UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Micro- et Nano-électronique**

Arrêté ministériel : 7 août 2006

Présentée par

Giorgio PALMA

Thèse dirigée par **Prof. Amara AMARA**

préparée au sein du **CEA-LETI**
et de l' **Ecole Doctorale de Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS), Grenoble**

Nouvelles Architectures Hybrides: Logique / Mémoires Non-Volatiles et technologies associées

Thèse soutenue publiquement le **29 Novembre 2013**,
devant le jury composé de :

M Gérard GHIBAUDO
Prof., Université de Grenoble (IMEP-LAHC), Président
M Jean-Michel PORTAL
Prof., Polytech' Marseille (IM2NP), Examinateur
M Ian O CONNOR
Prof., Ecole Centrale de Lyon (ECL), Examinateur
M Amara AMARA
Prof., Institut Supérieur d'électronique de Paris (ISEP), Directeur de thèse

Abstract

Novel approaches in the field of memory technology should enable backend integration, where individual storage nodes will be fabricated during the last fabrication steps of the VLSI circuit. In this case, memory operation is often based upon the use of active materials with resistive switching properties. A topology of resistive memory consists of silver (Ag) as electrochemically active metal and amorphous germanium disulfide (GeS_2) acting as electrolyte and relies on the reversible formation and dissolution of a conductive filament. The application potential of these new memories is not limited to stand-alone (ultra-high density), but is also suitable for embedded applications. By stacking these memories in the third dimension at the interconnection level of CMOS logic, new ultra-scalable hybrid architectures becomes possible which exploit low energy operation, fast write/read access and high performance with respect to endurance and retention. In this thesis, focusing on memory technology aspects in view of developing new architectures, the introduction of non-volatile functionality at the logic level is demonstrated through three hybrid (CMOS logic + ReRAM devices) circuits: nonvolatile routing switches in a Field Programmable Gate Array, nonvolatile 6T-SRAMs and stochastic neurons of an hardware neural network. To be competitive or even improve existing solutions, limitations on the memory devices performances are identified and solved by stack engineering of CBRAM devices or providing fault tolerant circuits.

Résumé

Nouvelles Architectures Hybrides: Logique / Mémoires Non-Volatiles et technologies associées

Les nouvelles approches de technologies mémoires permettront une intégration dite back-end, où les cellules élémentaires de stockage seront fabriquées lors des dernières étapes de réalisation à grande échelle du circuit. Ces approches innovantes sont souvent basées sur l'utilisation de matériaux actifs présentant deux états de résistance distincts. Le passage d'un état à l'autre est controlé en courant ou en tension donnant lieu à une caractéristique I-V hystérétique. Nos mémoires résistives sont composées d'argent (Ag) en métal électrochimiquement actif et de sulfure de germanium amorphe (GeS_2) agissant comme électrolyte. Leur fonctionnement repose sur la formation réversible et la dissolution d'un filament conducteur. Le potentiel d'application de ces nouveaux dispositifs n'est pas limité aux mémoires ultra-haute densité mais aussi aux circuits embarqués. En empilant ces mémoires dans la troisième dimension au niveau des interconnections des circuits logiques CMOS, de nouvelles architectures hybrides et innovantes deviennent possibles. Il serait alors envisageable d'exploiter un fonctionnement à basse énergie, à haute vitesse d'écriture/lecture et de haute performance telles que l'endurance et la rétention. Dans cette thèse, en se concentrant sur les aspects de la technologie de mémoire en vue de développer de nouvelles architectures, l'introduction d'une fonctionnalité non-volatile au niveau logique est démontrée par trois circuits hybrides: commutateurs de routage non volatiles dans un Field Programmable Gate Arrays, une 6T-SRAM non volatile et les neurones stochastiques pour un réseau neuronal. Pour améliorer les solutions existantes, les limitations des performances des dispositifs mémoires sont identifiés et résolus en utilisant de nouveaux empilements ou en fournissant des circuits qui utilisent la variabilité des cellules pour avoir de meilleures performances.

Acknowledgements

I am deeply thankful to many people (listed in alphabetical order): Amara A., Anghel C., Barci M., Becu S., Ben-Jamaa H., Bernard M., Bernard S., Bichler O., Billiot G., Blachier D., Blaise P., Boussey J., Bretegnier D., Buckley J., Cabout T., Cagli C., Carabasse C., Charbonneau M., Charpin C., Cibrario G., Clermidy F., Cluzel J., Commault C., Coquand R., Dahmani F., De Salvo B., Deleonibus S., Delaval G., Diokh T., Faynot O., Gaillardon P.-E., Garbin D., Gary M., Gaud P., Gely M., Ghibaudo G., Grosgeorges P., Guy J., Hraziia, Hubert Q., Jahan C., Jalaguier E., Jovanovic N., Kiouseloglou A., Klein J.-O., Larcher L., Liebault J., Longnos F., Lorenzi P., Makosiej A., Manzoni L., Martin O., Masoero L., Mazurier J., Molas G., Morvan S., Navarro G., Nodin J.-F., O Connor I., Onkaraiah S., Oucheick H., Ouerghi I., Perniola L., Persico A., Philippe J., Pirrotta O., Poiroux T., Portal J.-M., Poupinet L., Prouvee J., Querlioz D., Reimbold G., Reyboz M., Robert V., Rolandi P., Romano G., Roule A., Singh P., Souchier E., Souiki S., Sousa V., Suri M., Thomas O., Todorova T., Toffoli A., Traoré B., Turkyilmaz O. and last but not least Vianello E.

To my brother Vincenzo

Manuscript Outline

This manuscript was submitted in partial fulfillment of the requirements for obtaining the degree of Doctor of Philosophy of the Grenoble Institute of Technology (Grenoble INP). The topic addressed in this Ph.D. thesis deals with novel hybrid logic and nonvolatile memory architectures and associated technologies. The manuscript covers both emerging memory devices features, in particular related to Conductive Bridge RAM technology and design/simulation of hybrid circuits such as nonvolatile 6T-SRAM, nonvolatile routing switches in FPGAs embodiments and neuron circuits. Foreseen advantages in implementing hybrid architectures are discussed accordingly to electrical results obtained on several CBRAM technologies. In fact, a part of the thesis was devoted to electrically characterize several CBRAM cell stacks in view of developing nonvolatile memory solution for: renovating some of the FPGA blocks, design a nonvolatile SRAM, or create fault-tolerant designs, that could also exploit non optimized technologies.

It's well known that intrinsic variability of emerging technologies such as ReRAM complicates the physical understanding, the prediction of device behavior under different stresses (thermal, noise) and finally the development of reliable electrical models implemented in commercial IC design tools. Typically, the development of behavioral compact models is subjected to parameters extracted on a large number of devices but, nowadays, there is a limited number of hybrid architectures, that could validate, for example, the performances of a distributed (on logic) memory node. Hence, the integration of a CBRAM compact model in IC design tools is the critical step for proposing new concepts, for validating the main operations of new circuits and eventually for launching the fabrication on silicon.

Chap. 1 of this thesis introduces the international context, the state of the art for emerging nonvolatile memories, the limitations and the issues due to the aggressive scaling of 6T-SRAM or due to the excessive power consumption in FPGA. In a second part, we will discuss the most advanced solutions to design nonvolatile 6T-SRAMs and to renovate FPGA blocks. We will conclude the chapter with some basic concepts related to hardware neural networks.

In **Chap. 2** we will present some of the results of the electrical characterizations of Ag/GeS_2 based CBRAM devices (integrated in the 1R or the 1T-1R architecture) and the empirical model used to explain and predict the switching parameters. We will propose a statistical analysis on the switching probability based on measurements on 8×8 memory arrays to improve the model taking into account cell to cell and cycle to cycle variability.

Chap. 3 is devoted to find a suitable CBRAM technology in view of design and optimize a specific block of an FPGA. Electrical results related to several CBRAM stacks fabricated with a dual-layer based electrolyte will be presented. Among other CBRAMs technologies, our dual-layer electrolyte stack (2 nm HfO_2−30 nm GeS_2) is the most promising. We will demonstrate a resistance ratio ($R_{\mathrm{off}}/R_{\mathrm{on}}$) higher than 10^6, a reset current of 100 µA and no forming step.

In **Chap. 4** we will discuss a nonvolatile 6T-SRAM, a nonvolatile routing switch based on 1T-2R architecture and a stochastic neuron circuit as hybrid architectures. Operation schemes will be validated by means of Eldo simulations and the effect of V_T MOSFET variability will be introduced, for the 6T-SRAM design, by Monte-Carlo and worst case simulations. Next, advantages in implementing a 1T-2R nonvolatile element in FPGA using our dual-layer electrolyte stack will be elucidated. The chapter ends with a review on a Stochastic Integrate and Fire neuron circuit that exploit the variability in the time required to set CBRAM devices.

Chap. 5 provides the conclusions, the highlights and the perspective of the research conducted for this thesis. A description of the on-going activities is given.

Contents

1

Introduction

This chapter introduces some of the trends in the memory research both at device level and at the circuit integration level. There are two great motivations that push the research in finding new memory materials, new storage mechanism and new stacking architectures: the filling of the latency gap in the memory hierarchy and the replacement of flash memories for the sub-20 nm node. Both stand-alone and embedded applications will be renovated by the introduction of nonvolatile, byte addressable, and in place writable Random Access Memory (RAM), also called Storage Class Memory (SCM). Moreover, because of the low voltage, CMOS compatibility, projected 10 years retention at 110 °C and good cyclability other important markets such as low power mobile can be greatly improved. A brief survey of the main emerging non volatile memories such as: Phase Change RAM (PCM), Ferroelectric RAM (FeRAM), Spin Transfer Torque RAM (STT-RAM) and Resistive RAM (ReRAM) will be presented. Next, starting from a description of embedded volatile memories, we will introduce the benefits in implementing a nonvolatile 6T-SRAM, hence explaining motivations in investigating further this architecture. Novel 3D hybrid FPGA architectures leading to better performance than that of existing planar FPGA will also be discussed. We will describe some of the constraints at memory level in implementing novel architectures in the main blocks of an FPGA, boosting our research in finding the most suitable ReRAM technology expected to minimize the power consumption. Finally, a brief introduction on stochastic neural network will be given to propel the conception of a stochastic neuron.

Table 1.1: Comparison of metrics for memories on the market.

Metric	NAND Flash	NOR Flash	PCM	SRAM	DRAM
Technology Node [nm]	22	40	45	22	32
Cell Size	4 F^2	10 F^2	5.5 F^2	120 F^2	8 F^2
Chip Size	128 Gb	8 Gb	1-8 Gb	30 MB	4 Gb
Write Bandwidth	100 MB/s	2 MB/s	9 MB/s	500 GB/s	1 GB/s
Read Bandwidth	200 MB/s	100 MB/s	266 MB/s	500 GB/s	1 GB/s
Latency	30 us	70 ns	85 ns	5 ns	20 ns
Endurance	10^4	10^5	10^9	10^{15}	10^{15}

1.1 The Memory hierarchy

In recent years, Central Processor Units (CPUs) speeds have increased significantly [1]. On the other hand, memory improvements have mostly been in density rather than transfer rates (read/write bandwidth) (Fig. 1.1 (a)-(b)). As speeds have increased, the CPU has spent an increasing amount of time waiting for data to be fetched from memory. No matter how fast a given CPU can work, in some cases it is limited to the rate of transfer allowed by the bottleneck between memories and processor (Von Neumann bottleneck). Currently, state of the art CPUs operate at 5.7 GHz [1], whereas even the fastest off-chip memories operate between 1 GHz [2] and 2.66 GHz [3]. The memory hierarchy is an arrangement of different types of memories with different capacities and operation speeds to approximate the ideal memory request of a processor in a cost-efficient way [4].

Fig. 1.1(c) shows the typical access latency (in cycles assuming a 4 GHz machine) of different memory technologies, and their relative place in the memory hierarchy. As already demonstrated in the case of Phase Change Memories (Sec. 1.2.1) new emerging memories showing access latency between DRAMs and Hard Disks are expected to bridge completely the access latency gap (Fig. 1.1(b)). The ultimate target for Von-Neumann based computing would be replacing DRAMs using nonvolatile memories with the same bandwidths (1 GB/s for read and write), compatibility with high-speed wired interface schemes (DDR), endurance of 10^{15} cycles and exhibiting capacity of TBs [5]. Table 1.1 shows in more details several figures of merit for different kind of memories that we reported in Fig. 1.1(c).

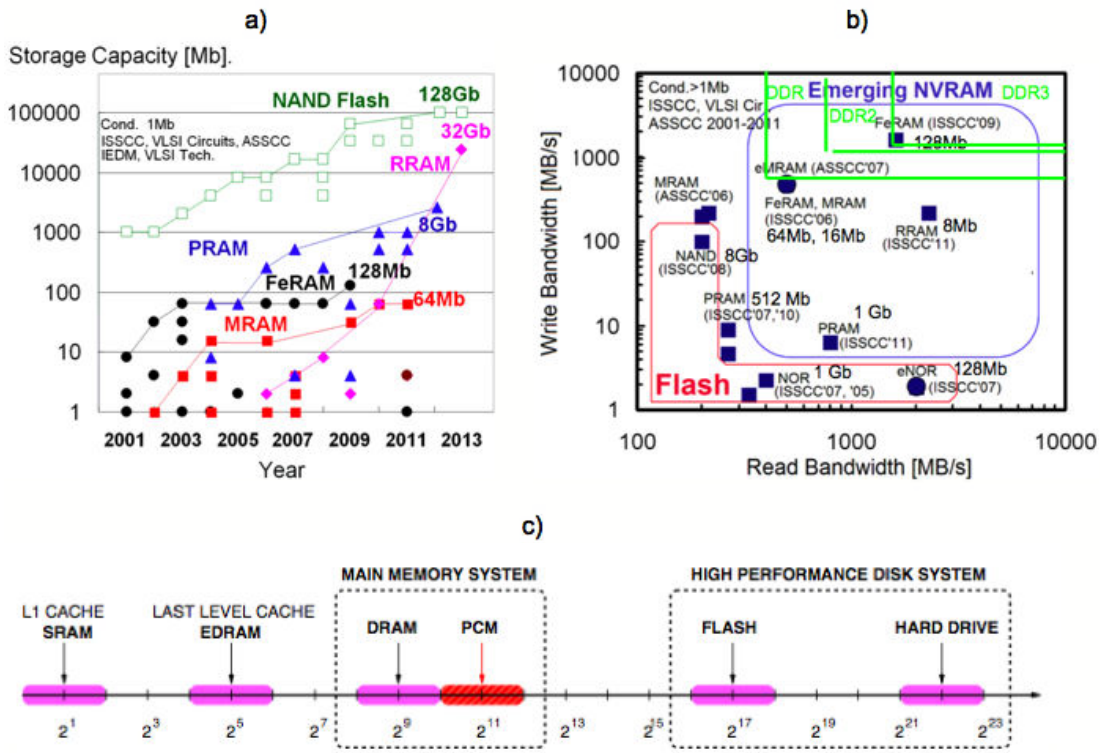


Figure 1.1: a) Memory capacity trend of different nonvolatile memory technologies. b) Comparison of different nonvolatile memory technology as a function of write and read bandwidth demonstrated at chip level. DDR, DDR2 and DDR3 are the high-speed interface scheme required for different DRAMs generation [1]. c) Access latency in terms of processor cycles for a 4 GHz processor [5]. Hybrid solutions Phase Change Memories + DRAM (two macro in a single package) have been introduced by Micron.

1.2 Overview of next generation of nonvolatile Random Access Memories (RAM)

As shown in Fig. 1.1(c) and Tab. 1.1 Flash and PCM do not fulfill all the requirements to compete with DRAMs. The main limitations are the latency for NAND and the write bandwidth for NOR and PCM. In addition the endurance is limited to 10^5 cycles for NOR, 10^4 cycles for NAND (considering the 22 nm technological node), and 10^9 for PCM, whereas DRAM requires 10^{15}. The high voltage needed for the operation of NOR and NAND memories is also a strong limitation for future integration with advanced CMOS logic in embedded systems thus making even more difficult an improvement in the bandwidth if the nonvolatile memory is used off-chip (i.e. with a serial interface)

(Sec. 1.3.2). As a result, other memories candidates are currently under investigation to provide lower latencies, higher bandwidths and compatibility with advanced (beyond 28 nm high-k MOSFETs) CMOS. This research paves the way for non volatility at several levels of the memory hierarchy and therefore provides a new abstraction on the memory systems. Integration of nonvolatile, byte addressable memories, supporting in-place writing is especially required for System on Chip (SoC) and could revolutionize consolidated architectures such as Field Programmable gate Array (FPGA) both in memory block and distributed memory cells at the interconnection level of CMOS logic. Advanced architectural studies have indicated that FPGA can be greatly improved by emerging memories distributed above the logic, as we will discuss in Sec. 1.9. Because a large number of novel hybrid (logic and SCM) architecture rely on different emerging non volatile memories, a brief survey of such technologies is presented in the next subsections.

1.2.1 Phase Change RAM

In Phase Change memory (PCM) the memory element is an alloy of $Ge_2Sb_2Te_5$, Ge_xTe_{1-x}, C doped Ge_xTe_{1-x} or N doped Ge_xTe_{1-x}. PCM utilizes the large resistivity difference between crystalline (low resistivity) and amorphous (high resistivity) phases of the phase change material. Set and reset state of PCM refers to low and high-resistance state, respectively. As fabricated, the phase change material is in the crystalline state. To reset the PCM cell into the amorphous phase, the programming region is first melted and then quenched rapidly by applying a large electrical current pulse for a short time period (Fig. 1.2 left). This operation generates a region of amorphous, highly resistive material in the PCM cell. To set the PCM cell into the crystalline phase, a current pulse is applied to anneal the programming region at a temperature between the crystallization temperature and the melting temperature for a time period long enough to crystallize (Fig. 1.2 left) [6]. Fig. 1.2 shows the current voltage I-V curve for the PCM. The set and reset states have two orders of magnitudes resistance difference for voltages below the threshold switching V_{th}. If a voltage higher than V_{th} is applied for longer than the crystallization time it leads to memory switching to low resistance state. When the PCM is in the reset state the resistance of the PCM is too high to conduct enough current to provide Joule heating to crystallize the PCM cell. In fact it is the electronic threshold switching effect that lower the resistance of the phase change

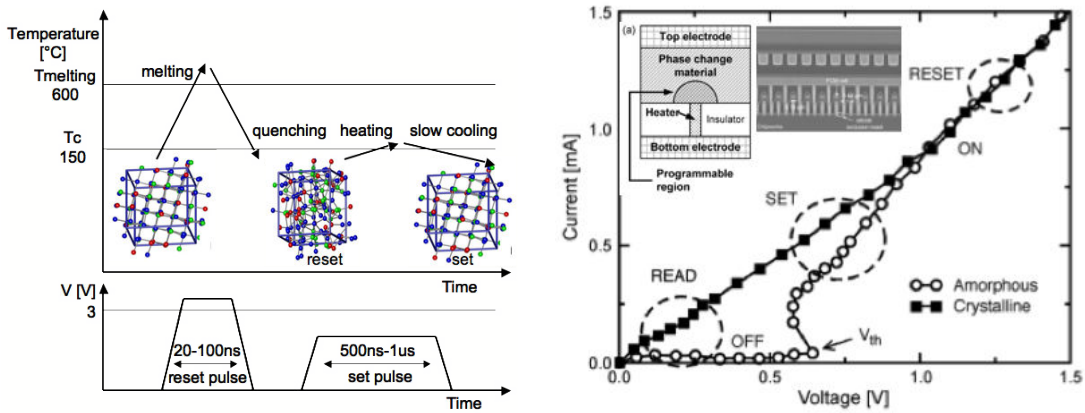


Figure 1.2: (Left) Temperature-time and voltage-time diagram to describe set and reset operations and consequences on the lattice structure of GST material. (Right) Current-Voltage characteristics of the cell. The electronic switching corresponds to the decrease of the voltage in the amorphous state and the current increase that leads to the crystallization of the amorphous region. In the inset: schematic of the PCM structure and TEM of the integrated structure.

material and enables set programming. Reset programming consumes the largest power since the cell needs to reach the melting temperature. Reset current is also determined by material properties (doping) or stoichiometry, but still is in the order of mA or hundreds of µA. Endurance of 10^9 and a projected data retention of 10 years at 85 °C have been reported [7]. PCM are fabricated with 20 nm technology with a density of 8 Gb showing 40 MB/s write bandwidth [8]. A nonvolatile 6T-SRAM was proposed in [9].

1.2.2 Ferroelectric RAM

In Ferroelectric RAM (FeRAM) the memory element consists of a ferro-electric material embedded between metallic electrodes made of Pt, Ir, or oxides of transition metals such as RuO_2 and IrO_2. The typical active ferroelectric materials are lead zirconate titanate ($PbZr_zTi_{1-x}O_3$) and strontium bismuth tantalite ($SrBi_2Ta_2O_9$) that shows two stable states of polarization corresponding to the two stable configuration of ions within the unit cell of the crystal lattice. This polarization does not vanish when the external field is removed (Fig. 1.3). Attainable cell sizes in 1T-1C configuration are in the range of 4-15 F^2. The main attractiveness of this technology is that the read and programming pulses can be less than 50 ns and it is possible to get endurance typically

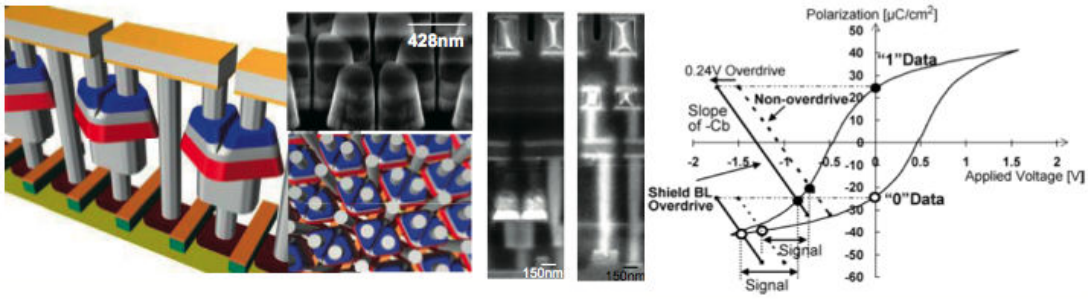


Figure 1.3: Schematic structure of FeRAM cell with capacitors connected in parallel (chain FeRAM) and TEM cross section of the integrated structure. Polarization-Voltage curve (P-V) that shows two stable states of remnant polarization ($25\,\mu\,\mathrm{C\,cm^{-2}}$) used to store the information [13], [10].

about 10^{12} cycles. Dynamic energy can be minimized till 200 fJ/bit and 1.6GB/s read/write bandwidth has been reported [10], [11]. Main limitations of FeRAM are related to the the read operation that is destructive, the scalability and the high thermal budget (600 °C - 700 °C) due to the crystallization temperature of these materials. 4 Mb embedded FeRAMs (e-FeRAMs) produced using a 180 nm process with a 1.8 V requirement have become widely deployed in IC cards and RFID tags for improving traceability and security. 16 Kb e-FeRAM for energy harvesting applications has been also developed, showing 82µA/MHz dynamic consumption (for comparison SuperFlash memories have 70µA/MHz dynamic consumption during read operation). The largest FeRAM chip was fabricated in 130 nm technology with a density of 128 Mb, and 75 ns read/write cycle [11]. A 512-byte nonvolatile SRAM using 250 nm double metal layer CMOS process with ferroelectric capacitors is described in [12].

