%	n/a	+11%

Table 4.4: Sensitivity of SRAM performances for worst case of 100 mV NBTI-induced ΔV_{th} degradation. -100 mV refer to 10 years @125 °C and 1.32 V.

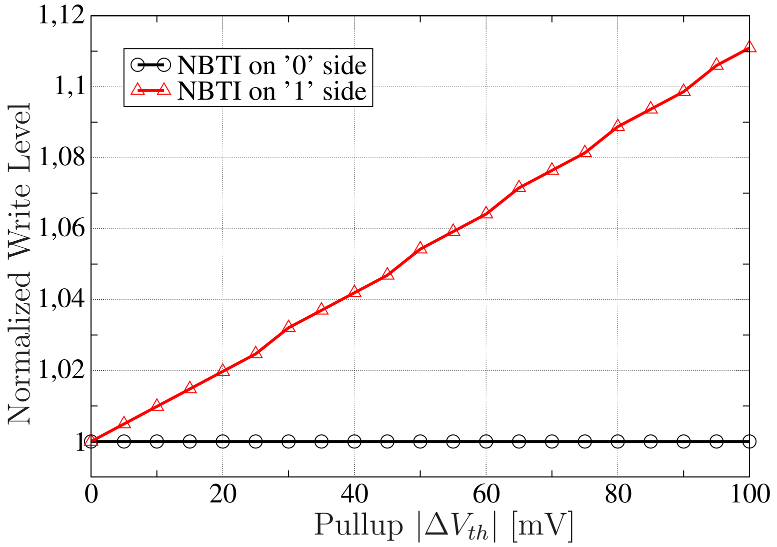


Fig. 4.19: Simulation of Write Level: if threshold voltage shift is applied on the '1' side of the memory cell, writeability can be improved.

4.3 Positive Bias Temperature Instability

The current and voltage characteristics show that PBTI is active during all cycles hold, read and write. This section will explore the sensitivity of SRAM performances to PBTI-degraded pulldown- and access-transistors. The pullup transistors are p-type and therefore do not show PBTI degradation in the SRAM modes of operation.

4.3.1 PBTI pulldown sensitivity

To simulate the PBTI degradation, the equivalent circuit in Fig. 4.20 is used. Fig. 4.21 shows the sensitivity simulation setup for pulldowns. Artificial threshold voltage drift ΔV_{th} is applied to one of both nMOS pulldowns. Then the 4 performance metrics SNM(read), SNM(hold), I(read) and WriteLevel are simulated. Again, ΔV_{th} is simulated for max. 100 mV under the assumption that PBTI will behave like NBTI in the future. Fig. 4.22 shows the impact of PBTI on pulldown transistor on read stability. 100 mV PBTI drift cause a read stability reduction of 20%. This is immense sensitivity! Fig. 4.23 shows the impact of PBTI on pulldown transistor on hold stability, which is less critical than read stability. Only 3% of hold stability reduction occurs on 100 mV PBTI drift. Fig. 4.24 shows the impact on I_{read}. Contrary to pullup NBTI degradation, PBTI on pulldown also impacts the speed of the cell. 9% speed is lost when 100 mV artificial PBTI drift are applied to one pulldown device. Fig. 4.25 shows the impact on Write Level. The writeability of the cell can be improved by 7% if the degradation occurs on the '1' side of the cell.

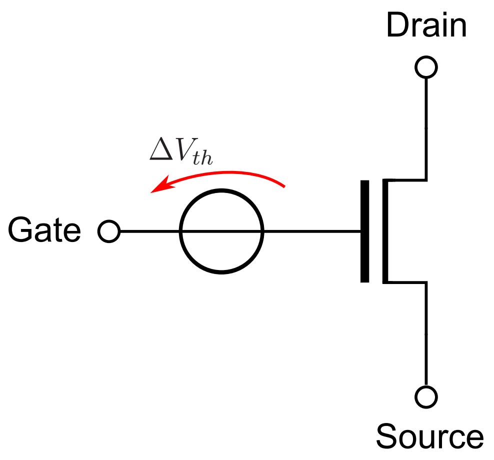


Fig. 4.20: Modeling PBTI on degraded nMOS transistors: shift in threshold voltage, $\Delta V_{th} > 0$

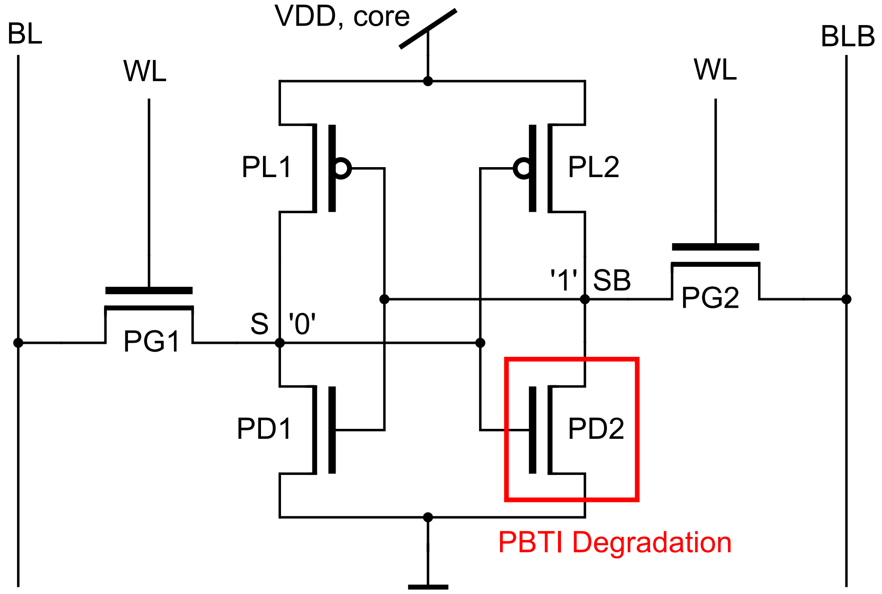


Fig. 4.21: Simulation setup of PBTI pulldown sensitivity: one pulldown is artificially PBTI degraded by applying a threshold voltage shift.

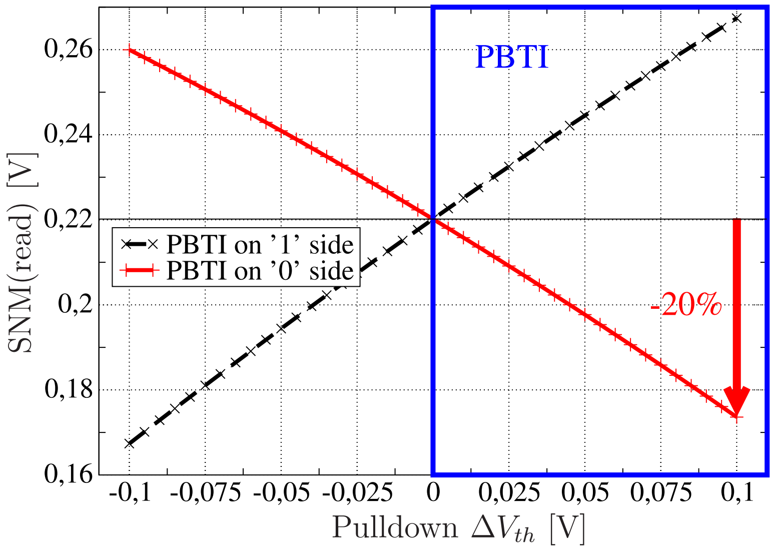


Fig. 4.22: Simulation of read stability SNM: if drift is applied on the '0' memory side, SRAM is losing 20% stability.

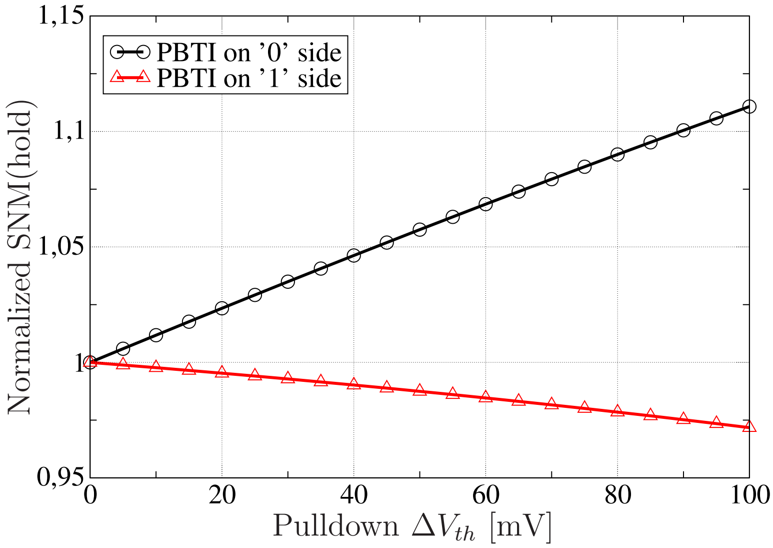


Fig. 4.23: Simulation of hold stability SNM: if drift is applied on the '0' memory side, SRAM is losing 3% stability.

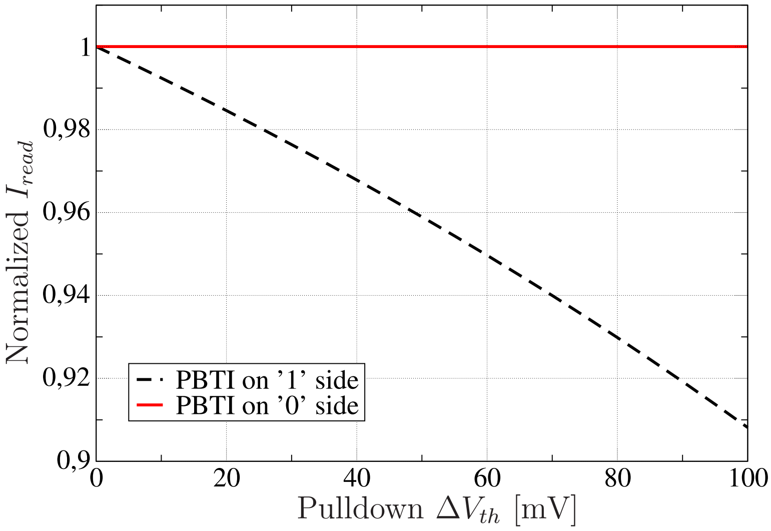


Fig. 4.24: Simulation of read current: if drift is applied on the '0' memory side, SRAM is losing 9% speed.

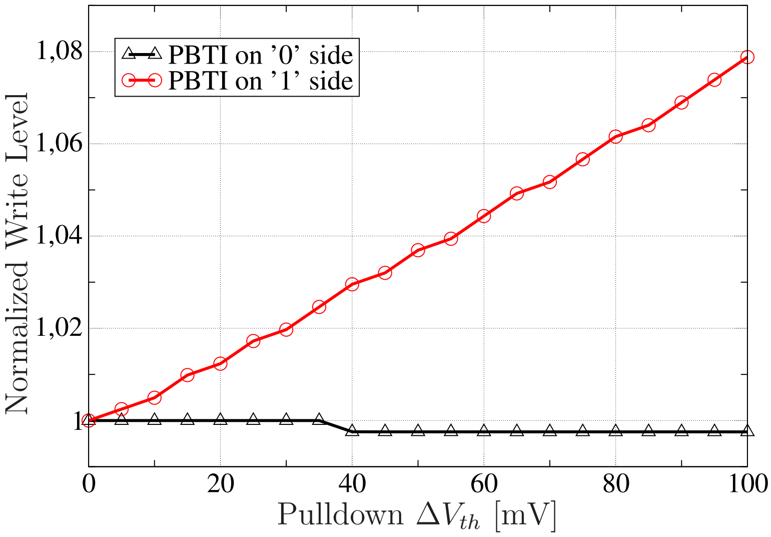


Fig. 4.25: Simulation of write level: if drift is applied on the '0' memory side, SRAM is gaining 7% writeability.

4.3.2 PBTI access sensitivity

Fig. 4.26 shows the sensitivity simulation setup for access transistors. Again, the equivalent circuit depicted in Fig. 4.20 is used. Fig. 4.27 shows the impact of PBTI on access transistor on read stability. Degradation of the access device improves read stability, because a weaker transistor allows less bitline influence on the memory node.

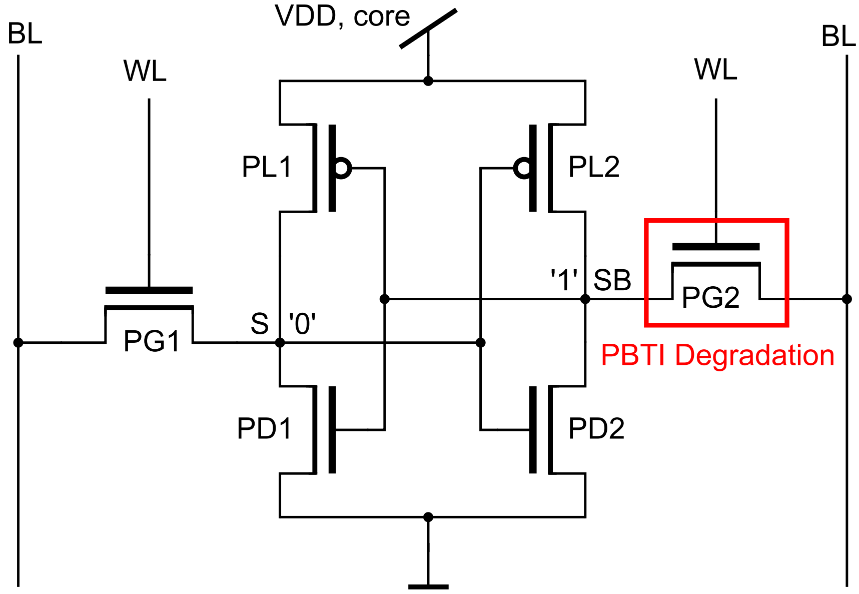


Fig. 4.26: Simulation setup of PBTI access sensitivity: one access is artificially PBTI degraded by applying a threshold voltage shift.

Hold stability is of course not affected by PBTI, because during ’hold’, the access transistor is switched off. This is why this simulation does not make sense here. Fig. 4.28 shows the impact on I_{read}. Almost 12% speed are lost with 100 mV artificial PBTI drift on the access transistor. Fig. 4.29 shows the impact on Write Level and therefore the writeability of the cell. 23% writeability are lost by 100 mV threshold voltage drift of the pulldown.

4.3.3 PBTI conclusion

Altogether, the SRAM cell is very sensitive to PBTI pullup- and access-degradation. An overview of the sensitivity is provided in Table 4.5.

100 mV degradation of the pulldown device results in 20% read stability degradation and 9% speed degradation. The stability is degraded twice as much as compared to pullup NBTI degradation.

100 mV degradation of the access device results in 23% degradation of speed and writeability, respectively.

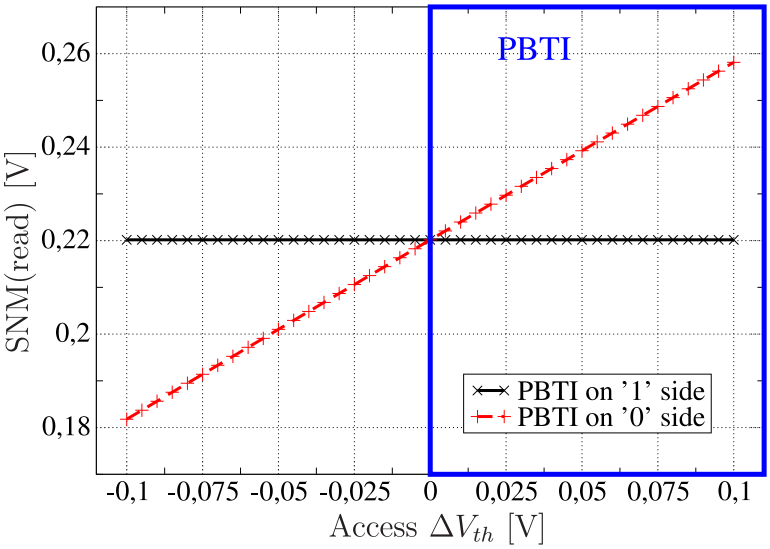


Fig. 4.27: Simulation of read stability SNM: if drift is applied on the '0' memory side, SRAM is gaining 18% stability.

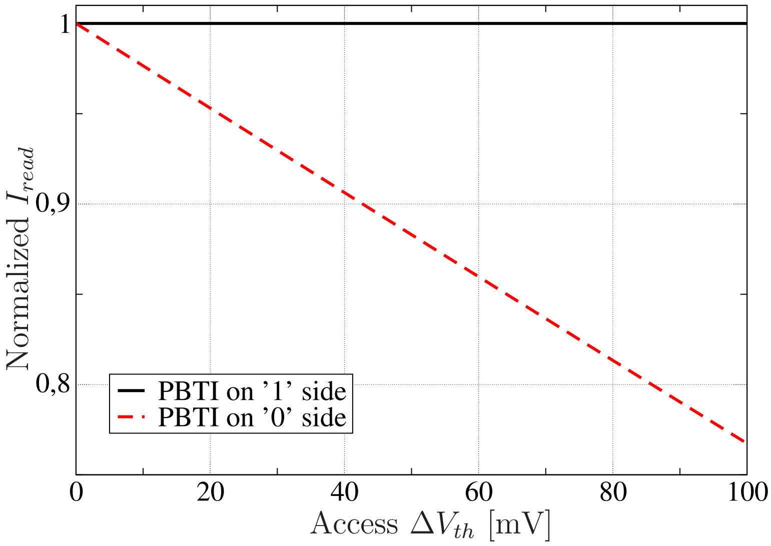


Fig. 4.28: Simulation of read current: if drift is applied on the '0' memory side, SRAM is losing 23% speed.

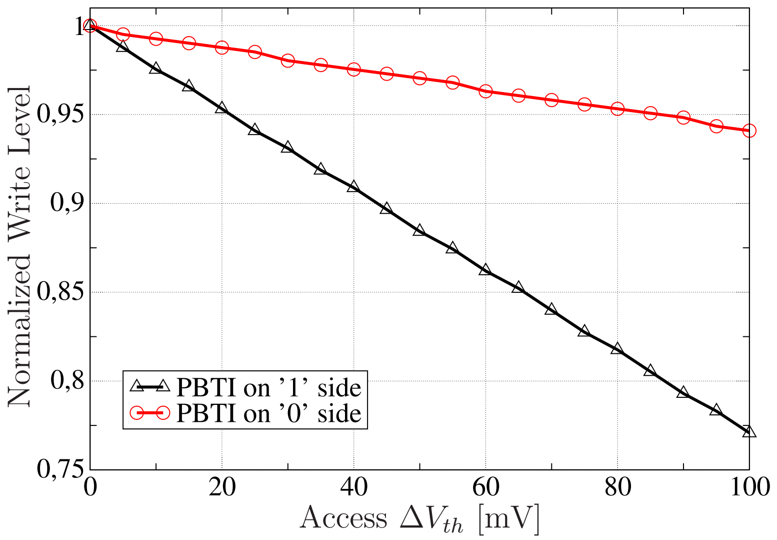


Fig. 4.29: Simulation of Write Level: if drift is applied on the '1' memory side, SRAM is losing 23% writeability.

	SNM(read)	SNM(hold)	I(read)	Write Level
pulldown	-20%	-3%	-9%	+8%
access	0%	n/a	-23%	-23%

Table 4.5: Sensitivity of SRAM performances for worst case of 100 mV PBTI-induced ΔV_{th} degradation of pulldown- and access transistor. 100 mV refer to 10 years @125 °C and 1.32 V, if PBTI is assumed to follow the same estimation as NBTI.

4.4 Hot Carrier Injection

HCI is a strong candidate for transistors with short gate lengths. Contrary to analog circuits with longer than minimum transistor dimensions, short gate lengths are realized in SRAM core cells. This is why this degradation effect might occur here.
The V/I characteristics of each transistor show that only the pullup- and access-transistors experience conditions in the Hot Carrier Injection region during normal lifetime operations. Therefore in this section, the sensitivity of the SRAM cell on HCI-degraded pullup- and access-transistors is examined.

4.4.1 HCI of pullup transistor

The worst HCI point in the pullup curves is during the write cycle in the region of V_{gs}=0.8 V, compare Figs. 4.8 and 4.9 on page 48. The current is maximum for this gate-source voltage and therefore, it is close to the HCI region in the output characteristics. If this voltage was applied for 10 years, it would result in 1% I_{on} degradation (using formula 3.7.7 on page 36). But this voltage is applied only very short-time (less than 1 ns) during each write cycle. If 10% of the complete switching time was assumed to be in this HCI region and the cell is re-written all the time like in a ring-osciallator, then there was a 0.5% I_{on} degradation, which is too small to be regarded here.

4.4.2 HCI sensitivity of access transistor

The only real HCI candidate therefore is the access transistor, because during read and during write, one or two access transistors are conducting in the saturation region with V_{ds} close to V_{DD} and appreciable I_d flowing, compare c and 4.13 on page 50. The sensitivity of the access transistor to HCI is reduced slightly because in 65 nm technology, it is not designed as a minimum length device: it has 70 nm instead of 65 nm. To simulate the impact of a HCI degraded access-transistor, the setup in Fig. 4.30 is used. To simulate the HCI degradation on one transistor, the equivalent circuit in Fig. 4.31 is used. It divides the I_{on} degradation, which is the output of the HCI simulator formula 3.7.7, into a threshold voltage drift and a drain current degradation. This is done to account for the V_{gs} dependency of the Hot Carrier effect: observed I_D degradation as a function of V_{GS} increases as V_{GS} decreases to V_{th}.
The 3 performance parameters Write Level, Iread and SNMread were now simulated as a function of the I_{on} current degradation, which was determined to be in the 10% range. SNM(hold) must not be determined, since the access transistor is not active during hold and therefore has no impact at all.
The results of the simulations are: Write Level is decreased as much as the Ion is decreased at the '1' sided access transistor, compare Fig. 4.32. Speed is reduced as much as the Ion is reduced at the '0' sided access transistor, compare Fig. 4.33. Stabiliy increases as much as the Ion is degraded at the '0' sided access transistor, compare Fig. 4.34.

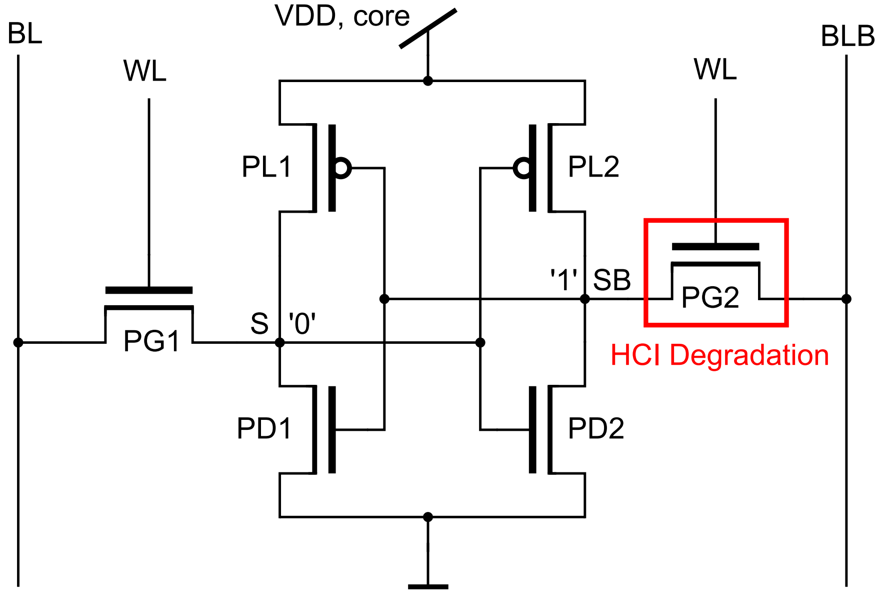


Fig. 4.30: Simulation setup for HCI access sensitivity. One access transistor is artificially HCI degraded by a mixture of threshold voltage- and Ion-drift

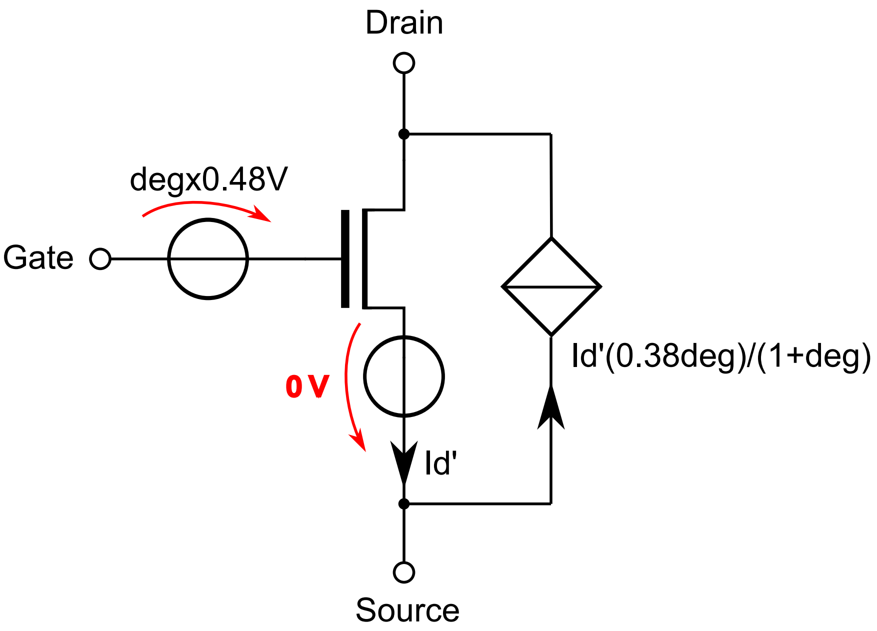


Fig. 4.31: Modeling a HCI degraded transistor: the degradation factor 'deg' is divided into a threshold voltage shift and a drain current degradation to account for the V_{gs} dependency of the Hot Carrier effect. The 'deg' value is the HCI degradation in the following plots.

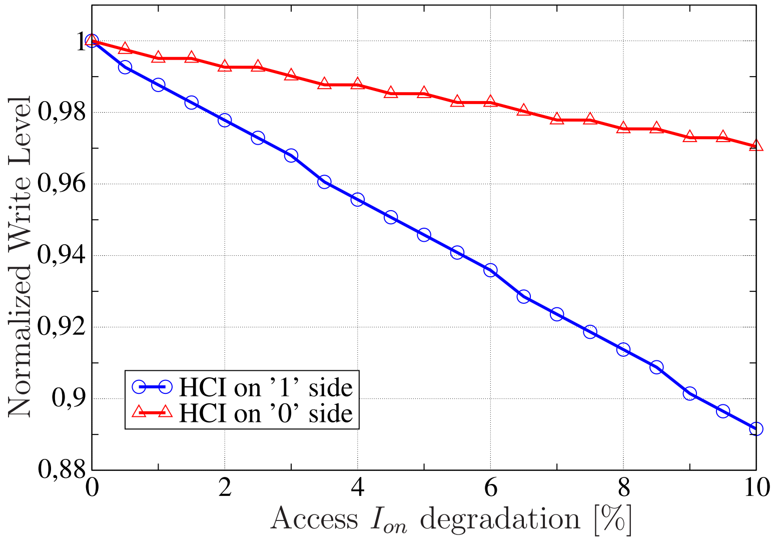


Fig. 4.32: Simulation of Write Level: if HCI degradation is applied to the '1' memory side, SRAM is losing 11% writeability.

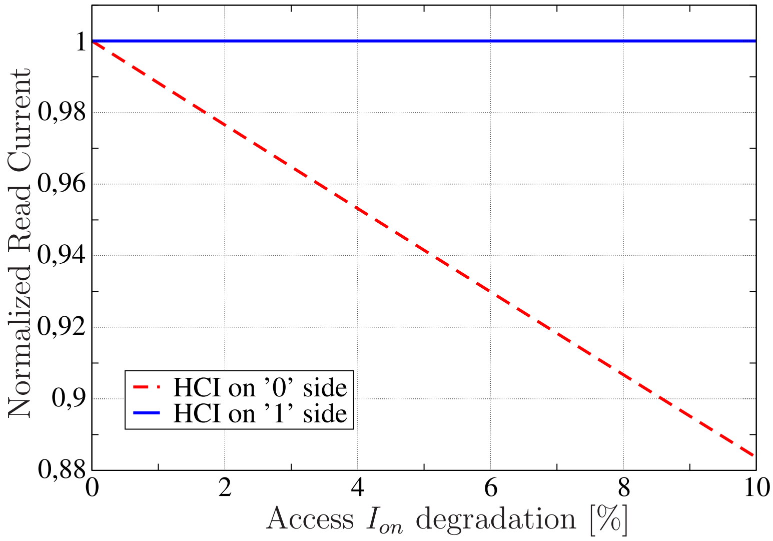


Fig. 4.33: Simulation of Iread: if HCI degradation is applied to the '0' memory side, SRAM is losing 12% speed.

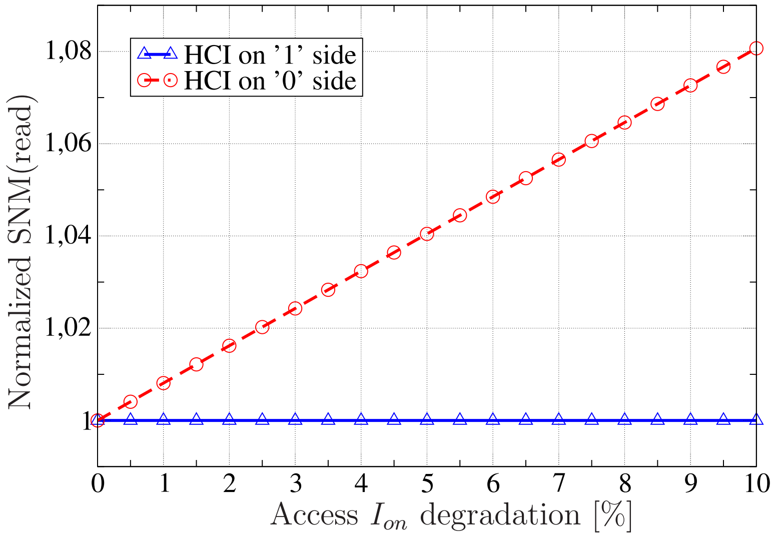


Fig. 4.34: Simulation of SNM(read): if HCI degradation is applied to the '0' memory side, SRAM is gaining 8% stability.

As a rule of thumb, the 3 performance parameters vary about the same fraction as the HCI degradation. This means, when access transistor is degraded by N % Id, ...
Write level is decreased about N %
I_{read} is decreased about N %
SNM(read) is increased about N % (but remains constant for the worse side).
Referred to the worst case of 10% Ion degradation, the 6T-SRAM circuit is remarkably sensitive to a HCI degraded access transistor.

