Design for variability of read circuitries for resistive memories

Salmen Mraihi

université PARIS-SACLAY

NNT : 2018SACLS218

Thèse de doctorat

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Prise en compte de la variabilité dans l'étude et la conception de circuits de lecture pour mémoires résistives

Thèse de doctorat de l'Université Paris-Saclay

préparée à l'Université Paris-Sud

École doctorale n°575 Electrical, Optical, Bio: physics and engineering (EOBE)

Spécialité Électronique et optoélectronique, nano- et microtechnologies

Thèse présentée et soutenue à Orsay, le 26 septembre 2018, par

M. Salmen MRAIHI

Composition du Jury :

M. Lionel Torres Professeur, Université Montpellier 2	Président
M. Jean-Michel Portal Professeur, Université d'Aix-Marseille	Rapporteur
Mme Lorena Anghel Professeur, Telecom ParisTech	Rapporteur
M. Hervé Mathias Maître de conférences, Université Paris-Sud	Examinateur
M. Guillaume Prenat Chargé de recherche, CEA Grenoble	Examinateur
M. Jacques-Olivier Klein Professeur, Université Paris-Sud	Directeur de thèse

Université Paris-Saclay

Espace Technologique / Immeuble Discovery

Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Acknowledgements

I would like to thank all the people that helped and supported me in both my thesis research and my well-being during these exhausting past four years.

First, I would like to thank the company 'Intel Mobile Communications' and the laboratory 'Centre de Nanosciences et Nanotechnologies (C2N)' for welcoming me and bringing all the support necessary to achieve my thesis in the best conditions.

I feel a lot of gratefulness towards my co-advisor, Elmehdi Boujamaa, for his patience and all his valued recommendations. I felt encouraged by his trust and was inspired by him to stay motivated.
I am also thankful for Cyrille Dray, for his helpful supervision and beneficial discussions during my thesis.
I would like also to express a particular acknowledgement to my thesis supervisor Jacques-Olivier Klein, for his precious support, especially during the so difficult step of the writing of this manuscript.

I am very grateful to the members of my thesis committee. Many thanks to Jean-Michel Portal and Lorena Anghel for reviewing and writing reports for this manuscript. I also thank Lionel Torres, Guillaume Prenat and Hervé Mathias for agreeing to examine this work.

I also thank all the members of the Intel team, and my managers (Christophe Chanussot, Axel Jahnke, Jean-Christophe Vial), who have been involved, from near and far, in achieving my thesis goals and helped me advancing with my work. Thanks to Hassan Lachkar, for his kindness and his friendly support. Thanks, among others, to Yohann Abiven, Axel Martinez, He Fan, Vincent Gouin, Loubna Hannati, Prasanth Pushparasah, Sophie Drean, Gregory Chaix, Yann Tellier, Martina Benvenuti, Linda Hirel, Ettore Amirante, Fabien Elleboode, Jean-Louis Scotto di Quaquero, Remy Girard, Laura Flogiarini, Sharath Seshadri …

My sincere thanks to all the members of IntegNano group in the C2N who welcomed me during my fourth year: Gefei Wang, Xiaochao Zhou, Xiaoxuan Zhao, Boyu Zhang, Nicolas Vernier, Sylvain Eimer, Damien Querlioz, Dafiné Ravelosona, Giancarlo Faini. I also thank Prof. Eric Cassan, Mme Sophie Bouchoule and Mme Laurence Stephen from Doctoral School for their assistance in the registration and thesis defense procedures.

A special gratitude for my friends (Taha, Ismail, Abdelaziz, Jean-Christophe, Steve, Jean Michel, Tariq …) and family (Dorsaf, Kawther, Dhafer, Aymen, …) for their continuous encouragement.

Finally, I allow myself terminating in french with acknowledgement to my parents : un grand merci à vous, chers papa et maman, pour ce soutien sans faille qui m'a permis d'affronter les très nombreux obstacles auxquels j'ai dû faire face ces dernières années.

Table of contents

List of Figures

List of Tables

List of Equations

Chapter I: General Introduction

1.1 Motivation

Nowadays, the microelectronic world lives a transition period linked to a miniaturization context. The economic Moore's law [1], dictating that the number of transistors in integrated circuits (IC) doubles every eighteen months, reaches its limitations. Indeed, technologic nodes downscaling leads to increased devices variability and huge manufacturing costs [2][3]. From year to year, new advances at both technologic and architecture levels allow it to be maintained (see Figure 1 [4]) or show that it still has years ahead in a new era called 'More than Moore' [5][6]. Nevertheless, researchers and industrialists are anticipating the changes to come, developing emerging technologies, particularly in the sector of memories. Phase-Change Memories (PCM), Resistive RAM (Resistive Random-Access Memory) and STT-MRAM (Spin-Transfer-Torque Magnetic Random-Access Memory) appear to be strong candidates to replace existing memories for some applications. Indeed, these non-volatile technologies feature better endurance and lower latency than flash memories, higher density compared to SRAM (Static Random-Access Memories) and process compatibility with current Complementary Metal-Oxide-Semiconductor (CMOS) technology [7]. However, many challenges have to be overcome in order to make these technologies profitable and competitive for commercialization. Among them, reducing their too high writing currents, or increasing their read yield made difficult by an ever-increasing variability with downscaling [8][9][10]. This thesis offers a detailed study of the variability problematic applied to the design of nanoscale resistive memories. It also lists and compares resistive memories read yield enhancement solutions, and finally proposes a circuit architecture and its implementation that improves significantly resistive memories read reliability.

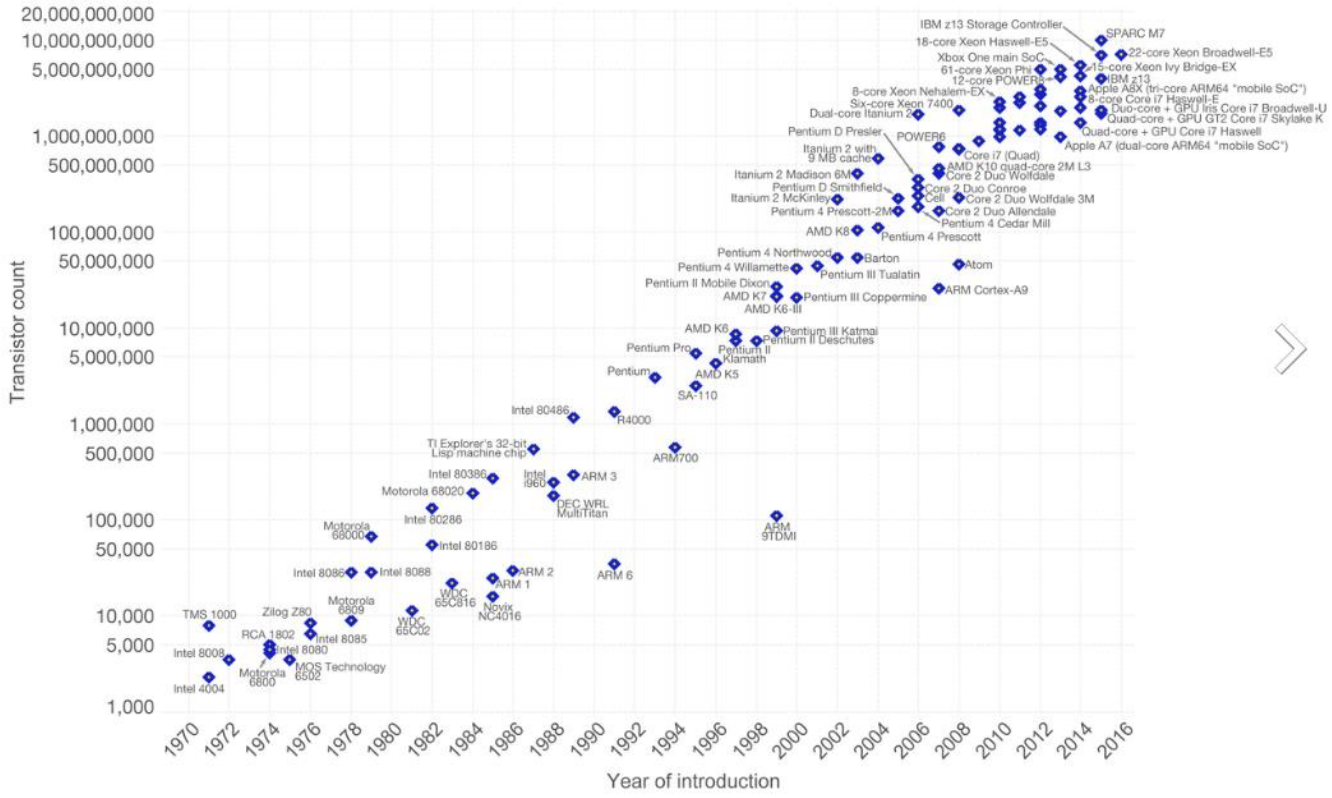


Figure 1 - Moore's law: the number of transistors on integrated circuit chips [4]

1.2. Goals

The main goal of this thesis is to understand and deeply analyze read performances of resistive memories, all along the sensing path of the memory: from the bit cell to the sense amplifier circuit-architecture via reference resistance scheme implementation. To reach this goal, some sub-studies were achieved:

- Literature review of the existing solutions for read yield enhancement, at both circuit and system level.
- Development of a statistical model evaluating the contributions on read yield of each component at the sense amplifier input, considering both resistive device and CMOS variability as well as parasitic elements.
- Analysis, characterization, modelling and optimization of the variability (quantified as offset) of a well-chosen sense amplifier architecture for resistive memories.
- Proposal of a sense amplifier architecture with reduced equivalent reference resistance variability and increased signal to offset ratio.
- Design of this circuit for a silicon test-structure production.

1.3. Outline

This manuscript first presents a state of the art about resistive memories. It briefly describes their working principle and their performances, before explaining the motivations to explore read capability of these memories. It then enumerates read yield enhancement techniques, at both system and circuit-level (Chapter 2).

The next chapter deals with a profound variability analysis along the sensing path of the resistive memory, from the bit cell alone to the sense amplifier circuit. The variability (quantified as offset) study and modelling of one well-chosen circuit architecture is achieved, in order to evidence the different offset optimization criteria (Chapter 3).

From this study, follows on (Chapter 4) the proposition of a sense amplifier architecture integrating some offset-optimization techniques stated in Chapter 2. This architecture also includes an accurate and easier-to-implement reference scheme giving a maximum signal-to-offset ratio.

The details of this type of sense amplifier test-structure design are described in chapter 5, and helps the reader to understand the different issues and their solutions for the design of memory-periphery circuits at very low technologic nodes (sub-30 nm).

The conclusion lists complementary research and application perspectives as well as additional promising read-yield enhancement ideas.

This version of the manuscript is submitted to confidentiality. Only readers that have signed a non-disclosure agreement are allowed to receive this version. The confidential section is integrated in Appendix I, and includes Chapter 4 and Chapter 5, as well as their conclusions in Chapter 6 and their French summary at the end of the manuscript.

Chapter 2: Background

2.1. Introduction

This chapter is a two-part literature review. On one hand a background on resistive memories is given, with the three main technologies' working principle, promises, challenges and their industrial roadmap. On the other hand, a focus on readability challenge is achieved, giving state-of-the-art of both system-level and circuit-level read yield enhancement techniques.

2.2. Resistive memories: background, promises and challenges

This section explains why searchers and industrials focused on resistive memory technologies these last years, by describing their advantages compared to existing memories. It then analyses current and future memory market and lists the main obstacles to massive commercialization of these emerging technologies, in particular read capability.

2.2.1.Context and motivation

As explained above, ensuring the extension of Moore's Law is becoming more and more complex, with the continual miniaturization of chip elements, especially CMOS memories.
Indeed, storage mechanisms of existing memories, mostly based on charge trapping, become more and more inefficient and unreliable with transistors' size reduction. For flash memory for instance, the number of electrons stored inside its floating gate is significantly decreasing, from more than one thousand to only a few tens in less than ten years [11] (see *Figure 2*).

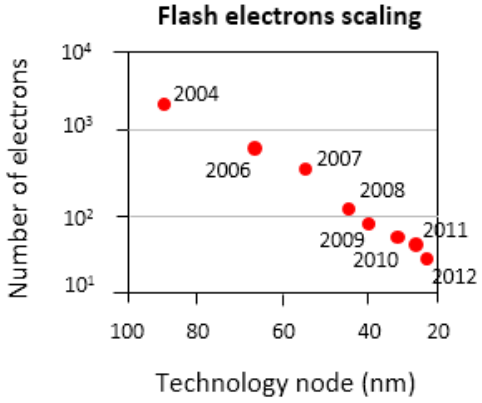


Figure 2 - Number of electrons stored in flash floating gate vs technology node

In light of these technologic issues, alternatives to existing memories appear unavoidable, in order to design nanometric-scale memories featuring high reliability and scalability along with better power, area and timing performances. Three new non-volatile technologies are emerging (PC-RAM, ReRAM, MRAM) and seem to be strong candidates to replace current memories (see *Figure 3*). Their data storage mechanism is not based on electron storage anymore. Data is stored in the form of resistive states, commonly either low resistive state (LRS) or high-resistive state (HRS). Their resistivity value depends on to the device material [12]. Some searchers came up to solutions to distinguish more than two resistive states, called multiple bit per cell [13][14][15], but it will not be the focal area of this thesis.

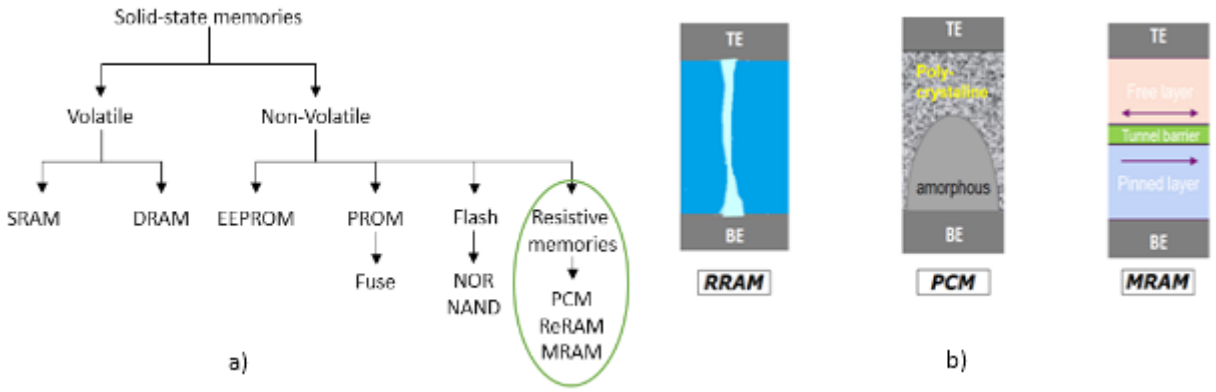


Figure 3 - a) position of resistive memories in solid-state memories classification b) the three main resistive memory technologies [12]

Resistive devices feature the benefit of lower power consumption, as their writing voltage is much lower than existing memories, floating gate flash in particular (up to a few volts for a current around 100 µA). This reduces the design complexity of the memory periphery, without the need of high voltage transistors or high current circuitries. Programing phase is also faster for resistive memories (under 10 ns). Besides, their simple structure and their CMOS process compatibility allow low cost manufacturing with only a small amount of additional masks to the core CMOS front-end. Finally, they are friendly with high density, scalability and design-flexibility [16].

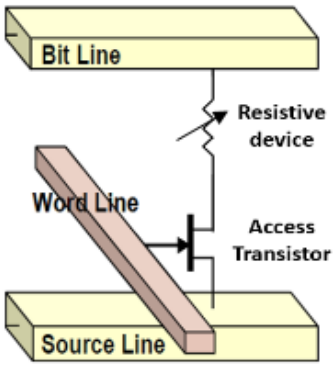


Figure 4 - Resistive memory bit cell

The resistive memory bit cell, or memristor, can thus be simply illustrated by a resistance in series with an access device that can be a transistor or a diode (see *Figure 4*). Some new advances come up with a resistive device integrating a selector and so without any separated access device [17][18][19]. The cell is accessed in a memory array by a correct polarization of their metallic bit, word and source lines.

Those demonstrated resistive memory performances arouse more and more interest of searchers and industrials to work in this field, with a possible view to compete with DRAM (Dynamic Random-Access Memories) and flash-NAND for some embedded applications [20] (see 2.2.3).

Below are detailed the working principles of these three promising technologies.

2.2.2.Resistive memory technologies

2.2.2.1. Phase-change memories (PCM)

PCM technologies use the property of a phase-change materials (usually chalcogenide glasses) to store data as resistive state. When the material is at the crystalline phase, its constituents (atoms, molecules or ions) are arranged in a highly ordered manner, forming a regular crystalline network extendable in all directions: it is the LRS. On the other hand, the amorphous material contains components that are randomly arranged without any allocation order: this is the HRS.

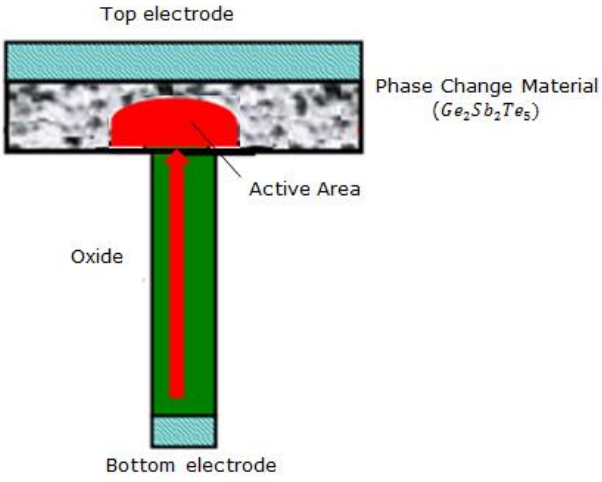


Figure 5 - PCM device structure: the Phase change material (generally an alloy of germanium, antimony and tellurium) and the oxide are sandwiched by two electrodes [21]

The PCM bit cell is made of the phase change material and a heater oxide sandwiched by two electrodes. Resistive states switching are achieved by passing a current pulse through this oxide. This current will help heat the phase change material, and generate an active phase-switched area, if the applied current exceeds the material critical current. In order to switch the material from the crystalline-low to the amorphous-high resistive state, a short pulse (under 10 nanoseconds) at high critical temperature (over 600 Celsius degrees) is sufficient since the crystalline phase is an instable phase, which network can be easily altered. Regarding HRS-to-LRS (set) switching, a current pulse with the same equivalent area than LRS-to-HRS (reset) switching is needed, in order to keep roughly the same energy cost. A longer pulse is required to ensure a well-mastered crystalline arrangement [21].

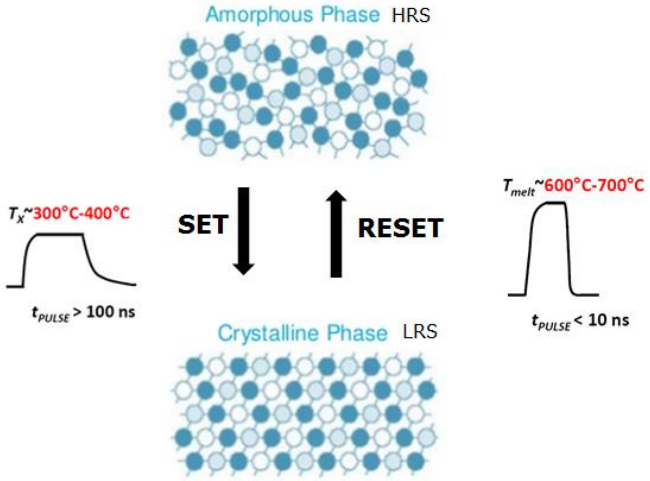


Figure 6 - PCM programming phase: illustrations of set and reset operations and the corresponding transienttemperature evolution of the phase change material [21]

This technology features limited performances in terms of power consumption, write speed and endurance, due to the power and timing-costly switching mechanism described above [22] [23][24].

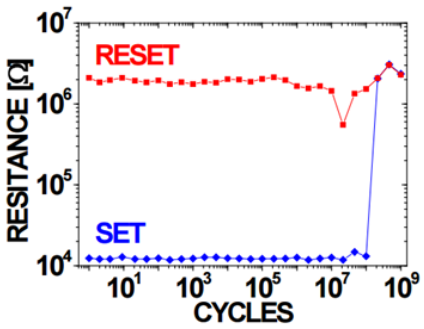


Figure 7 - PCM endurance is better than flash but limited compared to infinite SRAM endurance [24]

2.2.2.2. Resistive Random-Access Memories (RRAM)

The resistive RAM cell is made of one insulator (mainly oxide) sandwiched by two metallic electrodes. The data storage consists of the creation of a conductive path inside the oxide layer. Resistive state switching occurs when this path is formed (low resistive state: it is the set operation) or broken (high resistive state: it is the reset operation) [21].

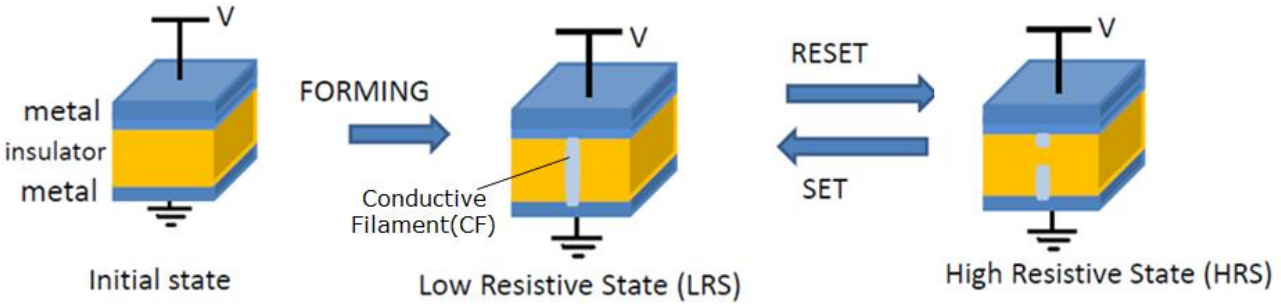


Figure 8 - Resistive RAM: forming, set and reset operation [21]

Two main resistive RAM families are distinguishable depending on their switching mechanism and more precisely the components of the formed conductive filament:

- Conductive-Bridge RAM (CB-RAM): the conductive filament is made of metal atoms. A positive bias between the top and the bottom electrode will generate oxidation-reduction reactions between those atoms and metallic ions all along the oxide. A negative bias will lead metal atoms to be oxidized and so generate the conductive filament rupture [25][26].

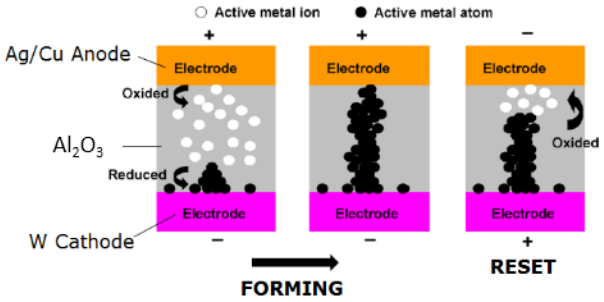


Figure 9 - Conductive-Bridge RAM: the conductive path is made of metal atoms

- Oxide-based resistive RAM (Ox-RAM): the conductive filament is composed of oxygen vacancies. These charges will follow the electric field resulting from the positive device biasing, while oxygen negative ions are moving in the opposite direction. A path of oxygen vacancies is therefore formed. Similarly, a negative bias can lead to the conductive filament rupture by generating the separation between the oxygen vacancies and oxygen ions [25][27].

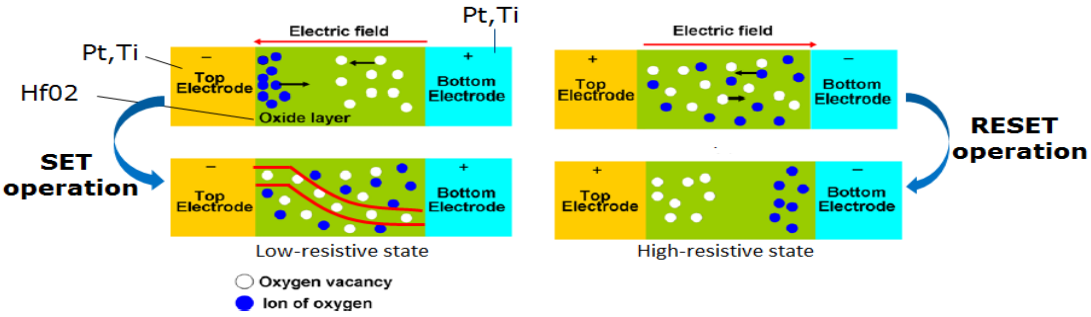


Figure 10 - Oxide-based resistive RAM: the conductive path is made of oxygen vacancies

The resistivity of this kind of device is very sensitive to voltage since the duration and intensity of the conductive filament forming depends on the device biasing. *Figure 11* depicts the typical evolution of ReRAM device current with voltage. Once the conductive filament is formed, the switching between LRS and HRS can either be bipolar or unipolar, according the direction of biasing of the top and bottom electrodes. The switching is called unipolar when only one direction of biasing is possible. Voltage is increased in order to set the low resistive state and reset to high resistive state is reached over a critical voltage. However, some materials (metal oxides like hafnium or titanium oxides) allow the ability to switch states with two directions of biasing, with either a positive or a negative voltage: it is the bipolar switching. Even if unipolar is easier to build in an array, bipolar switching is preferable, mainly because of lower voltage gap leading to increased probability of write disturb and limited endurance of unipolar switching. [12][16][21].

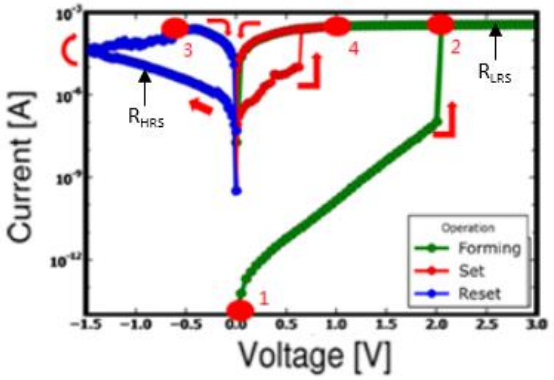


Figure 11 - Current-voltage bipolar characteristic of a resistive RAM putting forward the forming, the set and the reset operations [21]

Compared to PCM, resistive RAMs have lower switching currents and so lower power consumption because of the effect required in PCM to successively heat and cool the phase change material, which is mainly a chalcogenide glass. Similarly, switching mechanism of PCM is much slower than reduction-oxidation reactions or oxygen vacancies motions [22].

However, ReRAMs face to many challenges, notably high forming voltage requiring too much current limiting timing performances, scalability which can be much more limited when using a CMOS transistor as selector device or process-related issues limiting material control which can lead to read margin reduction and limiting the overall yield [16].

2.2.2.3. Magnetic Random Access Memories (MRAM)

Magnetic RAM relies on magnetism and exploits the principle of magnetoresistance. The magnetic memory element is called the magnetic tunnel junction (or MTJ). It is made of one thin non-magnetic oxide material

(insulating tunnel barrier) sandwiched by two ferromagnetic layers (FM1 and FM2). One of these two layers has its magnetic orientation fixed (called the pinned or fixed layer) and the other one can see its magnetic orientation being switched (called the free layer). Electrons have an angular momentum called either a spin-up or a spin-down depending on their rotation direction. When electrons travel from one FM layer to the other through the tunnel barrier, two resistive states can be distinguished: if the two-layer's magnetization are parallel, the probability for the electrons to cross the energetic barrier is low: it is the low resistive state. If the two layer's magnetization are not parallel, a higher energetic barrier is imposed by the fixed layer, it is the high resistive state [28][29].

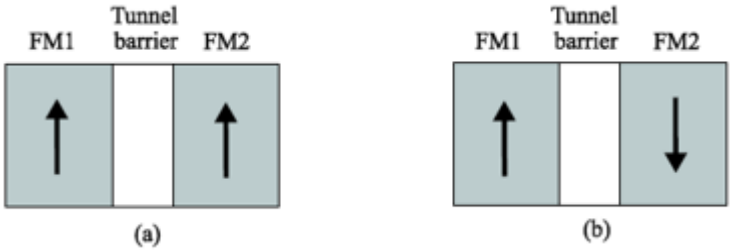


Figure 12 - Magnetic tunnel junction a) High resistive (or Anti-parallel) state b) Low resistive (or Parallel) state [29]

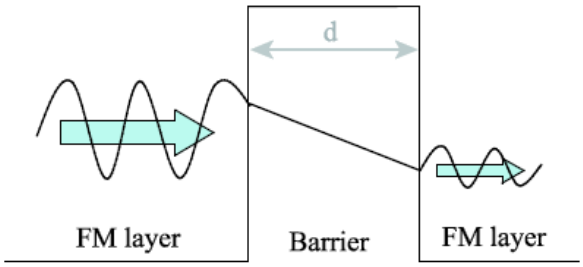


Figure 13 - Illustration of electron tunneling through the tunnel barrier [29]

The reading of the MRAM-cell is performed by electrically measuring the resistivity of the MTJ element. The lecture quality is quantified by the so-called tunneling magnetoresistance (or TMR), which evaluate the ability of the memory cell to distinguish the two complementary resistive states. This parameter is usually expressed in percentage and is a characteristic process parameter of the MRAM cell. A high TMR ratio is required to ensure a reliable sensing (default value: 100%, but more than 600% has been already demonstrated [30], see *Figure 14*):

$$TMR_0 = \frac{R_{HRS} - R_{LRS}}{R_{LRS}}$$ *Equation 1: TMR definition*

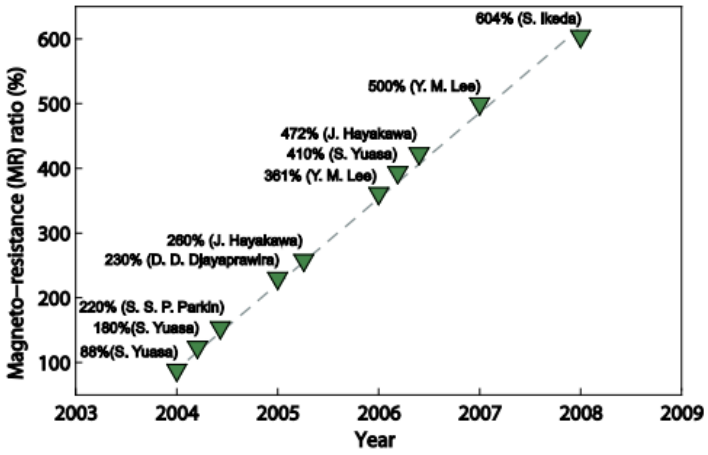


Figure 14 - Evolution of the TMR ratio for MTJs based on MgO tunneling barrier [29]

MTJ modelling [31][32][33] show that its resistivity, especially in HRS, varies non-linearly with voltage:

$$\mathrm{TMR(V)} = \frac{\mathrm{TMR_0}}{1 + \frac{\mathrm{V}}{\mathrm{V_h}}};$$

Equation 2: Non-linear dependence of TMR with voltage

Where $\mathrm{TMR_0}$ is the TMR ratio with zero bias voltage, $\mathrm{V_h}$ is the bias voltage when $TMR(V_h) = 0.5 * TMR_0$. Figure 15 depicts the typical resistance vs voltage characteristic putting forward this MTJ's resistance dependence.

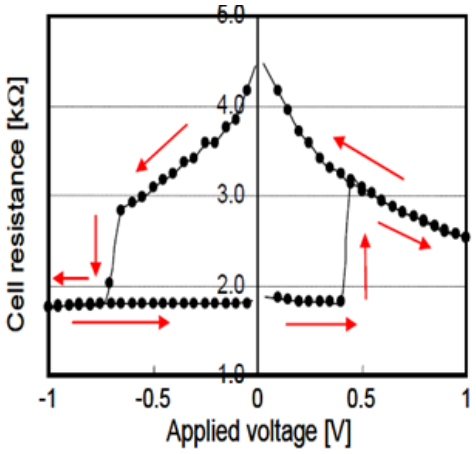


Figure 15 - R-V hysteresis curve of MRAM putting forward non-linearity of HRS

One the most-advanced writing mechanism of MRAM is called Spin Transfer Torque (or STT) effect [34][35][36][37]. Its working principle is described in *Figure 16* and *Figure 17*:

- If the MTJ is initially in anti-parallel (or high-resistive) state: the switching occurs when a spin-polarized current is applied from the free to the fixed. All electron spins penetrating the fixed layer will be fixed parallel to its magnetization, and so will be anti-parallel to the free layer's magnetization at the interface between the tunneling barrier and the free layer. If the accumulation of anti-parallel electron spins exceeds the free layer's magnetization threshold, they will exert a torque making the free layer switch to the parallel (low resistive) state (typical switching current = 300µA [38]).

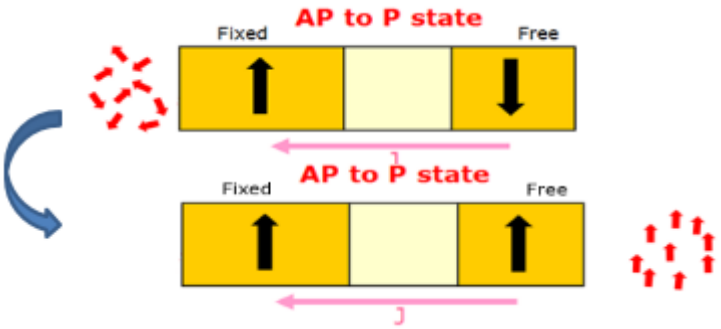


Figure 16 - Illustration of the spin transfer effect for the HRS-to-LRS switching

- The writing mechanism from the low to high resistive state is different: a spin-polarized current is applied from the fixed to the free layer. Both spin-up and spin-down electrons will cross through the free layer and the tunnel oxide. All electron spins opposite to fixed layer's strong magnetic field will be rejected back to the free layer because of the high energetic barrier of the fixed layer. As free layer's magnetization is parallel to the pinned layer's one, it

will be switched to the anti-parallel (high resistive) state by the accumulation of the rejected opposite electron spins (typical switching current = 400µA [38]).

Figure 17 - Illustration of the spin transfer effect for the LRS-to-HRS switching

Spin-Transfer Torque is a very fast writing operation (tens of nanoseconds demonstrated) and allows very high endurance performances (considered as infinite). However, it requires too much current (hundreds of µA), in order to increase the number of accumulated opposite electron spins at the interface with the free layer of the MTJ.

2.2.2.4. Performances comparison

This section summarizes resistive memories performances, compares them with existing technologies (see *Table 1*), and discusses about their limits and promises.

Table 1 - Demonstrated memory performances [22]

	Cell size (F²)	***Write speed***	***Power consumption***	***Endurance***	***Non-volatile***
STT-MRAM	*6 to 12*	*10 ns*	*Medium*	*10^{15} cycles*	*YES*
ReRAM	*6 to 12*	*20 ns*	*Low*	*10^{10} cycles*	*YES*
PCM	*6 to 12*	*75 ns*	*Medium*	*10^{8} cycles*	*YES*
SRAM	*100-150*	*5 to 10 ns*	*Low*	*10^{18} cycles*	*NO*
Flash	*4*	*10 µs*	*Very high*	*10^{5} cycles*	*YES*
DRAM	*6 to 10*	*10 ns*	*Low*	*10^{15} cycles*	*NO*

PCM needs to be further optimized at process-level in order to see it competing in memory market, by considerably improving its power, timing and endurance specifications. Nevertheless, the real attractiveness of PCM is its scalability. Indeed, its write operation is faster and easier when the chalcogenide material gets smaller. This is due to the fact that if the PCM scales down, the interface area of the oxide contact with the thermal active area of the phase-change material, and so its resistance R_{th}, decreases by a factor k. In order to keep the device temperature ΔT constant, the power P_d needed to switch phase has thus to be reduced by a factor 1/k as illustrated in *Equation 3* and *Equation 4*. This means that the resulting current I_M reduces as well when the device area goes down:

$$\Delta T = P_d . R_{th}$$

Equation 3: Illustration of the dependence of the required PCM writing power with scaling

$$P_d = R_{th}.I_M^2$$

Equation 4: Illustration of the reduction of the required PCM writing current with scaling

ReRAM is very promising in light of its high density, high endurance, fast writing operation and high power-efficiency. Compared to PCM, it requires high voltage for the forming operation, but this voltage can be reduced by inserting a nonstoichiometric oxide layer between the top electrode and the initial oxide [39]. Nevertheless, reset operation still requires too high current and its reduction affects reliability, while large noise is noticeable during set operation. Moreover, resistive memories good performances in terms of endurance are not sufficient. It is not only necessary to have a memory cell that can be highly programmed and erased; the materials must also be compatible with current semiconductor foundries in order to ensure cost-efficiency. ReRAM material control prove to be complex and so damages its manufacturing yield [16].

Finally, MRAM performances are summarized and compared with the other technologies in *Figure 18*. This illustrates well the fact that it has good global performances and that it scores everywhere.

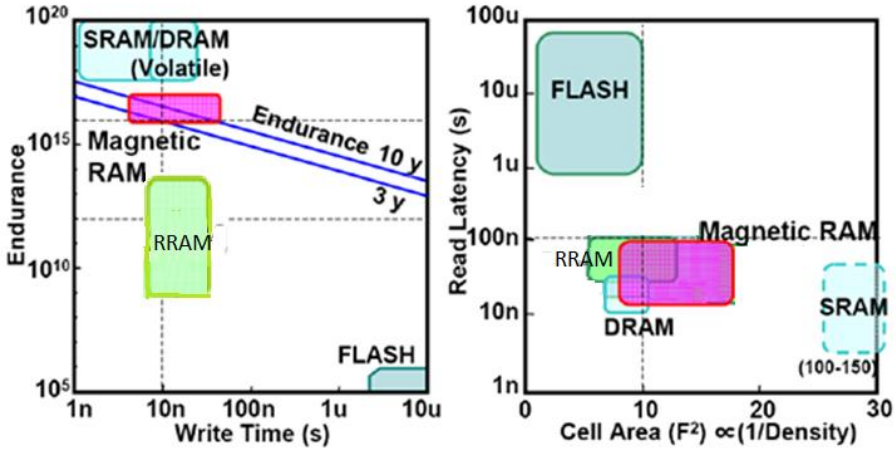


Figure 18 - Memory performances: MRAM scores everywhere!

Finally, the manufacturing cost-efficiency of MRAM is much better: its compatibility to standard front-end CMOS processes is understood and mastered (see *Figure 19*), and technological advances show extensive improvements regarding TMR ratio increase and writing current reduction [38].

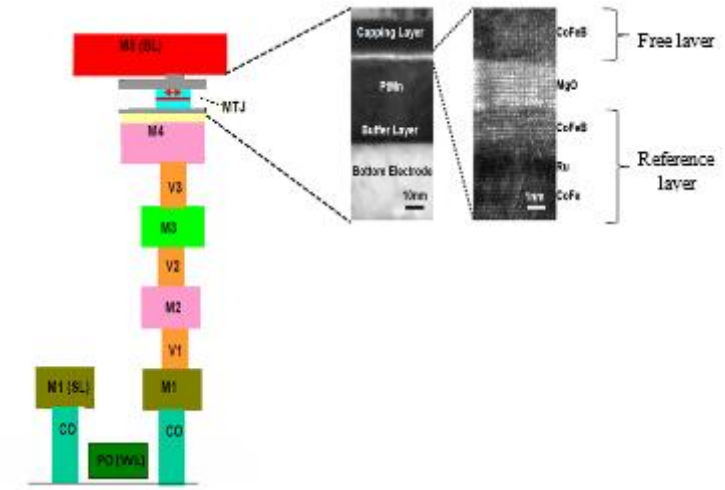


Figure 19 - CMOS compatibility of MRAM: the MTJ is integrated between two metal levels with only a few additional masks

2.2.3.Emerging memories market

It is now opportune to discuss about the current and future situation of resistive memories in the industrial global market [22].

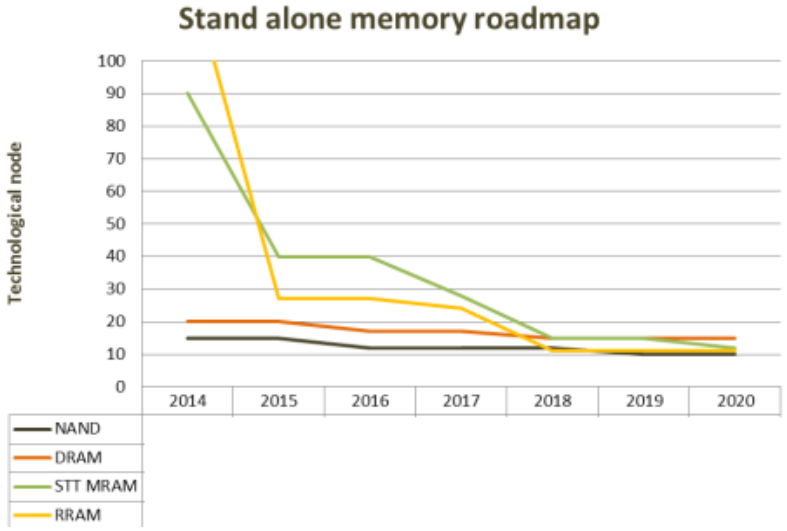
2.2.3.1. Current memory market

First, memory market is divided into standalone memories, implemented in a dedicated chip (mainly DRAM and flash-NAND), and embedded memories (mainly SRAM and flash-NOR), integrated in a shared chip with other elements for a given functionality (microcontrollers, microprocessors …). Memory market has been for a long time consecrated to computer products before being dedicated at the beginning of the 2000s to the mobile market. Nowadays, big data and Internet of Things (IoT) are emerging, and memory needs in these sectors are more and more increasing. In 2015, the five biggest current players in standalone memory sector were Samsung; Micron, SK Hynix, Toshiba and SanDisk. They concentrate 95% of the market. As for embedded memories, competition is more severe but the top five companies in 2015 were Renesas, STMicroelectronics, Microchip, NXP and Atmel.

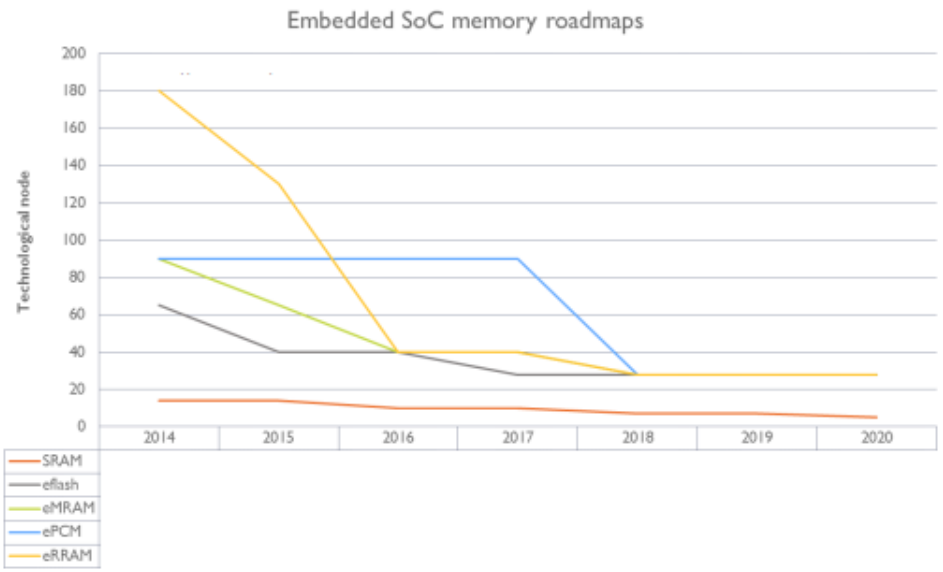
2.2.3.2. Emerging memories roadmap

Emerging memories seem first being able to replace embedded existing memories rather than standalone owing to less restricting embedded design needs (lower surface constraint and density in the order of tens to hundreds Mbits) with better performances. The main targeted sectors of embedded emerging memories are mobile communication modules and microcontrollers. Market actors are anticipating transition to newest memory technologies even if current memories still have years ahead thanks to process advances like 3D-approach of flash in order to go beyond 10nm technologic node.

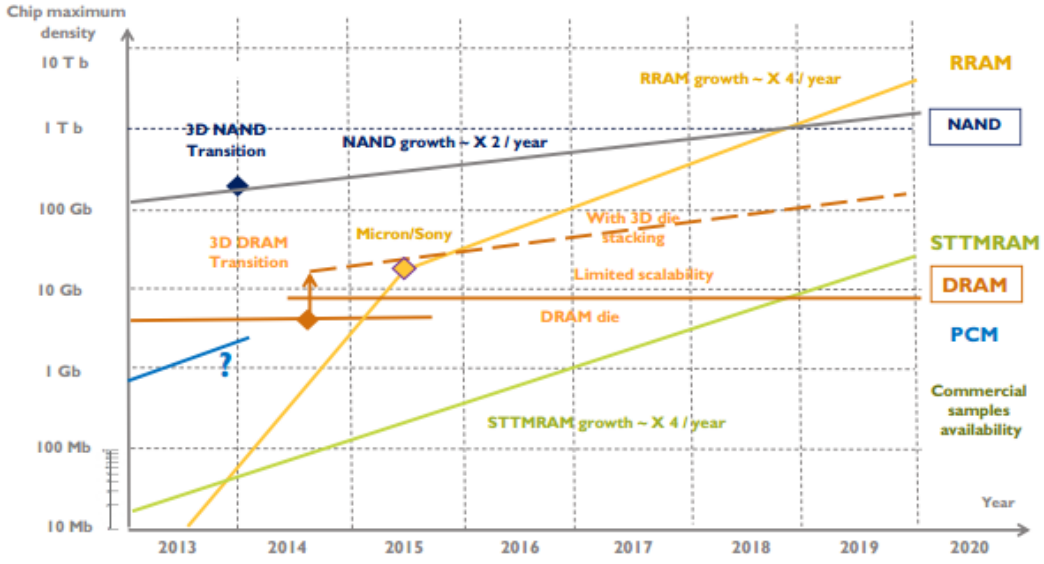
Regarding scalability, technologic nodes of standalone emerging memories should reach NAND and DRAM ones over the next three years. The same should be applied to chip density for standalone devices, and STT-MRAM appears to be the best successor to DRAM while ReRAM is reaching NAND-flash's density. On the downside, embedded emerging memories technologic nodes should not reach SRAM nodes (under 20nm) but this is counterbalanced by the huge cell size of SRAM (see *Figure 20*).



a)



b)



c)

Figure 20 - a) b) c) Standalone and embedded current and emerging memory roadmap in terms of technologic node and density [22]

2.2.4.Resistive memories: variability challenges

While resistive memories write yield enhancement seems well controlled, read yield is highly threatened by scaling, since both bit cell, memory array and memory periphery elements variability significantly increase. Variability can occur between different wafers, between different dies of the same wafer (interdie variability) or between different locations of the same die (intra-die variability) [40]. Variability results from resistive devices process uniformity but sense amplifiers also features high variability, due to CMOS transistors variations. Many sources of CMOS transistor variability exist, among them [41]:

- Random dopant fluctuations
- Interface State density fluctuations
- Line-edge roughness

Those sources correspond to process uniformity especially at the interfaces metal-oxide and oxide-substrate. This leads to severe parameters variations, mainly the transistor's threshold voltage and its effective channel length.
It is necessary to distinguish systematic and random variability. Systematic variations are reproducible inaccuracies that are consistently in the same direction. Even if they can feature a stochastic behavior, they are difficult to be analyzed statistically (e.g. DC offset). Random variations are unknown and unpredictable inaccuracies. They often have a Gaussian normal distribution, and so statistical methods are used to minimize them (e.g. CMOS mismatch) [42][43].
Variability is thus an important concern for robust circuit design. The following of this thesis will focus on variation-tolerant read circuits design for reliable resistive memories.

2.3. Statistical modelling and analysis

2.3.1. Statistics background

As variability is handled using statistical methods and as random variability is handled using statistical methods, as variability is mathematically defined as a standard deviation and yield depends on error probability, the statistics terms that will be developed thereafter have to be introduced in this section.
The read yield quantifies the percentage of memory bit cells that are correctly read in a given sample. It thus depends on the bit cell read failure probability. The intra-die and/or interdie variability corresponds to the distribution of a sensed parameter (reading current, input resistance …). Regarding the resistive device variability, LRS and HRS each follow a Gaussian distribution: $R_{LRS}(\mu(R_{LRS}), \sigma(R_{LRS}))$ and $R_{HRS}(\mu(R_{HRS}), \sigma(R_{HRS}))$ are so random variables. Similarly, CMOS variability is expressed through characteristic process parameters of a transistor: threshold voltage (V_T, σV_T) and channel length (L, σL) [41][44],[45] [46]).

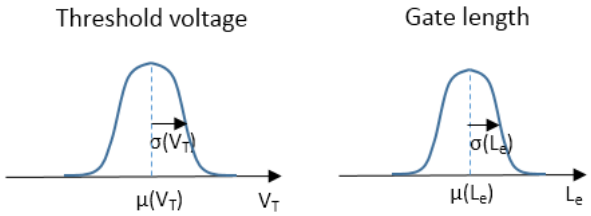


Figure 21 - Illustration of the Gaussian distributions of CMOS transistor's threshold voltage and channel length

As for the variability of analog CMOS circuit like sense amplifier, all its components' variability create a gap between the real value of the characteristic measure of the circuit (output voltage or current for example) and its expected value: it is the offset of the sense amplifier. It is better to take the input-referred offset (output measure divided by the gain of the circuit) since it will not be saturated compared to the output signal which is dependent of the gain of the circuit [47].
Moreover, it should be noted that the variability of several uncorrelated components is equal to the quadratic sum of each standard deviation, and that only the biggest terms of a quadratic sum are dominants, so that the terms that are three times inferior to the maximum value are negligible [48].
Evaluating the offset of a CMOS circuit consist on inserting some operations on individual transistor's random variables, like multiplication or division: the theorem of propagation of uncertainty helps compute the standard deviation of non-linear combinations of two random variables A and B, with respective standard deviation σ_A and σ_B and covariance σ_{AB}. It is detailed in [49][50]:

$$If\ \mathrm{f} = \mathrm{AB},\ \sigma_f \approx |\mathrm{f}|\sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2 + 2\frac{\sigma_{AB}}{AB}}$$

Equation 5: Standard deviation of the product of two random variables

$$If\ \mathrm{f} = \frac{\mathrm{A}}{\mathrm{B}},\ \sigma_f \approx |\mathrm{f}|\sqrt{\left(\frac{\sigma_\mathrm{A}}{\mathrm{A}}\right)^2 + \left(\frac{\sigma_\mathrm{B}}{\mathrm{B}}\right)^2 - 2\frac{\sigma_\mathrm{AB}}{\mathrm{AB}}}$$

Equation 6: Standard deviation of the division of two random variables

Nanoscale circuit design has to take care about what is happening in Gaussian tails (i.e. at four or five sigma far away from mean value) for good yield. Therefore, it is better to represent the cumulated distribution of the Gaussian law instead of the typical differential distribution, in order to depict tail effects [41].

2.3.2.Simulation methodologies

As variability has to be handled in circuit design, some statistical simulation algorithms are more and more integrated in computer-aided design (CAD) tools. This section lists the main statistical simulation methodologies.in order to characterize the variability of a circuit.

The most common used methodology is Monte Carlo (MC) simulation. It comprises a succession of random draws, and gives the mean and the standard deviation resulting from this draw. This method has two main drawbacks. First, it covers no more than three sigma overall variation and therefore it is not suitable for nanoscale circuit design, which has to cover more than four sigma to ensure sufficient yield. This is because two much draws are required (about ten thousands) and that simulation times are too long. Secondly, it does not output Pareto chart, which lists contributions of each circuit device on overall variation. Pareto results help optimize the circuit and reduce run time [51][52].