1.2.3 Spin Transfer Torque RAM

In Spin Transfer Torque RAM the memory element is a magnetic tunnel junction (MTJ) that consists of two magnetic electrodes embedding an insulating MgO tunnel barrier (Fig. 1.4). The resistance of these magnetic tunnel junctions depends on the relative orientation of the magnetic moments in the two magnetic electrodes interfacing with the tunnel barrier. When the magnetic moments of the two magnetic layers are antiparallel, the resistance of the tunnel junction is significantly higher than when they are in parallel. A trilayer known as synthetic antiferromagnetic layer is used as reference to fix the direction of the magnetic moment. Hence, the magnetic moment orientation

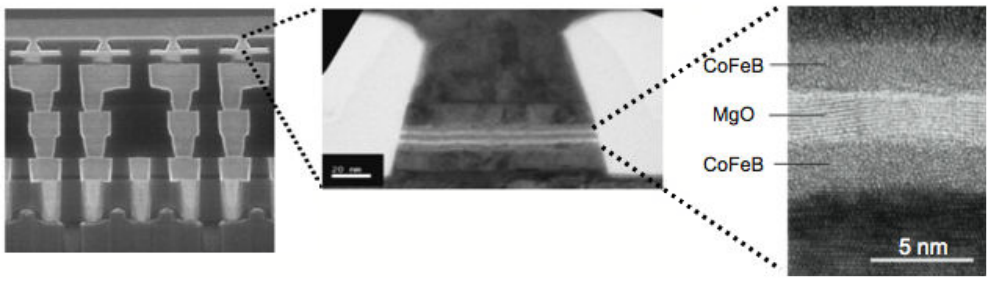


Figure 1.4: TEM cross section of the integrated MRAM structure [15].

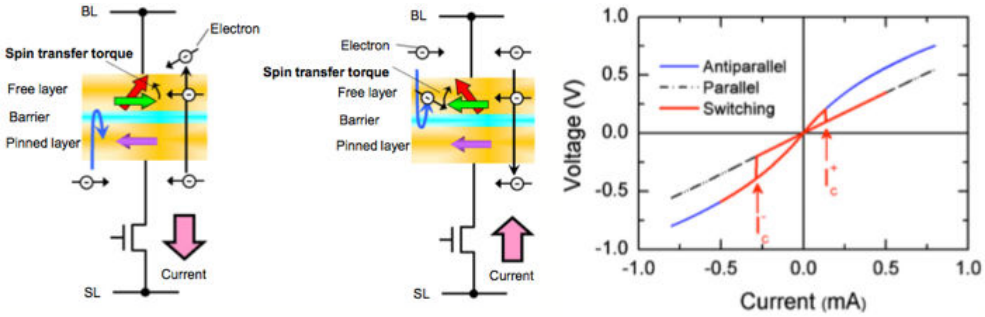


Figure 1.5: Spin transfer Torque Magnetization switching. From the left: Anti-Parallel to Parallel switching, and parallel to antiparallel switching. Example of current-induced switching. Quasi-static V-I curve show the existence of two available resistance states. The free layer of the MTJ can be switched parallel or antiparallel to the pinned layer depending on the direction of the current [15],[16].

of the storage layer will give rise to two states with distinctively different resistance values, thereby, the two states in binary bit [14]. Typical memory cells consist of two transistors connected in parallel with the MTJ to provide the current necessary for programming. The free layer of the MTJ can be switched parallel or antiparallel to the pinned layer depending on the direction of the current (Fig. 1.5). The injected current pulse is typically 200 µA in amplitude and 5 ns in duration, corresponding to a switching energy on the order of pJ [14]. The switching current density is of the order of $1\,\mathrm{M\,A\,cm^{-2}}$. The reported endurance is 10^{15}. STT-RAM suffers of poor CMOS compatibility and active layers roughness that can increase cell to cell variability. The largest STT-RAM chip was fabricated in 54 nm technology with a density of 64 Mb, employing a 2T-1MTJ cell; programming and read speeds of 20 ns were achieved [17]. Multi-Context Look-Up Table [18] and nonvolatile Flip-Flop [19] in FPGA have been proposed using TAS (Thermally Assisted Magnetic) and STT-RAM.

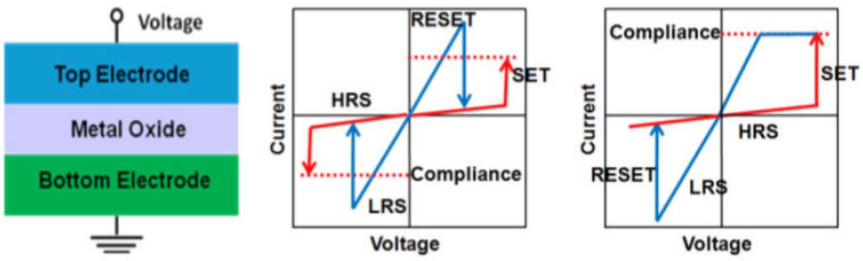


Figure 1.6: Schematic of MIM structure for metal oxide ReRAM and schematic DC I-V characteristics showing unipolar and bipolar behavior [20].

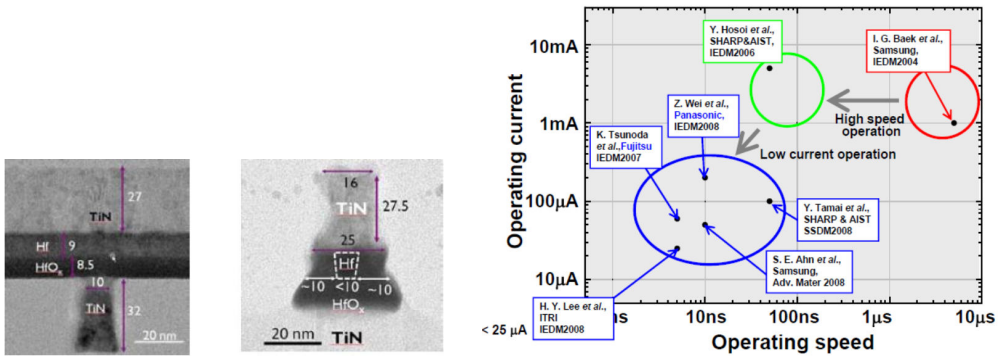


Figure 1.7: (Left) TEM of the integrated ReRAM. Both the bottom electrode and the top electrode have been scaled to 10 nm. Switching current vs operational speed.

1.2.4 Resistive RAM

In ReRAM memory elements are typically transition metal oxides embedded between two metal electrodes. The memory effect relies on the reversible transitions from high (reset) to low (set) resistive states upon the application of an electric field on the structure. Thermochemical and/or electrochemical reactions are responsible of the switching mechanism, which can be triggered by the amplitude of the applied electric field (unipolar switching mode) or by the polarity of the applied field (bipolar switching mode). In particular, in bipolar mode, set process can only occur at one polarity and reset can only occur at the reverse polarity (Fig. 1.6). The switching behavior is not only dependent on the oxide materials but it is also dependent on the choice of metal electrodes and the physics at the interfaces. Typical transition metal oxides are TaO_x, HfO_x and AlO_x that exploit the lattice sub-stoichiometry in oxygens atoms, hence the mobility of oxygen vacancies. In particular, it was observed that the low resistive state

Table 1.2: Comparison of metrics of largest chip demonstrated for various memory technologies with focus on evolution of ReRAM in 3 years. Note that the highest density has been achieved in ReRAM where the memory stack is 1D-1R where D is a bipolar diode to avoid sneak current.

Metric	**FeRAM**	**STT-RAM**	**PCM**	**ReRAM**	**ReRAM**	**ReRAM**
Year	2009	2010	2009	2010	2012	2013
Technology Node [nm]	130	54	45	130	180	24
Active Stack	SRO-PZT/Ir	CoFeB-MgO	GST	MeO_x 1D-1R	TaO_x/Ta_2O_5 1D-1R	MeO_x 1D-1R
Half-Pitch, F [nm]	225	54	52	200	N.A.	N.A.
Memory Area [µm^2]	0.252	0.041	0.015	0.168	0.144	N.A.
Cell Size	5F^2	14F^2	5.5F^2	4.2F^2	N.A.	N.A.
Chip Size	128Mb	64Mb	1Gb	64Mb	8Mb	32Gb
Write Speed	83 ns	15 ns	100-500ns	1-10ms	N.A.	230us
Read Speed	43ns	20ns	85 ns	100ms	25ns	40us
Vcc [V]	1.9	1.8	1.8	3-4	N.A.	N.A.
Reference	[11]	[17]	[7]	[26]	[27]	[23]

can be ascribed to a oxygen poor phase in the insulator (es. Magneli phase in TiO_2) that leads to d-orbitals overlap and hence metallic conduction. In this case the memory effect is classified as a valence change in the oxide and the ReRAM is usually called OxRRAM (Oxide Resistive RAM). Moreover, recently, has been shown that the oxygen vacancies concentration in the insulator can be controlled by an interfacial layer such as Ti or Zr greatly enhancing the performance of the ReRAM. In particular a switching speed of 10 ns and endurance of 10^{10} was reported in [21] using Ti/HfO_x. Hf/HfO_x based ReRAM has been demonstrated to be scalable to 10nm (Fig. 1.7) [22]. The largest chip shows a density of 32 Gb and it was fabricated in 24 nm technology [23]. Tab. 1.2 provides a comparison of the fabricated memory arrays that represents a metric to define the grade of maturity of the different technologies. In this manuscript we focus on Electrochemical Metallization Cells, also called Conductive Bridge or Programmable Metallization Cells that is a topology of bipolar ReRAM. The most frequently studied Conductive Bridge RAM (CBRAM) uses Ag or Cu as electrochemically active metals and amorphous selenides and sulfide as well as various oxides, acting as electrolytes [24]. Tab. 1.3 shows the main characteristics of fabricated arrays based on 1T-1CBRAM (1T-1R) architectures. The first 6T-NVSRAM based on Ag/$Zn_{0.4}Cd_{0.6}S$ CBRAM device was fabricated in 2006 [25].

Table 1.3: Characteristics of fabricated array with CBRAM technology.

Company	**NEC/JSTA**	**Quimonda**	**Sony**
Tested array size	1kbit	2Mbit	4kbit
Technology Node [nm]	250	90	180
Active Stack	Cu/Cu_2S	Ag/GeSe or GeS	$CuTe/GdO_x$
Memory structure	1T-1R	1T-1R	1T-1R
Programming HRS→LRS	1 V, 5-32 us	0.6 V, 10 uA, 50 ns	3 V, 110 uA, 5 ns
Erasing LRS→HRS	1.1 V, 5-32 us	0.2 V, 20 uA, 50 ns	1.7 V, 125 uA, 1 ns
Retention	3 months	10 years	10 years
Resistance Ratio	10	10^6	10^4
Operating temperature	NA	110 °C	130 °C
Endurance	10^4	10^6	10^7
Reference	[28]	[29]	[30]

1.3 Embedded nonvolatile memories

To meet the increasing demand for higher performance and lower power consumption in many different system applications, it is often required to have a large amount of embedded memory to support the need of data bandwidth in a system. The varieties of embedded memory in a given system range from static to dynamic, volatile to nonvolatile, one or many times programmable. Among various NVM technologies, floating-gate-based 1Tr-NOR flash has been the early technology choice for embedded logic applications. Along with the technology scaling several memories cells have been extensively explored, including alternative materials for the storage layers, different architectures and different programming mechanism.

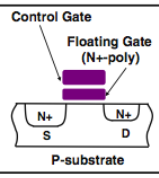
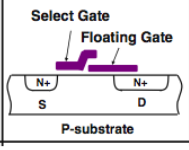
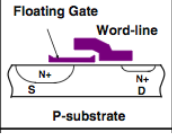
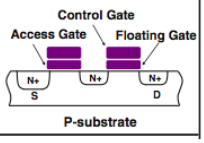
1.3.1 Mask-programmable ROM

For defined programs and data, which have no need of code-updating after manufacturing, embedded read-only-memory (ROM) provide a cost-effective and reliable solution for on-chip non volatile storage due to its small macro area, high reliability (data retention time) and process that does not require high voltage (HV) devices. There are three types of embedded mask-programmed read-only memories (ROM): NOR-ROM, NAND-ROM and flat-cell ROM. NOR-ROMs commonly appear in high-speed, small-capacity applications. NAND-ROM has the advantages of a small cell area, low power consumption, and a

small stand-by current. The flat-cell ROM is a popular solution for-low cost non high speed mid-capacity (1Mb - 32Mb) embedded applications due to its ultra-small cell area, and a logic compatible flat cell manufacturing process [31].

1.3.2 Electrically programmable FLASH

Electrically re-programmable, embedded non volatile memory technologies usually deviate from discrete memory technologies such as NOR and NAND flash memories, because of specific requirements like CMOS process full compatibility, low access time (up to 20 ns for automotive) for in place code execution, minimum area and power consumption. Moreover, the major constraint for embedded nonvolatile memory is minimum additional cost, including fewer process steps to the base CMOS logic [4]. The 1Tr-NOR cell structure has been widely used also for on chip code storage because of random access and reading latency of 70 ns with up to 10^5 times of program/erase covering different markets with different requirements from automotive to smart cards. The memory cell consists of a floating gate structure, which stores charge in the floating polysilicon gate to alter the threshold voltage of n-channel memory transistor. Nevertheless, the 1Tr-NOR floating-gate flash technology suffers from scalability issues beyond the 40 nm technological node. Innovations at the device level led to the introduction of Split gate memories 20 years ago. Fig. 1.8 shows different geometry for such memories compared to the 1Tr-NOR cell. Due to the process simplicity, high endurance (1M E/P cycles at 40 nm), and the use of low current source-side injection (SSI) (1 µA/bit and 40 p J/cell) split-gate discrete charge-trapping cell with TG-first structure are widespread used for low power applications and in smart card. Accordingly to Silicon Storage Technology (SST), 45% of all smartcard devices uses SST SuperFlash Technology. New generations of SuperFlash are expected to be scalable beyond 40 nm and compatible with LV CMOS such as FDSOI, high-k metal gate and FinFET due to the read voltage between 0.9 V and 1.2 V allowing also high performance CMOS devices to be used for the read path.

	1Tr NOR cell	1.5Tr cell (split-gate)	1.5Tr cell (SuperFlash™)	2Tr cell
Program	CHE	SSI	SSI	FN
Erase	FN (poly-sub)	–	FN (poly-poly)	FN (poly-sub)
Device Structure				

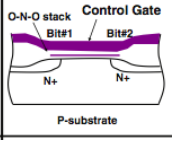
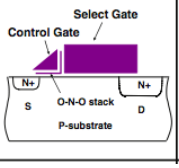
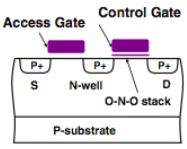
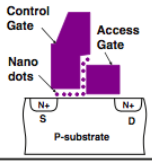
	1Tr SONOS (NROM™)	1.5Tr SONOS (split-gate)	2Tr SONOS (PMOS)	1.5Tr Nano-dot
Program	CHE	SSI	CHE	SSI
Erase	HH	HH	FN	FN
Device Structure				

Figure 1.8: Physical mechanism for program and erase and device geometry of 1Tr-NOR cell, split gates cell and Split gate memory technologies. Charge trap memories such as SONOS or nanodot based have been merged with split gate. It is presented the first generation of SuperFlash memories [4].

1.4 Embedded volatile memories

1.4.1 6T-Static Random Access Memories (6T-SRAM)

1.4.1.1 Architecture of a 6T-SRAM

The storage element of an SRAM consists of two inverters and two access NMOS. A pair of inverters is cross-coupled such that it has the output of one inverter going into the input of the other and vice-versa. The CMOS cross-coupled inverters can hold a logical 1 or a logical 0 state as long as SRAM is powered up. SRAM elements are arranged in an array of rows and columns. Each row of bit cells shares a common word-line (WL), while each column of bit cells shares a common bit-line (BL)(Fig. 1.9). SRAM is used for the register file and caches memories of all levels from first to third in the embedded memory system, providing the highest random access speed (up to 500 GB/s) and a

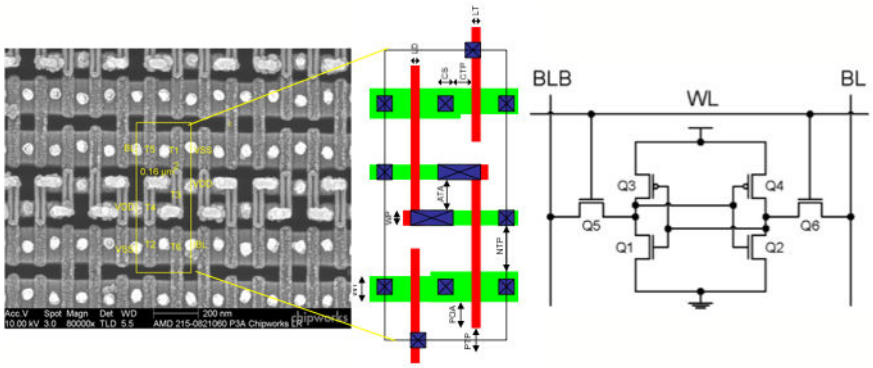


Figure 1.9: TEM image of an array of 6T-SRAM. Layout of a 6T-SRAM cell. Schematic structure of the cell. Adapted from [32].

seamless integration with logic circuits due to its compatibility of process and operating voltage. As multiple processing cores are being integrated into one chip, the demand for integrated on-chip SRAM has become even higher to provide a sufficient data stream and to keep up with an increasing demand of memory capacity and bandwidth. In particular, the total percentage of occupied SRAM area of overall chip size is increasing and has been reached 70%. In 2008, a 128 Mb SRAM was typically implemented in an area of 100 mm^2 at the 45 nm process node (Fig. 1.10). In today's high-end CPUs there are more than 30 MB total of SRAMs that occupies over 70% of the total chip area.

1.4.1.2 Scaling issues and power consumption

SRAM leakage has become very important with technology scaling for many reasons. In fact, since the geometry of the transistor keeps shrinking, higher leakage current in channel, gate and junction is measured. Leakage current can be limited using high-k MOSFET, FinFET or UTBB-FDSOI and/or reducing the nominal voltage. Nevertheless, the threshold voltage V_T variability also becomes a problem (Fig. 1.10 left down). This is mostly caused by Random Dopant Fluctuations (RDF) in the channel region, but also by oxide thickness variation, Line Edge Roughness and TiN grains with different work functions. The relation between process related parameters is typically defined as the Pelgrom coefficient (A_{VT}), which is proportional to the magnitude of threshold voltage variation. Since the σ_{VT} is also inversely proportional to the root square of the product of transistor size the SRAM becomes particularly susceptible to V_T variability as it typically uses smallest possible transistors to increase density. As a consequence both

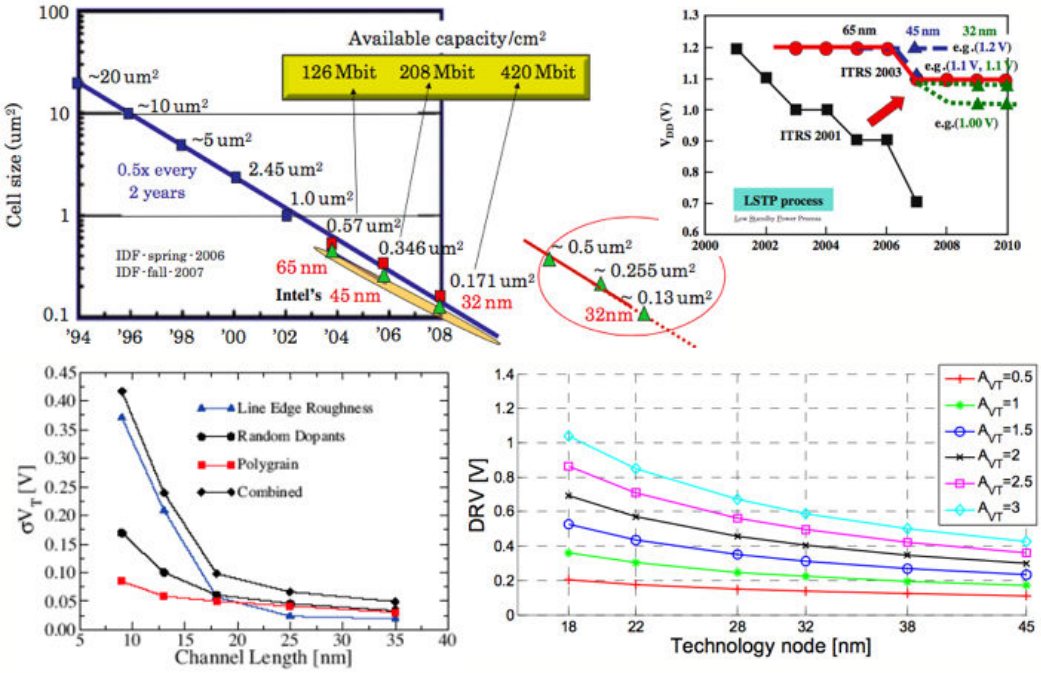


Figure 1.10: (Left up) SRAM memory cell scaling trend. (Right up) SRAM operating voltage scaling trend. (Left down) Standard deviation of threshold voltage variation vs channel length, for square planar bulk MOSFETs. Constant gate line edge roughness (4 nm) is assumed. (Right down) Data Retention Voltage (DRV) degradation vs. technology node for various Pelgrom coefficients A_{VT} values. Adapted from [32].

the nominal voltage scaling and the Data Retention Voltage (DRV) scaling are limited by stability constraints (Fig. 1.10 right down). Stability is measured in terms of the Signal to Noise Margin. Obviously, SRAM power savings are particularly important for battery-operated mobile applications. Scaling of the supply voltage greatly minimizes active and leakage power. Hence, highly energy-constrained systems, where performance requirements are secondary, benefit greatly from SRAMs that provide a DRV at the lowest possible voltage, particularly down to 0.3 V (Fig. 1.10). Moreover portable applications have relaxed workloads for the vast majority time, but can provide bursts of high performance right after the power up thus techniques such as Dynamic Voltage Scaling and ultra-dynamic voltage scaling are employed.