SNM(read)	SNM(hold)	I(read)	Write Level
0%	n/a	-12%	-11%

Table 4.6: Sensitivity of SRAM performances for -10% HCI degradation on access transistor

But the HCI-active region in the output characteristic is only active for a very short time in the sub-ns region. So the question is: how big is the ΔI_{on} degradation caused by regular SRAM operation?
To answer this question, 3 different estimation scenarios were assumed.

1. Artificially keep the cell in static read state, which is not realistic, but can be done under test conditions. The '0' memory side is getting degraded.

2. Artificially keep switching the cell between its two states like in a ring oscillator. Then the '0' as well as the '1' side is getting degraded.

3. Estimate real operating conditions based on the assumption of a 1 MBit array that uses all cells equally often.

4.4.3 Static reading for 10 years

This assumption is not a realistic case, but provides a worst case scenario. Formula 3.7.7 can be used to simulate a static read case for 10 years. For t=87600 hours (10 years), V_{ds}=1.083 V (the '0' memory node has approx. 0.1 V during read state) and l=70 nm, the ΔI_{on} degradation adds up to 0.43%. Even if the max. allowed voltage of 1.32 V is assumed, V_{ds}=1.181 V and the ΔI_{on} degradation only adds up to 2.14%.

4.4.4 Continuous switching for 10 years

This is the second worst case scenario, which does not represent a realistic case. Hot carrier effects occur in nFET devices when V_{ds} is close to V_{DD} and appreciable Id is flowing. For typical CMOS circuits, these conditions occur only during switching transients. In this case, equivalent stress time t(eq) can be approximated by the formula

$$t_{eq} = ITR \cdot F(TTR) \cdot SF \cdot t_{use} \tag{4.4.1}$$

with
ITR = ratio of gate voltage rise time (10-90%) to total cycle time
TTR = ratio of drain voltage fall time (10-90%) to gate voltage rise time (10-90%)
SF = switching factor (the fraction of cycles in which the NFET switches on; rising input, falling output)
t_{use} = actual use time in hours.

In this case,
input rise time = 100 ps when WL is raised.
output fall time = 200 ps when memory state changes
cycle time = 1 ns
switching factor = 1/4
V_{DD}=1.2 V, L=0.07 µm, ITR=0.1, TTR=2.

$$F(TTR) = A \cdot (\frac{TTR^{m1}}{1 + B * TTR^{m2}})^{n} \tag{4.4.2}$$

t_{eq} therefore is 657 h and the formula gives 0.36% for 10 years of switching and only 1.78% for 1.32 V and 10 years.
Both values are close to the static reading case, and again both are too low to be considered within this work.

4.4.5 Real operating conditions

After two worst case scenarios to determine the upper limit of HCI impact, this third use case is oriented on real operating conditions. After 10 years of operating lifetime with typically 128 wordlines per SRAM macro, each wordline was switched on for about 0.1 year. Two memory states exist, but there is only current flowing in the '0' memory

state, which divides the on-time by 2. The result is 0.05 years of current flowing in the cell, which is about 18 days. The typical user profile of a mobile phone nowadays is 8% talking, 8% listening to music, 4% writing SMS. During all these times, the processor voltage is reduced to 1 V, because there are no challenges for the processor and leakage can be reduced. Only 9% of the lifetime is used for videobrowsing or other computation-intensive tasks, where V_{DD} is increased by 10%. This means that in 10 years of lifetime, a mobile phone is HCI-degraded only for about 10% of the time, which is 1.8 days (approx. 40 hours). Only during 50% of this time, the wordline is really high, which means a total time of 20 hours in 10 years of lifetime.
Using the HCI formula 3.7.7, this translates to less than 0.1 % Ion degradation. This is totally neglectable.

4.4.6 HCI Conclusion

Estimating the HCI degradation with 3 different scenarios resulted in different Ion degradation values. The unrealistic worst-case scenarios for endless reading or endless switching showed maximum HCI Ion degradations of 2%. Realistic estimations for SRAM-like usage showed HCI degradation <0.1%. So although minimum gate lengths are used for SRAM transistors, which is a precondition for HCI problems, the times of switching are so short, that only non-realistic all-time switching scenarios lead to remarkable performance degradation.
This is why HCI in the end is not regarded a problem for SRAM compared to the much higher impact mechanisms NBTI and PBTI.

4.5 Off-State Stress

Three facts suggest that 6T-SRAM core cells might be a candidate for strong NCHCI degradation. First, V-I analyses have shown that some devices of the core cell are very long time under NCHCI conditions. During long hold of data, one pullup, one pulldown and one access transistor are suffering NCHCI stress. This behavior is contrary to HCI, which only shows up on one access transistor very shortly during read access.
Second, NCHCI is increasing with short channel lengths, which is true for SRAM cells. While the transistors are too large in analog applications, the tiny SRAM transistors are small enough to fulfill this condition.
Third, the SRAM circuit is sensitive to HCI-degraded transistors (and therefore also to NCHCI-degraded transistors), which was found in section 4.4 on page 63.
However, the impact on the current 65 nm technology is almost negligible. In section 3.7.2, it was found that p-channel transistors do not show any off-state stress degradation in this technology node. nMOS transistors at least provide a formula, but which is only valid for burn-in conditions. Therefore, this is only a candidate for smaller future technologies and is neglected within the focus of this work. Only in future, with even smaller gate lengths, the impact might increase so much that it will affect the performance of the cell.

4.6 NBTI plus PBTI

Examining each degradation mechanism alone has shown that in normal life, only NBTI and PBTI have serious impact on the performance of SRAM [43]. It is therefore important to check how both degradation mechanisms together impact SRAM. Long hold of one memory state degrades both the pullup on the '1' side by NBTI and the pulldown on the '0' side by PBTI. The problem then is a completely asymmetrical cell. These simulation results are made under the assumption that the absolute threshold voltage drift for NBTI and PBTI is equal. The sensitivity simulation is again done like for NBTI and PBTI alone, by artificially applying a threshold voltage drift to both pullup and pulldown transistors. Fig. 4.36 shows how NBTI and PBTI are adding to huge impact on read stability. For -100 mV NBTI and 100 mV PBTI degradation, read stability is reduced by almost 30%! Fig. 4.37 shows how NBTI and PBTI together have strong impact on hold stability, and Fig. 4.38 shows how NBTI and PBTI together have impact on read current. Fig. 4.39 shows how NBTI and PBTI together have impact on Write Level.

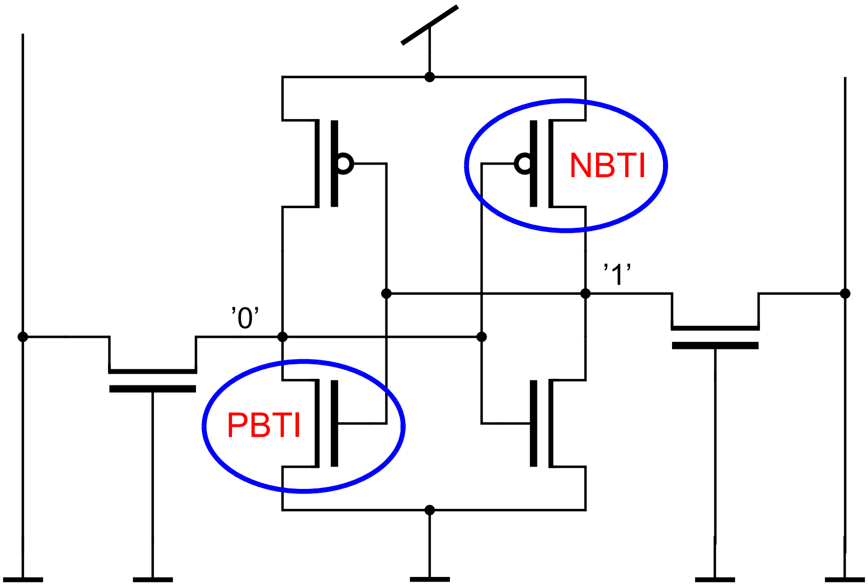


Fig. 4.35: Simulation setup for NBTI+PBTI sensitivity. One pullup plus one opposite pulldown are artificially BTI degraded by applying a threshold voltage drift.

Fig. 4.40 shows the impact of NBTI plus PBTI on the yield of a 256k SRAM array. The criterion is one failing cell. These calculations were done with the 'Worst Case Distance' algorithm [44], which gives the same result as Monte Carlo simulations, but with less computing effort. Fig. 4.41 shows the yield of a single core cell after the impact of NBTI plus PBTI.

Altogether, the pulldown is the most critical device in terms of degradation. It shows strong PBTI V_{th} shift and is very sensitive to SRAM performances. The pullup is the second critical device. It also shows strong NBTI V_{th} shift, but SRAM performances only show half the sensitivity compared to the pullup degradation. The access transistor is the least critical device. SRAM performances show strong sensitivity to its degradation,

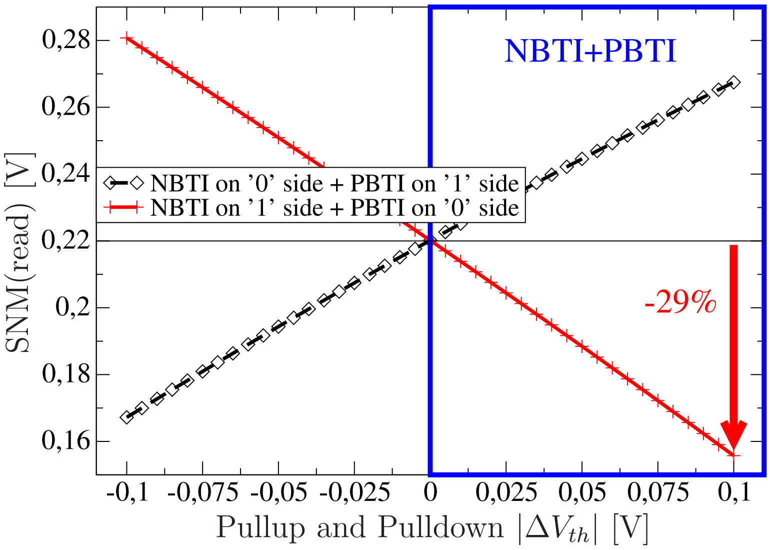


Fig. 4.36: Simulation of read stability for NBTI plus PBTI. Both effects are adding to 30% read stability loss (breite bbox noch entfernen)

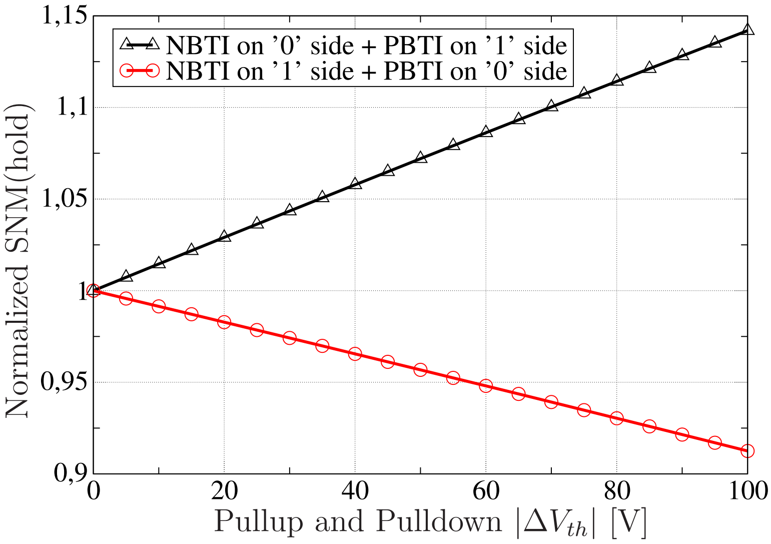


Fig. 4.37: Simulation of hold stability for NBTI plus PBTI.

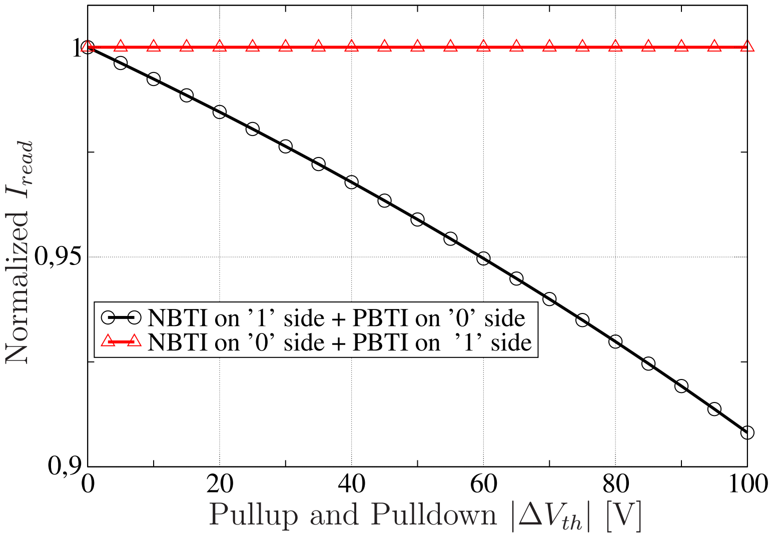


Fig. 4.38: Simulation of read current for NBTI plus PBTI.

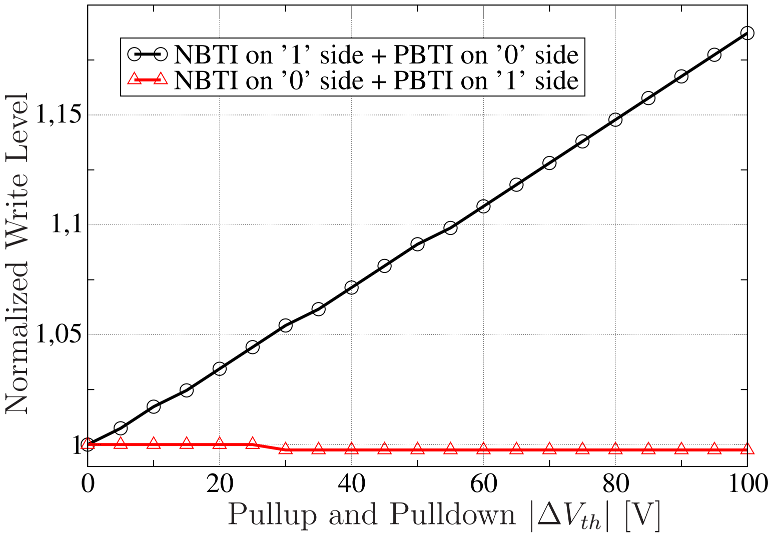


Fig. 4.39: Simulation of Write Level for NBTI plus PBTI.

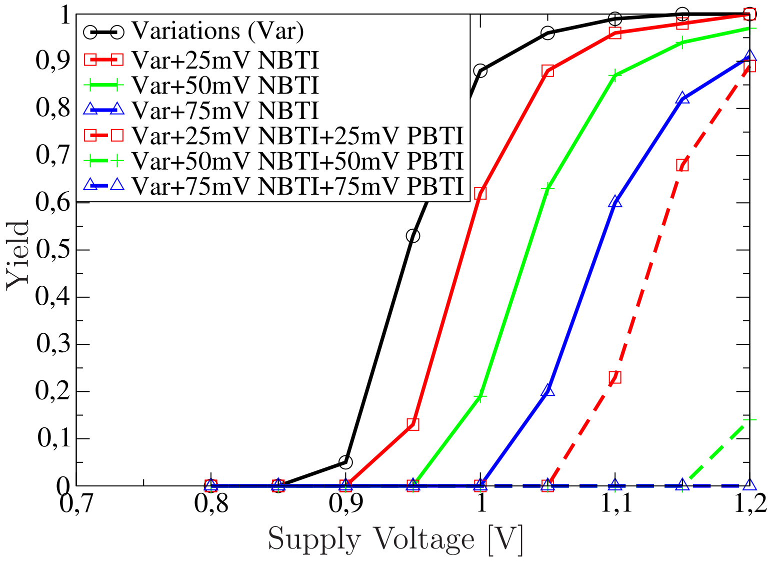


Fig. 4.40: Yield including global and local variations after long hold of one state with NBTI plus PBTI degradation for a 256 kBit cell array. If at least one cell in the complete array is failing, the whole array is considered as failing.

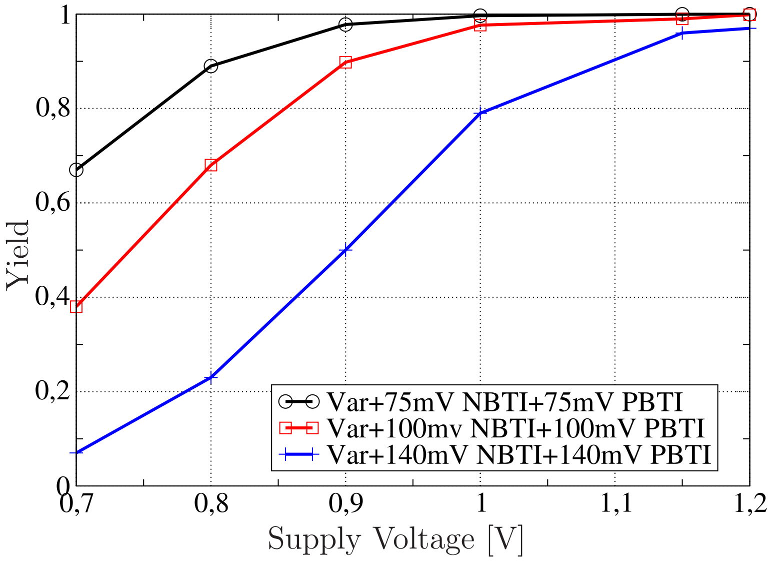


Fig. 4.41: Yield including global and local variations after long hold of one state with NBTI and PBTI degradation for a single bit cell.

but it is hardly PBTI stressed and therefore shows only litte V_{th} shift.
The ranking of degradation mechanisms clearly puts NBTI on the less critical position. Realistic worst case scenarios show a maximum SRAM performance degradation of approx. 10% on standard metrics. Compared to that, PBTI is much more critical for SRAM core cells. If the V_{th} shift is assumed to be in the same region like pMOS V_{th} shift, the performance degradaton with approx. 21% is about double. Regarding literature, this might be the case in the future. Combination of NBTI and PBTI on hold state for DF=1 is adding the problems: 29% SNM(read) degradation for -100 mV NBTI and +100 mV PBTI shift.

4.7 Conclusion

After examining the operating conditions of each transistor in the 6T-SRAM cell, the impact of the 4 existing degradation mechanisms can be summarized, including the findings of chapter 3.

- HCI does not affect the cell in normal use conditions because the time it is active is too short
- NC-HCI does not affect the cell because the degradation by off-state stress is too small
- PBTI does not affect the cell because in the used 65 nm technology in this work, this degradation mechanism almost does not exist. But in future technology nodes, it was shown that it will have tremendous impact.
- NBTI is the only degradation mechanism that exists for the used technology. It is most severe after long 'hold' states, and impacts the stability of the memory cell.

Table 4.7 shows the overview on the results.

Degradation Effect	Triggered	cell sensitivity	degradation in 65nm	Impact
NBTI	yes	yes	yes	HIGH
PBTI	yes	yes	no	LOW
HCI	short	yes	yes	LOW
NCHCI	yes	yes	no	LOW

Table 4.7: Overview of the impact of the 4 parametrical degradation mechanisms on 6T-SRAM

Therefore, in the rest of this work the focus will be on NBTI.
Simulations now answered many questions about the impact of parametric degradation mechanisms on 6T-SRAM core cells. But there are three reasons why measurements now need to be done additional to the simulations.

1. Degradation is adding to variability. Only the sum of both, where variability is worse than degradation, are critical for a circuit.

2. Degradation itself has tremendous variability. Especially in SRAM arrays with thousands or even millions of cells, this variability is a big concern.

3. NBTI shows strong and fast recovery behavior. The impact of this must be evaluated.

All these aspects are covered in chapters 5 and 6 of this work.

Chapter 5

Stability Analysis of SRAM Arrays

Abstract

Chapter 4 has proved that under real working conditions, only NBTI affects 65 nm 6T-SRAM core cells and degrades their stability. This chapter focuses on measuring this effect directly on the stability of SRAM arrays. While stability is often only simulated like before, a measurement approach is developed to characterize large-scale SRAM arrays very fast. This is necessary because of variability of the manufacturing process as well as variability of NBTI degradation.
The most important results of this chapters were reported in a joint session of ESSCIRC and ESSDERC and are published in:

S. Drapatz, T. Fischer, K. Hofmann, E. Amirante, P. Huber, M. Ostermayr, G. Georgakos, and D. Schmitt-Landsiedel:,"Fast stability analysis of large-scale SRAM arrays and the impact of NBTI degradation",

European Solid-State Circuits Conference ESSCIRC, 2009 and

European Solid-State Device Research Conference ESSDERC, 2009

5.1 State of the Art: SRAM Stability Analysis

As stated in chapter 2, all currently available Figures of Merit used for stability characterization are not well suited for in-field product measurements:
A generally accepted definition of memory cell stability was introduced with the Static Noise Margin (SNM) [12]. The smaller this metric, the easier the cell is flippable during read access. Unfortunately, this metric is hardly accessible to experimental array analysis and mostly suited for simulation, because it is given by the maximum eye opening of the 'butterfly' curve, which consists of two overlaid inverter curves. It is therefore difficult to measure with the need of dedicated test structures with access to the memory nodes and only of practical use for simulation, compare section 2.2.1.

Another stability criterion is the Read Margin (RM), which is detected by a bitline current drop with varying core operating voltage, compare section 2.2.1. This metric can be used much better for physical analysis and correlates to SNM [16] [15]. However, a dedicated test structure with multiplexed bitlines to external pads is also necessary, and analysis of a complete SRAM array takes long time, because V-I curves of each cell have to be measured individually.

One reason why the parameter V_{min} is so popular is because it can be measured without dedicated test structures, compare section 2.2.5. But it is not a FoM that can be used for stability analysis, it represents a voltage at which the complete SRAM circuit including periphery still works for hold, read and write.

With state-of-the-art analysis methods, a quick stability analysis of large-scale SRAM arrays is therefore not given. So a stability analysis method is needed which ...

1. ... does not require access to the memory nodes, because this would change the product core cell
2. ... does not require highly precise and slow measurement of V-I curves.

5.2 New Approach

In this work, we propose a new and fast method to analyze the stability of large SRAM arrays using the RM criterion without a dedicated test structure and without measurement of VI-curves. The only requirement is a two-rail V_{DD} design for core and periphery, then this method can be used even in-field on product chips. This technique is then applied to characterize the SRAM cell stability distribution pre- and post-NBTI-stress.

Simulations in Chapter 4 have shown that $-100\,\mathrm{mV}$ ΔV_{th} reduces the SNM(read) by approx. 10%. The effect of such a shift should be clearly visible in RM distribution.

All stability considerations mentioned so far were for the read state, which is the most critical state in 6T SRAM circuits at constant V_{DD}. But in nowadays' products, V_{DD} is typically reduced during 'hold' to keep leakage current low and decrease power dissipation. This so-called 'retention mode' is also critical with respect to stability. SNM is also defined for the hold case, but has the same disadvantages as SNM(read). RM measurement via bitline current does not work, because the crucial subthreshold currents are too small. The method presented in this work allows to analyze the hold stability as well.

The next section 5.3 will examine if Read Margin is a linear measure to characterize stability. In chapter 5.4, the required hardware as well as the measurement technique for read- and hold-stability are described. Also the NBTI stress conditions are explained there. Measurement results as well as stability distributions pre- and post-NBTI stress are shown in section 5.5. The chapter is finished with the conclusions in the last section.

5.3 Simulation of Read Margin

First, RM is simulated in a 65 nm process. Therefore, the cell is put to read conditions (WL=BL=BLB=1) and then the core voltage ($V_{DD,core}$ in Fig. 5.1) is decreased. The read current on the '0' memory side (through access transistor PG1) is measured. For about 50% of the cells, there will be a critical voltage $V_{DD,core}$ where the read current drops, compare section 2.2.1.

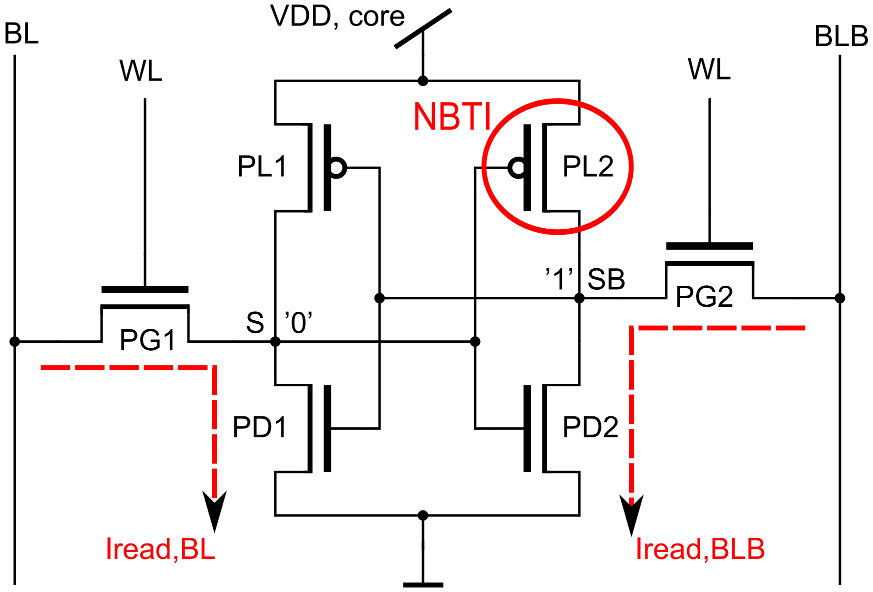


Fig. 5.1: Single 6T SRAM cell used in the Universal SRAM array. The read currents $I_{read,BL}$ and $I_{read,BLB}$ can be measured individually in each cell. This cell is in state '0', and pullup transistor PL2 is degraded during NBTI stress.

This happens when the manufacturing variations of the cell make it tend to the '1' cell state. So when the cell stability is artificially reduced (here by reducing the core voltage), the cell will flip to its preferred state. Since statistically 50% of all cells tend to the '0' side, these cells will not flip during core voltage reduction. This is why this simulation method only works for Monte Carlo simulations. Otherwise, a perfectly symmetrical cell would not have a preferred state and therefore show no flipping behavior.
10^4 Monte Carlo simulations with the measured V_{th} process variations (1 sigma of pullup W/L=90/65: 48 mV, 1 sigma of pulldown W/L=215/65: 35 mV, 1 sigma of pullup W/L=15/70: 45 mV) give the results in table 5.1 and the distribution in Figure 5.2. Without a pullup V_{th} shift, 5005 of 10000 cells showed the current drop like described in section 2.2 and were 'successful', the rest of 4995 did not show this current drop and therefore 'failed' (first column in Table 5.1). In the group of the 'succesful' cells, the mean value of the reduced core voltage when the current drop could be observed was at 852 mV, the minimum core voltage value at 737 mV and the maximum core voltage value at 1000 mV. If these varying core voltage values are plotted in a histogram, they result in the 'no V_{th} shift' curve of Fig. 5.2. Changing the V_{th} shift of the pullup device because of NBTI degradation causes the changes in the cell behavior like depicted in Table 5.1 and Fig. 5.2. So the simulations show that N mV V_{th} shift of the '1' pullup show about N/3 mV in RM shift. The standard deviation is not changed in a considerable way.

It is now important to find out if the RM criterion is a linear measure for cell stability. First, it is strongly correlated with SNM(read) [16]. Lowering the core voltage from 1.2 V

ΔV_{th}	successful/failed	min	max	mean	Δ mean	σ
w/o	5005/4995	737mV	1.000V	852 mV	w/o	34.69mV
20mV	5152/4848	741mV	1.017V	857 mV	5 mV	35.33mV
40mV	5231/4769	754mV	1.1018V	865 mV	13 mV	35.52mV
60mV	5403/4597	753mV	1.1027V	870 mV	18mV	35.10mV
80mV	5461/4539	765mV	1.1031V	876 mV	24mV	35.49mV
100mV	5635/4365	762mV	1.1048V	883 mV	31mV	35.50mV

Table 5.1: The RM simulation results after 10^4 MC simulations. N mV V_{th} shift on pullup cause about N/3 mV mean value shift in distribution. The standard deviation is not affected considerably.

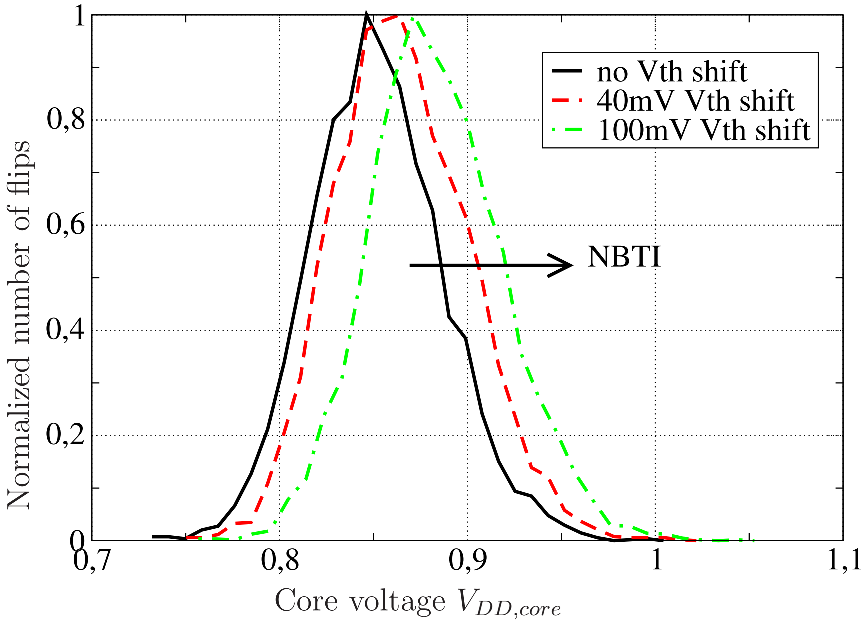


Fig. 5.2: RM distribution of 10^4 MC simulations with no ΔV_{th}, 40 mV ΔV_{th} and 100 mV ΔV_{th} on the '1' side pMOS pullup. The distribution gets shifted to the right, without a considerable standard deviation change.

to 0 V lets all cells which tend to the '1' side flip. Applying an artificial V_{th} shift to one pMOS pullup, i.e. make it NBTI-degraded, lets the number of flips increase linearly to the applied V_{th} shift. Fig. 5.3 shows the result of the 10^4 MC simulations described before with varied artificial pMOS V_{th} shift between 0 mV and -100 mV. The 'successful' count of Table 5.1 was depicted in Fig. 5.3. It shows that the number of flipping bits increases linearly with the V_{th} shift. Since the V_{th} shift has linear dependence on SNM (compare simulations in chapter 4.2), this means that lowering the core voltage to 0 V and counting the flipped cells is a linear Figure of Merit to measure cell stability.