The Response Surface Model (RSM) gives a representation of the different contributors to a circuit variation. For example, we can represent the evolution of the delay from the first to the last stage of an inverter chain with respect to V_t et L of each transistor of a circuit [53].

Finally, the principle of the Most Probable Point (MPP) methodology is to find the smallest combination $\{\sigma_1 \dots \sigma_N\}$ such that $\sigma_{tot} = \sqrt{\sigma_1 + \cdots + \sigma_N}$ for a given measurement. For example, MPP gives the smallest $\{\sigma V_T, \sigma L\}$ pair of a MOS transistor of an inverter chain giving a 200 ps delay through the chain. The solution can be illustrated by the intersect point between a straight line (depicting the targeted delay) and a circle (depicting all possible $\{\sigma V_T, \sigma L\}$ pairs). The most probable point can also be obtained by varying the measurement until finding the intersect point with a fixed circle depicting a fixed σ_{tot} (called MPP 2) [41].

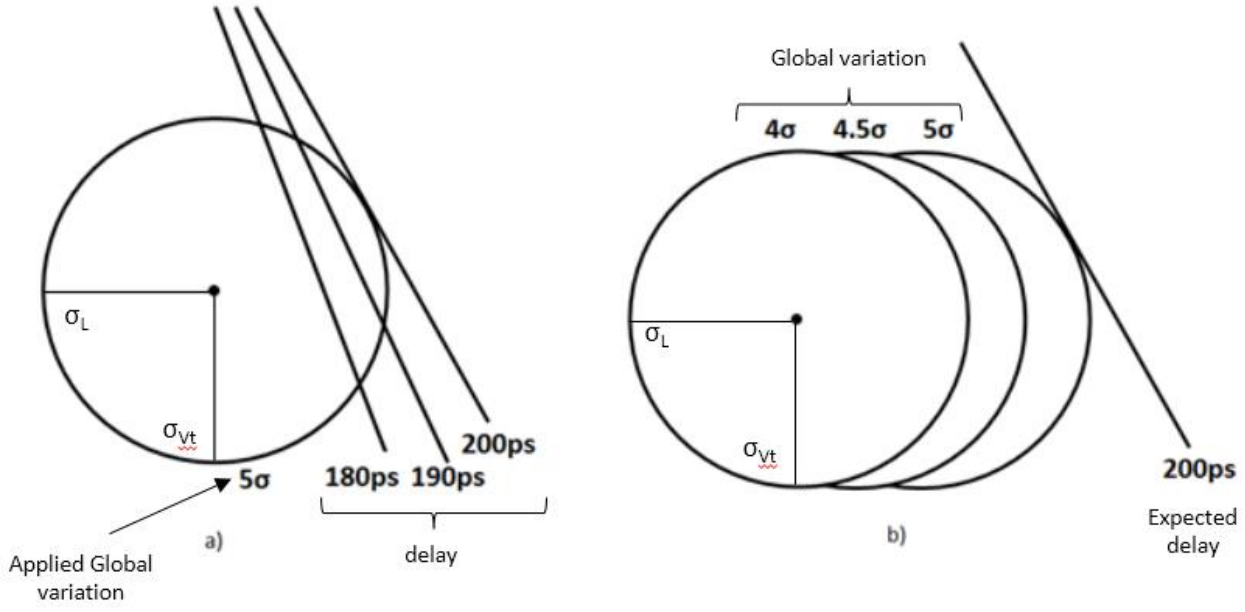


Figure 22 - Principle of MPP and MPP 2 methodologies

MPP features better characteristics than RSM or MC: it can covers more than six sigma of global variation (i.e. several Mbits in a memory array), reports a Pareto and handles high-order non-linear circuits, like level shifters or Schmitt trigger. However, it cannot completely replace the two aforementioned methods in all circuits, since it cannot afford more than 60 variables (i.e. 30 transistors by varying their channel length and threshold voltage), and since it noticeably depends on the measured circuit parameter: hence, it is suitable for delay measurements but not for memory state or retention evaluation.
RSM seems to be the best first choice for small analog circuits (which are mostly linear): it is better to replace a MC simulation of a big complex circuit by the quadratic sum of many RSM simulations of small sub blocks.

Method	Useful σ Range	Reports Sources of Variation	Non-linearities
Monte Carlo Simulations	**~ 2.7**	**No**	**High Order**
Design of Experiment Response Surface Model (DOE/RSM)	**~ 4 (with some insight to 6)**	**Yes**	**2nd Order**
Most Probable Point (MPP)	**> 6**	**Yes**	**High Order**

Figure 23 - Comparison of the different statistical simulation methodologies

2.4. Resistive memories readability: enhancement techniques

As aforementioned, readability, which can be defined as the ability to ensure high read yield or reduce read failure probability [54], appears to be the greatest challenge and the newest barrier for resistive memories massive commercialization. This thesis focuses on this topic, in order to better understand variability issues in analog read circuits design. This section reports state-of-the-art read yield enhancement techniques, both at system and circuit level.

2.4.1. System-level leveraging: ECC and redundancy

The most commonly used and well-mastered memory read-yield enhancement techniques are based on Error Correcting Codes (ECC). These codes are integrated in order to detect and correct soft errors, which are error occurrences that temporary change the data value but, contrary to hard errors, do not damage the system's hardware and can be rectified. These coding techniques are based on redundancy and has the advantages to reduce data transmission error rates without significantly increase the memory speed [55]. Mathematically speaking, the correcting code is represented by an injective function: ϕ: $A^n \rightarrow B^p$, which gives a coded word $M' \in B^p$ from a unique initial word $M \in A^n$, so that $M'=\phi(M)$. A^n and B^p are two alphabets of respective dimensions n and p. For example, if n=p=2, A and B are the binary alphabet. Parameter n is called the dimension of the code φ and p is the length of the code. Correcting codes are the most often systematic codes, meaning that the coded word corresponds to the initial word followed by a given number of redundancy symbols [56].

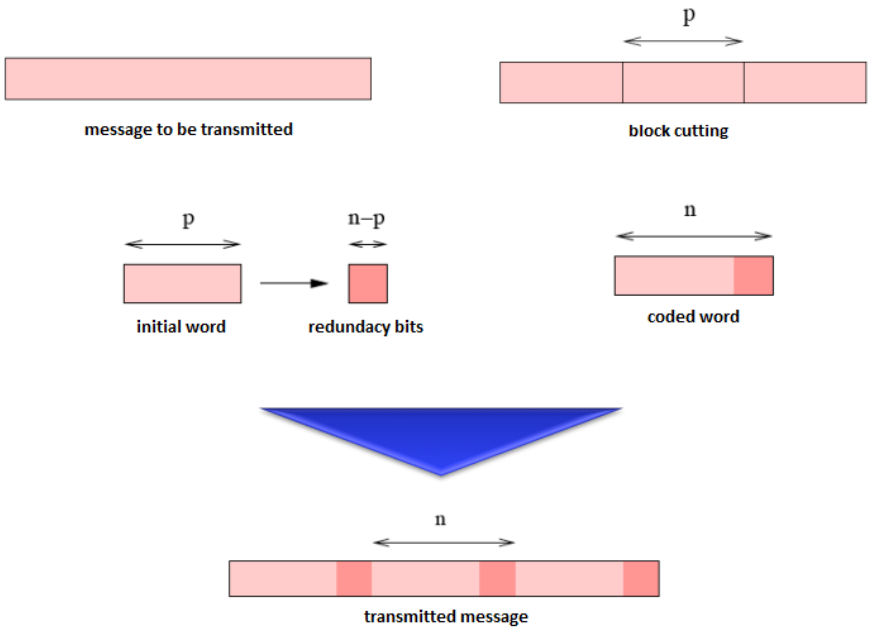


Figure 24 - Systematic coding: principle [56]

ECC are thus efficient techniques to improve memory read yield. However, they are costly in terms of area due to the use of redundancy (see Chapter 4). That is why circuit design-level enhancement techniques have to be proposed and circuit designs to be optimized for a moderate use of ECC [57][58][59].

2.4.2.Circuit-level leveraging: Sense Amplifier optimization

In this section, some state-of-the-art sense amplifier architectures are introduced and compared regarding their tolerance to variability.

2.4.2.1. Conventional Current Mode Sense Amplifier

The design of a variation-tolerant sense amplifier is critical for reliable resistive memories design. The sense amplifier is a memory periphery analog circuit (see *Figure 25*), needed to sense the data to be read due to the ever-decreasing margin between two complementary inputs with scaling.

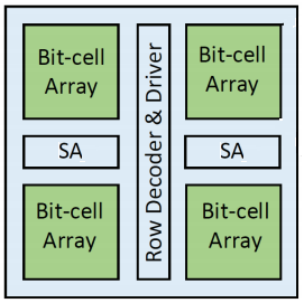


Figure 25 - Sense Amplifier (SA) position in a memory array [60]

Simplified schematics of current sensing circuit, commonly used for existing memory technologies (SRAM, DRAM …), are depicted in *Figure 26* where the resistive memory bit cell impedance is represented by R_{DATA} and R_{REF}. Dynamic current sense amplifiers are often chosen because they are faster than voltage ones [61]. Reading operation is achieved by an electrical R_{DATA} and R_{REF} comparison in order to output a digital signal via the latch-type circuit (either 0 or 1 depending on whether data resistance is superior or inferior to reference one).

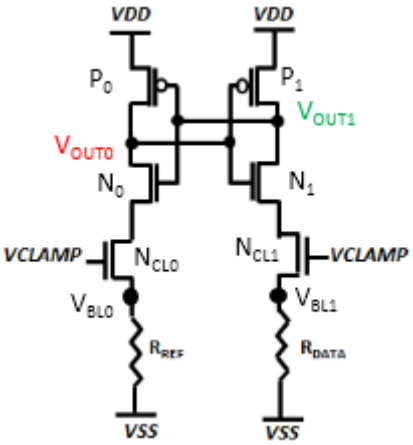


Figure 26 - Conventional sense amplifier for resistive memories

The working principle of the current-mode conventional sense amplifier is described now, assuming R_{DATA} is superior to R_{REF}. This circuit is typically is based on a latch stage and two NMOS transistors (NCL0 and N_{CL1}) that set the bit line voltages V_{BL0} and V_{BL1} the two output voltages V_{OUT0} and V_{OUT1} are first initialized to supply voltage V_{DD} so that the two upper PMOS transistors are off. As data resistance is superior to reference one, its current I_{DATA} is lower than I_{REF} and so V_{OUT0} will decrease faster than V_{OUT1} corresponding to a faster capacitive load discharge. This way, transistor P_1 turns on, leading to V_{OUT1} increase while P_0 is kept off. This leads to two complementary digital outputs. This behavior is illustrated in *Figure 27*.

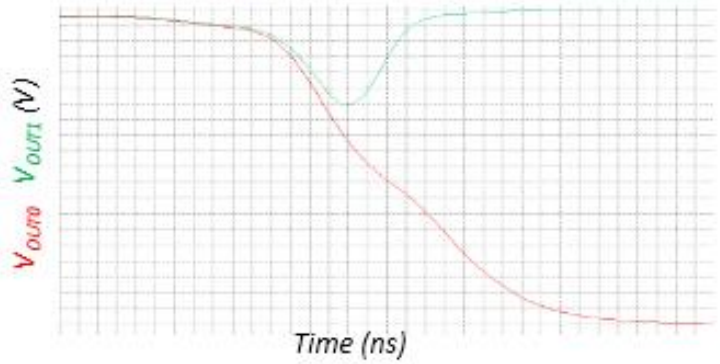


Figure 27 - Conventional current sense amplifier: output voltage characteristics

For reliable resistive memories design, the circuit in *Figure 26* is not suitable. Indeed, it is highly impacted by process variations such as random dopant fluctuations or interface-state density fluctuations (see 2.1.4) [62]. Indeed, in deep submicron technology, those physical variations severely impact within die transistor threshold voltage variability. This translates to high offset for a circuit based on matched transistors. Since circuit depicted in Fig.1 is based on matched sensing paths, its read window can be seriously narrowed down. Additionally, process variations impact severely the resistive states distributions, increasing even more overall system failure rate. Moreover, a complex mid-point reference scheme is required [63], leading to severe layout irregularities (see 2.4.2.2). This circuit helps understand the two main issues for a variation-tolerant sense amplifier design: reference implementation and process variations sensitivity. The biggest challenge is so to find an architecture handling both of these problematics. Next section details the reference scheme implementation issue.

2.4.2.2. Reference Scheme

The conventional current sense amplifier needs the implementation of a mid-reference, that can be either a current or a resistance. The different conventional reference schemes are depicted in *Figure 28*.

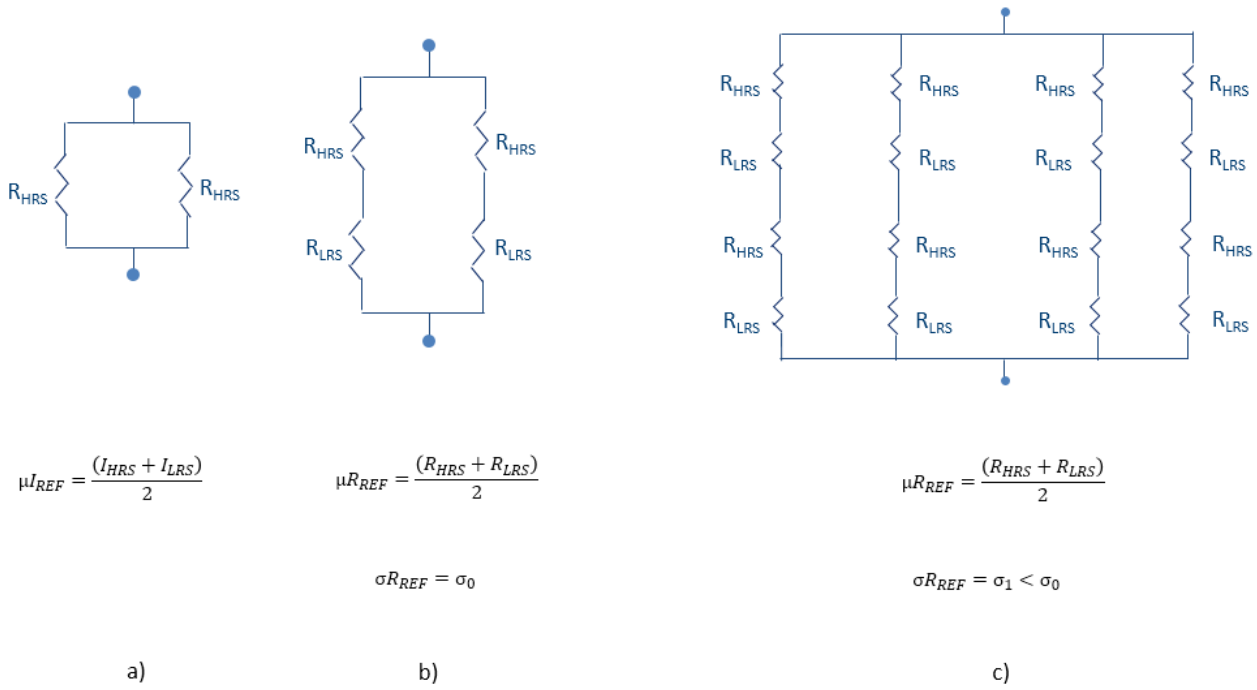
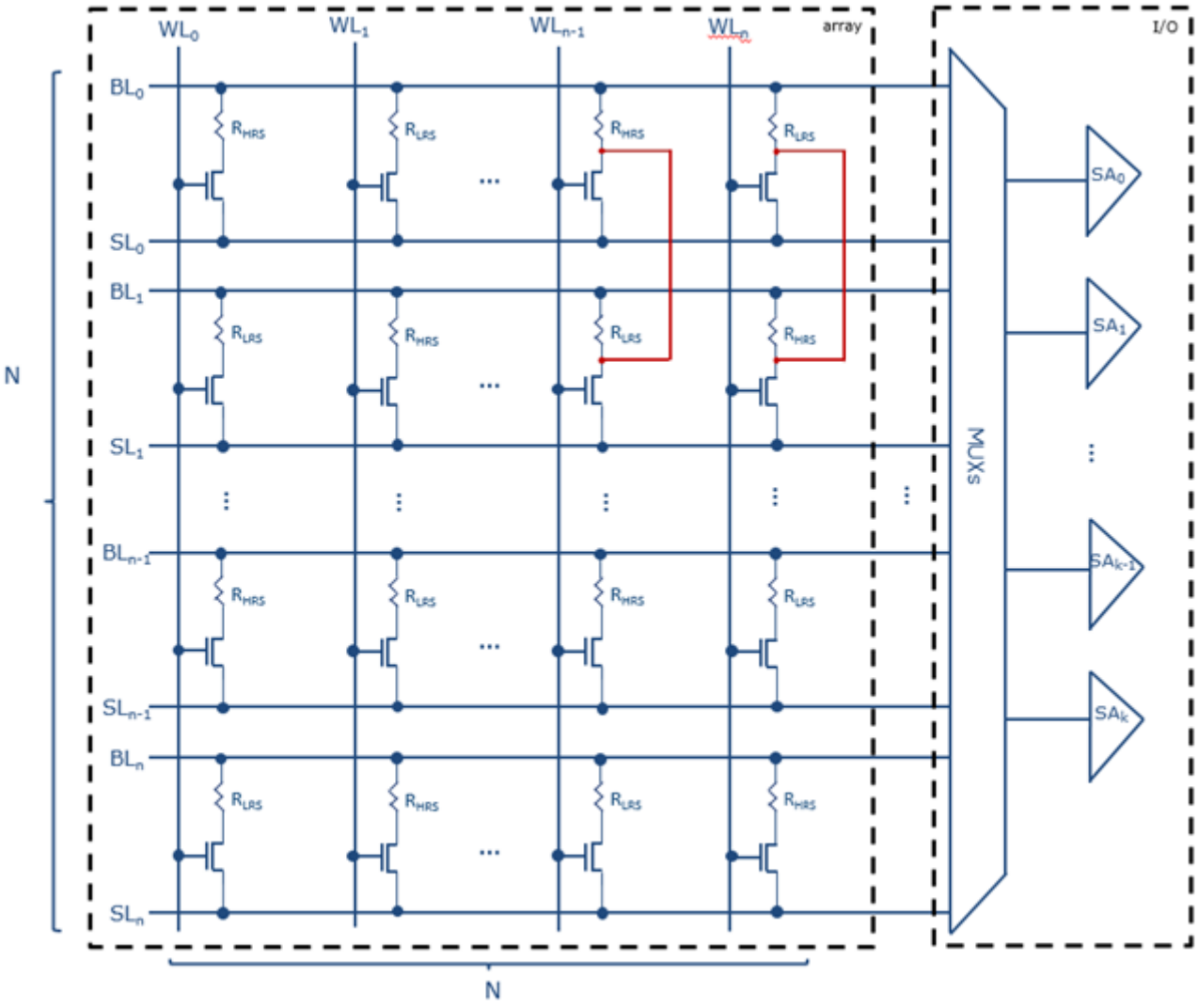


Figure 28 - a) Conventional reference current b) Conventional reference resistance c) Reference resistance with reduced variability

The current reference in a) is suitable for sense amplifiers like current mirrors. It has the advantage to include only two resistive devices. This includes high sensitivity to variations. The resistance reference in b) is convenient with latch-type circuits, while four minimum resistive cells are required. To reduce its variability, more cells are needed, like in c). All these implementations are complex to design in a memory to ensure both layout regularity and high accuracy [64][65].

Layout regularity is a design concept corresponding to the repetitiveness of layout blocks. The goal is to design with a reduced set of layout bricks, making small configuration changes to generate different circuits. Regular layouts require a simplified mask set, leading to reduced manufacturing cost, better process control and reduced design complexity [66].

Figure 29 illustrates two types of metallic connections in a memory array to build the conventional reference resistance scheme of *Figure 28*b): on the one hand, the resistive devices connections are made inside the bit-cell array and on the other hand, they are made on bit-cell array periphery (or inputs/outputs called I/Os). The first configuration is not suitable because it will not ensure homogeneous layout and cleaned design rules checking (DRC). Moreover, it is sensitive to sneak currents, leading to increased write disturb probabilities. The second configuration does not alleviate sneak currents issues unless using a switch device that adds layout complexity.


WL0
WL1
WLn-1
WLn
array
I/O
BL0
SL0
BL1
SL1
BLn-1
SLn-1
BLn
SLn
RHRS
RLRS
MUXs
SA0
SA1
SAk-1
SAk
N
N

a)

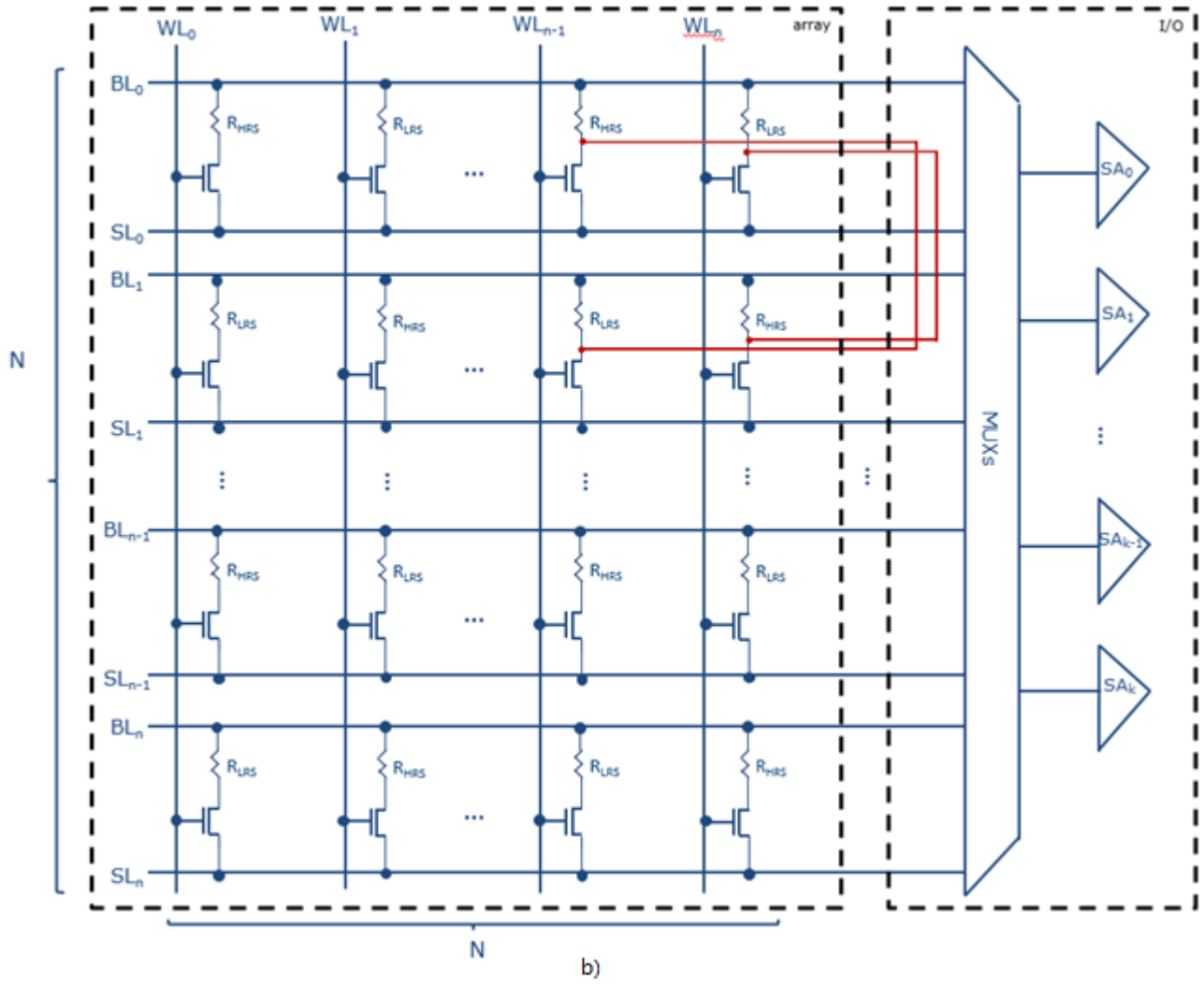


Figure 29 - Illustration of conventional reference layout irregularities a) connection through the array b) connection through the I/O

Another drawback of these schemes is that their layout irregularities do not allow them to track resistive devices variations with temperature and voltage, as we can see in Figure 30 [65].

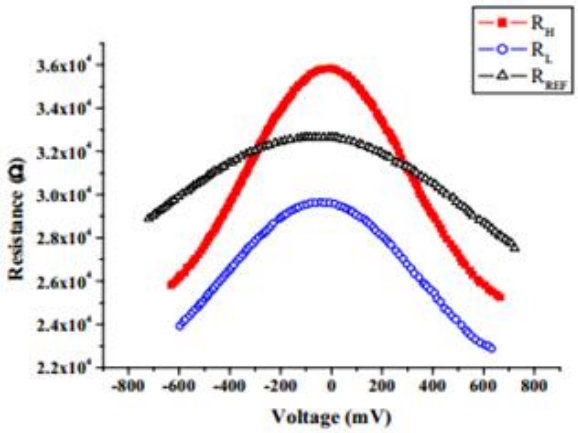


Figure 30 - Evolution of high, low and conventional reference resistive states with voltage [67]

Thereby, developing a well-averaged reference scheme appears to be complex to ensure both layout regularities and good equivalent resistance accuracy, and existing solutions as the multiple-cell reference scheme [63] (see *Figure 31*) which is based on redundant 1T-1R cells, do not manage to solve those issues.

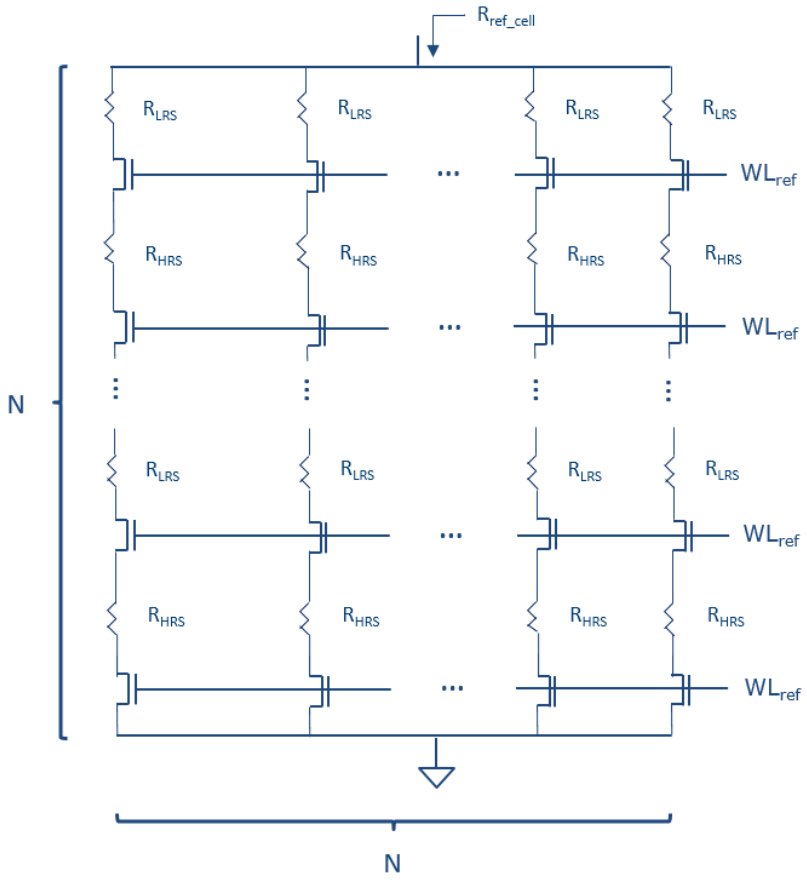


Figure 31 - Multiple-cell reference scheme

2.4.2.3. Covalent-Bonded Cross Coupled Current Mode Sense Amplifier

To alleviate reference issues, [64] proposes a double-latch type sense amplifier, called covalent-bonded cross-coupled current-mode sense amplifier (CBSA, see *Figure 32*). This circuit features an improved sensing margin in relation to conventional sense amplifier, thanks to the comparison of data resistance with two references. It is composed of three memory cells, two are reference cells with low resistive state (LRS) and high resistive state (HRS) and the third one is the data to be read (can either be LRS or HRS). Two latches are then necessary to compare separately the HRS reference to data and the LRS reference to data. The two latches are connected and so compared one another to get the final digital output. The pre-charge stage (PPR and PEQ transistors) initially sets output nodes to supply voltage V_{DD} (PRECH signal first equal to 0) and then turns OFF (PRECH=V_{DD}) to trigger the dynamic behavior of the circuit (output capacitor load discharges). Because the two latches share data current branch with each other, transistors P3HD, N3HD, P3LD, and N3LD are added to balance the current amount. Transistors P4HD, P4LD, N4HD and N4LD are MOS capacitors integrated to balance the systematic offset due to output capacitor mismatch between OUTA (or OUTB) and OUTM nodes. This circuit appears to be an alternative architecture of latch-based circuit adapted to resistive memories, i.e. with an adapted reference scheme. Indeed, no mid-reference scheme is integrated here: two intermediary output signals are first obtained through the two latches in order to evaluate the margins of the data with LRS and HRS simultaneously and separately, and the margin between these two outputs gives finally the digital output signal.

On the other hand, it is a proposal of a dynamic sense amplifier handling variability issues. It is important to evaluate this variability, understand precisely the different factors impacting it and how it can be reduced. Its working principle and the variability study is detailed in Chapter 3.

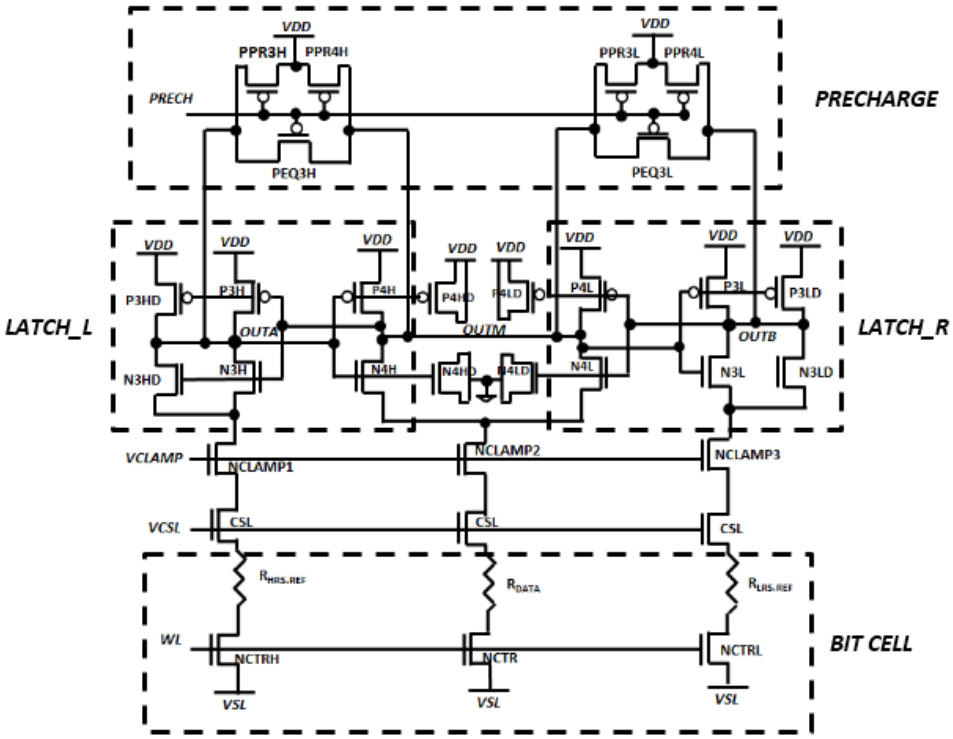


Figure 32 - Covalent-bonded cross-coupled current sense amplifier [65]

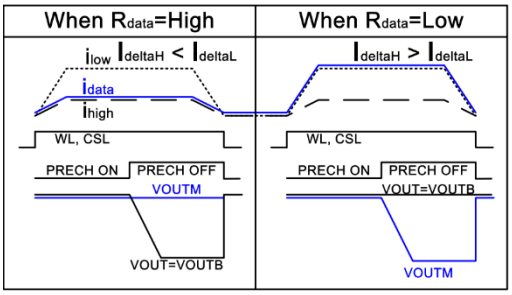


Figure 33 - CBSA: sensing principle

At this point, it is necessary to study how to reduce variability in analog circuits. This variability is quantified as an offset, which is defined as the differential DC amount (voltage, current or resistance) required between the inputs of an amplifier to make the output zero.
Below are listed some offset-cancellation techniques and their implementations in some sense amplifier architectures. Those techniques are more detailed in [68].

2.4.2.4. Offset Cancellation Techniques: Autozeroing and Chopper Stabilization

Autozeroing and chopper stabilization are two circuit techniques first used to reduce both dc offset (which can be defined as noise at a zero frequency) and noise of operational amplifiers. Indeed, those two parasitic effects can result in the reduction of the dynamic range and output swing of the amplifier, leading to a low gain amplifier and cascode would not be sufficient to significantly compensate for this. Those two techniques can be applicable not only to operational amplifiers but also to any voltage amplifiers, digital-to-analog and analog-to-digital converters, filters and integrators, sample and hold circuits, analog delay stages or comparators.

The basic principle of autozeroing is to use switches and capacitors in order to sample in a first phase only offset and noise contributions by shorting the input signal, i.e positioning it at a given common mode voltage. Then, in a second phase, the input signal is connected again for amplification and the offset and noise of the second phase is subtracted with the one registered by the sampling capacitors.

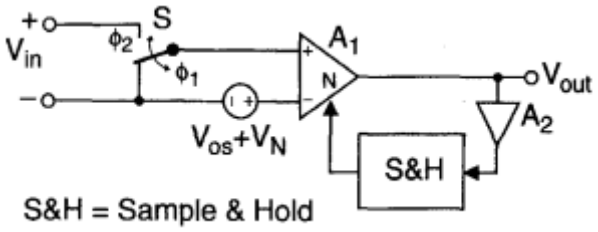


Figure 34 - Autozeroing basic principle [68]

This way, all dc components can be cancelled. Regarding noise, the 1/f noise is also cancelled, thanks to the introduction of a double zero term in the power baseband transfer function brought by the autozeroing stage. But the 1/f noise and above all the thermal noise spectrum is replied, and leads to the increase of the white noise component.

As autozeroing consists in registering offset and noise contributions through sampling capacitors by means of voltage V_C or current I_C (simply called control variable X_C), a residual offset can still be present, corresponding for example to the variation of the amount of switches injected charges. *Figure 35*.a) shows that residual offset can be deduced from the input referred offset vs X_C characteristic, which can be linear or nonlinear depending on the circuit used. *Figure 35*.b) shows that for a low initial offset, it is more convenient to adopt a nonlinear characteristic in order to reduce residual offset.

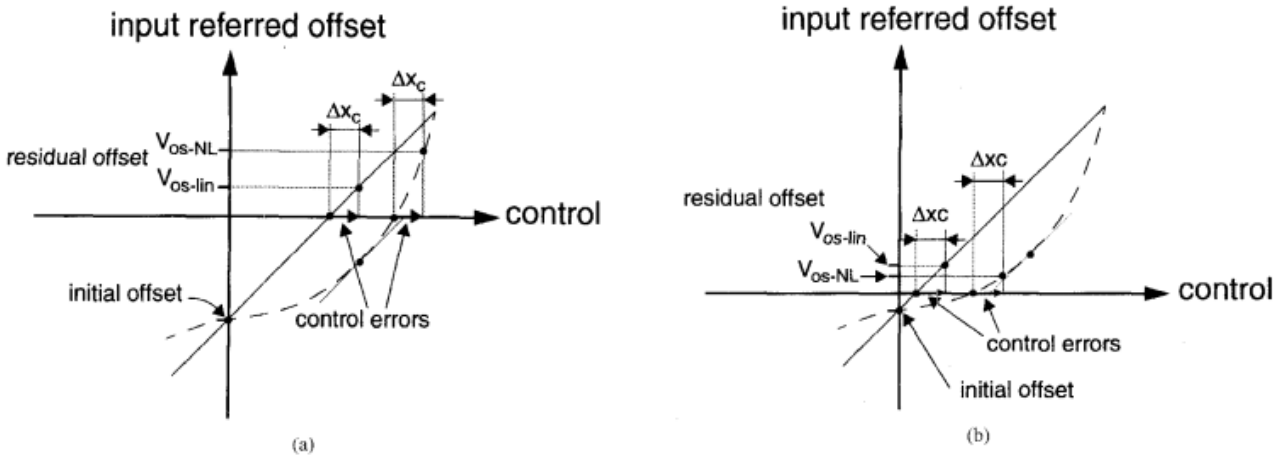


Figure 35 - Input-referred offset vs XC characteristic for a linear (a) or a nonlinear (b) circuit [68]

The Chopper Stabilization technique does not anymore consist on sampling offset and noise contributions but on frequency modulation. Instead, the input signal spectrum is translated to higher frequencies (where there is no 1/f noise), then amplified and so spectrum aliasing is avoided and finally demodulated in order to come back to baseband. In order to avoid aliasing, input signal frequency must not exceed the double of modulation (or chopping) frequency. This condition also allows a considerable reduction of the noise brought by the frequency modulator and demodulator. Indeed, the spectral noise density after demodulation is inferior to initial noise before modulation, for both white noise and 1/f noise, as depicted in Figure 36. This is the case for residual offset as well.

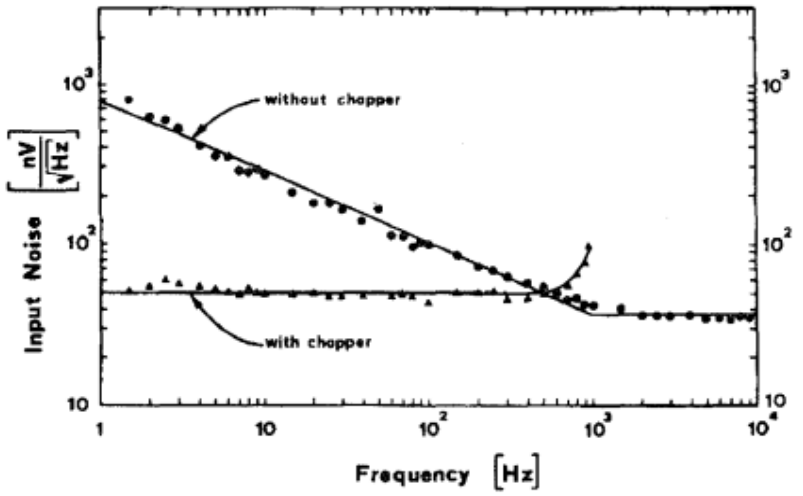


Figure 36 - Spectral noise density with and without chopper stabilization [68]

2.4.2.5. Offset-cancellation circuit implementations:

This section describes examples of offset-cancellation techniques seen above in some state-of-the-art circuits, and sense amplifiers in particular.

2.4.2.5.1. Open and Closed-Loop Offset Cancellation

The simplest implementation is the open-loop offset-cancellation circuit (see ***Figure 37***). During sampling phase, one input is connected to ground in order to register only amplifier's offset through sampling capacitance C. Input signal is then amplified and input offset voltage is cancelled with the one stored in C. A residual offset $V_{OS,res}$ is however still present, due to charge injection q_{inj} brought by non-ideal S1 switch: $V_{OS,res} = \frac{1}{A}.\frac{q_{inj}}{C}$. This technique can only be efficient if the amplifier gain is not high enough (A<10) to keep output voltage under minimum saturation voltage of the amplifier.

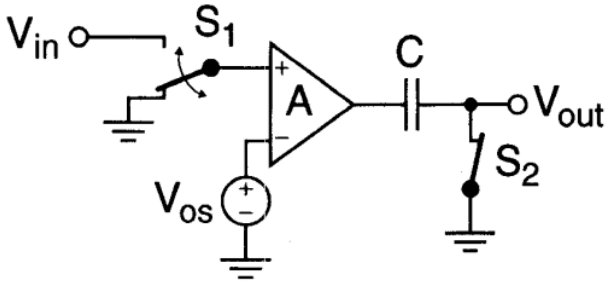


Figure 37 - Open-Loop offset cancellation configuration [68]

For high-gain amplifiers, it is more suitable to use the closed-loop configuration of ***Figure 38***. Saturation is in this case avoided by storing offset contributions through amplifier's input capacitance thanks to a feedback: $V_C = \frac{A}{1+A}.V_{OS} \cong V_{OS}\ since\ A \gg 1$. The efficiency of this technique is in this case limited by the following residual offset: $V_{OS,res} = \frac{V_{OS}}{A} + \frac{q_{inj}}{C}$. When a continuous-time amplification is required, it is more suitable to duplicate amplifier and use time-shared 'ping-pong' operation [69]. A closed-loop technique is thus advised for high-gain amplifiers whereas open-loop technique is more suitable for circuits like comparators.

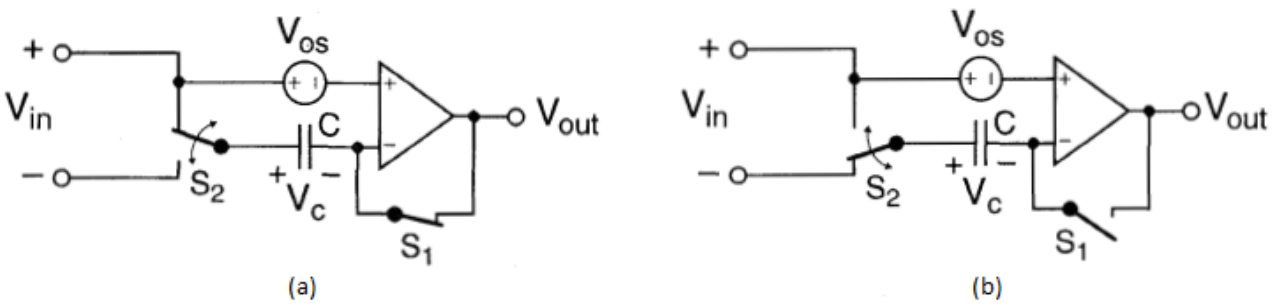


Figure 38 - Closed-Loop offset cancellation configuration (a) during sampling phase (b) amplification phase [68]

Residual offset can be even more reduced with a closed-loop offset compensation using an auxiliary input port (*Figure 39*). The idea is to reduce charge injection effects by storing offset contributions at an intermediate node instead of using a sampling capacitor at the amplifier's input.

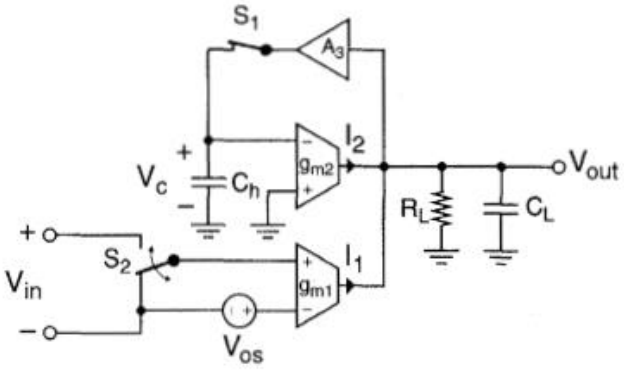


Figure 39 - Closed-Loop offset cancellation using an additional offset nulling input [68]

With this configuration, the voltage differential due to charge injection is reduced by A_1/A_2 where A_1 is the gain of the amplifier from input to output and A_2 the gain of the feedback amplifier (from the nulling input to output). That's why it is convenient to make A_1 much higher than A_2. These two amplifiers are often used as transconductance stages, as depicted in the figure. The amplifier of gain A_3 has the advantages to avoid autozeroing to slow down since g_{m1} amplifier output is no longer charged by means of voltage directly through C_h. Moreover, it leads to the reduction of S_1 switch charge injection effects and so the reduction of the overall residual offset $V_{OS\text{-}res}$, as described in the following equation:

$$V_{OS,res} \cong -\frac{V_{OS}}{A_2.A_3} - \frac{A_2}{A_1}.\frac{q_{inj}}{C_h}$$

Equation 7: Residual offset reduction of the closed-loop offset-cancellation configuration using an offset nulling input

2.4.2.5.2. Partial and Full Offset Cancellation Technique

This section presents applications of autozeroing technique specifically for sense amplifier offset reduction. Two state-of-the-art sense amplifier architectures with reduced offset are presented here, taken from [65].

As latch-type circuits are too sensitive to mismatch and presents a dynamic behavior, they are not convenient with offset-cancellation. A basic current mirror is thus introduced (*Figure 40* and *Figure 41*). Assuming for the moment a mid-reference, the sampling capacitor can be located between the two gates of PMOS transistors and inputs are short-circuited at the sampling phase. In this case, the following small signal analysis in both sampling and amplification phase shows that offset contributions brought by NMOS transistors cannot be

cancelled, as the current path used is different between the two phases. It is the partial offset cancellation technique (POCT):

$$V_{C,\varphi_1} = \frac{g_{m_N}}{g_{m_p}} . (\sigma V_{T_{n_2}} - \sigma V_{T_{n_1}})$$

Equation 8: Small signal sampling voltage during the first phase of the POCT-based sense amplifier

$$v_{out} = -g_{m_N} r_{out} . \left[\frac{\sigma V_{T_{n_1}}}{1 + g_{m_N} . R_{BIT}} - \frac{\sigma V_{T_{n_2}}}{1 + g_{m_N} . R_{REF}} - \left(\sigma V_{T_{n_1}} - \sigma V_{T_{n_2}} \right) \right]$$

Equation 9: Small signal output voltage at the end of the amplification phase of the POCT-based sense amplifier

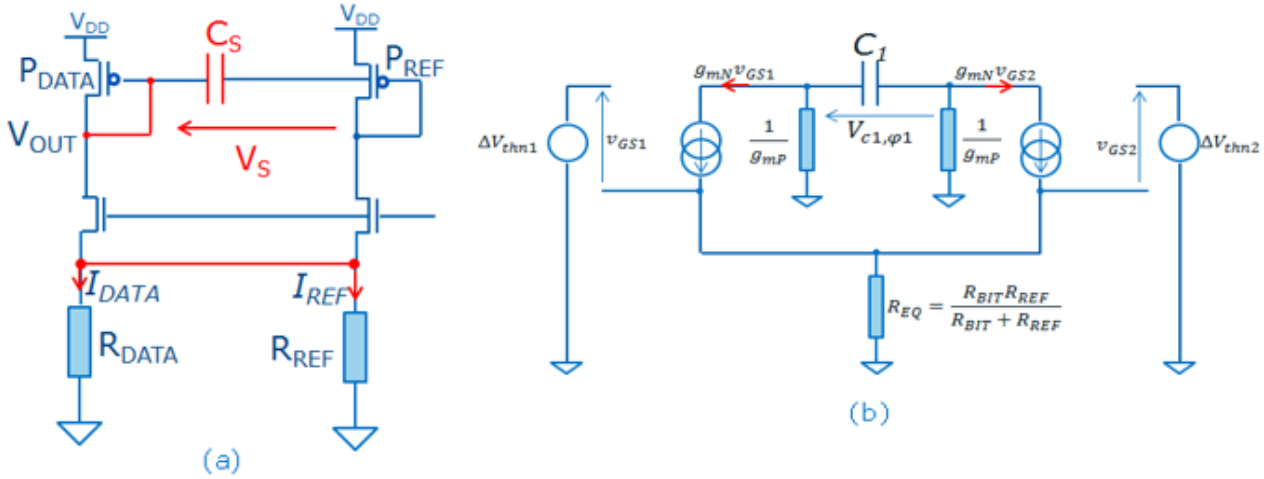


Figure 40 - (a) Current mirror-based sense amplifier = POCT implementation during sampling phase (b) small signal equivalent circuit

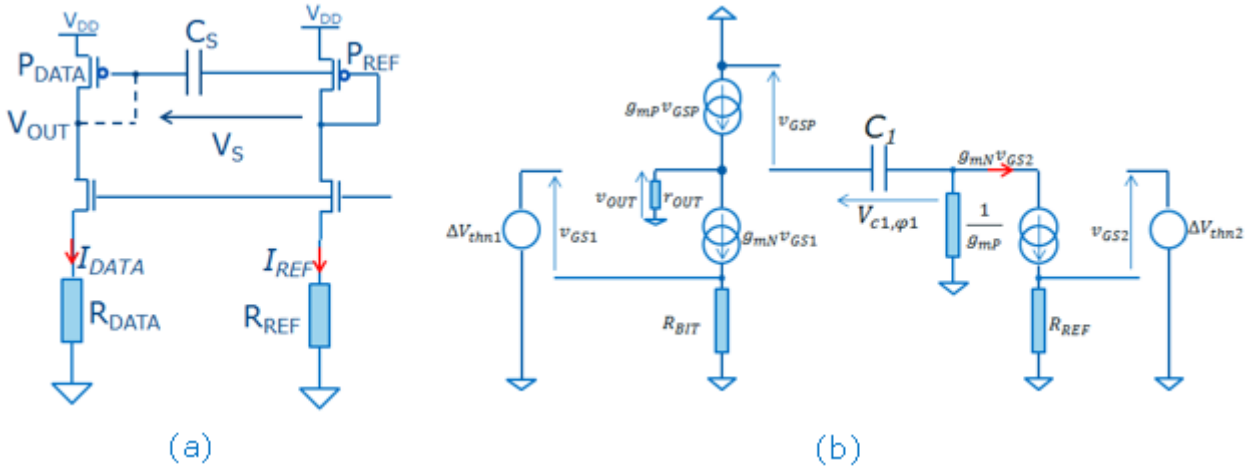


Figure 41 - (a) Current mirror-based sense amplifier = POCT implementation during amplification phase (b) small signal equivalent circuit

For a complete cancellation, the inputs are inverted between the two phases, and as demonstrated by the small signal analysis below, all transistors offset contributions are registered (when R_{data} gets closer to R_{ref}, i.e for a

narrow read window) and amplified through the same current path: it's the full offset cancellation technique (FOCT, see *Figure 42* and *Figure 43*).