1.5 Embedded nonvolatile SRAM (NV-SRAM) for mobile applications

Many portable chips employ power management techniques such as dynamic voltage scaling (DVS) scheme or low supply voltage to reduce power consumption and extend battery lifetime [33]. These approaches reduce dynamic power and active-mode leakage current. Hence, low minimum operation nominal voltage single eNVM macros are needed for these DVS and low voltage devices. Unfortunately, many eNVM designs cannot achieve low V_{DD} min due to read failure noises. In particular, the current mode sensing scheme, frequently used by embedded NVM, is not suitable for ultra low V_{DD} operations. This is because the eNVM I_{cell} is small ($< 1\,\mu\mathrm{A}$) at a low V_{DD} and the current-mode sensing scheme cannot increase the sensing margin with tradeoff in speed [31]. Additionally a large program current for channel hot electrons sets an obstacle in achieving low-power program operations. As a result eNVMs have become a bottleneck in achieving a lower V_{DD} min for low voltage chips. For these reasons, serial Flash and SRAMs in a two macro approach are preferred to a single embedded NVM macro in low power mobile applications. Flash memory uses a serial interface SPI which sequentially accesses data. This solution lead to a reduction in board space, less power consumption contributing to lower overall system cost. The two macro scheme also reduces the number of NVM accesses and relaxes the endurance requirements for NVM; however, it requires long store and restore time due to serial SRAM read write and long NVM write read procedures. This results in long power off-on time thus requiring a power intensive task during the transfer operation limiting the battery life time (Fig. 1.11 left). Nonetheless, this operation implies a power gain if the stand-by operation operated at 0.2 V take an amount of time larger than 10^3 s as reported in Fig. 1.11 (right) for 65 nm CMOS technology. Moreover, since the DRV typically occurs in subthreshold, its value, among cells in the array, is highly affected by process variations (i.e. by the Pelgrom Coefficient, Fig. 1.10) that increase with scaling [4]. For these reasons a NV-SRAM that integrates SRAM cells and NVM devices within a single cell, forming a direct bit-to-bit connection in a 3D or vertical arrangement to achieve fast parallel data transfer and fast power on/off speed has been proposed. Various types of NVM devices, such as SONOS [34], FeRAM [12], MRAM [35], and ReRAM [33], have been employed in NV-SRAM macro so far, but the use of ReRAM is the most promising. In particular, a 16Kb HfO_2

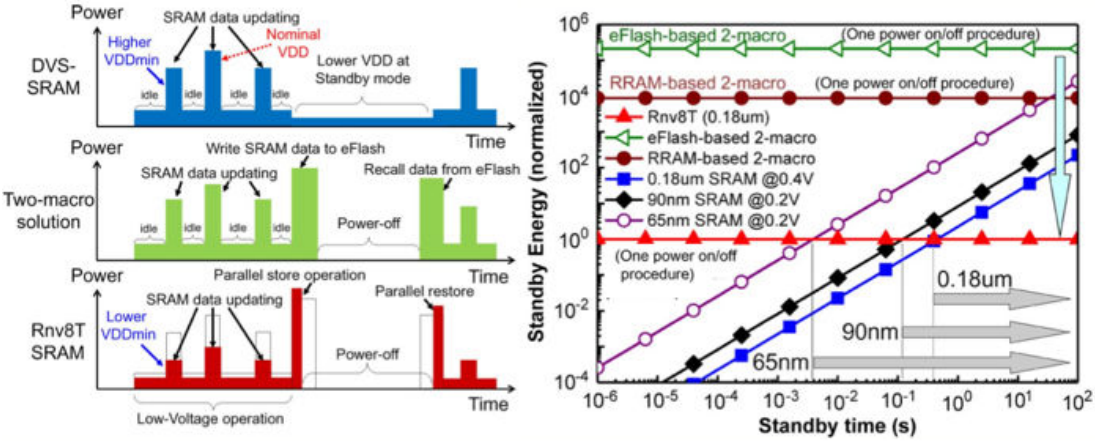


Figure 1.11: (Left) Power consumption of DVS-SRAM, two macro solution, and NV-SRAM (Rnv8TSRAM) during active and stand-by modes. (Right) Comparison of standby mode energy consumption [33].

based OxRRAM chip demonstrated that power off-on procedure (including store and restore operations) contribute to power gain, if the corresponding stand-by operation take an amount of time larger than 2 ms. This feature mainly depends on the switching energy to set/reset the non volatile device and hence on the inherent logic (i.e the transistors in the latch). The HfO_2 based OxRRAM used in this NV-SRAM shows 25 µA set/reset current, a set time of 1 ns at 3 V with a resistance ratio $R_{\text{off}}/R_{\text{on}}$ of only 5 and a switching energy of 0.1 p J [33].

1.6 Field Programmable Gate Array (FPGA)

A Field Programmable Gate Array (FPGA) consists of programmable logic resources embedded in an array of programmable interconnections. The programmable logic resources can be programmed to implement any logic function, while the interconnects provide the flexibility to connect any signal in the design to any logic resource. The memory technologies for the logic and interconnect resources can be: static random access memory (SRAM), flash or antifuse. SRAM-based FPGAs offer reconfigurability at the expense of being volatile, while antifuse are One Time Programmable (OTP) devices. Flash-based FPGAs provide an intermediate alternative by providing reconfigurability as well as non volatility at the cost of scalability issues [36].

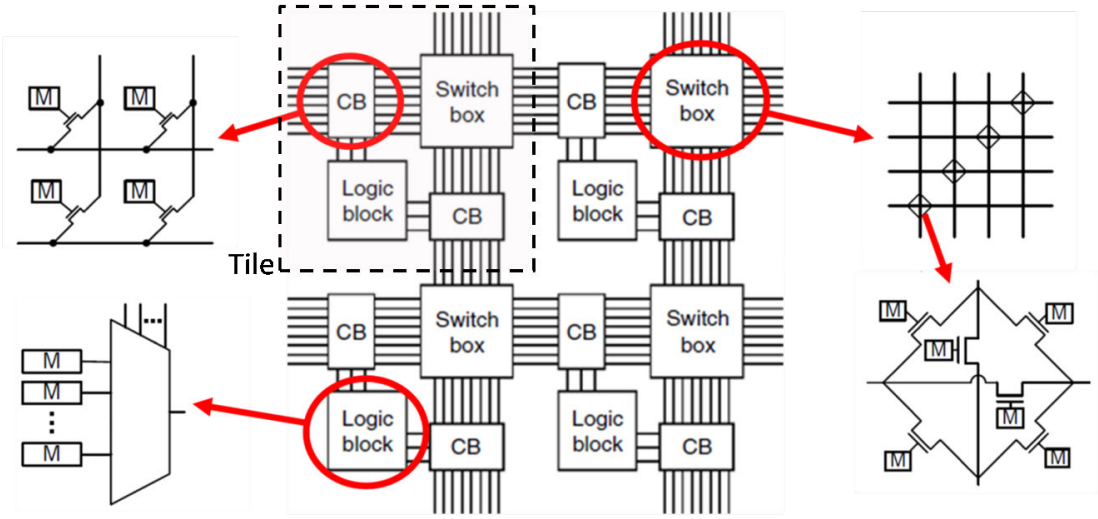


Figure 1.12: Structure of an island-type FPGA to show the periodic fabric. 6T-SRAM (M) controls both pass gates in Connection Blocks (CB) and Switching Blocks to define the general routing. Look up tables implemented with memories and multiplexer are used to store the result of a defined function.

FPGAs circuits are composed of switching blocks (SB), connection blocks (CB), logic blocks (LB) and embedded memories (Block memory). Every LB is formed by a group of basic logic elements (BLE). The most simple BLE includes a Look-up table (LUT) a D flip-flop (register) and a multiplexer.

A look-up table implements a truth table, that can be considered a function generator: in particular the truth table for a K-input logic function is stored in an array of dimension $2^k \times 1$ constituted by 6T-SRAM cells or by Flash cells or antifuse cells. Considering the truth table of the logic function f = ab+c; if this logic function is implemented using a three-input LUT, then the SRAM would have a 0 stored at address 000, a 1 at 001 and so on, as specified by the truth table; clearly specific inputs are used to select the corresponding memory cell to define the output. During the run time the LUT outputs the result of the function; the D flip-flop saves the result temporarily and synchronizes it with the global clock.

The conventional structure of a D Flip-Flop includes two elements: Master and slave, which are both clock-controlled Latch. The master part is used to write the information in Flip-Flop and the slave part is used to output the information toward a multiplexer.

To build large logic structures, FPGAs use vertical and horizontal routing signals

in a matrix arrangement that are paired with switch boxes at intersections to support FPGA element interconnection. Switch boxes are located at the intersection of rows and columns. A switch box is used to route between the inputs/outputs of a logic block to the general on-chip routing network. The switch box is also responsible for passing or stopping signals from wire segment to wire segment. The wire segments can be short (span a couple of logic blocks) or long (run the length of the chip). All interconnections on FPGAs are active: pass transistors behave as switches thus the configuration memory controlling the pass transistors is distributed over the chip. Hence, every time a signal crosses a switch it suffers of the RC delay. This adds up quickly on long routes. Thus in FPGA there is a fundamental trade off in the amount of flexibility, speed, area and power consumption.

1.7 6T-SRAM based FPGA

In 6T-SRAM based FPGA, 6T-SRAM cells are used as distributed configuration bits. Since 6T-SRAMs can be reprogrammed many times (10^{15}) this architecture gives the highest grade of flexibility among FPGA types, at the cost of volatility. In fact, SRAM based FPGAs have to be configured every time the system is powered up. Configuration is the process of downloading configuration data into an FPGA using an external, usually on board, source such as a flash memory device.

The configuration bitstream is loaded at every power-up into the device through a configuration interface that links the external nonvolatile device with the FPGA. Modern FPGAs such as Xilinx 7 series offers both serial and parallel configurations. The highest throughput and speed is achieved with a parallel scheme, while the serial interface reduces the number of pins of the flash device and surface on the board. The Master Byte Peripheral Interface (BPI) flash configuration mode with a 16x data bus is widespread used as a parallel interface.

For this programming solution, the user configures the FPGA with a bitstream that serves as a bridge between the JTAG bus and the parallel NOR flash bus interface. This solution is referred to as indirect programming because the flash is not directly programmed but is programmed through the FPGA itself. This programming scheme is also used by ALTERA [37]. Using a synchronous read of the NOR flash at 100MHz and an external oscillator at 80MHz the configuration time for a 16Mb bitstream loaded

from a NOR Flash Memory to the FPGA is 126 ms [38]. This time is quite high for application that requires instantly on capabilities.

The maximum FLASH NOR density used by Xilinx in 7 series is 1Gb while ALTERA provides also external NAND Flash memories to achieve very high capacity (up to 16Gb).

At the board level (i.e not integrated on the FPGA architecture) emerging technologies have been introduced: Xilinx Nexys3 development board uses both 16 MB Parallel and Quad-mode SPI PCM.

1.7.1 Power consumption in 6T-SRAM based FPGA

As the technology is scaled down, the oxide thickness t_{ox} in MOSFET is also shrunk. The scaling down of t_{ox} results in an exponential increase in the gate oxide leakage current. Moreover, to maintain the switching speed improvement of the scaled CMOS devices, the threshold voltage V_T of the devices is reduced to keep a constant device overdrive.

Decreasing V_T results in an exponential increase in the subthreshold leakage current. As a result of the continuous scaling of t_{ox} and V_{th}, the contribution of the total leakage power to the total chip power dissipation is increasing even if the nominal voltage is decreasing.

The average leakage power dissipation of a typical 90-nm SRAM based FPGA at 25°C and 85°C, was calculated to be 4.25 µW and 18.9 µW respectively [36], for a utilization of a single CLB of the 75%. Hence, the leakage power dissipation for a 1000-CLB FPGA would be in the range of 4.2 mW. If these FPGAs have to be used in a mobile application, which has a typical leakage current of 300 µA then the maximum number of CLBs that can be used would be 86 CLBs for the 25°C and only 20 CLBs at 85°C.

Furthermore in [36], it was reported that for a 50% CLB utilization, 56% of the leakage power was consumed in the unused part of the FPGA. The consumption in the unused part of the FPGA is a recurrent problem in such type of architecture, that is the cost of flexibility. In fact almost 40% of the FPGA consumes stand-by leakage current without delivering useful output. Hence solutions to suppress leakage currents are strongly required.

Moreover, FPGAs used in wireless applications can go into idle mode for long periods of time [36]. In such designs, even the utilized resources need to be forced into a low-power (standby) mode during their idle periods to save leakage power. One of the most popular techniques used in leakage power reduction is multi-threshold CMOS (MTCMOS) [36]. In an MTCMOS implementation, a high-V_T MOSFET called the sleep transistor connects the pull-down network using low-V_T devices of a circuit to the ground. When the sleep transistor is turned OFF, the circuit subthreshold leakage current is limited to that of the sleep transistor, which is significantly low. Hence, the circuit benefits from the high performance of the low-V_T pull-down network when the sleep transistor is turned ON, while limiting the circuit sub-threshold leakage current when the sleep transistor is turned OFF. This approach is utilized for CLBs but can not be used for the configuration bits in the programmable interconnect.

Moreover as reported in [39], 38% of the total leakage power in 1.2 V SRAM-based FPGAs built in 90 nm CMOS process is due to the configuration bits distributed in the FPGA and continuously read during FPGA run time. Obviously, inside the routing resources the SRAM cells are designed using high-V_T to minimize the leakage that can vary between 10 and 100 pA for a nominal voltage of 1.2 V on every single 6T-SRAM, thus there is a consistent research on the minimization of leakage at the device level and/or at the architecture level.

For example, at the device level the introduction of low power FinFET technology at the 16 nm technology node is foreseen by Xilinx for the 2014. FinFET, that replaces high-k 28nm CMOS technology in 6 series Xilinx, should enable a consistent diffusion in the mobile market, such as portable ultrasound equipment consuming less than 2 watts. Power consumption can be ascribed to five main components: Static power (leakage), Dynamic Power (frequency dependent), Power-up (inrush power), Configuration power, sleep mode power and can be calculated as:

$$\begin{aligned} P_{tot} =& P_{logic,dyn} + P_{mem,dyn} + P_{net,dyn} + P_{clk,dyn} = \\ =& P_{logic,dyn} + P_{mem,dyn} + \beta V_{DD}^2 f_{net} \sum_{allnets} C_{net} + C_{clk} V_{DD}^2 f_{clk} + I_{leak} V_{DD} \end{aligned} \tag{1.1}$$

where $P_{\text{logic,dyn}}$ is the dynamic energy consumed in the logic circuits (including the non configurable blocks), $P_{\text{mem,dyn}}$ in the memory elements, $P_{\text{net,dyn}}$ in the interconnection wires and $P_{\text{clk,dyn}}$ in the clock network. The ratio between $P_{\text{logic,dyn}}$, $P_{\text{mem,dyn}}$,

$P_{\text{net,dyn}}$, $P_{\text{clk,dyn}}$ is estimated to be 0.3 : 0.15 : 0.4 : 0.15 using Xilinx X-Power Tool [40]. Hence, the reduction of $P_{\text{net,dyn}}$ should be targeted with the shrinking of the FPGA architecture by using new emerging integration schemes (such as 3D) that uses new emerging memories fully compatible with CMOS process flow.

The advantage in using 3D integration is the reduction of C_{net} and C_{clk} because memory can be stacked in the third dimension, thus achieving a smaller die area reducing the power consumption or improving the circuit speed at the same power. For example, in monolithical stacking, electronic components and their connections are lithographically built in layers on a single wafer.

The predicted advantage in using emerging memories is still a reduction of C_{net} and C_{clk} but also the suppression of the static power when the circuit is IDLE, with data kept in the SRAMs applying the data retention voltage (strongly dependent on the technological node and Pelgrom coefficient). Keep the information in volatile SRAM/DRAM is essential to have a fast loading in the CPU of the last context saved in the memory. Thus, providing a massively parallel and fast bit transfer from nonvolatile memories to the logic, every IDLE operation can be avoided and substituted with a power off thus potentially reducing the power consumption. We used the term potentially because also the contribution of static power consumption during circuit activity should be analyzed and is strongly dependent on the architecture implemented for the configuration bits as will be explained in Sec. 1.9.

1.8 FPGA in mobile applications

In section 1.7.1 we have claimed that historically FPGA are not suitable for mobile applications, because of the very large power consumption and area. In fact, aggressive mobile applications require an on-chip nonvolatile memories with ideally zero stand-by current, minimum area and scalable with the same trend as the MOSFETs. Moreover a sufficient bandwidth to support execute in place operation thus avoiding shadowing the context in to DRAMs and enough storage capacity is also required. In the last years, one solution for mobile applications was provided by antifuse FPGA. Unlike SRAM-based devices, which are programmed by the user himself, antifuse-based devices are programmed directly by the foundry. In the 2T architecture used in Antifuse based FPGA the circuit is in high resistive state to begin with and is programmed by

applying electrical stress that creates a low resistance conductive path. These FPGAs are nonvolatile, do not require external Flash and are radiation tolerant. State of the art ultra power 40 nm FPGA, proposed by Lattice Semiconductor (SiliconBlue), consumes 42 µW in stand-by at the nominal voltage of 1.2 V. Dissipation is mainly due to the 640 logic cells (i.e. BLE 4 input LUT and Flip-Flop) and 32Kb of volatile RAM [41]. A One Time Programmable (OTP) based memory is used inside this FPGA. OTP antifuse non volatile memories are scalable till 20 nm, show densities up to 32 Mb and depending on the capacity and the size of the boot code multiple programming cycles can be obtained. In fact, though Antifuse is an OTP, it can emulate a Multiple Time Programmable MTP for a few cycles of endurance. In fact, 1Mb memory can be chosen if 128Kb of code may need to be updated up to 8 times in the field [42]. The area of an OTP memory is 0.8 mm^2 for 1 Mb density. State of the art OTP memories are slower than SRAM, DRAM, or even plain old ROM, but they are usually fast enough to allow firmware to execute in place (XIP), without copying the code into DRAM and in some cases even skip the SRAM based L2 Cache. Memory array connects to the system over a 32-bit parallel interface that runs synchronously with the on-chip bus, so it's four times faster than quad serial flash memory or serial EEPROM, giving 40 ns as random access time. Besides the low endurance, the OTP requires an internal charge pumping thus increasing the area consumption.

Flash-based FPGAs are similar to their SRAM counterparts in that their configuration cells are connected together in a long shift-register-style chain. These devices can be configured off line using a device programmer or by the user. State of the art Flash based FPGA are commercialized by Microsemi. In IGLOO nano FPGA a 130 nm, 7 metal layers, Flash-based CMOS process is used. This FPGA consumes only 5 µW in stand-by at the nominal voltage of 1.2 V with 10k LUT and FF [43]. Microsemi flash-based IGLOO nano allows 1us instant on. This feature helps in system component initialization, execution of critical tasks before the processor wakes up, setup and configuration of memory blocks, clock generation, and bus activity management. Others solutions are instead optimized to operate reactively or periodically. Rapid stopping and starting of the FPGA fabric and related IOs while preserving the state of the FPGA fabric is possible with Flash-based FPGA. In State of the art technologies 100 us are requested to enter in or exit from the sleep mode, moreover only 1 mW in sleep mode and 10 mW static power dissipation during operation of 50 K LUT and FF are

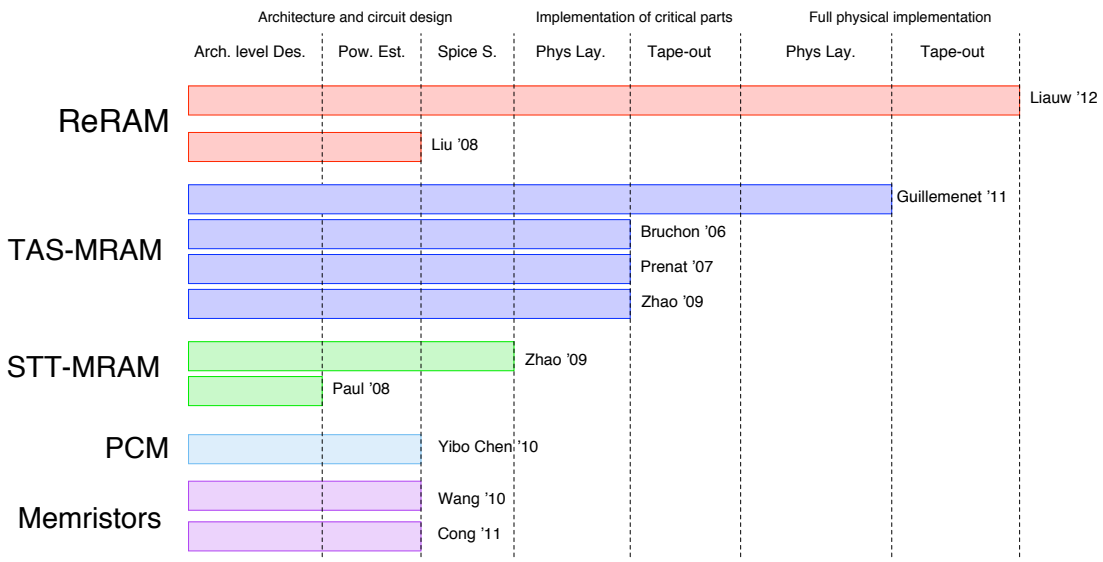


Figure 1.13: Grade of maturity in the process flow for many hybrid NVM FPGA with different technologies (adapted from [46]).

achieved. The main concerns related to the integration of Flash on chip are related both to the Flash cells that require high quality tunnel oxide, interpoly dielectric and two polisilicon layers (control gate and floating gate) that require additional thermal budget. Moreover high voltage transistors are required. Hybrid technology: SRAM and Flash have also been fabricated. State of the art Lattice XP2 FPGA offers instant-on and non volatility of Flash and the reconfigurability of SRAM in one chip [44]. Densities from 5K to 40K with 4-input Look-up Tables (LUTs) and instant on (1 ms) are achieved. Update logic configuration while equipment continues to operate is possible, thus giving the flexibility to load a new bitstream when the processor is running. This FPGA consumes 102 µW in stand-by at the nominal voltage of 1.2 V. Usually the Flash architecture is a NOR but also NAND is available [45].

1.9 Hybrid architectures: FPGA and ReRAM

The improvement of FPGA performances by using nonvolatile emerging memories is gaining importance in recent years and can be considered one of the key features of this thesis. The possibility of integrating memory and logic in a distributed way without silicon area overhead will give rise to disruptive hybrid architectures both inside the

FPGA BLE (such as: nonvolatile Flip-Flop [47] [18] or nonvolatile Look-up-Table) and/or in the routing resources (SBs or CBs). By stacking emerging memories with CMOS devices, new routing switches can be achieved to reduce FPGA area, remove external Flash, thus reducing board area and obtain instantly-on capability. As a consequence we expect to reduce the power consumption related to the in-rush and SRAM reconfiguration after power up, and dynamic power consumption due to the area shrinking (Sec. 1.7.1). In [48] it is estimated that a 3D-FPGA with the configuration memory stacked on top of FPGA logic and routing can achieve 57% smaller area than a baseline 2D-FPGA in 65nm CMOS technology. It is shown that the size of the configuration memory cell plays a key role in the degree of performance improvement achieved by a monolithically stacked 3-D FPGA. For a memory cell that is <0.7 the area of an SRAM cell, a 3D-FPGA can achieve 3.2 times higher logic density, 1.7 times lower critical path delay, and 1.7 times lower total dynamic power consumption than the baseline 2-D FPGA at the 65-nm technology node. The size of the configuration memory is determined by the design rules of the specific process design kit and usually is larger than the physical dimensions of the plug or the via between metal lines (Fig. 4.9 left). Fig. 1.13 shows the grade of maturity of hybrid FPGA architectures. A full physical implementation was demonstrated in [49], using a novel architecture to replace 6T-SRAM based configuration cells that will be explained in Sec. 1.9.2.