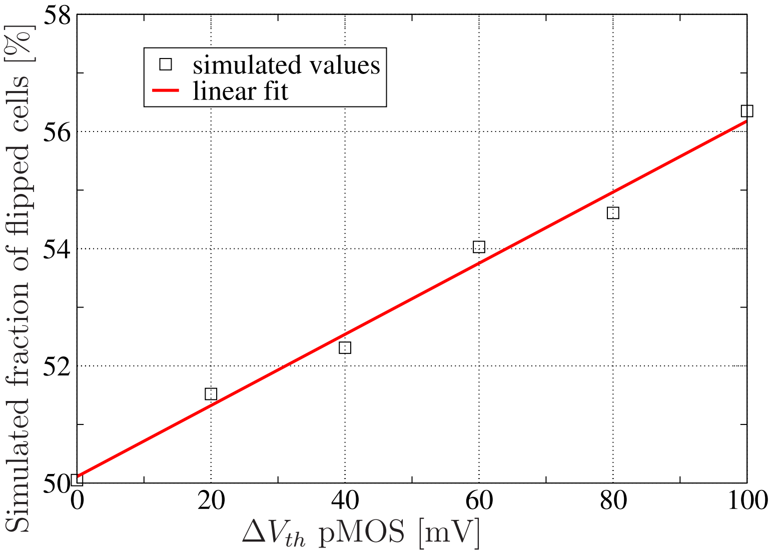


Fig. 5.3: Simulations show that pMOS pullup ΔV_{th} and number of flipping bits show linear behavior. For 40 mV ΔV_{th}, one can expect about 2.5% more flips.

If the core voltage is not reduced down to 0 V, but kept at some higher voltage (in this example 880 mV), then not all cells which tend to the 1 side flip, only the cells which are unstable enough to lose their state at a reduced core voltage of 880 mV. But 10^4 MC simulations show that also this is a linear measure for cell stability, see Fig. 5.4.

Hold Margin (i.e. Read Margin under hold conditions) cannot be simulated in this way. While in RM simulation, the current drop of the read current is detected, this read current does not exist during hold conditions. The modeling of subthreshold current is not precise enough to accurately describe switching in the subthreshold region.

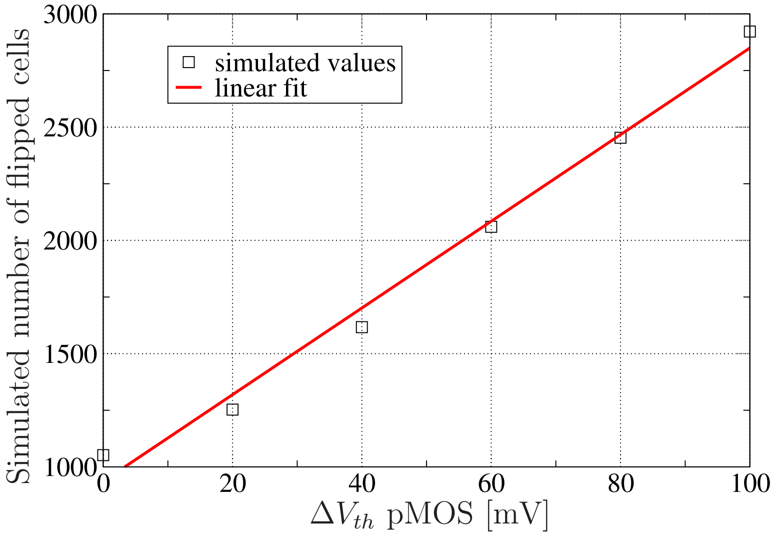


Fig. 5.4: Monte Carlo simulations indicate that pMOS pullup DeltaVth and number of flipping bits show linear behavior when supply voltage is lowered to a certain value, here 0.88 V. This shows that the number of flipping cells scales linearly with the V_{th} shift.

5.4 Experimental setup

5.4.1 Universal SRAM Testchip

All measurements in this work are based on a 1 Mbit SRAM Array test structure in a 65 nm low power technology [45] with 2.816 cells per bitline and 384 cells per wordline (Fig. 5.5). The SRAM array is completely fabricated like for a real SRAM product, only the periphery was adapted. Shift registers on wordlines (Fig. 5.6) and bitlines (Fig. 5.7) let the user select single cells in the array without a large number of address pins, and multiplexed bitline-pairs to 2 external pins on the chip (Fig. 5.8) allow read current measurement (Fig. 5.1) [29].

Selecting one bitline-pair to connect to the external pins causes all other bitlines being clamped to 1.2 V, compare Fig. 5.8. This structure enables to analyze the RM distribution of the memory array like in [16] [15] by lowering $V_{DD,core}$ and detect the drop of the bitline current. But in our approach we only want to detect the flipping of a cell, and BL current was only used to detect the cell state (Fig. 5.1): $>$ 1 µA read current means state '0' (typically 35-40 µA), $<$ 1 µA means state '1' (typically 0.1 µA). This test could of course be performed much faster in an ordinary SRAM array, as the bitline current measurement is no prerequisite to this approach. The only prerequisites are separate V_{DD} rails for the core array and for the periphery. Reduced core voltage is necessary for the noise margin analysis (section 5.4.2), enlarged core voltage is necessary for NBTI stress acceleration (section 5.4.3).

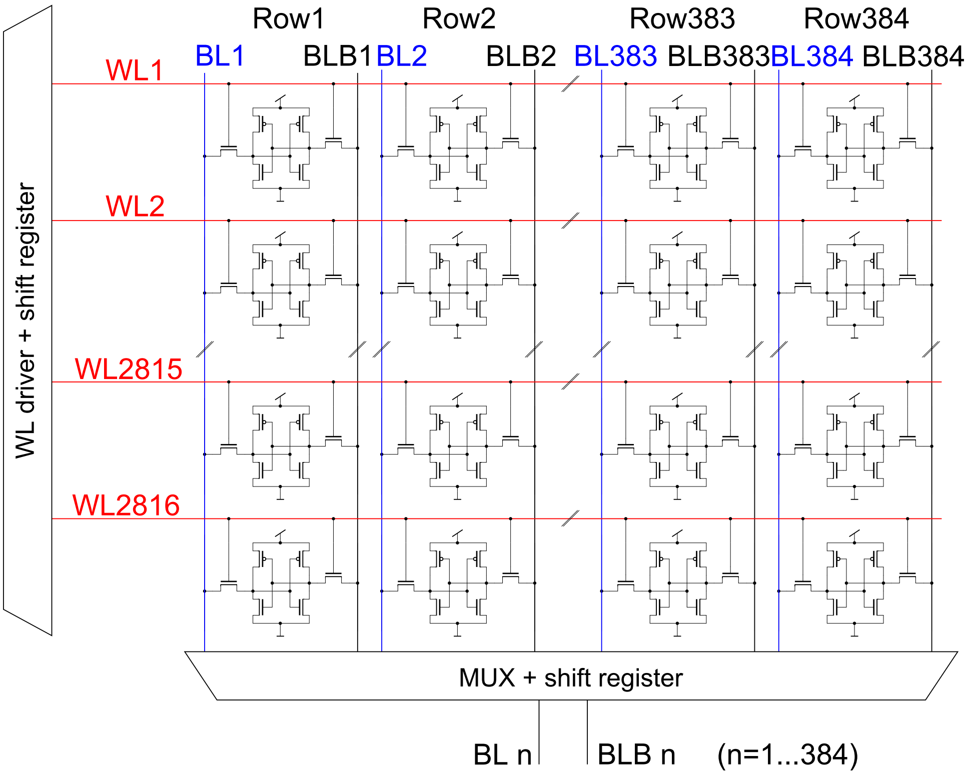


Fig. 5.5: 1 MBit SRAM array with Bitline- and Wordline-shift registers and multiplexed bitline pairs for read current measurement in any cell. Current measurement is used here for cell reading. It could also be done with sense amplifiers.

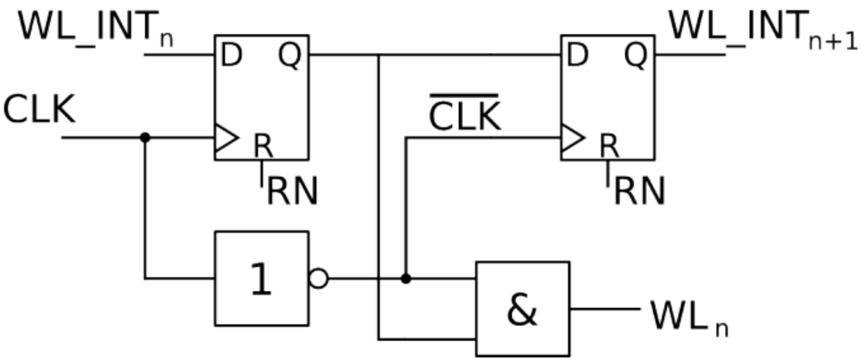


Fig. 5.6: One stage of the WL shift register which is used for the universal SRAM array instead of WL decoder to avoid many address pins. It consists of 2 latches that pass a starting bit from WL_INT_n terminal to the WL_INT_{n+1} terminal during one clock cycle. [29]

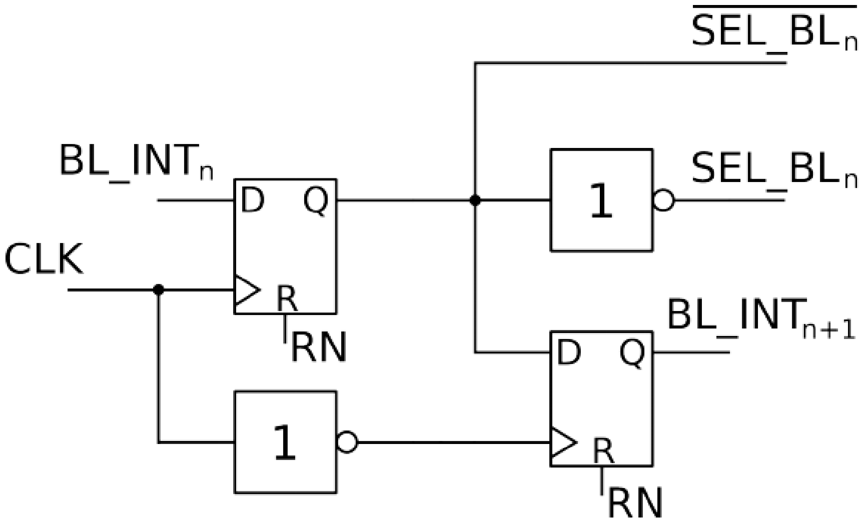


Fig. 5.7: BL shift register which is used for the universal SRAM array instead of BL decoder to avoid many address pins. Additionally to the shift register functionality, it provides the two signals SEL_BL and $\overline{SEL\_BL}$ to switch on the BL transfer gates in Fig. 5.8. [29]

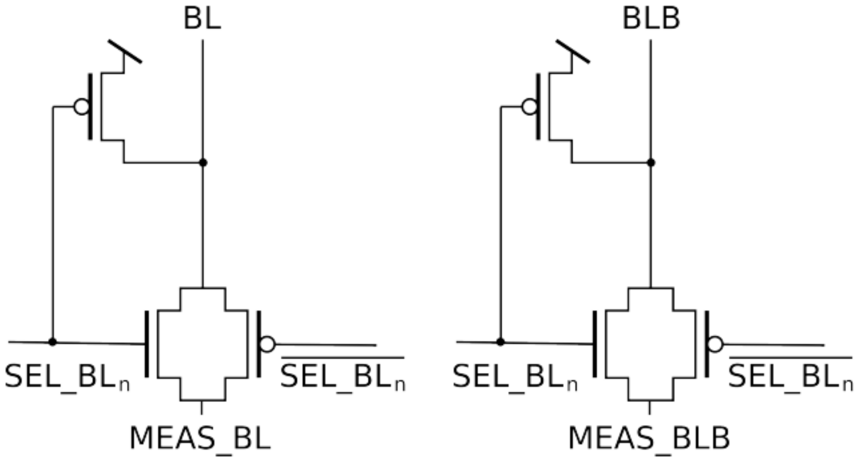


Fig. 5.8: Transfer gates at every BL and BLB to connect to the external pins MEAS_BL and MEAS_BLB. They have big dimensions so that the voltage drop is typically <1mV. The non-selected BLs and BLBs are connected to V_{DD} via pMOS precharge transistors. This enables read conditions for all 384 cells that are connected to one wordline. [29]

5.4.2 Measurement Methodology

The combination of this product-like test chip with the cell-flip based stability metrics provides statistical information on the SRAM array. It is not very well suited for single cell analysis.

First, all cells in the array are written to the same state (all '1' or all '0'). Then the core array voltage is lowered in steps of 10 mV, while the periphery with WL, BL, and BLB (Fig. 5.1) is kept at nominal V_{DD}=1.2 V. After 10 ms with lowered core voltage, it is raised again to nominal V_{DD}. This time of low V_{DD} is long enough to reach the static case. If this time was shorter than the RC constant of the array, the voltage would not reach the low value at all cells. This would correspond to a dynamic stability analysis which yields higher cell stability [46]. After each $V_{DD,core}$ lowering step, the array is read with nominal V_{DD} to count flipped cells. Then the core voltage is again lowered by another 10 mV. The number of flipped cells for each reduced core voltage is the direct output from this measurement technique. The RM stability distribution is simply the derivative of this curve.

It is important to note that the point where the cell flips is the same point where the bitline current drops. This is why the flip statistics in the end provides the same information as the measurement of the bitline currents. But by testing the flip statistics, many cells can be addressed in parallel, and no VI-curves have to be evaluated. This is the speedup factor in this measurement approach. The memory does not have to be re-written to the initial state after each voltage reduction step: we have found that the same behaviour occurs, no matter if the chip is rewritten or not.

Read stability RM

Read stability can only be measured along wordlines, because all cells have to be set to read conditions (WL=BL=BLB=$V_{DD,nom}$=1.2 V) at the same time. After the core voltage of the complete array is reduced to the specified value, the wordline is switched on for 10 ms. By use of the shift registers in our test structure or in general by a wordline decoder, wordlines can be set to read conditions sequentially. Then all cells along the activated wordlines are read with nominal V_{DD}. This could for speedup reasons also be done with the lowered voltage. We have chosen to step the core voltage between 1.1 V and 0.7 V in 41 steps and evaluate the average flip values over 50 bitlines, i.e. 19.200 cells.

Hold stability or minimum retention voltage $V_{min,ret}$

The same measurement like for read stability along wordlines, but without activating the wordline during low V_{DD}. Since the cells are much more stable in 'hold' state, the cells start flipping at a much lower core voltage, this is why it is stepped between 0.4 V and 0 V. But it is more effective to measure hold stability along bitlines, because they are much longer (2.816 cells compared to 384 cells), this improves the statistics. In this setup, the bitline conditions can be chosen: For our normal hold stability examinations they are set to read conditions, but also write conditions could be applied through the 2 pins at the MUX input (Fig. 5.5).

5.4.3 NBTI stress conditions

One motive for development of our method was fast characterization of NBTI stress impact on SRAM cell stability. Therefore, the complete array is written to '0' (Fig. 5.1), then the temperature of the chip is raised to 125 °C. Additionally, $V_{DD,core}$ is in the first experiment raised to 1.8 V for 1.000 s, then in the second experiment to 2.2 V for additional 10.000 s. Temperature is then decreased to 25 °C and all measurement is done at room temperature (the NBTI recovery effect therefore cannot be observed, because cool-down takes too long so that the fast-recovering component already disappeared). Pullup PL2 (which is on the '1' side of the memory cell, Fig. 5.1) degrades during this procedure, because it has full NBTI stress conditions applied (Fig. 3.5). Previous measurements on single pMOS SRAM transistors [29] have confirmed simulations about threshold voltage drifts of approx. 9 mV and 40 mV for these stress conditions, respectively. 9 mV ΔV_{th} correspond to a real user profile of 200 hours @ 125 °C and V_{DD}=1.2 V or 20 hours @ 125 °C and V_{DD}=1.32 V , while 40 mV ΔV_{th} correspond to 1/3 year @ 125 °C and 110% V_{DD} (=1.32 V) or 1.5 years @ 125 °C and 100% V_{DD} (=1.2 V). These V_{th} shifts can be translated to approx. 1% and 4% SNM(read) loss, respectively. Fig. 4.17 shows the normalized SNM(read) and SNM(hold) simulation results for pMOS V_{th} shifts: one pMOS pullup V_{th} was varied first on the '0' side, then on the '1' side of the memory cell. The cell stability is improved in one case, but is degraded in the other case.

9 mV ΔV_{th} happens for probably every real-product memory cell; results will show if this slight V_{th} shift is visible in the RM distribution. The 40 mV case is a 'some years of usage' case for memory cells: it should be obvious in the stability diagram.

5.5 Results

Lowering $V_{DD,core}$ reduces the stability of the cells: the lower $V_{dd,core}$, the more cells flip. The most unstable cells are those with high V_{th} mismatch. They are unsymmetrical and flip even at higher core voltages, but will of course also flip with low core voltage and then stay in this more stable state. This is why the flip-cell curve is a monotonic function. Only at the lower end of the flip curve, slight non-monotonicity is observed. This is due to the symmetric cells with very little mismatch. They flip only at very low core voltages, because they do not have a preferred state. Therefore they can reflip to the other state after they have changed once. But this does not distort the main information in this technique. There is no influence of the time duration of V_{min} or the slope of the V_{DD} reduction, as long as the voltage reaches V_{min}.

Unstressed arrays are statistically symmetrical: the same flip curves and therefore the same distributions (derivatives) are found for '0 to 1' and for '1 to 0' flips, as both have exactly 50% of the whole cell count for very low core voltages. But NBTI stress makes them asymmetrical: in our case, where the '0' state was stressed, the '0 to 1' flips increase, while the '1 to 0' flips decrease. This agrees with the SNM simulation results:

while degrading the pMOS on the '0' side of the memory cell reduces stability, degrading the pMOS on the '1' side of the memory increases stability (Fig. 4.17).

Read stability along wordlines

Fig. 5.9(a) shows the flip statistics including the RM distribution averaged over 50 wordlines with 384 cells each (=19.200 cells). The RM distribution is Gaussian distributed and follows the function

$$f(x) = a_1 exp[-(\frac{x-\mu}{\sigma})^2] \tag{5.5.1}$$

Fitting this distribution to a Normal distribution restulted in μ=902 mV and σ=55 mV. This means that the expectation value of RM was 1.2 V - 902 mV = 298 mV. The strong variations in only 19,200 cells caused the minimum RM value at 170 mV, the maximum RM value at 430 mV. This is a huge difference of 260 mV! Fig. 5.9(b) shows the flip statistics before and after NBTI stress. The derivative from this curve is the stability distribution pre- and post-NBTI stress, which is plotted in Fig. 5.10. Even at high NBTI threshold voltage drift, there is no dramatic shift in the flip statistics and therefore in the RM distribution. The NBTI-shifted curve has $\mu = 920\ mV$ and $\sigma = 54\ mV$. While 9 mV ΔV_{th} shift is hardly visible in the distribution, 40 mV ΔV_{th} shifts the read stability distribution only by approx. 17 mV, compare Fig 5.10. This measurement perfectly confirms the simulations. The distribution is still Normal distributed, the mean value was right-shifted by approx. $\Delta V_{th}/2$ and the standard deviation remains almost constant. This shows that under these conditions, the impact of NBTI to read stability is not critical if compared to the considerably larger production variability.

pre-stress				post-stress			
nominal	$+1\sigma$	$+3\sigma$	$+6\sigma$	nominal	$+1\sigma$	$+3\sigma$	$+6\sigma$
298 mV	55 mV	133 mV	-32 mV	280 mV	54 mV	117 mV	-45 mV

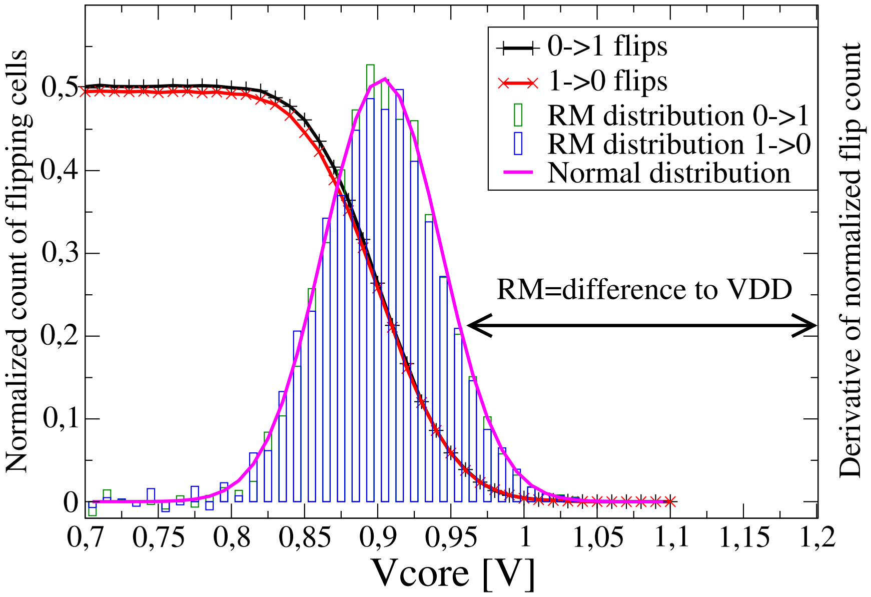
Table 5.2: Read Margin data. The distribution got right-shifted by 17 mV, while the standard deviation did not change considerably. This perfectly aggrees with simulations. Negative RM means real data loss during read operation, without lowering V_{DD}. NBTI does not change this value considerably.

Hold stability along wordlines

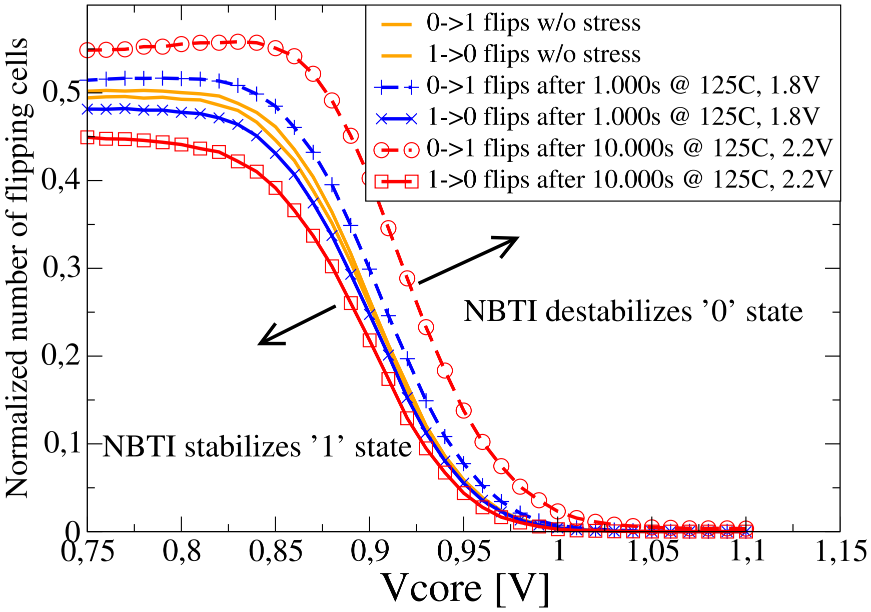
Measurements have proven the same behavior of hold stability along wordlines as along bitlines. Graphs are therefore only given for the hold stability along bitlines, see next section.

Hold stability along bitlines

Fig. 5.11(a) shows the flip statistics including the HM distribution averaged over 10 bitlines with 2,816 cells each (=28.160 cells). The HM distribution is also almost Gaussian



(a) Number of flipped cells over V_{core} and derived read-stability distribution averaged over 50 wordlines (=19.200 cells). The Read Margin distribution shows good correlation to Normal Distribution. This plot only considers production variability.



(b) Number of flipped cells over V_{core} pre- and post-NBTI over 19.200 cells stress

Fig. 5.9: Flip statistics and stability distribution for Read Case

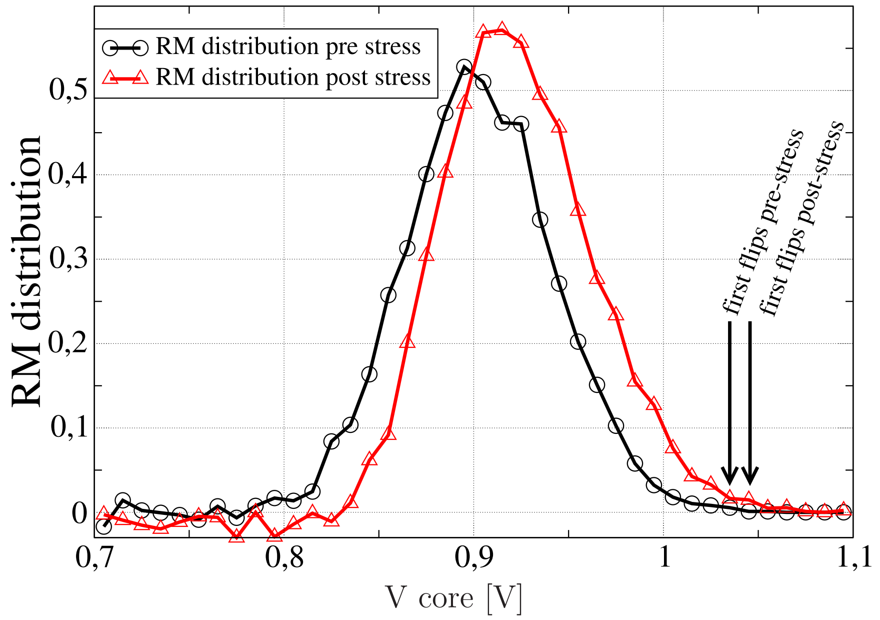


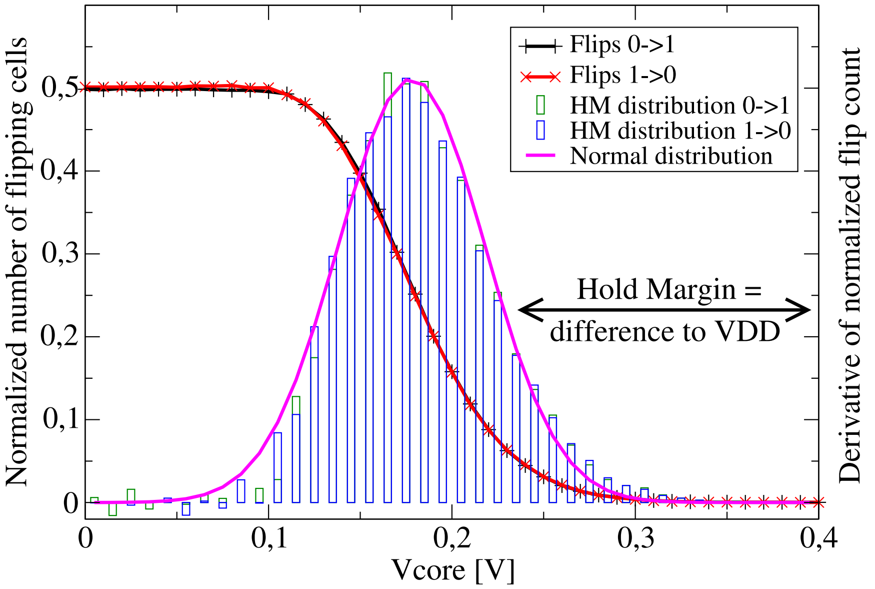
Fig. 5.10: Read stability distributions pre- and post-NBTI stress. Approx. 40 mV pMos ΔV_{th} results in a 17 mV RM distribution shift. The standard deviation remains constant, which confirms the simulation results.

distributed. Fitting this distribution to a Normal distribution gave $\mu = 178\ mV$ and $\sigma = 53\ mV$. This means that the expectation value of HM was 1.2 V - 178 mV = 1022 mV. The strong variations in only 28,160 cells caused the minimum HM value at 870 mV, the maximum HM value at 1150 mV. This is again a huge difference of 200 mV. Fig. 5.11(b) shows the flip statistics before and after NBTI stress. The derivative from this curve is the stability distribution pre- and post-NBTI stress, which is plotted in Fig. 5.12. This shifted distribution was fitted to Normal distribution and gave $\mu = 232\ mV$ and

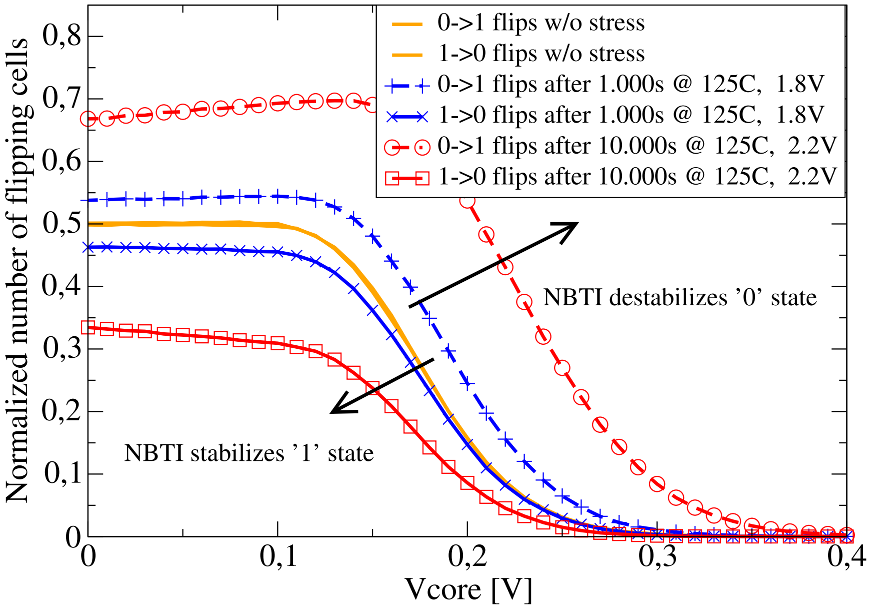
pre-stress				post-stress			
nominal	1σ	$+3\sigma$	$+6\sigma$	nominal	1σ	$+3\sigma$	$+6\sigma$
1.022 V	53 mV	0.863 V	0.704 V	0.968 V	69 mV	0.761 V	0.554 V

Table 5.3: Hold Margin data. The distribution got right-shifted by 55 mV, while the standard deviation increased by 16 mV. Contrary to the RM case, NBTI has a lot of impact on retention stability.