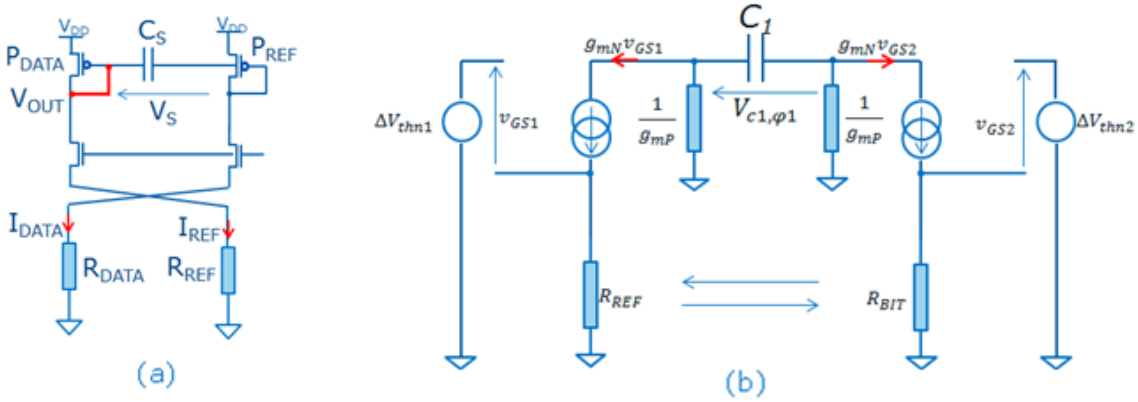


Figure 42 - (a) Full offset cancellation technique during sampling phase (b) Small signal equivalent circuit

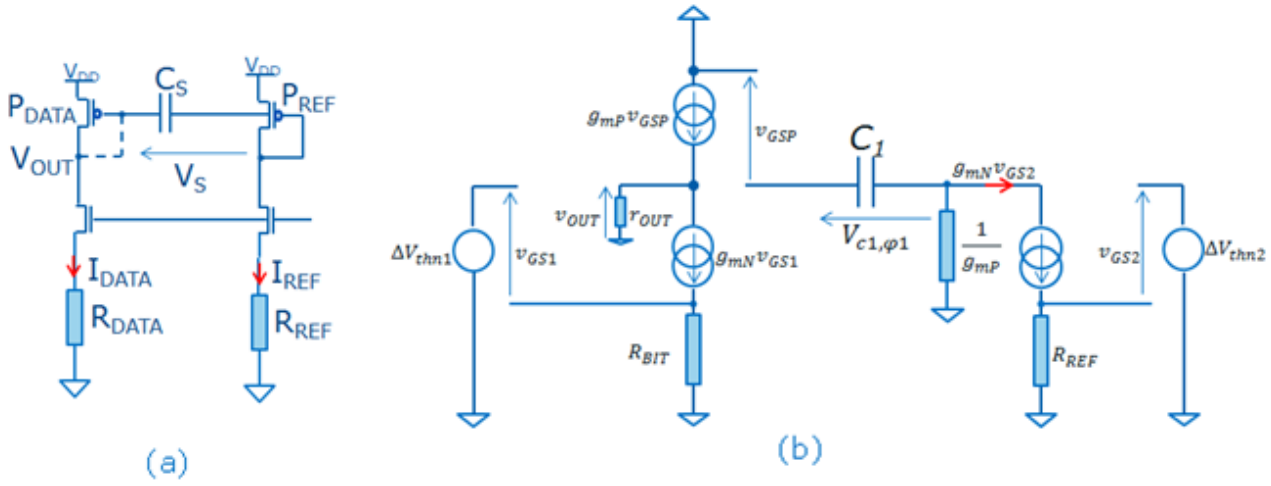


Figure 43 - (a) Full offset cancellation technique during amplification phase (b) Small signal equivalent circuit

$$V_{C_{1,\varphi_1}} = -\frac{g_{m_N}}{g_{m_P}}.\left(\frac{\sigma V_{T_{n_2}}}{1+g_{m_N}.R_{data}} - \frac{\sigma V_{T_{n_1}}}{1+g_{m_N}.R_{ref}}\right)$$

Equation 10: Small signal sampling voltage during the first phase of the FOCT-based sense amplifier

$$v_{out} = -g_{m_N}.r_{out}.\left(\begin{array}{l}\left[\frac{\sigma V_{T_{n_1}}}{1+g_{m_N}.R_{data}} - \frac{\sigma V_{T_{n_2}}}{1+g_{m_N}.R_{ref}}\right] \\ -\left[\frac{\sigma V_{T_{n_1}}}{1+g_{m_N}.R_{ref}} - \frac{\sigma V_{T_{n_2}}}{1+g_{m_N}.R_{data}}\right]\end{array}\right)$$

Equation 11: Small signal output voltage at the end of the amplification phase of the FOCT-based sense amplifier

The inversion of the inputs during the second phase also leads to a doubled signal as described by equations below:

$$V_{C_{1,\varphi_1}} = V_{T_{P_2}} - V_{T_{P_1}} + \sqrt{\frac{I_{data}}{K_P}} - \sqrt{\frac{I_{ref}}{K_P}}$$

Equation 12: Large signal sampling voltage during the first phase of the FOCT-based sense amplifier

$$v_{signal} = V_{dd} - V_{T_{P_1}} - \sqrt{\frac{I_{data}}{K_P}} - \left(V_{dd} - \sqrt{\frac{I_{ref}}{K_P}} - V_{T_{P_2}} + V_{C_{1,\varphi_1}} \right)$$
$$= \mathbf{2}. \sqrt{\frac{1}{K_P}} . \left[\sqrt{I_{ref}} - \sqrt{I_{data}} \right]$$

Equation 13: Large signal output voltage at the end of the amplification phase of the FOCT-based sense amplifier

The FOCT appears to be a suitable read-yield enhancement technique for a mid-reference scheme. However, this implementation does not allow dense and homogenous memory array integration, due to the resulting design complexity, as explained above.

Thus, current literature does not propose solutions that combine both layout regular and accurate mid-reference scheme on one hand, and reliable offset cancellation technique in order to get the best signal to offset ratio and so improve read yield of resistive memories.

2.5. Conclusion

This chapter explains why emerging resistive memories and MRAM in particular, are promising technologies in terms of standalone performance, compared to existing solutions. STT-MRAM is a well-advanced technology at process-level. One of the biggest challenge of resistive memories is their low design robustness to ensure higher read reliability. It is important to propose design-level solutions of sense amplifiers architectures that simultaneously reduce or cancel its offset and include an accurate and layout-regular reference scheme. Current literature proposes solutions that handle layout regularity of reference schemes like the CBSA circuit, and reliable offset cancellation techniques like FOCT in order to get the best signal to offset ratio. This thesis gives a design-level analysis of the variability problematic for the design of read circuits for resistive memories and proposes sense amplifier architecture-solutions that combine both offset-cancellation and reduced reference variability.

References of Chapters 1 and 2

[1] G. E. Moore, “Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.,” *IEEE Solid-State Circuits Soc. Newsl.*, vol. 11, no. 3, pp. 33–35, Sep. 2006.

[2] M. Bohr, “The new era of scaling in an SoC world,” in *2009 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, 2009, pp. 23–28.

[3] D. Sylvester, K. Agarwal, and S. Shah, “Variability in nanometer CMOS: Impact, analysis, and minimization,” *Integr. VLSI J.*, vol. 41, no. 3, pp. 319–339, May 2008.

[4] M.Roser and H.Ritchie, “Technological Progress,” *Online at OurWorldinData.org*, 2017. .

[5] E.Grochowski and T.Coughlin, “The Perennial Hard Disk Drive Storage Industry ‘Workhorse,’” in *SNIA Data Storage Innovation Conference (DSI)*, 2015.

[6] H. H. Takasu, “‘More than Moore’ expands the semiconductor world,” in *2016 International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*, 2016, pp. 1–2.

[7] Y. Fujisaki, “Review of Emerging New Solid-State Non-Volatile Memories,” *Jpn. J. Appl. Phys.*, vol. 52, no. 4R, p. 40001, Apr. 2013.

[8] W. Zhang and T. Li, “Characterizing and mitigating the impact of process variations on phase change based memory systems,” in *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture - Micro-42*, 2009, p. 2.

[9] D. Garbin, Q. Rafhay, E. Vianello, S. Jeannot, P. Candelier, B. DeSalvo, G. Ghibaudo, and L. Perniola, “Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network,” in *EUROSOI-ULIS 2015: 2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon*, 2015, pp. 125–128.

[10] Y. Zhang, Y. Li, Z. Sun, H. Li, Y. Chen, and A. K. Jones, “Read Performance: The Newest Barrier in Scaled STT-RAM,” *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 23, no. 6, pp. 1170–1174, Jun. 2015.

[11] K. Prall and K. Parat, “25nm 64Gb MLC NAND technology and scaling challenges invited paper,” in *2010 International Electron Devices Meeting*, 2010, p. 5.2.1-5.2.4.

[12] D. Wouters, “Resistive switching materials and devices for future memory applications,” in *43rd IEEE Semiconductor Interface Specialists Conference*, 2012, no. December.

[13] A. Gokce, I. Cinar, S. C. Ozdemir, E. Cogulu, B. Stipe, J. A. Katine, and O. Ozatay, “Toward Multiple-Bit-Per-Cell Memory Operation With Stable Resistance Levels in Phase Change Nanodevices,” *IEEE Trans. Electron Devices*, vol. 63, no. 8, pp. 3103–3108, Aug. 2016.

[14] P. Stoliar, P. Levy, M. J. Sanchez, A. G. Leyva, C. A. Albornoz, F. Gomez-Marlasca, A. Zanini, C. Toro Salazar, N. Ghenzi, and M. J. Rozenberg, “Nonvolatile multilevel resistive switching memory cell: A transition metal oxide-based circuit,” *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 61, no. 1, pp. 21–25, 2014.

[15] A. Vatankhahghadim and A. Sheikholeslami, “A Multi-level Cell for STT-MRAM with Biaxial Magnetic Tunnel Junction,” in *2015 IEEE International Symposium on Multiple-Valued Logic*, 2015, pp. 158–163.

[16] Y. Roizin, “Resistive Memories Promising for Industrial Applications,” in *ACRC Workshop on Memristors and Resistive Memory Devices*, 2012, no. February.

[17] M. Son, J. Lee, J. Park, J. Shin, G. Choi, S. Jung, W. Lee, S. Kim, S. Park, and H. Hwang, “Excellent Selector Characteristics of Nanoscale $ \hbox{VO}_{2}$ for High-Density Bipolar ReRAM Applications,” *IEEE Electron Device Lett.*, vol. 32, no. 11, pp. 1579–1581, Nov. 2011.

[18] W. Wang, “Transistor-Less Spin Torque Transfer Magnetic Random Access Memory Cell Design,” *IEEE Trans. Magn.*, vol. 51, no. 11, pp. 1–4, Nov. 2015.

[19] Intel and Micron, “Intel and Micron produce breakthrough memory technology,” *Intel Website*, 2015. [Online]. Available: http://newsroom.intel.com/%5Cncommunity/intel_newsroom/blog/2015/07/28/%5Cnintel-and-micron-produce-breakthrough-memory-technology.

[20] Mark Lapedus, “Sorting Out Next-Gen Memory,” *Semiconductor Engineering*, 2016. .

[21] L. Perniola, “Other resistive RAM technologies: PCRAM, OxRAM, CBRAM and respective applications,” in *InMRAM*, 2015.

[22] Y. de Charentenay, “Emerging NVM market: Technological choices are about to be made by key players - SSTMRAM, RRAM... or PCM? Focus on Embedded Markets,” in *e-NVM*, 2015.

[23] M.Borghi, “Endurance performances of Ge-rich chalcogenide for embedded automotive applications,” in *EPCOS*, 2015.

[24] V. Sousa, G. Navarro, N. Castellani, M. Coue, O. Cueto, C. Sabbione, P. Noe, L. Perniola, S. Blonkowski, P. Zuliani, and R. Annunziata, “Operation fundamentals in 12Mb Phase Change Memory based on innovative Ge-rich GST materials featuring high reliability performance,” in *2015 Symposium on VLSI Technology (VLSI Technology)*, 2015, pp. T98–T99.

[25] L.Wang, “L.Wan,” in *Nanchang University*, 2015.

[26] A. Belmonte, W. Kim, B. Chan, N. Heylen, A. Fantini, M. Houssa, M. Jurczak, and L. Goux, “90nm W\Al<inf>2</inf>O<inf>3</inf>\TiW\Cu 1T1R CBRAM cell showing low-power, fast and disturb-free operation,” in *2013 5th IEEE International Memory Workshop*, 2013, pp. 26–29.

[27] E. Vianello, O. Thomas, G. Molas, O. Turkyilmaz, N. Jovanovic, D. Garbin, G. Palma, M. Alayan, C. Nguyen, J. Coignus, B. Giraud, T. Benoist, M. Reyboz, A. Toffoli, C. Charpin, F. Clermidy, and L. Perniola, “Resistive Memories for Ultra-Low-Power embedded computing design,” in *2014 IEEE International Electron Devices Meeting*, 2014, p. 6.3.1-6.3.4.

[28] M. Julliere, “Tunneling between ferromagnetic films,” *Phys. Lett. A*, vol. 54, no. 3, pp. 225–226, Sep. 1975.

[29] E. Deng, “Design and development of low-power and reliable logic circuits based on spin-transfer torque magnetic tunnel junctions,” 2017.

[30] S. Ikeda, J. Hayakawa, Y. Ashizawa, Y. M. Lee, K. Miura, H. Hasegawa, M. Tsunoda, F. Matsukura, and H. Ohno, “Tunnel magnetoresistance of 604% at 300K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature,” *Appl. Phys. Lett.*, vol. 93, no. 8, p. 82508, Aug. 2008.

[31] W. Zhao, E. Belhaire, Q. Mistral, C. Chappert, V. Javerliac, B. Dieny, and E. Nicolle, “Macro-model of Spin-Transfer Torque based Magnetic Tunnel Junction device for hybrid Magnetic-CMOS design,” in *2006 IEEE International Behavioral Modeling and Simulation Workshop*, 2006, pp. 40–43.

[32] L.-B. Faber, W. Zhao, J.-O. Klein, T. Devolder, and C. Chappert, “Dynamic compact model of Spin-Transfer Torque based Magnetic Tunnel Junction (MTJ),” in *2009 4th International Conference on Design & Technology of Integrated Systems in Nanoscal Era*, 2009, pp. 130–135.

[33] W. Zhao, J. Duval, J.-O. Klein, and C. Chappert, “A compact model for magnetic tunnel junction (MTJ) switched by thermally assisted Spin transfer torque (TAS + STT),” *Nanoscale Res. Lett.*, vol. 6, no. 1, p. 368, 2011.

[34] J. C. Slonczewski, “Current-driven excitation of magnetic multilayers,” *J. Magn. Magn. Mater.*, vol. 159, no. 1–2, pp. L1–L7, Jun. 1996.

[35] L. Berger, “Emission of spin waves by a magnetic multilayer traversed by a current,” *Phys. Rev. B*, vol. 54, no. 13, pp. 9353–9358, Oct. 1996.

[36] A. V Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R. S. Beach, A. Ong, X. Tang, A. Driskill-Smith, W. H. Butler, P. B. Visscher, D. Lottis, E. Chen, V. Nikitin, and M. Krounbi, “Basic principles of STT-MRAM cell operation in memory arrays,” *J. Phys. D. Appl. Phys.*, vol. 46, no. 7, p. 74001, Feb. 2013.

[37] D. Apalkov, A. Ong, A. Driskill-Smith, M. Krounbi, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, and E. Chen, “Spin-transfer torque magnetic random access memory (STT-MRAM),” *ACM J. Emerg. Technol. Comput. Syst.*, vol. 9, no. 2, pp. 1–35, May 2013.

[38] B. Dieny, “STT-MRAM,” in *InMRAM*, 2015.

[39] N. Amir, “Phase-Change Memory,” in *ACRC Workshop on Memristors and Resistive Memory Devices*, 2012.

[40] A. Agarwal, D. Blaauw, and V. Zolotov, “Statistical timing analysis for intra-die process variations with spatial correlations,” *#Proc_Iccad#*, pp. 900–907, 2003.

[41] A. Singhee and R. A. Rutenbar, *Extreme Statistics in Nanoscale Memory Design*. Spinger US, 2010.

[42] R. Errors and S. Errors, “Random vs Systematic Error,” *Physics.umd.edu*, 2011. [Online]. Available: https://www.physics.umd.edu/courses/Phys276/Hill/Information/Notes/ErrorAnalysis.html.

[43] E. Healthcare and E. M. Guantai, “Random Error and Systematic Error,” 2015. [Online]. Available: https://www2.southeastern.edu/Academics/Faculty/rallain/plab193/labinfo/Error_Analysis/05_Random_vs_Systematic.html.

[44] Alexandre Levisse, “3D HIGH DENSITY MEMORY BASED ON EMERGING RESISTIVE TECHNOLOGIES: CIRCUIT AND ARCHITECTURE DESIGN,” 2017.

[45] Wikepedia, “Normal distribution,” *Wikipedia, The Free Encyclopedia*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Normal_distribution.

[46] Heechai Kang, Jisu Kim, Hanwool Jeong, Young Hwi Yang, and Seong-Ook Jung, “Architecture-Aware Analytical Yield Model for Read Access in Static Random Access Memory,” *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 23, no. 4, pp. 752–765, Apr. 2015.

[47] B. Razavi, *Design of analog CMOS integrated circuits*. McGRAW-HILL International Edition, Electrical Engineering Series, 2004.

[48] B. C. Berndt, R. J. Evans, and K. S. Williams, *Gauss and Jacobi Sums*. Wiley and Sons, 1998.

[49] I. Miller and A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, vol. 8, no. 2. McGRAW-HILL, 1966.

[50] Harvard University, "A summary of error propagation," in *Physical Sciences 2*, no. 3, Harvard University, 2007, p. 5.

[51] S. Volz, "Monte Carlo Method," *Flux*, 2007. [Online]. Available: https://en.wikipedia.org/wiki/Monte_Carlo_method.

[52] L. Wilkinson, "Revising the Pareto Chart," *Am. Stat.*, vol. 60, no. 4, pp. 332–334, Nov. 2006.

[53] A. I. Khuri and S. Mukhopadhyay, "Response surface methodology," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 2, pp. 128–149, 2010.

[54] Wang Kang, Liuyang Zhang, J.-O. Klein, Youguang Zhang, D. Ravelosona, and Weisheng Zhao, "Reconfigurable Codesign of STT-MRAM Under Process Variations in Deeply Scaled Technology," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1769–1777, Jun. 2015.

[55] E. Detection, "Error Detection and Correction," *Wikipedia*, pp. 1–7, 2008.

[56] G.Montcouquiol, "Codes détecteurs et correcteurs d'erreurs," 2007.

[57] "ECC memory," *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/ECC_memory.

[58] "AMD-762™ System Controller Software/BIOS Design Guide, p. 179." .

[59] "Benchmark of AMD-762/Athlon platform with and without ECC," *Wayback Machine*, 2013. .

[60] Mehdi Tahoori, "Beyond MRAM, CMOS/Mag integrated electronics," in *InMRAM*, 2015.

[61] T. M. Maffitt, J. K. DeBrosse, J. a. Gabric, E. T. Gow, M. C. Lamorey, J. S. Parenteau, D. R. Willmott, M. A. Wood, and W. J. Gallagher, "Design considerations for MRAM," *IBM J. Res. Dev.*, vol. 50, no. 1, pp. 25–39, Jan. 2006.

[62] A. Asenov, "Statistical Nano CMOS Variability and Its Impact on SRAM," in *Extreme Statistics in Nanoscale Memory Design*, Boston, MA: Springer US, 2010, pp. 17–49.

[63] T. Na, J. P. Kim, S. H. Kang, and S.-O. Jung, "Multiple-Cell Reference Scheme for Narrow Reference Resistance Distribution in Deep Submicrometer STT-RAM[1] T. Na, J. P. Kim, S. H. Kang, and S.-O. Jung, "Multiple-Cell Reference Scheme for Narrow Reference Resistance Distribution in Deep Submicrometer S," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 24, no. 9, pp. 2993–2997, Sep. 2016.

[64] C. Kim, K. Kwon, C. Park, S. Jang, and J. Choi, "7.4 A covalent-bonded cross-coupled current-mode sense amplifier for STT-MRAM with 1T1MTJ common source-line structure array," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1–3.

[65] E. M. Boujamaa and C. Dray, "System and Methods using a multiplexed reference for sense amplifiers," 2012.

[66] M. P. Sol, "Layout Regularity for Design and Manufacturability," 2012.

[67] and H. S. S. S. Y. Lee, H. J. Kim, S. J. Lee, "A New Reference Cell for 1T-1MTJ MRAM," *J. Semicond. Technol. Sci.*, vol. 4, pp. 110–116, 2004.

[68] C. C. Enz and G. C. Temes, “Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization,” *Proc. IEEE*, vol. 84, no. 11, pp. 1584–1614, 1996.

[69] Chong-Gun Yu and R. L. Geiger, “An automatic offset compensation scheme with ping-pong control for CMOS operational amplifiers,” *IEEE J. Solid-State Circuits*, vol. 29, no. 5, pp. 601–610, May 1994.

Chapter 3: Offset Analysis and Design Optimization of a Dynamic Sense Amplifier for resistive memories

3.1. Introduction

The previous chapter provides elements to understand the importance of variation-tolerance for the design of sense amplifier circuits.
This chapter offers a complete analysis of the read margin all along the sensing path of the resistive memory, and gives a methodology to optimize at design-level read window degradations:

- A statistical model is first proposed, evaluating read margins at the sense amplifier input, in order to minimize degradation due to variability. This model includes a design optimizer for an area-aware resizing of bit line voltage clamping transistors.
- The offset of the CBSA circuit introduced above is analyzed, modelled and optimized, in order to put forward the different variation-increase factors. An offset-characterization methodology is also proposed.

3.2. Statistical model describing read margin degradation at bit cell level

In this section, a statistical model is proposed, in order to evaluate the impact of variability on read margin degradation at the beginning of the resistive memory-sensing path. It is important at this point to remind the reader that this study is achieved from a design point of view and for a further optimization of sense amplifier architecture.
Figure 44 describes the circuit and the different parameters that are considered for the model:

- R_{DATA} is the memory bit cell resistance. The resistive technology can be one of the three described above (PCM, ReRAM or MRAM), by introducing a realistic resistive cell model.
- σR_{DATA}: the memory bit cell's variability is quantified by its resistance's standard deviation. The main source of variability of switching resistance cells is cycle-to-cycle caused by the stochastic behavior of the devices [70][71]. This behavior can be included in the model but it is not the focus of interest of this study, as the main point is to optimize circuit architectures considering the bit cell and its variability (i.e the random variable $R_{DATA}(\mu,\sigma)$ is sufficient here).
- $\Sigma R_{PAR} = R_{BL}+R_{SL}+R_{accesdevice}$: this parameter includes bit line and source line parasitics (respectively R_{BL} and R_{SL}) as well as the access-device equivalent resistance ($R_{accesdevice}$). This access device can either be a transistor or a diode or any other selector [17][18][19][72][73][74][75]. Hence, layout effects and their resulting variability are taking into account through this random variable (ΣR_{PAR}, $\sigma\Sigma R_{PAR}$).
- NCLAMP is bit line voltage clamping MOS transistor, working as a source follower (kept at saturated regime) and so setting via V_{CLAMP} signal the initial bit line voltages at the sense amplifier inputs. Its sources of variations are multiple, among them line-edge roughness, random dopant fluctuations and interface-state density fluctuations. This results in parameters variation, mainly its threshold voltage $V_T(\mu,\sigma)$.

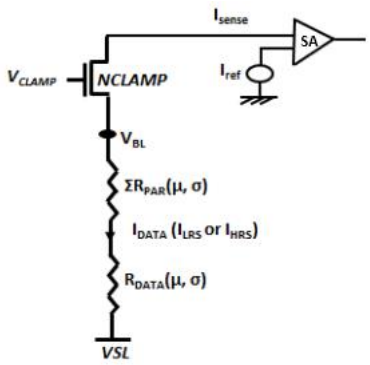


Figure 44 - Equivalent circuit of the sensing path

The memory read window can be defined as the difference between two current read margins, for respectively LRS and HRS:

$$RM_0 = \mu(\Delta I_{LRS}) - N.\sigma(\Delta I_{LRS})$$ *Equation 14: Read margin for LRS*

$$RM_1 = \mu(\Delta I_{HRS}) - N.\sigma(\Delta I_{HRS})$$ *Equation 15: Read margin for HRS*

with: $\mu(\Delta I_{LRS}) = I_{LRS} - I_{REF}$; $\mu(\Delta I_{HRS}) = I_{HRS} - I_{REF}$; $\sigma(\Delta I_{LRS}) = \sqrt{\sigma I_{LRS}^2 + \sigma I_{REF}^2}$; $\sigma(\Delta I_{HRS}) = \sqrt{\sigma I_{HRS}^2 + \sigma I_{REF}^2}$ (since data and reference resistances are uncorrelated) and $I_{REF} = \frac{I_{LRS}+I_{HRS}}{2}$. N quantifies the overall variation rate applied to the circuit (expressed in number of sigma).

3.2.1.Estimation of the read margin degradation

Assuming first an ideal NCLAMP ($\sigma V_{T=} \sigma V_{BL}=0$), and so a fixed bit line voltage, data current mean value can be expressed from *Figure 44* as:

$$I_{DATA} = \frac{V_{BL}}{R_{DATA} + \Sigma R_{PAR}}$$ *Equation 16: Current through the resistive bit cell*

Applying the theorem of propagation of uncertainty [49][50][76][77], and as R_{DATA} and ΣR_{PAR} are uncorrelated variables, its variability σI_{DATA} is:

$$\frac{\sigma I_{DATA}}{I_{DATA}} = \frac{\sqrt{\sigma^2 R_{DATA} + \sigma^2 \Sigma R_{PAR}}}{R_{DATA} + \Sigma R_{PAR}}$$ *Equation 17: Data current variability assuming ideal bit line clamping transistor*

In order to depict the read margins degradation, the non-linearity of the resistive device with voltage has to be included in the model. The following results show estimations with magnetic tunnel junctions modelling [31][32][33][78][79](see *Equation 18*) but all the following is applicable for any other memristor model [80][81]. The tunnelling magnetoresistance (TMR) ratio (dimensionless quantity) evaluates the ability of the memory cell to distinguish the two complementary resistive states: $TMR = \frac{R_{HRS}-R_{LRS}}{R_{LRS}}$.

$$TMR(V) = \frac{TMR_0}{1 + \frac{V}{V_h}}$$ *Equation 18: Non-linearity of magnetic tunnel junctions with voltage*

where TMR_0 is the TMR ratio with zero bias voltage anV_h is the bias voltage when $TMR(V_h) = 0.5 * TMR_0$.

The degradation of the read margins will be illustrated thereafter for specific parameter values:

- R_{LRS} is arbitrary assumed equal to 4 kΩ
- V_{BL} is first ideally assumed fixed and equal to 180 mV in order not to consider NCLAMP transistor dependence
- $\sigma\Sigma R_{PAR}$: Layout effects are considered less variable than the resistive device: $\frac{\sigma\Sigma R_{PAR}}{\Sigma R_{PAR}} = \frac{1}{6}.\frac{\sigma R_{DATA}}{R_{DATA}}$ is arbitrary chosen.

The degradation of the read margins given in *Equation 14* and *Equation 15* with respect to parasitic resistances is illustrated in *Figure 45* varying the global variation rate applied to the circuit in *Figure 44*, for an ideal non-variable clamping transistor. Memory read yield has to be optimized with specific care on Gaussian distribution tails and so for a maximum applied variation. This result helps targeting the maximum N parameter keeping a positive and acceptable read margin RM_0-RM_1, for specific resistive device parameters ($\sigma R_{LRS}/R_{LRS}$=5% and TMR=100% are arbitrary chosen here; note that $\sigma R_{HRS} = \sigma R_{LRS} * (1 + TMR)$).

A read margin stabilization is observed (for four sigma global variation) for LRS, which is less variable than HRS. Hence this result helps get into position at process level (taking into account data and access device technologic choice and layout effects) for a further design-level study. *Figure 46* shows this positioning regarding process parameters (σR_{DATA}, ΣR_{PAR}).

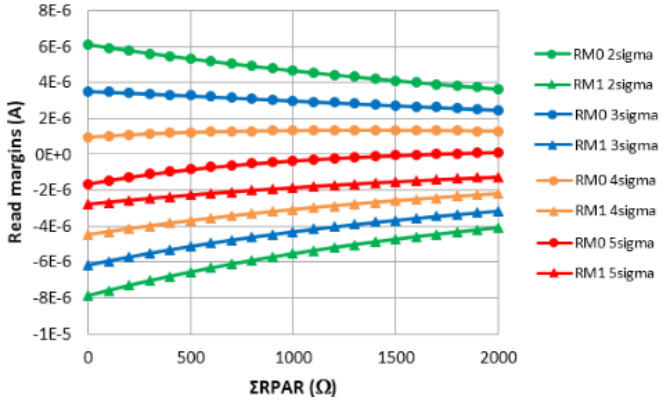


Figure 45 - Read margin degradation w.r.t parasitic resistances for different global variation rate (R_{LRS}=4kΩ, TMR=100%, V_{BL}=180mV, $\sigma R_{LRS}/R_{LRS}$=5%)

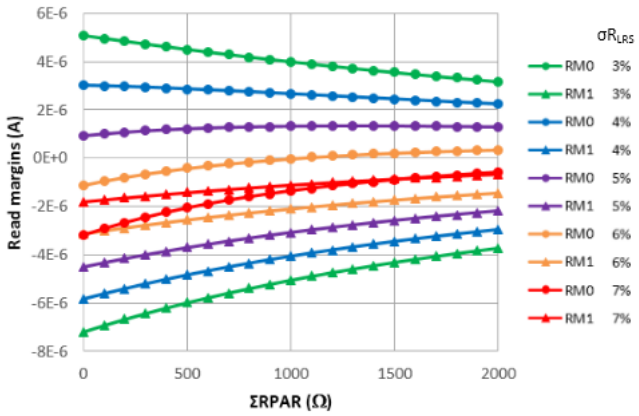


Figure 46 - Read margin degradation w.r.t parasitic resistances for different data resistance variations (R_{LRS}=4kΩ, TMR=100%, V_{BL}=180mV, N=4)

After this setting, it is important to choose a correct bit line voltagerange, taking into account the non-linearity of the resistive device with voltage (*Equation 18*). *Figure 47* shows for example that for $\sigma R_{LRS}/R_{LRS}$=5%, ΣR_{PAR}=500Ω, the bit line voltage has to be set through NCLAMP's gate voltage (V_{CLAMP}) under 200mV to minimize read margin degradation.

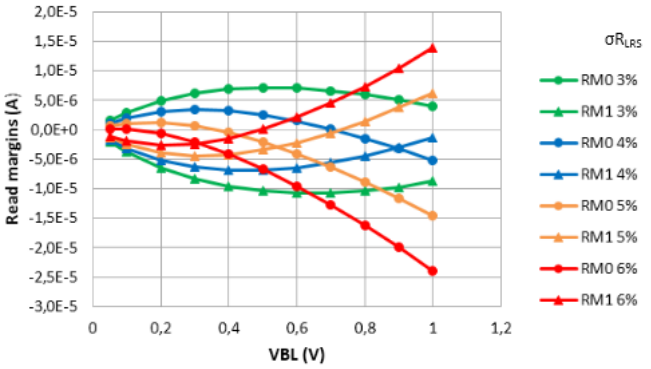


Figure 47 - Read margin degradation w.r.t bit line voltage for different data resistance variations (R_{LRS}=4kΩ, ΣR_{PAR}=500Ω, N=4)

A serious imbalance between RM0 and RM1 ($RM_0 \neq -RM_1$) is otherwise noticeable, that can be solved by modifying the division factor of the reference current, and thus, by changing the configuration of the reference scheme. Another solution is to use a sense amplifier increasing the read window (but with a reference scheme still leading to an imbalance), and cancel the imbalance with an appropriate offset stage (illustrated in *Figure 48*).

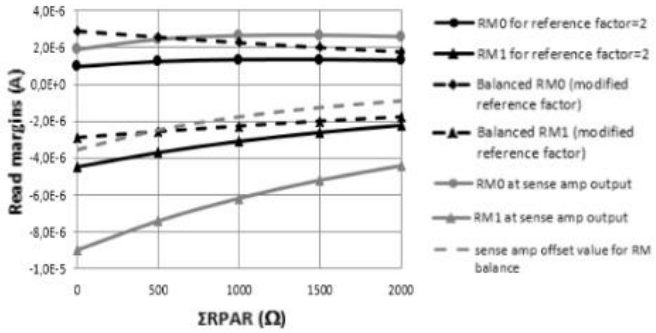


*Figure 48 - Read margin optimization by modifying the reference factor or using an appropriate sense amplifier transfer function (e.g: μ(ΔI)= 2*IDATA-ILRS – IHRS) (RLRS=4kΩ, TMR=100%, VBL=180mV, N=4, σRLRS/RLRS=5%)*

3.2.2. Impact of the bit line voltage clamping transistor's variability on read margin degradation

Taking now into account bit line voltage variability ($\sigma V_T \neq 0$), I_{DATA} is submitted to a negative feedback brought by NCLAMP: an increase in V_{BL} leads to an increase in the current through ΣR_{PAR} and R_{DATA}; while the saturated NCLAMP transistor makes it decrease as we can see in *Equation 19*.

$$I_{DATA} = \frac{V_{BL}}{R_{DATA} + \Sigma R_{PAR}} = \frac{\mu_n C_{ox}}{2} . \frac{W}{L} . (V_{CLAMP} - V_{BL} - V_T)^2$$

Equation 19: Negative feedback brought by NCLAMP on data current

Solving the resulting second order equation, the bit line voltage required is:

$$V_{BL} = \frac{2\alpha(V_{CLAMP} - V_T) + 1 - \sqrt{\Delta}}{2\alpha}$$

Equation 20: Bit line voltage assuming non-ideal NCLAMP

with: μ_n=electron mobility, C_{ox}=NCLAMP's oxide capacitance, W=NCLAMP's width, L=NCLAMP's channel length, $\alpha = \frac{\mu_n C_{ox}}{2}.\frac{W}{L}.(R_{MTJ} + \Sigma R_{PAR})$, $\Delta = 4\alpha(V_{CLAMP} - V_T) + 1$.

As standard deviation describes a small variation around a mean value, bit line voltage variability σV_{BL} can be expressed with small signal analysis, representing each variation source by a small signal voltage source. From *Figure 49* and using the superposition principle:

$$\sigma V_{BL} = \sqrt{v_{bl_1}^2 + v_{bl_2}^2 + v_{bl_3}^2}$$

Equation 21: Bit line voltage variability obtained by small signal analysis

with:

$$v_{bl_1} = \frac{r_{ds}}{r_{ds} + R_{DATA} + \Sigma R_{PAR} + g_m.r_{ds}.(R_{DATA} + \Sigma R_{PAR})}.\sigma\Sigma R_{PAR}.i_{DATA}$$

Equation 22: Impact of parasitics variations on bit line voltage variability

$$v_{bl_2} = \frac{r_{ds}}{r_{ds} + R_{DATA} + \Sigma R_{PAR} + g_m.r_{ds}.(R_{DATA} + \Sigma R_{PAR})}.\sigma R_{DATA}.i_{DATA}$$

Equation 23: Impact of data resistance variations on bit line voltage variability

$$v_{bl_3} = \frac{g_m.r_{ds}.(R_{DATA} + \Sigma R_{PAR})}{r_{ds} + R_{DATA} + \Sigma R_{PAR} + g_m.r_{ds}.(R_{DATA} + \Sigma R_{PAR})}.\sigma V_T$$

Equation 24: Impact of NCLAMP threshold voltage variations on bit line voltage variability

where v_{bl_1} is the small signal bit line voltage when $\sigma R_{DATA} = \sigma V_T = 0$; v_{bl_2} is the small signal bit line voltage when $\sigma\Sigma R_{PAR} = \sigma V_T = 0$; v_{bl_3} is the small signal bit line voltage when $\sigma\Sigma R_{PAR} = \sigma R_{DATA} = 0$.
g_m is NCLAMP transconductance and r_{ds} is NCLAMP output resistance.

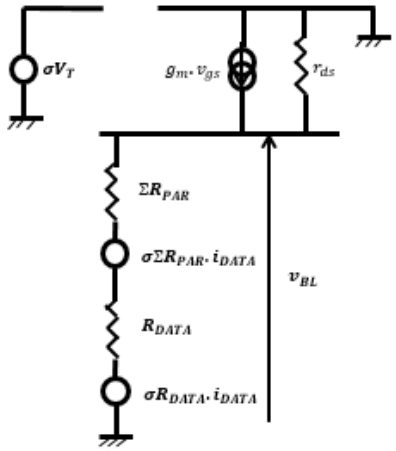


Figure 49 - Small signal equivalent circuit for bit line voltage's variability estimation

Equation 17 thus becomes:

$$\frac{\sigma I_{DATA}}{I_{DATA}} = \sqrt{\left(\frac{\sigma V_{BL}}{V_{BL}}\right)^2 + \frac{\sigma^2 R_{DATA} + \sigma^2 \Sigma R_{PAR}}{(R_{DATA} + \Sigma R_{PAR})^2}}$$

Equation 25: Data current variability assuming non-ideal bit line clamping transistor

σV_{BL} reflects the impact of resistive device (σR_{DATA}), parasitic resistance ($\sigma\Sigma R_{PAR}$) and NCLAMP (σV_T) variabilities on read margin degradation. *Equation 21* to *Equation 24* show that NCLAMP design optimization of σV_{BL} involves optimizing the area-dependent NCLAMP parameters g_m, r_{ds} and σV_T (see section below):

$$g_m = \mu_n C_{ox} \frac{W}{L} (V_{CLAMP} - V_{BL} - V_T)$$

Equation 26: NCLAMP's transconductance dependence with area

$$\sigma V_t = \frac{A_{V_T}}{\sqrt{W * L}}$$

Equation 27: NCLAMP's threshold voltage variability dependence with area

where $A_{V_T} = 3.19 * 10^{-8} \left(t_{ox} + \frac{z_0 \varepsilon_{ox}}{\varepsilon_{si}}\right) N_A^{0.4}$. t_{ox} is the gate dielectric thickness, z_0 is the inversion layer charge centroid, ε_{ox} and ε_{si} are the dielectric constants of the gate dielectric and the silicon, N_A is the channel doping concentration [62].

Figure 50 shows, for suitable parameters settings as described in previous subsection, a comparison of the impact on read margins degradations between resistive device and bit line clamping transistor's variabilities. Although from *Figure 50*, the memristor has the biggest impact on the degradation, it is important to reduce NCLAMP influence, as this optimization study is made from a design point a view. *Equation 22* to *Equation 24* highlight two design parameters playing a significant role in read margin degradation: g_m, which is proportional to NCLAMP's W/L ratio, and σV_T, which is inversely proportional to NCLAMP's area. The trade-off between these two parameters is discussed below.

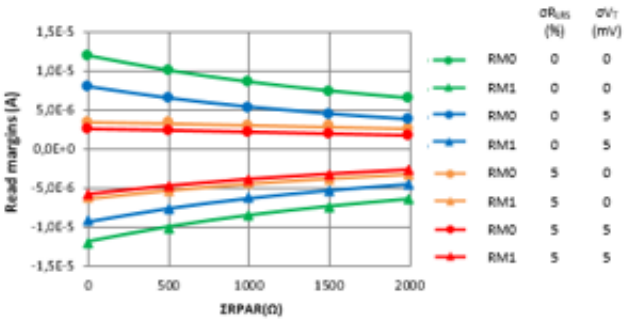


Figure 50 - Impact of σR_{DATA} and σV_T on read margin degradation (R_{LRS}=4kΩ, TMR=100%, N=3, $\sigma R_{DATA}/R_{DATA}$=5%, V_T=0.25V, V_{CLAMP}=0.5V, W/L=42)

3.2.3. Methodology for design optimization

This section defines a methodology of the different steps for the read margins evaluations and explains the principle of the design optimizer included in order to further boost NCLAMP for a minimum read window degradation.

The different steps to follow for a design optimization of the resistive memory read margin are:

- Choose the correct variation rate to apply to the whole circuit (bit cell stage plus sense amplifier)
- Process parameters settings: consider given process specifications (σR_{DATA}, $\sigma \Sigma R_{PAR}$) and position themselves regarding read margin degradation
- Consider non-linearity of the resistive device through bit line voltage clamping transistor
- Compare resistive device and NCLAMP's variability impact on read margin degradation
- Area-aware NCLAMP sizing that minimize read margin degradation.

A design optimizer is proposed to achieve the last step. The analysis of *Equation 14*, *Equation 15*, and *Equation 21* to *Equation 25* shows that for read margin reduction, σI_{DATA} and so σV_{BL} have to be reduced. σV_{BL} can be reduced either by increasing g_m to minimize resistances' variability contributions, or by decreasing σV_T to minimize NCLAMP's variability contribution. *Figure 51* describes this read margin dependence with g_m (proportional to W/L) and σV_T. As bit line voltage also depends on W/L, it is kept fixed, in order to focus only

on g_m variation with W/L, by adapting V_{CLAMP} value for each W/L variation. *Equation 21* to *Equation 24* show that σV_{BL} decreases with g_m until reaching a constant value. Thus, read margin increases with W/L and then saturates. Therefore, a maximum read margin is ideally obtained for the biggest and non-variable NCLAMP. Assuming $\sigma V_T \neq 0$, the design optimizer helps finding the best $\{\sigma V_T, W/L\}$ pair giving the optimum read margin/area trade-off.

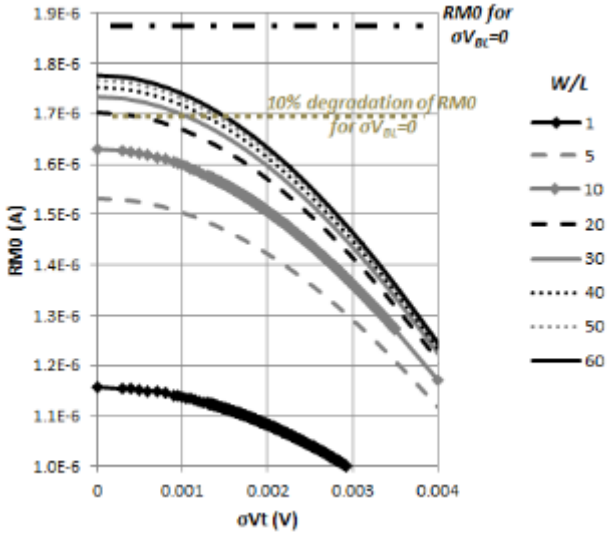


Figure 51 - Read margin dependence with NCLAMP's σV_T and g_m (e.g. for RM0; R_{LRS}=4kΩ, TMR=100%, N=3, $\sigma R_{DATA}/R_{DATA}$=5%, V_T=0.25V)

The principle of the area-aware read margin NCLAMP optimization is as follows:

- Target a given percentage of read margin degradation (e.g. 10% in *Figure 51*)
- Report, for each W/L variation, the required V_T variability for this percentage (see *Figure 52*)
- Take into account both W/L and σV_T NCLAMP's area contributions (see *Figure 53*):

$$(W * L)_{\sigma V_T} = \left(\frac{A_{V_T}}{\sigma V_T}\right)^2$$

Equation 28: Contribution of NCLAMP's threshold voltage on area

$$(W * L)_{\frac{W}{L}} = \frac{W}{L} * (K.L_{eff})^2$$

Equation 29: Contribution of NCLAMP's transconductance on area

where K is a shrink factor quantifying the ratio between the effective L_{eff} and the real drawn channel length. The whole area contribution is: $\max\left((W * L)_{\sigma V_T}; (W * L)_{\frac{W}{L}}\right)$

- Output the optimum $\{\sigma V_T, W/L\}$ pair from *Figure 53*: the best pair is minimum of $max\left((W * L)_{\sigma V_T}; (W * L)_{\frac{W}{L}}\right)$ curve. If this curve reaches a plateau, the best pair is the beginning of this plateau.

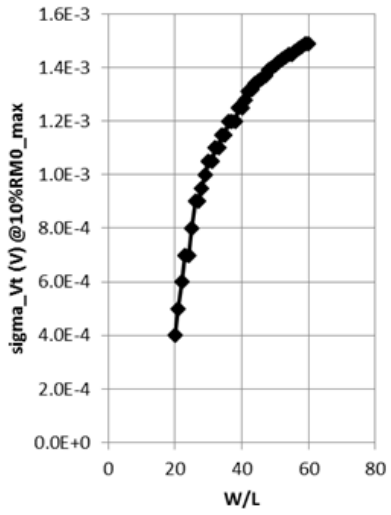


Figure 52 - List of all possible {σV_T, W/L} pairs for 10% tolerated maximum read margin degradation (obtained from Figure 51)

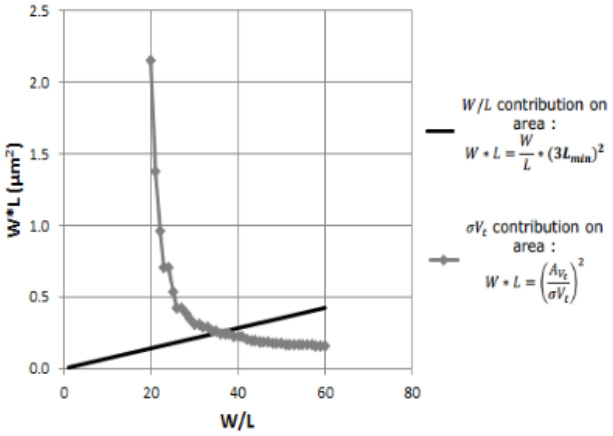


Figure 53 - Area contributions of NCLAMP's σV_T and g_m (e.g for RM0; R_{LRS}=4kΩ, TMR=100%, N=3, σR_{DATA}/R_{DATA}=5%, V_T=0.25V). For this particular case study, the best pair is {σV_T=1.2 mV, W/L=36}

A fully configurable statistical model of the beginning of the sensing path of a resistive memory is developed, taking into account process level parasitic and data resistances variabilities, for bit line clamping transistor's design optimization to minimize read margin degradations. A methodology is proposed to handle variability issues on read yield, according to the resistive device technology. This study is important to optimize read capability before focusing on reducing sense amplifier variability. The model perfectly matches with simulation (see *Figure 54*).

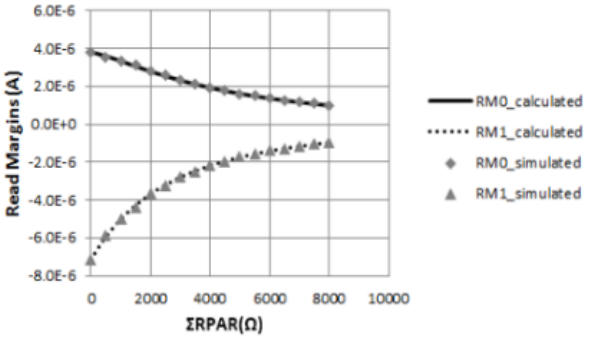


Figure 54 - Simulated vs modelled read margins (e.g. for R_{LRS}=4kΩ, TMR=100%, N=3, σR_{DATA}/R_{DATA}=5%, V_T=0.25V, V_{CLAMP}=0.5V, W/L=50, σV_T=2mV)

At this point, at the beginning of the sensing path and before the signal treatment by the sense amplifier, the attenuation of the degradation of the read margin is optimized as much as possible.

3.3. Circuit-level: Offset analysis of the Covalent-Bonded Cross-Coupled Current Mode Sense Amplifier (CBSA)

Regarding read reliability through the sense amplifier, it will depend on the chosen circuit architecture. As detailed in 0, a conventional latch circuit is not sufficient to ensure reliable read operation because of its transistor pairs' mismatch and its complex reference implementation.
The CBSA circuit (see 2.4.2.3) is a promising architecture for resistive memories since it handles reference problematic and variation-tolerance of latch-type circuits, as well as keeping the advantage of high-speed dynamic circuit.
In order to understand the different stakes of sense amplifier variability, it is judicious to analyze the mismatch (quantified by its offset) of this particular dynamic sense amplifier [82].

3.3.1.Offset Analysis

3.3.1.1. Systematic offset

In order to assess the circuit systematic offset, an input resistance difference equal to 0 (i.e $R_{HRS..REF}$=$R_{LRS.REF}$=R_{DATA}) has to be applied. In this configuration, the positive feedback leading to the latch dynamic behavior does not occur, and in an ideal case, maintain it in the meta-stability state (all outputs voltages are between ground and V_{DD}). However, since reference and data transistors are not identically sized, output capacitances of OUTA (or OUTB) and OUTM nodes are different. Therefore, when pre-charge turns OFF, output capacitive loads discharge with different discharge rates. This generates an input voltage while it should not. Output voltages thus decrease with time from their initial states (V_{DD}) with different slopes, triggering the sensing. Correctly sized dummy capacitors (P4HD, P4LD, N4HD, and N4LD) have to be added to balance this output load mismatch. Evaluating the output capacitances as a function of gate capacitances and assuming gate capacitances of latch transistors and dummy capacitors are respectively equal to C and C_{dummy}, we can write:

$$C_{OUTA} = C_{g(P4H)} + C_{g(P4HD)} + C_{g(N4H)} + C_{g(N4HD)} = 2C + 2C_{dummy}$$

Equation 30: Output capacitance at HRS reference current branch

$$C_{OUTB} = C_{g(P4L)} + C_{g(P4LD)} + C_{g(N4L)} + C_{g(N4LD)} = 2C + 2C_{dummy}$$

Equation 31: Output capacitance at LRS reference current branch

$$C_{OUTM} = 8C$$

Equation 32: Output capacitance at data current branch

The dummy transistors' capacitances and so their sizes can thus be deduced, for well-balanced outputs (i.e. for C_{OUTA} = C_{OUTB} = C_{OUTM}):

$$(W * L)_{dummy} = 3.(W * L)_{latch}$$

Equation 33: Dummy capacitors resizing for systematic offset cancellation

As their output capacitive mismatch counterpart, gate-drain coupling of output nodes is another factor influencing systematic offset (see *Figure 55*). Indeed, if latch transistors and thus their parasitic gate-drain capacitors are mismatched, a fall in V_{OUTA} will lead by coupling to a fall in V_{OUTM} (by a voltage shift equal to V_{SHIFT0}). Similarly, V_{OUTA} will decrease with V_{OUTM} but by a voltage shift V_{SHIFT1} different from V_{SHIFT0}. This

will generate an equivalent input voltage difference and thus trigger the sensing, even if dummy capacitors (P4HD, P4LD, N4HD, N4LD) are well-sized. This cross-coupling induced offset is treated in section III.

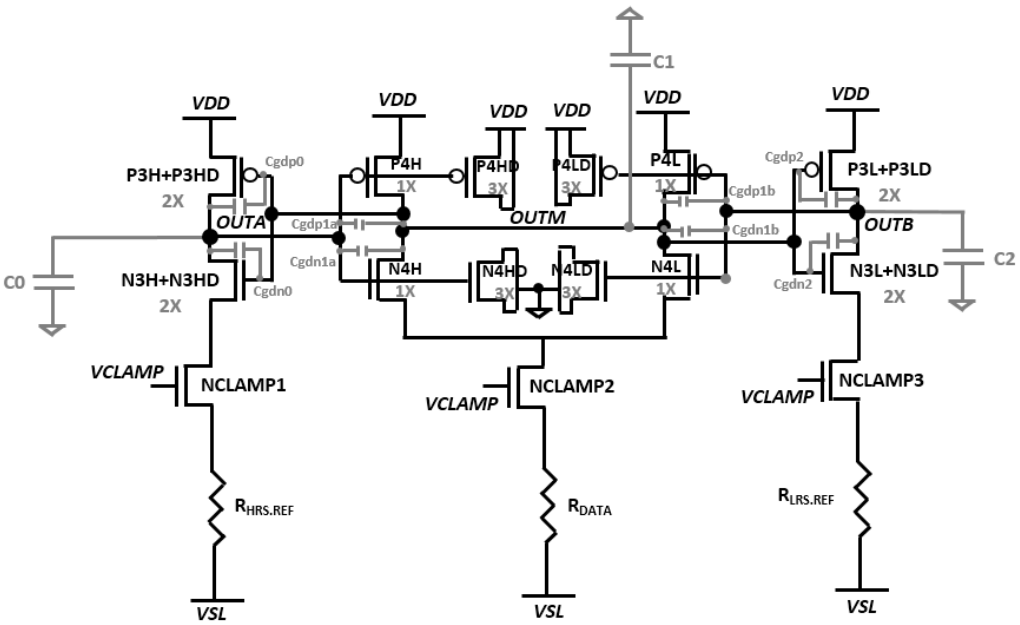


Figure 55 - Capacitive contributions (load and gate-drain coupling) on CBSA systematic offset

3.3.1.2. Design-level systematic offset reduction

The previous offset analysis emphasizes the impact of dummy capacitors as well as gate-to-drain coupling on systematic offset. This observation highlights some design improvements to further reduce systematic offset of the circuit in *Figure 55*. First, in order to accurately balance output capacitances, the dummy transistors P4HD, P4LD, N4HD and N4LD should be biased in the same way as P4H, P4L, N4H, and N4L. Therefore, connecting N4HD and N4LD source and drain to NCLAMP2 drain helps reducing systematic offset instead of connecting them to ground as it was proposed in [64].