1.9.1 ReRAM in switching blocks: 2T-1ReRAM architecture

As described in Sec. 1.6 currently used CMOS routing switch consist of a pass transistor controlled by a 6T-SRAM cell to provide the routing function. The 2T-1R switch utilizes a ReRAM cell in the signal path removing both the 6T-SRAM and the pass gate. The main advantage of this solution is the area gain, from $120F^2$ to $4F^2$ and the reduced RC if the ReRAM resistance in the low resistive state is below the on resistance of the pass gate (1 kΩ). Nonetheless, speed gain saturation is obtained since repeater buffers for driving interconnect circuit have on-resistance of order of 1 kΩ thus limiting the delay reduction [50]. The architecture that has been introduced using the 2T-1R structure is called crossbar. This architecture is composed of a bistable memory element (typically a ReRAM, Sec 1.2.4) embedded by two sets of parallel conductive interconnects crossing perpendicularly (Fig. 1.14). The crossbar architecture is programmed by applying voltage pulses to horizontal and vertical lines. In Fig. 1.14

switch S_{00} can be programmed to be on by applying V_{PP} to X_0, grounding Y_0 and applying $V_{\mathrm{PP}}/2$ to the other lines, where V_{PP} is the programming pulse for the ReRAM and has to be greater than the threshold to have the switching for a defined pulse width. $V_{\mathrm{PP}}/2$ is less than the threshold. Since the voltage difference between the two terminals of S_{00} would become V_{PP}, S_{00} will be turned on. With respect to all the other switches, since the voltage difference for them would be $V_{\mathrm{PP}}/2$ or zero their states will not change [28]. ReRAMs that do not show high non linearity between the off and the on state can be used to implement such architecture if the crossbar does not exceed a specific density (i.e the power consumption due to the leakage in the off path is acceptable). For instance a 32×32 crossbar has been fabricated leading to a 72% reduction in chip area [51]. Moreover, the programming conditions used to switch the ReRAM must be carefully chosen, because the amount of current that can flow after the programming time and before the end of the pulse can lead to an important power consumption. The choice of the programming conditions is even more difficult because the time required to program ReRAM devices usually follows a statistical distribution, thus the pulse width should cover the worst case with consequences on the power consumption ([28] and Sec. 2.3). Finally, it is worth to note that with this solution the programming current and the signal trasmission current share the same path. Thus, during the programming stage, the programming voltage propagates to the other circuits connected to the routing switch, which might cause degradations and failures [52]. Some studies [53], [54] propose also the use of crossbar arrays for high density storage applications without integrated bipolar diode. In this case several constraints relies on the switching device that must exhibits R_{on} higher than MΩ, non linearity, and a sufficient $R_{\mathrm{off}}/R_{\mathrm{on}}$ ratio.

1.9.2 ReRAM in switching and logic blocks: 1T-2ReRAM architecture

In [47], [49], [55] the 6T-SRAM cell is replaced by a structure consisting of one transistor and two ReRAM cells in a voltage divider configuration (1T-2R NVE) to control the pass gate or to store the data in a Look-Up Table (Fig. 1.15). Compared to the 6T-SRAM routing switch, 1T-2R NVE switch leads to a density enhancement, since the two ReRAM cells can be easily integrated between two metal levels in a standard CMOS process flow. In [49] it was demonstrated that a 3D-FPGA with stacked configuration memory, based on ReRAM technology, can achieve up to 40% smaller die area and

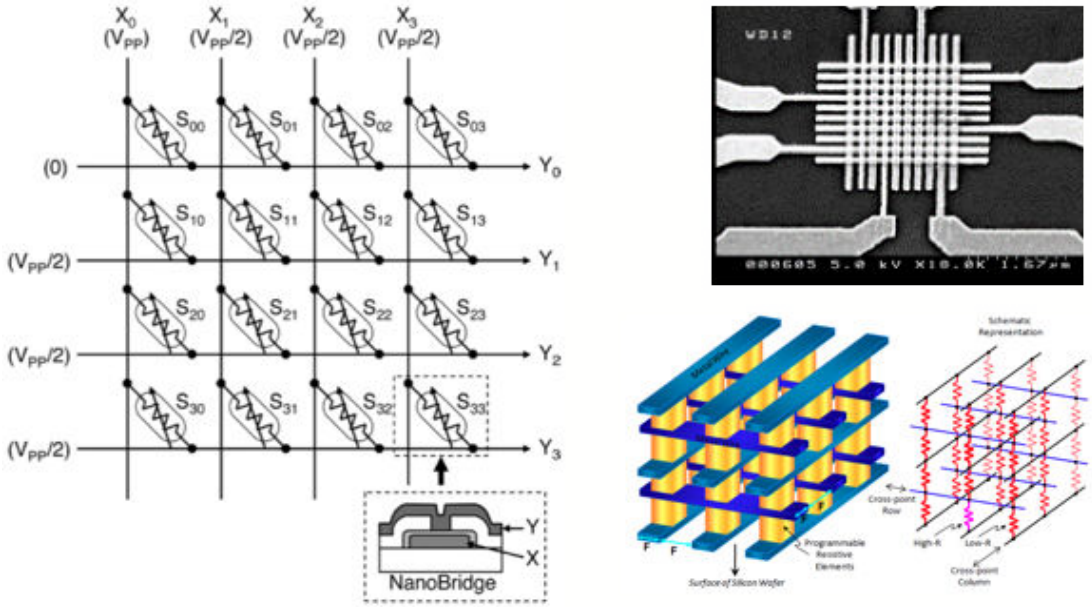


Figure 1.14: (Left) Crossbar switch [28]. (Right) TEM image and schematic of a 48nm × 48nm crossbar architecture.

up to 28% lower energy delay product than a baseline 2D-FPGA. Smaller area also allows a reduction of the interconnection capacitance thus reducing the dynamic power consumption and enabling same operational speeds at lower power [56]. Although the 1T-2R NVE solution eliminates stand-by power consumption, the leakage current through the ReRAM during run time (i.e. in continuous read operation) depends on the resistance of the high resistive state. Maximizing the high resistive state is essential to reduce the static power consumption during FPGA run time. Hence, materials engineering and/or specific programming conditions (compatible with logic) are required to provide a competitive solution with respect to SRAM based FPGA. In particular in [49] an optimized ReRAM stack composed of nitrogen-doped AlO_x was adopted showing sub-µA programming currents, 10 years retention at 125 °C, 10^5 switching cycles, and an R_{off} of 1GΩ (Fig. 4.13). Our research to satisfy the aggressive requirement on R_{off} will be discussed in Chap. 3. In 6T-SRAM leakage current strongly depend on V_{th}, the oxide thickness, and the feature size and vary in the range of pAs [57], [58]. This implies that an R_{off} value higher than $10^{12}\,\Omega$ at a read voltage of 1 V should be targeted to reduce the power consumption during FPGA run time.

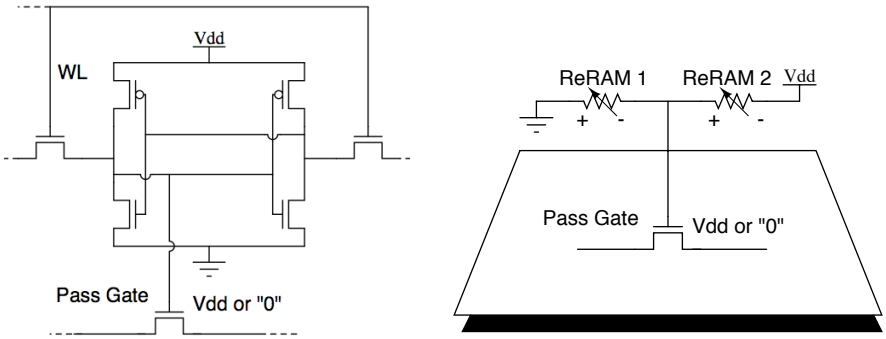


Figure 1.15: Configuration memory on a pass gate to establish connectivity between Logic blocks. (Left) 6T-SRAM based. (Right) 1T-2R based with Nonvolatile Voltage divider Element (NVE) using bipolar ReRAMs.

1.10 Main concepts on stochastic neural networks

In Sec. 1.1 we emphasized the poor efficiency of Von Neumann architecture because of the sequential processing of fetch, decode and execute instructions that relies on the read/write bandwidth between the processor and the memories. Not surprisingly, brains of biological creatures are configured differently from the Von Neumann architecture. The key to the high efficiency of biological systems is the large connectivity between neurons that offers highly parallel processing power. The brain is thus a large neural network. A neural network is composed of neurons and synapses. Neurons are the basic processing units of the brain (Fig. 1.16). Each neuron receives electrical inputs from about 1000 other neurons. Impulses arriving simultaneously are added together and if sufficiently strong (i.e above a defined threshold) lead to the generation of an action potential (i.e. neuron spike). Neurons communicate through structure called synapses in a process called synaptic transmission. When an action potential reaches a synapse, pores in the cell membrane are opened allowing an influx of calcium ions into the pre-synaptic terminal. This causes a small packet of a chemical neurotransmitter to be released into a small gap between the two cells. The neurotransmitter interacts with receptors that are embedded in the post synaptic membrane. These receptors are ion channels that allow certain types of ions to pass through a pore within their structure. The pore is opened following interaction with the neurotransmitter allowing an influx of ions into the post-synaptic terminal. It has been shown that a single voltage-dependent ion channel exhibits probabilistic gating which is recorded as random opening and

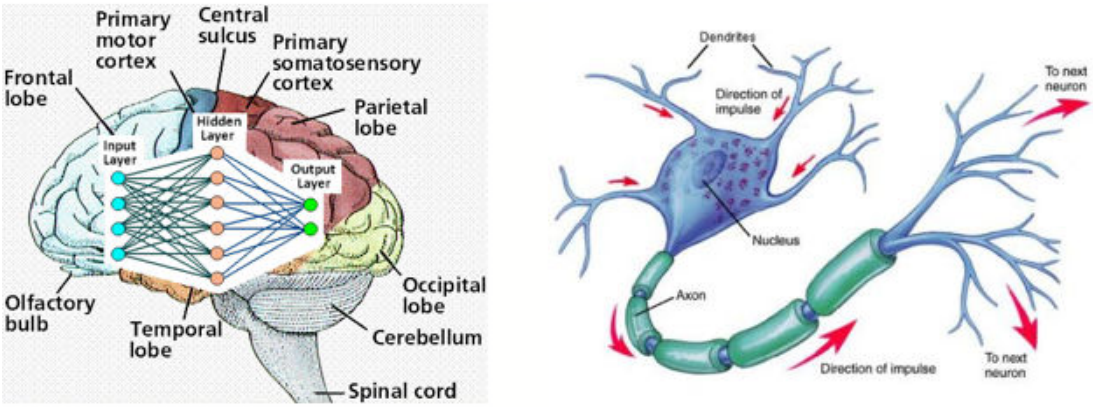


Figure 1.16: Human brain can be considered as a large neural network. Circle and connections represent the neurons and the synapses respectively. Sketch of a biological neuron showing the nucleus and the axon.

closing of an ion channel (Fig. 1.17). The opening of a single ion channel is capable of triggering action potentials spike. Considering that only 100 ion channels are open at the same time the stochastic behaviour of a single ion channel adds noise to the total membrane current of the neuron and can change the transmembrane voltage dynamics at and close to the threshold of firing [59]. This stochastic component motivates our research in designing a neuron intrinsically stochastic as we will discover in Sec. 4.6.

Neuromorphic hardware research has gained a lot of importance in recent years due to its promising low-power, fault-tolerant, and ultra-adaptative computing paradigms [60],[61],[62],[63],[64]. Neural networks are employed to classify patterns based on learning from examples. Learning rules define and allow a modification of the synaptic conductance on which relies the memory effect in biological brains. Long Term Potentiation (LTP)/Long Term Depression (LTD) rules define an enhancement/depression in signal transmissions between two neurons. Different neural network paradigms employ different learning rules, but all in some way determine pattern statistics from a set of training samples and then classify new patterns on the basis of these statistics. Current methods such as back propagation use heuristic approaches to discover the underlying class statistics. The heuristic approach usually involve many small modifications to the system parameters that gradually improve system performance. Besides requiring long computation times for training, the incremental adaption approach of back-propagation is susceptible to false minima [65]. To improve this approach, many algorithms exploits

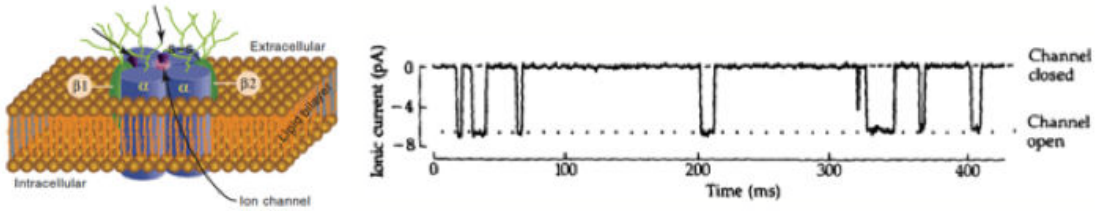


Figure 1.17: Ion channel and patch clamp recordings of single channel activity: the current flow through an ion channel over time shows a stochastic behaviour.

random numbers to improve learning. In particular, literature in the fields of neural networks [66],[67] and of biology [68] suggests that in many situations, actually providing a certain degree of stochastic, noisy or probabilistic behavior in their building blocks may enhance the capability and stability of neuroinspired systems. Some kind of neural networks even fundamentally rely on stochastic neurons, like Boltzmann machines [69].

In neuromorphic hardware, providing stochastic behavior to neurons using pseudo-random number generators or thermal noise amplifiers will lead to significant overheads. This explains interest in developing silicon neurons with an intrinsic stochastic behavior, but which may be controlled.

1.11 Conclusions

In this chapter, we reviewed emerging storage class memories such as PCM, FeRAM, STT-RAM and ReRAM as possible candidates both to fill the latency gap in the memory hierarchy and to introduce new hybrid architectures to improve, for example, existing FPGA embodiments. Concerning embedded volatile memories (i.e. 6T-SRAM) we emphasized the problems related to the scaling of the voltage supply, the power consumption due to leakage and variability due to advanced nodes. Advantages in implementing a nonvolatile SRAM based on ReRAM were discussed as a background to start our analysis on this architecture both related to the logic and to the switching properties of ReRAM devices. We also provided an overview of several types of FPGA, to identify possible paths of innovations on some blocks of the FPGA leading to area and power gain. The implementation of new architectures such as 1T-2R or 2T-1R was critically discussed to outline the constraints at memory level, hence define a strategy for developing new CBRAM stacks that could satisfy these requirements. Finally,

motivations and advantages in using stochasticity in hardware neural network were briefly introduced in view of developing an extremely compact hybrid neuron that exploits unavoidable variability in CBRAM devices.

2

Characterization and modeling of Ag-GeS_2 based CBRAM devices

In this chapter, we introduce the empirical model used to explain and predict the main switching parameters in CBRAM devices measured by electrical characterization. To this aim a thermally activated hopping model is chosen to describe the ion migration and the consequent filament growth/dissolution during set/reset process. The dependence of R_{on} resistance (LRS) and the reset current on the compliance current is also take into account. Parameters of the equations were mainly extracted by fitting of electrical characteristics obtained both by DC quasi-static measurements and by pulse measurements. Temperature effects will be also considered. Next, results of the electrical tests on 8×8 1T-1R NOR memory array will be introduced.

A statistical analysis on the correlation between the programming conditions and the percentage of CBRAM devices in the memory array that reversibly switch from high resistive state (HRS) to low resistive state (LRS) will be performed. The cell to cell and cycle to cycle variability of the switching parameters observed among different CBRAM will be empirically explained. We conclude the chapter with a comparison between the electrical performance of W-GeS_2-Ag and Ta-GeS_2-Ag based CBRAM to gain new insights on the role of the inert electrode that rules the electron transfer and the nucleation of the Ag phase leading to the filament growth and the switching from HRS to LRS.

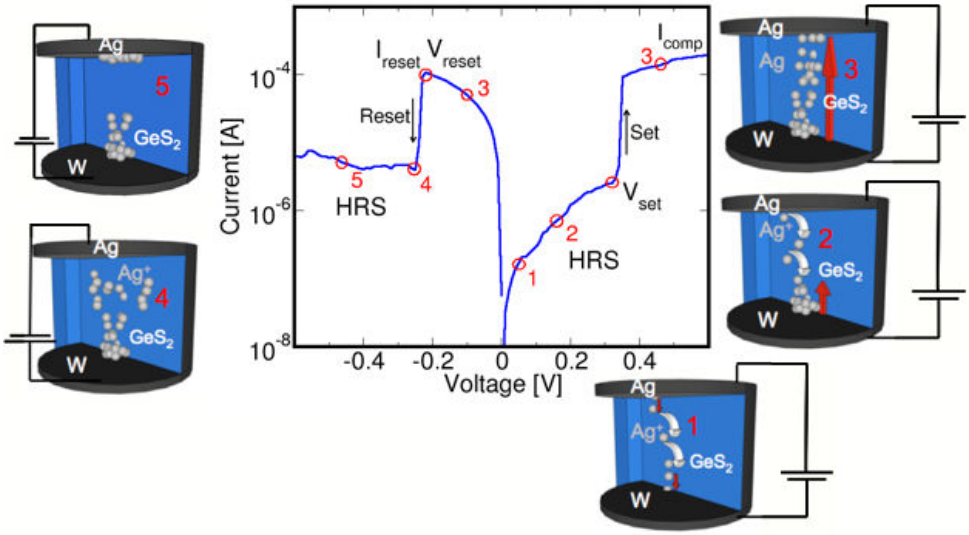


Figure 2.1: Steps of the switching process for W-GeS$_2$-Ag based CBRAM devices and corresponding DC Current-Voltage characteristic (main switching parameters are indicated). First step: oxidation of the Ag top electrode and diffusion into the GeS$_2$ electrolyte. Second step: Reduction of Ag$^+$ ions at the bottom electrode and nucleation of the new phase. Third step: Ag-rich CF formation at the set event and switching from high resistive state (HRS) to low resistive state (LRS). Fourth step: dissolution of the CF with re-oxidization of Ag during the reset event. Fifth step: reduction of Ag$^+$ ions at the top electrode.

2.1 W-GeS$_2$-Ag based CBRAM devices

Fig. 2.1 displays a typical current-voltage characteristic of a Ag-GeS$_2$ CBRAM device obtained in DC regime applying a voltage sweep. At the beginning the cell is in the high resistive state (HRS). To switch the cell from HRS to low resistive state (LRS) a positive voltage is applied to the silver anode which oxidizes, generating Ag$^+$ ions (step 1). These cations under the influence of the electric field, migrate by hopping to the W cathode where they are reduced (gain electrons) and nucleate, building-up the Ag-rich CF (step 2). After the CF has grown to make a metallic contact to the opposite electrode, the cell switches to the LRS at the set voltage (V_{set}). The on conductance is limited by the compliance current I_{comp}, that can be applied with the external semiconductor parameter analyzer (SPA) or an integrated MOSFET. The cell retains the LRS unless a sufficient voltage of opposite polarity is applied. To switch the cell from LRS to HRS a negative voltage is applied. During the reset process, besides an electronic current

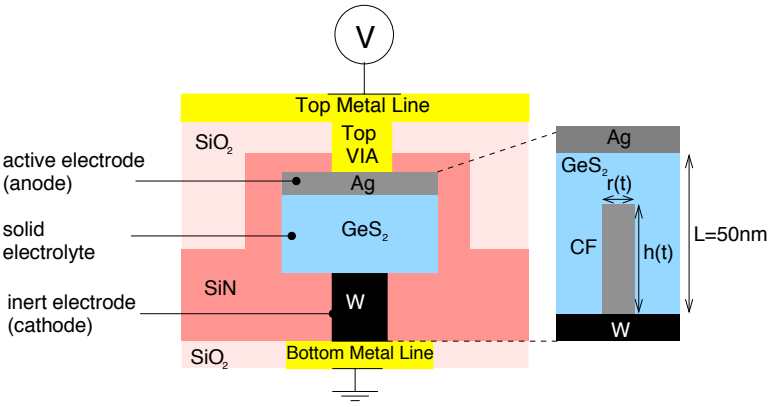


Figure 2.2: Schematic representation of W-GeS$_2$-Ag based CBRAM device. A Tungsten (W) plug is used as bottom electrode. The electrolyte consists of a 50 nm thick GeS$_2$ layer deposited by RF-PVD and a thin layer of Ag deposited by a DC PVD process. The thin Ag layer is dissolved into the GeS$_2$ using the photo-diffusion process, as described in [75]. Then a 2nd layer of Ag is deposited to act as top electrode. The Ag-rich CF is considered cylindrical with height $h(t)$ and radius $r(t)$.

flowing in the CF (step 3), an electrochemical current gives rise to Ag$^+$ ions that, not contributing at the metallic conduction in the CF, replate on the Ag electrode. In this phase, the cell switches from LRS to HRS. Conventionally, V_{reset} and I_{reset} are defined as the peak of the reset curve. In the HRS the CF can be partially or completely dissolved inside the GeS$_2$ depending on several factors (step 4) [70]. During this thesis, both isolated 1CBRAM cell (1R) and 1T-1CBRAM (1T-1R) cells were electrically measured.

2.1.1 Empirical model of the resistive switching in W-GeS$_2$-Ag

The CF formation is determined by the mass redistribution associated to the Ag$^+$ ion current $J(t)$. Different approaches exist to describe $J(t)$ depending on the step that rule the kinetics of the filament growth [71, 72, 73, 74]. If the switching speed during the program/erase process is related to the reaction rate at the cathode, Butler-Volmer equation applies [71, 72]. The ion diffusion in the electrolyte is another possible growth limiting step [72, 73, 74]. In this case the Mott-Gurney ionic hopping current can be adopted to describe $J(t)$. In this manuscript, we consider the CF as cylindrical, with a radius $r(t)$ and height $h(t)$ (Fig. 2.2) we assume that the vertical and lateral time evolution of the CF are proportional to the ion current density. Namely, $dh/dt \propto J_h(t)$ and $dr/dt \propto J_r(t)$ [72].