$\sigma = 69\ mV$. A strong shift in flip statistics is already visible for small ΔV_{th}. This means that the mean value of the distribution got right-shifted by 55 mV, while the standard deviation increased by 16 mV. It is also important to notice that after NBTI stress, the hold stability distribution is not Gaussian distributed anymore but shows a larger tail towards high $V_{DD,core}$ values (Fig. 5.12). There are more instable cells on the right side of the distribution, which is bad for the minimum retention voltage. RM distribution post-stress is between 790 mV and 1070 mV. The first flips pre-stress were detected at 0.3 V,



(a) Number of flipped cells over V_{DDcore} and derived hold-stability distribution over 10 Bitlines (28.160 cells). The Hold Margin distribution shows good correlation to Normal Distribution. This plot only considers production variability.



(b) Number of flipped cells over V_{DDcore} pre- and post-NBTI over 28.160 cells stress

Fig. 5.11: Flip statistics and stability distribution for Hold Case

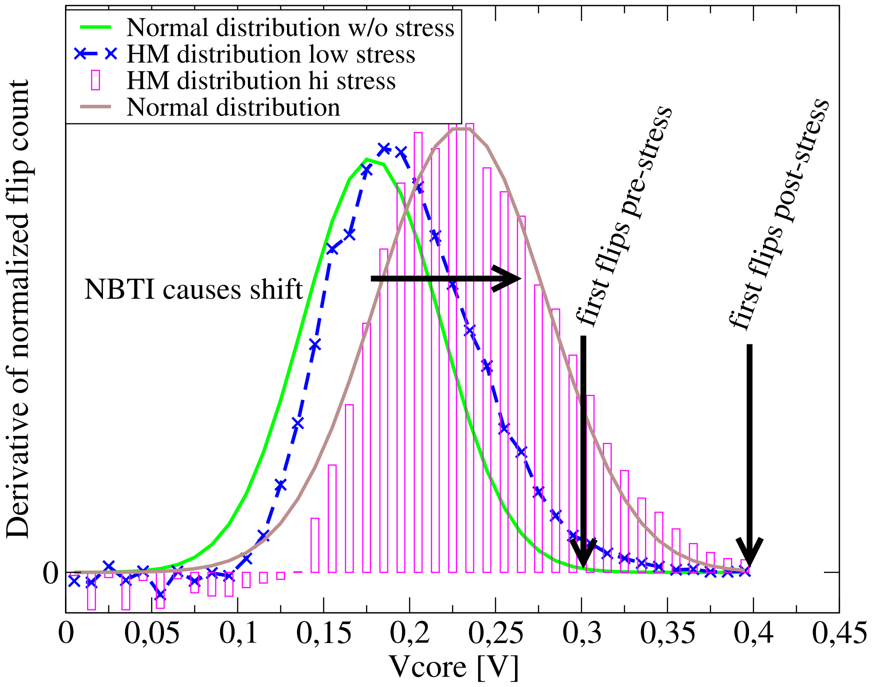


Fig. 5.12: Hold stability distributions pre- and post-NBTI stress. Approx. 9 mV as well as 40 mV pMos ΔV_{th} are both clearly visible in a 15 mV and 55 mV distribution shift, respectively. After long NBTI stress, hold stability distribution does not show good correlation to Normal Distribution anymore.

this value got shifted post-stress to 0.4 V. This means that 40 mV V_{th} rise increases the minimum retention voltage $V_{min,ret}$ by 100 mV!

5.6 Discussion

Hold stability, or minimum retention voltage, is more critical than read stability. But simulations showed that read stability is degraded by 10%, while hold stability is only degraded by 7%, compare Fig. 4.17. However, these values were simulated for nominal V_{DD}. If the supply voltage is not lowered, then hold stability is uncritical. Only because in today's systems, voltages are reduced especially during retention mode, NBTI becomes critical in hold case. The reason is that cells flip at approx. 0.9 V in read case compared to approx. 0.4 V in hold case, and the threshold voltage drift of 9 mV / 40 mV is much worse relative to the lower V_{DD} value.

SRAM suffers most from NBTI degradation during retention case. V_{th} sensitivity is maximum for the low supply voltage, but the distribution gets also wider and not Gaussian distributed anymore. This is because the NBTI degradation variations are maximum in SRAM circuits. They have the smallest transistors, and only one transistor in the circuit, namely the pullup, is the critical one for stability. In typical logic, the transistors are bigger and have more logic depth, so one strongly degraded transistor is normally balanced by a hardly degraded transistor, which is not the case for SRAM. This fact is depicted in Figs. 5.13: the simulation of the variation of the NBTI degradation is calibrated to 1 for minimal logic transistor width employed in a 40 nm standard library. The expected NBTI degradation variation is severely reduced for wider transistors. Additionally, in typical critical paths of combinatorial logic, several physical gates build up the logic path, which also attenuates the statistical NBTI degradation. Fig. 5.14 supports this simulation with measurements. The measured variance/mean decreases with the increased gate area and

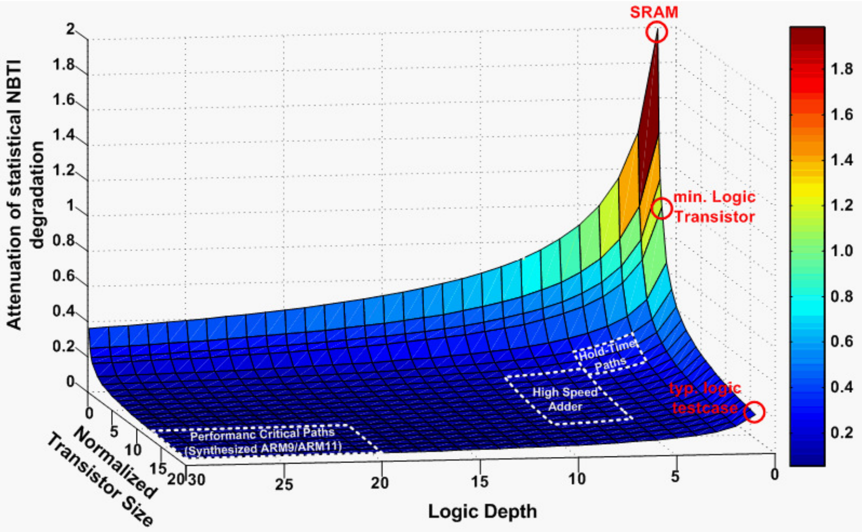


Fig. 5.13: SRAM transistors have smallest area and only logic depth=1. Therefore they show the strongest NBTI variations in a SoC.

confirms the qualitative picture of Fig. 5.13. The observed V_{th} variation is worst for SRAM cells.

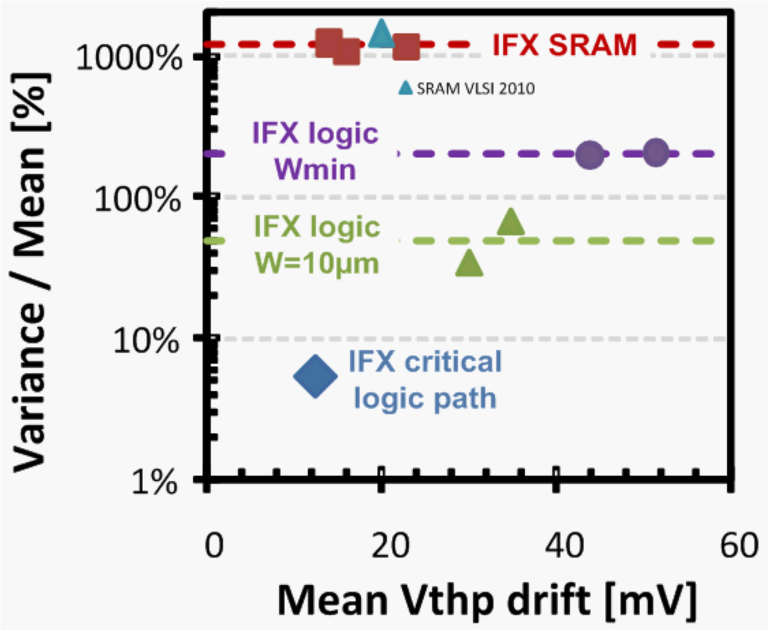


Fig. 5.14: SRAM has strongest NBTI variations of all transistors used in a SoC. [Graphics provided by Infineon Report]

5.7 Conclusion

A fast and simple approach to measure and analyze the read- as well as the hold-stability of large-scale SRAM arrays was presented. This approach, which works for any SRAM implementations (6 transistors or more), can be even implemented in product chips by separating V_{DD} between core and periphery, which allows a fast stability analysis in-field. With this approach, the stability distribution of a 65 nm 6T-SRAM cell array was analyzed directly after production: it shows good correlation to Normal Distribution. Additionally, the impact of NBTI on the stability distribution was demonstrated.
The results for read stability (measured by Read Margin distribution) are:

- Right shift of the whole distribution by approx. 1/2 pMOS ΔV_{th}, which corresponds to RM reduction by 1/2 pMOS ΔV_{th}
- Almost no widening of the distribution, i.e. standard deviation is not changed
- Good fit to Normal distribution pre- and post-NBTI stress
- RM is a linear measure for pMOS V_{th} shift and therefore cell stability. This is also true if the core voltage is not reduced to 0, but to some higher voltage. At least some bits must flip for the most stable state, otherwise the linear region is not yet reached.

- Measurements confirmed the simulation results

The results for hold stability (measured by Hold Margin distribution) are:

- Right shift of the whole distribution by approx. pMOS ΔV_{th}
- Widening of the distribution
- Not ideally Normal distributed anymore
- Rule of thumb: N mV pMOS NBTI induced ΔV_{th} leads to $2 \cdot N$ mV HM reduction

So the main killer effect of static NBTI is the strongly increasing minimum retention voltage. But on the other hand, if the supply voltage is reduced always when the cell is in retention mode, then the NBTI degradation will be very small due to the small electric field. This worst case will then be no worst case after all.
The read case is not dramatically influenced by static NBTI, but this will be examined in the next section for recovering NBTI.

Chapter 6

Impact of Recovering NBTI on SRAM Arrays

Abstract

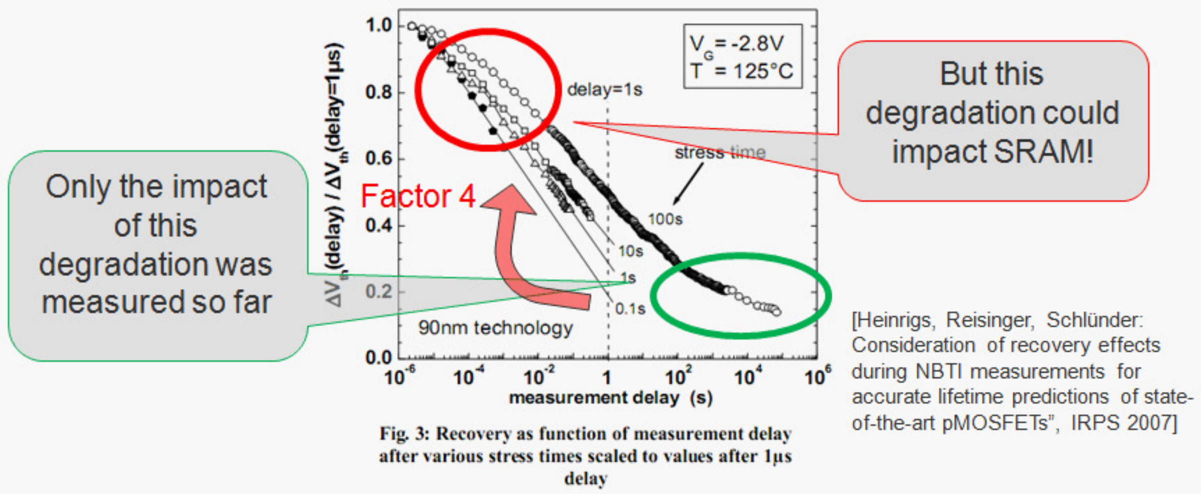
After addressing variability of manufacturing and of degradation, the second challenging problem of NBTI must be examined: the recovery effect. This chapter presents stability analysis of large-scale SRAM arrays directly after terminating NBTI stress. While the impact of static NBTI is well examined for cells and arrays, the fast-recovering component was not yet measured on SRAM arrays. The novel method presented here analyzes the flipping of cells directly after the supply voltage was lowered to a specific value where the structure is most sensitive for NBTI induced cell flips. Thus, read margin criterion is used to characterize the decreasing cell stability due to NBTI degradation with a resolution down to 1 ms. Applying this method, the impact of static and dynamic NBTI is measured on a 1 MBit product-like SRAM array fabricated in a 65 nm low power CMOS technology. The most important results are reported in:

S. Drapatz, K. Hofmann, G. Georgakos, and D. Schmitt-Landsiedel:,"A method to analyze the impact of fast-recovering NBTI degradation on stability of large-scale SRAM arrays", Solid-State Electronics, 2011

S. Drapatz, K. Hofmann, G. Georgakos, and D. Schmitt-Landsiedel: "Impact of fast-recovering NBTI degradation on stability of large-scale SRAM arrays", European Solid-State Device Research Conference ESSDERC, 2010

6.1 State of the Art: Measure Recovering NBTI on SRAM Cells

As already stated in section 3.5.3 on page 31, NBTI is partly a static, but even more a dynamic process [47]. In Fig. 6.1 the impact of the recovering component can be seen.



Only the impact of this degradation was measured so far

But this degradation could impact SRAM!

Fig. 3: Recovery as function of measurement delay after various stress times scaled to values after 1µs delay

[Heinrigs, Reisinger, Schlünder: Consideration of recovery effects during NBTI measurements for accurate lifetime predictions of state-of-the-art pMOSFETs", IRPS 2007]

Fig. 6.1: All measurements on SRAM stability have been performed on already recovered NBTI long after end of stress. It is necessary to measure the stability of SRAM cells directly after termination of stress conditions to examine the full impact in real life.

The annealing time constant and the ratio between static and dynamic V_{th} shift depends on stress time and stress voltage. Generally, with high stress voltages this annealing process is very fast, so that after a couple of seconds a major part of the voltage shift has disappeared. SRAM NBTI stability measurements are typically done after accelerating stress conditions (high T, high V_{DD}), otherwise the V_{th} shift would be too small. Cooling down after accelerating stress takes some time, and the threshold voltage shift meanwhile decreases to a quasi-static value which does not change considerably within the next hours or days, compare Fig. 6.1. This value therefore is called the permanent or static NBTI, although it will decrease further on a logarithmic time scale, with very long time constants (weeks to months).

Altogether, this means that with nominal measurement techniques at room temperature, it is impossible to detect the dynamic part of NBTI degradation effects on SRAM stability. This is why all investigations on SRAM arrays only address the static NBTI effect, the dynamic part has not been measured on SRAM arrays so far.

Especially in SRAM, storing one value for a long time represents the NBTI DC worst case, while in logic circuits, the dominating AC case allows continuous recovery. Fig. 6.2 shows the so-called 'S-Curve' [48], which is a qualitative plot of measured V_{th} shift over the duty factor of the stressing signal. But in real SRAM products, there is no time for recovery. In hold state, one pullup experiences NBTI stress. When the cell is read, the NBTI stress is not interrupted, since the voltage conditions on the pullup transistors stay the same. This is even true for half-selected cells, which means that the WL and the BLs are high, compare Fig. 6.3.

The problem only appears when the NBTI effect is accelerated. Therefore, the supply voltage is raised, and to perform measurement, this increased supply voltage must be lowered to nominal supply voltage before. This is where the problem appears: only 1 s after reducing the acceleration voltage, already 50% of the drifted threshold voltage have

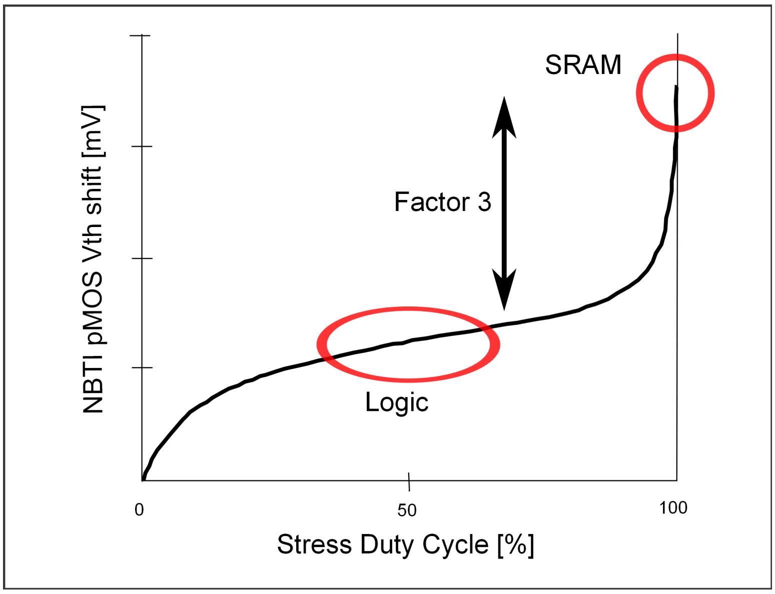


Fig. 6.2: Dependency of V_{th} shift on the duty factor of the stressing signal: DF=1 (long hold at SRAM) gives 3 times more V_{th} shift compared to DF=0.5 (typical clocked logic) [48]

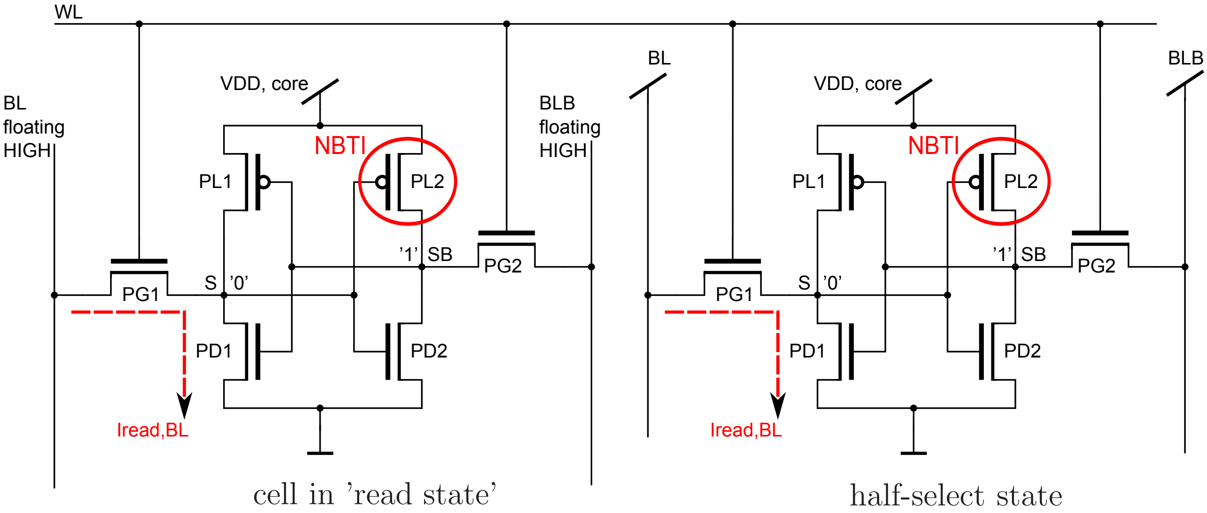


Fig. 6.3: Even half-selected cells do not allow recovery. This means that as long as a cell is not written, it is under NBTI conditions without recovery phases.

disappeared. So when the cell is measured some hours later, only a fraction of the ΔV_{th} is left. Therefore it is of big importance to know this worst effect of NBTI on SRAM stability, i.e. what is the impact of the maximum V_{th} shift directly after end of stress.
This is why measurements are needed which are done extremely fast after reducing the voltage. But to the best of our knowledge, such measurements have not been performed on SRAM arrays.
In the past, several approaches were developed to characterize the stability of SRAM cells and SRAM arrays, which could be and partly were used to measure the impact of static NBTI. Static Noise Margin (SNM) [12] as a stability metrics is suitable for simulation. Other techniques, e.g. Read Margin (RM = difference between reduced core voltage where the cell flips and nominal voltage) determined by current measurement [16] [15], are suitable for single cell analysis. The flipping cell analysis in chapter 5 is particularly suitable for fast array measurement. Therefore it was extended for the task of dynamic NBTI characterization.
The test chip and the measurement setup with its specific challenges are described in the next section. Results are presented in section 6.3, followed by the conclusion in the last section.

6.2 New Approach

6.2.1 Universal SRAM Testchip

All measurements are performed on a 1 MBit SRAM array test structure in a 65 nm low power technology [45] as described in section 5.4 on page 80. Shift registers on bitlines and wordlines allow to select single cells in the array without a large number of address pins, and bitline-pairs are multiplexed to 2 external pins for direct read current measurement (Fig. 5.1 on page 77) [29]. As in the static NBTI experiments in chapter 5, the current measurement is not necessary for this analysis method. It is only done because of the simplified periphery in this test chip to detect the cell state, which could also be done faster with conventional sense amplifiers.
An important feature for this technique is the fact that selecting one bitline-pair causes all other bitlines being clamped to 1.2 V. Thus, all 384 cells along the selected WL are in read condition, which is the so-called half-select state (WL=BL=BLB=1). All other $2815 \cdot 384$ cells in the array are in hold condition.

6.2.2 New Measurement Methodology

Stability analysis is based on the flipping behavior of the cells. The goal is to detect cell stability over time after stress with an ordinary SRAM array. Therefore, only a part of the cells should be sensitive to NBTI degradation in each time slot, while the rest is not affected at all. We start from the fact that read state is less stable than hold state, and that all 384 cells along one wordline (WL) can be set to read state, while the rest of 2815 WLs is kept in hold state. By sequentially activating only one WL at a time, the complete array serves as a time-dependent NBTI sensor: the cells in read state flip according to the actual NBTI degradation, while the cells in hold state are more stable and do not flip. This technique is illustrated in Fig. 6.4. After writing all cells in the array to one state

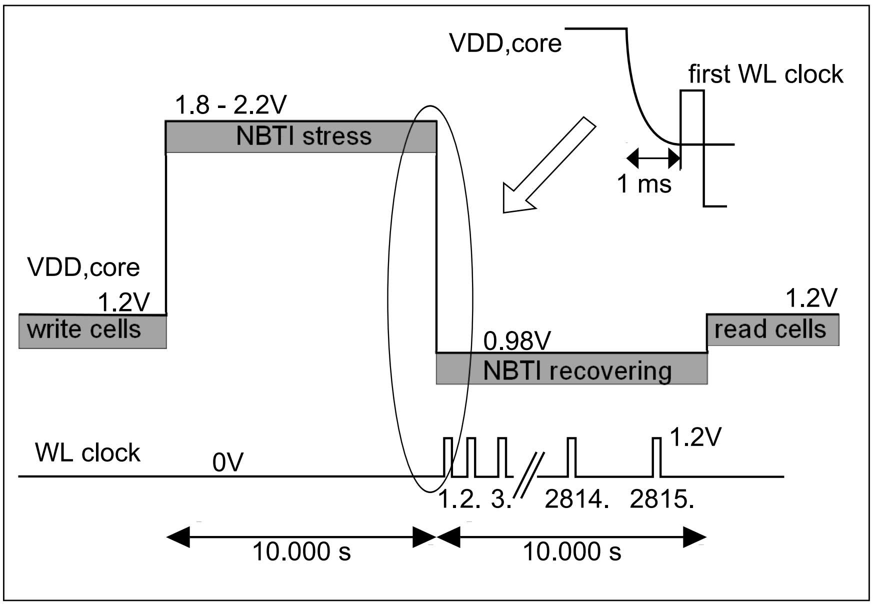


Fig. 6.4: Schematic overview of measurement approach with the two important signals V_{DDcore} and WL clock

(here '0'), $V_{DD,core}$ is ramped up to the NBTI acceleration voltage (in our case between 1.8 V and 2.2 V) for 10.000 s to stress pullup transistor PL2 (Fig. 5.1). To get sufficient NBTI acceleration conditions, the complete measurement is performed at 125 °C. The periphery voltage, i.e. the supply to all WL and BL drivers, is kept at $V_{DD,nom}$=1.2 V. After stressing, $V_{DD,core}$ is lowered to a critical voltage of 0.98 V, and 1 ms later, the first WL clock activates the first WL, setting the first 384 cells to read condition. Now these cells flip according to the actual NBTI-caused V_{th} drift.

Another 1 ms later, a shift register creates the next WL clock, switching off the first WL and activating the second WL. Then, the first WL is set back to hold condition and the second WL now is set to read condition. This is repeated until the WL clock was activated 2815 times and the whole array was sequentially set to read condition for a short fraction of time.

Switching the WLs is done with exponentially increasing delay between the WL activations, leading to a quasi-logarithmic time scale. (It is called 'quasi'-logarithmic because the very small time values are 'compressed', i.e. the sub-second decades do not show equidistant behavior.) The goal is to achieve a complete recovery time of 10.000 s, which is as long as stress time. Therefore, after waiting $1.0038^{\sharp WL}$ ms at each WL, the shift register switches to the next WL. Time between WL ♯1 and WL ♯2 therefore is 1.0038 ms, time between WL ♯2799 and ♯2800 is approx. 41 s. The left half of Table 6.1 provides the timing information. The cells are not read out during this procedure, this would take much too long. They are only set to read conditions all in parallel, which is a kind of 'simulated read'. After WL deactivation, the flip pattern is kept in the stable hold state,

$delay = 1.0038^{\sharp WL}ms$			$delay = 1.007^{\sharp WL}ms$	
#WL	t betw. 2 WL	t after stress	t betw. 2 WL	#WL
1	1.0038 ms	1 ms	1.007 ms	1
10	1.038 ms	10 ms	1.07 ms	10
85	1.38 ms	100 ms	1.7 ms	76
413	4.78 ms	1 s	8 ms	298
965	39 ms	10 s	70 ms	610
1566	0.38 s	100 s	0.7 s	938
2173	3.8 s	1 ks	7 s	1268
2800	41 s	10 ks	70 s	1600

Table 6.1: Relation between WL number and time after stress. For times >0.1 s, the WL results in an almost exponential x axis. Only very short times are compressed.

after all it is a memory array! Reading of the flipped cells is done after the relaxation process with nominal $V_{DD,core}$, when reading does not influence the cell states anymore. In our approach, we only want to detect the flipping of a cell, so the measured BL current was used only to detect the cell state. This could be performed even faster in an ordinary SRAM array, as the bitline current measurement is no prerequisite of this approach. The only prerequisites are separate V_{DD} rails for the core array and for the periphery. Reduced core voltage is necessary for the flip-cell analysis (section 5.4.2), enlarged core voltage is necessary for accelerated NBTI stress.

Read margin analysis (stepwise decreasing V_{core} and counting the number of flips, compare chapter 5) of the unstressed and of the stressed+recovered array has given the flip curve in Fig. 6.5. From the 125 °C curves, the most NBTI sensitive core voltage was determined to 0.98 V: there the NBTI sensitivity was maximum ($\partial Flips/\partial Vth, NBTI \approx 1.5Flips/mV$), together with only 20% of flips in unstressed condition, leaving enough headroom for additional flips due to NBTI-induced V_{th}-shift.

6.2.3 Measurement challenges

Temperature is kept stable at 125 °C ± 0.05 °C during write, stress, recover and read. Otherwise, the flip count would vary due to temperature fluctuations. Also writing at 25 °C and heating up quickly can cause random cell flips. On Fig. 6.5 the shift of the flip curve by 50 mV Read Margin between 25 °C and 125 °C can be seen. Simulations confirm 20% less stability at 125 °C compared to room temperature. A Peltier Element together with a PID regulator enables the circuit to keep the temperature in such narrow margins.

A fast falling voltage slope must be generated at switching from stress to recovery voltage to reach a high resolution time in the ms range. This is done with an active filter with a time constant of 100 µs. This guarantees a slope time of about 0.5 ms independent of the impedance of the array (which changes with temperature alteration). Using a switch was avoided to guarantee a well defined timing behavior without voltage peaks or drops.

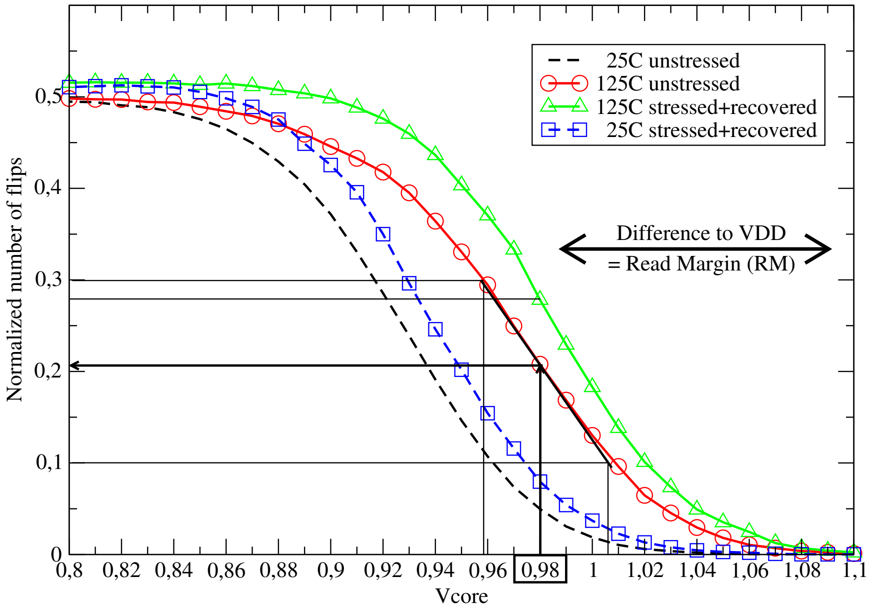


Fig. 6.5: Flip curve of WL ♯1500 at 25 °C and 125 °C in unstressed and stressed+recovered state. The highest NBTI sensitivity at 125 °C together with a low number of flips in unstressed condition could be found at 0.98 V.

Furthermore unintended flips could occur if the slope was too fast. This does not happen with the used slope, which was verified via experiments. A stable voltage is required especially during the recovery period. The sensitivity to $V_{DD,core}$ is comparable to the sensitivity to V_{th} drift: $(\partial Flips/\partial VDD \approx \partial Flips/\partial Vth, NBTI \approx 1.5 Flips/mV)$. So if a V_{th} drift of some tenths of Volts is to be measured, $V_{DD,core}$ must be stable to some mV.

6.2.4 Testchip challenges

In Fig. 6.6 the flip curve of the unstressed array is plotted (low-pass filtered over 10 WLs using a median filter). The unstressed 20% flips at 0.98 V in Fig. 6.5 refer to a total of 384 cells, which equals approx. 80 flips ($384 \cdot 0.2 \approx 80$) that are visible at WL ♯1500. Ideally, this should show constant behavior over the WLs without NBTI stress, but a gradient is visible. Examinations have shown that this is due to the design of the test chip. It was reported also at read current measurements with the same test chip in [29]. The reason is IR drop along the wordlines resulting in a gradient in access-transistor gate-voltages. This could be avoided by dividing the 1 MBit array into smaller subarrays. Here, the simple countermeasure was only to use the 1600 WLs ♯1150 to ♯2750 (in the circle of Fig. 6.6) for the measurements. The x axis is therefore sampled with the law $1.007^{\sharp WL}$ ms to cover 10.000 s of recovery (right half of Table 6.1).