Besides, gate-to-drain coupling unbalance between N4H and N3H+N3HD (and between N4L and N3L+N3LD) can be corrected by adding MOS capacitors N4H_CGD (and N4L_CGD) connected to gate and drain of transistors N4H (and N4L) and so cancel the cross-coupling induced systematic offset. As N3H+N3HD (N3L+N3LD) has twice the size than N4H (N4L), N4H_CGD (N4L_CGD) must have half the size of N4H (N4L) since their gate and drain are connected each other. *Figure 56* depicts the design improvements brought to *Figure 55* (highlighted in grey) and *Table 2* shows the corresponding systematic offset reduction, demonstrating that this architecture is quite suitable to handle systematic offset reduction.

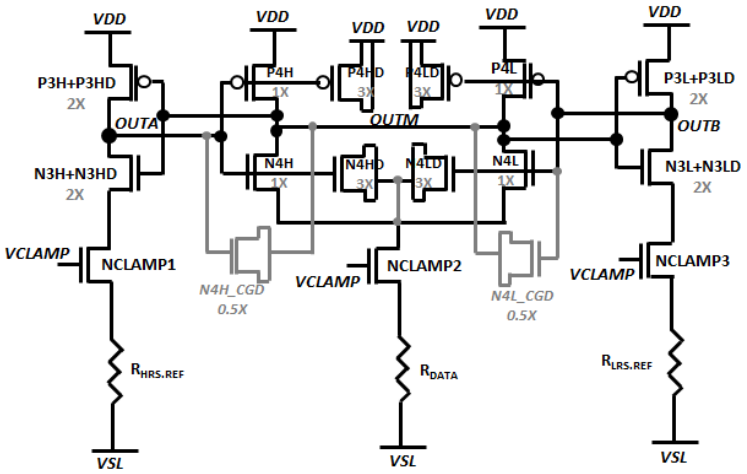


Figure 56 - CBSA design optimization

Table 2 - Systematic offset cancellation of the CBSA after design optimization

With N4HD (N4LD) to ground ?	*No*	*Yes*	*Yes*
With gate-drain dummy capacitors ?	*No*	*No*	*Yes*
Input-referred systematic offset (Ω)	*742*	*495*	*3*

3.3.1.3. Random offset

The random offset due to device mismatches can now be analyzed for further optimization. First, analyzing the main transistors contributing to random offset is required. Once pre-charge turns off, PMOS transistors of the two latches remain off, leading to output capacitors discharge through data and reference branches, whose currents are set by bit line clamping NMOS transistors. The offset value therefore depends on two main contributors: bit line clamping NMOS transistors and PMOS transistors of the latches. Indeed, output voltage decreases (output capacitors discharge) with different speeds depending on data and reference current set by the clamping NMOSFETs. After that, the output voltage will reach latch PMOS transistors' threshold voltages and so those transistors will trigger positive feedback.
Consequently, NMOS transistors of the latches do not influence the offset of the circuit as they are cascoded and so the clamping transistors only dictate their drain current. This observation will be exploited in section 3.3.2. *Figure 57* shows a pareto chart from statistical simulation results (using most probable point methodology [10]) quantifying the main contributors of the offset, for four sigma overall variations. It confirms that latch PMOSFETs (P3H+P3HD, P4H, P4L, P3L+P3LD) and bit line clamping NMOSFETs (NCLAMP1, NCLAMP2, NCLAMP3) are the main contributors to offset.

Moreover, capacitive loads are also important random offset contributors. Indeed, offset decreases with the increase of output capacitors. This is because intrinsic capacitive effects of the circuit (gate capacitors mismatch and gate-drain coupling mismatch, in the range of a few femtofarads) have much less impact in presence of high loads (in the range of ten femtofarads).

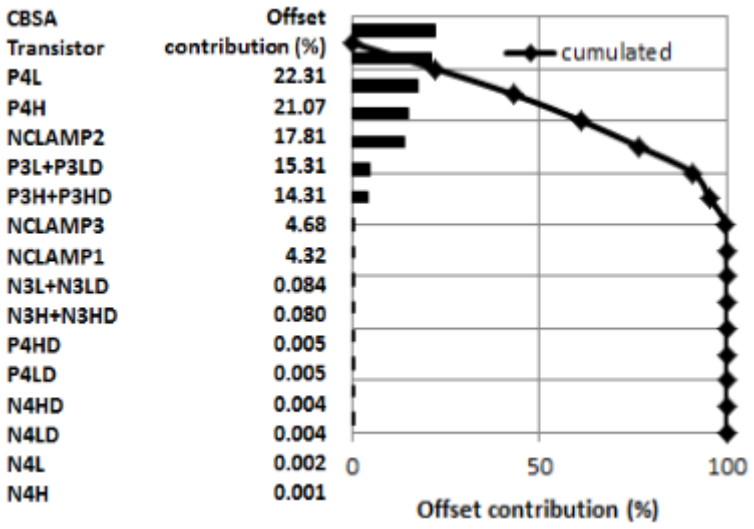


Figure 57 - Pareto Chart: Contribution (in percentage) on the offset of each CBSA transistor, for 4 sigma overall variation

3.3.2.Offset modelling

An analytical model for offset (applied for both systematic and random) is derived for this circuit. Analytic computations of the offset of dynamic sense amplifier that can be found on literature are mostly dedicated to SRAM or DRAM memories using small or large signal analysis [83][84]. For simplicity and as the circuit is symmetric, the calculation was made for half of the circuit (only one latch, R_{DATA} and $R_{HRS.REF}$ sensing paths for example) with correct transistor sizes.

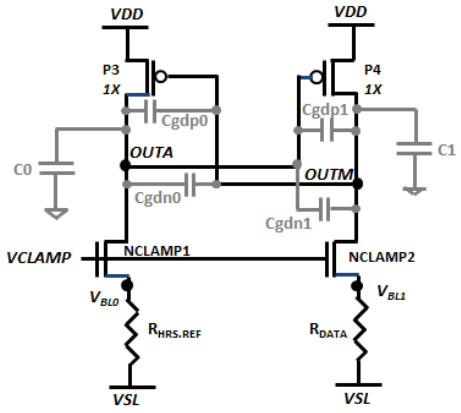


Figure 58 - Equivalent circuit of left-hand latch and 1T-1R stage for offset modelling

The equivalent circuit used for the analytical model is thus depicted in *Figure 58*. Circuit random variables taken into account for this model are: V_{tp0}= threshold voltage of transistor P3, V_{tp1}= threshold voltage of transistor P4, C_{gdn0}= gate-drain coupling capacitor of transistor N3H+N3HD, C_{gdp0}= gate-drain coupling capacitor of transistor P3, C_{gdn1}= gate-drain coupling capacitor of transistor N4H, C_{gdp1}= gate-drain coupling capacitor of transistor P4, C_0 = output capacitor at OUTA node, C_1 = output capacitor at OUTM node, C_{g0} = gate capacitance at node OUTA, C_{g1} = gate capacitance at node OUTM, V_{bl0} = data bit line voltage, V_{bl1} = reference bit line voltage, $V_0 = V_{OUTA}$, $V1 = V_{OUTM}$. The variables V_{bl0} and V_{bl1} take into account threshold voltages' of bit line clamping transistors as well as bit cell resistive states. This modelling thus includes the statistical computations described in section 0. Kirchhoff law at output nodes gives:

$$\left(C_0 + C_{g1}\right) * \frac{dV_0}{dt} + \frac{V_{bl0}}{R_{REF}} + \left(C_{gdp0} + C_{gdn0}\right) * \frac{d(V_0 - V_1)}{dt} = 0$$

Equation 34: Kirchhoff's law at reference branch output node

$$\left(C_1 + C_{g0}\right) * \frac{dV_1}{dt} + \frac{V_{bl1}}{R_{DATA}} + \left(C_{gdp1} + C_{gdn1}\right) * \frac{d(V_1 - V_0)}{dt} = 0$$

Equation 35: Kirchhoff's law at data branch output node

This leads to:

$$\frac{dV_0}{dt} = -\left(c * \frac{V_{bl0}}{R_{REF}} + d * \frac{V_{bl1}}{R_{DATA}}\right) = X_0$$

Equation 36: Reference output voltage slope as a function of input resistances and capacitive contributions

$$\frac{dV_1}{dt} = -\left(a * \frac{V_{bl0}}{R_{REF}} + b * \frac{V_{bl1}}{R_{DATA}}\right) = X_1$$

Equation 37: Data output voltage slope as a function of input resistances and capacitive contributions

with:

$$\frac{1}{a} = \frac{\left(C_1 + C_{g0} + C_{gdp1} + C_{gdn1}\right) * \left(C_0 + C_{g1} + C_{gdp0} + C_{gdn0}\right)}{C_{gdp1} + C_{gdn1}} - \left(C_{gdp0} + C_{gdn0}\right)$$

Equation 38: Capacitive contribution of the reference current on data output voltage decrease

$$\frac{1}{b} = \left(C_1 + C_{g0} + C_{gdp1} + C_{gdn1}\right) - \frac{\left(C_{gdp1} + C_{gdn1}\right) * \left(C_{gdp0} + C_{gdn0}\right)}{C_0 + C_{g1} + C_{gdp0} + C_{gdn0}}$$

Equation 39: Capacitive contribution of the data current on data output voltage decrease

$$c = \frac{\left(C_{gdp1} + C_{gdn1}\right) * \left(C_{gdp0} + C_{gdn0}\right)}{\left(C_1 + C_{g0} + C_{gdp1} + C_{gdn1}\right) * \left(C_0 + C_{g1} + C_{gdp0} + C_{gdn0}\right)} + \frac{1}{C_0 + C_{g1} + C_{gdp0} + C_{gdn0}}$$

Equation 40: Capacitive contribution of the reference current on reference output voltage decrease

$$\frac{1}{d} = \frac{\left(C_1 + C_{g0} + C_{gdp1} + C_{gdn1}\right) * \left(C_0 + C_{g1} + C_{gdp0} + C_{gdn0}\right)}{C_{gdp0} + C_{gdn0}} - \left(C_{gdp1} + C_{gdn1}\right)$$

Equation 41: Capacitive contribution of the data current on reference output voltage decrease

Output voltages can then be expressed as followed:

$$V_0(t) = V_{DD} - X_0 * t$$

Equation 42: Reference output voltage

$$V_1(t) = V_{DD} - X_1 * t$$

Equation 43: Data output voltage

At $t=t_1$, i.e once voltage V_1 reaches P3 threshold voltage, $V_1 = V_{DD} - V_{TP0}$. Likewise, at $t=t_0$, $V_0 = V_{DD} - V_{TP1}$. When $t_0=t_1$, PMOSs threshold voltages are reached at the exact same moment. Thus, no positive feedback triggering occurs. The offset can therefore be calculated assuming this equality, which represents a perfectly balanced circuit:

$$\frac{X_1}{X_0} = \frac{V_{tP0}}{V_{tP1}}$$

Equation 44: Output voltage slope condition for metastable state

Offset can consequently be deduced:

$$\frac{R_{REF}}{R_{DATA}} = \frac{V_{bl0}}{V_{bl1}} * \frac{a.V_{tP1} - c.V_{tP0}}{d.V_{tP0} - b.V_{tP1}}$$

Equation 45: Input-referred systematic offset

Using the theory of propagation of uncertainty [49][50], the random offset is:

$$\frac{\sigma^2\left(\frac{R_{REF}}{R_{DATA}}\right)}{\left(\frac{R_{REF}}{R_{DATA}}\right)^2} = \frac{(aV_{TP1})^2.\left(\frac{\sigma_a^2}{a^2} + \frac{\sigma^2 V_{TP1}}{V_{TP1}^2}\right) + (cV_{TP0})^2.\left(\frac{\sigma_c^2}{c^2} + \frac{\sigma^2 V_{TP0}}{V_{TP0}^2}\right)}{(aV_{TP1} - cV_{TP0})^2} + \frac{(dV_{TP0})^2.\left(\frac{\sigma_d^2}{d^2} + \frac{\sigma^2 V_{TP0}}{V_{TP0}^2}\right) + (bV_{TP1})^2.\left(\frac{\sigma_b^2}{b^2} + \frac{\sigma^2 V_{TP1}}{V_{TP1}^2}\right)}{(dV_{TP0} - bV_{TP1})^2} + \frac{\sigma^2 V_{BL0}}{V_{BL0}^2} + \frac{\sigma^2 V_{BL1}}{V_{BL1}^2}$$

Equation 46: Input-referred random offset

Similarly, the offset of the whole CBSA can be obtained from the circuit depicted in *Figure 59*. Kirchhoff law at the three output nodes of the CBSA gives (with $V_2 = V_{OUTB}$):

$$(C_0 + C_{g1a}) * \frac{dV_0}{dt} + \frac{V_{BL0}}{R_{HRS.REF}} + (C_{gdp0} + C_{gdn0}) * \frac{d(V_0 - V_1)}{dt} = 0$$

Equation 47: Kirchhoff's law at HRS reference branch output node of the CBSA

$$(C_2 + C_{g1b}) * \frac{dV_2}{dt} + \frac{V_{bl2}}{R_{LRS.REF}} + (C_{gdp2} + C_{gdn2}) * \frac{d(V_2 - V_1)}{dt} = 0$$

Equation 48: Kirchhoff's law at LRS reference branch output node of the CBSA

$$(C_1 + C_{g0} + C_{g2}) * \frac{dV_1}{dt} + (C_{gdp1a} + C_{gdn1a}) * \frac{d(V_1 - V_0)}{dt} + (C_{gdp1b} + C_{gdn1b}) * \frac{d(V_1 - V_2)}{dt} + \frac{V_{bl1}}{2 * R_{DATA}} + \frac{V_{bl1}}{2 * R_{DATA}} = 0$$

Equation 49: Kirchhoff's law at data branch output node of the CBSA

Therefore, the conformity of *Equation 46* with simulations for the whole sense amplifier can be obtained by assuming for simplicity $R_{HRS.REF}$=$R_{LRS.REF}$. Thus, the CBSA is electrically equivalent to the circuit in *Figure 58* by applying the following changes in *Equation 34* and *Equation 35*:

-In Equation 34: replace (C_{gdp0} + C_{gdn0}) by 2*(C_{gdp0} + C_{gdn0}) as P3H+P3HD is twice the size of P3.

-In Equation 35: replace C_{g0} by 4*C_{g0} to take into account all gate capacitances at V_{OUTM}, and replace (C_{gdp1} + C_{gdn1}) by 2*(C_{gdp1} + C_{gdn1}) as V_0=V_2 is assumed (see Equation 49).

In terms of offset contribution, we have: $\sigma V_T(P4H) = \frac{\sigma V_T(P4)}{\sqrt{2}}$ and $\sigma V_T(P3H + P3HD) = \frac{\sigma V_T(P3)}{\sqrt{2}}$.

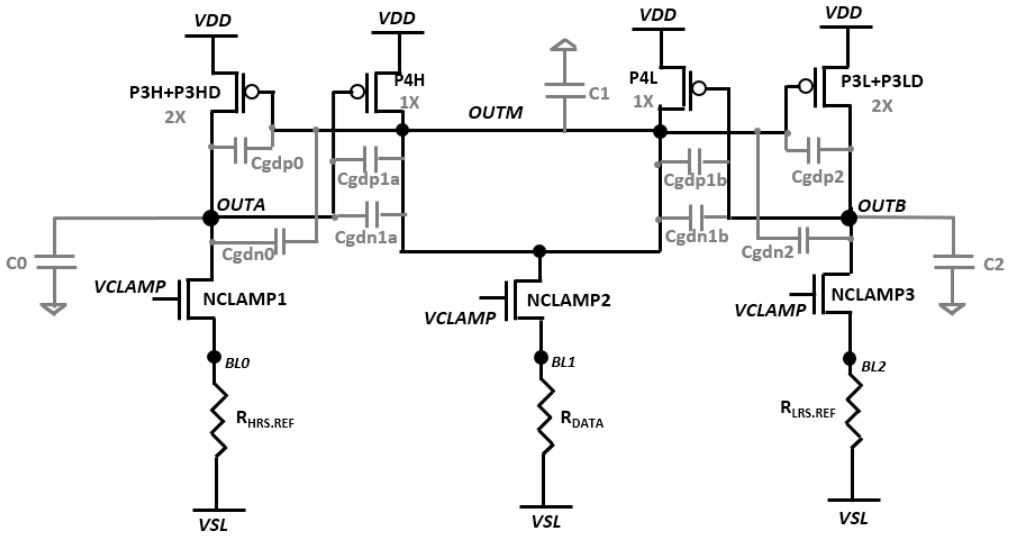


Figure 59 - CBSA equivalent circuit for offset modelling

3.3.3.Offset Characterization

In order to compare the modeled-based calculated offset of the CBSA with simulation, we have to choose the statistical simulation methodology to use among those proposed in 2.2.2. Monte Carlo simulation is not suitable here too much runs would be needed to ensure good read yield accuracy (at 5 or 6 sigma range). Moreover, DOE/RSM method cannot be used here since it does not cover highly nonlinear circuits as this dynamic CBSA. Therefore, MPP characterization is adopted here. By varying the data input resistance, the maximum time delay between the sensing triggering (latch PMOS transistors turning ON) and the creation of a quite sufficient output voltage differential (for example $V_{out} = V_{dd}/2$) is measured. The minimum R_{data}-R_{ref} difference giving the first measurable delay is so called the input referred offset, as described in the dichotomy flow chart in *Figure 60*. This flow was used to develop a Skrill coded script for offset characterization, which can be found in Appendix B.

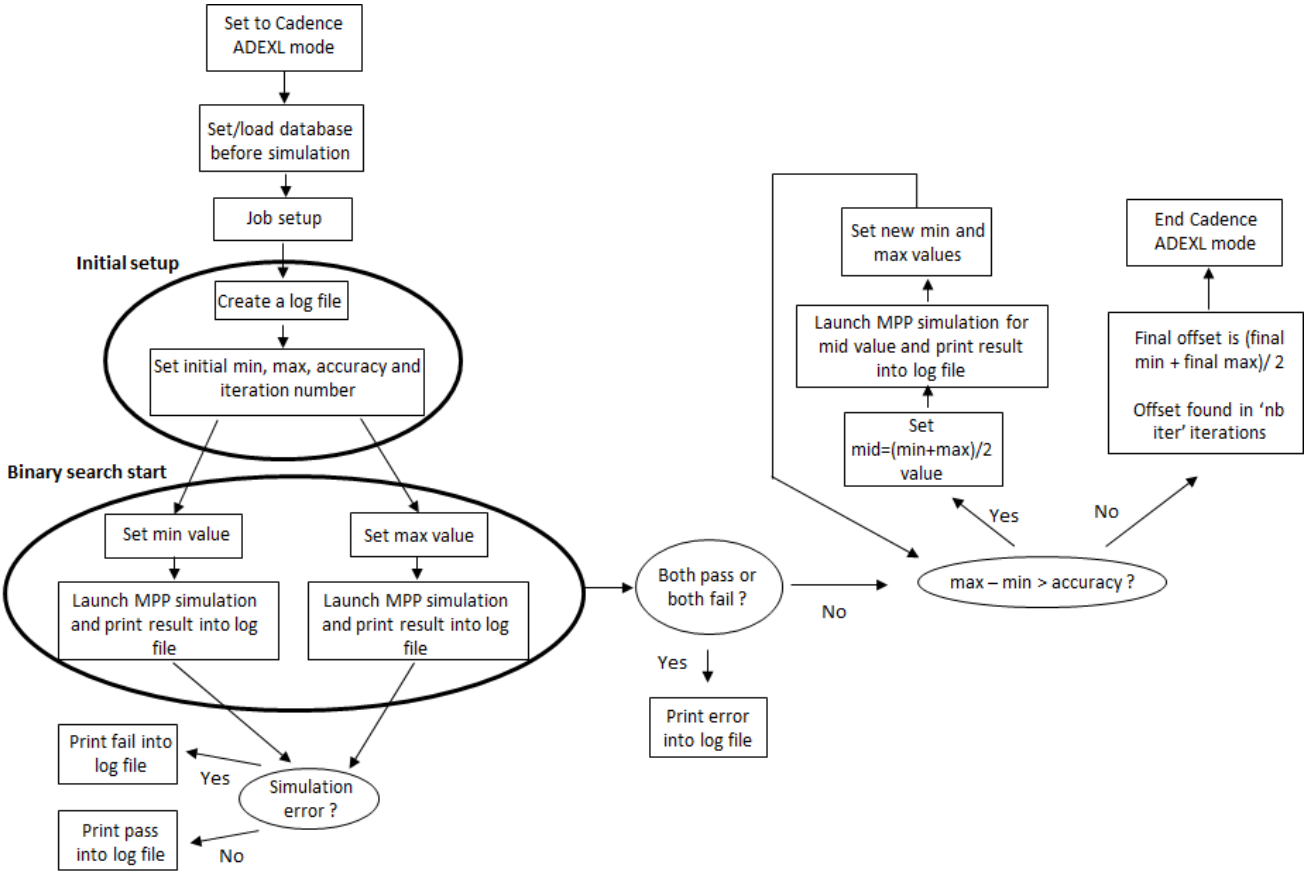


Figure 60 - Offset characterization methodology using MPP simulations

3.3.4.Discussion

Simulated values of the CBSA offset and that given by *Equation 46* are plotted in *Figure 61* for different load capacitances (C_0=C_1=C_{load} in the figure) and assuming $R_{HRS.REF}$=$R_{LRS.REF}$. σC_0=σC_1=0 is also assume as a best case to put forward the ideal minimum obtainable offset. Note that the mismatch value of capacitors contributions on offset, namely σ_a, σ_b, σ_c, and σ_d are obtained by statistical computation method and integrated in *Equation 46*. This model shows the importance of load capacitances on offset reduction, and takes into account its mismatch value. A good matching is found to first order approximation. Dotted line curve, which represents the calculated offset for C_{gd}=0, shows that for high loads, intrinsic capacitive effects are strongly reduced or even cancelled.

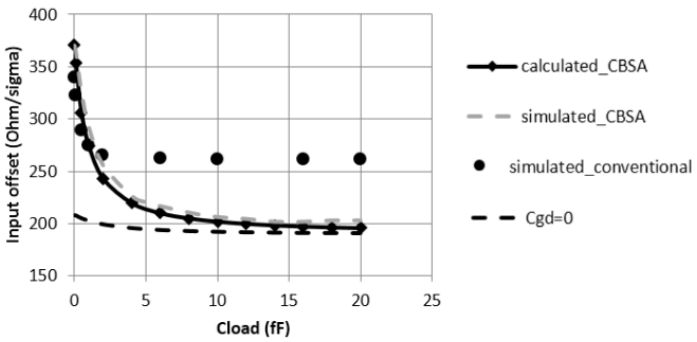


Figure 61 - Simulated and analytically modeled CBSA offset compared to conventional sense amplifier, for $\sigma C_0=\sigma C_1=0$ and $R_{HRS.REF}=R_{LRS.REF}$. Dotted line curve shows gate-drain coupling impact on offset for low load capacitance

The CBSA offset reduction, linked to the use of two symmetrical latch stages sharing one common current path, is of the order of more than fifty ohms for one sigma overall variation compared to the conventional latch sense amplifier. Although it appears insufficient, it seems to be the optimum attainable reduction for circuits based on transistors matching. *Table 3* shows CBSA improvements in terms of offset reduction, reference layout regularity and sensing speed, along with the area overhead that is required in return.

Table 3 - CBSA features compared to conventional sense amplifier

	Minimum input offset (Ohms/sigma)	***Layout regular reference?***	***Sensing Speed***	***Area overhead: # transistors***
Optimized CBSA	***200***	***Yes***	***~9 ns***	***Minimum 11***
Conventional sense amplifier	***260***	***No***	*>10 ns*	*6*

For further offset reduction, some alternatives are possible by combining dynamic stages. [85] proposes a configuration able to divide the offset of a sense amplifier by a factor of 3. The principle is to split one small but variable sense amplifier into two or more bigger but less variable stages. Depending on the desired configuration, these stages are combined by either putting in parallel their inputs or switching them (see *Figure 62*).

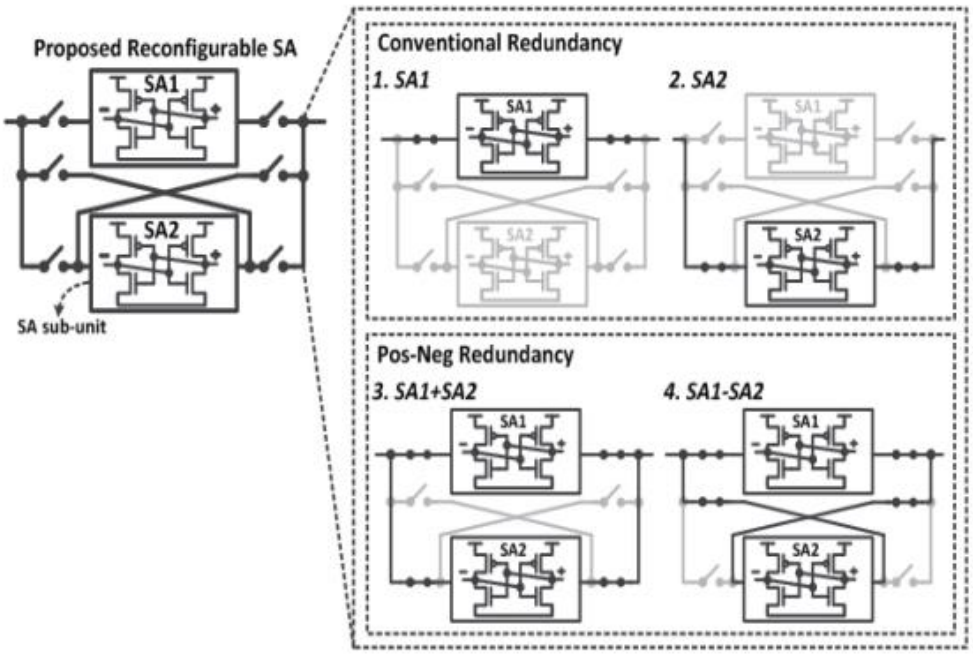


Figure 62 - Reconfigurable Sense Amplifier for 3X offset reduction [85]

A statistical model was thus developed to describe the factors degrading sense amplifier offset. This model takes into account both bit cell and loads variation. It was applied to a sense amplifier for resistive memories, but can be derived for any other latch-type circuit. A summary of model parameters is shown in *Table 4*.

Table 4 - Dynamic sense amplifier offset model: main parameters

		Bit cell			***Capacitive contributions***		
Parameter	N_{sigma}	R_{LRS}, σR_{LRS}	$K_{on\text{-}off}$	ΣR_{PAR}, $\sigma\Sigma R_{PAR}$	C_{g_PLATCH} σC_{g_PLATCH}	C_{gd_PLATCH} σC_{gd_PLATCH}	C_{load} σC_{load}
Description	*Overall applied variation*	*Bit cell Resistance and variation*	*On-off ratio*	*Parasitics and layout effects*	*Latch P transistor gate capacitance and its variation*	*Latch P transistor gate-to-drain coupling capacitance and its variation*	*Load capacitance and its variation*
Unit	-	*Ω, %*	*%*	*Ω, %*	*fF, fF*	*fF, fF*	*fF, fF*
	Sense amplifier						
Parameter	V_{CLAMP}	W_{NCLAMP}, L_{NCLAMP}	V_{T_NCLAMP}, σV_{T_NCLAMP}	W_{PLATCH}, L_{PLATCH}	V_{T_PLATCH}, σV_{T_PLATCH}		
Description	*Bit line clamping voltage*	*Clamping transistor length and width*	*Clamping transistor threshold voltage*	*Latch P transistor length and width*	*Latch P transistor threshold voltage and*		

			and its variation		*its variation*
Unit	*V*	*µm, µm*	*mV, mV*	*µm, µm*	*mV, mV*

3.4. Conclusion

A comprehensive analysis of the variability of the sensing path of resistive memories is performed in this chapter. A design optimization methodology is proposed to minimize read margin degradation with variability at the input of sense amplifier. It is important to position process parameters and their variability regarding read window reduction, as well as non-linear dependency with voltage of the resistive bit cell. The resizing of bit line voltage clamping transistor is then automated for a minimum read margin decrease at design-level. An algorithm for offset characterization of a design-optimized dynamic sense amplifier (CBSA) is moreover described. This offset is additionally modelled taking into account bit cell statistics. This study highlights the necessity of non-dynamic sense amplifier architectures that combines offset-cancellation techniques and layout regular reference scheme.

References of Chapter 3

[70] C. Baeumer, R. Valenta, C. Schmitz, A. Locatelli, T. O. Menteş, S. P. Rogers, A. Sala, N. Raab, S. Nemsak, M. Shim, C. M. Schneider, S. Menzel, R. Waser, and R. Dittmann, "Subfilamentary Networks Cause Cycle-to-Cycle Variability in Memristive Devices," *ACS Nano*, vol. 11, no. 7, pp. 6921–6929, Jul. 2017.

[71] T. Devolder, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, P. Crozat, N. Zerounian, J.-V. Kim, C. Chappert, and H. Ohno, "Single-Shot Time-Resolved Measurements of Nanosecond-Scale Spin-Transfer Induced Switching: Stochastic Versus Deterministic Aspects," *Phys. Rev. Lett.*, vol. 100, no. 5, p. 57206, Feb. 2008.

[72] E. Cha, J. Woo, D. Lee, S. Lee, and H. Hwang, "Selector devices for 3-D cross-point ReRAM," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2014, pp. 428–431.

[73] R. Aluguri and T.-Y. Tseng, "Overview of Selector Devices for 3-D Stackable Cross Point RRAM Arrays," *IEEE J. Electron Devices Soc.*, vol. 4, no. 5, pp. 294–306, Sep. 2016.

[74] H. Yang, X. Hao, Z. Wang, R. Malmhall, H. Gan, K. Satoh, J. Zhang, D. H. Jung, X. Wang, Y. Zhou, B. K. Yen, and Y. Huai, "Threshold switching selector and 1S1R integration development for 3D cross-point STT-MRAM," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, p. 38.1.1-38.1.4.

[75] DerChang Kau, S. Tang, I. V. Karpov, R. Dodge, B. Klehn, J. A. Kalb, J. Strand, A. Diaz, N. Leung, J. Wu, Sean Lee, T. Langtry, Kuo-wei Chang, C. Papagianni, Jinwook Lee, J. Hirst, S. Erra, E. Flores, N. Righos, H. Castro, and G. Spadini, "A stackable cross point Phase Change Memory," in *2009 IEEE International Electron Devices Meeting (IEDM)*, 2009, pp. 1–4.

[76] M. Palmer, "Propagation of Uncertainty through Mathematical Operations," *MIT Modul.*, pp. 1–7, 2003.

[77] E. S. Lee and R. E. Forthofer, "Strategies for Variance Estimation," in *Analyzing Complex Servey Data*, vol. 22, 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc., 2006, pp. 22–39.

[78] Y. Zhang, W. Zhao, Y. Lakys, J.-O. Klein, J.-V. Kim, D. Ravelosona, and C. Chappert, "Compact Modeling of Perpendicular-Anisotropy CoFeB/MgO Magnetic Tunnel Junctions," *IEEE Trans. Electron Devices*, vol. 59, no. 3, pp. 819–826, Mar. 2012.

[79] J. Kim, A. Chen, B. Behin-Aein, S. Kumar, J.-P. Wang, and C. H. Kim, "A technology-agnostic MTJ SPICE model with user-defined dimensions for STT-MRAM scalability studies," in *2015 IEEE Custom Integrated Circuits Conference (CICC)*, 2015, pp. 1–4.

[80] M. Bocquet, D. Deleruyelle, H. Aziza, C. Muller, J.-M. Portal, T. Cabout, and E. Jalaguier, "Robust Compact Model for Bipolar Oxide-Based Resistive Switching Memories," *IEEE Trans. Electron Devices*, vol. 61, no. 3, pp. 674–681, Mar. 2014.

[81] E. Covi, A. Kiouseloglou, A. Cabrini, and G. Torelli, "Compact model for phase change memory cells," in *2014 10th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*, 2014, pp. 1–4.

[82] S. Mraihi, E. M. Boujamaa, C. Dray, and J.-O. Klein, “Offset Analysis and Design Optimization of a Dynamic Sense Amplifier for Resistive Memories,” in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2017, pp. 326–331.

[83] Y. Li, H. Schneider, F. Schnabel, R. Thewes, and D. Schmitt-Landsiedel, “Latched CMOS DRAM Sense Amplifier Yield Analysis and Optimization,” 2009, pp. 126–135.

[84] R. Singh and N. Bhat, “An offset compensation technique for latch type sense amplifiers in high-speed low-power SRAMs,” *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 12, no. 6, pp. 652–657, Jun. 2004.

[85] M. Khayatzadeh, F. Frustaci, D. Blaauw, D. Sylvester, and M. Alioto, “A reconfigurable sense amplifier with 3X offset reduction in 28nm FDSOI CMOS,” in *2015 Symposium on VLSI Circuits (VLSI Circuits)*, 2015, pp. C270–C271.

Appendix I: Confidential

The next two chapters are part of the confidential appendix of this manuscript. They contain relevant information about methods and performances used by the company.
The reader must be aware that the possession of this section is submitted to a confidentiality agreement.
Chapter 6 as well as the French summary of this manuscript contain conclusions of these two chapters, and so are part of the confidential section.

Chapter 4: Proposed Offset-cancelled Sense Amplifier with reduced reference variability

4.1. Introduction

0 helps understand that dynamic circuits are not convenient in order to improve more resistive memories read reliability, even when a layout regular reference scheme is included. This chapter proposes a sense amplifier architecture that combines offset-cancellation techniques (introduced in 2.4.2.5) and easy-to-implement reference scheme with reduced variability, allowing an improved read yield.

4.2. Two-references: time-multiplexed reference scheme

The Full Offset-Cancellation Technique (FOCT) described in Chapter 2 (see 2.4.2.5.2) uses the principle of multiple-phase gate voltage register through sampling capacitors in order to set the required current signal. *Figure 63* illustrates this technique using two different sampling capacitors for separate current branches (for non-current-mirrors applications).

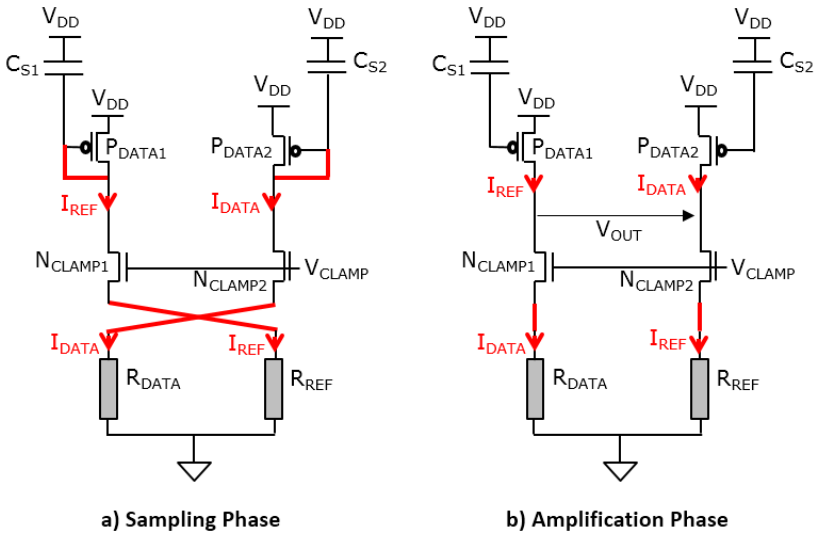
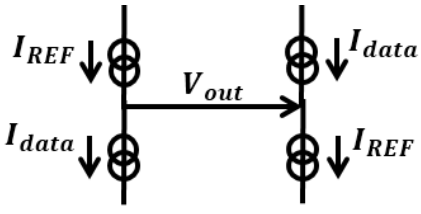


Figure 63 - FOCT using two sampling capacitors a) during sampling phase b) during amplification phase

From *Figure 63*, the currents set in each branch at the end of the amplification phase are:



$$\mu(\Delta I) = 2(I_{data} - I_{REF})$$

$$\sigma(\Delta I) = 2\sqrt{\sigma I_{data}^2 + \sigma I_{REF}^2} = \sqrt{4\sigma I_{data}^2 + \sigma I_{LRS}^2 + \sigma I_{HRS}^2}$$

b)

Figure 64 - a) FOCT current scheme at the end of the amplification phase b) Current mean and variance transfer function

[65] proposes a sense amplifier architecture integrating the FOCT as well as a so called time-multiplexed reference scheme (see

Figure 65). The principle is to implement two complementary resistive devices as references. Each of these references is connected to the circuit during either the sampling or the amplification phase. The resulting current transfer function is similar to the one of the FOCT (see *Figure 64*), but with two uncorrelated

complementary resistive devices. This feature allows a much easier reference implementation. Moreover, separating current branches allow the use of more references in order to reduce their variability, as it will be described below.

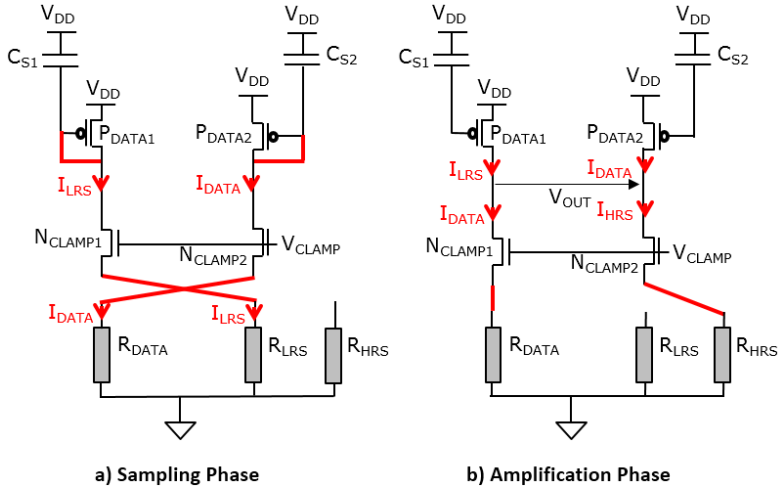


Figure 65 - Offset-cancelled sense amplifier with time-multiplexed reference scheme

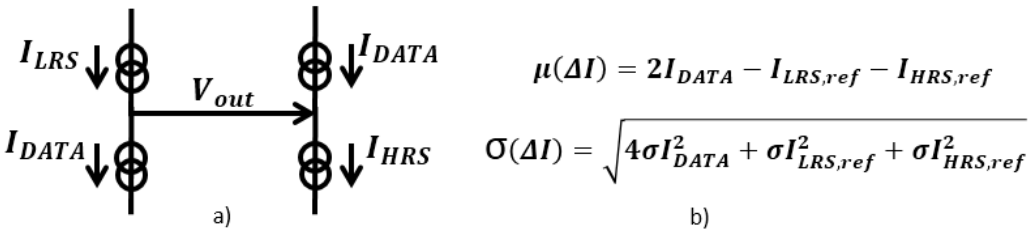


Figure 66 - Offset-cancelled sense amplifier with time-multiplexed reference a) current scheme at the end of the amplification phase b) Current mean and variance transfer function

The small signal analysis of this circuit architecture (see *Figure 67*) demonstrates that in case of low read margin (R_{LRS} close to R_{HRS}), the offset of this sense amplifier can be almost cancelled:

$$v_{C_{S_1}} = -\frac{g_{m_N}}{g_{m_P}} \cdot \frac{\sigma V_{T_{n_1}}}{1 + g_{m_N} \cdot R_{LRS}}$$

Equation 50: Small signal first sampling voltage during the sampling phase of the FOCT-based sense amplifier featuring time-multiplexed reference scheme

$$v_{C_{S_2}} = -\frac{g_{m_N}}{g_{m_P}} \cdot \frac{\sigma V_{T_{n_2}}}{1 + g_{m_N} \cdot R_{DATA}}$$

Equation 51: Small signal second sampling voltage during the sampling phase of the FOCT-based sense amplifier featuring time-multiplexed reference scheme

$$v_{out} = v_{out_2} - v_{out_1}$$
$$= -g_{m_N}.r_{out}.\left(\left[\frac{\sigma V_{T_{n_2}}}{1+g_{m_N}.R_{HRS}} - \frac{\sigma V_{T_{n_2}}}{1+g_{m_N}.R_{DATA}}\right] - \left[\frac{\sigma V_{T_{n_1}}}{1+g_{m_N}.R_{DATA}} - \frac{\sigma V_{T_{n_1}}}{1+g_{m_N}.R_{LRS}}\right]\right)$$

Equation 52: Small signal output voltage at the end of the amplification phase of the FOCT-based sense amplifier featuring time-multiplexed reference scheme

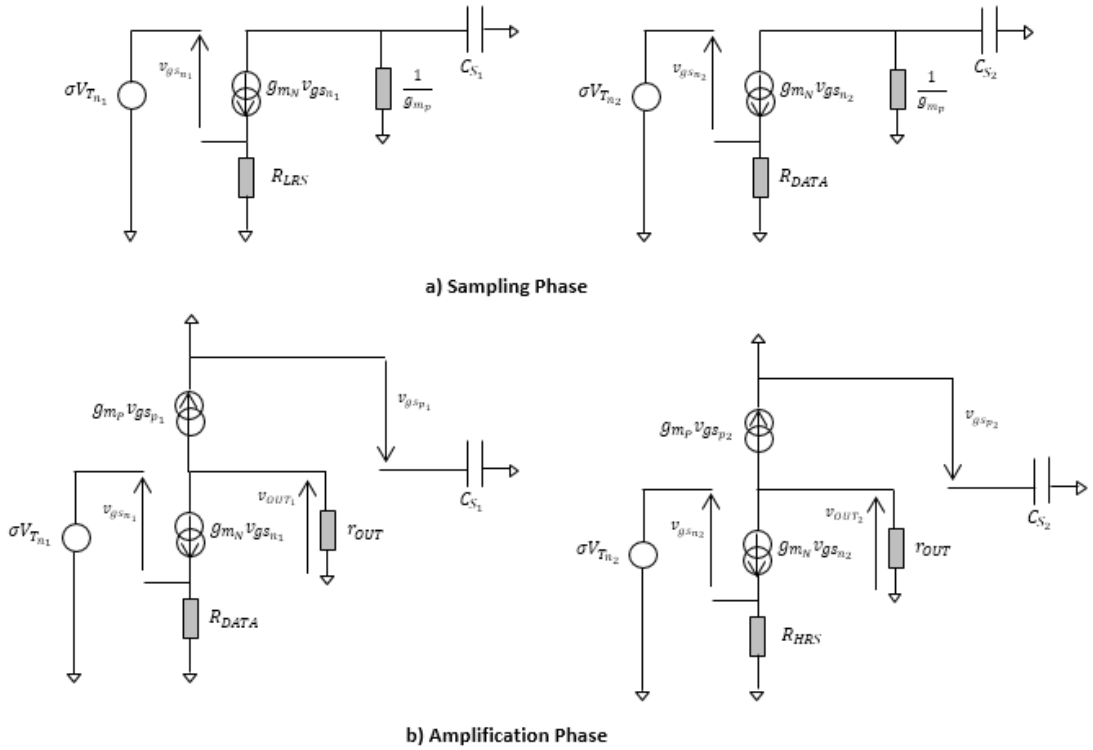


Figure 67- Full offset cancellation technique featuring time-multiplexed reference scheme: Small signal equivalent circuit during (a) sampling phase (b) amplification phase

This double principle of offset cancellation featuring increased signal to offset ratio by reducing the reference variability should be extended for N references, by evaluating both read margin improvements and the resulting power, timing and area costs. This is described below.

4.3. Proposed offset-cancelled sense amplifier with reduced reference variability

4.3.1.Proposed architecture for three references

4.3.1.1. Current scheme

In following lines, an innovative approach that combines both offset-cancellation and reference variability reduction is presented.

First, the proposed circuit maintains the Full Offset-Cancellation technique (FOCT) for higher output signal. Therefore, references and data current paths are identical between sampling and amplification phases (same transistors used). The reference resistances' variability reduction will then consist of adding sampling phases per additional uncorrelated reference. In order to illustrate that, *Figure 68* describes the proposed current scheme for three references, at the end of the amplification phase.

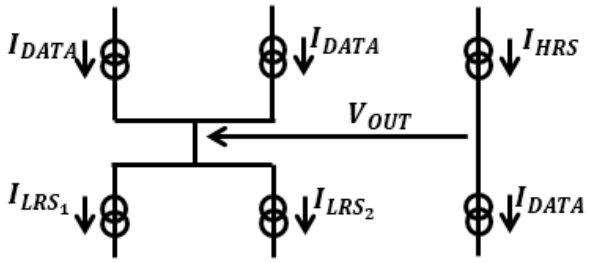


Figure 68 - Proposed current scheme for improved signal to offset ratio with three references

The current transfer function below shows that the signal output is still doubled ($2I_{data}$), but the LRS reference variability is reduced thanks to the use of two uncorrelated resistive devices:

$$V_{OUT} = \frac{r_{out}}{2}\left(2I_{DATA} - I_{LRS1,ref} - I_{LRS2,ref}\right) - r_{out}(I_{HRS,ref} - I_{DATA})$$

Equation 53: Output voltage of the proposed current scheme with three uncorrelated references, at the end of the amplification phase

$$\mu(\Delta I) = 2I_{DATA} - \frac{I_{LRS1,ref}}{2} - \frac{I_{LRS2,ref}}{2} - I_{HRS,ref}$$

Equation 54: Current transfer function of the proposed current scheme with three uncorrelated references

$$\sigma(\Delta I) = \sqrt{4\sigma I_{DATA}^2 + \frac{\sigma I_{LRS,ref}^2}{2} + \sigma I_{HRS,ref}^2}$$

Equation 55: Illustration of reference variability reduction of the proposed current scheme with three uncorrelated references

4.3.1.2. Circuit implementation

Figure 69 illustrates the circuit implementation of the proposed current scheme for three references. The use of one additional uncorrelated reference requires one additional offset sampling phase and one additional current branch. In total, four resistive devices, two sampling phases to register the two pull-up data currents and one amplification phase are implemented to result to an offset-cancelled with improved signal to offset ratio sense amplifier.