To reproduce some experimental evidences, we introduce an empirical parameter (Δ) in the expression of the Mott-Gurney ionic hopping current (Eq. 2.1). In particular, if $V(t) < \Delta$ some processes or combination of them related to the silver oxidation, the migration of the ions through the chalcogenide, the electron transfer between the incoming ions and the cathode or the nucleation of a new phase on the cathode are not sufficiently activated to allow the switching from HRS to LRS. To obtain simulated DC I-V characteristics, we describe the CF evolution with three stages: 1) the vertical growth; 2) the lateral growth; 3) the lateral dissolution. The CF vertical (dh/dt) and lateral (dr/dt) time evolutions are thus described as follow:

$$\frac{dh}{dt} = \frac{J_h(t)}{qN_i} = v_h \exp\left(\frac{-E_A}{k_B T}\right) \sinh\left(\alpha q \frac{V_c(t) - \Delta}{k_B T}\right) \tag{2.1}$$

$$\frac{dr}{dt} = \frac{J_r(t)}{qN_i} = v_r \exp\left(\frac{-E_A}{k_B T}\right) \sinh\left(\beta q \frac{V_c(t) - \Delta}{k_B T}\right) \tag{2.2}$$

where q is the elementary charge, N_i is the density of the metal ions in the solid electrolyte, v_h and v_r are fitting parameters for the vertical and lateral evolution velocities, E_A is the activation energy for overcoming energy barriers in the electrolyte lattice (considered isotropic), k_B is the Boltzmann constant, T is the temperature and α and β are fitting parameters to take into account vertical and lateral electric field dependencies. Once $h(t)$ and $r(t)$ of the conductive filament are evaluated, the resistance of the cell is simply calculated as the sum of two series resistors:

$$R_c = \frac{\rho_\mathrm{on} h(t) + \rho_\mathrm{off}(L - h(t))}{\pi r^2(t)} \tag{2.3}$$

where ρ_on is the resistivity of the Ag-rich nanofilament, ρ_off is the resistivity of the chalcogenide and L is the chalcogenide thickness. Since our measured values for R_set are in the order of few kΩ, Joule heating effect is neglected. Indeed, as shown in [73], the temperature increase in the CF becomes relevant for filament resistances in the order of few tens of Ohm. Moreover, in [72] the Joule heating effect was introduced with an empirical formula to take into account the non-simmetry of the DC I-V set and reset curves, in our equations this feature is introduced by the parameter Δ itself. Table 2.1 shows the parameters used in our simulations extracted by fitting of the electrical measurements (both quasi-static and dynamic) of GeS$_2$ (50 nm) based CBRAM devices. The value of ρ_off can be estimated considering that in the pristine state the chalcogenide

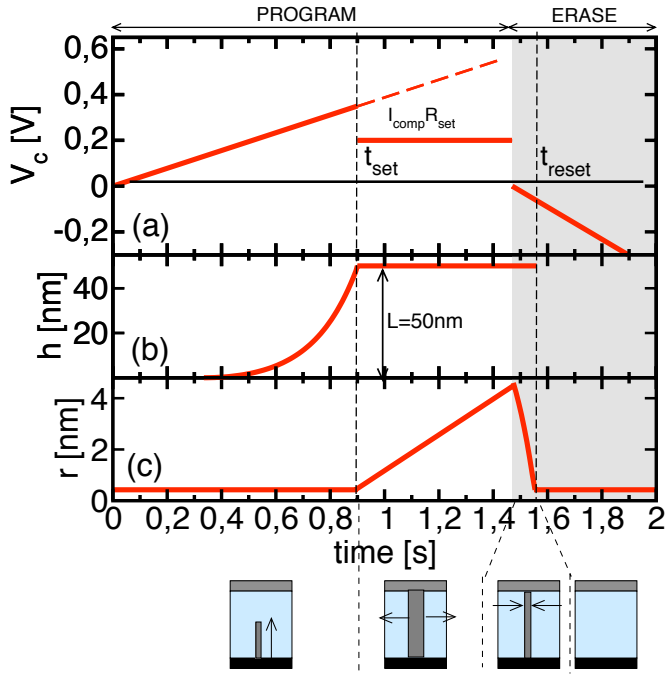


Figure 2.3: Simulated sequence of the set and reset transients: (a) Up-down voltage sweep applied to the GeS_2 (50 nm) based CBRAM cell; (b) vertical and (c) lateral evolution of the corresponding CF. The set occurs when the CF reaches the top electrode $h(t) = L$. Since the compliance current is enforced into the device, the applied voltage, V_c, decreases abruptly to a constant value: V_c=$R_{set}I_{comp}$ and the CF radius is expected to grow laterally [79]. At the beginning of the reset process the CF starts to laterally dissolve and the reset occurs when the CF radius shrinks to zero.

measured resistance is above 1 GΩ, while the value of ρ_{on} has been extrapolated assuming a CF radius ≃1 nm from resistivity measurements of Ag nanofilaments reported in [76]. Similar values for ρ_{on} were also reported in [77] for GeSe and [78] for GeS_2. Note that in the pristine state R_{off} is higher with respect to the measured values obtained after few cycles. This can be due to a modification of the chalcogenide resistive properties due to diffusion of the Ag cations. Hence, other values of ρ_{off} will be used in simulating the electrical behavior of devices previously cycled (Sec. 4.6).

2.1.2 Set and reset operation in quasi-static mode

Fig. 2.3 illustrates the procedure adopted to simulate a DC (or quasi-static) set/reset transient: the double voltage sweep $V_c(t)$ applied to the GeS_2 (50 nm) based CBRAM is

Table 2.1: Parameters used in the simulations for Ag-GeS$_2$(50 nm)-W 1R CBRAM devices

Parameter	Value	Parameter	Value
v_{h}	0.07 m/s	v_{r}	0.05 m/s
ρ_{on}	$2.3 \times 10^{-6}\ \Omega\ m$	ρ_{off}	$8 \times 10^{3}\ \Omega\ m$
α	0.4	β	0.35
E_{A}	0.4 eV	L	50 nm
A	0.2 V	Δ	0.15 V

illustrated in (a) as well as the corresponding simulated vertical $h(t)$ (b) and lateral $r(t)$ (c) evolution of the CF according to Eqs. 2.1 and 2.2 solved analytically. During the program or set phase a positive staircase from 0 V to 0.55 V is applied to the top electrode of the CBRAM. Under the positive bias the CF starts to grow according to Eq. 2.1 and the set event occurs when the filament reaches the top electrode (i.e. $h(t) = L$). After the set, the external Semiconductor Parameter Analyzer (SPA) regulates the applied voltage to provide a compliance current (I_{comp}) until the end of the positive ramp. Consequently, V_{c} decreases abruptly to the value $V_{\mathrm{c}} = R_{\mathrm{set}} I_{\mathrm{comp}}$. In this second phase, the CF grows laterally according to Eq. (2.2) starting from a initial value dependent on the compliance current. Then a staircase down signal is applied to the top electrode to enable the reset process: the CF tends to laterally dissolve by oxidation of Ag atoms of the CF and the reset occurs when the radius shrinks to zero.

Fig. 2.4 (left) shows experimental and simulated quasi-static I-V characteristic obtained when applying to the top electrode the double sweep voltage shown in Fig. 2.3. An asymmetry of the set and reset voltages appears, in particular ($V_{\mathrm{set}} \simeq 350$ mV and $V_{\mathrm{reset}} \simeq$ -80 mV), which was also reported for the same stack in [73]. Simulations with different Δ values were performed to fit the experimental V_{set} and I_{reset}. An increase of the parameter Δ reduces the effective voltage drop inside the solid electrolyte, thus decelerating the vertical and lateral growth of the CF. Therefore, the higher Δ is, the higher set voltages and lower R_{set} values are, if the filament is assumed to grow laterally. As a consequence of the lower R_{set}, a larger current and voltage need to be applied to dissolve the conductive filament during the reset (Fig. 2.4). Simulations well reproduce data when Δ is fixed to 0.15 V. Fig. 2.4 (right a) shows the set resistance obtained by programming our devices with different values of external I_{comp}. For these measurements the positive voltage ramp is stopped at compliance, preventing the expected lateral

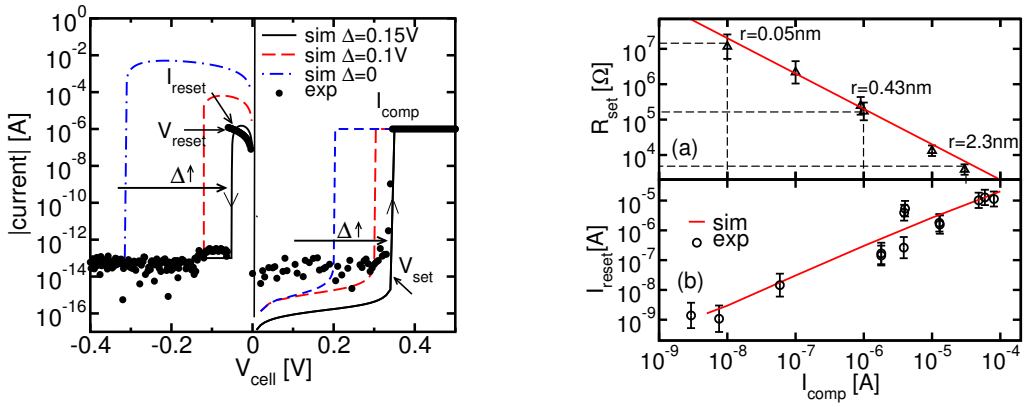


Figure 2.4: (Left) Experimental (symbols) and simulated (lines) current-voltage curves obtained by applying a voltage sweep as illustrated in Fig. 2.3. The compliance current was 1 μA (I_{comp}= 1 μA). Experimental data shows a strong asimmetry for the set and reset voltage. Different Δ values impact on V_{set}, V_{reset} and I_{reset}. Data are best fitted fixing Δ equals 0.15 V. (Right a) R_{set} dependence on I_{comp}. The experimental data (symbols) are fitted with the inversely proportional relationship R_{set}=A/I^n_{comp} where A = 0.2 V and n = 1. The value of the radius corresponding to the obtained R_{set} is also shown. (Right b) I_{reset} as a function of I_{comp} enforced during the previous set operation.

growth of the CF. The dependence of R_{set} on the compliance current follows an empirical power-law relationship [80]:

$$R_{\text{set}} = \frac{A}{I^n_{\text{comp}}} \tag{2.4}$$

A and n being the fitting parameters. It is noteworthy that our experimental data at room temperature are best fitted with $A = 0.2$ V and $n = 1$. Note that A corresponds to the voltage applied on the cell after the set occurrence during quasi-static programming (Fig. 2.3 (a)). The CF radius at $t = t_{\text{set}}$ can be calculated by introducing Eq. 2.4 in Eq. 2.3 as illustrated in Fig. 2.4 (right). Moreover, experimental data show that larger I_{comp} determines a larger reset current (I_{reset}), which is defined as the maximum current flowing into the CBRAM during the reset phase. Fig. 2.4 (right b) shows that the simulated I_{reset} reproduces the experimental data for several orders of magnitude of I_{comp}. Due to the very low value of the reset current that we achieved, the reset mechanism should be based on electrochemical reactions, and a thermal rupture of the CF should be excluded (at least for low I_{comp}) [81]. In order to investigate the

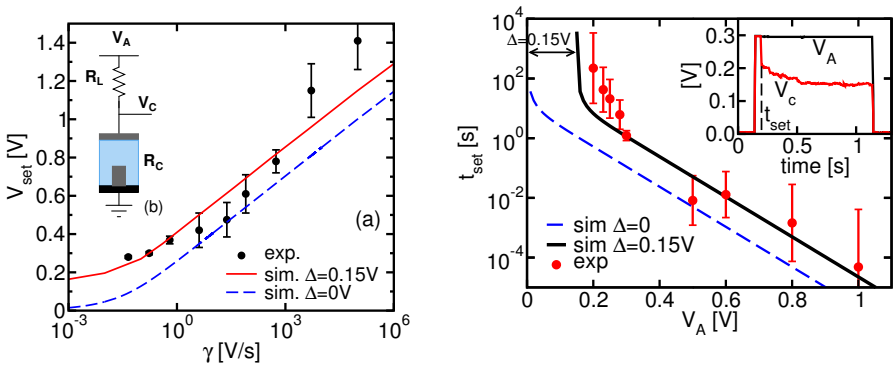


Figure 2.5: (Left a) Experimental (symbols) and simulated (lines) switching voltage V_{set} versus sweep voltage rate. The inclusion of Δ in the model allows to reproduce the saturation of V_{set} for low sweep voltage rates. (Left b) Experimental setup to perform pulsed test. (Right) Experimental (symbols) and simulated (lines) switching time t_{set} as a function of the applied voltage amplitude. Simulation carried with $\Delta = 0.15$ V reproduces the abrupt increasing of t_{set} when $V_{\text{A}} \simeq 0.2$. In the inset typical oscilloscope trace of V_{c} and V_{A} during a set operation.

dependence of V_{set} with the kinetic of the switching process, we varied the sweep rate γ of the ramp voltage signal over seven orders of magnitude recording the obtained set voltages. Tests performed with a ramp rate $\gamma < 1$ V/s were conducted with a parameter analyzer, while a pulse generator was used for higher ramp rate. We used a series resistance of 1 kΩ is added in series to the cell while an active probe measures the voltage drop on the memory device (Fig. 2.5 left b). The series resistance limits the current overshoot during set transient and allows the detection of the set event as a drop of voltage detected by the probe with respect to the input signal. Fig. 2.5 (left) shows that V_{set} increases with the ramp speed of the applied voltage, but for ramp rate below 1 V/s, $V_{\text{set}} \simeq 0.2$ V approaching a saturation value. This behavior is captured by our simulations by fixing Δ equal to 0.15 V.

2.1.3 Set and reset operation in pulse mode

In this section results of the pulse mode programming are presented. Inset of Fig. 2.5 (right) shows a typical programming pulse. Right after the set, which is evidenced by a sudden drop of the signal on the device because of the configuration reported in Fig. 2.5 (left b), the cell voltage slightly decreases as a consequence of the supposed

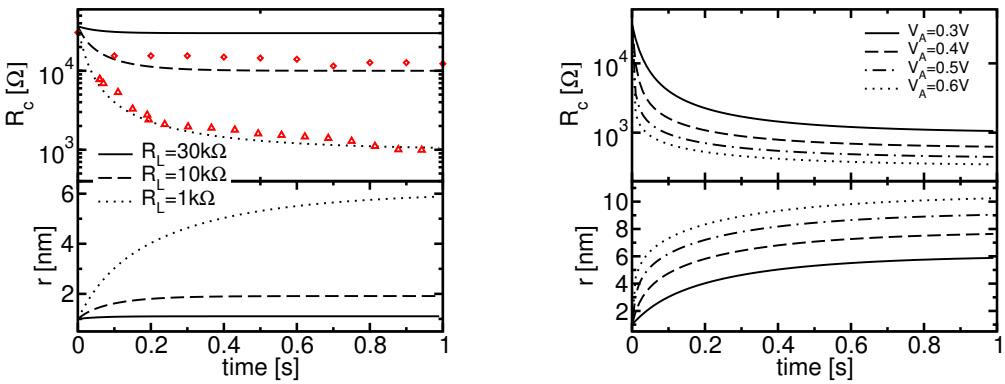


Figure 2.6: (Left Up) Measured (symbols) and simulated (lines) CBRAM resistance R_c versus time for different load resistances R_L. (Left Down) Simulated filament radius versus time for different load resistances R_L. (Right Up) Simulated CBRAM resistance R_c versus time for different applied voltages V_A. (Right Down) Simulated filament radius versus time for different applied voltage V_A.

lateral growth.

To simulate the AC transient we calculate the vertical and lateral growth of the CF when a rectangular pulse $V_A(t)$ is applied to the series of the CBRAM and the load resistor. Before the set, as the cell resistance is in the order of GΩ we can assume that $V_c = V_A$. After the set, the voltage on the CBRAM cell can be calculated with the voltage divider formula. In this case $V_c(t)$ is a function of both $r(t)$ and $V_A(t)$ so the lateral evolution is the solution of the system of Eqs. 2.2 and 2.3 and must be solved numerically. Fig. 2.5 (right) displays the switching time as a function of the voltage applied to the cell. The inverse of the switching time exponentially depends on the applied voltage when V_A is above 0.2 V. However, at lower voltages t_{set} increases much faster, indicating that a much longer time is required to set the cell. This behaviour has been widely reported in the literature [73, 74, 80]. Fig. 2.5 (right) also shows that simulations well reproduce the experimental data. For voltage value between 0.2 V and 1 V, the dependence of t_{set} versus V_A (i.e. the slope of the curve $d(\ln t)/dV$) is determined by the fitting parameter α in Eq. 2.1, while at lower voltages the introduction of the Δ parameter allows to reproduce the abrupt increase of the switching time. Fig. 2.6 (left) shows the evolution of the CBRAM resistance R_c after the set event for different load resistances (R_L). By decreasing R_L a lower R_c is obtained. Indeed the effect of R_L is to

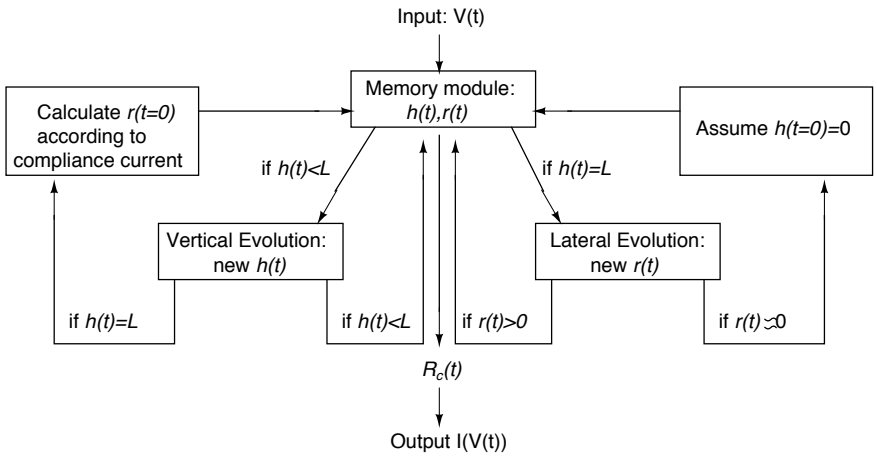


Figure 2.7: Flow chart of the compact model for Ag/GeS$_2$ CBRAM cells.

reduce $V_c(t)$ thus preventing the radius growth (Fig. 2.6 left down). We also show that simulations well reproduce both the kinetics of the R_c evolution and the asymptotic resistance values obtained with different R_L. Fig. 2.6 (right) reports the time evolution of the simulated R_C (up) and $r(t)$ (down) for different applied voltages V_A. A load resistance of 1 kΩ has been used. As expected V_A influences the value of R_c.

2.1.4 Flow chart of the compact model

The proposed model explains the time dependent switching process of CBRAM cells and it has been implemented in Verilog-A language for electrical simulations in the process design kit of ALTIS Semiconductor [82]. Fig. 2.7 shows the flow chart of the implemented model: the input of the device is the voltage $V(t)$ and the output is the current $I(V(t))$. A memory module stores the quantities $h(t)$, $r(t)$. The value of $r(t = 0)$ is determined by the current compliance, and the initial height ($h(t = 0)$) is assumed to be zero. The evolution of these two internal quantities is used to calculate $R_c(t)$ according to Eq. 2.3. At each time step, if $h(t) < L$ (i.e. the cell is in the *off* state), Eq. 2.1 is used to calculate the new $h(t)$. If $h(t) = L$ (i.e. the cell is in the *on* state), Eq. 2.2 is used to calculate the new $r(t)$. When the RESET occurs ($r(t) = 0$), h and r have to be reinitialized: h is set to zero and r is evaluated substituting Eq. 2.4 in Eq. 2.3. The inclusion of the cell to cell variability in the model will be considered in Sec. 2.2, after measurements of 8×8 memories array.

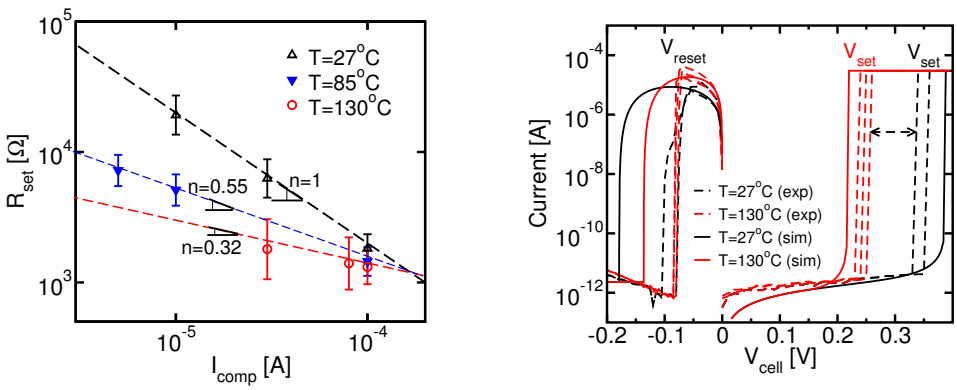


Figure 2.8: (Left) Set resistance (R_{set}) dependence on current compliance (I_{comp}) measured at 27, 85 and 130 °C. Only at room temperature data can be fitted with the law $R_{\text{set}}=A/I_{\text{comp}}^{n}$ where $A = 0.2$ V and $n = 1$. Parameter n accounts for the slope of the fitting curves. (Right) Experimental (dashed) and simulated (lines) I-V set and reset electrical characteristics obtained for the sample with thickness 30 nm at 27 °C and 130 °C. The compliance current was 30 μA. Lower V_{set} were observed at 130 °C.

2.1.5 Temperature effects on the switching kinetics

In this section, we will describe the temperature effects on the switching parameters of Ag/GeS$_2$ CBRAM devices. As described in Sec. 2.1.2 and in [80, 70, 83] the set Resistance (R_{set}) can be tailored over a wide range using the compliance current of an external semiconductor parameter analyzer (SPA) or using an integrated transistor [84, 85] during the set operation. At room temperature this dependence can be fitted with Eq. 2.4 and it was reported in Fig. 2.4 (right a). This equation has been demonstrated to be independent of the material composition and the set voltage. Fig. 2.8 (left) displays R_{set} obtained when programming a W-GeS$_2$(30 nm)-Ag based CBRAM device with different external compliance currents. The test has been performed at three temperatures, namely 27, 85 and 130 °C, and indicate that for a defined value of the compliance current, lower resistance values are obtained when increasing the temperature; hence the exponent n, that gives the slope of the fitting curve, is no longer equal to 1. It is worth noting that the pristine resistance is around 1 GΩ for all temperatures and current compliance values. Fig. 2.8 (right) shows the measured and simulated current voltage (I-V) switching characteristics performed at 27 °C and 130 °C in W-GeS$_2$(30 nm)-Ag. These curves were obtained by applying a triangular voltage sweep with ramp

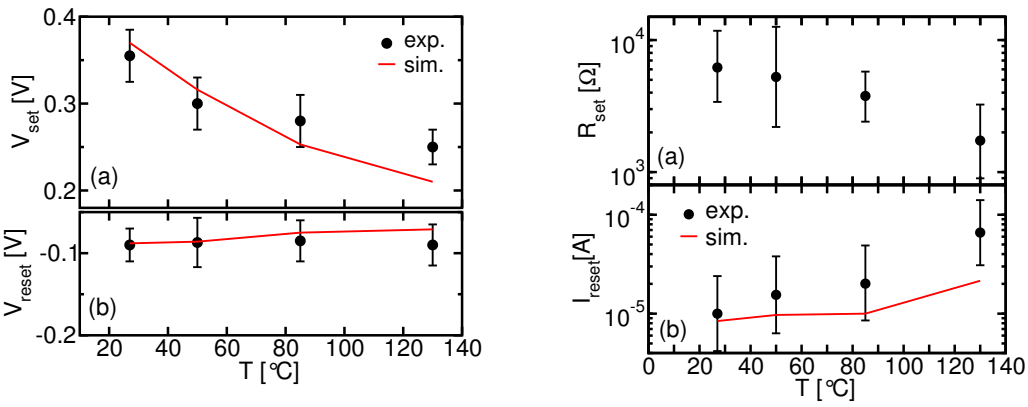


Figure 2.9: (Left) Experimental (symbols) and simulated (lines) V_{set} (a) and V_{reset} (b) versus temperature. (Right) Experimental (symbols) and simulated (line) R_{set} (a) and I_{reset} (b) versus temperature obtained applying on the cell a compliance current of 30 μA.

Table 2.2: Parameters used in the simulations for Ag-GeS$_2$(30 nm)-W 1R CBRAM devices to take into account temperature effects.