To check the reproducibility of the obtained data with this approach, the described setup was then used twice on different days. The first impression of almost identical data was confirmed with the differential plot in Fig. 6.7. It proves that 2 measurements typically

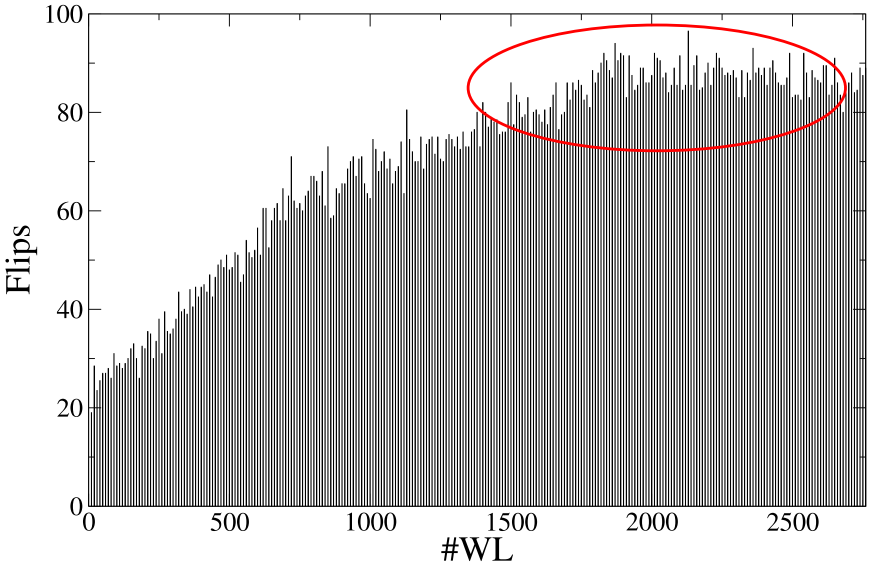


Fig. 6.6: Flips along 2800 WLs for unstressed array show non-constant behavior due to test chip design. The quasi-constant WLs ♯1150 to ♯2750 in the ellipse are used for measurements within this paper.

have only 1-2 flips difference between each other, while the mean value of this ’flip noise’ equals zero. This confirms that this flipping bit technique allows to obtain reliable data. It also shows the small noise component in the recovery behavior.

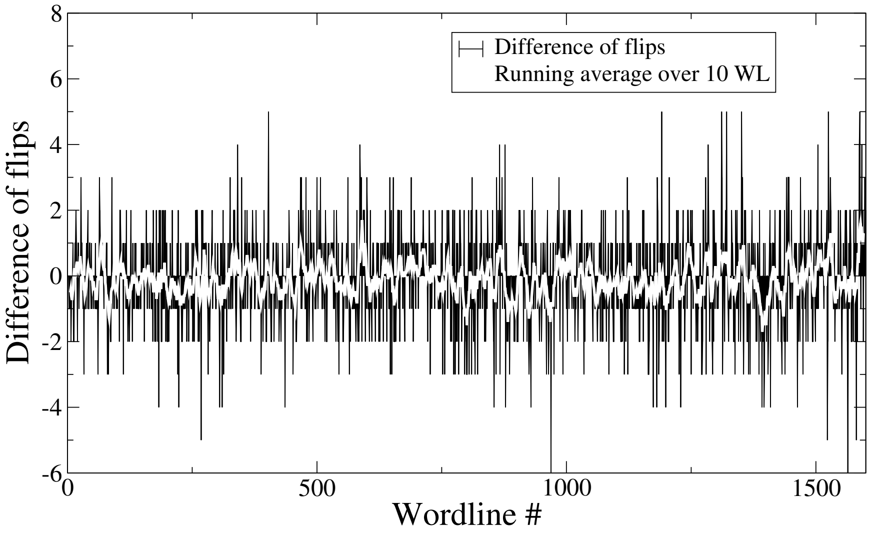


Fig. 6.7: The flip noise between 2 measurements on different days is approx. 2-3 flips, the mean value of the flip noise is zero.

6.3 Results

The flip curves between WL ♯1150 and ♯2750 after stressing the array for 10.000 s with various stress voltages at 125 °C are presented. All graphs are low-pass filtered over 10 WLs using a median filter. This is done to properly display the results in a bar plot and to minimize the impact of mavericks to see a clear trend in data. Additionally, the trend of these filtered data obtained by another moving-average low pass filter is drawn in each diagram.

The first stress experiment was done with $V_{DD,core}$=1.8 V. Fig. 6.8 shows the flips directly after stress minus the flips before stress. The almost constant gradient towards less flips, i.e. more stability due to recovering NBTI-caused V_{th} shift, can be seen clearly. The same recovery curve is plotted in Fig. 6.9 for a stress voltage of 2.0 V and in Fig. 6.10 for a stress voltage of 2.2 V.

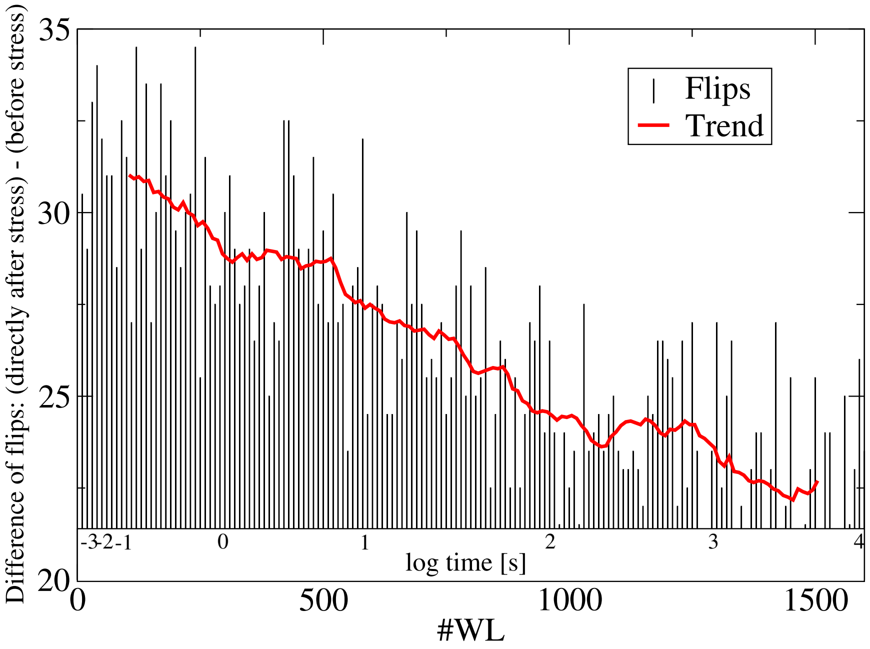


Fig. 6.8: Flips directly after stress minus flips before stress with 1.8 V stress voltage show recovery behavior. The difference between 1 ms and 10.000 s after stress is approx. 10 flips, which equals about 1/3 of the max. flip count.

After measurements directly after end of stress were finished, the array was disconnected from the power supply (V_{DD}=0 V) but kept at 125 °C. Since recovery behaviour also seems to depend on the applied gate voltage V_{GS}, this 0 V case provides maximum recovery. Then after approx. 16 hours without power supply (cells of course lost their states), the complete experiment like depicted in Fig. 6.4 incl. writing the cells to ’0’ state was repeated, only the stress step with higher $V_{DD,core}$ was skipped. It checks if the number of flips has decreased according to the measured gradient before. Figs. 6.11 and 6.12 confirm a constant flip curve of approx. 17 and 39 flips difference between long after stress and

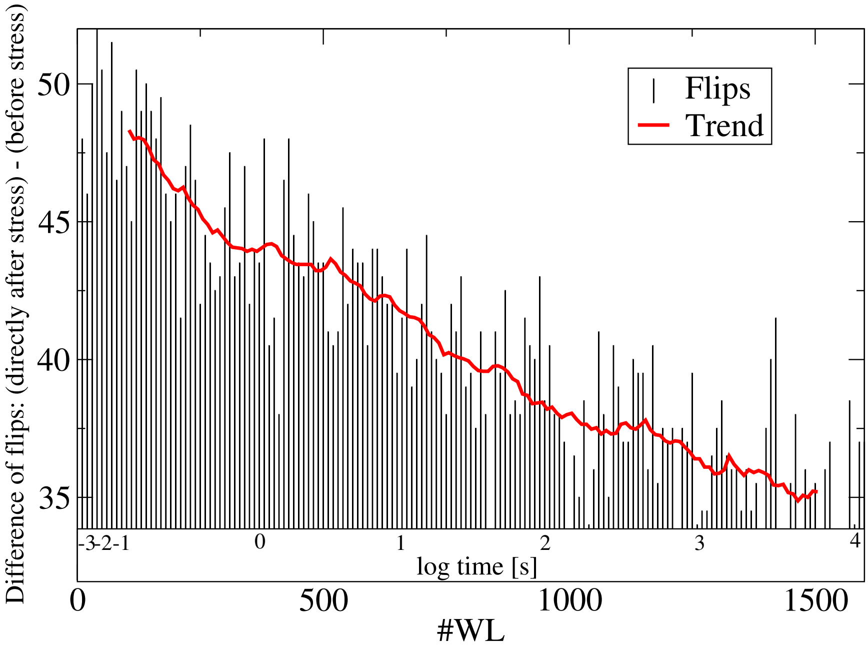


Fig. 6.9: Flips directly after stress minus flips before stress with 2.0 V stress voltage show recovery behavior. The difference between 1 ms and 10.000 s after stress is approx. 15 flips, which equals about 1/3 of the max. flip count.

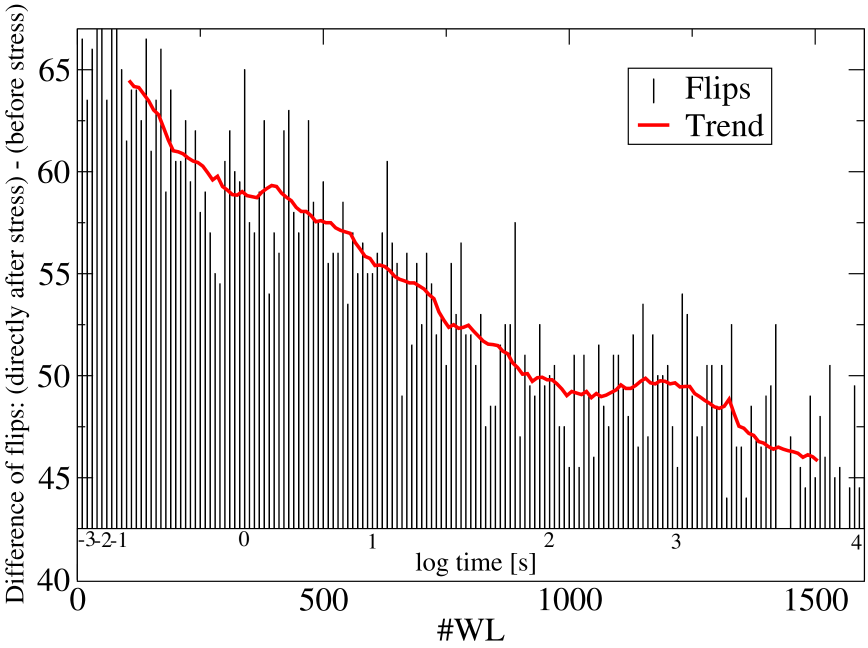


Fig. 6.10: Flips directly after stress minus flips before stress with 2.2 V stress voltage show recovery behavior. The difference between 1 ms and 10.000 s after stress is approx. 20 flips, which equals about 1/3 of the max. flip count.

before stress. This shows that during 10.000 s recovery time, the flips decrease about 30%, then saturate at about 50% (compare Table 6.2).

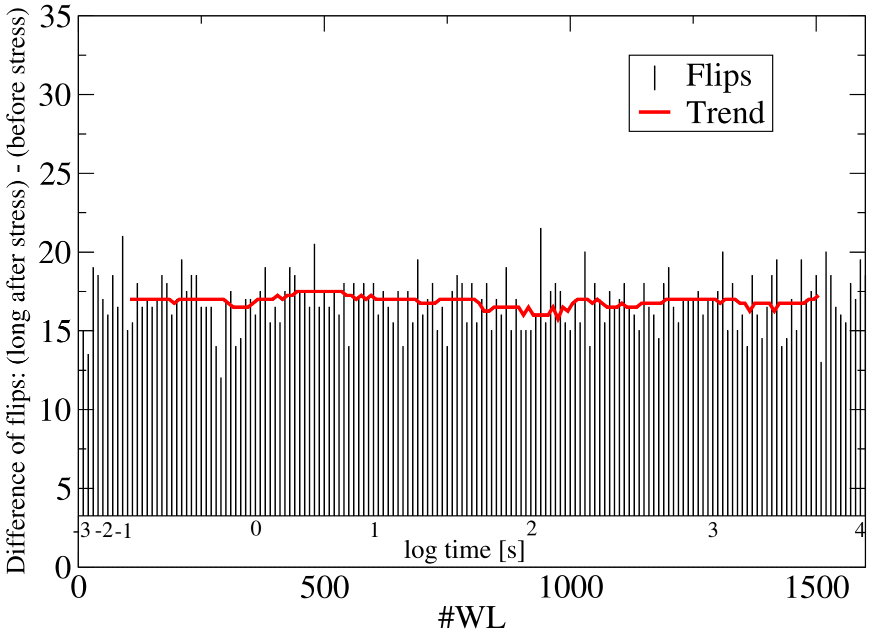


Fig. 6.11: Flips ≫10.000 s after stress minus flips before stress show constant behavior. Approx. 17 flips more in each WL due to static NBTI long after stress.

$V_{DD,core}$ for 10^4 s	flips after 1 ms @ 0.98 V	flips after 10^4 s @ 0.98 V	flips after 16 h @ 0 V
1.8V	32	22 (69%)	17 (53%)
2.0V	48	35 (73%)	N/A
2.2V	65	45 (69%)	39 (60%)

Table 6.2: Comparison of flip count differences pre- and post NBTI stress after different stress voltages and recovery times.

6.4 Plausibility checks

After development, implementation and execution of the new measurement technique, the next necessary step is to evaluate if the results really make sense and if they measured the recovering NBTI effect and nothing else.

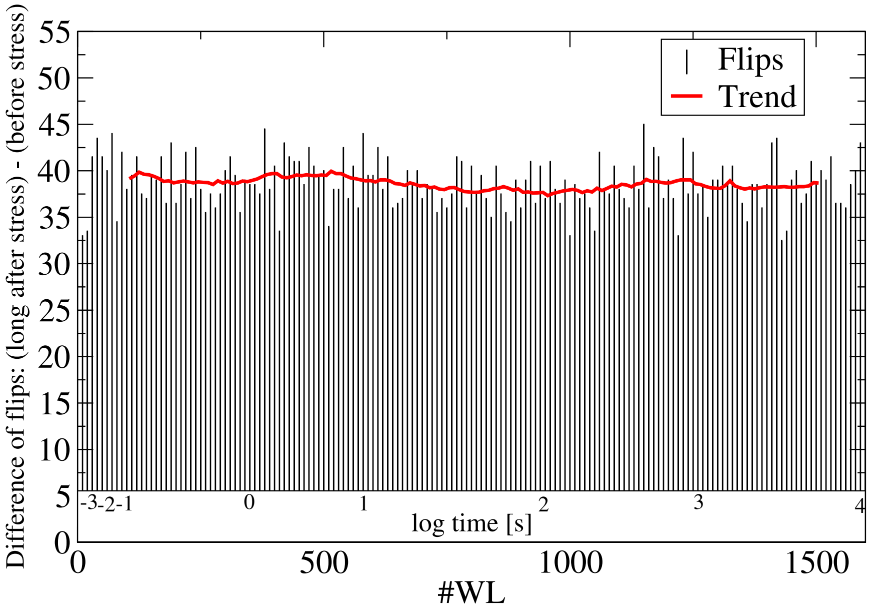


Fig. 6.12: Flips ≫10.000 s after stress minus flips before stress show constant behavior. Approx. 39 flips more in each WL due to static NBTI long after stress.

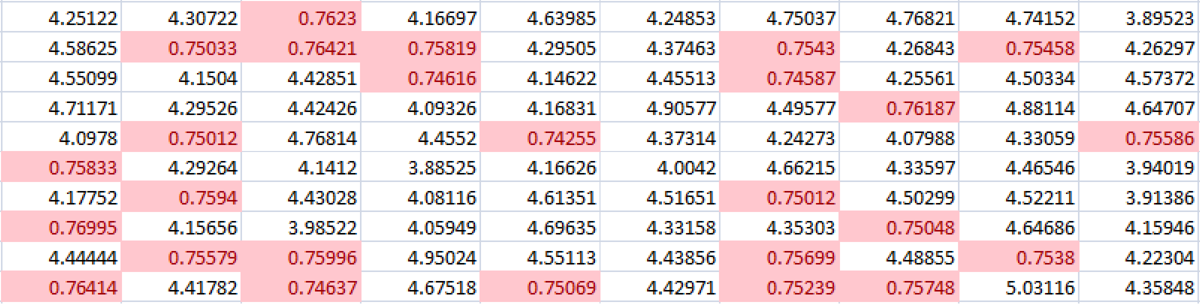
6.4.1 Flip curve withouth stress

The measurements long after end of stress were not only done to check for the static NBTI component long after stress, but also to prove the validity of this technique, i.e. that the measured gradient has no other reason than NBTI recovery. A constant flip curve of 17 and 40 flips over the complete time was seen in Figs. 6.11 and 6.12, respectively. This proves that the gradient in the flip curve before had no other reason than the recovering effect of NBTI, because otherwise the gradient would still be visible.

6.4.2 Bitmap analysis

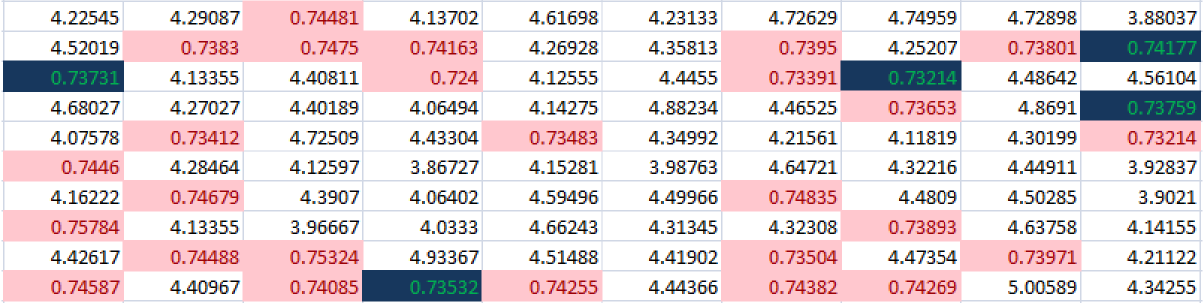
Additionally, a bitmap analysis was performed in order to investigate local distribution of the measured flip data. So far, only the number of flipping bits per wordline was detected. The bitmap analysis checks which bits have flipped to see if there are patterns or other local inhomogenities. It also provides the information if the same bits show flipping behavior and to make clear that this is not a statistical effect.

Each bitmap of 10x10 cells somewhere in the array shows the measured read current in deca-micro ampere (4.25122 means a current of 42.5 µA). Without stress, there are 26 highlighted cells (out of 100) in Fig. 6.13 that flipped to the '1' side only because of the decreased core voltage. It is clear to see that these cells are randomly distributed over the cell area and do not follow an obvious rule or reveal a hardware problem. After 10^4 s stress with 1.8 V, the same cells were analyzed 100 ms after end of stress in Fig. 6.14. Now there were 31 flipped cells instead of 26 cells before. So 5 more cells have flipped



4.25122	4.30722	0.7623	4.16697	4.63985	4.24853	4.75037	4.76821	4.74152	3.89523
4.58625	0.75033	0.76421	0.75819	4.29505	4.37463	0.7543	4.26843	0.75458	4.26297
4.55099	4.1504	4.42851	0.74616	4.14622	4.45513	0.74587	4.25561	4.50334	4.57372
4.71171	4.29526	4.42426	4.09326	4.16831	4.90577	4.49577	0.76187	4.88114	4.64707
4.0978	0.75012	4.76814	4.4552	0.74255	4.37314	4.24273	4.07988	4.33059	0.75586
0.75833	4.29264	4.1412	3.88525	4.16626	4.0042	4.66215	4.33597	4.46546	3.94019
4.17752	0.7594	4.43028	4.08116	4.61351	4.51651	0.75012	4.50299	4.52211	3.91386
0.76995	4.15656	3.98522	4.05949	4.69635	4.33158	4.35303	0.75048	4.64686	4.15946
4.44444	0.75579	0.75996	4.95024	4.55113	4.43856	0.75699	4.48855	0.7538	4.22304
0.76414	4.41782	0.74637	4.67518	0.75069	4.42971	0.75239	0.75748	5.03116	4.35848

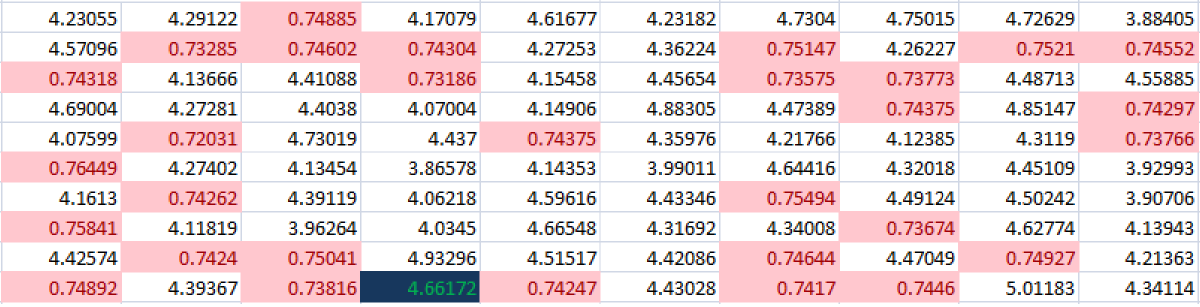
Fig. 6.13: 26 cells of a 10x10 cell array have flipped only because of lowered core voltage. No degradation effect was triggered so far.



4.22545	4.29087	0.74481	4.13702	4.61698	4.23133	4.72629	4.74959	4.72898	3.88037
4.52019	0.7383	0.7475	0.74163	4.26928	4.35813	0.7395	4.25207	0.73801	0.74177
0.73731	4.13355	4.40811	0.724	4.12555	4.4455	0.73391	0.73214	4.48642	4.56104
4.68027	4.27027	4.40189	4.06494	4.14275	4.88234	4.46525	0.73653	4.8691	0.73759
4.07578	0.73412	4.72509	4.43304	0.73483	4.34992	4.21561	4.11819	4.30199	0.73214
0.7446	4.28464	4.12597	3.86727	4.15281	3.98763	4.64721	4.32216	4.44911	3.92837
4.16222	0.74679	4.3907	4.06402	4.59496	4.49966	0.74835	4.4809	4.50285	3.9021
0.75784	4.13355	3.96667	4.0333	4.66243	4.31345	4.32308	0.73893	4.63758	4.14155
4.42617	0.74488	0.75324	4.93367	4.51488	4.41902	0.73504	4.47354	0.73971	4.21122
0.74587	4.40967	0.74085	0.73532	0.74255	4.44366	0.74382	0.74269	5.00589	4.34255

Fig. 6.14: 100 ms after end of stress, 5 cells more have flipped. These 5 new cells did not flip before. The other 26 cells are the same like pre-stress.

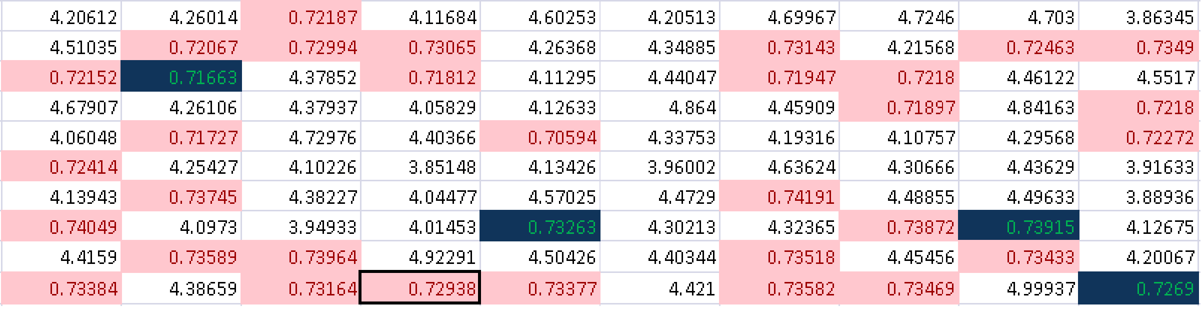
because of NBTI stress including the not yet recovered component. It is also important to note that the same 26 cells as before have flipped again, plus 5 new ones that did not flip before. Some hours after end of the 1.8 V stress, the same cells were analyzed again in Fig. 6.15. Now 30 cells have flipped instead of 31 directly after end of stress. This one cell is one of the 5 cells which flipped first time after end of stress. So recovery in this cell was big enough not to flip the cell anymore.



4.23055	4.29122	0.74885	4.17079	4.61677	4.23182	4.7304	4.75015	4.72629	3.88405
4.57096	0.73285	0.74602	0.74304	4.27253	4.36224	0.75147	4.26227	0.7521	0.74552
0.74318	4.13666	4.41088	0.73186	4.15458	4.45654	0.73575	0.73773	4.48713	4.55885
4.69004	4.27281	4.4038	4.07004	4.14906	4.88305	4.47389	0.74375	4.85147	0.74297
4.07599	0.72031	4.73019	4.437	0.74375	4.35976	4.21766	4.12385	4.3119	0.73766
0.76449	4.27402	4.13454	3.86578	4.14353	3.99011	4.64416	4.32018	4.45109	3.92993
4.1613	0.74262	4.39119	4.06218	4.59616	4.43346	0.75494	4.49124	4.50242	3.90706
0.75841	4.11819	3.96264	4.0345	4.66548	4.31692	4.34008	0.73674	4.62774	4.13943
4.42574	0.7424	0.75041	4.93296	4.51517	4.42086	0.74644	4.47049	0.74927	4.21363
0.74892	4.39367	0.73816	4.66172	0.74247	4.43028	0.7417	0.7446	5.01183	4.34114

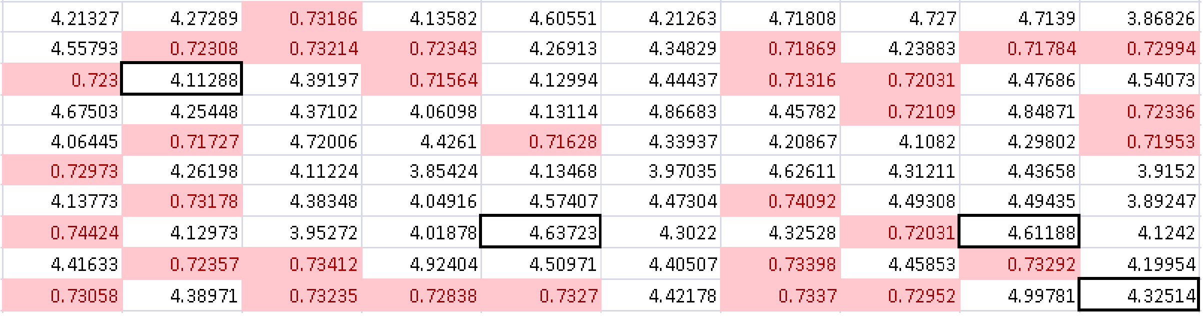
Fig. 6.15: After some hours of relaxation, 1 of the 5 NBTI-induced cell flips now does not flip anymore. The recovering component has reduced Vth shift so much that this cell does not flip anymore.

Now the cells were stressed again, this time with 2.2 V for 10^4 s. 100 ms after end of stress, the bitflip pattern was analyzed in Fig. 6.16. Now 35 cells have flipped, that is 9 more than without stress and 4 more than directly after 1.8 V stress. Exactly the same cells as pre stress and directly after 1.8 V stress have flipped, plus 4 new. The cell that recovered after 1.8 V stress flipped again. The next bitmap analaysis in Fig. 6.17 was taken long after 2.2 V stress. NBTI has recovered again, and only 31 cells flip now. This difference of 4 flips is from the group of cells that flipped new after 2.2 V stress. Finally, the bitmap analysis in Fig. 6.18 was taken the next day to check the results. The flip pattern did not change compared to Fig. 6.17. This proves that all these flips are stable results.



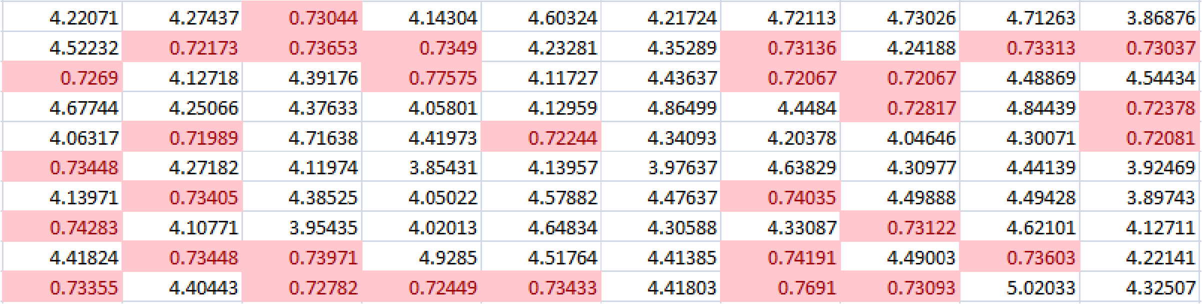
4.20612	4.26014	0.72187	4.11684	4.60253	4.20513	4.69967	4.7246	4.703	3.86345
4.51035	0.72067	0.72994	0.73065	4.26368	4.34885	0.73143	4.21568	0.72463	0.7349
0.72152	0.71663	4.37852	0.71812	4.11295	4.44047	0.71947	0.7218	4.46122	4.5517
4.67907	4.26106	4.37937	4.05829	4.12633	4.864	4.45909	0.71897	4.84163	0.7218
4.06048	0.71727	4.72976	4.40366	0.70594	4.33753	4.19316	4.10757	4.29568	0.72272
0.72414	4.25427	4.10226	3.85148	4.13426	3.96002	4.63624	4.30666	4.43629	3.91633
4.13943	0.73745	4.38227	4.04477	4.57025	4.4729	0.74191	4.48855	4.49633	3.88936
0.74049	4.0973	3.94933	4.01453	0.73263	4.30213	4.32365	0.73872	0.73915	4.12675
4.4159	0.73589	0.73964	4.92291	4.50426	4.40344	0.73518	4.45456	0.73433	4.20067
0.73384	4.38659	0.73164	0.72938	0.73377	4.421	0.73582	0.73469	4.99937	0.7269

Fig. 6.16: Directly after 2.2 V stress, 4 new cells have flipped. The cell that recovered after 1.8 V stress flipped again, giving a total of 35 flips.