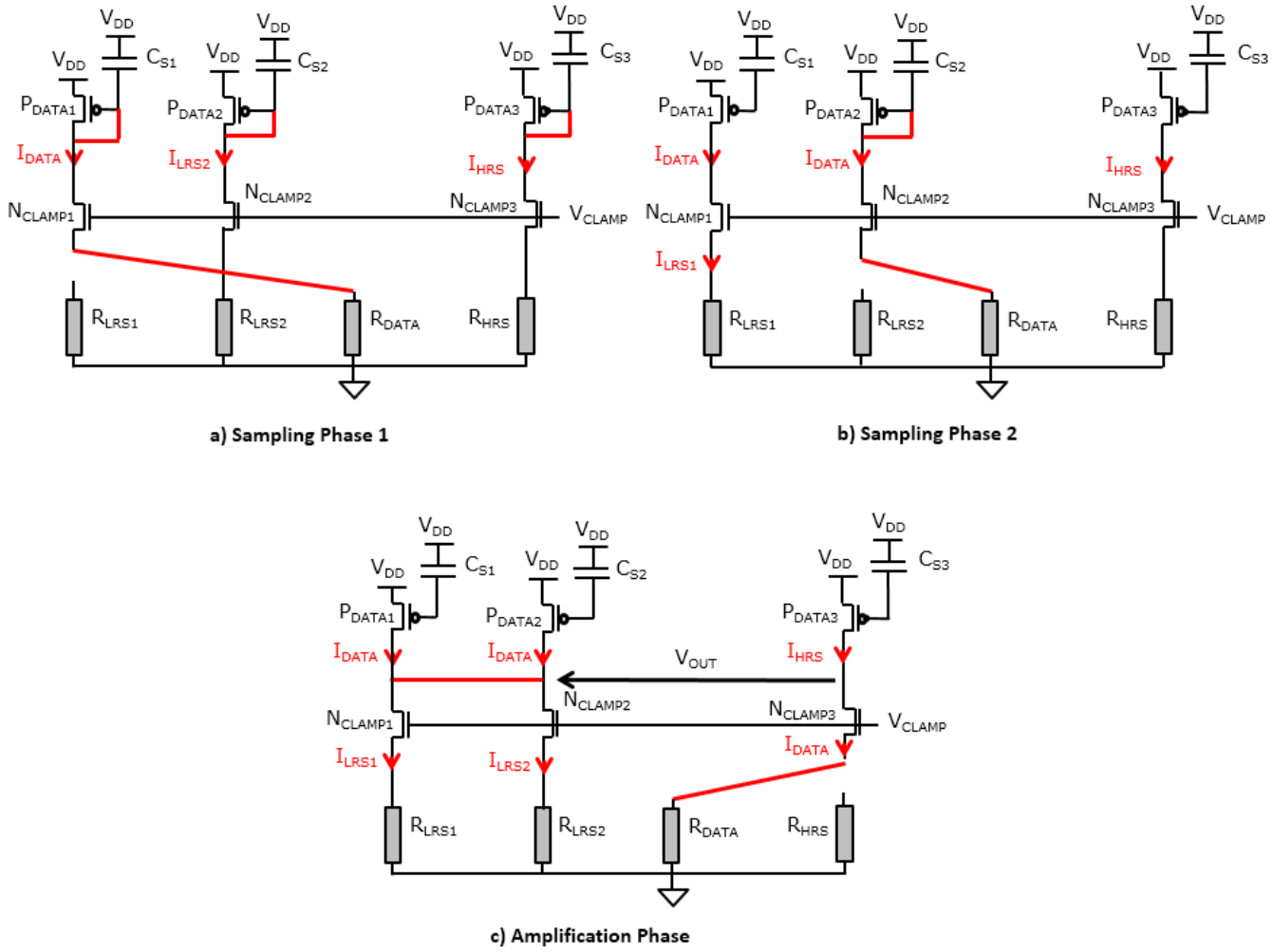


Figure 69 - Proposed sense amplifier architecture, for three references (e.g. for 2 LRS and one HRS)

4.3.2.Proposed architecture for N references

4.3.2.1. Current scheme for four references

With four reference resistances, according to the same principle, the current scheme and the corresponding variability reduction would be as below, adding another uncorrelated resistive device:

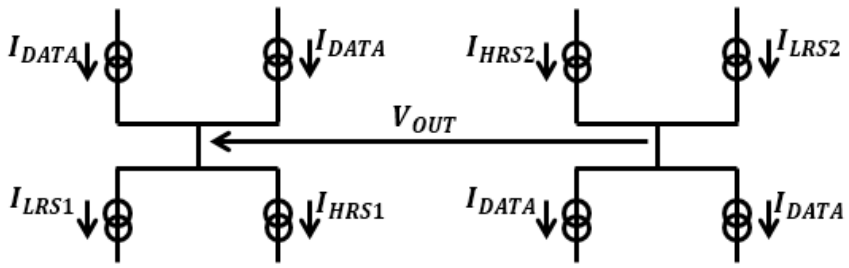


Figure 70 - Proposed current scheme for improved signal to offset ratio with four references

$$V_{OUT} = \frac{r_{out}}{2}\left(2I_{DATA} - I_{LRS1,ref} - I_{HRS1,ref}\right) - \frac{r_{out}}{2}\left(I_{HRS2,ref} + I_{LRS2,ref} - 2I_{DATA}\right)$$

Equation 56: Output voltage of the proposed current scheme with four uncorrelated references, at the end of the amplification phase

$$\mu(\Delta I) = 2I_{DATA} - \frac{I_{LRS1,ref}}{2} - \frac{I_{LRS2,ref}}{2} - \frac{I_{HRS1,ref}}{2} - \frac{I_{HRS2,ref}}{2}$$

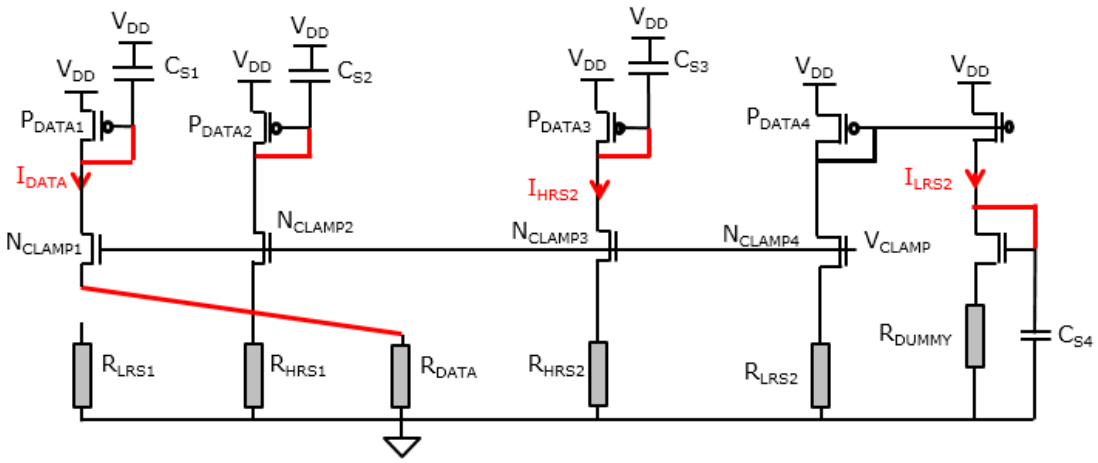
Equation 57: Current transfer function of the proposed current scheme with four uncorrelated references

$$\sigma(\Delta I) = \sqrt{4\sigma I_{DATA}^2 + \frac{\sigma I_{LRS,ref}^2}{2} + \frac{\sigma I_{HRS,ref}^2}{2}}$$

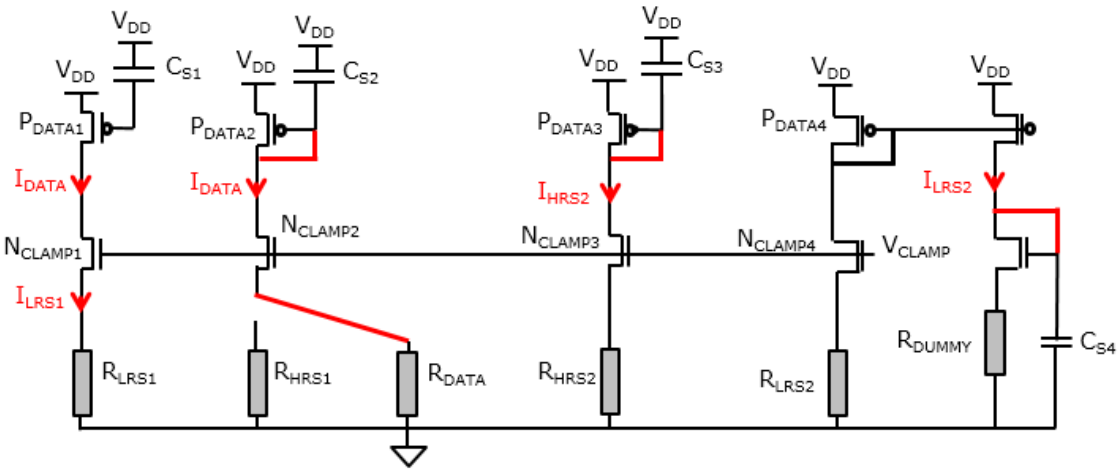
Equation 58: Illustration of reference variability reduction of the proposed current scheme with four uncorrelated references

4.3.2.2. Circuit implementation for four references

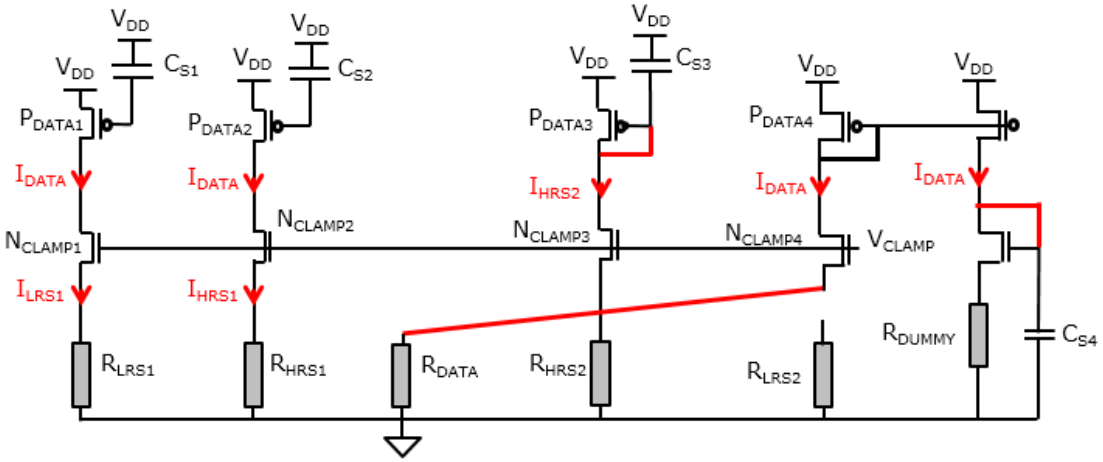
This scheme adds one design complexity. Indeed, two pulling down currents from the same resistive device (data cell) are needed here. It means that one of these two currents has to be recorded in one phase and then pulled down in order to generate in another phase the second pulling down data current. This is achieved with a current mirror (see *Figure 71*). However, this stage is an additional offset source. FOCT helps cancelling this additional offset since the same circuit is used between the sampling and the amplification phases. The role of resistance R_{DUMMY} depicted is to counterbalance output resistance mismatch created by the current mirror.



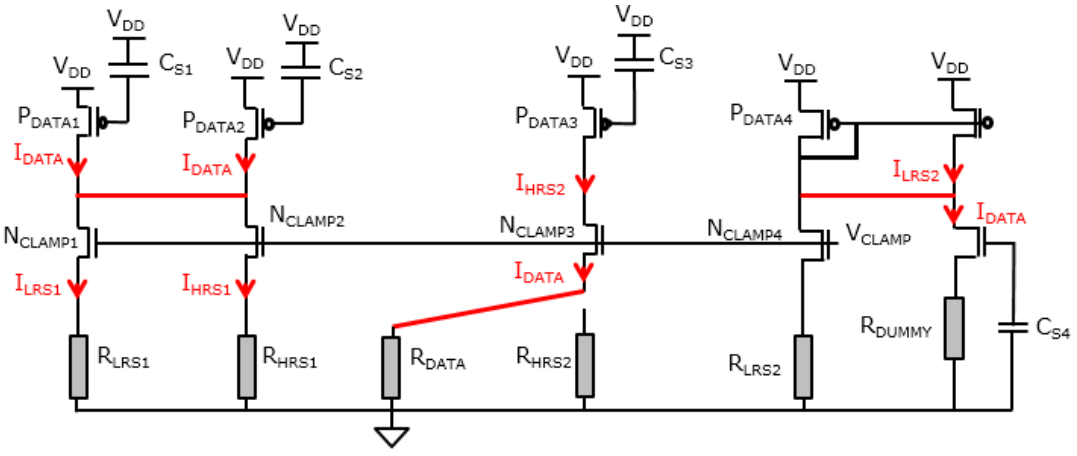
a) Sampling Phase 1



b) Sampling Phase 2



c) Sampling Phase 3



d) Amplification Phase

Figure 71 - Proposed sense amplifier architecture, for four references

This way, we can extend the idea to N references, and establish an offset-cancelled sense amplifier with the suitable signal to offset ratio improvement.

4.4. Read margin improvement and power, timing, area costs

4.4.1. Read margin improvement

In order to evaluate the signal to offset for N references, an iterative process is carried out. As the current scheme of the proposed architecture is different depending on whether the number of implemented references is odd or even, two expressions of the output signal variability are deduced:

$$\sigma(\Delta I) = \sqrt{4\sigma I_{DATA}^2 + \frac{\sigma I_{LRS,ref}^2}{N/2} + \frac{\sigma I_{HRS,ref}^2}{N/2}}, \text{ if } N \text{ is even}$$
$$= \sqrt{4\sigma I_{DATA}^2 + \frac{\sigma I_{LRS,ref}^2}{(N+1)/2} + \frac{\sigma I_{HRS,ref}^2}{(N-1)/2}}, \text{ if } N \text{ is odd}$$

Equation 59: Illustration of the reference variability reduction of the proposed sense amplifier for N references

Figure 72 shows the evolution of the signal to offset ratio improvement with the number of references used, for different process parameters (data resistance variability and on-off ratio of MTJs chosen as an example). The read margin results are compared to the read margin of a typical sense amplifier for a 2T-2R cell ($\mu(\Delta I) = 2.(I_{LRS} - I_{HRS})$), which represents the maximum obtainable read window.

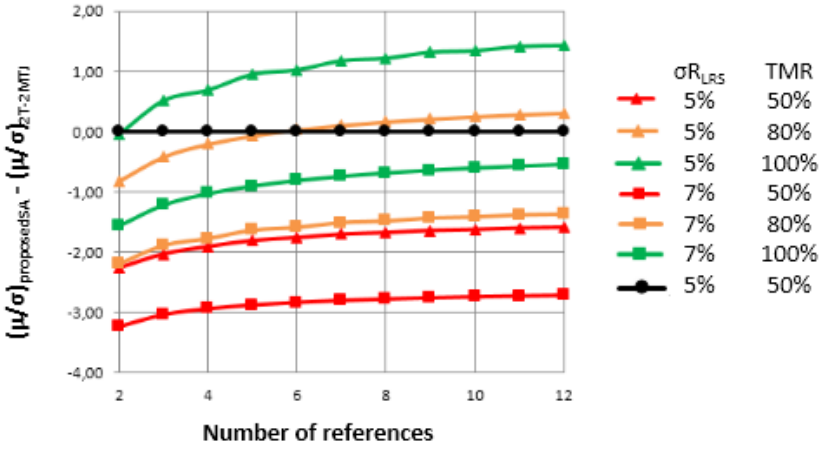


Figure 72 - Evolution of read margin improvement of the proposed sense amplifier with number of references for different process parameters (R_{LRS}=2.5 kΩ, V_{BL}=100mV)

This curve gives the equivalent resistive process to adopt in order to reach a given 2T-2R read margin. Above all, it shows a significant read margin gain from two to four references, regardless of the process selected. For σR_{LRS}=7% and TMR=100% (same process settings than in 3.2.3), around 0.5 sigma gain in terms of read margin is obtained, equivalent to a 10 factor gain in terms of bit error rate (see *Figure 73*).

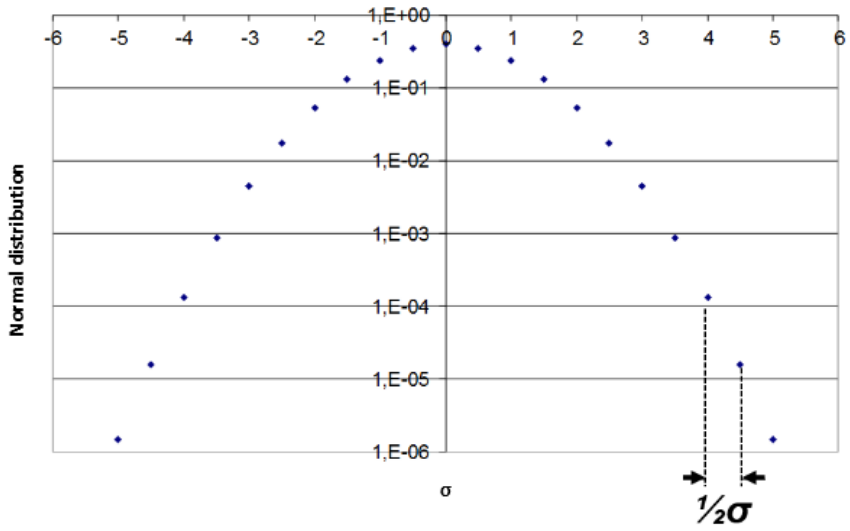


Figure 73 - Probability density function of the Normal distribution ($= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$) on a logarithmic scale

4.4.2.Power, timing and area costs

The circuit implementations proposed above highlight the non-negligible costs in terms of timing, area and power consumption, as additional phases and current paths are needed per additional references.
Regarding timing cost, adding one reference leads to one additional sampling phase of the order of ten nanoseconds. Therefore, for a 0.5 sigma read margin gain (four references for σR_{LRS}=7% and TMR=100%, see *Figure 72*), around twenty additional nanoseconds are necessary to read the resistive bit cell:

$$N_{phase} = N$$

Equation 60: Timing cost of the proposed sense amplifier for N references

As for power consumption, the energy required by the addition of two current branches and two sampling phases is non-negligible (5 times higher than a 2T-2MJ sense amplifier from two to four references, corresponding to a 10-factor read margin gain), to reach nevertheless less than ten Pico joules (see *Figure 74*):

$$E = \sum_i \sum_j E_{branch_i, phase_j} = \sum_i \sum_j V_{DD} \cdot \frac{V_{bl_{i,j}}}{R_{i,j}} dt_j$$

Equation 61: Energy required by the proposed sense amplifier for N references

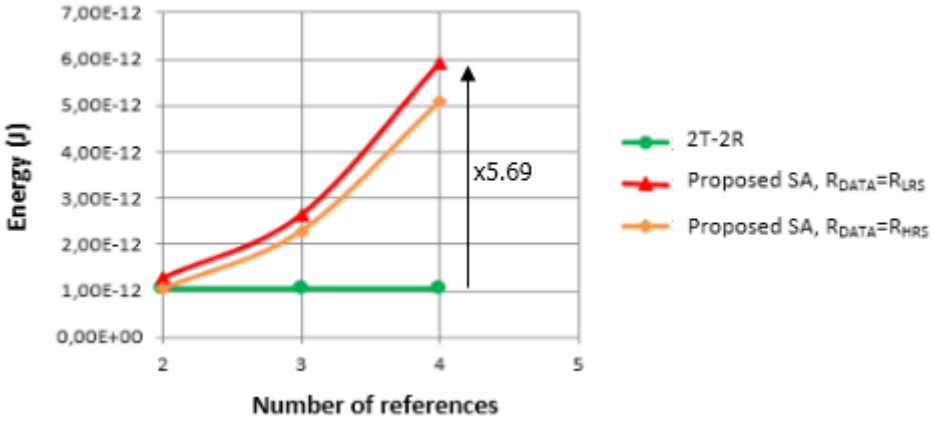


Figure 74 – Energy required by the proposed sense amplifier and comparison to a typical 2T-2R sense amplifier (R_{LRS}=2.5 kΩ, V_{BL}=100mV, TMR=100%)

Finally, about area cost, the increasing number of transistors w.r.t the number of references needs to be put into perspective by evaluating the area cost at array level. The number of bit lines per multiplexer for the proposed sense amplifier with four references still remains much lower than a 2T-2MTJ sense amplifier (44% gain, see *Figure 75*). These results justify the utility of this sense amplifier architecture for some timing-tolerant (tens of ns) and/or power-tolerant (tens of Pico joules) applications.

$$N_{branches} = \frac{3N}{2} - 1, \textit{ if N is even}$$
$$= \frac{3}{2}(N-1), \textit{ if N is odd}$$

Equation 62: Area cost of the proposed sense amplifier for N references

$$N_{BL} = N_{MUX} + N$$

Equation 63: Area cost of the proposed sense amplifier at array level for N_{MUX} data bit lines and N references

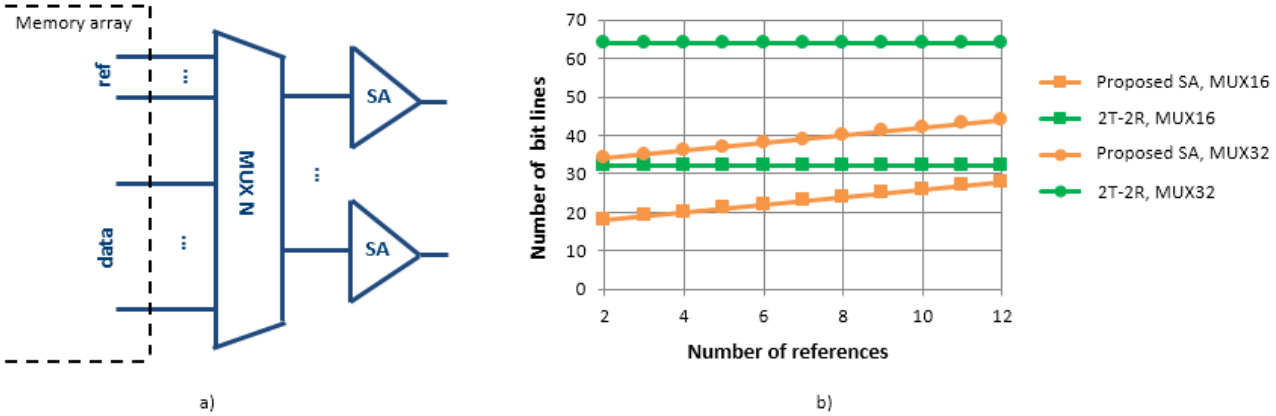
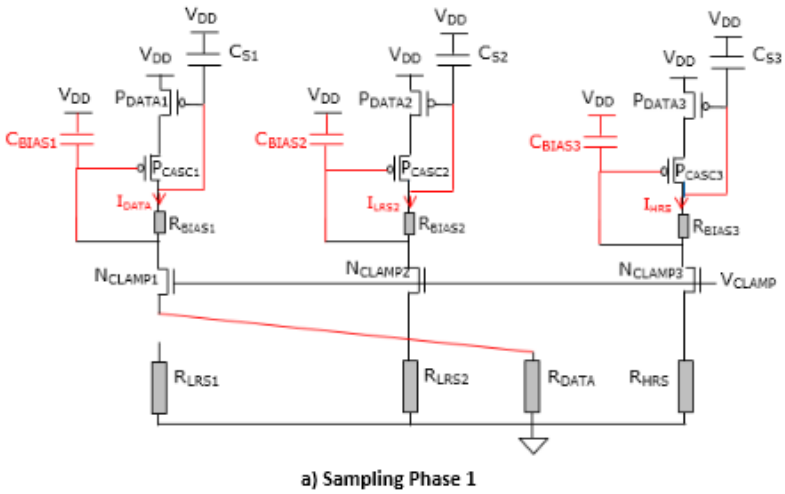


Figure 75 – a) SA area cost positioning in the memory array b) Array-level area cost of the proposed sense amplifier and comparison to a typical 2T-2R sense amplifier

4.5. Alternative implementations

Some alternative implementations and/or improvements of this proposed architecture are provided. First, in order to improve the signal to offset ratio, the gain of the proposed sense amplifier can be boosted using cascode devices. These devices have to be self-biased in order to alleviate complexity, area and power consumption issues of an external voltage bias. In order to self-bias the cascode stage during both sampling and amplification phase, capacitors C_{BIASi} are added, which sample the voltage bias needed for the cascode structure.



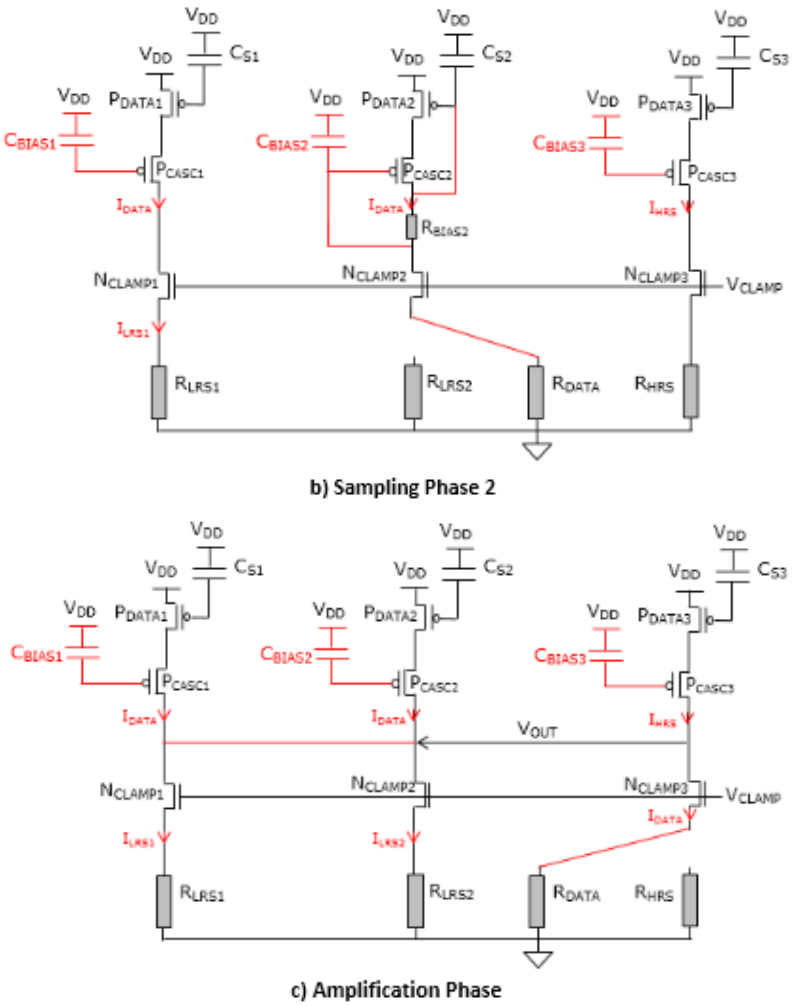
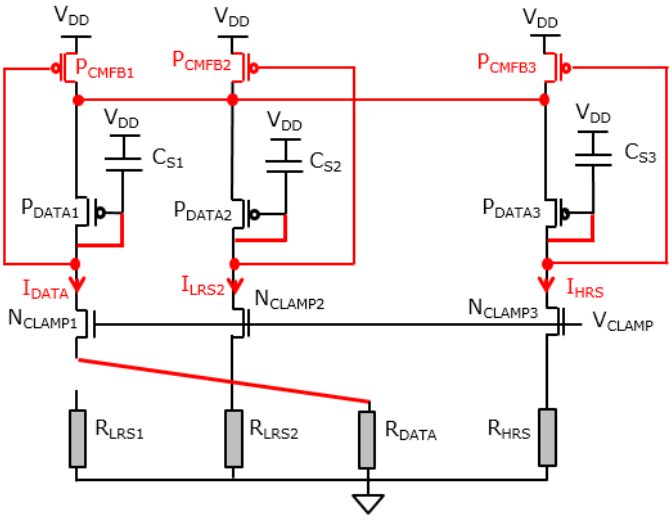
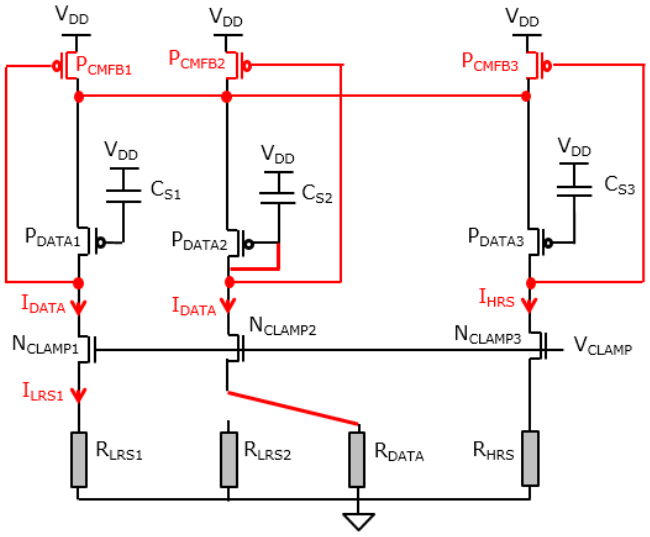


Figure 76 - Improvement of proposed sense amplifier using sampled self-biased cascode stage, during sampling and amplification phases (e.g for three references)

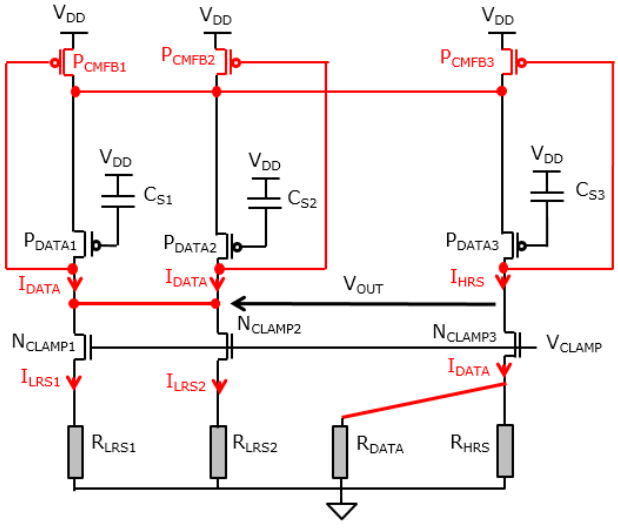
The circuit in *Figure 76* has a high gain, and its output voltage is taken between two high impedance nodes $V_{OUT+/-}$ (fully differential operation [47][86]). Therefore, a moderate common mode input signal variation (i.e moderate R_{HRS}/R_{LRS} variation) can lead to a large output common mode swing that consequently completely saturates the differential signal. In this case, the data resistive state is not detectable anymore. In order to improve the robustness of the circuit, a common-mode feedback (CMFB) configuration has to be implemented. A well-known CMFB circuit is based on two parallel transistors biased in triode region and connected to output nodes [47]. The CMFB structure is described in *Figure 77* where the "sampled self-biased cascode" is not shown (for the sake of circuit legibility).



a) Sampling Phase 1



b) Sampling Phase 2



c) Amplification Phase

Figure 77 - Proposed sense amplifier with common mode feedback implementation during sampling and amplification phases (e.g for 3 references)

Finally, we can notice that for each operating phase of the proposed sense amplifier, one of the N references is not exploited and can thus be used for a neighboring sense amplifier. This leads to an important reduction of the circuit footprint. *Figure 78* shows an example of timing and area optimization (similar to a pipelining scheme) of the proposed sense amplifier, describing the use of nine sense amps instead of 27 in only 10 clock cycles.

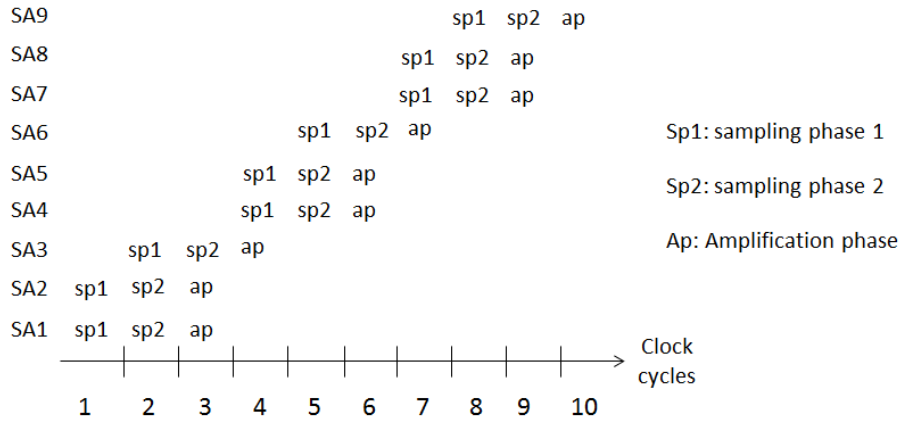


Figure 78 - Example of 'pipelining' sequencing for the proposed sense amplifier (e.g for 3 references)

4.6. Comparison with ECC and repair

The circuit architecture that was presented features non-negligible costs. They nevertheless should be compared to common system-level techniques of read yield enhancement (ECC and repair: see 2.4.1). A model of comparison between the proposed sense amplifier, ECC and repair is introduced below, in terms of bit error rate (BER) gain and area costs. Its parameters are summarized in *Table 5*:

Table 5 - BER model: input parameters

Bit cell	***Proposed Sense amplifier***	***ECC***	***Repair***
V_{BL}	*Number of references*	*Number ECC bits*	*Number of redundant words*
R_{LRS}	*MUX type*	*Word width*	*Number of redundant word lines*
On-off ratio		*Word depth*	*Number of redundant I/Os*
σR_{LRS}		*Memory size*	

4.6.1.Comparison with ECC

On a one hand is evaluated the equivalent ECC configuration, meaning the number of bits required, to generate the correcting code leading to the same read margin improvement than the proposed sense amplifier (for example, the 10 factor BER gain for four references). No redundancy is first assumed here.
Figure 79 depicts the evolution of a 10 kb memory-array failure probability as a function of the error rate for one bit in a 32 bits word, for different ECC configurations. This output results from the following equations:

$$P_{fail}(\text{memory}) = 1 - \binom{N_{rows}}{0} \left(P_{fail}(\text{row})\right)^{N_{rows}}$$

Equation 64: Failure probability of the memory array as a function

of failure probability of one memory row

$$P_{fail}(row) = 1 - \binom{N_{words/row}}{0} (P_{fail}(word))^{N_{words/row}}$$

Equation 65: Failure probability of one memory row as a function of failure probability of one data word

$$P_{fail}(word) = 1 - \binom{wordwidth}{N_{ECCbits}} (BER)^{wordwidth - N_{ECCbits}} * (1 - BER)^{N_{ECCbits}}$$

Equation 66: Failure probability of one data word as a function of failure probability of one bit and ECC configuration

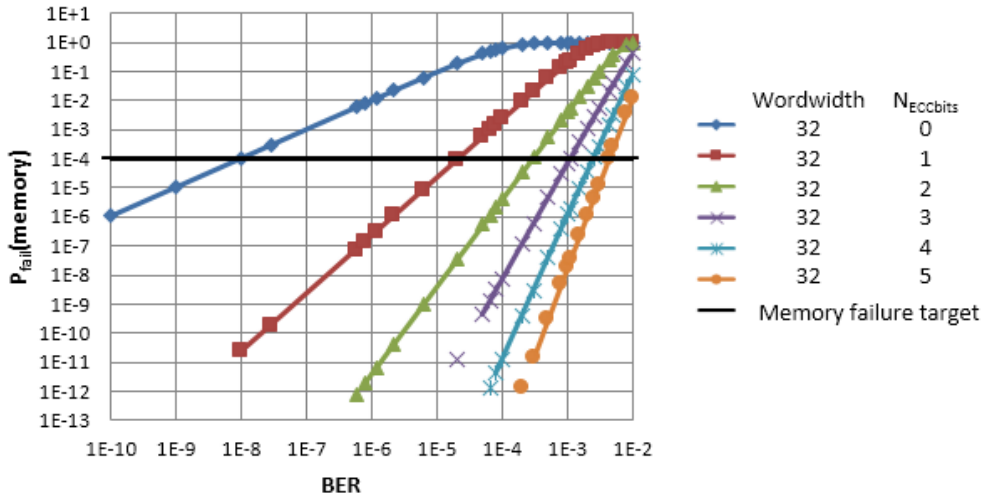


Figure 79 – Failure probability of the memory array as a function of the bit error rate for different ECC configurations (Memory array size = 10 kb, no redundancy)

For a given target of memory failure probability (10^{-4} in *Figure 79*), the BER gain obtained by the addition of one ECC-bit can be extracted. *Figure 80* demonstrates that no more than two ECC-bits are required to obtain the same BER gain than the proposed sense amplifier. This means that, in terms of read yield gain, it would be preferable to use the proposed sense amplifier rather than more than two bits ECC.

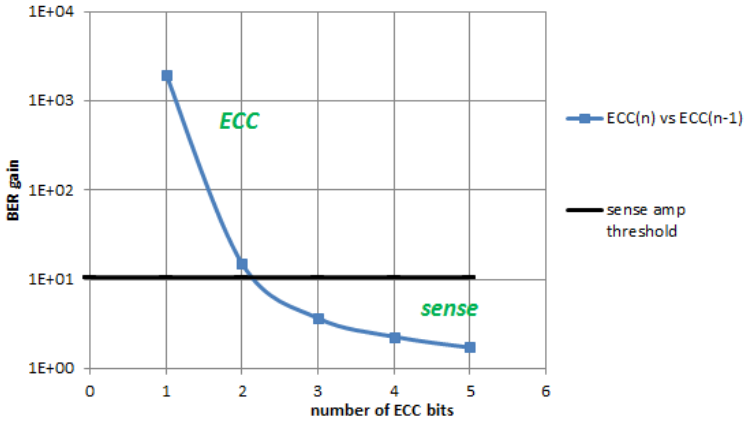


Figure 80 - Evolution of the BER gain for adding one ECC bit for different number of initial ECC bits, and comparison to the BER gain of the proposed sense amplifier for four references (V_{BL}=100 mV, R_{LRS}=2.5 kΩ, TMR=100%, σR_{LRS}=7%, memory array size = 10 kb, wordwidth = 32, no redundancy, targeted memory failure probability=10^{-4})

The total area overhead of the proposed sense amplifier is compared with ECC (*Table 6* and *Table 7*). It shows it has a better result for a 10-factor BER gain (four references or two ECC bits).

Table 6 - Proposed sense amplifier area overhead

	total sense area overhead (array+sense level)			
MUX	8	16	32	64
nb ref	%	%	%	%
3	*9,9*	*7,76*	*6,47*	*5,76*
4	*24,7*	*20,51*	*17,94*	*16,51*
5	*29,6*	*23,27*	*19,40*	*17,27*
6	*34,4*	*26,03*	*20,87*	*18,03*
8	*44,2*	*31,54*	*23,81*	*19,54*
10	*53,9*	*37,06*	*26,74*	*21,06*
12	*63,6*	*42,57*	*29,68*	*22,57*

Table 7 - ECC area overhead

ECC overhead (%)	ECC bits		
Data wordwidth	1	2	3
8	*50*	*100*	*150*
16	*31,25*	*62,5*	*93,75*
32	*18,75*	*37,5*	*56,25*
64	*10,94*	*21,88*	*32,81*
128	*6,25*	*12,5*	*18,75*
256	*3,52*	*7,03*	*10,55*
512	*1,95*	*3,91*	*5,86*

4.6.2.Comparison with repair

Repair is another technique of read yield enhancement that is based on memory-level redundancy. Three types of memory repair are studied here:

- IO repair (or row redundancy): consists in extra IO(s) redundancy in the memory array, including the corresponding bit lines, multiplexers and sense amplifiers;
- Word Line repair (or column redundancy): consists in extra word line(s) in the memory array;
- 2D repair: simultaneous row and column redundancy.

4.6.2.1. IO repair

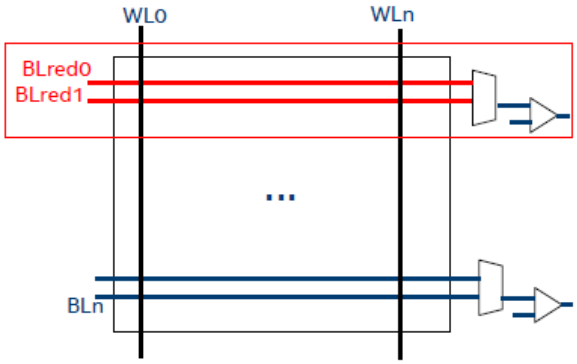


Figure 81 - Illustration of IO repair

The memory failure probability due to IO redundancy is described in *Equation 67* and *Equation 68*, using the same principle as memory failure probability due to ECC:

$$P_{fail}(\text{memory}) = 1 - \binom{N_{IO}}{N_{redIO}} \left(P_{fail}(IO)\right)^{N_{IO}-N_{redIO}} * (1 - P_{fail}(IO))^{N_{redIO}}$$

Equation 67: Failure probability of the memory array as a function of failure probability of one memory IO

$$P_{fail}(IO) = 1 - \binom{N_{bits/IO}}{0} (BER)^{N_{bits/IO}}$$

Equation 68: Failure probability of one memory IO as a function of failure probability of one bit

4.6.2.2. Word Line (WL) repair

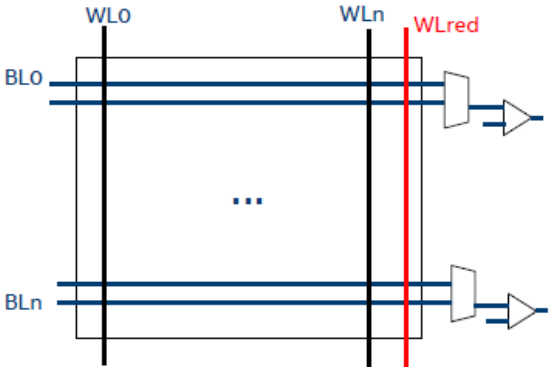


Figure 82 - Illustration of WL repair

In the same way as above is described the memory failure probability due to WL redundancy:

$$P_{fail}(\text{memory}) = 1 - \binom{N_{WL}}{N_{redWL}} \left(P_{fail}(WL)\right)^{N_{WL}-N_{redWL}} * (1 - P_{fail}(WL))^{N_{redWL}}$$

Equation 69: Failure probability of the memory array as a function of failure probability of one word line

$$P_{fail}(WL) = 1 - \binom{N_{bits/row}}{0} (BER)^{N_{bits/row}}$$

Equation 70: Failure probability of one word line as a function of failure probability of bit

Figure 83 shows that the equivalent repair configuration required to reach the BER gain of the proposed sense amplifier for four references is to integer no more than one redundant WL or IO.

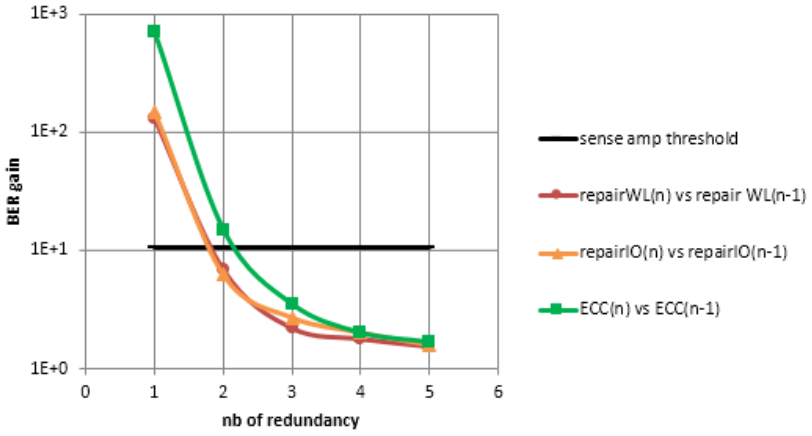


Figure 83 - Comparison of proposed sense amplifier for four references with ECC and IO and word line repair for a 10 kb memory array and a MUX16 in terms of read margin gain (V_{BL}=100 mV, R_{LRS}=2.5 kΩ, TMR=100%, σR_{LRS}=7%, wordwidth = 32, targeted memory failure probability=10^{-4})

Table 8 shows that one WL and IO redundancy requires non-negligible area and timing costs, justifying the use of the proposed sense amplifier as an alternative.

Table 8 - Comparison between IO and word line repair for a 10kb memory array with a MUX16 in terms of area and timing costs (wordwidth=32)

Redundancy type	***Repairable memory cells***	***Area***	***Timing***
1 WL	*128 (= 4 * 32) cells*	*1 additional column + special decoders + comparator*	*High due to the comparator before the memory access*
1 IO	*2048 (= 4*32*16) cells*	*32 Mux for the shift*	*Low because static*

4.6.2.3. IO + WL repair

[87] proposes a mathematical methodology for calculation of memory yield when two-dimensional redundancy is required, as we can see in *Figure 84* (simultaneous IO and WL repair).

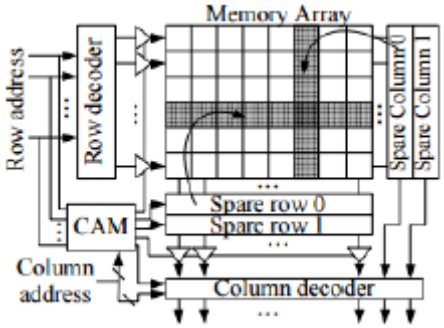


Figure 84 - Two-dimensional redundancy structure for memory repair [87]

It offers an algorithm of computation in order to find the best redundancy configuration for memory repair. The proposed algorithm assumes that defects allocation are known, meaning that a defect mapping algorithm is initially integrated for example in a memory Built-In Self Test (BIST) [88][89][90].
The probability YR that all defective cells on the memory array can be successfully repaired by its redundancy is defined as:

$$YR = \sum_{x=0}^{M} \sum_{rd=0}^{N} \sum_{cd=0} Dp(x, rd, cd).DSR(x, rd, cd)$$

Equation 71: Yield after repair assuming two-dimensional redundancy

where M is the total number of spare rows; N is the total number of spare columns; Dp(x,rd,cd) is the probability that x single defects, rd row defects and cd column defects occur; DSR(x,rd,cd) is the probability that the x single defects, rd row defects and cd column defects can be repaired by the M spare rows and N spare columns.

The probability Dp(x,rd,cd) is well known:

$$Dp(x, rd, cd) = \frac{\lambda_{SD}^{x} e^{-\lambda_{SD}}}{x!} . \frac{\lambda_{RD}^{rd} e^{-\lambda_{RD}}}{rd!} . \frac{\lambda_{CD}^{cd} e^{-\lambda_{CD}}}{cd!}$$

Equation 72: Probability that x single defects, rd row defects and cd column defects occur

where x, rd and cd are assumed independent random variables and modeled by a Poisson distribution. λ_{SD}, λ_{RD}, λ_{CD} are respectively the total number of single defects, row defect and column defects [87][91][92][93][94].

The probability DSR(x,0,0) (or DSR(x)) is the interest of our study since it focuses on single defect repair, in other words linked to the bit error rate. It is calculated in an iterative way through the four-dimensional parameter $S^{(x)}[m][n][z]$ representing the probability that x defects can be repaired by m spare rows, n spare columns and z spare units. A spare unit is a spare element that is not yet defined (either a spare row or a spare column) depending on the defect to repair (see *Figure 85* to *Figure 88*).

$$DSR(x) = \sum_{m=0}^{M} \sum_{n=0}^{N} \sum_{z=0}^{M+N-m-n} S^{(x)}[m][n][z]$$

Equation 73: Probability that x single defects can be repaired by M spare rows and N spare columns

The (i+1)th defect probability $S^{(i+1)}[m][n][z]$ as a function of the (i)th one depends on four possibilities of allocation for this new defect:

$$\begin{aligned} S^{(i+1)}[m][n][z] = {} & S^{(i)}[m][n][z] * p_1(i,m,n,z) \\ & + S^{(i)}[m-1][n][z+1] * p_2(i,m-1,n,z+1) \\ & + S^{(i)}[m][n-1][z+1] * p_3(i,m,n-1,z+1) \\ & + S^{(i)}[m][n][z-1] * p_4(i,m,n,z-1) \end{aligned}$$

with: $S^{(1)}[0][0][1] = 1$ and $S^{(1)}[m][n][z] = 0$ otherwise.

Equation 74: Probability of repair for adding a new defect in a memory array

The event associated with p_1 occurs when the extra (i + 1)th defect falls in a location covered by the m spare rows or the n spare columns repairing the original i defects. Hence, no extra spare row, column, or unit needs to be used to repair this (i+ 1) defect. *Figure 85*-a) shows an instance of this probability event. The computation of $p_1(i,m,n,z)$ is described in *Equation 75*. It is based from the illustration of *Figure 85*-b), which illustrates the area of repairable defects by m spare rows and n spare columns (R is the total number of rows, C the total number of columns):

$$p_1 = \frac{(R*n + C*m - m*n - (i-z))}{R*C - i}$$

Equation 75: Probability that the new defect's location is already covered by the existing spare rows or columns

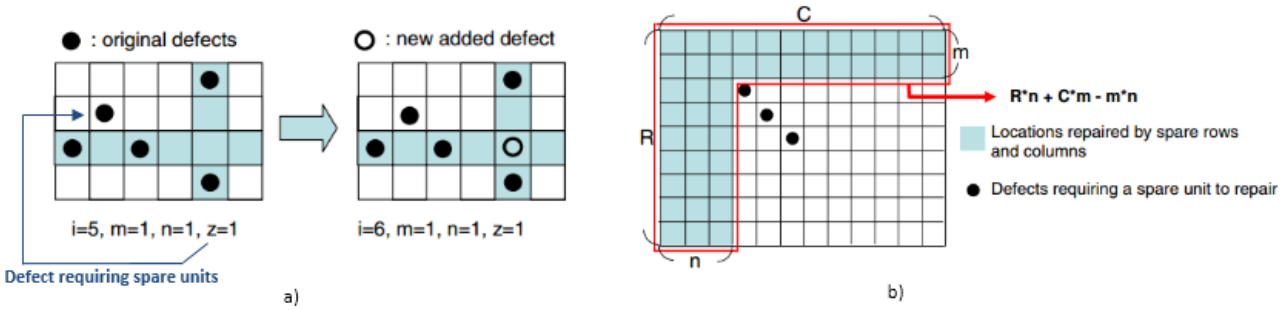


Figure 85 – a) An instance of the probability event associated with p_1 b) Illustration for the computation of $p_1(i,m,n,z)$ [87]

The event associated with p_2 occurs when the extra (i+ 1)th defect falls in a location where we need to fix the usage of a spare unit as a spare row to repair this defect. Hence, the number of used spare rows m is increased

by 1 and the number of spare units z is decreased by 1. *Figure 86*-a) shows an instance of this probability event. The computation of $p_2(i,m,n,z)$ is described in *Equation 76*. It is based from the illustration of *Figure 86*-b), which illustrates the area where a spare unit has to be turned into a spare row to repair the extra defect:

$$p_2 = \frac{\big((C - n - z) * z + 0.5 * z * (z - 1)\big)}{R * C - i}$$

Equation 76: Probability that the new defect's location requires a spare unit to be turned into a spare row

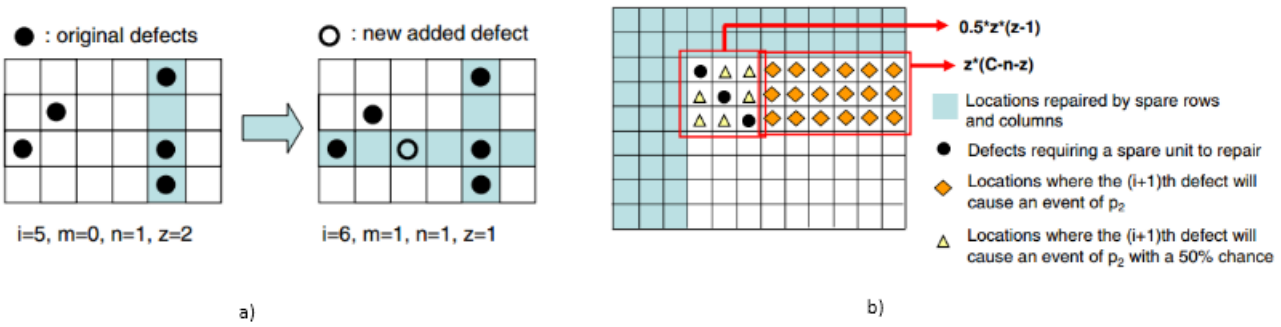


Figure 86 - a) An instance of the probability event associated with p_2 b) Illustration for the computation of $p_2(i,m,n,z)$ [87]

The event associated with p_3 occurs when the extra (i + 1)th defect may fall in a location where we need to fix the usage of a spare unit as a spare column to repair this defect. Hence, the number of used spare column n is increased by 1 and the number of spare units z is decreased by 1. *Figure 87*-a) shows an instance of this probability event. The computation of $p_3(i,m,n,z)$ is described in *Equation 77*. It is based from the illustration of *Figure 87*-b), which illustrates the area where a spare unit has to be turned into a spare column to repair the extra defect:

$$p_3 = \frac{\big((R - m - z) * z + 0.5 * z * (z - 1)\big)}{R * C - i}$$

Equation 77: Probability that the new defect's location requires a spare unit to be turned into a spare column

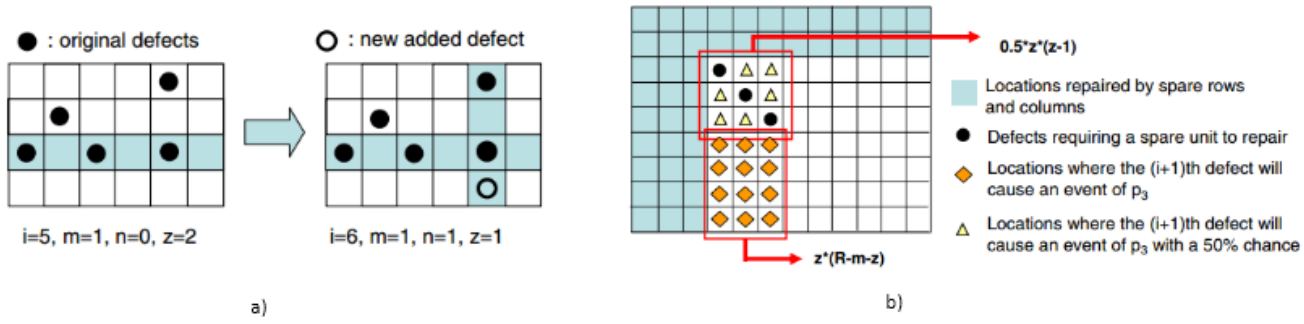


Figure 87 - a) An instance of the probability event associated with p_3 b) Illustration for the computation of $p_3(i,m,n,z)$ [87]

Finally, the event associated with p_4 occurs when the extra (i + 1)th defect may fall in a location where we can use another spare unit to repair this defect. Hence, the number of spare units z is increased by 1. *Figure 88*-a) shows an instance of this probability event. The computation of $p_4(i,m,n,z)$ is described in *Equation 78*. It is based from the illustration of *Figure 88*-b), which illustrates the area where an extra spare unit is needed to repair the extra defect:

$$p_4 = \frac{\big((R - m - z) * (C - n - z)\big)}{R * C - i}$$

Equation 78: Probability that the new defect's location requires an extra spare unit

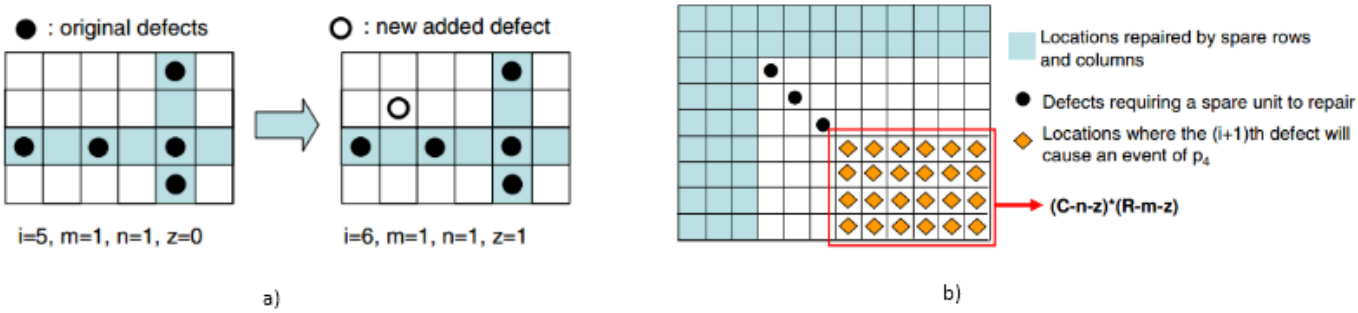


Figure 88 - a) An instance of the probability event associated with p_4 b) Illustration for the computation of $p_4(i,m,n,z)$ [87]

Table 9 depicts the equivalent BER (obtained from the single defects number λ_{SD}, see *Equation 79*) according the number of spare rows or spare columns for a given memory failure target (obtained from the yield after repair YR of *Equation 71*, see *Equation 80*) target. This result show that the equivalent IO+WL configuration to get a similar gain than the proposed sense amplifier from two to four references (10x) corresponds to using more than 10 (or 14 or 18) instead of 2 spare rows and 22 (or 18 or 14) instead of 2 spare columns. The corresponding area and timing costs for this configuration can be deduced from *Table 8*: adding only one column actually requires additional decoders and one more comparator, while adding one row leads to 32 additional multiplexers (for a word width equal to 32). This demonstrates the interest of the proposed periphery circuit-level alternative to improve read yield of resistive memories and alleviate costly system-level solutions.

$$BER = \frac{\lambda_{SD}}{mem\_size}$$

Equation 79: BER due to IO+WL repair as a function of single defects number

$$P_{fail}(memory) = 1 - YR$$

Equation 80: Failure probability of the memory array as a function of the yield after IO+WL repair

Table 9 - BER (obtained fr65 λ_{SD}) according the IO+WL repair configuration for a 10kb memory array (128 rows, 128 columns)

# of spare rows (M) / # of spare columns (N)	2	6	10	14	18	22
	λ_{SD} for YR=99.99% (i.e. $P_{fail}(memory)=10^{-4}$)					
2	**2.036**	4.222	6.096	7.602	7.956	7.963
6	4.222	6.124	8.022	10.02	12.418	15.548
10	6.096	8.022	10.02	12.424	15.788	**19.805**
14	7.602	10.021	12.427	15.74	**19.75**	24.13
18	7.956	12.418	15.788	**19.75**	24.15	28.93
22	7.963	15.548	**19.805**	24.15	28.93	34.11

4.7. Conclusion

Variation-tolerance has to be taken into account at circuit-level by implementing offset-cancelled sense amplifier architectures that also handles layout regular reference schemes. A sense amplifier architecture was proposed above, featuring offset-cancellation and time-multiplexed reference scheme. This circuit allows significant signal to offset ratio enhancement by considerably reducing reference variability. Read margins estimation shows a factor 10 improvement in the BER when implemented four references to this architecture compared to two references. The non-negligible timing and above all area costs of this sense amplifier were compared to system-level read yield improvement techniques (ECC, word line repair, IO repair, 2D IO and WL repair) and show that the proposed scheme can position as a circuit-level alternative solution, in particular for some timing and power-tolerant applications. Indeed, for example, more than 10 spare rows instead of two and more than 14 spare columns instead of two would be required to get a factor 10 improvement in the BER using 2D repair techniques, which is a considerably area-costly configuration.
The following deals with the design of this architecture and details pre and post-layout simulated offset results.