Parameter	Value	Parameter	Value
v_{h} (m/s)	0.07	v_{r} (m/s)	0.005
$\rho_{\text{on (T = 300 K)}}$ (Ω m)	2.3×10^{-6}	α	0.4
$\rho_{\text{on (T = 400 K)}}$ (Ω m)	93×10^{-6}	β	0.35
E_{A} (eV)	0.4	ρ_{off} (Ω m)	8×10^{2}
A (V)	0.2	Δ (V)	0.15

rate $dV/dt = 1$ V/s and compliance current equal to 30 μA. Fig. 2.9 (left) displays the measured (symbols) and calculated (lines) set and reset voltages (V_{set}, V_{reset}) as a function of the temperature. As the temperature increases, the V_{set} decreases while V_{reset} remains almost constant. The experimental results are well reproduced by the simulations. The model predicts an increase of V_{set} for lower temperature. Experimental results reported in [86] for Cu-Cu$_2$S based CBRAM devices show the same trend.

Clearly, in the proposed model the ions diffusion is enhanced at higher temperatures (Eq. 2.1) thus speeding up the set process and explaining the V_{set} decreasing. The weak dependence on temperature measured for V_{reset} is attributed to the lower set resistance values collected at high temperatures (Fig. 2.9 right a), that could veil the impact of the temperature on V_{reset}. It is worth noting that in order to capture the experimental

I_{reset} values we used a resistivity ρ_{on} of the CF increasing with temperature (Tab. 2.2). This seems reasonable since a metallic behavior for the conduction has been measured for both CBRAM and even OxRRAM in the low resistive state [87, 88, 85].

2.2 Switching probability in W-GeS$_2$-Ag based CBRAM devices

After having investigated isolated 1R and 1T-1R devices, we characterized the switching behavior of 64 cells organized in a 1T-1R NOR 8×8 memory array. The transistor in series has length L=140 nm and width W=500 nm. The aim of this section is to provide a statistical analysis on the correlation between the programming conditions and the percentage of CBRAM devices in the memory array that reversibly switch from high resistive state (HRS) to low resistive state (LRS) and vice-versa. In fact, CBRAM devices exhibits cell to cell variability in the set and reset voltage/time both inside the same memory array and among memories array of different dies in the wafer. A statistical analysis on the switching conditions is required to improve the model described in Sec. 2.1.4 taking into account the cell to cell variability. In this case, the model describes the switching parameters of the memory defining the programming conditions that lead to the highest switching percentage on a total of 64 devices.

The cell to cell variability is generally corroborated by the cycle to cycle variability both for the set and the reset operation, thus complicating the statistical analysis or the development of compact models.

On 1T-1R structures organized in a memory array the programming conditions include: the polarization of the transistor i.e. the pulse amplitude (V_g) and width applied at the gate (t_{pw}), the voltage on the bitline (VBL) and the polarization of the anode (top electrode) (V_a) of the CBRAM device (Fig. 2.10). During the set operation, the bitline is grounded, the anode is polarized and the 8 wordlines are sequentially pulsed. During the read operation, the bitline and the wordline are polarized and the current is measured at the anode of the CBRAM. During the reset operation, the anode is grounded, the bitline is polarized and the 8 wordlines are sequentially pulsed (Fig. 2.10). Programming conditions can be defined as strong (high voltage, long pulse width) or weak (low voltage, short pulse width). Weak programming conditions determine a percentage of devices reversibly switched that is lower than 100%. To characterize the

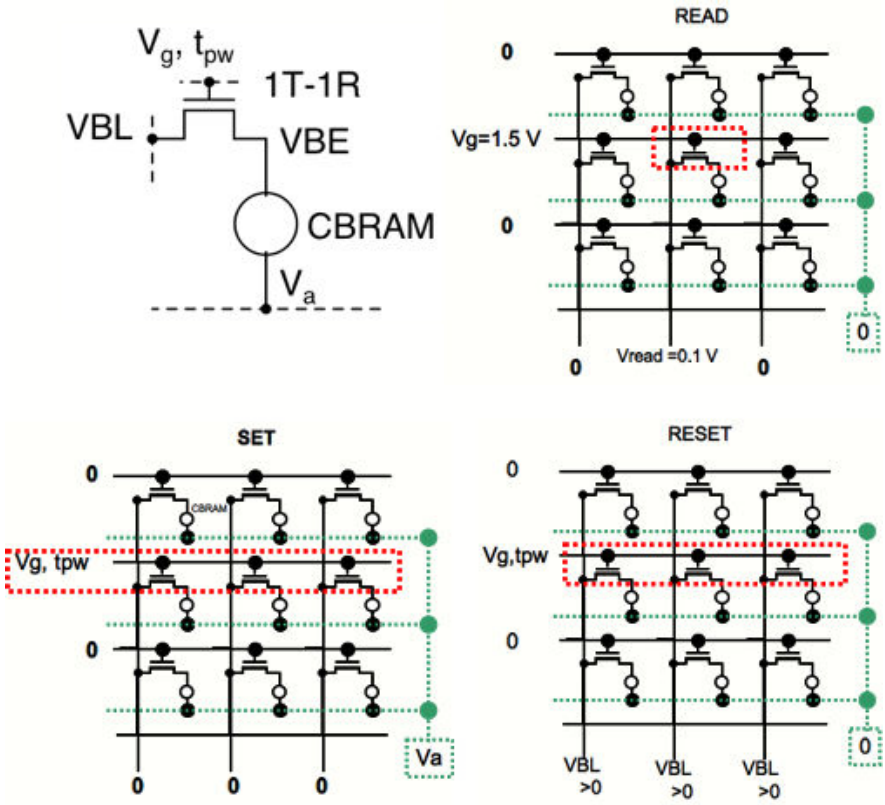


Figure 2.10: (Left up) Schematic of the 1T-1R structure. (Right up) Schematic of the 8×8 NOR memory array (only three lines represented). During the read operation (right up), the anode (V_a) is grounded, the bitline (VBL) is polarized to 0.1 V and the wordline to 1.5 V. During the set operation (left down), the bitline is grounded, the gate is pulsed (V_g, t_{pw}) and the anode is polarized. During the reset operation (right down), the anode is grounded, the gate is pulsed and the bitline is polarized.

cell to cell and the cycle to cycle variability related to the set process we proceed as follows: the resistance values of the 64 devices in the memory array in the HRS are measured. Next, we switch the 64 devices applying a pulse amplitude on the wordline of 1.5 V, a defined pulse width (t_{pw}) and voltage on the anode (V_a) and we measure the final resistance of all the devices (first set). A reset with a strong condition is then applied. Using the same set condition of the first cycle we set the devices again. Finally a last reset is performed. The strong condition used in the reset ensure the switching from LRS to HRS for all the previously switched devices in the memory array. Fig. 2.11 reports the Cumulated Distribution Functions (CDF) of the resistance values obtained

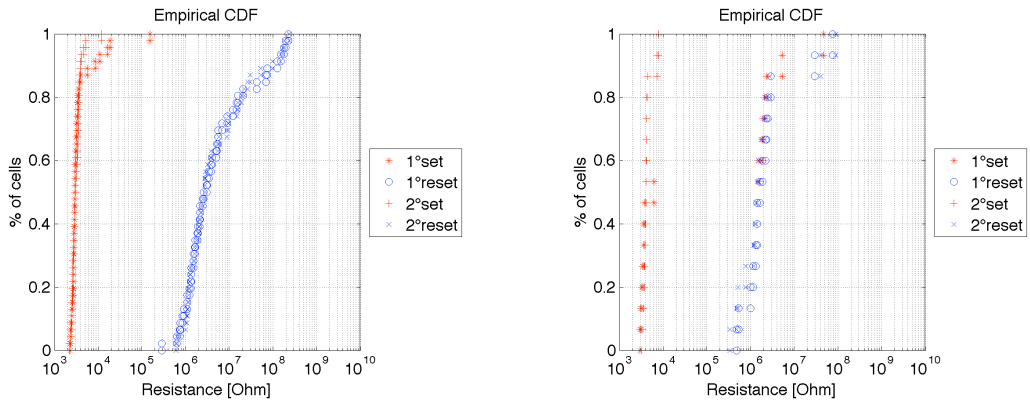


Figure 2.11: (Left) Empirical CDF on 52 devices of the LRS and the HRS after two set cycles and two reset cycles. A pulse of 100 µs and a voltage on the anode of 2 V were applied. (Right) Empirical CDF on 22 devices of the LRS and the HRS after two set cycles and two reset cycles. A pulse of 400 ns and a voltage on the anode of 1.5 V were applied.

after the two set/reset cycles for the cells switched in a reversible way. In particular, in Fig. 2.11 (left) a strong set programming conditions was used corresponding to a pulse width t_{pw} of 100 µs and a voltage on the anode of 2 V (V_g=1.5 V). In Fig. 2.11 (right) a weak condition was used corresponding to a pulse width t_{pw} of 400 ns and a voltage on the anode of 2 V (V_g=1.5 V). Accordingly to the these distributions the number of two times switched devices is lower if weaker conditions are applied. Mean R_{on} and R_{off} values are reported in Fig. 2.12. Note that these values were calculated considering the population of the two times switched devices (as reported by the CDF), thus the total number varies depending on the applied condition (strong or weak). Several set conditions, specified on the x-axis of Fig. 2.12, were applied to investigate the set efficiency, the R_{on} and R_{off} values. Between every test, strong programming conditions were applied to the memory array to reinitialize the 64 CBRAM devices in the HRS. Fig. 2.13 shows the mean R_{on} and R_{off} values when 1.1 V were applied on the wordline (V_g). In this case slightly higher R_{on} values with respect to values of Fig. 2.12 were measured, regardless of the conditions used on the pulse width and on the anode. This is because of the lower current value that flows into the device. The efficiency in the set operation is reported in Fig. 2.14. The set efficiency is defined as the mean of the percentage of the cells that switched during the first (or the second cycle) and the percentage of cells that switched in both cycles. For example on 64 cells, a total of 42 cells switched in the first set or in the second set operation, while 36 switched in

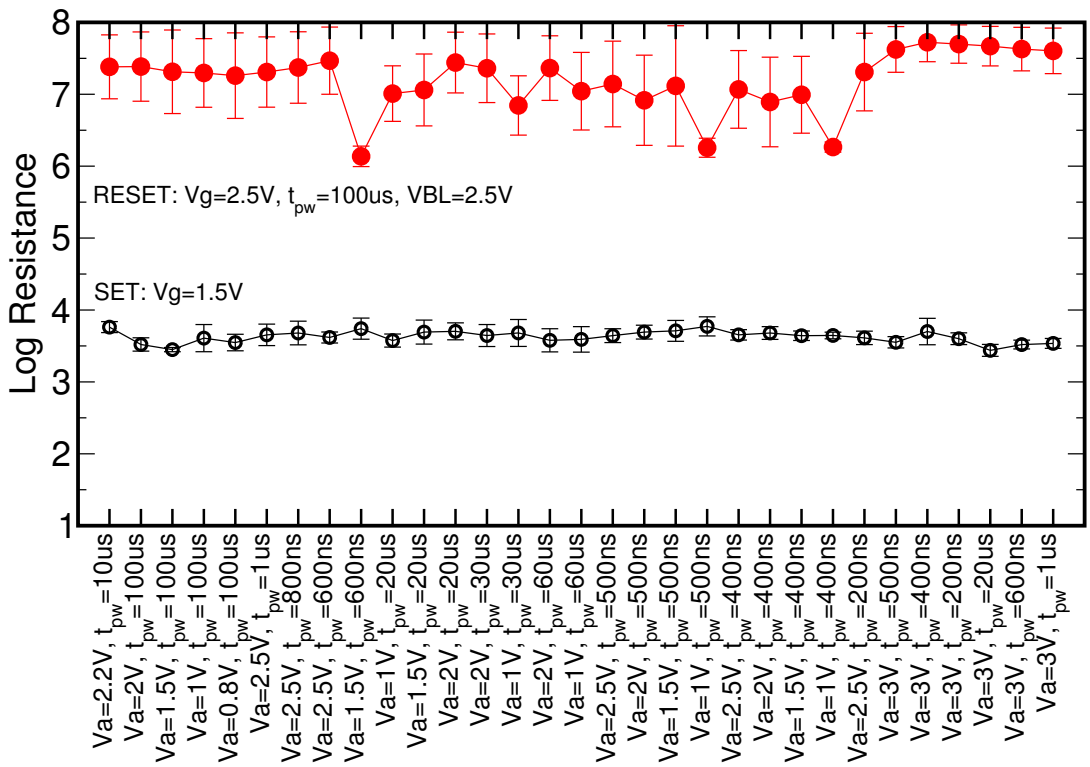


Figure 2.12: Mean R_{off} and R_{on} after two set/reset cycles for the switched devices. On the x axis the conditions used during the set operation are reported (voltage on the gate 1.5 V). A fixed strong reset conditions was applied.

both cycles. The set efficiency will be 60%. Conventionally, cells are considered in LRS if the resistance is lower than 20 kΩ, while are considered in HRS if the resistance is higher than 200 kΩ. Fig. 2.14 displays the set efficiency as a function of the voltage on the anode for a constant wordline voltage of 1.5 V (left) and 1.1 V (right) (pulse width on the gate specified in the legend, t_{pw}). The efficiency raises with the voltage on the anode and also the pulse width, even if the impact of the voltage is stronger. Longer pulses increase the set efficiency, when the same V_a is applied. Next, we investigate the effects of programming conditions on the reset process analyzing the reset efficiency, the R_{on} and R_{off} values. Mean R_{on} and R_{off} values are reported in Fig. 2.15 applying a voltage of 2.5 V (left) and 2 V (right) at the wordline. Both the pulse width and the bitline voltage (VBL) were varied accordingly to the x-axis of Fig. 2.15 while a fixed

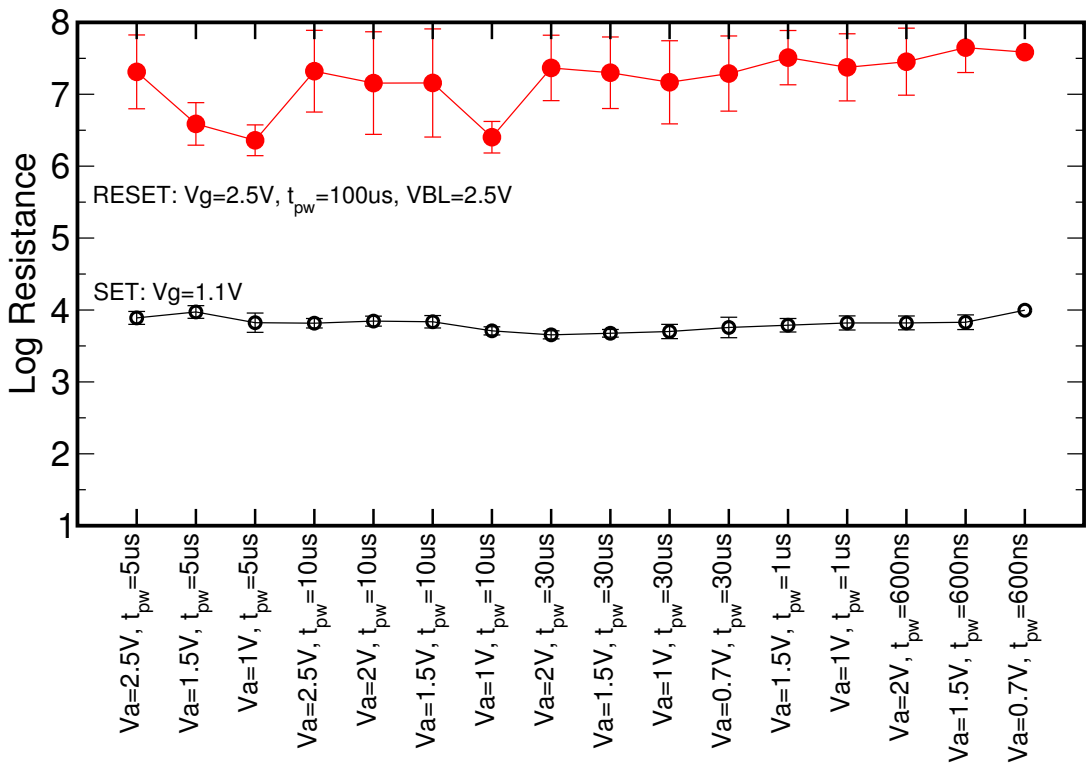


Figure 2.13: Mean R_{off} and R_{on} after two set/reset cycles for the switched devices. On the x axis the conditions used during the set operation are reported (voltage on the gate 1.1 V). A fixed strong reset conditions was applied.

strong set conditions was applied to ensure the switching from HRS to LRS. The R_{off} value does not depend on the V_g, but the reset efficiency is strongly dependent on this value, because it raises the voltage on the bottom electrode (Fig. 2.16). Note also that the reset efficiency is constant when bitline voltages higher than V_g are applied. Longer pulses increase the reset efficiency, when the same VBL is applied (Fig. 2.16).

2.3 Phenomenological explanation of variability in the switching time

In Sec. 2.2 we claim that the applied set/reset conditions can control the set/reset efficiency values. These values are related to the cell to cell and cycle to cycle dispersion

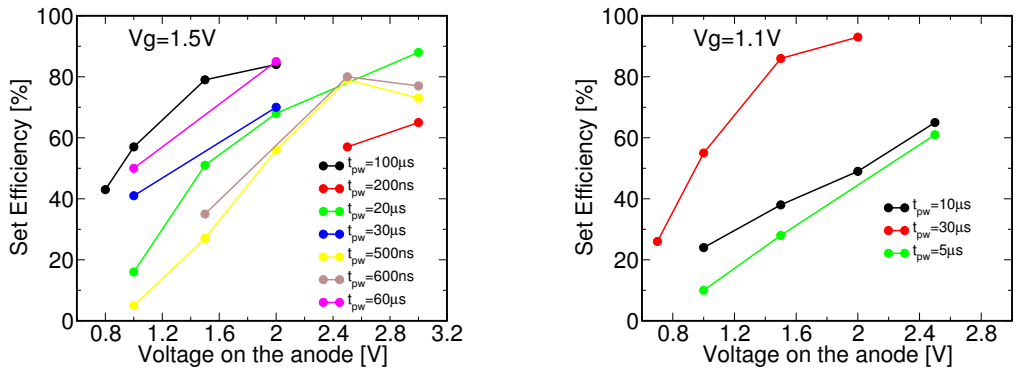


Figure 2.14: Set efficiency (Switching probability) for 64 devices of the matrix varying the voltage on the anode (voltage on the gate 1.5 V (left) or 1.1 V (right)). Switching being considered successful if R_{on} lower than 20 kΩ and R_{off} higher than 200 kΩ.

of the switching parameters. In fact, measured quantities such as R_{off} or t_{set} shows a distribution both when the same device is cycled many times and when many devices are cycled one time with the same condition. In the model that we proposed in Sec. 2.1, Eq. 2.3 provides the most simply way to link the resistance to the filament shape. By inserting, the measured distribution of R_{off} on a single cell cycled many times, we were able to compute a list of the initial filament height of the CF. In this calculation the resistivity of the chalcogenide ($\rho_{off} = 10^{-3}\,\Omega\,\mathrm{m}$) was used to provide a mean filament height between 0 nm and 27 nm (L=30 nm). The filament radius was assumed to be 2.2 nm. The calculated filament height distribution was used as the initial condition to solve Eq. 2.1 to evaluate the distribution of set time/voltage. The set process was simulated both in quasi-static and pulse mode. In quasi-static mode the simulated ramp rate was 0.6 V/s. Similarly, the distribution of the time required to set the CBRAM was computed assuming a pulse amplitude of 1 V. The dispersion of t_{set} is consistent with results reported in Fig. 3.2 (right) for GeS_2 thickness (L) of 30 nm. Calculations results are reported in Fig. 2.17 and provide an empirical explanation of the distributions of measured quantities in CBRAM devices. Clearly, also other parameters of the model are influenced by statistical fluctuations with an effect on the measured quantities. Moreover, we lumped the dispersion in R_{off} with the filament height, but we can expect also a distribution of radii of the filament or different geometries. All these phenomena should be taken into account for a more realistic description of the source of variability

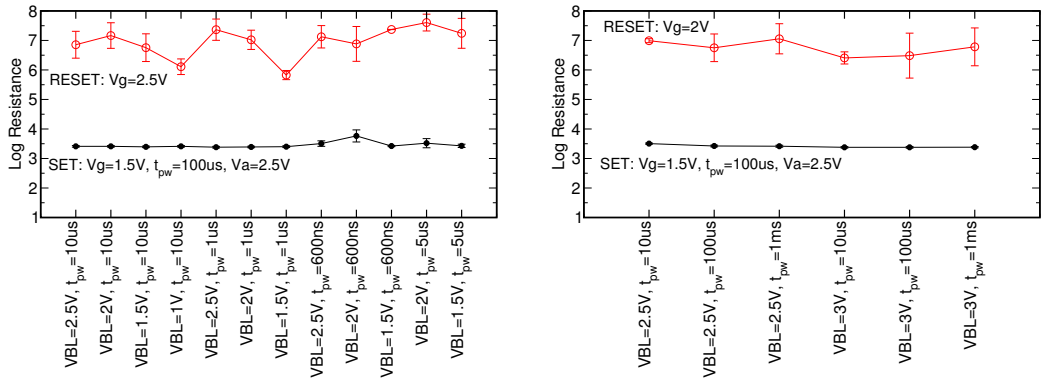


Figure 2.15: Mean R_{off} and R_{on} after two set/reset cycles for the switched devices. On the x axis the conditions used during the reset operation are reported (wordline voltage V_g=2.5 V (left) or 2 V (right)). A fixed strong set conditions was applied.

in GeS_2 based CBRAM devices.

2.4 Ta-GeS$_2$-Ag based CBRAM

In this section a Ta-GeS$_2$-Ag based CBRAM is presented. A Tantalum plug was fabricated instead of the Tungsten one of the CBRAM discussed so far. Deposition of Tantalum gives a better control of the surface roughness at the interface with the GeS_2, thus is expected to improve uniformity in the switching parameters from device to device and from cycle to cycle. The aim of this section is the comparison of the electrical performances of the new stack with respect to the previous one. In both the stacks the GeS_2 has a thickness of 30 nm. DC sweep I-V characteristic for 1R devices are reported in Fig. 2.18 using a compliance current of 30 nA. A mean reset current of only 6 nA was measured. Similar results were obtained for W-GeS$_2$-Ag as shown in Fig. 2.4 (right). This demonstrates that external current compliance is effective in limiting the current also in this stack and the potential interest in developing both the technologies for ultra-low power applications. Fig. 2.19 (left) displays the I-V characteristics for 1T-1R devices (100 cycles). Set/reset voltage of 0.3 V/−0.2 V were measured. Fig. 2.19 (right) reports the R_{off} and R_{on} distributions measured on 100 cycles. A mean R_{off}/R_{on} of 500 kΩ/5 kΩ is reported, similarly to the W-GeS$_2$-Ag CBRAM (Fig. 3.8).