4.21327	4.27289	0.73186	4.13582	4.60551	4.21263	4.71808	4.727	4.7139	3.86826
4.55793	0.72308	0.73214	0.72343	4.26913	4.34829	0.71869	4.23883	0.71784	0.72994
0.723	4.11288	4.39197	0.71564	4.12994	4.44437	0.71316	0.72031	4.47686	4.54073
4.67503	4.25448	4.37102	4.06098	4.13114	4.86683	4.45782	0.72109	4.84871	0.72336
4.06445	0.71727	4.72006	4.4261	0.71628	4.33937	4.20867	4.1082	4.29802	0.71953
0.72973	4.26198	4.11224	3.85424	4.13468	3.97035	4.62611	4.31211	4.43658	3.9152
4.13773	0.73178	4.38348	4.04916	4.57407	4.47304	0.74092	4.49308	4.49435	3.89247
0.74424	4.12973	3.95272	4.01878	4.63723	4.3022	4.32528	0.72031	4.61188	4.1242
4.41633	0.72357	0.73412	4.92404	4.50971	4.40507	0.73398	4.45853	0.73292	4.19954
0.73058	4.38971	0.73235	0.72838	0.7327	4.42178	0.7337	0.72952	4.99781	4.32514

Fig. 6.17: Long after 2.2 V stress, NBTI has recovered, and the number of flips got reduced back to 31.



4.22071	4.27437	0.73044	4.14304	4.60324	4.21724	4.72113	4.73026	4.71263	3.86876
4.52232	0.72173	0.73653	0.7349	4.23281	4.35289	0.73136	4.24188	0.73313	0.73037
0.7269	4.12718	4.39176	0.77575	4.11727	4.43637	0.72067	0.72067	4.48869	4.54434
4.67744	4.25066	4.37633	4.05801	4.12959	4.86499	4.4484	0.72817	4.84439	0.72378
4.06317	0.71989	4.71638	4.41973	0.72244	4.34093	4.20378	4.04646	4.30071	0.72081
0.73448	4.27182	4.11974	3.85431	4.13957	3.97637	4.63829	4.30977	4.44139	3.92469
4.13971	0.73405	4.38525	4.05022	4.57882	4.47637	0.74035	4.49888	4.49428	3.89743
0.74283	4.10771	3.95435	4.02013	4.64834	4.30588	4.33087	0.73122	4.62101	4.12711
4.41824	0.73448	0.73971	4.9285	4.51764	4.41385	0.74191	4.49003	0.73603	4.22141
0.73355	4.40443	0.72782	0.72449	0.73433	4.41803	0.7691	0.73093	5.02033	4.32507

Fig. 6.18: Final measurement long after the last stress phase. No changes happened compared to the last recovered measurement.

Altogether, it can be stated that the pattern of flipped bits totally matches the theory of SRAM stability and recovery behavior.

6.4.3 Measurements on single transistors

Reisinger et al. have developed a method to measure the V_{th} shift of single transistors ultra-fast after termination of NBTI stress conditions [49]. The schematic of this method is depicted in Fig. 6.19. The measurement results on big transistors were already shown in Fig. 3.11 on page 32. The inventor of this method now used his setup to measure SRAM-sized single transistors in order to compare the results.

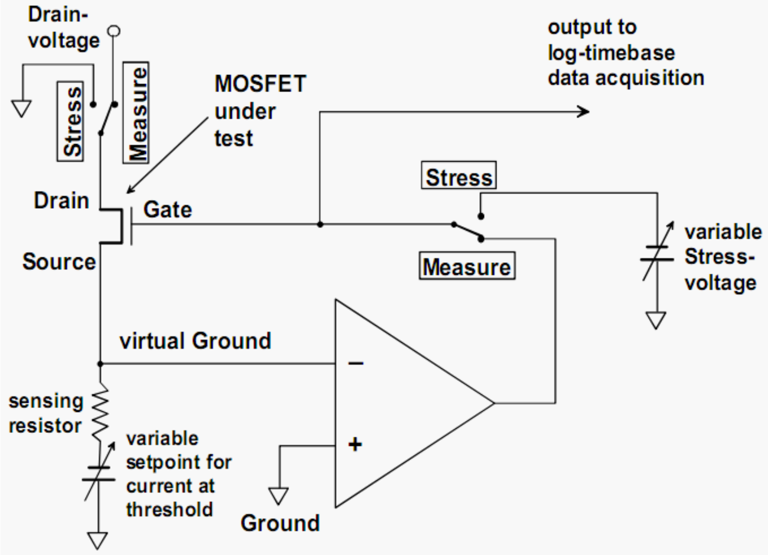


Fig. 6.19: Setup for single transistor characterization [49]. This setup was now used for SRAM sized transistors. The biasing conditions were set equal to the recovery conditions of the newly developed test method in order to compare the results. Original plot taken from [49].

9 single pMOS transistors of the original pullup size were measured directly after end of NBTI stress. The recovery conditions were chosen equal to the SRAM stability analysis method with 0.98 V. The result is shown in Fig. 6.20: the variation of the initial V_{th} shift as well as of the recovery behavior is huge. Inital V_{th} shift varies between 50 mV and 150 mV, and recovery is taking place either almost linearly or in some big steps. The mean value of these 9 transistor characteristics is also plotted and shows almost linear behavior over logarithmic time with a slope of approx. 10 mV per decade.

Fig. 6.21 shows the comparison of the mean value single transistor recovery (from plot 6.20) and the SRAM stability recovery (from plot 6.10). Both curves do not fit perfectly excellent, but they show very similar behavior. Some reasons why they do not fit excellently: 1. the mean value of only 9 single transistors is not fully representative. 2. the bias point of recovery (0.98 V) might be chosen too low so that the upper limit of stability degradation reaches the non-linear region. This means that the cells with strongest V_{th} shift in the pullup do not tend to flip more than with less V_{th} shift.

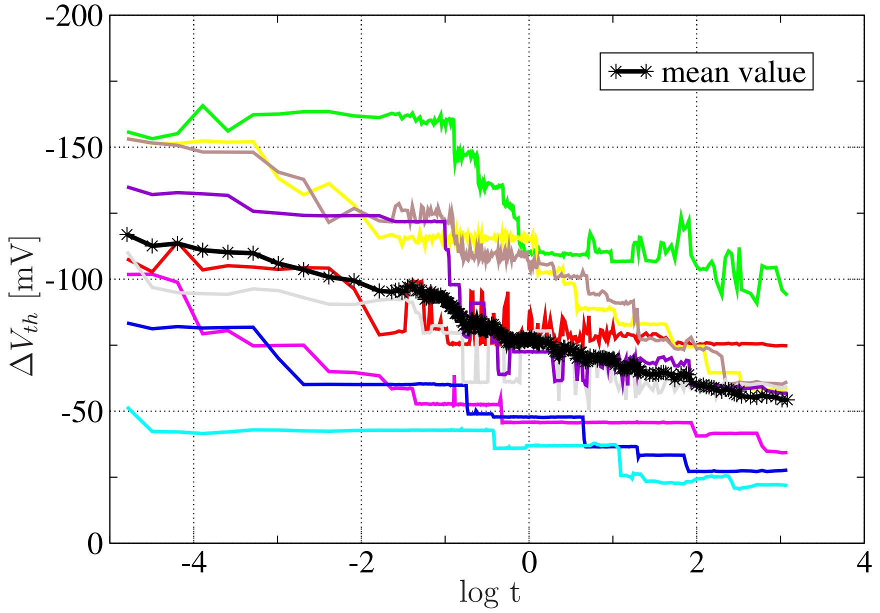


Fig. 6.20: Recovery of 9 different single SRAM size transistors shows strong variability. The mean value of these 9 transistors shows almost linear behavior over log t. Since SRAM cells are depending on single pullups, variability in stability is huge.

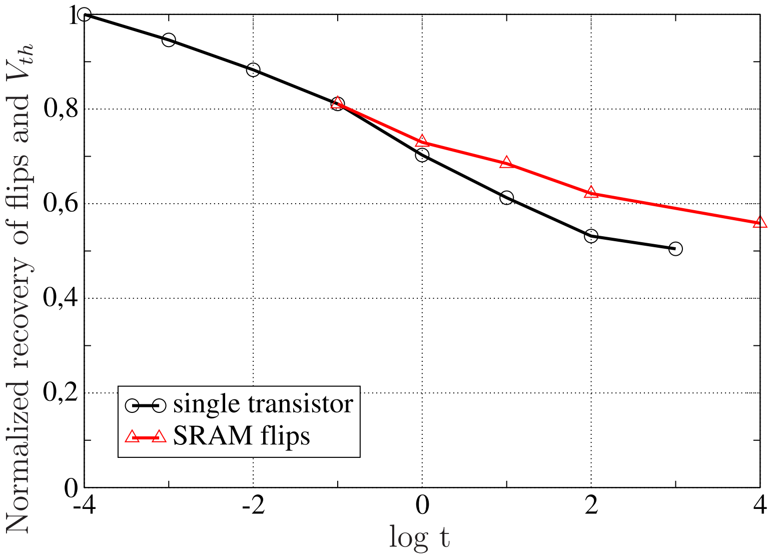


Fig. 6.21: Comparison of single transistor threshold voltage recovery and SRAM stability recovery after termination of NBTI stress conditions.

With slightly bigger transistors, the variations on single transistors are already much smaller (Fig. 6.22). This proves that SRAM products are NBTI worst case. They do not benefit from recovery and they suffer most from variability.

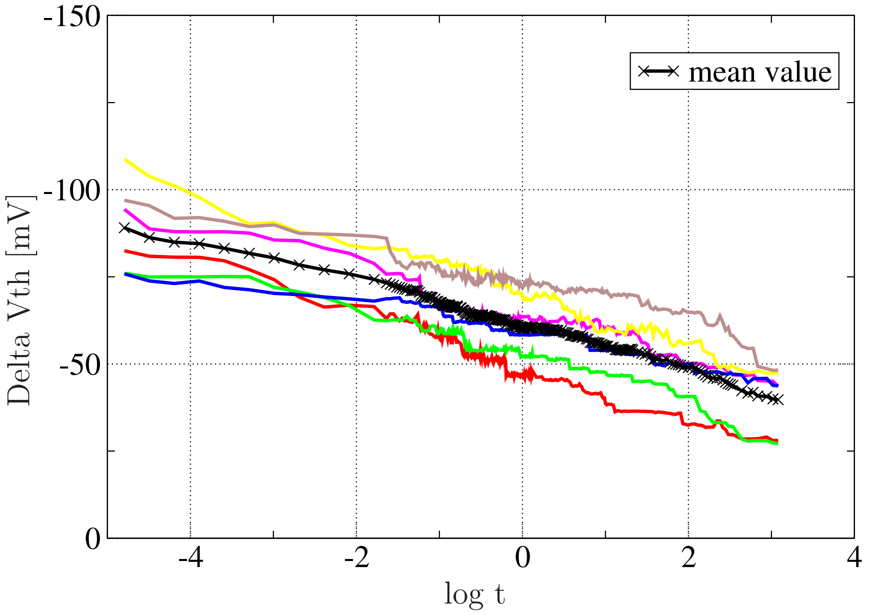


Fig. 6.22: Measurement of 2.35 times larger transistors (W/L=250/55nm) show less variability in recovery behavior.

6.5 Temperature dependency of NBTI recovery

To check the temperature dependency of the NBTI recovery effect, the static HM measurement from chapter 5 was used. The question was if the NBTI recovery effect can be frozen. Therefore, the SRAM array was NBTI stressed with 2.2 V for 10^4 s. While the increased stress voltage was still applied, temperature was already reduced from 125 °C to room temperature. Only when the temperature reached 25 °C, the stress voltage was reduced to nominal supply voltage. The flip distribution was then measured at room temperature. Then the SRAM array was kept for many days at 25 °C, and the same flip distribution was taken again. It had almost not changed, compare Fig. 6.23. Then, the temperature was increased again up to 125 °C again for 10^4 s. After cooling down again to room temperature, the flip curve was taken again. A clear step in degradation recovery was visible now. So degradation can be frozen. The recovery process is very strongly depending on temperature. At 125 °C, recovery is about a factor of 10^3 to 10^4 faster than at room temperature. So with directly cooling down the chip when the accelerating increased supply voltage is still applied, freezes the NBTI recovery. If then some 10^3 s

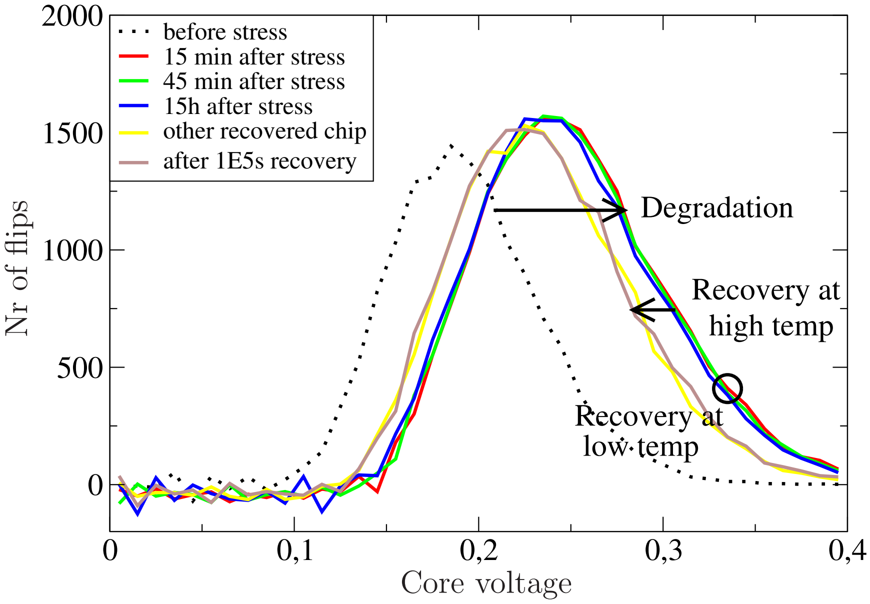


Fig. 6.23: Recovery can be frozen at room temperature. Waiting for 15 hours does not change the distribution considerably. But heating up to 125 °C accelerates recovery by a factor of 10^3 to 10^4, then after 10^4 s, a clear recovery step is visible.

later the chip is analyzed, this is about the degradation at about 1 s after end of stress. Unfortunately, 50% of the maximum ΔV_{th} is lost in the first second, so the worst case cannot be frozen for long time.

Keeping the circuit then for many days at room temperature does not change the result considerably. But after heating up to 125 °C, recovery is again accelerated.

The point is: with 125 °C, recovery is 10^3 to 10^4 times faster than at nominal room temperature. This means that 10^3 s of waiting is then equal to 1 s after stress at high temperature. Also in this short time range, ΔV_{th} has recovered about 50%.

6.6 Determination of Read Margin Distribution directly after end of Stress

After applying the new measurement approach and verifying its output data, the results will be used to determine the stability distribution of the complete SRAM array @ 1 ms after end of stress. This has never been done before. It is a combination of both stability analysis techniques in chapters 5 and 6 that were developed in this work. This is the recipe, depicted in Figs. 6.24 and 6.25.

1. Determine the pre-stress static Read Margin distribution like described in chapter 5 @ 125 °C on WL 1500 (dotted line in the plots).

2. Determine the post-stress static Read Margin distribution like described in chap-

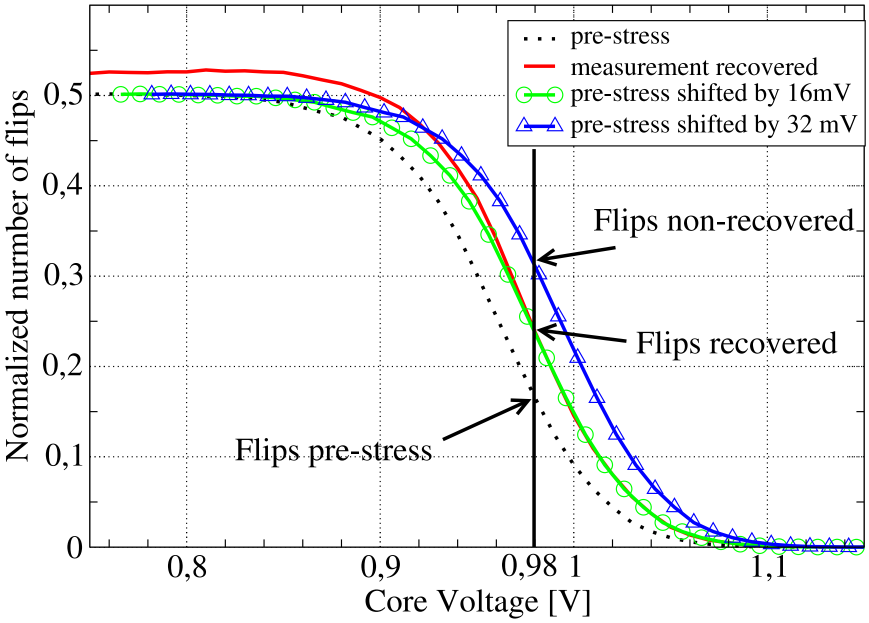


Fig. 6.24: The pre-stress flip curve is shifted by 16 mV and 32 mV to approximate the recovered flip curve and the flip curve 1 ms after end of stress. The 16 mV shifted curve exactly matches the recovered measurement.

ter 5 @ 125 °C on WL 1500. This technique is only able to detect the static component of NBTI and therefore does not provide the stability @1 ms after end of stress, but the recovered values @ many hours after end of stress. This measured curve exactly fits the flip curve that is right-shifted by 16 mV.

3. Determine the fast recovery characteristics like described in this chapter @ 125 °C after 10^4 s stress with 2.2 V. This is depicted in Fig. 6.10. For a core voltage of 0.98 V, it fits the number of recovered flips in the previous step. This is the link between the slow and the fast measurement.

4. Now the maximum number of flips is determined @1 ms after end of stress. The flip-curve is again right-shifted by another 16 mV. This results in the RM distribution at 1 ms after end of stress, which was not possible to measure so far.

With this method, also an estimation of the RM distribution @ 1 ns after end of stress could be done. Assuming further linearity of the Vth shift measurements down to 1 ns (which is only an assumption and cannot be proven here), another right-shift of the stability distribution by 16 mV can be done, resulting in 48 mV total shift. This would result in the worst-case scenario of reading an SRAM cell with the first GHz-clock, allowing only a recovery time of 1 ns. This shifted curve is also plotted in Fig. 6.25.

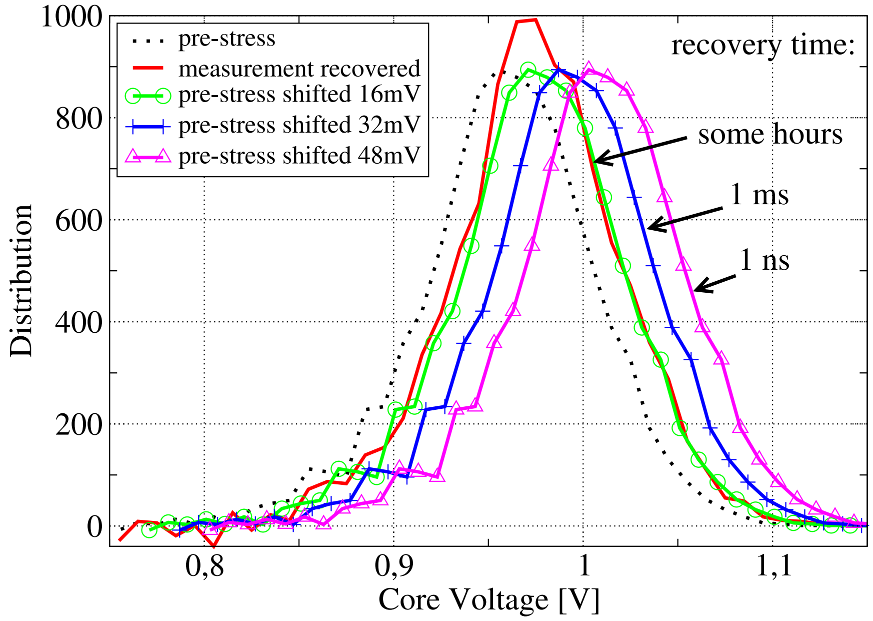


Fig. 6.25: Read stability distribution pre-stress, post-stress recovered and 1 ms after end of stress. The 16 mV shifted distribution perfectly matches the measured distribution. An estimation of the distribution 1 ns after end of stress is also done.

6.7 Conclusion

A new concept to measure fast-recovering NBTI directly on large SRAM arrays was presented. With a resolution of 1 ms, the stability of a 1 MBit 65 nm low power SRAM array was analyzed directly after end of stress. The recovering part of NBTI could be identified, which was not possible with former stability measurements. The stability decreasing effect of NBTI was continuously measured from 1 ms after end of stress until 10.000 s later, and then again after half a day with switched-off voltage.

The least cell stability, caused from worst NBTI degradation, appears directly after end of stress and starts to recover. This results in more stable SRAM cells over time, which was identified by an approx. 50% decrease of cell flips. This confirms and quantifies that the worst effect of NBTI on SRAM is reading a cell directly after keeping one value for a long time.

The new concept was confirmed by flip-curves without stress, bitmap flip-pattern analysis and single transistor measurements. After it turned out in Chapter 5 that hold stability is maximally degraded for SRAM arrays with adapted supply voltage, this Chapter now examined the worst case impact for SRAM arrays with constant supply voltage. Only read stability can then be dangerous, but the impact of the fast-recovering component has never been measured before. For the first time, this work now allowed a stability distribution measurement of the whole chip @ 1 ms after end of stress. An estimation of the worst-case scenario @ 1 ns after end of stress was also done.

Part III

Countermeasures

Chapter 7

Countermeasures

Abstract

Examining the impact of parametrical degradation mechanisms, it got clear that only NBTI has remarkable impact on the performance of 6T-SRAM core cells. Stability is reduced due to V_{th} shift induced by long NBTI stress during hold of data.

This chapter gives an overview on different known techniques to fight this impact. Simulations in 90 nm and 65 nm technology are used to quantify the advantages and disadvantages of these countermeasures. Based on these simulation results, the best candidates are chosen and a recommendation based on the priority of the user is given.

The most important results are reported in

E. Glocker, D. Schmitt-Landsiedel, and S. Drapatz: ”Countermeasures against NBTI degradation on 6T-SRAM cells”, Advances in Radio Science, 2011

7.1 The different levels of countermeasures

Fighting the impact of parametrical degradation mechanisms can be done on various levels. Starting at the bottom, the first approach can be done on device level. The goal would be to avoid shift of transistor parameters under the impact of degrading voltage- and current-conditions. Since BTI is caused by trapping of charges, research in technology is pushed heavily to at least minimize this effect.

Also on top of the design hierarchy, the impact of degradation can be fought on system level. This addresses ideas like redundancy (compare section 3.1) and Error Correction Codes (ECC). Generally, Error Correction Codes are based on the idea of adding some redundant information to the basic information so that, if some cells lose their state, this cannot only be detected, but also corrected. It is good for statistical failures with a low probability. But for NBTI induced failure there is a high possibility that a large number of cells, which all have seen equal stress conditions during long hold, fail all at the same time. Unfortunately, to cover this by redundancy would cause a huge overhead.

This work focuses on countermeasures on circuit level. In the following sections, two general approaches are covered: 1. fighthing the instability and 2. minimizing V_{th} shift.

7.2 Countermeasures against instability

As in Chapter 4 it turned out that NBTI degrades cell stability, the first class of countermeasures handled in this chapter are to generally improve cell stability. They are not especially developed for degradation scenarios, but were also the choice in the past to fight e.g. variations.

7.2.1 Guard Banding

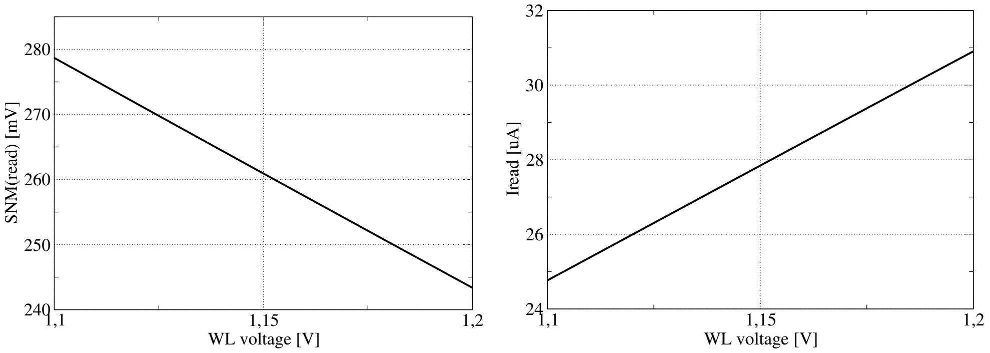
Degradation is especially dangerous for SRAM cells when the supply voltage is lowered. V_{DD} is decreased to V_{min} (or $V_{min,ret}$ during retention mode) to save leakage power. Unfortunately, lowering the supply voltage also decreases stability. The simplest way to increase stability and avoid loss of data is not to lower the supply voltage to the theoretical limit. Some 'Guard Band' is kept between the lowest possible voltage and the real choice of V_{DD}. The advantage is that this way is very simple and can be done without any alteration of the system. It is therefore commonly used in industry nowadays. Unfortunately, the optimum power saving due to lowering the supply voltage is not reached.

7.2.2 Lowering WL level

Reducing the wordline (WL) voltage during a read process reduces the impact of this read process on the memory nodes. The danger to increase the '0' memory node to the switching level of the opposite inverter is reduced even if the WL voltage is only reduced some ten millivolts. This increases read stability drastically. 80 mV less V_{WL} result in 10% more read stability, compare Fig. 7.1(a). The problem of this procedure is the dramatic loss of performance. Read current is reduced, so the read time for a cell increases: 50 mV less V_{WL} result in 10% less speed (compare Fig. 7.1(b)). So read stability can be increased at the expense of speed, while hold stability is not affected with this approach. Since speed is the killer feature of SRAM, one has to be aware of this important performance reduction. If the writeability must not be decreased as well, the periphery must provide two WL voltage levels: one normal WL level for writing and one reduced WL level for reading. This also means some periphery overhead.

7.2.3 Increasing core voltage - 'Core Boosting'

The Read Margin (RM) stability criterion is based on the reduction of the core voltage, because this degrades cell stability. Contrary to this, core voltage is now raised to increase stability, which is depicted in Fig. 7.2. This method can be done if the circuit provides a dual-V_{DD} wiring for core and periphery, since periphery is kept on nominal supply



(a) Read stability is increased drastically with lowering WL voltage during read. 80 mV less V_{WL} result in 10% more read stability.

(b) Speed is reduced drastically with lowering WL voltage during read. 50 mV less V_{WL} result in 10% less speed.

Fig. 7.1: Impact of WL boosting on stability and speed.

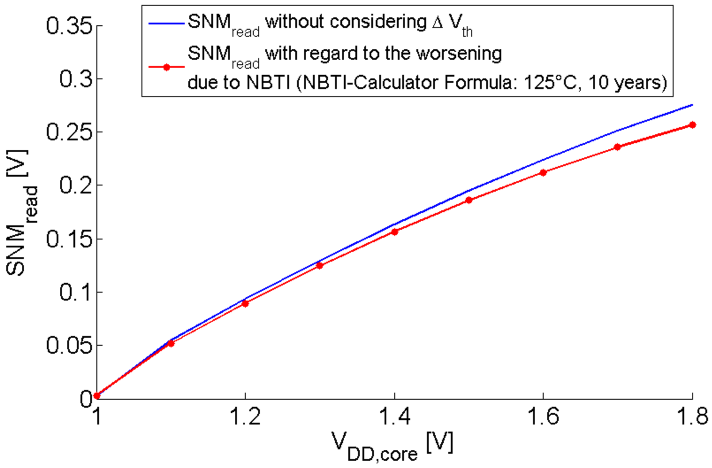


Fig. 7.2: Read stability is increased with core voltage

voltage.

Figs. 7.3 to 7.6 show the comparison of nominal core voltage with enhanced core voltage up to 1.5 V. Read stability is enhanced drastically by almost 80%, while hold stabiliy is increased slightly. Also speed is increased slightly, only writeability suffers from this method. So increased core voltage helps during read of data, while during write it reduces writeability.

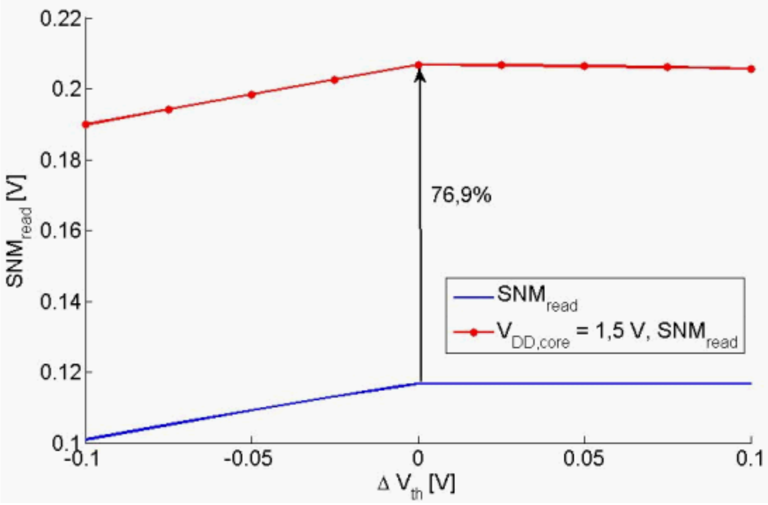


Fig. 7.3: Core boosting increases read stability drastically.

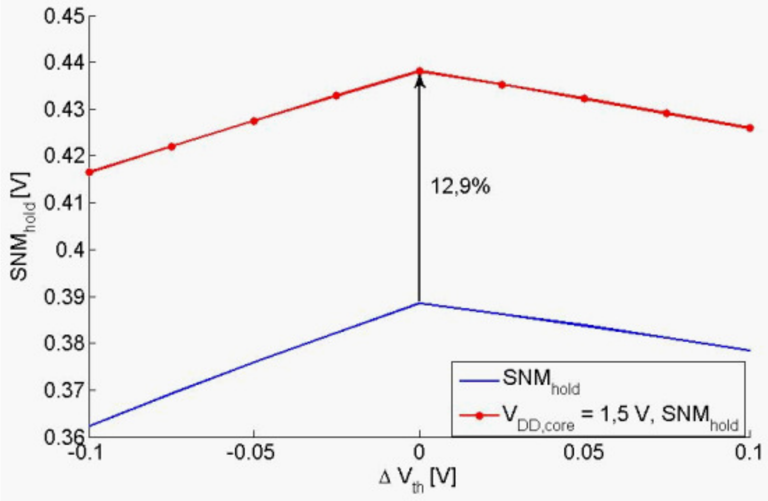


Fig. 7.4: Core boosting increases hold stability slightly.