References of Chapter 4

[86] D. R. Holberg and P. E. Allen:, *CMOS Analog Circuit Design: 2nd (Second) edition*, Oxford Uni. 2002.

[87] M. C.-T. Chao, C.-Y. Chin, and C.-W. Lin, “Mathematical yield estimation for two-dimensional-redundancy memory arrays,” in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2010, pp. 235–240.

[88] Tsu-Wei Tseng, Jin-Fu Li, and Da-Ming Chang, “A built-in redundancy-analysis scheme for RAMs with 2D redundancy using 1D local bitmap,” in *Proceedings of the Design Automation & Test in Europe Conference*, 2006, p. 6 pp.

[89] S. Babu M, S. Kumar Reddy, and S. V V Sateesh, “BuiltIn SelfRepair for multiple RAMs with different Redundancies in a SOC,” *Int. J. Comput. Appl.*, vol. 24, no. 8, pp. 26–29, Jun. 2011.

[90] Mentor, “Tessent SilliconInsight.” [Online]. Available: https://www.mentor.com/products/silicon-yield/products/silicon-insight.

[91] N. H. Ramadan, S. I. Yield, and I. Corp, “Redundancy Yield Model for SRAMS,” *Intel Technol. J.*, no. 3, pp. 1–8, 1997.

[92] S. Shoukourian, V. Vardanian, and Y. Zorian, “SoC yield optimization via an embedded-memory test and repair infrastructure,” *IEEE Des. Test Comput.*, vol. 21, no. 3, pp. 200–207, May 2004.

[93] L. Youngs and S. Paramanandam, “Mapping and repairing embedded-memory defects,” *IEEE Des. Test Comput.*, vol. 14, no. 1, pp. 18–24, 1997.

[94] R.-F. Huang, R.-F. Huang, R.-F. Huang, R.-F. Huang, J.-F. Li, J.-F. Li, J.-F. Li, J.-F. Li, J.-C. Yeh, J.-C. Yeh, J.-C. Yeh, J.-C. Yeh, C.-W. Wu, C.-W. Wu, C.-W. Wu, and C.-W. Wu, “Raisin: Redundancy Analysis Algorithm Simulation,” *IEEE Des. Test Comput.*, vol. 24, no. 4, pp. 386–396, Apr. 2007.

Chapter 5: Proposed Sense Amplifier: Test Structure Implementation and Simulation Results

5.1. Introduction

Within the scope of this thesis and as a proof a concept, a test structure of the proposed sense amplifier architecture has to be designed to demonstrate offset cancellation for very low technologic node (sub-30 nm). For sake of simplicity and legibility, the proposed architecture design is only described for two references (see *Figure 89*) but the principle can be extended for more.

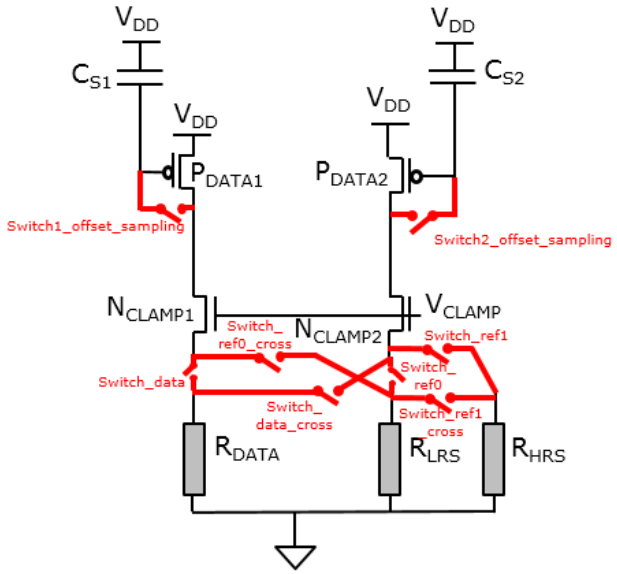


Figure 89 - Proposed sense amplifier architecture chosen for test structure design

5.2. Test Structure description

Figure 90 illustrates the input and output signals that result from the proposed sense amplifier architecture, in order to design a test structure. The clock signal (clk) generates two complementary signals to handle offset sampling and logic block switches. The details of the pads chosen for this structure are described and explained in 5.2.1.4.

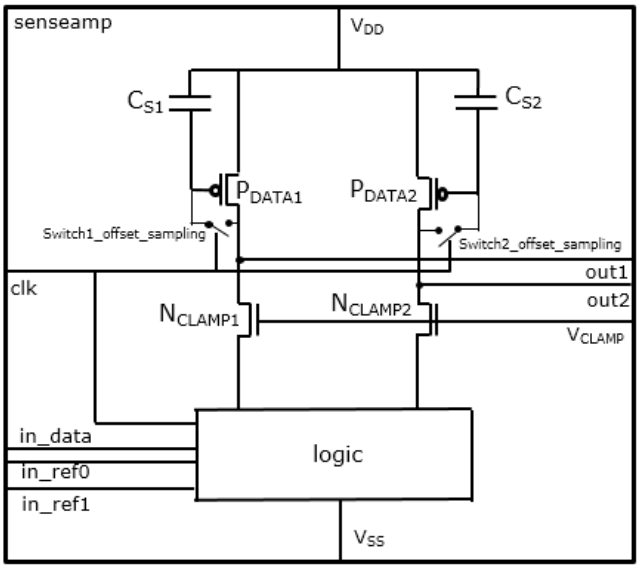


Figure 90 –Simplified description of the input/output signals of the proposed sense amplifier test-structure

The design of this structure is divided into different blocks that are detailed below.

5.2.1.Design sub blocks

The test structure is made of three main stages:

- The logic stage: based on NMOS switches, in order to invert when necessary the inputs during either the sampling or the amplification phase.
- The offset cancellation stage: mainly based on stacked (i.e. in-series) bit line voltages (leading to reduced mismatch) clamping NMOSs and symmetric-and-cascoded sampling MOS capacitors.
- The latch stage: perfectly symmetric and based on stacked transistors in order to minimize mismatch effects. The efficiency of stacked transistors on variation-tolerance is analyzed and discussed in [95][96][97][98].

5.2.1.1. Logic stage

This stage operates the inputs switching required to achieve the time-multiplexed reference scheme. It is made of four main NMOS low power switches, in order to minimize leakage during the phase they are turned off. Two more switches are added in order to initialize the bit line voltage of the reference that is not connected to the sense amplifier in the first phase (high reference 'ref1' in our case) and so reduce amplification phase timing overhead (*Figure 91*).

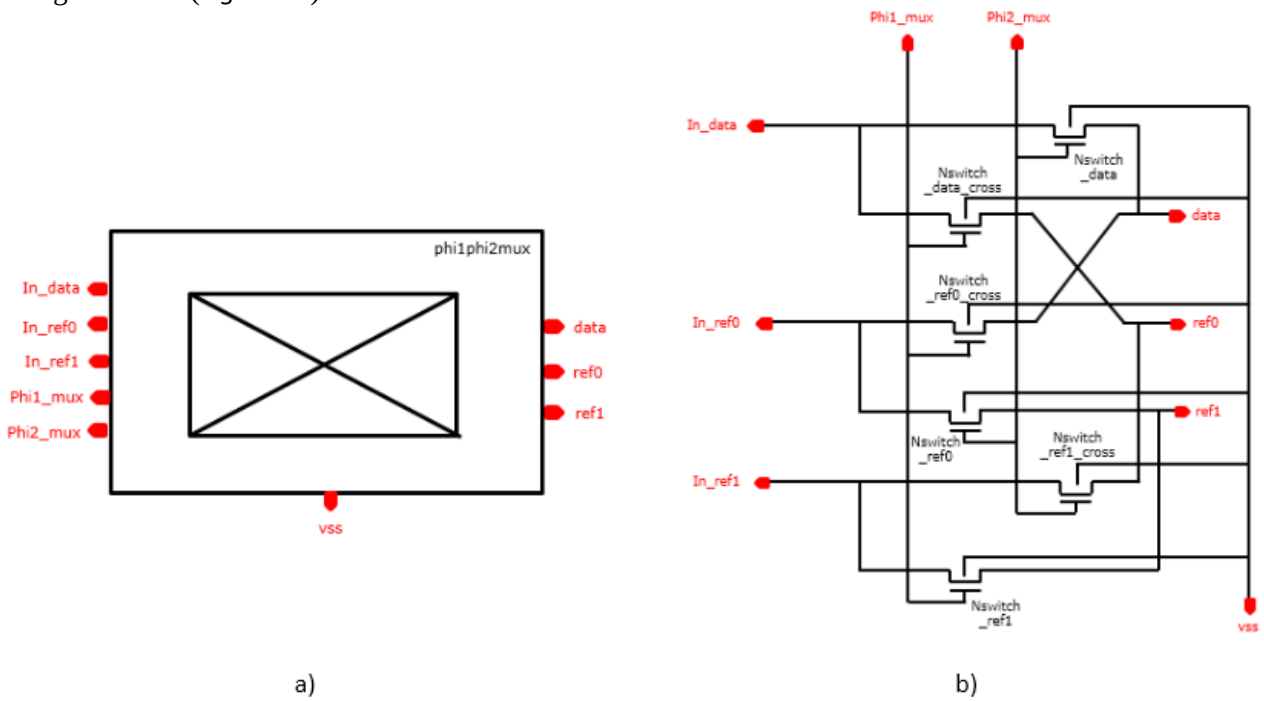


Figure 91 - Reference Time-multiplexing block a) symbol b) schematic

5.2.1.2. Offset cancellation stage

Figure 92 describes on one hand the transistors required in one current branch for offset-cancellation (oc_half stage). This includes PMOSs to trigger and achieve common-mode feedback (Pcmfb0, Pcmfb1, Pswitch_cmfb), PMOS for cascode (Pcascode), stacked data PMOSs for mismatch reduction (Pdata0, Pdata1, Pdata2), and finally pass gates (Nswitch_vbias, Pswitch_vbias, Nswitch_offset, Pswitch_offset) and MOS capacitors (Psampling_vbias, Psampling_offset) for cascode bias voltage and offset contribution sampling.

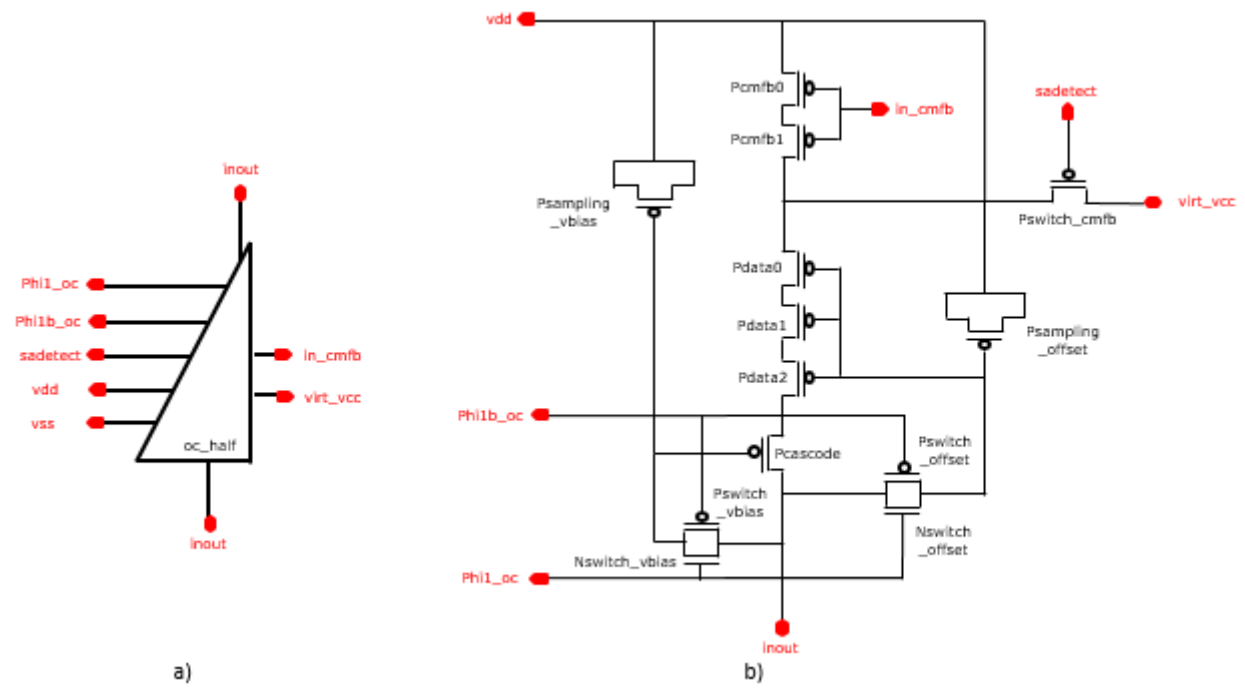


Figure 92 - Offset-cancellation half stage a) symbol b) schematic

The whole offset-cancellation stage is built with two symmetric 'oc_half' stages to connect to the two input current branches, as well as stacked bit line voltage clamping NMOSs (nclamp_data(0:7), nclamp_ref0(0:7)). To ensure layout regularity, an additional current branch for 'ref1' input is implemented. Similarly, dummies of this branch is added to ensure layout symmetry (*Figure 93*).

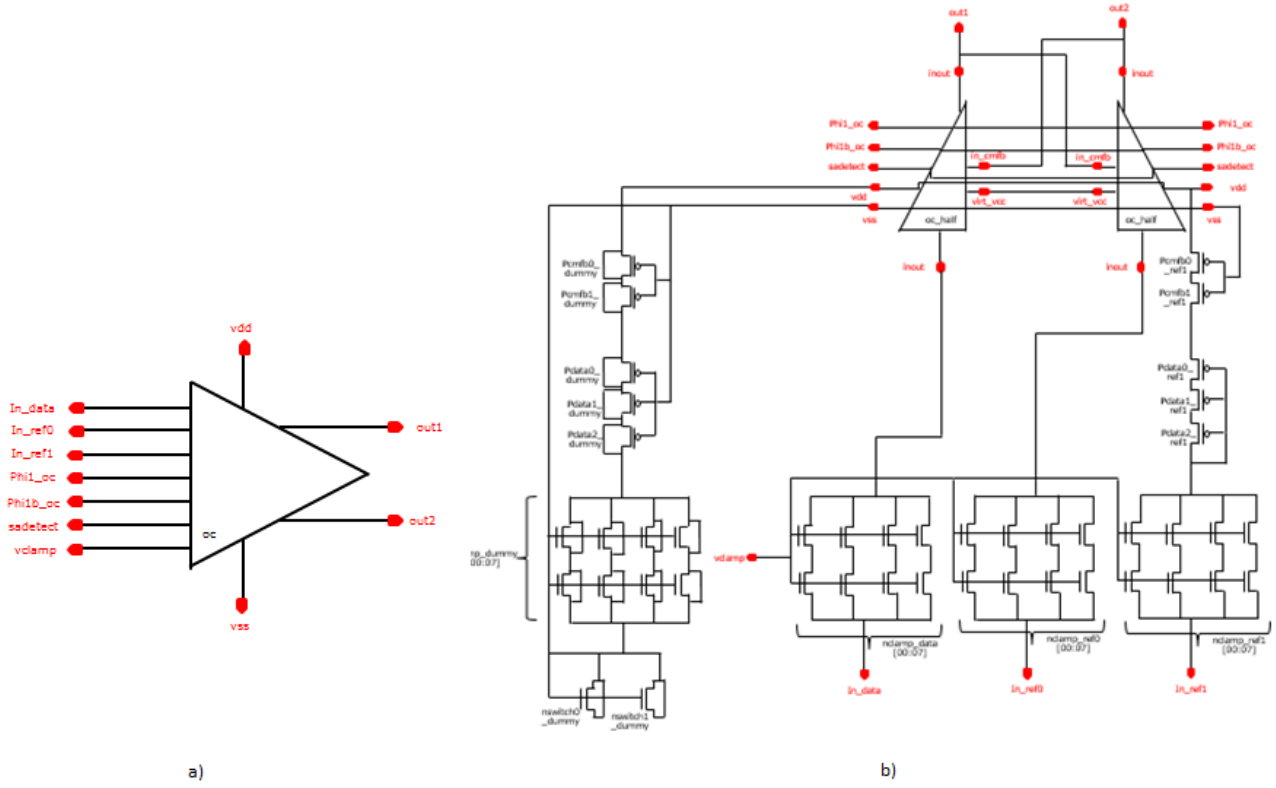


Figure 93 - Offset cancellation stage a) symbol b) schematic

5.2.1.3. Latch stage

An optimized latch circuit is also connected. As this is the last implemented stage, the offset of this variation-sensitive architecture has considerably less impact than the first stage. The PMOS (Platch[0:2]_left, Platch[0:2]_right) and NMOS (Nlatch[0:2]_left, Nlatch[0:2]_right) latch transistors as well as PMOS (Psaen[0:2]_left, Psaen[0:2]_right) and NMOS (Nsaen[0:2]_left, Nsaen[0:2]_right) switches to enable the sensing, are also stacked. A NAND gate is used to trigger the common-mode feedback (thanks to 'Sadetect' signal) of the offset-cancellation when a non-null output voltage is detected by the latch stage (*Figure 94*).

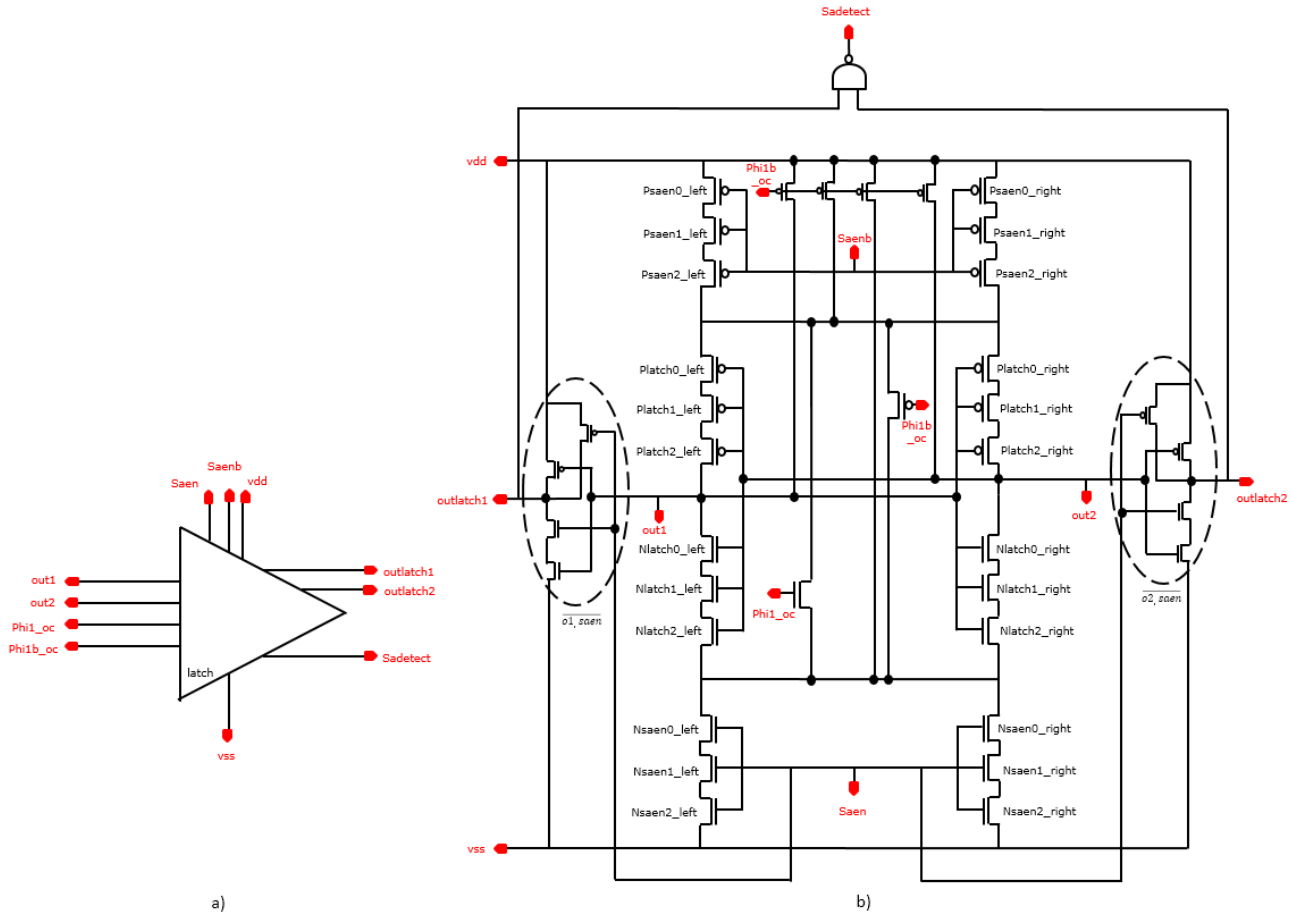


Figure 94 - Latch stage a) symbol b) schematic

To complete the whole sense amplifier structure (*Figure 95*), some inverter chains are required in addition to the three sub-blocks described above. These chains help straighten up the rising or falling edges duration for the phase generation signals (phi1_mux, phi2_mux, phi1_oc, phi1b_oc) and the sensing-triggering signal (latch/saen). Some output buffers gives the digital output signal (out_buf) and helps handling high load capacitances (around 50 pF).

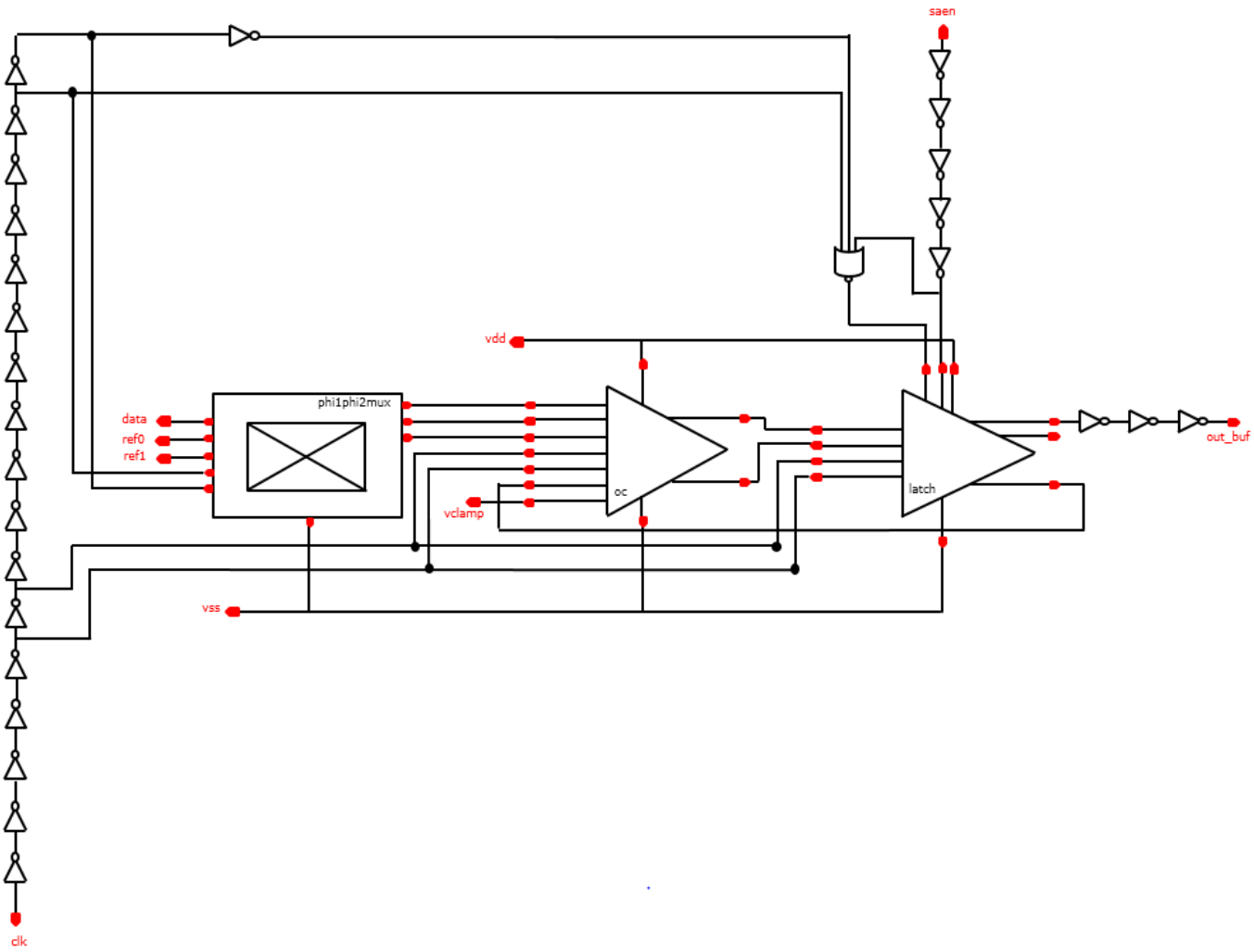


Figure 95 - Proposed sense amplifier ('testrow_sa'): schematic

5.2.1.4. In/out pads

From the final sense amplifier schematic can thus be extracted the in/out pads of the test structure (Figure 96). The three input signals (data, LRS and HRS reference) are chosen as DC input currents since the resistive devices are not provided and in order to only handle sense amplifier variability (and so to pursue the study independently of the resistive memory technology). 'Saen' and 'clk' are voltage pulses that generate the signals triggering sampling and amplification phases as well as sensing latch stage. It is important for offset-cancellation phases to generate non-overlapping signals so that sampling and amplification phases are decoupled. Only one digital output ('out_buf') is sufficient, as the outputs of the latch stage are complementary.

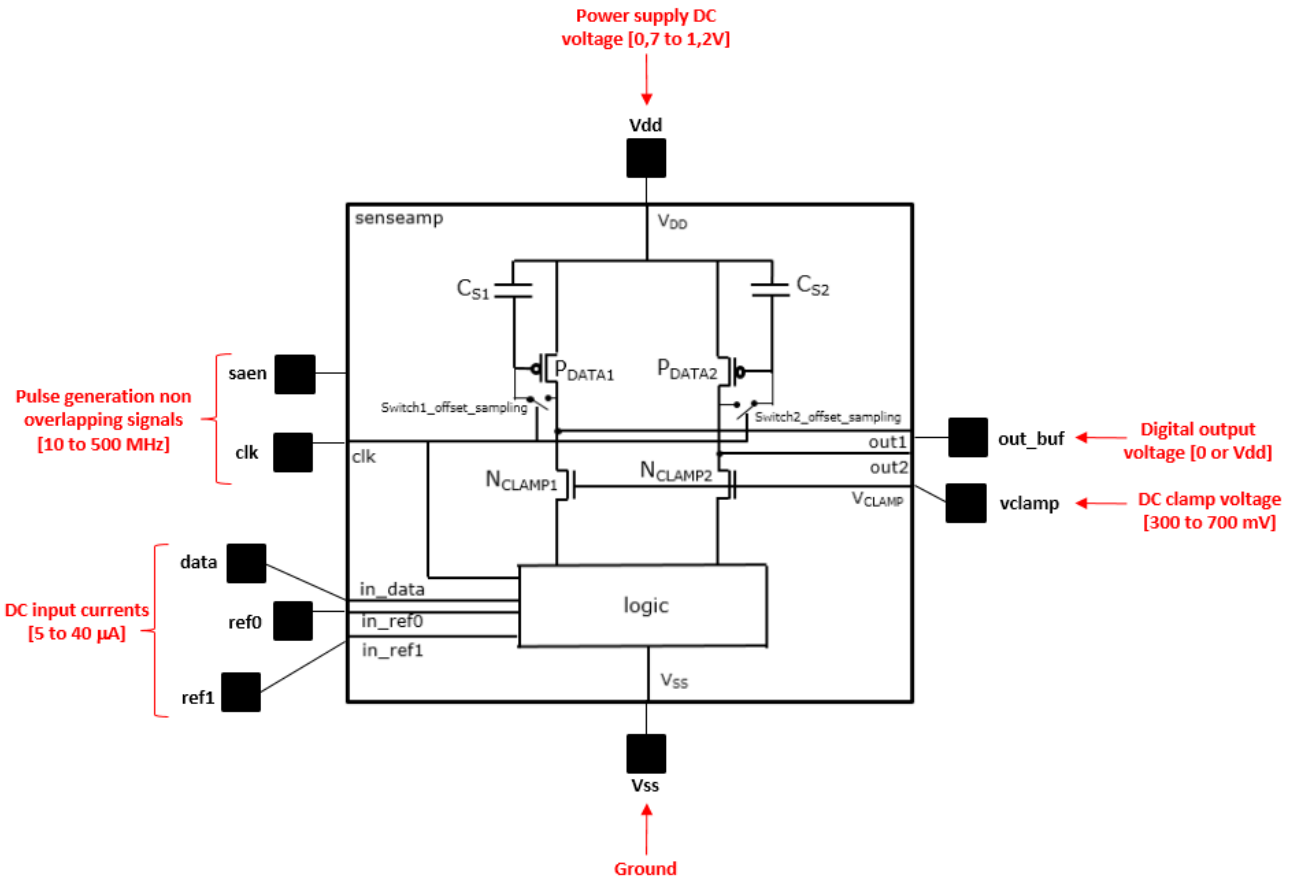


Figure 96 - Proposed sense amplifier test structure: in/out pads and signals description

To connect the sense amplifier structure and its high-level metallic pads, a test row is designed, in which many instantiations of the same structure are possible (Figure 97). The sense amplifier has to be placed under power supply for short routing. The three input pads locations are close to the structure to minimize parasitic resistances and capacitances, as it was demonstrated that parasitics have a significant impact on read margin.

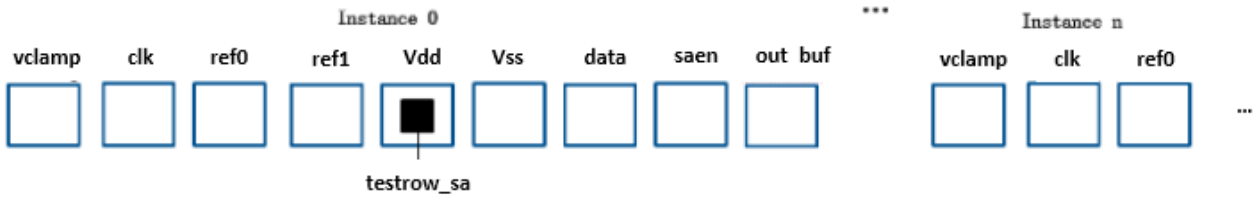


Figure 97 - Proposed sense amplifier test row: pads assignment

5.2.2.Waveforms

The proposed sense amplifier test-structure is first simulated to check the correct working operations of the different stages.

Figure 98 and *Figure 99* describe the sampling and amplification phase as well as the latch stage-triggering signal. It also highlight the importance of inverting input signals after the end of offset contributions sampling, and of taking into account the time required for switches charge injection.

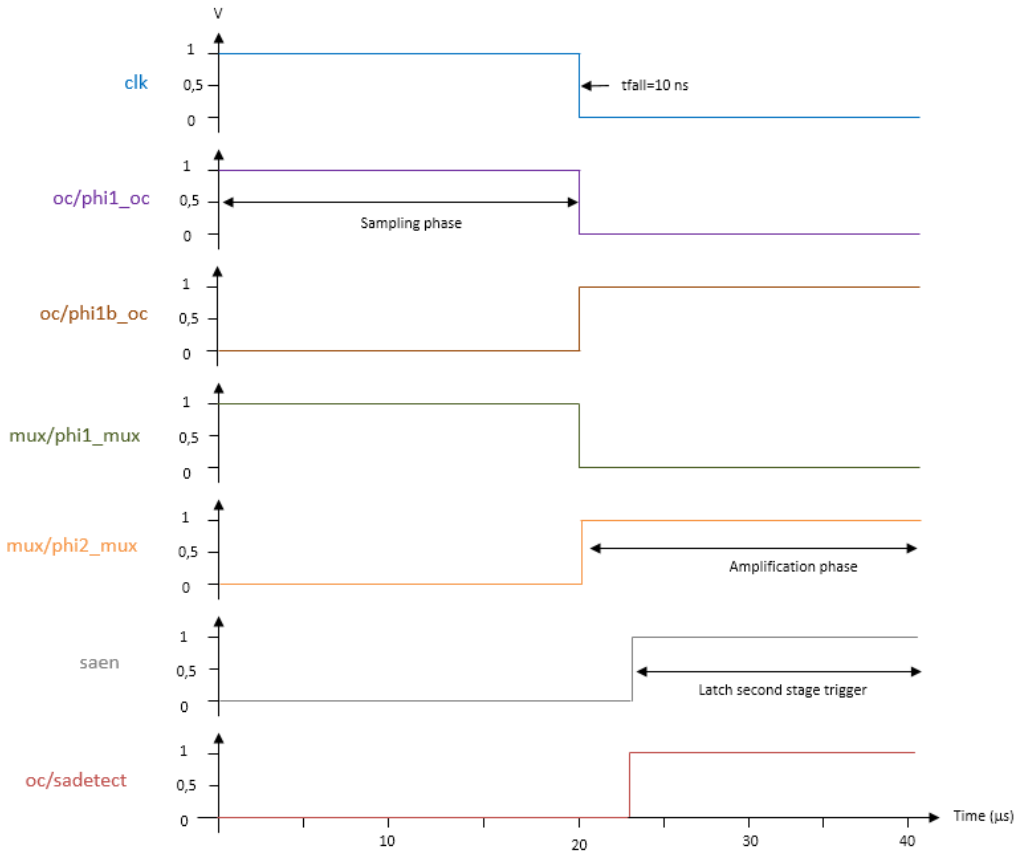


Figure 98 - Pre-layout simulations of pulse generation non-overlapping signals: Sampling and amplification phases

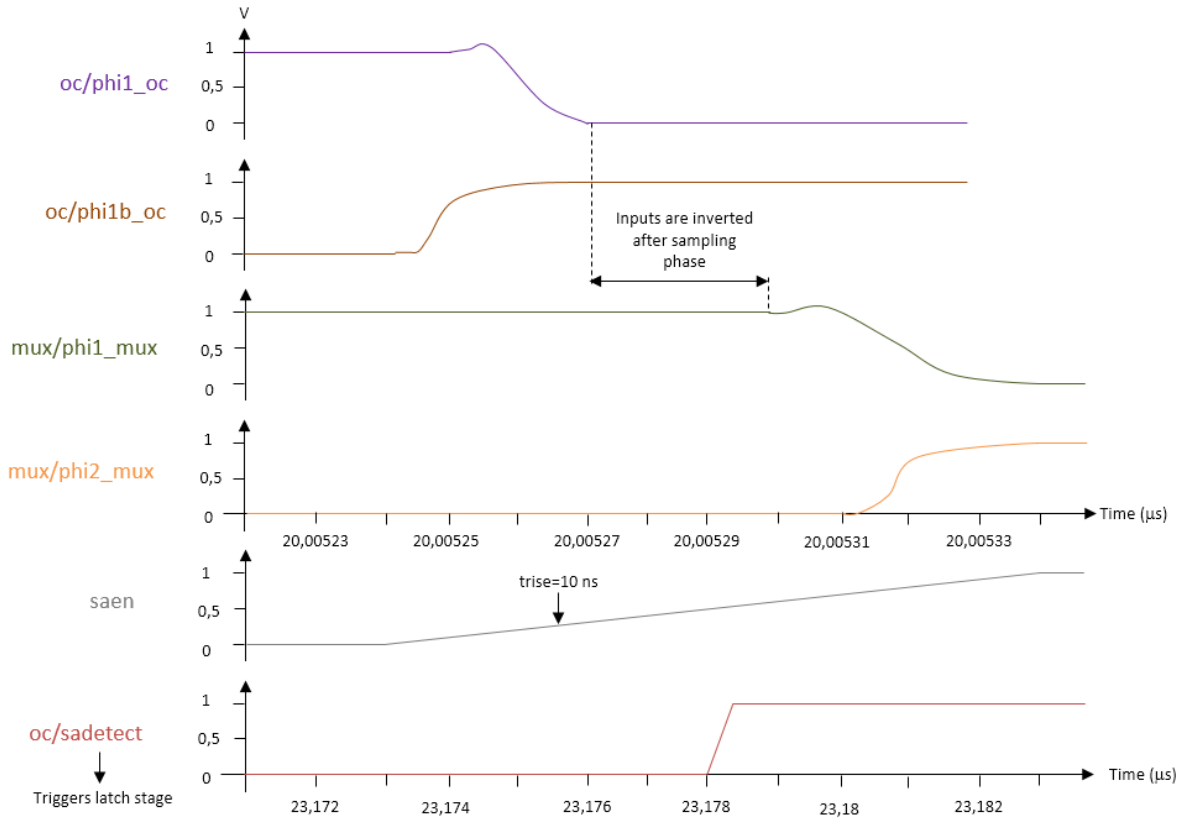


Figure 99 – Pre-layout simulations of pulse generation signals: non-overlapping checking at the beginning of the amplification phase

A particular attention has to be given to 'sadetect' signal. The triggering of the sensing (for latch stage) and the common-mode feedback operation (for offset-cancellation stage) start only after total stabilization of the output signals of the offset-cancellation stage, corresponding to about 3.2 µs after the beginning of the amplification phase in our case study. This leads to a sampling capacitors voltage difference of 17 mV due to switches leakage that are disconnected between the beginning of the amplification phase and the sensing triggering (see *Figure 100* and *Figure 101*).

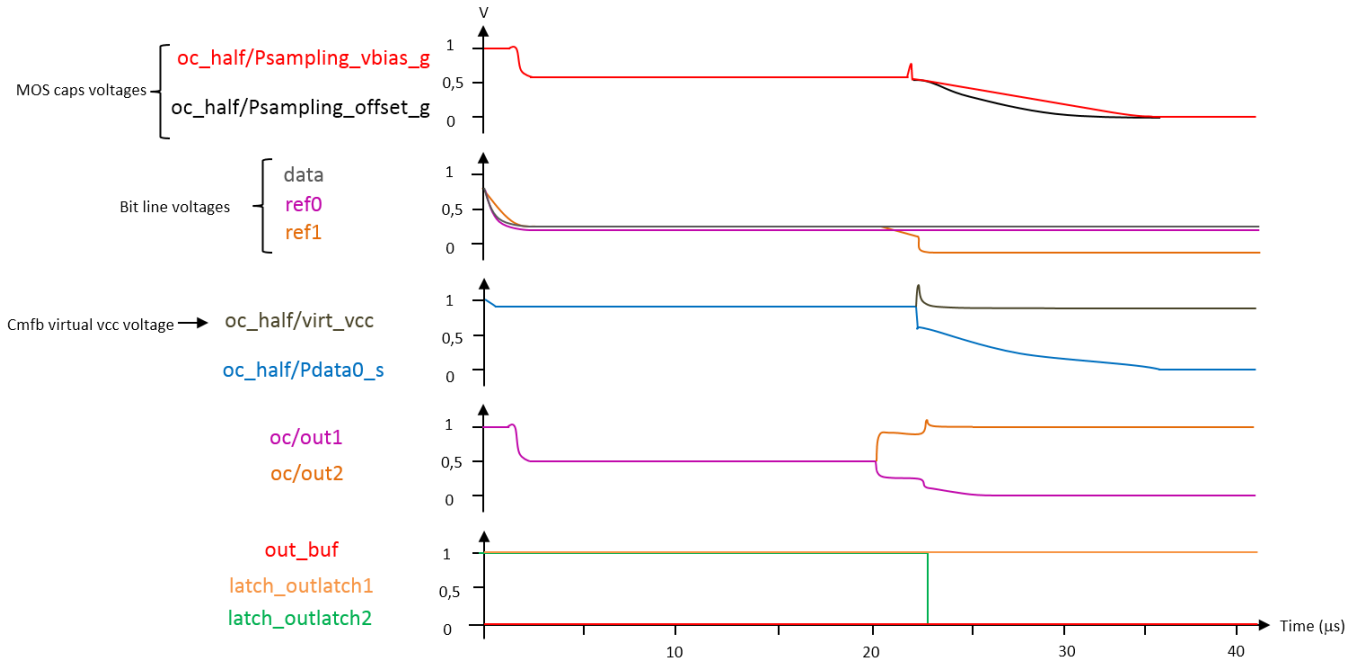


Figure 100 - Pre-layout simulations: description of sampling capacitors, bit line, CMFB and output voltages (with I_{DATA}=20.54 µA, I_{HRS}=20.54 µA, I_{LRS}=33.58 µA, bit line capacitances = 50 pF, load capacitance = 50 pF)

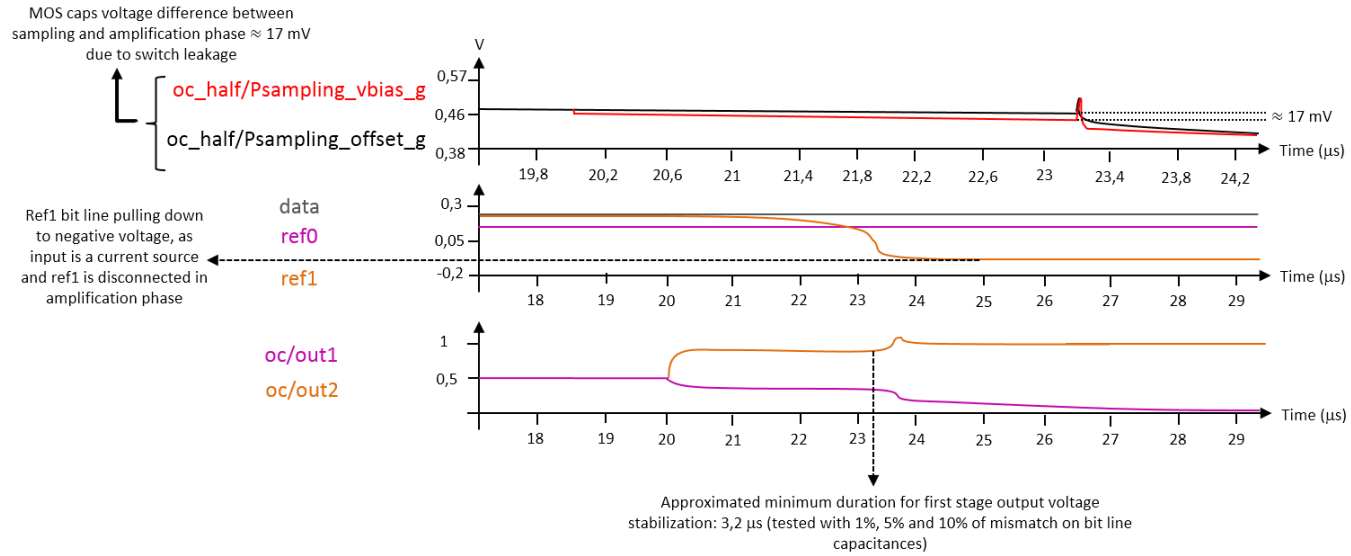


Figure 101 - Pre-layout simulations: illustration of the duration between the beginning of the amplification phase and sensing triggering and its effects (sampling switches leakage and negative bit line voltage; with I_{DATA}=20.54 µA, I_{HRS}=20.54 µA, I_{LRS}=33.58 µA, bit line capacitances = 50 pF, load capacitance = 50 pF)

5.2.3.Sub-block and test row layout

This section describes the different steps for the layout of the proposed sense amplifier test-structure and the test row. This step is fundamental to design variation-tolerant circuits [99][100].
The following pictures show layout screenshots of the three main sub-blocks and the final structure.

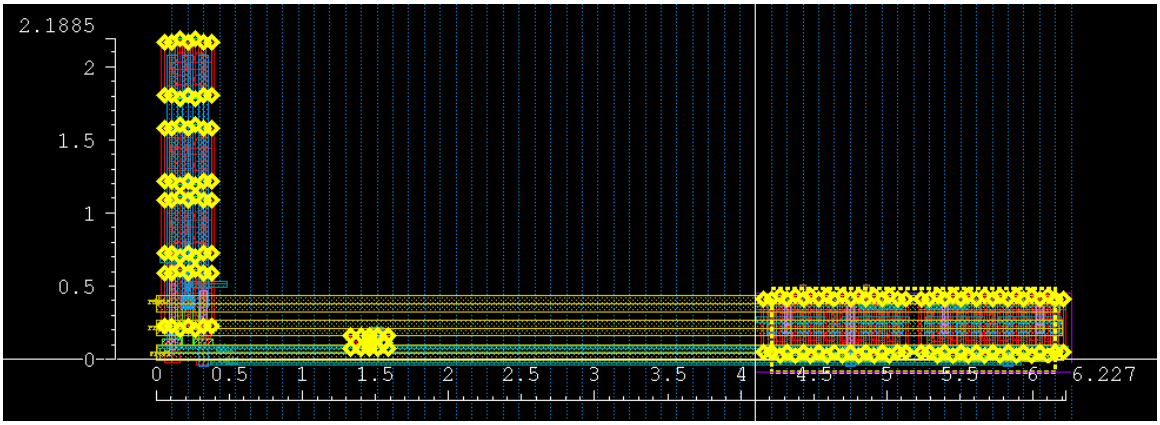


Figure 102 - Reference Time-multiplexing block: layout (dimensions are in μm)

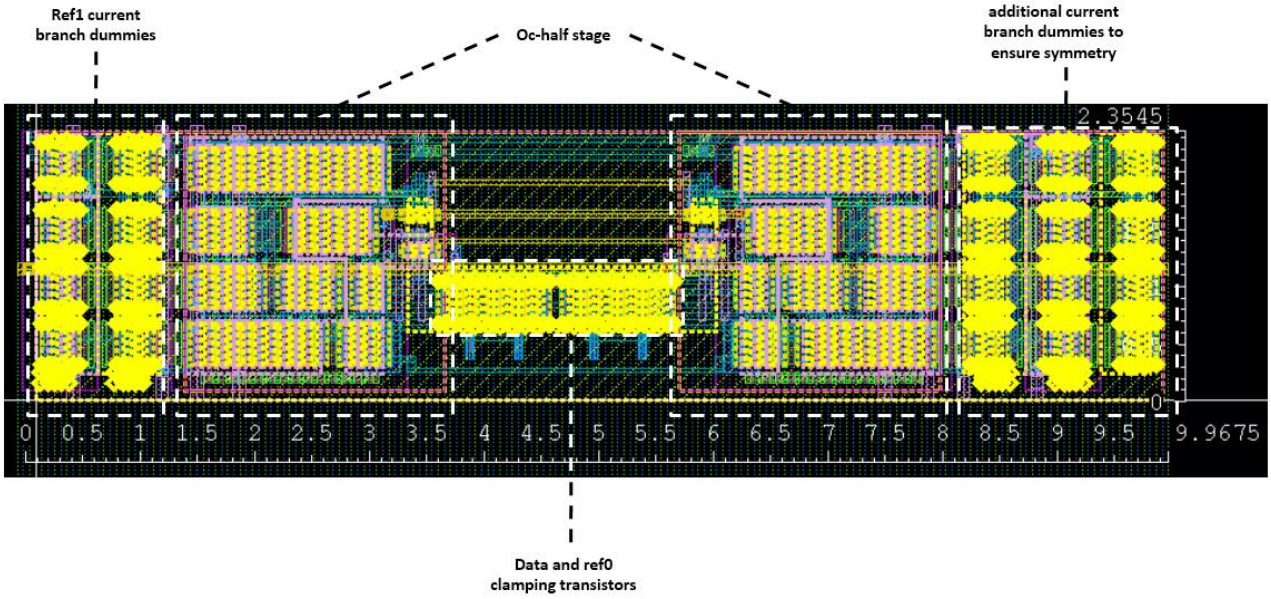


Figure 103 - Offset cancellation stage: layout (dimensions are in μm)

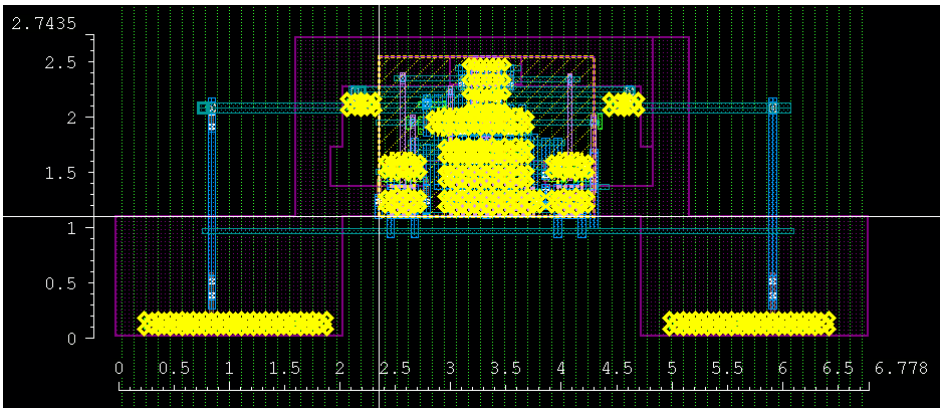


Figure 104 - Latch stage: layout (dimensions are in μm)

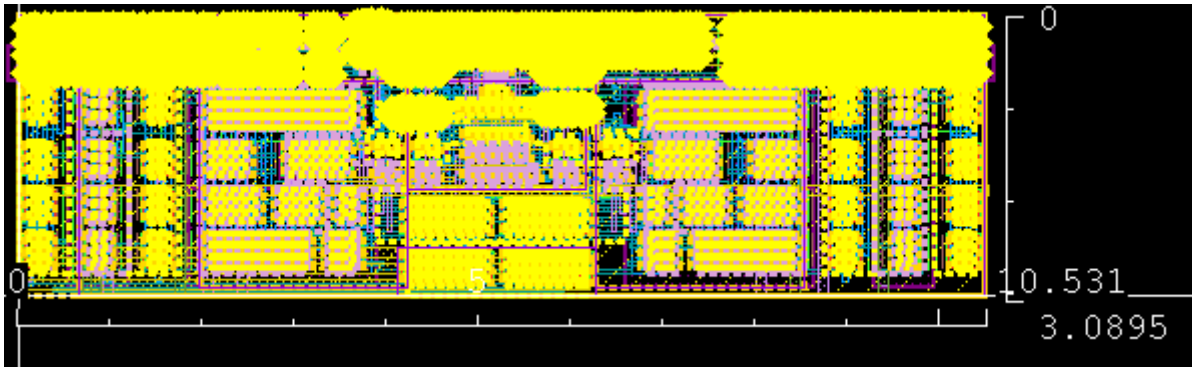


Figure 105 - Proposed sense amplifier: layout (dimensions are in µm)

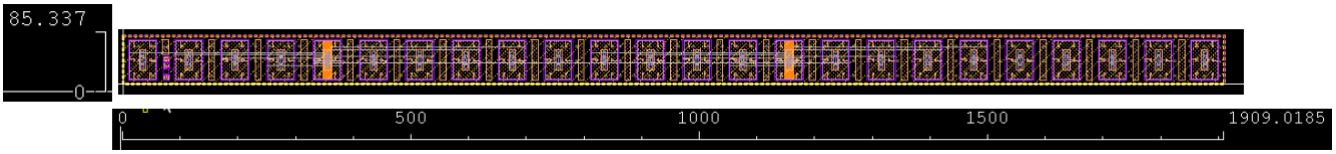


Figure 106 - Proposed sense amplifier test-row: layout (dimensions are in µm)

The main requirements for the layout of the whole structure are:

- Ensure layout regularity of all interconnections.
- Ensure layout symmetry (especially for the offset-cancellation stage and latch stage).
- Shielding of the three inputs are advised as they are highly sensitive to parasitics. In particular, it is important to reduce capacitive coupling between the inputs and crossing 'clk' or 'saen' lines.
- For the routing of the sense amplifier to the pads (level-8 metals), it is important, especially for the inputs, that the equivalent resistances per square are identical. Therefore, as ref1 is closer to the sense amplifier than ref0, the length of the level-7 metal for ref0 routing has to be twice bigger than the one for ref1.