In Sec. 2.1 we introduced the parameter Δ to take into account several features of the electrical behavior of our devices such as: the asymmetry in the measured I-V

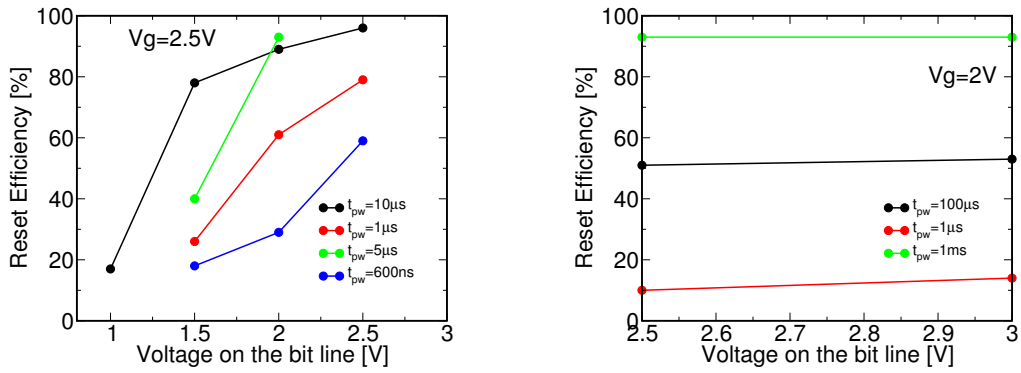


Figure 2.16: Reset efficiency (Switching probability) for 64 devices of the matrix varying the voltage on the bitline (voltage on the gate 2.5 V (left) or 2 V (right)). Switching being considered successful if R_{on} lower than 20 kΩ and R_{off} higher than 200 kΩ.

curves (Fig. 2.4), the increase in t_{set} in dynamic measurements for voltage lower than 0.2 V (Fig. 2.5 right) and the voltage saturation for low ramp speed (Fig. 2.5 left). We claimed that parameter Δ can be related to the overpotential at the cathode to allow the reduction of incoming Ag ions, but also on the nucleation and the growth of the new phase on the bottom electrode. Quasi-static I-V measurements showed same set/reset voltages for both Ta-GeS$_2$-Ag and W-GeS$_2$-Ag (Fig. 2.19 and Fig. 3.7 left up) with a memory window of 10^2. For this reason, we decided to perform dynamic measurements as described in Sec. 2.1.3. Results are reported in Fig. 2.20 and compared with the W-GeS$_2$-Ag stack. Note that for voltage applied higher than 0.3 V, the experimental data overlap among each other, thus no difference in t_{set} exists between the two stacks. We also investigate further at very low voltages (at the t_{set} saturation limit) using the electrical protocol described in Sec. 2.4.1.

2.4.1 Low-voltage pulse measurements

We characterized Ta-GeS$_2$-Ag based 1T-1R devices organized in the 8×8 memory array to obtain statistical results on the t_{set}. We measured the resistance of the devices in their pristine state. We chose the pristine state because, at our knowledge, the mechanism for the switching would be the diffusion by hopping of the silver ions towards the cathode and the reduction, thus we expect to investigate only the reduction and the nucleation process. The migration component should be equal between Ta-GeS$_2$-Ag and

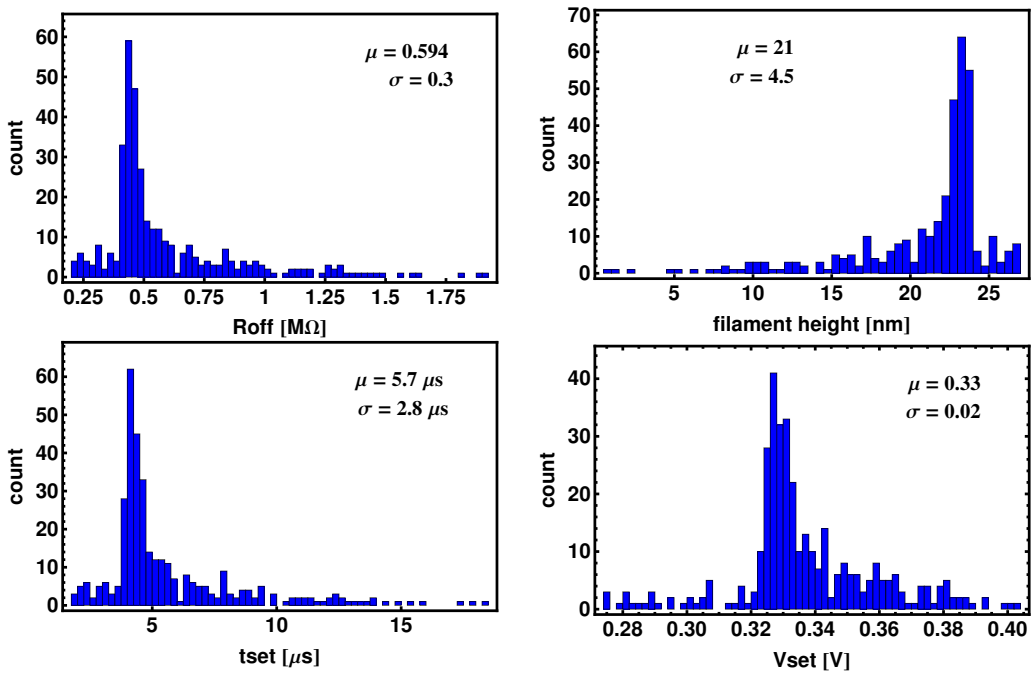


Figure 2.17: (Left up) R_{off} distributions as obtained in cycling many times GeS_2 based CBRAM devices. A log-normal distributions was used to fit the experimental values. (Right up) Calculated left over filament height as could be obtained after a reset operation. (Left down) Distribution of the time required to set the device starting from different left over filament heights the pulse amplitude used was 1 V. (Right down) Distribution of the voltage required to set the device starting from different left over filament heights, when a staircase ramp of 0.6 V s is used.

W-GeS$_2$-Ag because the GeS$_2$ and the top electrode thicknesses are the same. Even the photodissolution of Ag inside the GeS$_2$ was kept equal (i.e same process parameters) for the two stacks. In a first part of the test a constant voltage is applied at the anode (VPL in Fig. 2.21), the in series transistor is on (V_g=2 V) and the bitline is grounded (Fig. 2.10). We measure the current through the devices with logarithmically spaced reading operation. The total time of the measure depends on the applied voltage stress. The test ends with a functional check that consists in resetting/setting the devices 2 times each. Fig. 2.21 shows a schematic of the electrical protocol used. The set time (t_{set}) was defined as the time required to decrease the resistance by a factor of 5 but also 10 was considered. The mean t_{set} for a given applied stress was extracted at 30% and 45% of the t_{set} distribution obtained on the 64 cells of the memory array (Fig. 2.22). Fig. 2.22 (right) shows that when 190 mV are applied at the anode t_{set} is in the order

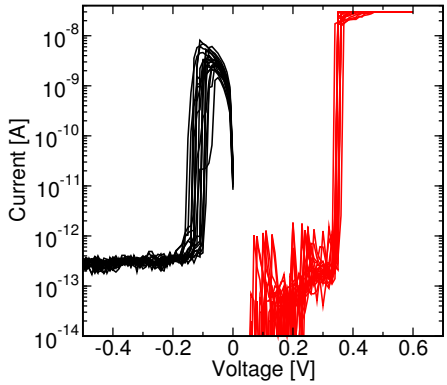


Figure 2.18: DC I-V characteristics for Ta-GeS$_2$-Ag 1R CBRAM devices. A current compliance of 30 nA was used.

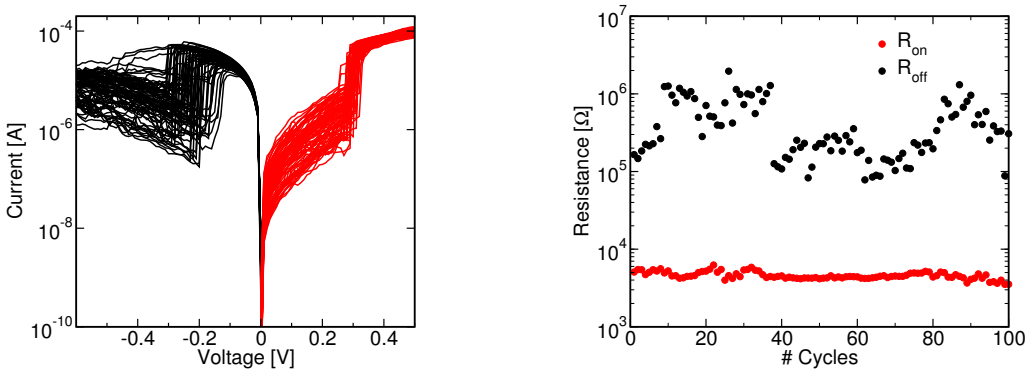


Figure 2.19: (Left) DC I-V characteristics for Ta-GeS$_2$-Ag 1T-1R CBRAM devices. (Right) Resistance values R$_{on}$ and R_{off} after set and reset respectively.

of 10^5 s considering the 45% of the distribution of the switched W-GeS$_2$-Ag. When the same stress was applied to Ta-GeS$_2$-Ag devices we measured a t_{set} in the order of 3×10^3 s. We can conclude that the t_{set} in W-GeS$_2$-Ag is slightly higher with respect to Ta-GeS$_2$-Ag if very low voltages are considered (below 200 mV). In other words, considering the same set time a voltage difference of 10 mV difference is required to switch the same percentage of devices for the two studied stacks.

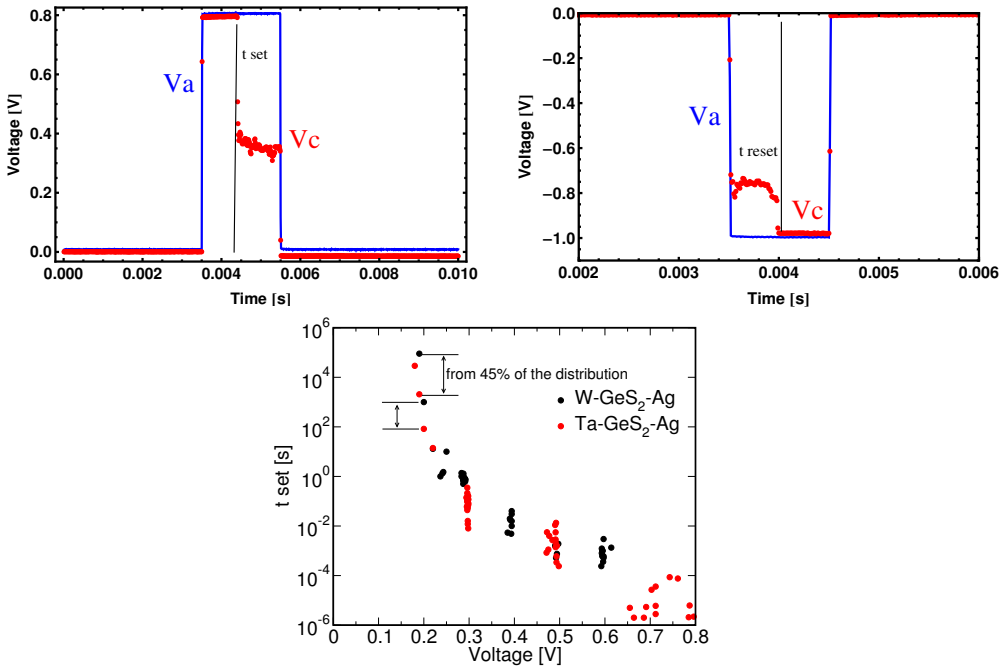


Figure 2.20: Oscilloscope trace from dynamic measurements of Ta-GeS_2-Ag: set (left) reset (right). Time required to set versus voltage applied for 1R Ta-GeS_2-Ag and W-GeS_2-Ag based CBRAM devices. Arrows indicate the difference in t_{set} of the two stacks to achieve the same percentage of CBRAM in LRS.

2.5 Conclusions

In this chapter, we provided a thermally activated hopping model to describe the ion migration and the consequent filament growth/dissolution during set/reset process in W-GeS_2-Ag based CBRAM devices. The dependence of LRS resistance and the reset current on the compliance current and the impact of the temperature on the switching parameters was also explained by the model corroborating the electrical results on 1R and 1T-1R devices. The extracted parameters were used to implement a compact model written in Verilog-A [82]. Next, by using a statistical analysis we analyzed the correlation between the programming conditions used in 8×8 memory array and the switching efficiency, providing an empirical explanation of the causes of variability in the switching parameters. Results based on the statistical analysis were used to provide a more reliable compact model that include the cell to cell variability in the memory array. The chapter ends with a comparison between W-GeS_2-Ag and Ta-GeS_2-Ag based

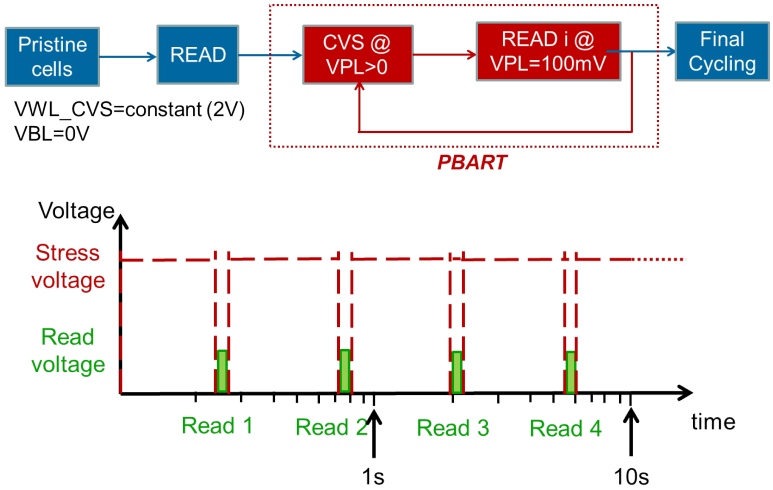


Figure 2.21: Schematic of the electrical test performed both on Ta-GeS$_2$-Ag and W-GeS$_2$-Ag 1T-1R devices in the memory array to understand the role of different cathodes in the t_{set}. A constant voltage stress is applied at the plate line (anode) to switch the devices from pristine state to LRS. Current is measured through logarithmically spaced readings.

CBRAM to gain new insights on the role of the inert electrode that rules the electron transfer and the nucleation of the Ag phase that lead to the filament growth and the switching from HRS to LRS.

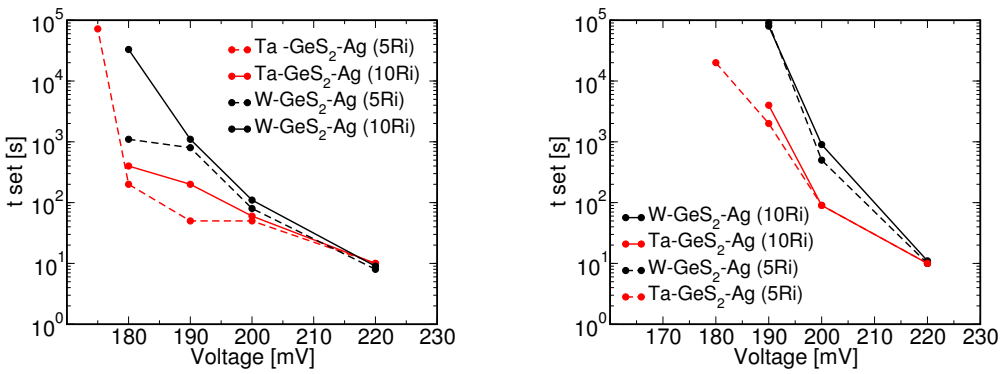


Figure 2.22: Time to set versus applied voltage in W-GeS_2-Ag and Ta-GeS_2-Ag based CBRAM devices extracted at 30% (left) and 45% (right) of the t_{set} distribution obtained on the 64 cells of the matrix. A changing of 5 (10) times the initial resistance (Ri) have been considered as switching criteria from HRS to LRS.

3

CBRAM stack engineering for increasing R_{off}

In hybrid (nonvolatile memory and logic) systems the circuit functionality is strongly dependent on the switching characteristics of the integrated ReRAM cells that in turn depend on the logic (i.e the type of the access MOSFETs) used to program the resistive device. The circuit performance can be enhanced by material engineering of the resistive switching memory cell, to satisfy the specific requirements of the targeted application. For example, a nonvolatile SRAM should integrate ReRAM cells with very low switching power both for set and reset, to be competitive with the consumption of a SRAM kept in stand-by at the Data Retention Voltage (DRV) (Sec. 1.5). According to our simulations (Sec. 4.1.2) a ratio $R_{\mathrm{off}}/R_{\mathrm{on}}$ of 100 between the HRS and LRS is enough for a reliable recovery operation considering our NV-SRAM architecture. Hence, Ag-GeS_2(30 nm)-W based CBRAM, characterized in Chap. 2, can be adopted for the development of NV-SRAM designs, because of a resistance ratio of 200 and no forming step. On the other hand, in the 1T-2CBRAM non volatile routing switch, proposed to control the pass gate in FPGA, the maximization of R_{off} is mandatory while the switching power consumption is not a primary concern (Sec. 1.9.2). This architecture requires research at device level to increase the R_{off} value, otherwise can not be competitive with existing (SRAM based) solutions. In the literature it is reported that the scaling of the device in the horizontal dimension is effective in increasing the R_{off} value [89]. Our test structures have a 200 nm plug (Fig. 2.2), thus the aim of this chapter is to find alternative solutions to maximize the resistance ratio in view of developing hybrid routing switches. Increase

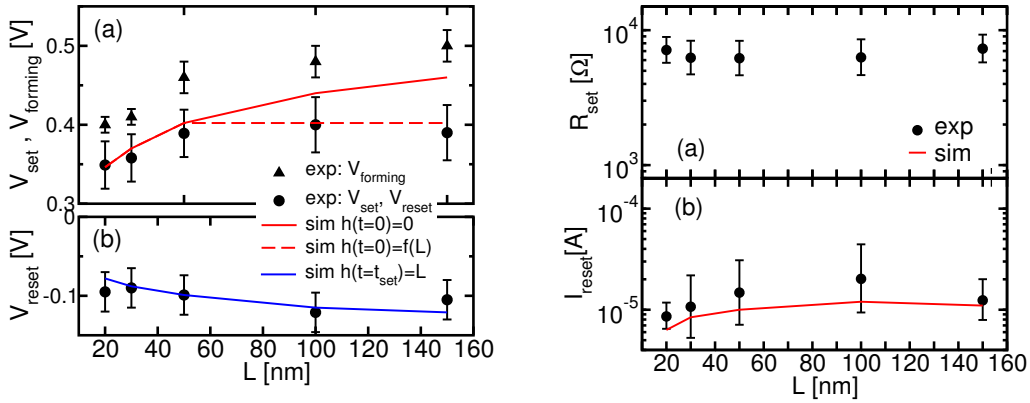


Figure 3.1: (Left) Experimental (symbols) and simulated (lines), V_{forming}, V_{set} (a) and V_{reset} (b) versus thickness. Two different initial conditions for solving Eq. (2.1) are used to take into account the increase of V_{forming} and the saturation of V_{set} observed for thicknesses between 50 and 150 nm. (Right) Experimental (symbols) and simulated (line) R_{set} (a) and I_{reset} (b) versus thickness. A compliance current of 30 μA is used. Constant R_{set} values are observed for the different samples.

of the chalcogenide thickness and the W-GeS_2 or Ta-GeS_2 interface engineering are the two solutions explored in this chapter. In particular, four different CBRAM cell stacks are investigated: i) W-GeS_2-Ag, ii) Ta-TaO_x-GeS_2-Ag, iii) W-SiO_x-GeS_2-Ag and iv) W-HfO_2-GeS_2-Ag.

3.1 GeS_2 thickness effects on the switching kinetics

First we investigated the impact of GeS_2 layer thickness (L) on the switching characteristics of W-GeS_2-Ag 1R based CBRAM cell both DC quasi-static and pulsed measurements were performed. Fig. 3.1 (left a) shows that V_{forming}, defined as the first V_{set} on a virgin cell, monotonically increase from 0.4 V for the device with $L = 20$ nm up to 0.5 V for the sample with $L = 150$ nm, while V_{set}, measured in the following cycles, saturates to approximately 0.4 V for thicknesses between 50 and 150 nm. These results may be interpreted assuming that for $L < 50$ nm, the CF in the reset state is almost completely dissolved, leading to a dependence between V_{set} and the thickness, while for higher L values a portion of the filament might subsist on the W electrode, or in the electrolyte as dispersed Ag-rich conductive clusters, acting as a new cathode in the following cycles [90, 70]: thus reducing the effective distance the ions have to

cover in the following set cycle to shunt bottom and top electrode. In particular, the set voltage saturation seems to indicate that during this reset process, up to $\simeq$ 50 nm of filament is dissolved, regardless of the GeS_2 thickness. To simulate the saturation on V_{set} (Fig. 3.1 left a) we analytically solve Eq. 2.1 with the following initial conditions:

$$h(t=0) = f(L) = \begin{cases} 0, & 0 < L < 50\ nm, \\ 50, & L = 100\ nm, \\ 100, & L = 150\ nm, \end{cases}$$

V_{reset} is less sensitive to L that is in agreement with the constant measured R_{set} (Fig. 3.1 right a). To validate our hypothesis of a non complete dissolution of the CF in the cells with thick active layer, we reset the cells in the sample with $L = 100$ nm applying two different DC conditions. Reset states were obtained through a reset sweep starting from the same initial set state (same compliance current) and interrupting the sweep at two different voltages (V_{stop}): namely, -0.5 and -1 V. I-V curves acquired are displayed in Fig. 3.2 (left) where we observed, in some cases, that for increasing V_{stop}, R_{off} increases from 3 MΩ to 50 MΩ and the following V_{set} from 0.3 V to 0.5 V. The growth of V_{set} with respect to V_{stop} for the sample with $L = 100$ nm can be explained by an increasing gap between the top electrode and the residual portion of the CF, due to the increasing final voltage in the reset sweep. On the contrary, sample with $L = 20$ nm exhibits a V_{set} almost indipendent of V_{stop}, thus confirming that the CF is dissolved for both the applied reset conditions. According to this test we observe a modulation of the off state in CBRAM devices based on GeS_2 thicknesses of 100 nm and 150 nm, but not for 20 nm, 30 nm and 50 nm. This result obtained in quasi-static measurements was not observed performing reset operation in pulse mode, where the R_{off} was around 5 MΩ for all the considered stacks. The possibility to tune the R_{off} value in quasi-static mode with a very long negative ramp is not of practical interest in designing hybrid architectures.