Unfortunately, increased core voltage was also used in chapters 5 and 6 to accelerate aging. So if this method is used for long time during hold of data, it will let the circuit age faster, which of course is counterproductive (see Fig. 7.2, lower curve). Altogether, if this method is only applied during read of data, it helps to increase stability without the disadvantages of accelerated aging and decreased writeability. But it must be kept in

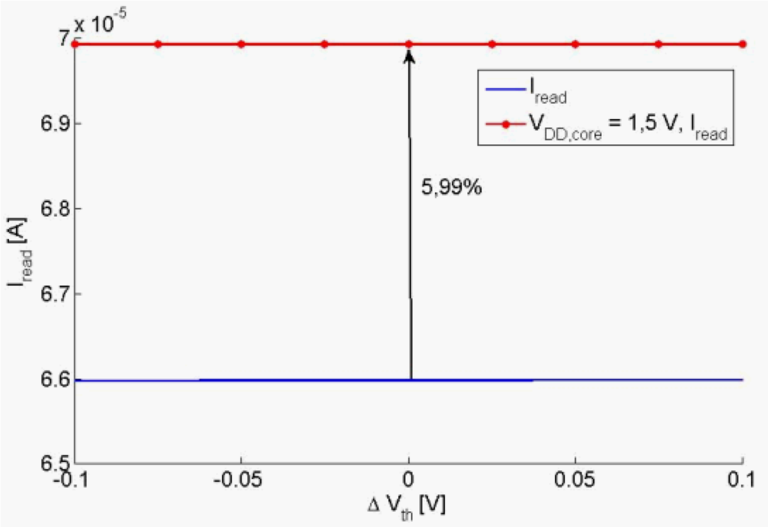


Fig. 7.5: Core boosting increases speed slightly.

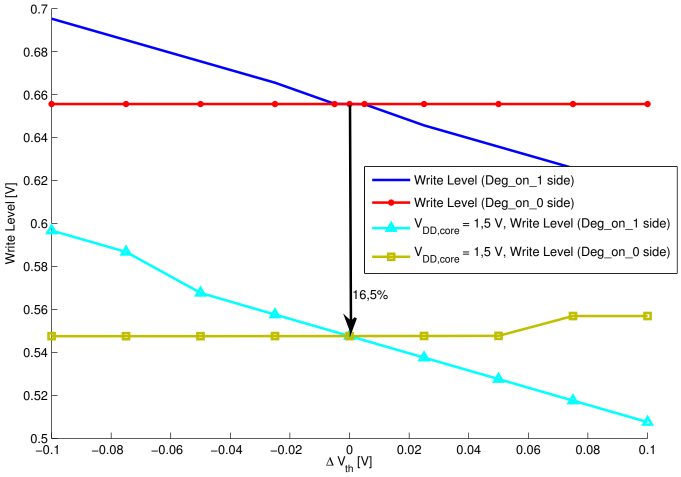


Fig. 7.6: Core boosting increases writeability slightly.

mind that increasing core voltage before a read cycle and lowering core voltage after a read cycle is a huge periphery and time overhead compared to nominal functionality.

7.2.4 8T core cell

Different proposals of core cells exist, from 4T over 6T, 7T and 8T to 10, 12 and even more transistors per cell. The elementary idea of more than 6 transistors per cell is to decouple hold, read and write modes. Since read stability is the main problem of parametric degradations, the 8T cell is the approach with the least circuit overhead: it decouples read from hold mode.
The schematic of the 8T SRAM core cell is depicted in Fig. 7.7. The basic idea is that the memory nodes are not attacked by the BLs during read, by decoupling the Read Bitline from the internal memory nodes. In the 8T approach, reading is done by only connecting a high-resistance gate node of an nFET to the memory node SB. So during read access, the memory nodes are not influenced by high loads, the voltages of the memory nodes do not change and therefore the cell is as stable as in hold mode [50] [51].

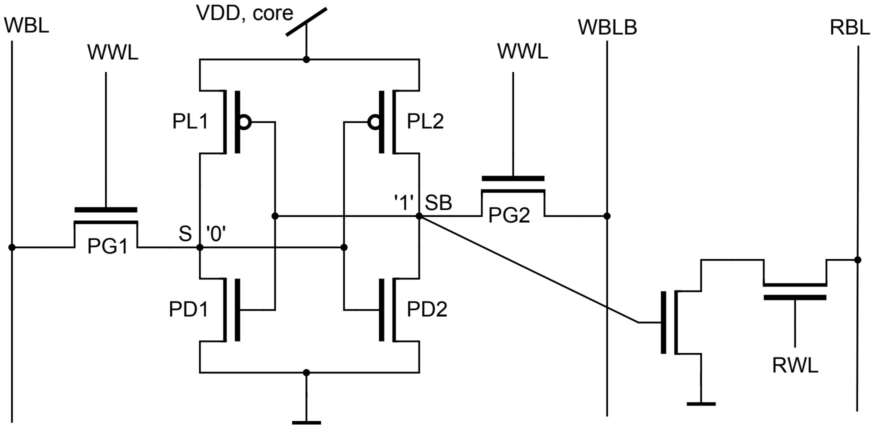


Fig. 7.7: 8T SRAM circuit with read decoupled from hold. The two bitlines WBL and WBLB are for write cycles only, the third bitline RBL is for read cycles only. Reading does not influence the memory nodes anymore.

It is important to note that during write operation, the 8T-SRAM array is disturbed by a parasitic read operation. The voltages of the wordline and bitlines of the 6 core transistors of the 8T cell for the half-selected columns in write operation are identical to the voltages of the 6T cell in read operation (Fig. 7.8).
As a result the half-selected columns experience the same loss of stability in the write operation as the 6T cell in read operation. To prevent this loss of stability and the potential data loss, an array architecture without a bitline multiplexer (MUX) is needed, where all cells connected to the same wordline are written at the same time. This MUX-free array architecture leads to further increase of the required area.
The main drawback of course is increased area consumption of about 30% because of 2 additional transistors, one WL and one BL compared to the 6T cell. But the advantage of getting rid of the read stability problem maybe worth it [52]. It lowers V_{min} directly,

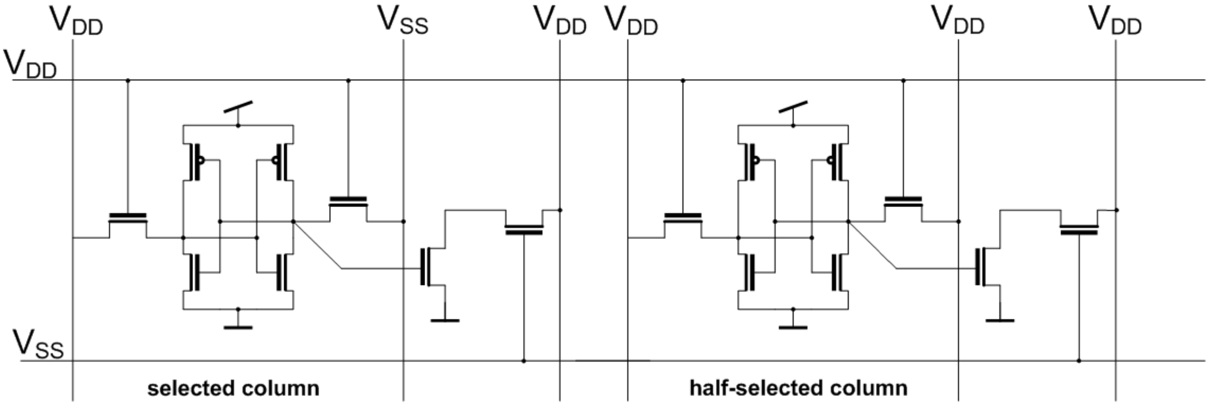


Fig. 7.8: 8T-SRAM memory cell write operation for the selected (left) and the half-selected (right) column. The half-selected column experiences the same loss of stability as the 6T cell in read operation, due to a parasitic read operation.

since this minimum operating voltage has to provide hold, write and read mode. Reading now is not less stable than hold, so this has a very positive effect on V_{min}.
This was the official reason why for the 45 nm Nehalem processor, Intel switched from 6T to 8T cells. Being able to reduce V_{min} decreases leakage power. But the investigations in this work suggest that also NBTI and PBTI (the 45 nm process is a high-κ metal gate process) might have had an influence on this decision.

7.3 Countermeasures against NBTI- and PBTI specific V_{th} drift

After some techniques to generally increase cell stability, this section is about some special approaches to minimize NBTI- and PBTI-specific V_{th} drift.
A perfect SRAM cell is a perfectly centered cell: it has the best trade-off between reading and writing. Both inverters must be seen individually: if one inverter is perfectly centered, this memory state will be fine. But if the opposite inverter is not centered at all, which may happen due to degradation and variations, the other memory state is strongly degraded. And since a cell must of course work in both directions and is therefore only as good as its worst state, one must keep both inverters in mind.
This is why the approaches in this section try to keep the degradation-induced threshold voltage shift as low as possible, which also minimizes the degradation.

7.3.1 Symmetrical hold

The worst case of NBTI for SRAM concluded from chapter 5 was 'reading after long time keeping data in one state'. 'Keeping data in one state' means that the pullup of one inverter is maximally degraded, while the opposite inverter does not see any stress at all. The idea is now to balance the load between the 2 possible memory states and therefore age each inverter for only lifetime/2 instead of aging one inverter for the whole lifetime. This is done by switching the cell from time to time. Fig. 3.7 showed the log-like behavior of V_{th} drift over linear time. This explains why after 5 years of aging, the V_{th} drift is already about 82% compared to the drift after 10 years. Only considering the static NBTI drift, the impact of this method is limited by this behavior, compare Table 7.1. Read stability can be improved by 1%, hold stability can be improved by 2.7% compared to the worst case of keeping one state for 10 years.

memory usage	ΔV_{th} after 10 years pullup1 / pullup2	SNM(read)	SNM(hold)
keep one memory state (DF=1)	53mV / 0mV	206mV	459mV
balanced memory states (DF=0.5)	48mV / 48mV	208mV	472mV

Table 7.1: Only considering mean value static NBTI, the worst case difference between keeping only one memory state and balancing the memory states is not big. Read stability can be improved by 1%, hold stability can be improved by 2.7% compared to the worst case of keeping one state for 10 years.

But the research on the dynamic component of NBTI in the last years re-raised the discussion of this approach and led to several publications on this topic. Considering the recovering NBTI component, the idea of 'symmetrical hold' not only balances V_{th} drift, but allows to get rid of 50% of the complete drift (which is the fraction of recovering NBTI). First, it was suggested to switch the cell once a day [53]. Then the toggle frequency was increased to 1000 s [54]. Last year, this approach was optimized on real-life register files [55]. Unfortunately, all these publications are based on simulations, so it is still not really clear how big is the impact of this approach.

7.3.2 Preferred state

The basic idea of this approach is the correlation between power-up state of an SRAM cell and its stability [56]. During power-up, a cell flips to one of both possible states. Contrary to former belief, this is not an arbitrary state, which changes from one power-up cycle to the other, but it is the 'preferred state' of the memory cell, which turned out to be the more stable one. Examinations in chapter 3 have shown that keeping an SRAM cell in one state decreases its stability. So powering up the cells makes it flip to its more stable side, and keeping it there makes it more centered. With this approach, stability distribution of an array could already be improved [56]. But this approach should also work in-field to re-center already degraded cells.

Like the symmetrical hold approach in the former section, it was not used so far for in-field re-centering of degraded cells, so the impact is not yet investigated.

7.3.3 Body Biasing

By applying a voltage V_{BS} to the substrate of the pMOS device, the threshold voltage can be adjusted [57]. The threshold voltage is defined by

$$V_{th,p} = V_{FB} + 2\phi_F - \gamma_p\sqrt{-2\phi_F + V_{BS}} - \frac{Q_{SS}}{C_{OX}} \tag{7.3.1}$$

with

$$\gamma_P = \frac{\sqrt{2eN_D\epsilon_0\epsilon_{Si}}}{C_{OX}} \tag{7.3.2}$$

Applying a positive voltage V_{BS} therefore shifts the threshold voltage to more positive values. Under normal SRAM conditions, V_{BS}=0 because source and bulk are both connected to V_{DD}. For V_{BS} >0, the threshold voltage $|V_{th}|$ can be decreased, so that an NBTI-degraded transistor can be re-adjusted to a not-aged one. For $|V_{BS}|$=0.7V, an NBTI degradation of ΔV_{th}=50 mV could be equalized.

In an SRAM array, all pMOS transistors have one common n-well, which means that V_{BS} must be adjusted equally to all pMOS transistors. This could only be avoided by a triple-well process, which is adding additional area consumption. Additionally, each pMOS transistor would need its own body bias control.

So this approach could generally be used to compensate the aging process. Unfortunately, this requires lots of effort. An on-chip monitor would need to detect the mean ΔV_{th} of the pMOS transistors and generate the according bulk-source voltage. Then the asymmetric behavior still is a problem. Additionally, one must be aware of the parasitic consequences, like diode-behavior because of p-n layers.

7.3.4 Burn-in

To prevent failure of circuits in the field, Burn-In is typically used to make those circuits fail before they go to the customer. This is achieved by applying higher supply voltage and temperature for a defined period. Burn-In can now be adopted as a NBTI countermeasure in the sense of pre-aging: If a specific ΔV_{th} can be artificially achieved directly after production, this ΔV_{th} will increase only a little bit over the operating lifetime, because it rises with a logarithmic dependence over time, compare Fig. 3.7 on page 30. The SRAM cell is centered without the Burn-In i.e. in matters of stability and writeability the cell has the best possible performance. After the Burn-In, the cell is not centered anymore.

To achieve both, best possible performance and the use of Burn-In, the cell has to be adapted. This means that the pullup devices must be designed with slightly increased width. After the burn-in step, which decreases the conductivity of the pullups, the artificially degraded transistors are equal to the not-aged transistors in the original design, which means that the cell is centered again. The exact Burn-In and Enhancement parameters must be chosen for each particular case with respect to the desired accuracy.

The 2 drawbacks of this approach are the increased area because of slighlty bigger pullup devices and the burn-in step, which creates cost.

7.3.5 Limited operating temperature

Limited temperature leads to improvement of the hold and read stability as well as lower threshold voltage drift. For T = -100 °C, the degradation formula would yield ΔV_{th}=0 mV, so the pMOS transistors would not be degraded by NBTI. However, this is far away from the operating conditions, where the measurements for fitting of the formula were performed, and this operating temperature is not practicable anyhow. But in general it has to be considered that at higher temperatures the cells are less stable. So the temperature can be limited, although this narrows down the SRAM operating range (compare Fig. 7.9). Additionally, the NBTI degradation is getting worse with higher temperature. This can be seen in Fig. 7.9, where a second plot was added to the nominal plot. It considers an NBTI-related ΔV_{th} (worst case, calculated for 10 years at 1.32 V) that occurs due to the increased temperature. For this voltage and time the decrease in stability by raised T is so large that the NBTI degradation, also increased with T, does not affect the result much. It is not possible to determine an optimal temperature limit: the lower the temperature, the higher the stability. So a suitable temperature limit for each particular case must be chosen.

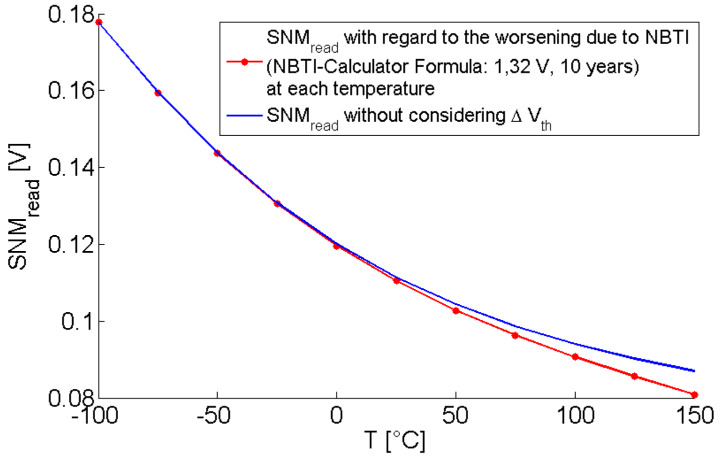


Fig. 7.9: Read stability over temperature considering both the simulation results at a constant ΔV_{th} and the potential rise of ΔV_{th} with V_{DD} The lower the temperature, the higher the stability.

7.4 Comparison of countermeasure techniques

Table 7.2 provides an overview on all techniques considered in the last sections.

Countermeasure	Effect	Advantage	Disadvantage
Guard banding	+	simple	wasting energy
Lowering WL level	+	little V_{WL} reduction needed	reduced speed
Core boosting	+	works well	strong effort
8T core cell	++	improve read to hold SNM	effort, area
Symmetrical hold	o		limited effect
Preferred state	o	can be used in-field	limited effect
Body biasing	++	eliminate aging	strong effort
Burn-in	+	works	cost during production
Limit temperature	o	simple	not suitable for real world

Table 7.2: Overview on countermeasure techniques listed in this work. The best 5 countermeasures are compared in the next section.

7.5 The best countermeasure techniques

After introduction of 9 different techniques to fight the impact of NBTI on 6T-SRAM core cells, five techniques will now be compared more in detail. Although body biasing got a '++' ranking for effectivity, it is not considered in the top five countermeasures. This is because this technique does not only increase stability like the other methods, but it can also decrease the other performances if the applied body voltage does not fit the requirements. These requirements would have to be adapted to each core cell and its individual age, which requires huge periphery overhead and some age monitoring technique. Core Boosting, WL Boosting, Burn-In, 8T SRAM Design und Guard Band have been chosen because they were the most reasonable approaches in terms of positive effect and practicability. They are compiled in the double-sided Table 7.3 for a 90 nm technology and compared to each other in the following. The best countermeasure technique shall be chosen afterwards.

The 8T-SRAM design is the most powerful NBTI countermeasure: The read stability problem of the 6T-SRAM cell does not occur anymore, this got reduced to hold stability values, which are not critical. The drawback of this best technique is that 2 additional transistors and periphery (required additional area is approx. 30%) are needed. The Burn-In ensures that ΔV_{th} only rises a little during operating time by an additional assembly step at higher T and V_{DD}. So the stability is approximately constant and the SRAM memory cell stays functional over operating time. The required area rises, because the pMOS transistors must be widened. For implementation of the Core and WL Boosting an additional voltage (plus additional periphery and wire connection) to supply the cell is necessary in each case. Both countermeasures increase the read stability. The Core Boosting also increases the hold stability. The WL Boosting lowers the writeability and read speed, while the Core Boosting leads to higher power consumption and bigger leakage current. An optimal $V_{DD,core}$ or V_{WL} respectively, cannot be chosen because of the approximately linear dependency of the metrics, i.e. no optimal point exists. Only a trade-off between stabiliy, speed and degradation can be done, so a suitable voltage must be chosen for each particular case. It is not recommended to implement the Core and the WL Boosting at the same time, because three voltages to supply the array would

	Performance parameters @ $\Delta V_{th} = 0mV$
Nominal point (V_{DD} = 1,2 V , T=25 °C)	SNM_{read} = 0,117 V SNM_{hold} = 0,388 V $I_{read} = 6,5982 \cdot 10^{-5}$ A Write Level = 0,656 V

Countermeasure	**Result**
Core Boosting ($V_{DD,core}$ = 1,5 V)	best $V_{DD,core}$ depends on applicaton $SNM_{read} = 0,207V$ (77% increase) $SNM_{hold} = 0,438V$ (13% increase) $I_{read} = 6,9932 \cdot 10^{-5}A$ (6% increase) $WriteLevel = 0,548V$ (17% decrease)
WL Boosting (V_{WL} = 0,7 V)	opt. V_{WL} depends on application $SNM_{read} = 0,286V$ (144% increase) SNM_{hold} no change $I\_\,read = 2,4 \cdot 10^{-5}A$ (64% decrease) $WriteLevel = 0,1479V$ (77% decrease)
Burn-In	pMOS 5nm bigger ⇒ approx. 1% more area ”‘Burn-In”’ depends on application, e.g. for usage at $V_{DD} = 1,2V$, $T = 25C$: e.g. 5s at 2 V, 175C
Guard Band	opt. V_{DD}: depends on application e.g. $\Delta V_{th} = -50mV$: $V_{min} \approx 0,8V$
8T SRAM Design	approx. 30 % more area than 6T, no decreased SNM like 6T SNM$_{read}$ approx. 3x bigger than 6T

Table 7.3: Best countermeasures against NBTI degradation

Performance parameters @ $\Delta V_{th} = -100mV$	
Decrease by 0,016 V Decrease by 0,026 V Decrease by $7 \cdot 10^{-9}A$ Increase by 0,039 V	

positive aspects	**negative aspects**
Stability improvement; simple to implement	higher $V_{DD,core} \Rightarrow$ energy/leakage $\uparrow$; dual voltage supply for $V_{DD,core}$ and $V_{WL}/V_{BL(B)}$ necessary Decreased Writeability

Read stability improvement;
simple to implement;
V_{WL} smaller $\Rightarrow$ energy/leakage $\downarrow$
	Decrease of
	writeability/read access speed
	dual supply voltage of
	V_{WL} and $V_{DD,core}/V_{BL(B)}$ necessary

ΔV_{th} decreases over lifetime almost no more $\Rightarrow$ stability almost constant	center cell before burn-in $\Rightarrow$ more area consumption additional production step
actual method because only V_{DD} must be adapted	higher $V_{DD} \Rightarrow$ energy/leakage $\uparrow$; Lower stability (more degradation)
Separation of hold & read solves read stability problem; growing profit with smaller technology	MUX-free architecture necessary; more area consumption

Table 7.4: Best countermeasures against NBTI degradation

be needed. The Guard Band is the easiest countermeasure in terms of implementation: Only the minimal V_{DD} must be guarded to be above $V_{DD,min}$. This narrows down the operating range of the SRAM and increases power consumption and leakage current. A suitable minimal V_{DD} must be chosen for each particular use case.

7.6 Conclusion

In this chapter countermeasures against NBTI degradation that most impact the stability of the cell were presented. With regard to simulation results and practicability the best candidates were chosen and compared to each other. Since it is not possible to generally determine the optimal countermeasure, the best technique, depending on the individual preferences in memory design, is recommended as follows:

1. Reliability and leakage, but not area is the first priority: 8T SRAM Design
If the increased area consumption of 30% is acceptable, the 8T-SRAM core cell design provides maximum impact for reliability and leakage. The reduced stability during read access can be improved to the much stronger hold stability. Therefore, V_{min} can be reduced drastically, which also lowers leakage current. This design also helps with variability challenges, and both together might be the reason, why it is already used in modern high-κ SRAM designs.

2. Reliability is important, and additional periphery and wiring is acceptable: WL or Core Boosting
Wordline- and core boosting are techniques that have been already used in the past to fight variations. So the required additional effort for wiring and periphery are often already implemented, this is why this was not considered as a show-stopper for these approaches. Both techniques not only impact variability, but also improve read stability and therefore help a lot against parametric variation and degradation. While the impact of both approaches to read stability is high, the drawbacks of slower read access (WL boosting) and read cycle overhead because of charging and discharging the pMOS-wells (core boosting) must be traded off.

3. No design-change of the cell array is preferred: Guard Band
Guard band is the choice if the reliability aspect is not given high priority. This might be the case if e.g. the array is exclusively used in a low temperature environment or operated with drastically reduced supply voltage during a major fraction of lifetime. There is no additional effort to implement this technique, this is why it is the actually used countermeasure for most products in industry nowadays.

Chapter 8

Conclusion and Outlook

In this work, the impact of the four parametric degradation mechanisms 'Negative Bias Temperature Instability' (NBTI), 'Positive Bias Temperature Instability' (PBTI), 'Hot Carrier Injection' (HCI) and 'Non-conducting Hot Carrier Injection' (NCHCI or Off-state stress) on 6-transistor SRAM core cell arrays was investigated. The periphery of SRAM macros was not in the focus of this work and might show different behavior [58].

In actual 65 nm technology with conventional SiO_2 gate oxide, only NBTI has remarkable impact on 6T SRAM core cells. PBTI is only active in high-κ dielectrics, HCI is not active long enough during normal SRAM operations, and NCHCI does not show enough degradation in this technology.
Generally, **NBTI has degrading impact only on stability**, while write-ability and speed are not negatively influenced. As a rule of thumb in this technology, 100 mV NBTI-induced threshold voltage drift cause approx. 10% SNM read stability loss. This ΔV_{th} value is the simulated expectation value of maximum real-life aging: holding one memory state for 10 years with 125 °C and 110% V_{DD}.
The worst case of NBTI to SRAM is keeping one memory state for long time (in the range of years). This leads to maximum V_{th} drift of one pullup transistor, while the other pullup remains unstressed. This again leads to maximum instability of the cell in this memory state.

NBTI shows strong variability and recovery, adding to the manufacturing variability. Variations are extremely large because of the tiny SRAM cells, following Pelgrom's law of doubling the standard deviation with dividing the transistor area by the factor of 4. With not yet fully understood variability and recovery behavior, no sufficiently accurate models are available to simulate the real impact of NBTI to SRAM stability. This is why measurements were necessary to characterize product-like SRAM arrays to examine a real world scenario. **Therefore in this work, unconventional new stability measurement techniques were developed.** Contrary to state-of-the-art methods, they are faster, do not need dedicated test chips which do not represent mass product design, do not need highly accurate V-I measurements and therefore can be used in-field in products with the only precondition of dual-V_{DD} power routing. Using these new techniques, the impact of variability and recovery on the NBTI-affected stability was directly measured.

The new measurement techniques are all based on the flipping behaviour of SRAM cells when the core voltage is lowered. Since the best known stability metric SNM needs access to the memory nodes which is not possible in SRAM products, the stability metric 'Read Margin' was adapted to quickly measure hold and read stability distributions on a 1 MBit SRAM array. Simulations have shown that this is a linear metric to characterize stability. After development, implementation and execution of measurements, the following results were achieved.

Read Stability of recovered NBTI drift:
The manufacturing variability alone shows Gaussian-distributed RM values between 170 mV and 430 mV, which is a difference of 260 mV. Accelerated 1.5 years of operation with 1.2 V @ 125 °C (approx. 40 mV V_{th} drift) shift these values by approx. 17 mV to again Gaussian distribution with almost the same standard deviation. **N mV of threshold voltage drift decreases the Normal-distributed read stability by N/2 mV.** Compared to the huge impact of manufacturing variability of 260 mV, these additional 17 mV are not extremely critical.

Hold Stability of recovered NBTI drift:
The same metric was applied to Hold Stability, which represents the retention case for keeping data. Normal-distributed values between 870 mV and 1150 mV were measured, which means that this is only a critical case when the supply voltage is lowered during retention phases ('adaptive supply voltage'). It showed that the minimum retention voltage $V_{min,ret}$ is raised very strongly by NBTI, i.e. **N mV pMOS NBTI induced V_{th} shift leads to $2 \cdot N$ mV $V_{min,ret}$ rise.** This is because the distribution was not only right-shifted like in read case, but also got wider and not exactly Gaussian Distributed anymore. **Strong $V_{min,ret}$ raise is the worst result of NBTI to SRAM in real-life products, when the supply voltage is dynamically adapted.** On the other hand, if the supply voltage is reduced for long periods during retention mode, there will be no strong NBTI degradation, which limits the impact drastically.

Stability directly after end of stress:
With conventional measurement techniques, only the impact of recovered NBTI could be analyzed so far. But the V_{th} shift is maximum directly after end of stress and starts to recover with extremely short time constants. Therefore, ΔV_{th} has lost about 50% after only 1 s of recovery time. It is therefore extremely critical to know the impact of (almost) non-recovered NBTI. Especially on SRAM cells, the impact of the recovering component is especially high due to long hold times that do not allow recovery like in typical clocked logic. **In this work, for the first time, the impact of this recovering component of NBTI was directly measured on SRAM arrays at 1 ms after end of stress.** It could be shown that the number of destructive read events 1 ms after end of stress is approx. double compared to the recovered values some hours

later. Therefore it was proved that **the worst case of NBTI to real-life SRAM products without dynamically adapted core voltage is reading a cell directly after long time keeping data.** This does not allow recovery and attacks the cell in the least stable state. **Additionally, this work now allowed to characterize the stability distribution of the whole chip @ 1 ms after end of stress.**

Finally, known countermeasures against variability, instability and V_{th} drift were examined. **Comparing countermeasures against BTI, technically the 8T core cell is the best candidate.** Although it needs about 30% more area and still has the problem of increased retention voltage, it solves the problem of read stability and V_{min}. This is because read and hold are separated, so that reading is not a strong attack to cell stability anymore. Together with reduced V_{min}, this might be the reason why 8T core cells are already chosen in Intel's 45 nm high-κ 'Nehalem' processor.

The overall result of all these simulations and measurements is:
SRAM products are NBTI worst case: they do not benefit from recovery but suffer most from variability. However, compared to the strong manufacturing variability and temperature dependency, parametric degradation is not a show-stopper in non-high-κ technologies.

A view into the future shows that the next sub-65 nm generation of SRAM cells including high-κ metal gate stacks is much more vulnerable. Then, not only NBTI, but also PBTI and maybe NCHCI will impact 6T-SRAM core cells. PBTI is active in high-κ dielectrics, and NCHCI might have enough degradation in this technology. Only HCI still is not active long enough during normal SRAM operations. This means that **not only the pullup will be degraded, but also the pulldown, which has about double the impact on SRAM performance compared to pullup**. Simulations have proved that **both effects are adding their impacts and result in a dramatic loss of stability** (approx. 30% SNM loss @ 100 mV NBTI- and PBTI-induced V_{th} Drift, respectively) and also of speed. Additionally, both NBTI and PBTI show recovering behavior, which does not happen in SRAM: directly after long hold, the read cycle is the worst attack to SRAM stability.
Altogether, this means that both transistors will experience strong V_{th} shifts directly after long hold, which will exceed the mentioned 30% stability penalty. Variations are not yet included in this scenario, but will again increase with smaller transistor dimensions.

So if from technological side both BTI effects cannot be minimized, this might be a real threat for 6T-SRAM core cell arrays. It could lead to failures in-field after about 1 year of usage. Fortunately, several countermeasures can be considered in the design phase to still produce reliable SRAM arrays.