Finally, cleaned (Layout Versus Schematic) and Design Rules Check (DRC) are achieved after the addition of diffusion layers (to fulfill DRC requirements, in yellow in layout pictures above) and the filling of the whole test row.

5.3. Offset results

Both systematic and random offset (as defined in 3.3.1.1 and 3.3.1.3) of this structure is characterized using MPP statistical simulation methodology, as detailed in section 3.3.3.

The pre-layout systematic and random input-referred offset results of the proposed sense amplifier are respectively 2.3Ω and 30Ω per sigma variation (with I_{HRS}=20.54 µA, I_{LRS}=33.58 µA, bit line capacitances = 50 pF, load capacitance = 50 pF). As for post-layout offset results, *Table 10* depicts the systematic and random input offsets after parasitic resistances and/or capacitances extraction. All these very low values demonstrate the interest for the design of the proposed architecture.

Table 10 - Post-layout input-referred offset results of the proposed sense amplifier test structure (with I_{HRS}=20.54 µA, I_{LRS}=33.58 µA, bit line capacitances = 50 pF, load capacitance = 50 pF)

Extraction type	***Systematic offset***	***Random offset***
R only	*12 Ω*	*54 Ω/sigma*
C only	*3 Ω*	*24 Ω/sigma*
RC	*15 Ω*	*60 Ω/sigma*

5.4. Conclusion

A proof of concept of the proposed sense amplifier architecture is achieved. The layout design of the proposed test structure and post-layout offset results confirm the interest of this circuit, in terms of regularity and read reliability.

Table 11 summarizes the advantages and corresponding costs of the proposed sense amplifier for two references that are explained in 0 and compares them to dynamic CBSA and conventional sense amplifier. It highlights the two main features of the proposed scheme: an almost-cancelled offset (about 6.7 of factor gain compared to CBSA) and a layout regular reference scheme with reduced variability. When using more than two references, the unavoidable area, latency and consumption costs of the proposed scheme need to be put into perspective when compared to ECC and repair system-level techniques.

Table 11 – Proposed sense amplifier features compared to CBSA and conventional sense amplifier (e.g. for two references)

	Minimum input random offset (Ohms/sigma)	***Layout regular reference?***	***Sensing Speed***	***Area overhead: # transistors***
Proposed SA	***Pre-layout = 30*** ***Post-layout = 60***	***Yes***	***Medium***	***Minimum 4***
Optimized CBSA	***Pre-layout=200***	***Yes***	***Very fast***	***Minimum 11***
Conventional sense amplifier	***Pre-layout=260***	***No***	***Fast***	***6***

References of Chapter 5

[95] M. Alioto, G. Palumbo, and M. Pennisi, “Understanding the effect of process variations on the delay of static and domino logic,” *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 18, no. 5, pp. 697–710, May 2010.

[96] M. Alioto, G. Palumbo, and M. Pennisi, “Analysis of the impact of process variations on static logic circuits versus fan-in,” in *2008 15th IEEE International Conference on Electronics, Circuits and Systems*, 2008, pp. 137–140.

[97] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, Min Zhao, K. Gala, and R. Panda, “Statistical delay computation considering spatial correlations,” in *Proceedings of the ASP-DAC Asia and South Pacific Design Automation Conference, 2003.*, pp. 271–276.

[98] A. Agarwal, D. Blaauw, and V. Zolotov, “Statistical timing analysis for intra-die process variations with spatial correlations,” in *ICCAD-2003. International Conference on Computer Aided Design (IEEE Cat. No.03CH37486)*, 2003, pp. 900–907.

[99] A. Balasinski, “Layout techniques and rules to reduce process-related variability,” *J. Micro/Nanolithography, MEMS, MOEMS*, vol. 6, no. 3, p. 31009, Jul. 2007.

[100] J. Morris, P. Prabhat, J. Myers, and A. Yakovlev, “Unconventional Layout Techniques for a High Performance, Low Variability Subthreshold Standard Cell Library,” in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2017, pp. 19–24.

Chapter 6: Conclusion, Discussion and prospects

6.1. General Conclusion

This thesis puts forward the importance of variation-tolerance in emerging nanoscale circuits design, in particular resistive memories sense amplifiers. It gives a circuit-level analysis of the variability problematic for the design of read circuits for resistive memories and proposes sense amplifier architecture-solutions that combine both offset-cancellation and reduced reference variability.

Emerging resistive memories and MRAM in particular, are promising technologies in terms of standalone performance compared to existing solutions. STT-MRAM is a well-advanced technology at process-level. One of the biggest challenge of resistive memories is their low design robustness to ensure high read reliability. It is important to propose design-level solutions of sense amplifiers architectures that simultaneously reduce or cancel variability and include an accurate and layout-regular reference scheme. Current literature proposes solutions that handle either layout regularity of reference schemes like the CBSA circuit, either reliable offset cancellation techniques (like FOCT or [101]), but fail to handle the two issues in order to get the best signal to variability ratio of sense amplifiers.

A comprehensive analysis of the variability of the beginning of the sensing path of resistive memories is first performed. A design optimization methodology is proposed to minimize read margin degradation with variability. It is important to position process parameters and their variability regarding read window reduction, as well as non-linear dependency with voltage of the resistive bit cell. The resizing of bit line voltage clamping transistor is automated for a minimum read margin decrease at design-level. An algorithm for offset characterization of a design-optimized dynamic sense amplifier (CBSA) is moreover described. This offset is additionally modelled taking into account bit cell statistics. The high best-case demonstrated offset highlights the necessity of non-dynamic sense amplifier architectures that combines offset-cancellation techniques and layout regular reference scheme, alleviating the impact of load and coupling capacitive mismatch of latch stages.

A sense amplifier architecture is thus proposed, featuring offset-cancellation and time-multiplexed reference scheme. It extends the principle of FOCT to N references. This circuit allows significant signal to offset ratio enhancement by considerably reducing reference variability. Read margins estimation shows an improvement in the BER by a factor of 10 when implementing four references to this architecture compared to two references. The non-negligible timing and above all area costs of this sense amplifier are compared to system-level read yield improvement techniques (especially 2D IO and WL repair, requiring at least 10 spare rows and 14 spare columns to get similar BER gain than the proposed scheme). It shows that the proposed scheme can position as a circuit-level alternative solution, in particular for some timing and power-tolerant applications.

The interest of this circuit is demonstrated by the design of a test structure. Its layout design and post-layout offset results confirm its high features in terms of regularity and read reliability (about 6.7 of factor gain on input offset compared to CBSA).

6.2. Perspectives

This section introduces both design and technologic-level other solutions to better improve read reliability of resistive memories, while not degrading significantly its area or power consumption performances especially.

6.2.1.Design level

6.2.1.1. Mid-reference scheme

The development of a mid-reference scheme ($R_{REF}=(R_{LRS}+R_{HRS})/2$) to include in an offset-cancelled sense amplifier would be an optimum solution to offer high read margin (mainly due to sense amplifier offset cancellation and no more due to reference variability as proposed above) with maintained area cost (since two current branches would be sufficient in this case). However, proposed literature solutions do not manage either to solve layout regularity issues and do not track voltage variations [67][63][102], either ensure suitable reference scheme but fail to include offset-cancellation [64][103][104].

6.2.1.2. Ordered element matching

A convenient way to improve even more read reliability is to implement an 'intelligent' arrangement of the different components of the sensing path of the resistive memory.

Ordered element matching (OEM) technology was firstly proposed in [105] and then rigorously proven in [106]. It was developed and optimized in [107]. *Figure 107* shows the process of this optimized OEM technology. Each rectangle denotes a component (that could be the resistive cell or the transistor) with random mismatch error in the unary-weighted segment, where R_{AVG} is the average resistance amplitude (if the component is a resistor for example). Firstly, all resistors are measured and sorted according to their amplitudes. The next step is to choose the resistor with resistance amplitude closest to R_{AVG}. Then the complementary ordered resistors are paired which is called single "folding". In this way, the original 3-bit unary-coded resistor array is converted into a 2-bit unary-weighted and 1-bit binary-weighted array. In details, the resistor value of each 2-bit unary weighted array is nearly twice of the R_{AVG}, and the random variations in resistors are reduced. The mismatch errors are consistently diminishing after each choosing and folding operation. As shown in *Figure 107*, if the choosing and single folding operations are repeated until the 3-bit unary-weighted array is converted into 3-bit binary-weighted array, the mismatch errors are further reduced. This process is so called "complete-folding".

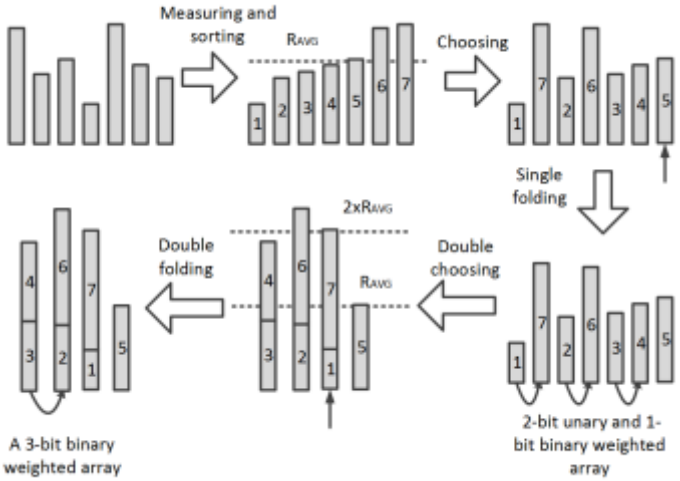


Figure 107 - OEM process [108]

6.2.2.Technologic level

Some developments and extensions of spintronic-based memory technologies can help improve read reliability. As the methodologies detailed in this manuscript are not dependent on the chosen technology, they are suitable with the technologies introduced below.

6.2.2.1. Racetrack memory

To overcome the difficulties of conventional memory technologies (large cell area, high leakage current, technology reliability and scalability issues [109]) and resistive memories presented in 0 (see 2.2.2), researchers have turned to explore alternative solutions.

Domain Wall Memory (DWM, also known as Racetrack Memory) that features high integration density, low operational power consumption, and fast access time [110][111][112][113][114][115] is another resistive technology of particular interest. DWM integrates many memory cells (or domains) along a single ferromagnetic nanowire (or a track). By sharing one or a few access ports among all the memory cells on the same track, DWM can achieve a very high cell integration density. Particularly, a shift operation is required to align the memory cell to be accessed below one read/write port during read/write operations. Prior-arts have demonstrated that DWM can achieve a read/write performance close to that of SRAM [114]. However, frequent shift operations seriously degrade the system performance and introduce significant power consumption. Hence, many architectural techniques have been proposed to overcome the disadvantages of shift operations in DWM designs [110][111][112][113]. The results showed that DWM cache can achieve 25% performance improvement and 62% energy reduction over Spin-Transfer Torque (STT-RAM) caches. Nonetheless, shift operations still consume more than 50% cache energy, becoming the dominant contributor to the overall power and thermal budget. Moreover, as racetrack memory is normally employed as lower-level caches with large capacity, variability has a significant impact on the reliability of DWM caches. To address these issues, [116] presents the use of skyrmions racetrack memory (SKM) as on-chip caches instead of domain walls, and proposes a variation aware data management technique to minimize the performance degradation of SKM cache incurred by variability. Experimental results show 2× density increase and 23% energy reduction compared to DWM under the same area constraint. In addition, the proposed variation-aware data management technique can further improve the system IPC (Instruction Per Cycle) by 25%.

6.2.2.2. Spin-Orbit Torque MRAM

Spin-orbit-torque (SOT) -MRAM is effective because of its fast write operation and good energy efficiency due to lower critical current and voltage compared to STT-MRAM. Additionally, read and write current paths in SOT-MRAM are decoupled to allow separate optimization for read and write [117].

The memory cell is a three-terminal device, where the free layer (FL) of the MTJ is in contact with a nonmagnetic heavy metal (HM) with a strong spin-orbit interaction. In such FL/HM systems, when current is injected through HM, the magnetization of the FL is reversed due to the injection of a pure spin current into the FL as a result of the spin-hall effect. During read operations, current is injected into HM, and the current in HM results in an antidumping torque on the FL owing to SOT. The read reliability is disturbed by the read current. Therefore, the read reliability in SOT-MRAM is lower than that in STT-MRAM.

SOT-MRAM can thus operate with a low power consumption while maintaining read reliability. However, we clarified that read and write operations are not completely decoupled because a part of the inversion power created by SOT occurs during the read operation.

[118] introduces a model of read reliability of SOT-MRAM, showing that the effect of both STT and SOT on read disturbance has to be taken into account, instead of considering only STT for read and SOT for write disturbance as conventionally. This paper also proposes a new read current path of the SOT-MRAM cell to remove any read disturbance caused by p-SOT: the path in HM now flows in bilateral directions [119], leading to a 10-times higher critical current due to SOT. This way, the magnetic anisotropy of the FL is broken by this high critical current, and the read disturbance due to SOT is removed. Therefore, SOT-MRAM can operate with lower writing power consumption than STT-MRAM while maintaining read reliability.

References of Chapter 6

[101] B. Song, T. Na, J. Kim, J. P. Kim, S. H. Kang, and S.-O. Jung, “Latch Offset Cancellation Sense Amplifier for Deep Submicrometer STT-RAM,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 62, no. 7, pp. 1776–1784, Jul. 2015.

[102] W. Kang, T. Pang, Y. Zhang, D. Ravelosona, and W. Zhao, “Dynamic Reference Sensing Scheme for Deeply Scaled STT-MRAM,” in *2015 IEEE International Memory Workshop (IMW)*, 2015, pp. 1–4.

[103] W. Kang, T. Pang, W. Lv, and W. Zhao, “Dynamic Dual-Reference Sensing Scheme for Deep Submicrometer STT-MRAM,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 64, no. 1, pp. 122–132, Jan. 2017.

[104] H. Zhang, W. Kang, and T. Pang, “Dual Reference Sensing Scheme with Triple Steady States for Deeply Scaled STT-MRAM,” *Nanoscale Archit. (NANOARCH), 2016 IEEE/ACM Int. Symp. Nanoscale Archit.*, pp. 1–6, 2016.

[105] T. Zeng and D. Chen, “New calibration technique for current-steering DACs,” in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 573–576.

[106] T. Zeng and D. Chen, “An Order-Statistics Based Matching Strategy for Circuit Components in Data Converters,” *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 60, no. 1, pp. 11–24, Jan. 2013.

[107] H. Van de Vel, J. Briaire, C. Bastiaansen, P. van Beek, G. Geelen, H. Gunnink, Y. Jin, M. Kaba, K. Luo, E. Paulus, B. Pham, W. Relyveld, and P. Zijlstra, “11.7 A 240mW 16b 3.2GS/s DAC in 65nm CMOS with <-80dBc IM3 up to 600MHz,” in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 206–207.

[108] Y. Li and D. Chen, “A novel 20-bit R-2R DAC structure based on ordered element matching,” in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1030–1033.

[109] Y. G. Choi, S. Yoo, S. Lee, and J. H. Ahn, “Matching cache access behavior and bit error pattern for high performance low Vcc L1 cache,” in *Proceedings of the 48th Design Automation Conference on - DAC ’11*, 2011, p. 978.

[110] Z. Sun, W. Wu, and H. (Helen) Li, “Cross-layer racetrack memory design for ultra high density and low power consumption,” in *Proceedings of the 50th Annual Design Automation Conference on - DAC ’13*, 2013, p. 1.

[111] R. Venkatesan, V. Kozhikkottu, C. Augustine, A. Raychowdhury, K. Roy, and A. Raghunathan, “TapeCache: A High Density, Energy Efficient Cache Based on Domain Wall Memory,” in *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design - ISLPED ’12*, 2012, p. 185.

[112] R. Venkatesan, S. G. Ramasubramanian, S. Venkataramani, K. Roy, and A. Raghunathan, “STAG: Spintronic-Tape Architecture for GPGPU cache hierarchies,” in *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, 2014, pp. 253–264.

[113] M. Mao, W. Wen, Y. Zhang, Y. Chen, and H. Li, “Exploration of GPGPU register file architecture using domain-wall-shift-write based racetrack memory,” in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2014, pp. 1–6.

[114] R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, “DWM-TAPESTRI - An Energy Efficient All-Spin Cache Using Domain Wall Shift Based Writes,” in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2013*, 2013, pp. 1825–1830.

[115] Shuo Wang, Yun Liang, Chao Zhang, Xiaolong Xie, Guangyu Sun, Yongpan Liu, Yu Wang, and Xiuhong Li, “Performance-centric register file design for GPUs using racetrack memory,” in *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2016, pp. 25–30.

[116] F. Chen, Z. Li, W. Kang, W. Zhao, H. Li, and Y. Chen, “Process variation aware data management for magnetic skyrmions racetrack memory,” in *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2018, pp. 221–226.

[117] Y. Kim, X. Fong, K.-W. Kwon, M.-C. Chen, and K. Roy, “Multilevel Spin-Orbit Torque MRAMs,” *IEEE Trans. Electron Devices*, vol. 62, no. 2, pp. 561–568, Feb. 2015.

[118] H. Kazama and T. Kawahara, “Spin-orbit torque MRAM read reliability,” in *2017 IEEE International Magnetics Conference (INTERMAG)*, 2017, pp. 1–2.

[119] S. Manipatruni, D. E. Nikonov, and I. A. Young, “Energy-delay performance of giant spin Hall effect switching for dense magnetic memory,” *Appl. Phys. Express*, vol. 7, no. 10, p. 103001, Oct. 2014.

Appendix A: Excel Visual-basic code of the statistical read margin model of section 0

```
Private Sub CommandButton22_Click()

Dim i As Double
Dim j As Double
Dim k As Double

'clear all contents

ThisWorkbook.Sheets("Sheet3").Range("B4:R63").ClearContents
ThisWorkbook.Sheets("Sheet3").Range("B69:IP128").ClearContents
ThisWorkbook.Sheets("Sheet3").Range("B132:H191").ClearContents
ThisWorkbook.Sheets("Sheet3").Range("AB6:AC22").ClearContents
ThisWorkbook.Sheets("MAX_RM").Range("M28:N28").ClearContents
ThisWorkbook.Sheets("MAX_RM").Range("F193:F194").ClearContents
ThisWorkbook.Sheets("MAX_RM").Range("E197:F197").ClearContents

For Each cht In ThisWorkbook.Sheets("Sheet3").ChartObjects
    cht.Delete
Next
Application.ScreenUpdating = False

'create gm(W/L) chart for different R_PAR values

For i = 1 To 60
      ThisWorkbook.Sheets("MAX_RM").Cells(25, 7) = i
  For k = 1 To 17
      ThisWorkbook.Sheets("MAX_RM").Cells(k + 5, 14) = ThisWorkbook.Sheets("Sheet1").Cells(i + 27, 3)
  Next k
  For j = 0 To 16
      ThisWorkbook.Sheets("Sheet3").Cells(i + 3, j + 2) = ThisWorkbook.Sheets("MAX_RM").Cells(j + 6, 18)
  Next j
Next i

Set
cht1=ThisWorkbook.Sheets("Sheet3").ChartObjects.Add(Left:=1150,Width:=300,Top:=10, Height:=300)
With cht1
```

```
For i = 1 To 17
      .Chart.SeriesCollection.NewSeries
      .Chart.SeriesCollection(i).Name = "=Sheet3!$AI" & i
      .Chart.SeriesCollection(i).XValues = ThisWorkbook.Sheets("Sheet3").Range("A4:A63")
          .Chart.SeriesCollection(i).Values=
          ThisWorkbook.Sheets("Sheet3").Range(ThisWorkbook.Sheets("Sheet3").Cells(4,i+1),ThisWorkboo
          k.Sheets("Sheet3").Cells(63, i + 1))
Next i

.Chart.ChartType = Excel.XlChartType.xlXYScatterLines
.Chart.Axes(xlCategory).HasTitle = True
.Chart.Axes(xlCategory).AxisTitle.Text = "W/L"
.Chart.Axes(xlCategory).HasMajorGridlines = True

.Chart.Axes(xlValue).HasTitle = True
.Chart.Axes(xlValue).AxisTitle.Text = "gm_bit0 (S)"
.Chart.Axes(xlValue).TickLabels.NumberFormat = "0.0E+0"

.Chart.Shapes.AddTextbox(msoTextOrientationHorizontal,400,0,75,21).TextFrame.Characters.Text="RPAR
(Ohms)"

End With
```

'Adaptive Vbl : set correct Vbl per W/L variation

```
For i = 1 To 60
  For k = 1 To 17
    ThisWorkbook.Sheets("MAX_RM").Cells(k + 5, 14) = ThisWorkbook.Sheets("Sheet1").Cells(i + 27, 3)
  Next k
  For j = 1 To 249
     ThisWorkbook.Sheets("MAX_RM").Cells(25, 7) = ThisWorkbook.Sheets("Sheet3").Cells(i + 68, 1)
     ThisWorkbook.Sheets("MAX_RM").Cells(29, 6) = ThisWorkbook.Sheets("Sheet3").Cells(66, j + 1)
     ThisWorkbook.Sheets("Sheet3").Cells(i + 68, j + 1) = ThisWorkbook.Sheets("MAX_RM").Cells(40, 22)
   Next j
Next i
```

' create RM0(σV_T) chart for different W/L values

```
Set
cht2 = ThisWorkbook.Sheets("Sheet3").ChartObjects.Add(Left:=1150, Width:=300, Top:=310, Height:=300)
With cht2

For i = 1 To 60
          .Chart.SeriesCollection.NewSeries
          .Chart.SeriesCollection(i).Name = "=Sheet3!$K" & i + 132
          .Chart.SeriesCollection(i).XValues = ThisWorkbook.Sheets("Sheet3").Range("L133:L382")
          .Chart.SeriesCollection(i).Values=ThisWorkbook.Sheets("Sheet3").Range(ThisWorkbook.Sheets("S
          heet3").Cells(i+68,2),ThisWorkbook.Sheets("Sheet3").Cells(i + 68, 250))
```

```
Next i

.Chart.ChartType = Excel.XlChartType.xlXYScatterLines
.Chart.Axes(xlCategory).HasTitle = True
.Chart.Axes(xlCategory).AxisTitle.Text = "sigma_Vt (V)"
.Chart.Axes(xlCategory).HasMajorGridlines = True
.Chart.Axes(xlCategory).TickLabelPosition = xlTickLabelPositionLow
.Chart.Axes(xlValue).HasTitle = True
.Chart.Axes(xlValue).AxisTitle.Text = "Read Margin 0 (A)"
.Chart.Axes(xlValue).TickLabels.NumberFormat = "0.0E+0"

.Chart.Shapes.AddTextbox(msoTextOrientationHorizontal,400,0,40,21).TextFrame.Characters.Text = "W/L"

End With
```

'Find all the pairs {σV_T, W/L} giving 10% tolerance on RM0

```
For i = 1 To 60
  For k = 1 To 17
    ThisWorkbook.Sheets("MAX_RM").Cells(k + 5, 14) = ThisWorkbook.Sheets("Sheet1").Cells(i + 27, 3)
    ThisWorkbook.Sheets("MAX_RMnoclamp").Cells(k+8, 4)=ThisWorkbook.Sheets("Sheet1").Cells(3,22)
  Next k

ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 2) = ThisWorkbook.Sheets("MAX_RM noclamp").Cells(17,
20) - 0.1 * ThisWorkbook.Sheets("MAX_RM noclamp").Cells(17, 20)
ThisWorkbook.Sheets("MAX_RM").Cells(25, 7) = ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 1)

ThisWorkbook.Sheets("MAX_RM").Cells(25, 7) = i
ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 8) = ThisWorkbook.Sheets("MAX_RM").Cells(25, 9)

If
IsError(Application.Match(ThisWorkbook.Sheets("Sheet3").Cells(i+131,2).Value,ThisWorkbook.Sheets("Sh
eet3").Range(ThisWorkbook.Sheets("Sheet3").Cells(i+68,2),      ThisWorkbook.Sheets("Sheet3").Cells(i+68,
250)), -1))
Then
  Var = "not found"
 Else

Var=Application.Match(ThisWorkbook.Sheets("Sheet3").Cells(i+131,2).Value,ThisWorkbook.Sheets("Shee
t3").Range(ThisWorkbook.Sheets("Sheet3").Cells(i + 68, 2), ThisWorkbook.Sheets("Sheet3").Cells(i + 68,
250)), -1) + 1
ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 4) = Var

If ThisWorkbook.Sheets("Sheet3").Cells(66, Var) = 0
Then
  ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 3) = ""
Else
  ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 3) = ThisWorkbook.Sheets("Sheet3").Cells(66, Var)
```

```
End If

ThisWorkbook.Sheets("MAX_RM").Cells(29, 6) = ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 3)

If
IsError(ThisWorkbook.Sheets("MAX_RM").Cells(25, 8))
Then
  ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 5) = ""
Else
  ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 5) = ThisWorkbook.Sheets("MAX_RM").Cells(25, 8)
End If
```

End If

```
Next i

For i = 15 To 60
        ThisWorkbook.Sheets("Sheet3").Cells(i + 131, 6) = (Abs(ThisWorkbook.Sheets("Sheet3").Cells(i +
        1 + 131, 5).Value - ThisWorkbook.Sheets("Sheet3").Cells(i - 1 + 131, 5).Value)) / 2
        ThisWorkbook.Sheets("Sheet3").Cells(i+131,7)=ThisWorkbook.Sheets("Sheet3").Cells(i+131,5)-
        ThisWorkbook.Sheets("Sheet3").Cells(i+131,8)
Next i
```

' create σV_T(W/L) chart giving 10% tolerance on RM0

```
Set
cht3 = ThisWorkbook.Sheets("Sheet3").ChartObjects.Add(Left:=1150, Width:=300, Top:=610, Height:=300)
With cht3

  .Chart.SeriesCollection.NewSeries
  .Chart.SeriesCollection(1).Name = "sigma_Vt (V)"
  .Chart.SeriesCollection(1).XValues = ThisWorkbook.Sheets("Sheet3").Range("A132:A191")
  .Chart.SeriesCollection(1).Values = ThisWorkbook.Sheets("Sheet3").Range("C132:C191")

  .Chart.ChartType = Excel.XlChartType.xlXYScatterLines
  .Chart.Axes(xlCategory).HasTitle = True
  .Chart.Axes(xlCategory).AxisTitle.Text = "W/L"
  .Chart.Axes(xlCategory).HasMajorGridlines = True
  .Chart.Axes(xlCategory).TickLabelPosition = xlTickLabelPositionLow

  .Chart.Axes(xlValue).HasTitle = True
  .Chart.Axes(xlValue).AxisTitle.Text = "sigma_Vt (V) @10%RM0_max"
  .Chart.Axes(xlValue).TickLabels.NumberFormat = "0.0E+0"
  .Chart.Axes(xlValue).MinimumScale =
  Application.Min(ThisWorkbook.Sheets("Sheet3").Range("C132:C91"))
  .Chart.Axes(xlValue).MaximumScale =
  Application.Max(ThisWorkbook.Sheets("Sheet3").Range("C132:C191"))
```

```
End With

' create area contribution charts of σV_T and W/L

Set
cht4 = ThisWorkbook.Sheets("Sheet3").ChartObjects.Add(Left:=1450, Width:=300, Top:=610, Height:=300)
With cht4

  .Chart.SeriesCollection.NewSeries
  .Chart.SeriesCollection(1).Name = "W*L (um^2)=(A(Vt)/sigmaVt)^2"
  .Chart.SeriesCollection(1).XValues = ThisWorkbook.Sheets("Sheet3").Range("A132:A191")
  .Chart.SeriesCollection(1).Values = ThisWorkbook.Sheets("Sheet3").Range("E132:E191")

  .Chart.SeriesCollection.NewSeries
  .Chart.SeriesCollection(2).Name = "W*L (um^2)=(W/L)*L^2"
  .Chart.SeriesCollection(2).XValues = ThisWorkbook.Sheets("Sheet3").Range("A132:A191")
  .Chart.SeriesCollection(2).Values = ThisWorkbook.Sheets("Sheet3").Range("H132:H191")

  .Chart.ChartType = Excel.XlChartType.xlXYScatterLines
  .Chart.Axes(xlCategory).HasTitle = True
  .Chart.Axes(xlCategory).AxisTitle.Text = "W/L"
  .Chart.Axes(xlCategory).HasMajorGridlines = True
  .Chart.Axes(xlCategory).TickLabelPosition = xlTickLabelPositionLow

  .Chart.Axes(xlValue).HasTitle = True
  .Chart.Axes(xlValue).AxisTitle.Text = "W*L (um^2)"
  .Chart.Axes(xlValue).TickLabels.NumberFormat = "0.0E+0"

End With

'Extract the best {σVt,W/L} giving the minimum area contribution for 10% tolerance on RM0

If
IsError(Application.Match(0.001, ThisWorkbook.Sheets("Sheet3").Range("F148:F191"), -1))
Then
  Var = "not found"
Else
  Var = Application.Match(0.001, ThisWorkbook.Sheets("Sheet3").Range("F148:F191"), -1)
  ThisWorkbook.Sheets("Sheet3").Cells(193, 6) = Var
End If

If
IsError(Application.Match(0, ThisWorkbook.Sheets("Sheet3").Range("G148:G191"), -1))
Then
  Var = "not found"
Else
  Var2 = Application.Match(0, ThisWorkbook.Sheets("Sheet3").Range("G148:G191"), -1)
  ThisWorkbook.Sheets("Sheet3").Cells(194, 6) = Var2
```

```
End If

If
IsError(Var < Var2)
Then
  ThisWorkbook.Sheets("Sheet3").Cells(197, 5) = "not found"
  ThisWorkbook.Sheets("Sheet3").Cells(197, 6) = "not found"
Else
If Var < Var2 Then
  ThisWorkbook.Sheets("Sheet3").Cells(197, 5) = ThisWorkbook.Sheets("Sheet3").Cells(Var + 1 + 147, 1)
  ThisWorkbook.Sheets("Sheet3").Cells(6, 28) = ThisWorkbook.Sheets("Sheet3").Cells(197, 5)

  ThisWorkbook.Sheets("Sheet3").Cells(197, 6) = ThisWorkbook.Sheets("Sheet3").Cells(Var + 1 + 147, 3)
  ThisWorkbook.Sheets("Sheet3").Cells(6, 29) = ThisWorkbook.Sheets("Sheet3").Cells(197, 6)
Else
  ThisWorkbook.Sheets("Sheet3").Cells(197, 5) = ThisWorkbook.Sheets("Sheet3").Cells(Var2 + 1 + 147, 1)
  ThisWorkbook.Sheets("Sheet3").Cells(6, 28) = ThisWorkbook.Sheets("Sheet3").Cells(197, 5)

  ThisWorkbook.Sheets("Sheet3").Cells(197, 6) = ThisWorkbook.Sheets("Sheet3").Cells(Var2 + 1 + 147, 3)
  ThisWorkbook.Sheets("Sheet3").Cells(6, 29) = ThisWorkbook.Sheets("Sheet3").Cells(197, 6)
End If

End If

'Print RM for best {σVt,W/L} pair

For k = 1 To 17
  ThisWorkbook.Sheets("MAX_RM").Cells(k+5,14) =
  ThisWorkbook.Sheets("Sheet1").Cells(ThisWorkbook.Sheets("Sheet3").Cells(6, 28) + 27, 3)
  ThisWorkbook.Sheets("MAX_RM noclamp").Cells(k + 8, 4) = ThisWorkbook.Sheets("Sheet1").Cells(3, 22)
Next k

ThisWorkbook.Sheets("MAX_RM").Cells(25, 7) = ThisWorkbook.Sheets("Sheet3").Cells(6, 28)
ThisWorkbook.Sheets("MAX_RM").Cells(29, 6) = ThisWorkbook.Sheets("Sheet3").Cells(6, 29)

For i = 1 To 17
  ThisWorkbook.Sheets("Sheet3").Cells(i + 5, 25) = ThisWorkbook.Sheets("MAX_RM").Cells(i + 31, 22)
  ThisWorkbook.Sheets("Sheet3").Cells(i + 5, 26) = ThisWorkbook.Sheets("MAX_RM").Cells(i + 31, 23)
Next i

ThisWorkbook.Sheets("Sheet3").Cells(57, 37) = ThisWorkbook.Sheets("Sheet3").Cells(6, 28)
ThisWorkbook.Sheets("Sheet3").Cells(57, 38) = ThisWorkbook.Sheets("Sheet3").Cells(6, 29)

ThisWorkbook.Sheets("Sheet3").Cells(57, 40) = ThisWorkbook.Sheets("Sheet3").Cells(10, 25)
ThisWorkbook.Sheets("Sheet3").Cells(57, 41) = ThisWorkbook.Sheets("Sheet3").Cells(10, 26)

'compare with RM at fixed Vbl (i.e. σV_BL=0)
```

```
For i = 1 To 17
  ThisWorkbook.Sheets("Sheet3").Cells(i+5,30)=ThisWorkbook.Sheets("MAX_RMnoclamp").Cells(i+8, 20)
  ThisWorkbook.Sheets("Sheet3").Cells(i+5,31)=ThisWorkbook.Sheets("MAX_RMnoclamp").Cells(i+8, 21)
Next i

Set
cht5 = ThisWorkbook.Sheets("Sheet3").ChartObjects.Add(Left:=1750, Width:=300, Top:=610, Height:=300)
With cht5

  .Chart.SeriesCollection.NewSeries
  .Chart.SeriesCollection(1).Name = "RM0 @varVbl"
  .Chart.SeriesCollection(1).XValues = ThisWorkbook.Sheets("Sheet3").Range("AA6:AA22")
  .Chart.SeriesCollection(1).Values = ThisWorkbook.Sheets("Sheet3").Range("Y6:Y22")
  .Chart.SeriesCollection.NewSeries
  .Chart.SeriesCollection(2).Name = "RM0 @fixedVbl"
  .Chart.SeriesCollection(2).XValues = ThisWorkbook.Sheets("Sheet3").Range("AA6:AA22")
  .Chart.SeriesCollection(2).Values = ThisWorkbook.Sheets("Sheet3").Range("AD6:AD22")

  .Chart.SeriesCollection.NewSeries
  .Chart.SeriesCollection(3).Name = "RM1 @varVbl"
  .Chart.SeriesCollection(3).XValues = ThisWorkbook.Sheets("Sheet3").Range("AA6:AA22")
  .Chart.SeriesCollection(3).Values = ThisWorkbook.Sheets("Sheet3").Range("Z6:Z22")

  .Chart.SeriesCollection.NewSeries
  .Chart.SeriesCollection(4).Name = "RM1 @fixedVbl"
  .Chart.SeriesCollection(4).XValues = ThisWorkbook.Sheets("Sheet3").Range("AA6:AA22")
  .Chart.SeriesCollection(4).Values = ThisWorkbook.Sheets("Sheet3").Range("AE6:AE22")

  .Chart.ChartType = Excel.XlChartType.xlXYScatterLines
  .Chart.Axes(xlCategory).HasTitle = True
  .Chart.Axes(xlCategory).AxisTitle.Text = "RPAR (Ohms)"
  .Chart.Axes(xlCategory).HasMajorGridlines = True
  .Chart.Axes(xlCategory).TickLabelPosition = xlTickLabelPositionLow

  .Chart.Axes(xlValue).HasTitle = True
  .Chart.Axes(xlValue).AxisTitle.Text = "Read Margins (A)"
  .Chart.Axes(xlValue).TickLabels.NumberFormat = "0.0E+0"

End With

Application.ScreenUpdating = True

End Sub
```

Appendix B: MPP sense amplifier offset characterization: Cadence Skrill script (see Section 3.3.3)

```
;;*****************************************************************************************
;;
;;
;;                          THIS IS THE BISECTION (BASED ON DICHOTOMY) SCRIPT
;;                          TO FIND INPUT OFFSET VALUE OF A CIRCUIT
;;
;;
;;*****************************************************************************************

; Define the directories of the sense amplifier view for statistical simulation
libname   = "work_salmen"
cellname  = "cbsa"
viewname  = "adexl"

;========================== Set to XL mode ================ ====================
flag =  nil
ocnSetXLMode()
ocnxlProjectDir( "~/simulation" )
flag = ocnxlTargetCellView( libname cellname viewname ?mode "r")
if(( flag==nil) then printf("ERROR_OCEAN: Could not open the target view %s %s %s \n" libname cellname
viewname))

ocnxlResultsLocation( "/share/tmp2/mem1273-dp09-vob/units/hdc_spsram/simulation/boujamaa" )
ocnxlSimResultsLocation( "" )

;====================Set/load data base for the statistical simulation tool ==================

sessionname = ocnxlGetSession()
sdb = axlSetMainSetupDBLCV( sessionname libname cellname viewname ?mode "r")
nvInitNovaFromDB( sessionname )

when( axlIsValidAXLSession(sessionname)
  (stringToSymbol sessionname)->ignoreDesignChangesDuringRunUIDoNotDisplay = t)

;========================Setup for simulation runs ================================
ocnxlJobSetup( '(
                          "blockemail" "1"
```

```
                    "configuretimeout" "300"
                    "distributionmethod" "Command"
                    "jobsubmitcommand" "iwADEXL"
                    "lingertimeout" "300"
                    "maxjobs" "50"
                    "name" "ifxCommandMode"
                    "preemptivestart" "1"
                    "reconfigureimmediately" "1"
                    "runtimeout" "-1"
                    "showerrorwhenretrying" "1"
                    "showoutputlogerror" "0"
                    "startmaxjobsimmed" "1"
                    "starttimeout" "300"
) )

;============================ Start initial setup ================================
;; open the log file port for writing and start the offset search

LogBisection=outfile("/vobs/envm-dp093
vob/units/mram/simulation/spectre/work_salmen/cbsa/ocean/logCBSAOFF.txt" "w")
fprintf(LogBisection "##############################################################\n" )
fprintf(LogBisection "# Binary Search OCEAN script LOG file \n" )
fprintf(LogBisection "# Beta Version \n" )
fprintf(LogBisection "# Owner: Salmen MRAIHI " )
fprintf(LogBisection "##############################################################\n" )
fprintf(LogBisection "\n Dichotomy Search Start =======> " )
fprintf(LogBisection getCurrentTime() )
drain(LogBisection)

;; Setup for dichotomy search algorithm

MAX = 20000             ;; maximum R_DATA value (Ohms)
MIN = 10                ;; minimum R_DATA value (Ohms)
MID = (MAX+MIN)/2
accuracy = 10           ;; accuracy of the input referred offset (Ohms)
nbrIter=0               ;; number of iterations

;============================ End initial setup ================================

;======================== Start binary search algorithm ========================

;;  Flags are used to detect an error in the script
;range_flag=0
;flagRun_MIN=0
;flagRun_MAX=0
;flagRun_MID=0

;;set MIN value
```

```
axlSetVarValue(axlGetVar( sdb "Rdata" ) sprintf(nil "%L" MIN))
fprintf(LogBisection "\n\n_____ initial min = %L (s)" axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
fprintf(LogBisection getCurrentTime() )
printf("\n\n_____ initial min = %L (s)\n" axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
drain(LogBisection)
```

;; launch statistical simulation tool and print simulation result into log file

```
flagBi = ocnxlRun()  ;;Run the simulation, the function returns nil if an error occurs
if(( flagBi==nil) then
   printf("#####  ERROR_OCEAN: Simulation run failed ###### \n" )
   fprintf(LogBisection "\n#####  ERROR_OCEAN: Simulation run        failed ###### \n" )
   drain(LogBisection)
)

flagRun_MIN=
infile("/tmp/smraihi/adenova/work_salmen/cbsa/adexl/work_salmen:cbsa:1/nom/reports/result.mpp.report")
```

;; Check if a simulation file is generated for MIN value, meaning an MPP successful run

```
if((flagRun_MIN != nil) then
   printf("Output summary SUCCESSFULLY generated for MIN = %L  \n" axlGetVarValue( axlGetVar( sdb
"Rdata" ) ))
   fprintf(LogBisection  "Output  summary  SUCCESSFULLY  generated  for  MIN  =  %L      \n"
axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
   fprintf(LogBisection "_____ MPP report :  \n" )
   while( (gets(s flagRun_MIN) != nil) fprintf(LogBisection  s) )
   drain(LogBisection)
else
   fprintf(LogBisection "\n#####  ERROR_MIN: Cannot evaluate if the target value is between MIN and
MAX ###### \n")   ;; if MPP failed, input offset is not between MIN and MAX)
)

fprintf(LogBisection "End time:")
fprintf(LogBisection getCurrentTime() )
fprintf(LogBisection "\n")
fprintf(LogBisection "End initial min\n")
fprintf(LogBisection "**********\n")
drain(LogBisection)
```

;;This block is the same as the MIN one, except it runs the simulation for the MAX value

;;set MAX value

```
axlSetVarValue(axlGetVar( sdb "Rdata" ) sprintf(nil "%L" MAX))
fprintf(LogBisection "\n\n_____ initial max = %L (s)" axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
fprintf(LogBisection getCurrentTime() )
printf("\n\n_____ initial max = %L (s)\n" axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
```

```
drain(LogBisection)

;; launch statistical simulation tool and print simulation result into log file

flagBi = ocnxlRun()
if(( flagBi==nil) then
  printf("#####  ERROR_OCEAN: Simulation run failed ###### \n" )
  fprintf(LogBisection "\n#####  ERROR_OCEAN: Simulation run          failed ###### \n" )
  drain(LogBisection)
)

flagRun_MAX=
infile("/tmp/smraihi/adenova/work_salmen/cbsa/adexl/work_salmen:cbsa:1/nom/reports/result.mpp.report")
 ;;  Check if a simulation file is generated for MAX value, meaning an MPP successful run

if((flagRun_MAX !=nil) then
   printf("_____________Output summary SUCCESSFULLY generated for MAX = %L      \n"
axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
   fprintf(LogBisection "_____________ Output summary SUCCESSFULLY generated for MIN = %L  \n"
axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
   fprintf(LogBisection "_____________ MPP report :  \n" )
   while( (gets(s flagRun_MAX) != nil) fprintf(LogBisection  s) )
   drain(LogBisection)
else
   fprintf(LogBisection "\n#####  ERROR_MAX: Cannot evaluate if the target value is between MIN and
MAX ###### \n") ;; if MPP failed, input offset is not between MIN and MAX)
)

fprintf(LogBisection "End time:")
fprintf(LogBisection getCurrentTime() )
fprintf(LogBisection "\n")
fprintf(LogBisection "End initial max\n")
fprintf(LogBisection "**********\n")
drain(LogBisection)

;;verify initial range before starting Binary Search

range_flag=0
if( flagRun_MAX != nil && flagRun_MIN != nil then  ;;; offset out of initial range: do not run
dichotomy
  range_flag = 1
  fprintf(LogBisection "\n ERROR => offset not in initial range => Both range boundaries   leads to a PASS
run\n")
)

if( flagRun_MAX == nil && flagRun_MIN == nil then  ;;; offset out of initial range: do not run dichotomy
  range_flag = 1
```

```
  fprintf(LogBisection "\n ERROR => offset not in initial range => Both range boundaries leads to a FAIL run\n") )

;;Binary search loop starts here

if(range_flag==0 ;if one of the previous simulation failed and the other succeed start the dichotomy search
then

while( (MAX-MIN) >= accuracy
nbrIter++
MID=(MAX+MIN)/2.0                    ;calculation of mid point range
 axlSetVarValue(axlGetVar( sdb "Rdata" ) sprintf(nil "%L" MID))
 fprintf(LogBisection "\n\n*****************************\n")
 fprintf(LogBisection "Start iteration %L  : " nbrIter)
 fprintf(LogBisection getCurrentTime() )
 fprintf(LogBisection "\nMID range of iteration %L = %L \n" nbrIter axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
 printf( "Start iteration %L  \n" nbrIter)
 printf( "MID range of iteration %L = %L \n" nbrIter sprintf(nil "%L" axlGetVarValue( axlGetVar( sdb "Rdata" ) )))
 drain(LogBisection)

;;;;; run statistical simulation for MID and print result

flagBi = ocnxlRun()
if(( flagBi==nil) then
  printf("\n#####  ERROR_OCEAN: Simulation run failed ###### \n" )
  fprintf(LogBisection "\n#####  ERROR_OCEAN: Simulation run failed ###### \n" )
)

flagRun_MID=
infile("/tmp/smraihi/adenova/work_salmen/cbsa/adexl/work_salmen:cbsa:1/nom/reports/result.mpp.report")

if(flagRun_MID !=nil then
   printf("_____Output summary SUCCESSFULLY generated for MID = %L (s) \n" axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
   fprintf(LogBisection "_____ Output summary SUCCESSFULLY generated for MID = %L (s) \n" axlGetVarValue( axlGetVar( sdb "Rdata" ) ))
   fprintf(LogBisection "_____ MPP report :  \n" )
   while( (gets(s flagRun_MID) != nil)
   fprintf(LogBisection  s) )
   drain(LogBisection)
else
fprintf(LogBisection "\n#####  ERROR_MID: No value computed during the simulation ###### \n")
)

if(( flagRun_MID != nil) then    ;; new MIN and MAX for next simulation
```

```
if(( flagRun_MAX != nil) then MAX=MID
else MIN=MID
)
else
if(( flagRun_MAX != 0) then MIN=MID
else MAX=MID)
)

  fprintf(LogBisection "updated MAX = %L \n" MAX)
  fprintf(LogBisection "updated MIN = %L \n" MIN)
  drain(LogBisection)
)

 final_offset=(MAX+MIN)/2  ;;;; when accuracy is reached, print the final value

fprintf(LogBisection
"\n\n#################################################################################\n")
    fprintf(LogBisection "\n_____ Offset found in %L iterations \n" nbrIter)
    fprintf(LogBisection "\n_____ extracted offset is %L \n" final_offset)
    fprintf(LogBisection
"\n\n#################################################################################\n")
drain(LogBisection)
printf("_____ Offset found in %L iterations \n" nbrIter)
printf("_____ extracted offset is %L \n" final_offset)

else

fprintf(LogBisection
"\n\n#################################################################################\n")
fprintf(LogBisection "\nPLEASE UPDATE THE RANGE AND RESTART DICHOTOMY SCRIPT\n ")
fprintf(LogBisection
"\n\n#################################################################################\n")

printf( "\n\n#################################################################################\n")
printf( "\n_PLEASE UPDATE THE RANGE AND RESTART DICHOTOMY SCRIPT ____      \n ")
printf( "\n\n#################################################################################\n")
)
;========================= End binary search algorithm ==========================

fprintf(LogBisection "\n Binary Search End =======> " )
fprintf(LogBisection getCurrentTime() )
fprintf(LogBisection "\n\n\n")

close(LogBisection)

;============================= End XL mode ====================================
ocnxlEndXLMode()
```

Appendix C: Cadence Spectre MTJ Model

```
simulator lang=spectre

**************************************
** Option and parameters settings **
**************************************

* TMR is the percent increase of RMTJ => (HRS-LRS)/LRS
parameters TMR = 1

* Vh is the voltage drop (Volt) across the MTJ which leads to TMR = TMR_0/2
parameters Vh = 0.43

*R_LRS is the value of the low resistive state (Ohm)
parameters R_LRS = 4000

*Process_skew is the max variation rate impacting low MTJ

parameters pct=5 Process_skew = pct/100
parameters  RMTJ0=R_LRS RMTJ1=R_LRS*(1+TMR)

**************************************
***** Sub circuits settings ********
**************************************

model R_MTJ0  resistor res = RMTJ0
model R_MTJ1  resistor res = RMTJ1

**************************************
*******    APPLY VARIATIONS       *****
**************************************

statistics{
        process {
                  vary RMTJ0 dist=gauss std=R_LRS*Process_skew
                  vary RMTJ1 dist=gauss std=R_LRS*Process_skew*(1+TMR)
                }
}
```

Appendix D: C-language model for estimating memory-array yield after 2D-repair (section 4.6.2.3)s

```
#include <stdio.h>
#include <assert.h>
#include<math.h>
#include<stdlib.h>

// factorielle function
double factorielle(int y){
   int x;
   double resultat=1;
   if (y==0){resultat=1;}
   for (x=1;x<=y;x++)
   {
      resultat=resultat*x;
   }
   return resultat;
}

//initialization

#define MAX_i 120
#define MAX_lambdaSD 50
#define MAX_M 40
#define MAX_N 30
#define MAX_Z (MAX_M+MAX_N)

double p1[MAX_i+1][MAX_M+1][MAX_N+1][MAX_Z+1];
double p2[MAX_i+1][MAX_M+1][MAX_N+1][MAX_Z+1];
double p3[MAX_i+1][MAX_M+1][MAX_N+1][MAX_Z+1];
double p4[MAX_i+1][MAX_M+1][MAX_N+1][MAX_Z+1];

double S1[MAX_i+1][MAX_M+1][MAX_N+1][MAX_Z+1];
double DSR[MAX_i+1] ;
double Dp[MAX_i+1];
double YR;

double fp1(int i,int m,int n,int z){if (i<0 || m<0 || n<0 || z<0 ) {return 0;} else  {return p1[i][m][n][z];} }
double fp2(int i,int m,int n,int z){if (i<0 || m<0 || n<0 || z<0 ) {return 0;} else  {return p2[i][m][n][z];} }
double fp3(int i,int m,int n,int z){if (i<0 || m<0 || n<0 || z<0 ) {return 0;} else  {return p3[i][m][n][z];} }
double fp4(int i,int m,int n,int z){if (i<0 || m<0 || n<0 || z<0 ) {return 0;} else  {return p4[i][m][n][z];} }
```

```
int main(void){
   int M=10;
   int N=10;
   int Z=M+N;
   int R=128;
   int C=128;
   double lambdaSD=10.02;
   double x=0., sp1=0.,sp2=0.,sp3=0.,sp4=0;
   int min_i, min_m, min_n, min_z;
   double r;
   double f;
   double g;
   int i;
   int m;
   int n;
   int z;
   int k=3;
   int max_i=k*lambdaSD+1;

// calculation of YR for i=1

   i=1;
   DSR[i]= S1[i][0][0][1]=1.0;

   r=pow(lambdaSD,i);
   f=exp(-lambdaSD);
   g=factorielle(i);
   Dp[i]=(r*f)/g;

   YR += DSR[i]*Dp[i];
   printf("DSR[%d]:%lg * Dp[%d]:%lg = %lg\n\n",i,DSR[i],i, Dp[i], DSR[i]*Dp[i] );

// definition of p1, p2, p3, p4

   for (i=1;i<max_i+1;i++){
      for (m=0;m<M+1;m++){
         for (n=0;n<N+1;n++){
               int maxz=M+N-m-n;
            for (z=0;z<maxz+1;z++){
               x=(R*n+C*m-m*n-(i-z))/(double)(R*C-i);      x      =      (x!=x)?0.0:x;      p1[i][m][n][z]      =
(x<0.)?0.:(x>1.)?1.:x;
               x=((C-n-z)*z+0.5*z*(z-1))/(double)(R*C-i);      x      =      (x!=x)?0.0:x;p2[i][m][n][z]      =
(x<0.)?0.:(x>1.)?1.:x;
               x=((R-m-z)*z+0.5*z*(z-1))/(double)(R*C-i);x          =          (x!=x)?0.0:x;p3[i][m][n][z]          =
(x<0.)?0.:(x>1.)?1.: x;
               x=((R-m-z)*(C-n-z))/(double)(R*C-i);x = (x!=x)?0.0:x;p4[i][m][n][z] = (x<0.)?0.:(x>1.)?1.:x;
            }//end z loop
```

```
        }//end of n loop
      }//end of m loop
   }//end of i loop
```

// calculation of YR starting from i=2

```
   for (i=1;i<max_i;i++){
      for (m=0;m<M+1;m++){
         for (n=0;n<N+1;n++){
               int maxz=M+N-m-n;
            for (z=0;z<maxz+1;z++){
               DSR[i+1] +=
                S1[i+1][m][n][z]=S1[i][m][n][z]*fp1(i,m,n,z)+S1[i][m-1][n][z+1]*fp2(i,m-
1,n,z+1)+S1[i][m][n-1][z+1]*fp3(i,m,n-1,z+1)+S1[i][m][n][z-1]*fp4(i,m,n,z-1);
            }//end of z loop
         }//end of n loop
      }//end of m loop

      r=pow(lambdaSD,i);
      f=exp(-lambdaSD);
      g=factorielle(i);
      Dp[i+1]=(r*f)/g;

      YR += DSR[i+1]*Dp[i+1];
      printf("DSR[%d+1]:%lg * Dp[%d+1]:%lg = %lg\n\n",i,DSR[i+1],i, Dp[i+1], DSR[i+1]*Dp[i+1] );
   }//end of i loop
   printf("YR=%lg\n",YR);
   exit(0);
```

List of Acronyms

SoC	System on Chip
RRAM	Resistive Random-Access Memory
RAM	Random-Access Memory
PCM	Phase-Change Memory
MRAM	Magnetic Random-Access Memory
LRS	Low Resistive State
HRS	High Resistive State
STT	Spin-Transfer Torque
SOT	Spin-Orbit Torque
MTJ	Magnetic Tunnel Junction
TMR	Tunneling MagnetoResistance
FM	FerroMagnetic
DWM	Domain Wall Motion
SKM	SKyrmion Motion
CMOS	Complementary Metal Oxide Semiconductor
BER	Bit Error Rate
SRAM	Static Random-Access Memory
DRAM	Dynamic Random-Access Memory
CB-RAM	Conductive-Bridge resistive Random-Access Memory
Ox-RAM	Oxide-based resistive Random-Access Memory
IoT	Internet of Things

MC	Monte Carlo
RSM	Response Surface Model
MPP	Most Probable Point
ECC	Error Correcting Code
POCT	Partial Offset-Cancellation Technique
FOCT	Full Offset-Cancellation Technique
CMFB	Common-Mode FeedBack
CBSA	Covalent-Bonded cross-coupled current-mode Sense Amplifier
FL	Free Layer
HM	Heavy Metal
OEM	Ordered Element Matching

List of Publications

Journal

Salmen Mraihi, Elmehdi Boujamaa, Cyrille Dray, and Jacques-Olivier Klein, "Variability analysis of the sensing path for resistive memories for circuit design optimization" *(Revised); Microelectronics Journal*, 2018.

International Conference with publication

Salmen Mraihi, Elmehdi Boujamaa, Cyrille Dray, and Jacques-Olivier Klein, "Offset Analysis and Design Optimization of a Dynamic Sense Amplifier for Resistive Memories"; *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 326–331, 2017.

Synthèse en Français

Chapitre 1: Introduction générale

De nos jours, la conception des systèmes sur puce devient de plus en plus complexe, et requiert des densités de mémoire sans cesse grandissantes. Pour ce faire, une forte miniaturisation des nœuds technologiques s'opère. Les mémoires non-volatiles résistives, tels que les RRAM, PC-RAM ou MRAM se présentent comme des solutions prometteuses afin d'assurer à la fois une densité suffisante et des faibles contraintes en surface, en latence, et en consommation à l'échelle nanométrique. Cependant, la variabilité croissante des cellules mémoires ainsi que des circuits en périphérie, tels que des circuits de lecture, est un problème majeur à prendre en considération. Cette thèse consiste en une étude détaillée et une aide à la compréhension de la problématique de variabilité appliquée aux circuits de lecture pour mémoires résistives, et propose des solutions d'amélioration du rendement de lecture de ces mémoires. L'objectif principal de cette thèse est de comprendre et d'analyser en profondeur la performance en lecture des mémoires résistives, tout au long du chemin de lecture de la mémoire : de la cellule mémoire à l'amplificateur de lecture en passant par l'architecture de référence de cet amplificateur. Pour ce faire, diverses études ont été réalisées : a) revue générale des solutions existantes d'amélioration du rendement de lecture, à la fois au niveau circuit et système, b) développement d'un modèle statistique évaluant la contribution à la marge de lecture de chaque composant à l'entrée de l'amplificateur de lecture en prenant en compte les éléments parasites et leurs variabilités, c) analyse, caractérisation, modélisation et optimisation de l'offset d'un amplificateur de lecture dynamique pour mémoires résistives, et d) proposition d'architecture d'amplificateur de lecture à annulation d'offset et réduction de la variabilité de la référence. Ce manuscrit présente en premier lieu un état de l'art sur les mémoires résistives puis sur les techniques d'amélioration du rendement de lecture de ces mémoires déjà présentes dans la littérature, tant au niveau système qu'au niveau circuit. Le chapitre suivant traite de l'analyse profonde et détaillée de la variabilité tout au long du chemin de lecture la mémoire, de la cellule seule au circuit amplificateur de lecture avec l'analyse d'une architecture en particulier, afin de mettre en évidence les différents critères en jeu pour une future optimisation. De cette analyse découle par la suite une proposition d'architecture d'amplificateur de lecture prenant en compte l'ensemble des facteurs de réduction du rendement de lecture énoncé précédemment. Cette architecture présente un schème de référence précis et facile à implémenter ainsi qu'un algorithme d'annulation d'offset donnant un rapport signal à offset maximum. Les détails de la conception d'une structure de test d'architecture d'amplificateur de lecture de ce type sont décrits au chapitre 5, et permet au lecteur de cerner les paramètres clés et les problématiques à prendre en compte lors de la phase de conception de ce genre de circuit à des nœuds technologiques extrêmement bas (22nm). La conclusion liste quant à elle des perspectives d'études complémentaires à cette thèse, ainsi que des idées supplémentaires d'amélioration du rendement de lecture.

Chapitre 2: Etat de l'art

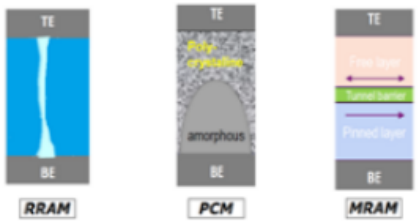
Depuis des décennies, la loi économique de Moore, qui prédit que la densité des systèmes sur puce (SoC=System-On-Chip) doit doubler tous les 18 mois, régit le secteur industriel de la microélectronique. L'extension de cette loi s'avère de plus en plus complexe à maintenir dans les années à venir du fait de la continuelle miniaturisation des composants électroniques, en particulier les mémoires à technologies CMOS. En prenant l'exemple de la mémoire flash, nous remarquons que la réduction des nœuds technologiques a pour conséquence de complexifier l'efficacité des mécanismes de stockage : seulement une centaine d'électrons peuvent être stockées dans la grille flottante de la mémoire flash en dessous de 20 nm [11].

Au vu de cela, les chercheurs s'orientent vers le développement d'alternatives technologiques, dont les caractéristiques sont égales voire supérieures aux mémoires actuelles. Trois d'entre elles émergent et apparaissent comme des candidats potentiels au remplacement de la mémoire flash ou de la DRAM pour certaines applications. Leur principe de fonctionnement n'est plus basé sur du stockage de charge. L'information est cette fois enregistrée sous la forme d'états résistifs dépendants du matériau utilisé, communément un état résistif bas (ou LRS-Low Resistive State) ou un état résistif haut (ou HRS-High Resistive State).

La mémoire à changement de phase (ou PCM) est une mémoire résistive composée de deux électrodes métalliques prenant en sandwich un oxyde et un matériau à changement de phase (principalement un chalcogénure). Ce matériau peut passer d'une phase cristalline (correspondante à un LRS) à une phase amorphe (HRS) et vice-versa. L'opération d'écriture de cette mémoire consiste à appliquer un pulse de courant à travers l'oxyde afin de chauffer le chalcogénure. Selon la procédure de refroidissement, une zone cristalline ou amorphe se crée alors à l'intérieur du matériau après une phase de recuit (ou annealing). Cette technologie de memristor permet une densité d'intégration élevée et une vitesse d'écriture moyenne (environ 75 ns), mais aussi une endurance limitée (typiquement 10^8 cycles d'écriture).

La ReRAM (ou Resistive RAM) est basée sur un oxyde placé entre deux métaux. La formation d'un filament conducteur à l'intérieur de l'oxyde génère un état résistif bas, alors que la rupture de celui-ci crée un état résistif haut. Le filament conducteur peut être composé d'atomes de métaux générés par réactions d'oxydoréduction (dans la cas de la conductive-bridge RAM) ou de lacunes d'oxygène (dans le cas de l'OxRAM). Comparée à la PCM, la ReRAM présente une plus grande densité d'intégration, une endurance élevée (10^{10} cycles), mais aussi une grande rapidité d'écriture (environ 20 ns) et une faible consommation d'énergie. [21][22][26][27].

Enfin, le fonctionnement de la MRAM (ou Magnetic RAM) s'appuie sur des propriétés magnétiques et exploite le principe de magnétorésistance. L'élément mémoire magnétique est appelé la jonction tunnel magnétique (ou MTJ), et consiste en une couche d'oxyde tunnel déposée entre deux couches ferromagnétiques. Une de ces deux couches présente une orientation magnétique fixe (appelée couche fixe) alors que l'orientation magnétique de la seconde couche peut être modifiée (appelée couche libre). Un courant d'électrons, qui possèdent un moment angulaire appelé spin-up ou spin-down, est appliqué à travers la MTJ. Si l'orientation magnétique des deux couches ferromagnétiques est parallèle, la probabilité pour que les électrons dépassent la barrière énergétique entre les douches couches est faible, c'est l'état résistif bas. Si l'orientation magnétique des deux couches ferromagnétiques est antiparallèle, une barrière énergétique plus élevée est imposée par la couche fixe, c'est l'état résistif haut [28][29]. L'un des mécanismes d'écriture de la MRAM les plus avancées est l'effet STT (ou Spin-Transfer Torque). Il présente l'avantage d'une très grande vitesse d'écriture (une dizaine de nanosecondes ont été démontrées) et d'une endurance considérée comme infinie. En revanche, il requiert des courants trop élevées (centaines de microampères) afin d'atteindre le seuil d'accumulation des spin d'électrons nécessaires à l'inversion de l'orientation magnétique de la couche libre de la MTJ [34][35][36][37][38].



a)

	Cell size (F^2)	Write speed	Power consumption	Endurance
STT-MRAM	6 to 12	10ns	Medium	10^{15}
PCM	6 to 12	75ns	Medium	10^{8}
RRAM	6 to 12	20ns	Low	10^{10}
SRAM	100-150	2ns	Low	10^{18}
Flash	4	10µs	Very high	10^{5}

b)

Figure 108 - a) Illustrations des trois principales technologies de mémoire résistive b) Comparaison des performances stand-alone des mémoires résistives avec les technologies actuelles

Alors que les techniques d'amélioration du rendement d'écriture des mémoires résistives semblent maitrisées, leur rendement de lecture est grandement menacé par la miniaturisation, en raison de leur sensibilité croissante à la variabilité à la fois de la cellule de mémoire résistive, de la matrice mémoire et de ses éléments en périphérie.

La variabilité d'un paramètre est observable entre différents *wafers*, entre différents *die* d'un même *wafer* (variabilité *interdie*) ou entre différents emplacements d'une même *die* (variabilité *intra-die*) [40]. On entend par variabilité celle résultante de non-uniformités du process de fabrication des matériaux résistifs, mais aussi celle résultante de variations CMOS dans des circuits basés sur des pairs de transistor, comme les amplificateurs de lecture. Les principales sources de variations CMOS sont les fluctuations de dopant aléatoires, les fluctuations de densité aux états d'interface ou encore des rugosités de bord de ligne [41]. Ceci se traduit par de sévère variations de paramètres caractéristiques du composant MOS comme la tension de seuil ou la longueur de canal effective. La variabilité suscite donc un important intérêt dans le but de concevoir des circuits analogiques fiables.

Les techniques d'amélioration de la tolérance aux variations des mémoires les plus communément utilisées et maitrisées sont les codes correcteurs d'erreur et les redondances lignes et/ou colonnes. Ceux-ci sont cependant couteux en surface notamment. Il convient ainsi de proposer une alternative au niveau circuit en optimisant l'amélioration de la tolérance aux variations des amplificateurs de lecture (ou sense amplifier).

Les architectures conventionnelles d'amplificateur de lecture consistent en une comparaison en tension ou en courant de la cellule mémoire avec une référence moyenne par l'intermédiaire d'un simple étage dynamique de type *latch*. Ceux-ci ne conviennent pas aux mémoires résistives : l'implémentation des connexions pour construire la référence moyenne s'avère complexe afin de s'assurer de respecter les règles de layout ; les circuits dynamiques présentent une forte sensibilité aux variations. [64] propose une architecture traitant ces deux problématiques. Une analyse détaillée de la variabilité de ce circuit est proposée dans le chapitre 3.

Par ailleurs, des techniques d'annulation d'offset comme la FOCT (technique d'annulation d'offset totale) au niveau circuit émergent pour des architectures de type amplificateur opérationnel. Ces techniques consistent en des opérations multi-phases : au moins une phase d'échantillonnage de la variabilité de chacun des composants du circuit (σV_T) à travers une ou des capacités switchées, ainsi qu'une phase d'amplification du signal d'entrée dans lequel le même chemin de courant que lors de la phase d'échantillonnage est conservé, afin d'annuler la variabilité de chaque transistor de l'amplificateur. Une inversion des entrées lors de cette phase d'amplification permet de doubler le signal et d'augmenter ainsi la marge de lecture.

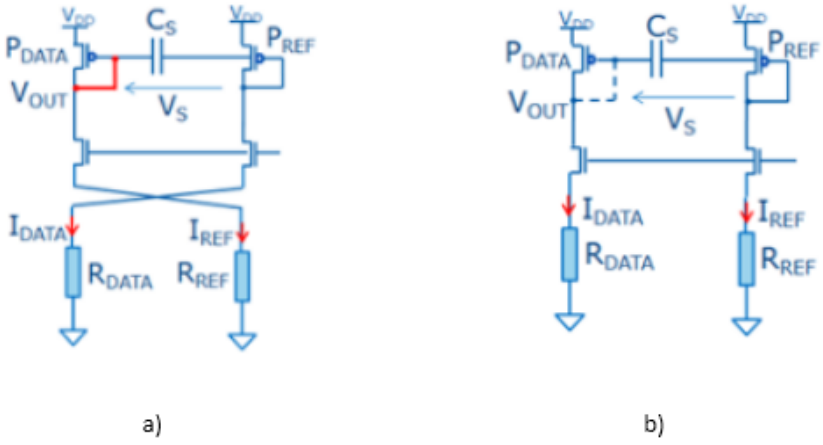


Figure 109 – Amplificateur de lecture à annulation totale d'offset a) phase d'échantillonnage b) phase d'amplification

Chapitre 3: Analyse d'offset et optimisation au niveau circuit d'un amplificateur de lecture dynamique pour mémoires résistives

Ce chapitre offre une analyse complète de la marge de lecture tout au long du chemin de lecture de la cellule mémoire résistive, et propose une méthodologie d'optimisation au niveau circuit de la dégradation de cette même marge de lecture.

Dans un premier temps, un modèle statistique est décrit, permettant d'évaluer l'impact de la variabilité sur la dégradation de la marge de lecture à l'entrée du chemin de lecture de la mémoire résistive. Ce modèle inclut les principaux paramètres influençant la variabilité du chemin de lecture, tant au niveau technologique (σR_{DATA}: : variabilité de la bit cell résistive; $\sigma \Sigma R_{PAR}$: variabilité de la *bit line*, de la *source line* et du dispositif d'accès) que circuit ($\sigma V_{T,NCLAMP}$: variabilité du transistor dit de 'clamp', permettant d'initialiser les tensions de *bit line* aux entrées de l'amplificateur de lecture). Les différentes étapes pour une optimisation au niveau circuit de la fiabilité de lecture sont:

- Choisir le taux de variation à appliquer à l'ensemble du circuit (cellule mémoire plus amplificateur de lecture)
- Considérer des paramètres process données (σR_{DATA}, $\sigma \Sigma R_{PAR}$) et les positioner vis-à-vis de la dégradation en marge de lecture correspondante
- Considérer la non-linéarité de la mémoire résistive à travers le transistor de 'clamp'
- Comparer l'impact de la variabilité de la mémoire résistive et du transistor de 'clamp' sur la dégradation de la marge de lecture
- Redimensionner le transistor de 'clamp' en prenant en compte sa contribution en surface afin de minimiser la dégradation de la marge de lecture.

Le principe du redimensionnement du transistor de 'clamp' est le suivant:

- Cibler un certain pourcentage de tolérance de la dégradation de la marge de lecture
- Rapporter l'ensemble des paires {W/L, σV_T} permettant d'obtenir ce pourcentage
- Prendre en compte les contributions en surface de chacune des paires
- Choisir la paire {σV_T, W/L} donnant une contribution en surface minimum

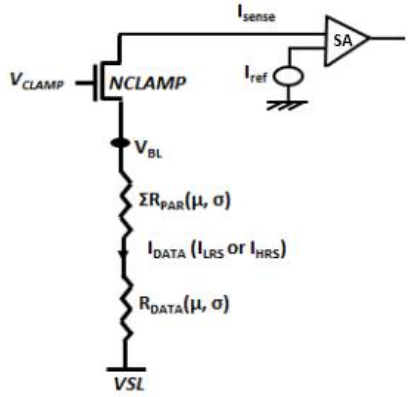


Figure 110 – Circuit équivalent du chemin de lecture de la mémoire résistive

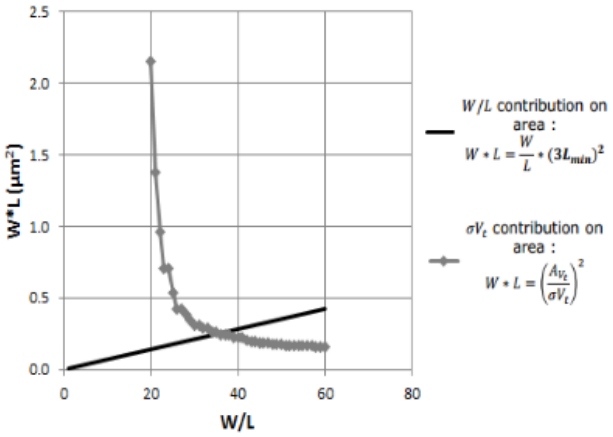


Figure 111 – Contributions en surface des paramètres du transistor de 'clamp' (respectivement g_m et σV_T) (exemple pour 10% de dégradation de RM0 acceptée; R_{LRS}=4kΩ, TMR=100%, N=3, $\sigma R_{DATA}/R_{DATA}$=5%, V_T=0.25V). Pour cette étude de cas en particulier, le redimensionnement correspond à {σV_T=1.2 mV, W/L=36}

Ainsi, à ce stade, c'est-à-dire avant le traitement du signal d'entrée par l'amplificateur de lecture, l'atténuation de la dégradation de la marge de lecture est optimisée au maximum.

La fiabilité de lecture de la mémoire résistive à travers l'amplificateur de lecture dépend de l'architecture circuit de celui-ci. Le circuit CBSA introduit au Chapitre 2 se présente comme un circuit dynamique prometteur de par sa symétrie (réduction possible d'offset) et son implémentation de la référence (deux cellules résistives complémentaires : régularité layout). Il convient d'analyser en profondeur sa variabilité (quantifiée par son offset).

L'offset global de ce circuit est constitué d'une part d'un offset systématique dû au déséquilibre des capacités de couplage des *latchs* et des charges capacitives, et d'autre part d'un offset aléatoire dû à la variabilité des transistors P des *latchs* et des transistors de 'clamp' des tensions de *bit line*, qui sont responsables du déclenchement de l'amplification du signal de lecture. L'ajout de capacités MOS correctement dimensionnées permet d'annuler l'offset systématique. La modélisation de l'offset aléatoire ainsi que sa caractérisation rend compte d'une valeur minimale de 200 Ohms ramenés en entrée de l'amplificateur de lecture. Cette valeur, obtenue pour un cas idéal correspondant à une variabilité nulle des charges capacitives, reste élevée et démontre la nécessité d'architectures non-dynamiques intégrant des techniques d'annulation d'offset, tout en étant compatibles avec des implémentations de références respectant la régularité layout.

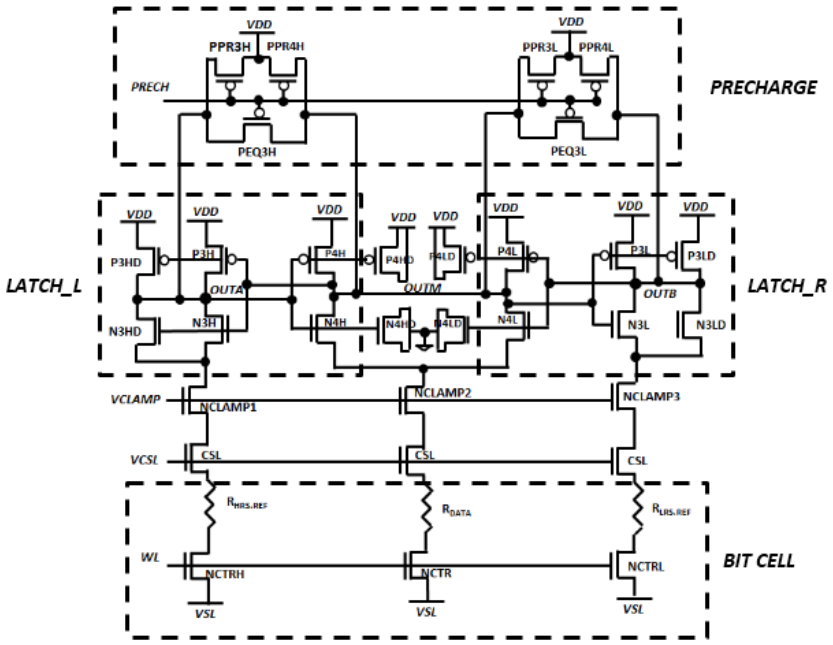


Figure 112 – 'Covalent-Bonded Sense Amplifier' (CBSA) [64]

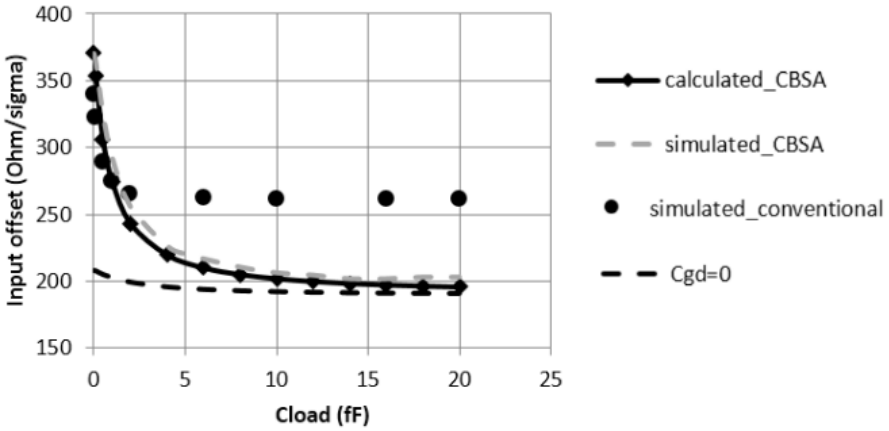


Figure 113 – Comparaison de l'offset aléatoire ramené en entrée du CBSA (modèle et simulation) avec un amplificateur de lecture conventionnel pour une variabilité de charge capacitive nulle. La courbe en pointillés montre l'influence du couplage capacitif des différents latchs pour de faibles charges capacitives

Chapitre 4: Proposition d'architecture circuit d'amplificateur de lecture à annulation d'offset et variabilité réduite du memristor de référence

[65] propose une architecture circuit d'amplificateur de lecture intégrant la technique FOCT ainsi qu'une implémentation de référence à multiplexage temporel (*Figure 114*). Le principe est de connecter alternativement au circuit deux dispositifs résistifs complémentaires et décorrélés qui serviront de référence. Deux phases d'opération permettront alors dans un premier temps l'échantillonnage des contributions à l'offset (phase d'échantillonnage) et dans un second temps l'amplification du signal d'entrée. Cette implémentation permet à la fois de concevoir une référence beaucoup moins complexe vis-à-vis des contraintes layout et de réduire la variabilité de celle-ci afin d'augmenter la marge de lecture de la mémoire résistive.

Cette thèse propose d'étendre la technique proposée à N références pour augmenter le rapport signal à offset correspondant. Ainsi, les chemins de courant de la donnée à lire et de la référence restent identiques entre les différentes phases d'opération afin de soustraire l'offset à lui-même à la fin de la phase d'amplification. La réduction de la variabilité de la référence consiste alors en l'ajout d'autant de phases d'échantillonnages par référence décorrélée supplémentaire.

La *Figure 115* illustre l'implémentation proposée pour trois références. L'ajout d'une référence décorrélée nécessite une branche de courant et une phase d'échantillonnage supplémentaires.

L'implémentation circuit pour un nombre de références paire rajoute une complexité design. En effet, pour quatre références par exemple, deux courants 'pull down' provenant de la même cellule résistive (R_{DATA}) doivent être générées. Un miroir de courant est alors nécessaire afin de pouvoir enregistrer le premier courant I_{DATA} lors d'une phase d'échantillonnage et pouvoir, lors de la phase d'amplification, générer les deux courants en 'pull down' (*Figure 116*). Cet étage devient néanmoins une source d'offset supplémentaire, qui peut être annulée car le même miroir de courant est utilisé entre les phases d'échantillonnage et d'amplification. La résistance R_{DUMMY} de la *Figure 116* permet de compenser le déséquilibre d'impédance de sortie crée par le miroir de courant.

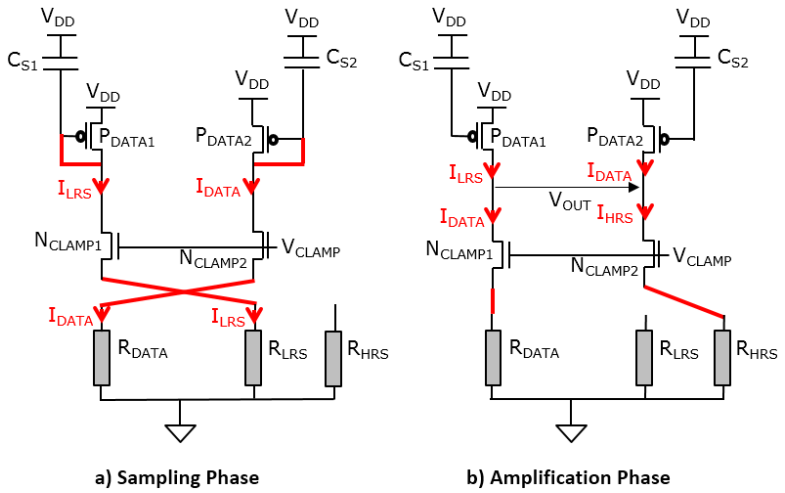


Figure 114 – Amplificateur de lecture à annulation d'offset et implémentation de référence à multiplexage temporel

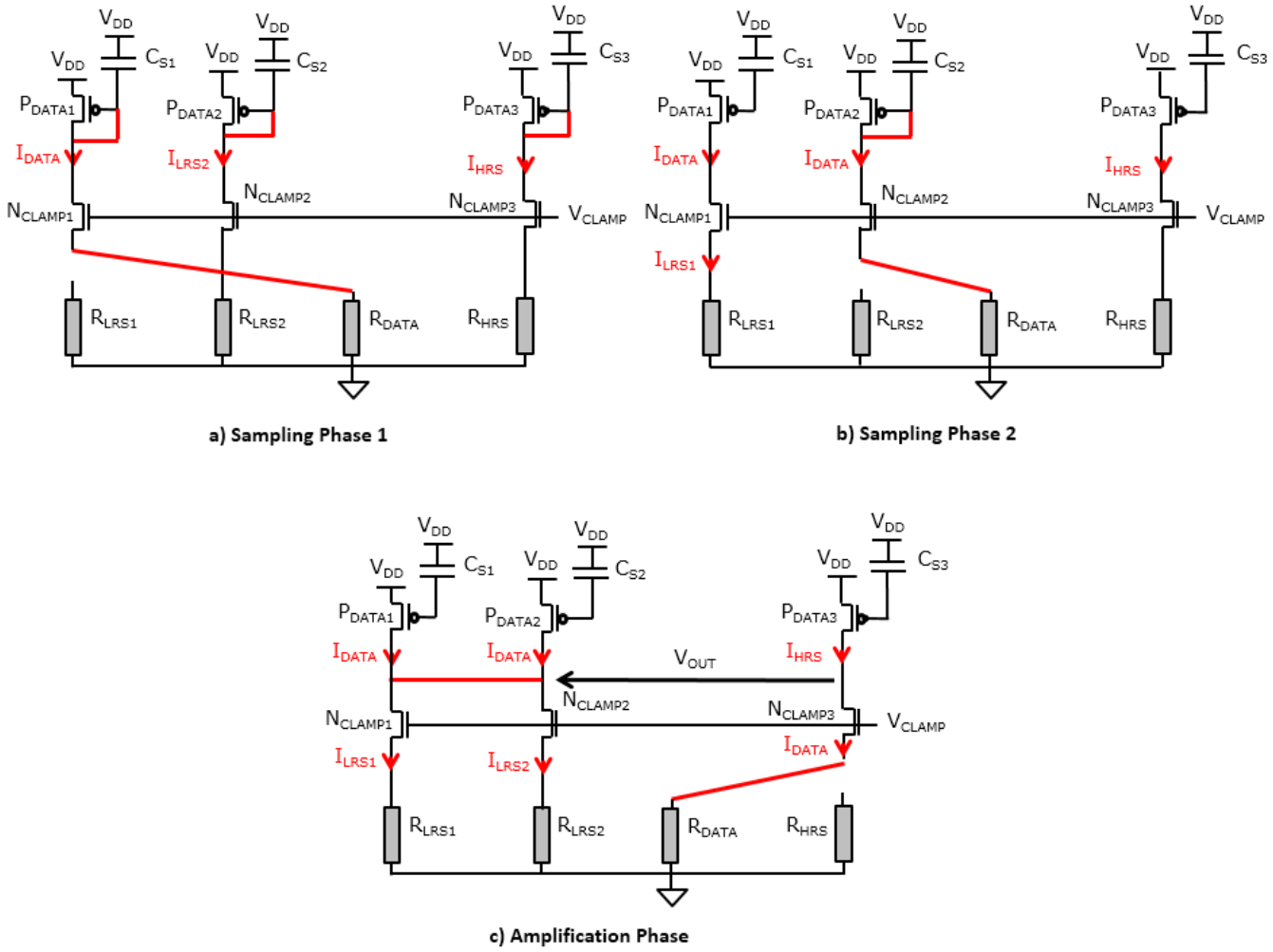


Figure 115 – Architecture d'amplificateur de lecture proposée, pour trois références (2 LRS et un HRS par exemple) a) Phase d'échantillonnage 1 b) Phase d'échantillonnage 2 c) Phase d'amplification

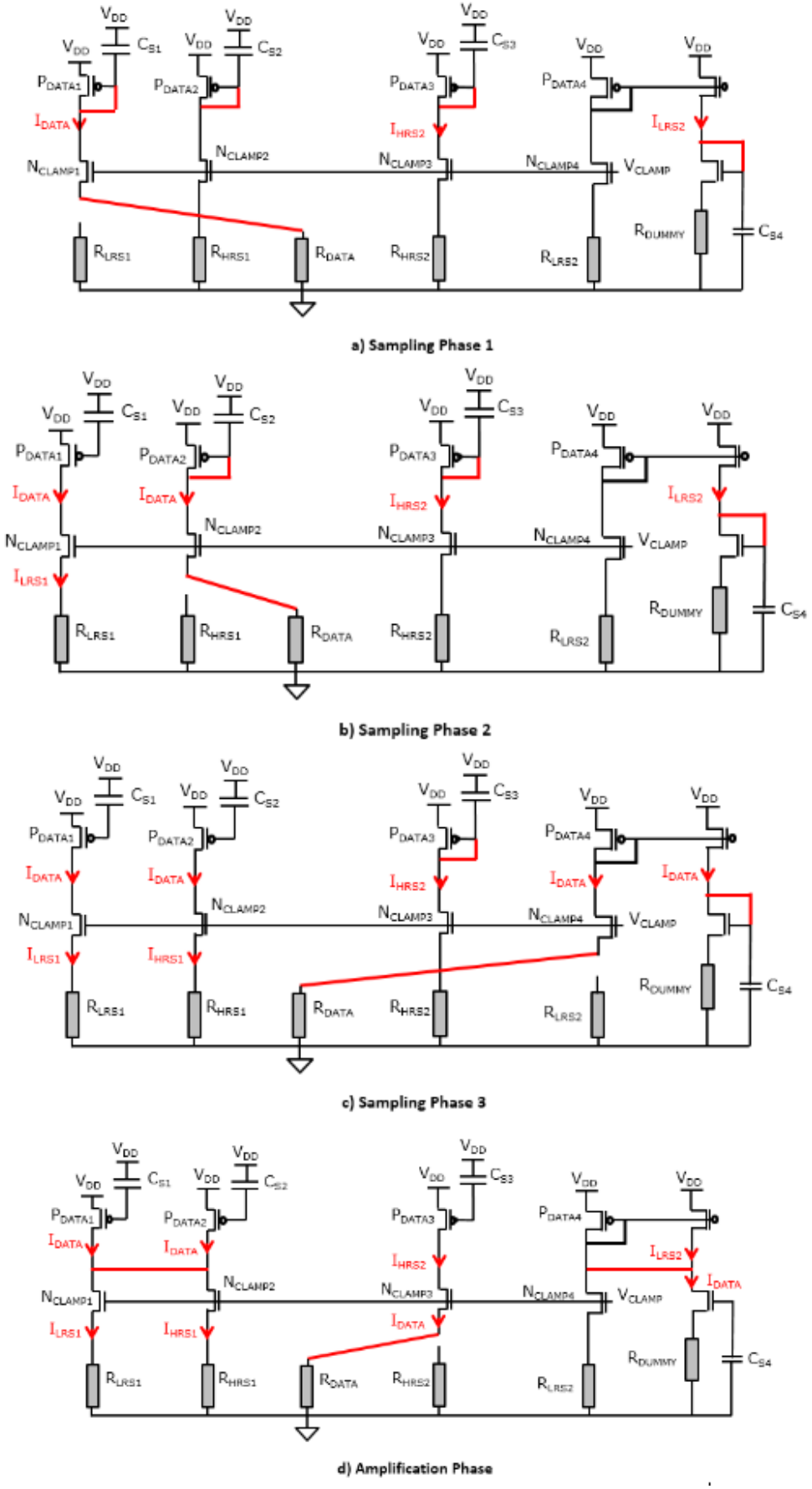


Figure 116 - Architecture d'amplificateur de lecture proposée, pour quatre références a) Phase d'échantillonnage 1 b) Phase d'échantillonnage 2 c) Phase d'échantillonnage 3 d) Phase d'amplification

La courbe de la *Figure 117* représente l'amélioration correspondante de la marge de lecture apportée par l'amplificateur de lecture proposé, en fonction du nombre de références implémentés et des paramètres process du dispositif résistif (MTJ par exemple avec $TMR = \frac{R_{HRS}-R_{LRS}}{R_{LRS}}$). Cette marge de lecture est comparée à celle d'un amplificateur de lecture typique d'une cellule 2T-2R, qui donne une marge maximum ($\mu(\Delta I) = 2.(I_{LRS} - I_{HRS})$). Un gain significatif est obtenu lors du passage de deux à quatre références implémentées. Pour σR_{LRS}=7% et TMR=100%, ce gain est d'un facteur 10 en terme de taux d'erreur bit (BER) (ou 0.5 sigma en terme de taux de variation global appliqué au circuit)

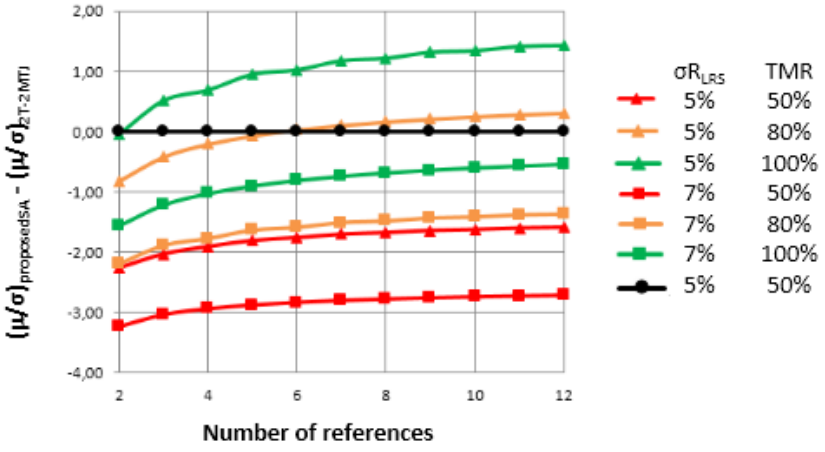


Figure 117 - Evolution du gain en marge de lecture apporté par l'amplificateur de lecture proposé en fonction du nombre de références implémentés et du dispositif résistif choisi (R_{LRS}=2.5 kΩ, V_{BL}=100mV)

Ce gain en marge de lecture est comparé à celui qu'apporterait l'utilisation de techniques d'amélioration du rendement de lecture au niveau système, comme les codes correcteurs d'erreur (ECC) ou la redondance ligne et/ou colonne. Il est démontré que 2 bits d'ECC ou une colonne (ou ligne) redondante sont requis pour obtenir un gain de marge de lecture équivalent à celui de l'amplificateur proposé pour quatre références. Il est aussi démontré que le coût en surface de ces configurations est supérieur à celui apporté par l'architecture circuit proposé.

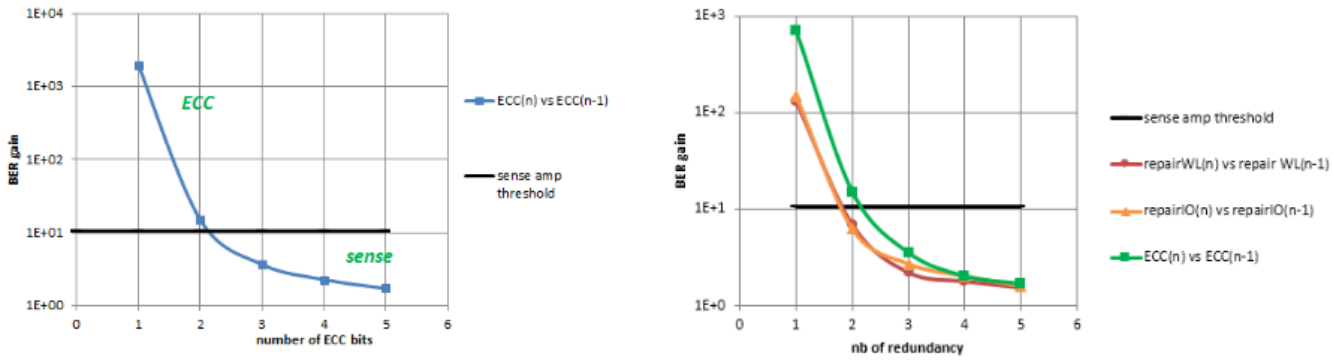


Figure 118 – Evolution du gain BER apporté par : a) l'ajout d'un bit ECC pour différents bits d'ECC initiaux b) l'ajout d'une ligne ou colonne redondante pour différentes lignes ou colonnes initiales, et comparaison avec le gain de l'amplificateur de lecture proposé pour quatre références (V_{BL}=100 mV, R_{LRS}=2.5 kΩ, TMR=100%, σR_{LRS}=7%, taille de la matrice mémoire = 10 kb, largeur de mot = 32, redondance nulle, probabilité d'erreur mémoire ciblée=10^{-4})

Pour ce qui est d'une redondance simultanée WL et IO, le *Tableau 12* permet d'extraire le gain en BER selon la configuration 2D choisie. Il est démontré que pour obtenir un gain équivalent à celui de l'amplificateur de

lecture proposé (environ égal à 10), il faut au moins passer de deux à plus de dix lignes et colonnes redondantes. Sachant que l'ajout d'une WL correspond à l'ajout d'une colonne supplémentaire mais aussi de décodeurs et d'un comparateur, alors que l'ajout d'une IO correspond à l'ajout de 32 multiplexeurs (pour un mot de 32 bits), le coût surfacique est très élevé comparé à celui de l'amplificateur proposé.

Tableau 12 – Extraction du BER (=λ_{SD}/memory_size) selon la configuration de redondance IO+WL choisie pour une matrice mémoire de 10kb (128 lignes, 128 colonnes)

# of spare rows (M) / # of spare columns (N)	2	6	10	14	18	22
	λ_{SD} for YR=99.99% (i.e. P_{fail}(memory)=10^{-4})					
2	**2.036**	4.222	6.096	7.602	7.956	7.963
6	4.222	6.124	8.022	10.02	12.418	15.548
10	6.096	8.022	10.02	12.424	15.788	**19.805**
14	7.602	10.021	12.427	15.74	**19.75**	24.13
18	7.956	12.418	15.788	**19.75**	24.15	28.93
22	7.963	15.548	**19.805**	24.15	28.93	34.11

Chapitre 5: Implémentation de l'architecture proposée sur structure de test et résultats de simulation

L'intérêt de l'architecture est démontré par la conception d'une structure de test dédiée à nœud technologique très bas (en dessous de 30 nm). Les principales contraintes layout de ce circuit consistent en une parfaite symétrie et régularité vis-à-vis des interconnections, ainsi qu'une attention toute particulière au routage des entrées vers les pads, qui est très sensible aux capacités et résistances parasites. Les très faibles valeurs d'offset obtenues par simulation post-layout (*Tableau 13*) confirment l'utilisation d'architectures similaires pour la conception de mémoires résistives à échelle nanométrique.

Tableau 13 – Résultats d'offset post-layout de la structure de test de l'amplificateur de lecture proposé (avec I_{HRS}=20.54 µA, I_{LRS}=33.58 µA, capacités de bit line = 50 pF, charge capacitive = 50 pF)

Extraction type	***Systematic offset***	***Random offset***
R only	*12 Ω*	*54 Ω*
C only	*3 Ω*	*24 Ω*
RC	*15 Ω*	*60 Ω*

Chapitre 6: Conclusion et perspectives

Cette thèse donne une analyse approfondie de la problématique de variabilité appliquée à la conception de circuits de lecture pour mémoires résistives. Elle propose des solutions d'architecture d'amplificateurs de lecture combinant annulation d'offset et réduction de la variabilité de la référence.

Les technologies de memristor tels que la STT-MRAM présentent des performances *standalone* prometteuses (vitesse d'écriture, endurance, miniaturisation …) comparées aux solutions (flash, SRAM …) existantes. L'un des plus grands freins à sa commercialisation est sa faible marge de lecture due à sa faible tolérance aux

variations. La littérature actuelle propose des solutions d'implémentation de référence à variabilité réduite (circuit CBSA), ou des techniques d'annulation d'offset pour les amplificateurs [101]. Il convient de proposer des architectures circuit traitant ces deux problématiques.

Pour ce faire, une analyse de variabilité du début du chemin de lecture de la mémoire est présentée. Cette étude inclut une méthodologie d'optimisation design de la dégradation de la marge de lecture prenant en compte les paramètres process de la cellule résistive. L'offset du circuit CBSA est modélisé et caractérisé, afin de rendre compte des facteurs de dégradation de la marge de lecture. La non-négligeable valeur de l'offset de ce circuit confirme la nécessité d'amplificateurs de lecture non-dynamiques, à annulation d'offset, et à implémentation de référence régulière en terme de layout et de faible variabilité.

Une architecture de ce type est ainsi proposée. Elle consiste en l'extension de la technique d'annulation totale d'offset (FOCT) à N références. Ce circuit combine annulation d'offset et réduction de variabilité de la référence, ce qui résulte en une amélioration significative du rapport signal à offset. L'estimation de la marge de lecture montre un gain d'un facteur 10 du taux d'erreur bit lorsque de deux à quatre références sont intégrés à ce circuit. Les coûts non-négligeables en surface résultant de l'augmentation du nombre de références sont comparés à ceux des techniques d'amélioration de rendement de lecture au niveau système (ECC, redondances ligne et/ou colonnes). Ceux-ci démontrent l'intérêt de l'utilisation du circuit proposé, notamment pour des applications tolérantes en latence et en consommation énergétique. Une structure de test de l'amplificateur de lecture proposé pour deux références a été conçu et son annulation d'offset a été démontrée par simulations post-layout.

Parmi les solutions prometteuses présentes dans la littérature proposant le meilleur compromis entre fiabilité de lecture des mémoires résistives et coût en surface, latence et consommation, peuvent être citées :

- Solutions au niveau design :
    - « Ordered element matching » : cette technique consiste en une disposition « intelligente » des différents éléments d'un circuit. Les paramètres de ces élements (valeur moyenne de la résistance de la cellule mémoire ou de la tension de seuil d'un transistor de l'amplificateur de lecture) peuvent être triés et regroupés afin de générés des éléments dont la variabilité est réduite [105][106][107][108].
- Solutions au niveau process (technologies de mémoires dérivées de la MRAM) - les méthodologies présentées dans ce manuscrit étant indépendantes de la technologie de mémoire résistive choisie, elles restent applicables à ces solutions :
    - La « racetrack memory » : cette technologie présente de meilleures performances que la STT-MRAM (densité, consommation, latence) grâce au mouvement de parois de domaine pour écrire la cellule [110][111][112][113][114][115]. [116] démontre que sa fiabilité de lecture peut être maitrisée.
    - La Spin-Orbit-Torque (SOT) –MRAM : la cellule mémoire est un dispositif à trois terminaux, ce qui permet de découpler les chemins de courant de lecture et d'écriture, et de réduire les courants et tensions d'écriture. Pour ce qui est de sa fiabilité de lecture, [118] introduit un modèle démontrant que l'effet SOT n'a pas seulement un impact sur la fiabilité d'écriture mais aussi sur la marge de lecture de la cellule mémoire. Un nouveau chemin de lecture à directions bilatérales est également proposé, permettant d'augmenter le courant critique dû à l'effet SOT d'un facteur 10, et ainsi de maintenir une fiabilité de lecture similaire à la STT-MRAM [119]

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : électrons, photons, sciences du vivant (EOBE)

Titre : *Prise en compte de la variabilité dans l'étude et la conception de circuits de lecture pour mémoires résistives*

Mots-clés : *mémoires résistives, conception de circuits analogiques, amplificateurs de lecture, variabilité, offset*

Résumé : De nos jours, la conception des systèmes sur puce devient de plus en plus complexe, et requiert des densités de mémoire sans cesse grandissantes. Pour ce faire, une forte miniaturisation des nœuds technologiques s'opère. Les mémoires non-volatiles résistives, tels que les RRAM, PC-RAM ou MRAM se présentent comme des alternatives technologiques afin d'assurer à la fois une densité suffisante et des faibles contraintes en surface, en latence, et en consommation à l'échelle nanométrique. Cependant, la variabilité croissante de ces cellules mémoires ainsi que des circuits en périphérie, tels que des circuits de lecture, est un problème majeur à prendre en considération. Cette thèse consiste en une étude détaillée et une aide à la compréhension de la problématique de variabilité appliquée aux circuits de lecture pour mémoires résistives. Elle propose des solutions d'amélioration de la fiabilité de lecture de ces mémoires. Pour ce faire, diverses études ont été réalisées : revue générale des solutions existantes d'amélioration du rendement de lecture, au niveau circuit et système ; développement d'un modèle statistique évaluant la contribution à la marge de lecture de la variabilité de chaque composante du chemin de lecture de la mémoire résistive ; analyse, caractérisation, modélisation et optimisation de l'offset d'un amplificateur de lecture dynamique pour mémoires résistives ; proposition d'architecture d'amplificateur de lecture permettant un rapport signal à offset optimum.

Title : *Design for variability of read circuitries for resistive memories*

Keywords : *resistive memories, design of analog circuits, sense amplifiers, variability, offset*

Abstract : Nowadays, Systems on chip (SoCs) conception is becoming more and more complex and demand an ever-increasing amount of memory capacity. This leads to aggressive bit cell technology scaling. Nonvolatile resistive memories (PC-RAM, RRAM, MRAM) are promising technologic alternatives to ensure both high density, low power consumption, low area and low latencies. However, scaling lead to significant memory cell and/or memory periphery variability. This thesis aims to address variability issues in read circuitries of resistive memories and propose solutions for read yield enhancement of these memories. To this end, several sub-studies were achieved: overall review of the existing solutions for read yield enhancement, at both circuit and system level; development of a statistical model evaluating the contributions to read margin of the variability of each component of the resistive memory sensing path; analysis, characterization modelling and optimization of the offset of one particular dynamic sense amplifier for resistive memories; proposal of a sense amplifier architecture that features an optimum signal to offset ratio.

Université Paris-Saclay

Espace Technologique / Immeuble Discovery

Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France