Appendix

Appendix A

Determination of Static Noise Margin SNM and Read N-Curve

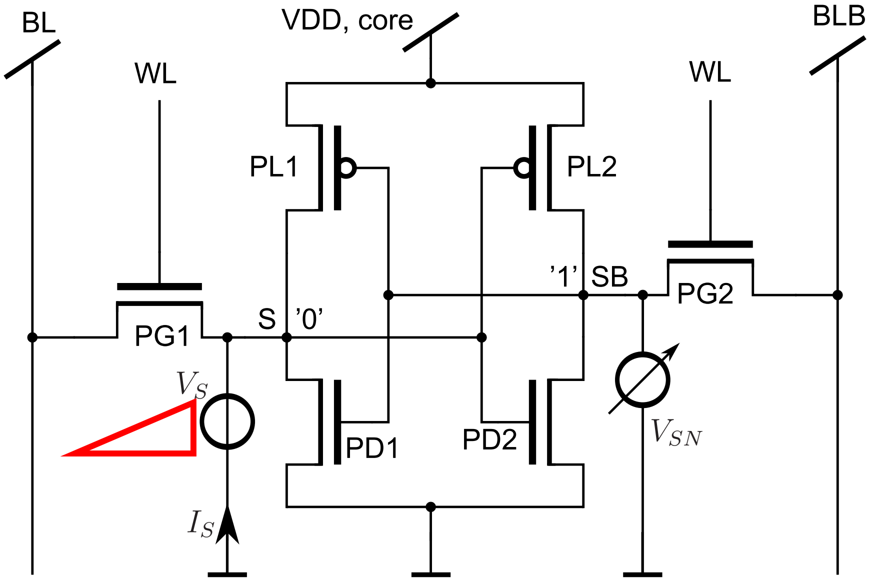


Fig. A.1: Setup for SNM(read), SNM(hold) and Read N-Curve [13] [12]. Both stability metrics require access to the memory nodes.

Mode	WL	Meaning
SNM(read)	1	Raise V_S, measure V_{SN} + vice versa
SNM(hold)	0	Raise V_S, measure V_{SN} + vice versa
Read N-Curve	1	Raise V_S, measure I_S + vice versa

Table A.1: Different modes of using the test setup in Fig. A.1

N-Curve for hold case is not supported, because the current is too small.

Appendix B

Determination of Read Margin RM

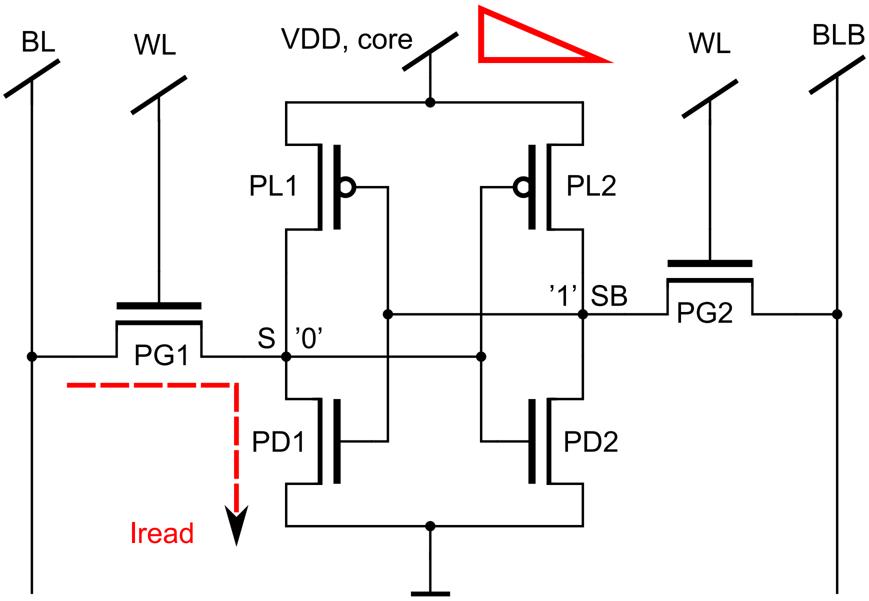


Fig. B.1: Setup for Read Margin

Lower the core voltage and measure the read current on the '0' memory side. At which core voltage does the bitline current drop? The difference between this voltage and the nominal voltage is called the Read Margin.

Appendix C

Determination of Write Level or Write-Trip Point

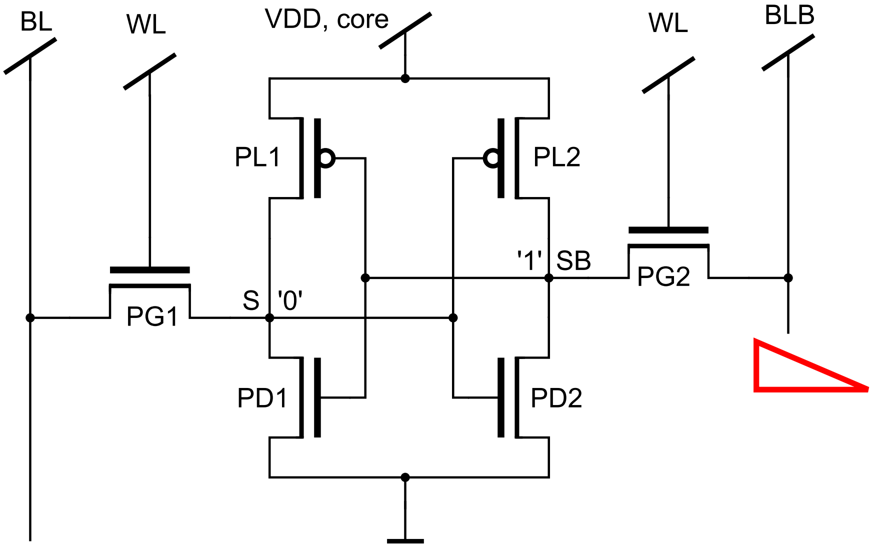


Fig. C.1: Setup for Write Level

Starting from read conditions, lower the BL voltage on the ’1’ memory side. At which BL voltage does the cell flip?

Appendix D

Determination of Write N-Curve

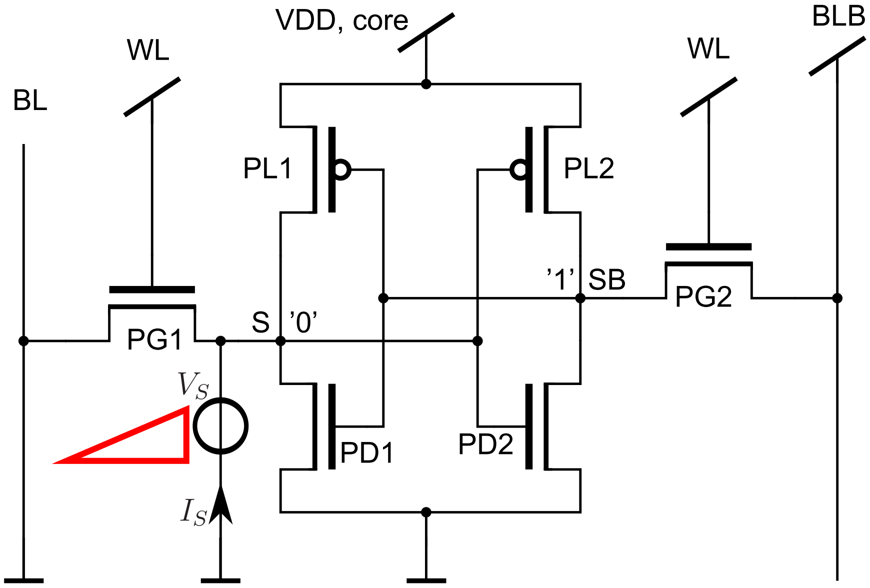


Fig. D.1: Setup for Write N-Curve

Measurement setup is similar to Read N-curve, but the BL on '0' memory side is connected to ground.

Appendix E

Determination of Read Current I_{read}

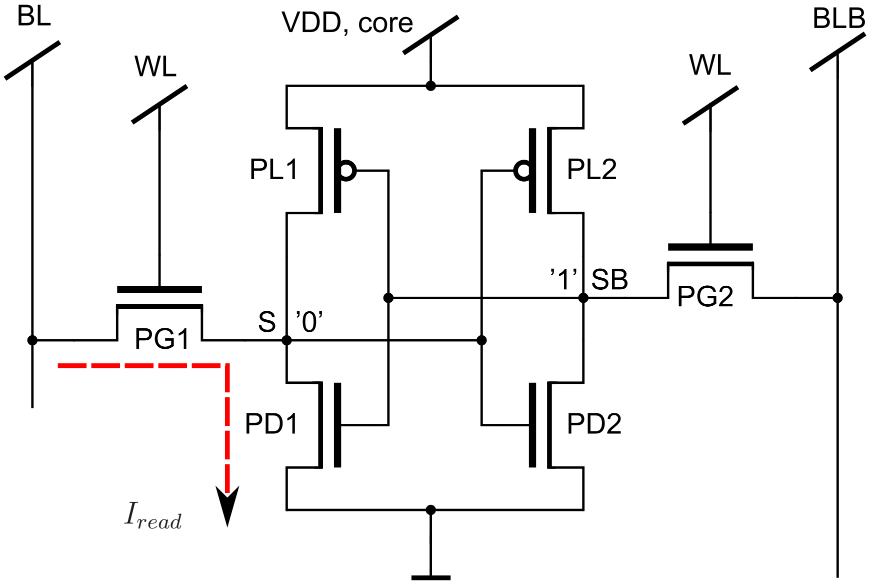


Fig. E.1: Setup for read current

Static read current flowing on '0' memory side.

Appendix F

List of Symbols and Abbreviations

6T	6 transistor
ADM	Access Disturb Margin
bit	Binary digit
BL, BLB	Bit Line, Bit Line Bar (aka $\overline{BL}$)
BTI	Bias Temperature Instability
CDF	Cumulative Density Function
CMOS	Complementary Metal Oxide Semiconductor
d	distance
DF	Duty Factor
DRAM	Dynamic Random Access Memory
ϵ	Dielectric constant, sometimes κ is used
ECC	Error Correction Codes
EUV	Extreme Ultra Violet
FET	Field Effect Transistor
FIT	Failure In Time - equivalent to 1 error per 10^9 hours of device operation
FoM	Figure of Merit
g_m	Small signal transconductance (or mutual conductance)
GB	Giga Byte = 2^{30} Byte = 1,073,741,824 Byte
HCI	Hot Carrier Injection
HM	Hold Margin
I	Current
IC	Integrated Circuit
I_{crit}	Critical Current
κ	Dielectric constant, sometimes ϵ is used
kB	Kilo Byte = 2^{10} Byte = 1,024 Byte
kBit	Kilo Bit = 2^{10} Bit = 1,024 Bit
L	Length of a device
LDD	Lightly Doped Drain
μ	Mean value, expectation value
μ_0	Carrier mobility
MB	Mega Byte = 2^{20} Byte = 1,048,576 Byte

MBit	Mega Bit = 2^{20} Bit = 1,048,576 Bit
MC	Monte Carlo
MOSFET	Metal Oxide Semiconductor FET
MTBF	Mean time between failure; 1 year MTBF is equal to approx. 114 FIT
MuGFET	Multi Gate FET
MUX	Multiplexer
NBTI	Negative Bias Temperature Instability
NCHCI	Non conducting Hot Carrier Injection (aka Off-State Stress)
PBTI	Positive Bias Temperature Instability
PC	Personal Computer
PDF	Probability Density Function
PVT	Process, Voltage, Temperature
RDF	Random Dopant Fluctuation
r_{out}	Small signal output resistance
RM	Read Margin
S, SB	Memory nodes S, S Bar (aka $\overline{S}$)
σ	Standard Deviation
SINM	Static Current Noise Margin
SNM	Static Noise Margin
SNM(hold)	Static Noise Margin for retention case
SNM(read)	Static Noise Margin for read case
SoC	System on chip
SOI	Silicon On Insulator
SRAM	Static Random Access Memory
SVNM	Static Voltage Noise Margin
t	time
V	Voltage
V_{ds}	Drain-Source voltage
V_{gs}	Gate-Source voltage
V_{min}	Minimum supply voltage that provides full SRAM functionality
$V_{min,ret}$	Minimum supply voltage that only provides safe hold (WL=0)
V_{th}	Threshold voltage
VTC	Voltage Transfer Characteristic
W	Width of a device
WL	Word Line or Write Level (depending on situation)

Appendix G

Publications by the author

Articles in Journals

S. Drapatz, K. Hofmann, G. Georgakos, and D. Schmitt-Landsiedel:
"A method to analyze the impact of fast-recovering NBTI degradation on the stability of large-scale SRAM arrays"
Solid-State Electronics, 2011 (invited)

S. Drapatz, G. Georgakos, and D. Schmitt-Landsiedel:
"Impact of negative and positive bias temperature stress on 6T-SRAM cells"
Advances in Radio Science, 2009

E. Glocker, D. Schmitt-Landsiedel, and S. Drapatz:
"Countermeasures against NBTI degradation on 6T SRAM memory cells"
Advances in Radio Science, 2011

Articles in the Proceedings of International Conferences

S. Drapatz, T. Fischer, K. Hofmann, E. Amirante, P. Huber, M. Ostermayr, G. Georgakos, and D. Schmitt-Landsiedel:
"Fast stability analysis of large-scale SRAM arrays and the impact of NBTI degradation"
European Solid-State Cicruits Conference ESSCIRC, 2009
and
European Solid-State Device Research Conference ESSDERC, 2009 (Joint Session)

S. Drapatz, K. Hofmann, G. Georgakos, and D. Schmitt-Landsiedel:
"Impact of fast-recovering NBTI degradation on stability of large-scale SRAM arrays"
European Solid-State Device Research Conference ESSDERC, 2010

Talks on International Conferences

"Fast stability analysis of large-scale SRAM arrays and the impact of NBTI degradation"
European Solid-State Cicruits Conference ESSCIRC, Athens, Greece, September 2009

"Impact of fast-recovering NBTI degradation on stability of large-scale SRAM arrays"
European Solid-State Device Research Conference ESSDERC, Seville, Spain, September 2010

Talks on National Conferences and Workshops

"NBTI+PBTI: Show-stopper for SRAM?"
TU München Reliability Workshop, Munich, July 2008

"Einfluss von Negative und Positive Bias Temperature Stress auf 6T-SRAM Zellen"
Kleinheubacher Tagung, Miltenberg, September 2009

"Impact of constant and fast recovering NBTI on stability of large-scale SRAM arrays"
HONEY Project Präsentation, Munich, February 2010

"Fast stability analysis of large-scale SRAM arrays and the impact of NBTI degradation"
TU München Reliability Workshop, Munich, May 2010

"Zuverlässigkeitscharakterisierung von embedded SRAM und der Einfluss von NBTI"
VDE-ITG Fachgruppe 8.5.6fWLR, Erfurt, May 2010 (invited)

"Zuverlässigkeitscharakterisierung an SRAM Speicherarrays"
Sommer-Kolloquium TU Darmstadt "Zuverlässigkeit - vom elektronischen Bauelement bis zur Schaltung", Darmstadt, July 2010 (invited)

"Fast stability analysis of large-scale SRAM arrays and the impact of NBTI degradation"
TU München Reliability Workshop, Munich, February 2011

List of Figures

Bibliography

[1] *ITRS. International Technology Roadmap for Semiconductors*, 2009.

[2] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, 1965.

[3] E. Grossar, *Techology-aware design of SRAM memory circuits.* Leuven-Heverlee, Belgium: Katholieke Universiteit Leuven, 2007.

[4] A. Pavlov and M. Sachdev, *CMOS SRAM Circuit Design and Parametric Test in Nano-Scaled Technologies.* Netherlands: Springer, 2008.

[5] D. A. Pelgrom, M. and A. Welbers, "Matching properties of mos transistors," *Solid-State Circuits, IEEE Journal of*, vol. 24, 1989.

[6] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital integrated circuits - A design perspective.* Pearson Education International, 2003.

[7] K. Itoh, *VLSI Memory Chip Design.* Berlin, Germany: Springer, 2001.

[8] K. Zhang, *Embedded Memories for Nano-Scale VLSIs.* Germany: Springer, 2009.

[9] K. Itoh, "Low-voltage memories for power-aware systems," in *Low Power Electronics and Design, 2002. ISLPED '02. Proceedings of the 2002 International Symposium on*, 2002, pp. 1 – 6.

[10] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective.* Addison Wesley.

[11] Hill, "Noise margin and noise immunity in logical circuits," *Microelectronics*, vol. 1, pp. 16–21, 1968.

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

[13] C. Wann, R. Wong, D. Frank, R. Mann, S.-B. Ko, P. Croce, D. Lea, D. Hoyniak, Y.-M. Lee, J. Toomey, M. Weybright, and J. Sudijono, "Sram cell design for stability methodology," in *VLSI Technology, 2005. (VLSI-TSA-Tech). 2005 IEEE VLSI-TSA International Symposium on*, april 2005, pp. 21 – 22.

[14] F. Bauer, *SRAM core-cell concepts for embedded SoC memories.* Aachen, Germany: Shaker, 2011.

[15] Z. Guo, A. Carlson, L.-T. Pang, K. T. Duong, T.-J. K. Liu, and B. Nikolic, "Large-scale sram variability characterization in 45 nm cmos," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 11, pp. 3174–3192, Nov. 2009.

[16] Z. Guo, A. Carlson, L.-T. Pang, K. Duong, T.-J. K. Liu, and B. Nikolic, "Large-scale read/write margin measurement in 45nm cmos sram arrays," in *VLSI Circuits, 2008 IEEE Symposium on*, 2008, pp. 42 –43.

[17] Z. Guo, "Large-scale variability characterization and robust design techniques for nanoscale sram," in *VLSI Circuits, 2008 IEEE Symposium on*, 2009.

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

[19] O. Hallberg, B. Eriksson, R. Francis, R. Hjortendal, L.-I. Lindberg, and B. Saevstroem, "Hardware reliability assurance and field experience in a telecom environment," *Quality and Reliability Engineering International*, vol. 10, no. 3, pp. 195–200, 1994. [Online]. Available: http://dx.doi.org/10.1002/qre.4680100310

[20] W. Gerling, *Zuverlaessigkeit mikroelektronischer Bauelemente*, 2005.

[21] T. Fischer, A. Olbrich, G. Georgakos, B. Lemaitre, and D. Schmitt-Landsiedel, "Impact of process variations and long term degradation on 6t-sram cells," *Advances in Radio Science*, vol. 5, pp. 321–325, 2007. [Online]. Available: http://www.adv-radio-sci.net/5/321/2007/

[22] J. McPherson, *Reliability Physics and Engineering.* Berlin: Springer, 2010.

[23] A. W. Strong, E. Y. Wu, R.-P. Vollertsen, J. Su, G. L. Rosa, S. E. Rauch, and T. D. Sullivan, "Reliability wearout mechanisms in advanced cmos technologies." Wiley, 2009.

[24] H. F. D.Widmann, H. Mader, *Technolgie integrierter Schaltungen.* Springer, 1996.

[25] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *J. Appl. Phys.*, vol. 94, no. 1, pp. 1–18, Jul. 2003. [Online]. Available: http://link.aip.org/link/?JAP/94/1/1

[26] J. Stathis and S. Zafar, "The negative bias temperature instability in mos devices: A review," *Microelectronics and Reliability*, vol. 46, no. 2-4, pp. 270 – 286, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0026271405003008

[27] F. Chouard, C. Werner, D. Schmitt-Landsiedel, and M. Fulde, “A test concept for circuit level aging demonstrated by a differential amplifier,” in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, may 2010, pp. 826 –829.

[28] V. Huard, C. Parthasarathy, A. Bravaix, C. Guerin, and E. Pion, “Cmos device design-in reliability approach in advanced nodes,” in *Reliability Physics Symposium, 2009 IEEE International*, April 2009, pp. 624–633.

[29] T. Fischer, E. Amirante, P. Huber, T. Nirschl, A. Olbrich, M. Ostermayr, and D. Schmitt-Landsiedel, “Analysis of read current and write trip voltage variability from a 1-mb sram test structure,” *Semiconductor Manufacturing, IEEE Transactions on*, vol. 21, no. 4, pp. 534 –541, 2008.

[30] W. Heinrigs, H. Reisinger, W. Gustin, and C. Schlunder, “Consideration of recovery effects during nbti measurements for accurate lifetime predictions of state-of-the-art pmosfets,” in *Reliability physics symposium, 2007. proceedings. 45th annual. ieee international*, April 2007, pp. 288–292.

[31] T. Grasser, H. Reisinger, P. Wagner, F. Schanovsky, W. Goes, and B. Kaczer, “The time dependent defect spectroscopy (tdds) for the characterization of the bias temperature instability,” in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, may 2010, pp. 16 –25.

[32] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, and M. Nelhiebel, “A two-stage model for negative bias temperature instability,” in *Reliability Physics Symposium, 2009 IEEE International*, april 2009, pp. 33 –44.

[33] H. Reisinger, T. Grasser, W. Gustin, and C. Schlunder, “The statistical analysis of individual defects constituting nbti and its implications for modeling dc- and ac-stress,” in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, may 2010, pp. 7 –15.

[34] S. Zafar, A. Kumar, E. Gusev, and E. Cartier, “Threshold voltage instabilities in high- kappa; gate dielectric stacks,” *Device and Materials Reliability, IEEE Transactions on*, vol. 5, no. 1, pp. 45 – 64, march 2005.

[35] S. Zafar, Y. Kim, V. Narayanan, C. Cabral, V. Paruchuri, B. Doris, J. Stathis, A. Callegari, and M. Chudzik, “A comparative study of nbti and pbti (charge trapping) in sio2/hfo2 stacks with fusi, tin, re gates,” in *VLSI Technology, 2006. Digest of Technical Papers. 2006 Symposium on*, 0-0 2006, pp. 23 –25.

[36] K. Zhao, J. Stathis, A. Kerber, and E. Cartier, “Pbti relaxation dynamics after ac vs. dc stress in high-k/metal gate stacks,” in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, may 2010, pp. 50 –54.

[37] E. Takeda and N. Suzuki, “An empirical model for device degradation due to hot-carrier injection,” *Electron Device Letters, IEEE*, vol. 4, no. 4, pp. 111 – 113, apr 1983.

[38] B. Yan, Q. Fan, J. Bernstein, J. Qin, and J. Dai, "Reliability simulation and circuit-failure analysis in analog and mixed-signal applications," *Device and Materials Reliability, IEEE Transactions on*, vol. 9, no. 3, pp. 339 –347, sept. 2009.

[39] T. Furukawa, D. Turner, S. Mittl, M. Maloney, R. Serafin, W. Clark, J. Bialas, L. Longenbach, and J. Howard, "Accelerated gate-oxide breakdown in mixed-voltage i/o circuits," in *Reliability Physics Symposium, 1997. 35th Annual Proceedings., IEEE International*, apr 1997, pp. 169 –173.

[40] S. Holzhauser and A. Narr, "Off-state-degradation of 170 nm and 140 nm buried ldd pmosfets with different halo implants," in *Integrated Reliability Workshop Final Report, 2000 IEEE International*, 2000, pp. 158–160.

[41] A. Muehlhoff, "An extrapolation model for lifetime prediction for off-state degradation of mos-fets," 2001.

[42] K. Hofmann, S. Holzhauser, and C. Kuo, "A comprehensive analysis of nfet degradation due to off-state stress," in *Integrated Reliability Workshop Final Report, 2004 IEEE International*, Oct. 2004, pp. 94–98.

[43] A. Bansal, R. Rao, J.-J. Kim, S. Zafar, J. Stathis, and C.-T. Chuang, "Impact of nbti and pbti in sram bit-cells: Relative sensitivities and guidelines for application-specific target stability/performance," in *Reliability Physics Symposium, 2009 IEEE International*, April 2009, pp. 745–749.

[44] K. Antreich, H. Graeb, and C. Wieser, "Circuit analysis and optimization driven by worst-case distances," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 13, no. 1, pp. 57 –71, jan 1994.

[45] Z. Luo, A. Steegen, M. Eller, R. Mann, C. Baiocco, P. Nguyen, L. Kim, M. Hoinkis, V. Ku, V. Klee, F. Jamin, P. Wrschka, P. Shafer, W. Lin, S. Fang, A. Ajmera, W. Tan, D. Park, R. Mo, J. Lian, D. Vietzke, C. Coppock, A. Vayshenker, T. Hook, V. Chan, K. Kim, A. Cowley, S. Kim, E. Kaltalioglu, B. Zhang, S. Marokkey, Y. Lin, K. Lee, H. Zhu, M. Weybright, R. Rengarajan, J. Ku, T. Schiml, J. Sudijono, I. Yang, and C. Wann, "High performance and low power transistors integrated in 65nm bulk cmos technology," in *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, 2004.

[46] M. Sharifkhani and M. Sachdev, "Sram cell stability: A dynamic perspective," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 2, pp. 609 –619, 2009.

[47] G. Chen, K. Chuah, M. Li, D. Chan, C. Ang, J. Zheng, Y. Jin, and D. Kwong, "Dynamic nbti of pmos transistors and its impact on device lifetime," in *Reliability Physics Symposium Proceedings, 2003. 41st Annual. 2003 IEEE International*, 2003.

[48] K. Hofmann, H. Reisinger, K. Ermisch, C. Schlunder, W. Gustin, T. Pompl, G. Georgakos, K. Arnim, J. Hatsch, T. Kodytek, T. Baumann, and C. Pacha, "Highly accurate product-level aging monitoring in 40nm cmos," in *VLSI Technology (VLSIT), 2010 Symposium on*, 2010, pp. 27 –28.

[49] H. Reisinger, O. Blank, W. Heinrigs, A. Muhlhoff, W. Gustin, and C. Schlunder, "Analysis of nbti degradation- and recovery-behavior based on ultra fast vt-measurements," in *Reliability Physics Symposium Proceedings, 2006. 44th Annual., IEEE International*, 2006, pp. 448 –453.

[50] F. Bauer, G. Georgakos, and D. Schmitt-Landsiedel, "A design space comparison of 6t and 8t sram core-cells," in *Integrated Circuit and System Design*, 2009.

[51] L. Chang, D. Fried, J. Hergenrother, J. Sleight, R. Dennard, R. Montoye, L. Sekaric, S. McNab, A. Topol, C. Adams, K. Guarini, and W. Haensch, "Stable sram cell design for the 32 nm node and beyond," in *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on*, june 2005, pp. 128 – 129.

[52] S. Krishnappa and H. Mahmoodi, "Comparative bti reliability analysis of sram cell designs in nano-scale cmos technology," in *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, march 2011, pp. 1 –6.

[53] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Impact of nbti on sram read stability and design for reliability," in *In International Symposium on Quality Electronic Design*, 2006, pp. 27–29.

[54] Y. Kunitake, T. Sato, and H. Yasuura, "Signal probability control for relieving nbti in sram cells," in *Quality Electronic Design (ISQED), 2010 11th International Symposium on*, march 2010, pp. 660 –666.

[55] S. Kothawade, K. Chakraborty, and S. Roy, "Analysis and mitigation of nbti aging in register file: An end-to-end approach," in *Quality Electronic Design (ISQED), 2011 12th International Symposium on*, march 2011, pp. 1 –7.

[56] J. Wang, S. Nalam, Z. Qi, R. Mann, M. Stan, and B. Calhoun, "Improving sram vmin and yield by using variation-aware bti stress," in *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, sept. 2010, pp. 1 –4.

[57] K. von Arnim, E. Borinski, P. Seegebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, and C. Pacha, "Efficiency of body biasing in 90-nm cmos for low-power digital circuits," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 7, pp. 1549 – 1556, july 2005.

[58] V. Huard, R. Chevallier, C. Parthasarathy, A. Mishra, N. Ruiz-Amador, F. Persin, V. Robert, A. Chimeno, E. Pion, N. Planes, D. Ney, F. Cacho, N. Kapoor, V. Kulshrestha, S. Chopra, and N. Vialle, "Managing sram reliability from bitcell to library level," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, may 2010, pp. 655 –664.

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit als Akademischer Rat am Lehrstuhl für Technische Elektronik der Technischen Universität München in Zusammenarbeit mit Infineon Technologies AG München.

Mein besonderer Dank gilt Frau Prof. Dr. Doris Schmitt-Landsiedel, die mir die Möglichkeit zur Durchführung dieser Arbeit gab. Mich begeisterte neben ihrer fachlichen und sozialen Kompetenz speziell ihr Führungsstil, der mir viel Verantwortung gab und viele Freiheiten ließ. Dadurch konnte ich einige sehr lehrreiche Jahre an ihrem Lehrstuhl verbringen, die ich auf keinen Fall missen möchte.

Herrn Prof. Dr. Walter Stechele danke ich für sein Interesse an der Thematik, der Erstellung des Zweitgutachtens und der damit verbundenen Mühen.

Meinem Ansprechpartner bei Infineon AG, Herrn Georg Georgakos, möchte ich für seine hervorragende Betreuung seitens der Industrie danken. Seine umfassende fachliche Expertise zusammen mit seiner menschlichen Kompetenz waren gleichermaßen angenehm wie förderlich. Ich erinnere mich gerne an viele fachlich weiterführende Diskussionen auf dem Gebiet der statischen Speicher mit meinem Ansprechpartner in der Fachabteilung, Herrn Dr. Karl Hofmann. Herrn Dr. Hans Reisinger möchte ich für die messtechnische Beratung und Durchführung der Vergleichsmessungen an Einzeltransistoren danken.

Einen großen Anteil am Erfolg dieser Arbeit haben alle meinen Kolleginnen und Kollegen am Lehrstuhl für Technische Elektronik. Viele gemeinsame Diskussionen unter den Doktoranden bei der Kaffeerunde und immer vorhandene Hilfsbereitschaft und Teamgeist haben die Jahre am Lehrstuhl zu einem tollen Lebensabschnitt gemacht. Die wissenschaftliche Kerntruppe während meiner Zeit bestand dabei aus den Herren Dr. Markus Becherer, Stephan Breitkreutz, Florian Chouard, Rainer Emling, Thomas Fischer, Dr. Christoph Friederich, Josef Kiermaier, Dr. Werner Kraus, Michael Lüders, Shailesh More, Dr. Philip Teichmann, Dr. Marcus Weis und Martin Wirnshofer. Herrn Dr. Uwe Penning, zu dessen Vorlesung 'Technische Elektrizitätslehre' ich die Übung bestreiten durfte, danke ich für sein herzliches Engagement und die Begeisterung, mit der er hinter den Kulissen des Lehrstuhls für einen reibungslosen Ablauf sorgte.

Abschliessend möchte ich meinen Eltern danken, die mir nicht nur eine hervorragende Ausbildung ermöglicht haben, sondern auch stets an mich glaubten und mir mit wertvollen Tips und dem Gefühl einer intakten Familie den Rücken gestärkt haben.