We also investigated the dependence of the t_{set} with respect to the chalcogenide thickness. Fig. 3.2 (right) shows the set time versus the applied voltage as obtained during programming in the pulsed mode. A reduction of t_{set} of about a factor 10 occurs when decreasing the GeS_2 thickness from 50 to 30 nm and similar results have been reported in [74]. This enforce the hypothesis that the CF after reset is completely dissolved, thus leading to the thickness dependence that is also confirmed by the model, that takes into account the distance that ions have to cover. Considering valid the

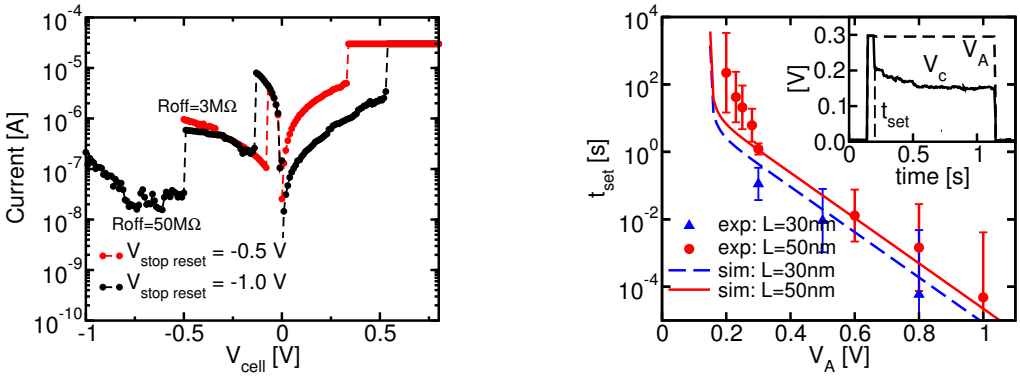


Figure 3.2: (Left) Experimental I-V electrical characteristics for GeS_2 (L = 100 nm) obtained stopping the reset sweep to -0.5 V (red dashed line) and -1 V (black dashed line). (Right) Experimental (symbols) and simulated (lines) switching time as a function of the applied voltage amplitude V_A on the cell. Inset: oscilloscope trace of V_A and V_c during a pulse-mode set operation. A reduction of t_{set} of about a factor 10 occurs when decreasing the GeS_2 thickness from 50 nm to 30 nm.

hypothesis of a non complete dissolution of the CF, we expect that the t_{set} would be almost constant for 100 nm and 150 nm even if dynamic measurements were not performed. To conclude, the increase of chalcogenide thickness in W-GeS_2-Ag based CBRAM devices seems not a suitable approach to increase the R_{off}, hence we were forced to search for other solutions.

It is worth to note that in the literature several reports show that dual-layer electrolytes based on buffer oxide layers and chalcogenides [91], [89], [81], [92] or buffer oxidizable metal layers and oxide [93] employed in ReRAM suppress leakage current (I_{off}) in the HRS, reduce the variability in set/reset voltages and improve data retention characteristics. In cycled Cu-$GeSe_x$-TaO_x-W based CBRAM, Cu ions were found in the buffer layer and the reduction of variability was due to an enhanced control and confinement of the nanofilament inside the TaO_x layer. Average resistance ratios of 85, 6.5×10^4 and 1.6×10^4 at compliance current of 1 nA, 50 µA and 500 µA were measured [77]. In [91] and [92], R_{off} of $2 \times 10^5\,\Omega$ and $10^7\,\Omega$ in Cu-GeSe-Ta_2-O_5-W and Cu-Cu:GeSe-SiO_x-Pt were measured respectively. From these promising results we decided to investigate the impact of buffer layers at the Ta-GeS_2 and W-GeS_2 interfaces.

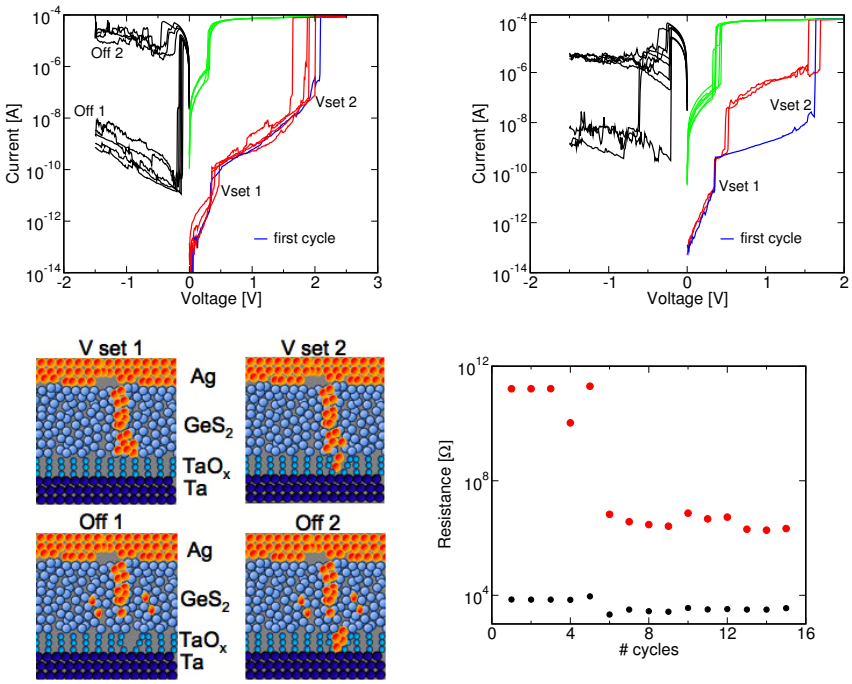


Figure 3.3: (Up) DC I-V characteristics for 1T-1R Ta-TaO$_x$-GeS$_2$-Ag based CBRAM devices. A compliance current of 100 µA (left) and 130 µA (right) was used. (Down) Schematic of the double set process (left). R_{off} and R_{on} values, obtained through low-field measurements, corresponding to the I-V characteristics obtained with a compliance current of 100 µA (right).

3.2 Ta-TaO$_x$-GeS$_2$-Ag based CBRAM

In this section we present a CBRAM stack that consists of Ta-TaO$_x$-GeS$_2$-Ag as possible solution to increase the R_{off}. After the definition of the plug the Tantalum was oxidized and the final thickness of the TaO$_x$ was 2 nm. A 30 nm thick GeS$_2$ electrolyte was deposited by RF-PVD. DC I-V characteristics are reported in Fig. 3.3 for 1T-1R devices using a current compliance of 100 µA (up left) and 130 µA (up right). It is interesting to note that during the first cycles a first set appears at around 0.5 V and a second set around 2 V or 1.5 V. The following cycles show that the device switches at the same voltage of the first set ($V_{\text{set 1}}$). Only few reset cycles show a very low I_{off}, then the ratio $R_{\text{off}}/R_{\text{on}}$ is strongly reduced because of the decrease of R_{off} (Fig. 3.3 right down). The double set voltage could be explained with a soft breakdown of the GeS$_2$ layer followed by the breakdown of the TaO$_x$ layer at higher voltage. We believe that Ag ions

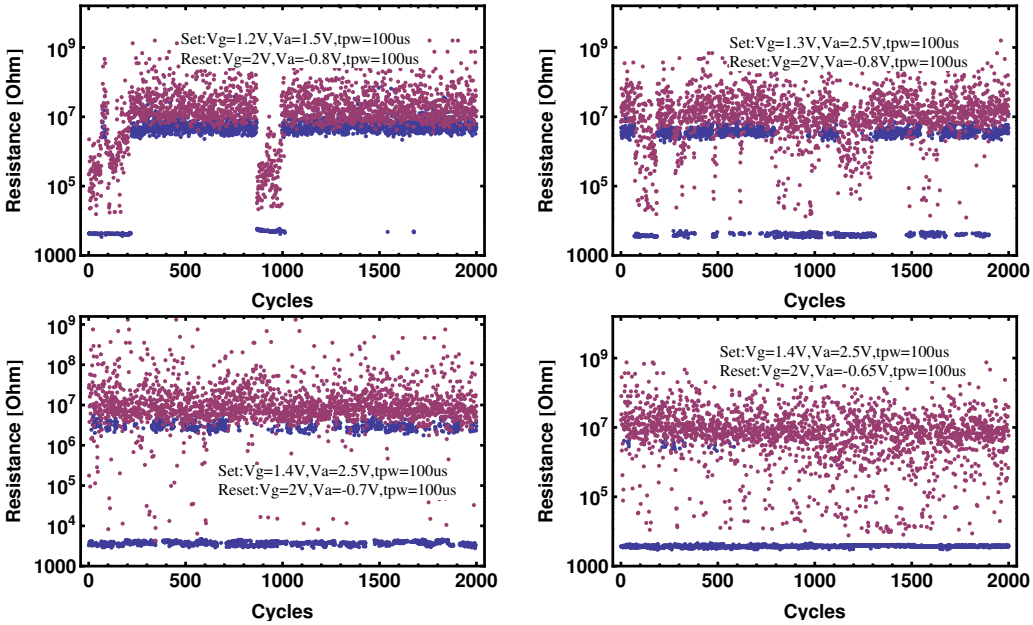


Figure 3.4: Stochastic switching of 1T-1R Ta-TaO_x-GeS_2-Ag based CBRAM devices during 2000 cycles using progressively stronger set conditions (or weaker reset conditions). The set probability increases from 16% (left up) to 99% (right down).

diffuse into the TaO_x layer degrading the oxide after few set/reset cycles, thus the role of the TaO_x as a barrier for the conduction in the off state is ineffective (Fig. 3.3 left down). The effect of weak and strong conditions was also investigated on isolated 1T-1R devices through endurance tests in pulse mode to corroborate the results obtained on the memory arrays of W-GeS_2-Ag. Fig. 3.4 shows a progressive increasing of successful set operations applying 4 different set/reset conditions on the same device. Note the increase in the set probability depending on the programming conditions: from weak (left up) to strong (right down) that lead to a set probability of 99%.

3.3 W-SiO_x-GeS_2-Ag based CBRAM

In [92] a Cu-Cu:GeSe-SiO_x-Pt based CBRAM was fabricated and the typical $R_{\text{off}}/R_{\text{on}}$ ratio of the memory cells written with programming current 100 µA was found around 10^4. For this reason, we decided to fabricate a CBRAM stack that consists of W-SiO_x-GeS_2-Ag. The thickness of the GeS_2 and SiO_x layer were 30 nm and 3 nm respectively. Both layers were deposited by RF-PVD. DC I-V characteristics are reported in Fig. 3.5

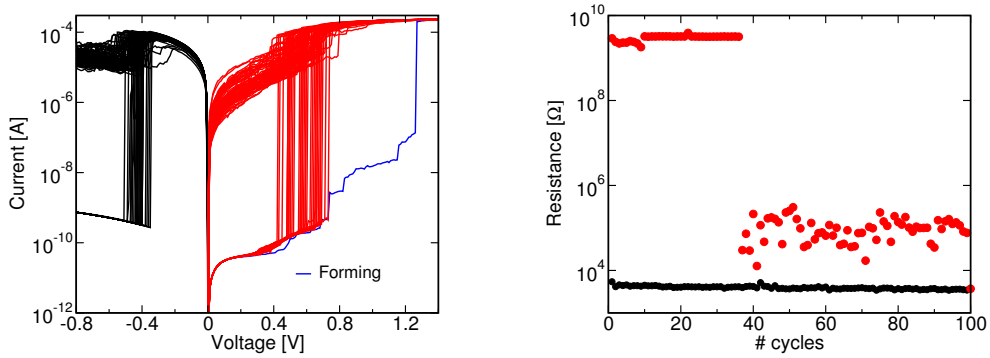


Figure 3.5: (Left) DC I-V characteristics for 1T-1R W-SiO$_x$-GeS$_2$-Ag based CBRAM devices (100 cycles). (Right) Evolution of R_{on} and R_{off} values by cycling of the device. Low-field measurements.

(left) for 1T-1R devices using a current compliance of 200 µA. The CBRAM cell requires a forming operation at a voltage of 1.3 V. The first 40 cycles shows a very high $R_{\text{off}}/R_{\text{on}}$ ratio of 10^6, then the memory window is reduced and the R_{off} stabilizes around 100 kΩ. Fig. 3.5 (right) also shows that the on resistance slightly decreases because of the degradation of the oxide. Forming free CBRAM stacks are more suitable for embedded applications, because of the fully logic compatibility that can avoid charge pumping circuitry, thus gaining area and reducing cost.

3.4 W-HfO$_2$-GeS$_2$-Ag based CBRAM

In the last section of this chapter we investigate three different CBRAM stacks. The first stack is the canonical W-GeS$_2$-Ag based CBRAM here characterized as a reference. The second and the third stacks consisted of an additional 1 nm and 2 nm HfO$_2$ layer inserted between the W-plug and the 30 nm GeS$_2$ layer respectively (Fig. 3.6). The HfO$_2$ layer was deposited by atomic layer deposition (ALD). We performed electrical measurements both on isolated 1T-1R devices and in 8×8 memory array. Typical DC-sweep current-voltage (I-V) curves are shown in Fig. 3.7. Mean set/reset voltages of 0.35 V/−0.2 V (GeS$_2$) (Fig. 3.7 left up), 0.4 V/−0.3 V(HfO$_2$(1 nm)+GeS$_2$) (Fig. 3.7 right up), 0.5 V/−0.4 V (HfO$_2$(2 nm)+GeS$_2$) (Fig. 3.7 left down) are reported. Interestingly no forming step was required even for the HfO$_2$(2 nm)+GeS$_2$ sample. The compliance current (I_{comp}) was fixed to 240 µA during the set operation and the measured reset current (I_{reset})

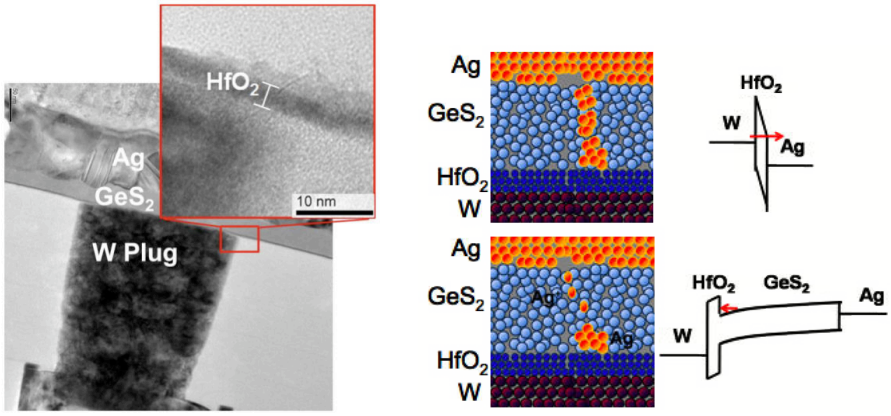


Figure 3.6: (Left) TEM cross section of HfO_2(2 nm)+GeS_2 based CBRAM device. (Right) Schematic of the set and reset operations and qualitative band diagram assuming GeS_2 a p-type semiconductor and Φ(W)>Φ(Ag).

was 100 µA. This implies that the MOSFET is effective in limiting the current through the CBRAM during the set event avoiding current overshoot. While R_{on} is around 5 kΩ for the three stacks, R_{off} significantly increases with the HfO_2 barrier insertion. A resistance ratio of two, five and more than six orders of magnitude were obtained in GeS_2, HfO_2(1 nm)+GeS_2 and HfO_2(2 nm)+GeS_2 respectively, through low-field (0.1 V) ramp measurement (Fig. 3.8 left). On 65 quasi-static and more than 1k cycles in pulse mode we did not observe any degradation of the R_{off} value in HfO_2(2 nm)+GeS_2 memory device. We believe that the working principle of the proposed memory stack relies on the reversible formation of the CF inside the GeS_2 layer without diffusion of Ag^+ ions in the HfO_2 layer. The Ag^+ ion hopping process in the HfO_2 could be suppressed by higher energy barriers in the potential energy surface of HfO_2 with respect to GeS_2. We used a physical model to gain better insight on the improved memory ratio of the optimized dual layer CBRAM devices. The model takes into account both direct tunneling (DT) and the multi phonon Trap Assisted Tunneling (TAT) as conduction mechanisms in the HRS simulated as a metal (W) - insulator (HfO_2) - semiconductor (GeS_2) (MIS) structure [94], [95]. The GeS_2 was assumed to be a p-type semiconductor with a carrier concentration of $6 \times 10^{17}\,\text{cm}^3$. In the LRS, the structure relies on a metal (W) - insulator (HfO_2) - Metal (Ag), (MIM) structure, because of the metallic behavior of the CF inside the GeS_2. The DT current was calculated through the semi-classical approach in the WKB approximation [96]. In the TAT model, the capture and emission

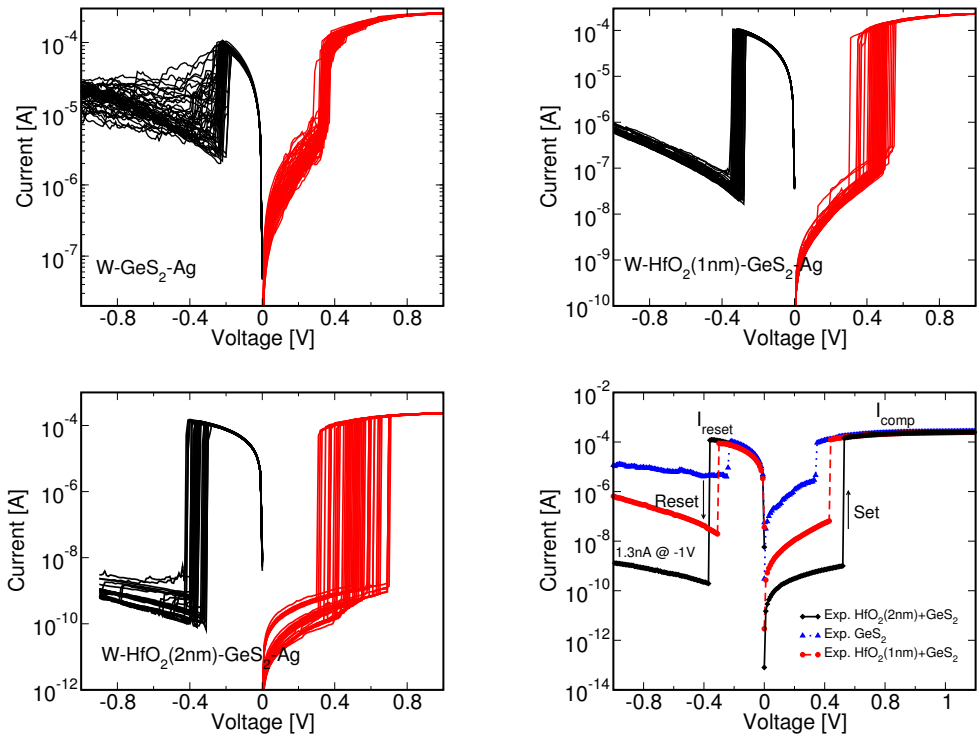


Figure 3.7: DC I-V characteristics for GeS_2 (left up), HfO_2(1 nm)+GeS_2 (right up) and HfO_2(2 nm)+GeS_2 (left down) based CBRAM devices. Comparison of the three stacks (right down).

rates are calculated by accounting for both the electron-phonon coupling and the lattice relaxation required to accommodate the trapped charge during capture and emission events [94], [97]. Fig. 3.8 (right) shows the simulated and experimental I-V curves of the HfO_2(1 nm)+GeS_2 and HfO_2(2 nm)+GeS_2 cells in the HRS and in the LRS. Calculations to evaluate the total (DT+TAT) current revealed that the increase of HfO_2 barrier from 1 nm to 2 nm reduces the I_{off} current of 1.5 orders of magnitude. On the other hand, the contribution of the HfO_2 barrier on the total current in the on-state is almost negligible because of the increased carriers concentration. For this reason, in the on-state HfO_2 acts as a transparent barrier with respect to the conduction.

Ramped voltage-pulse measurements on 8×8 array of 1T-1R devices were performed to identify the set/reset voltages conditions. In all cases, the pulse width (t_{pw}) used was 100 µs. Note that between every applied set pulse, we reset the devices with a constant reset pulse and vice-versa. Applied voltages/time and polarization lines are

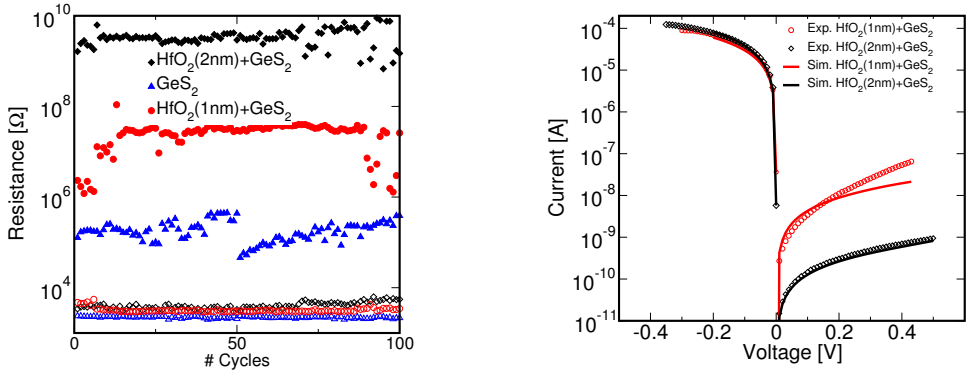


Figure 3.8: (Left) Resistance values R_{on} and R_{off} after set and reset respectively. $R_{\text{off}}/R_{\text{on}}$ increases accordingly to the thickness of HfO_2 barrier. In HfO_2(2 nm)+GeS_2 $R_{\text{off}}/R_{\text{on}}$ of 10^6 is demonstrated. (Right) Experimental and simulated current in the HRS and in the LRS before the set and the reset event in HfO_2(1 nm)+GeS_2 and HfO_2(2 nm)+GeS_2 devices. In the HRS, simulated current is the sum of direct tunneling (DT) and trap assisted tunneling (TAT) contributions, through a MIS structure and increasing the barrier thickness reduces the leakage current of 1.5 orders of magnitudes. HfO_2 traps have been modeled according to [94], [97]. In the simulated LRS, because of the increased carriers concentration, the difference in barrier thicknesses do not lead to distinct current levels.

summarized in Tab. 3.1. A read operation was performed to verify the resistance value of the CBRAM after every set/reset operation. The purpose of these measurements was to identify the programming conditions for an optimized resistance ratio. First, we study the impact of the set conditions on the LRS values. Fig. 3.9 (left up) shows that, once the LRS state achieved, the increasing of V_{anode} from 1.4 V to 2.4 V is not effective in modulating the R_{on} values in GeS_2 and HfO_2(1 nm)+GeS_2 CBRAM devices. However, in HfO_2(2 nm)+GeS_2, R_{on} slightly decreases with the applied voltage, showing a weak modulation from 7 kΩ to 3 kΩ. Fig. 3.9 (right down) shows that using different gate voltages (V_{gate}) LRS can be modulated in a range of values from 100 kΩ to 2 kΩ for all the three stacks. A slightly higher R_{on} was measured in HfO_2(2 nm)+GeS_2 when V_{gate} ranges from 1 V to 2 V. The decreasing of R_{on} in Fig. 3.9 (right down) can be explained, considering that a larger compliance current would supply more electrons to reduce more Ag^+ in the electrolyte, thus creating a larger CF during the set process [72]. An estimation of the current flowing through the CBRAM just after the set event is reported on the top axis of Fig. 3.9 (right down). Further, we

Table 3.1: Programming conditions used in 8×8 NOR memory array and isolated 1T-1R devices.

	8×8 Matrix							
	SET				RESET			
	V_{gate} [V]	V_{anode} [V]	t_{pw} [µs]	V_{bitline} [V]	V_{gate} [V]	V_{bitline} [V]	t_{pw} [µs]	V_{anode} [V]
Fig. 3.9 (left up)	1.5	1.4→2.4	100	0	2.5	2	100	0
Fig. 3.9 (right down)	0.6→2	2.5	100	0	2.5	2	100	0
	SET				**RESET**			
Fig. 3.9 (left down)	1.5	2.5	100	0	1.6→2.6	2.5	100	0
Fig. 3.9 (right up)	1.5	2.5	100	0	2.5	1→2.4	100	0
	1T-1R							
Fig. 4.12	1	2.5	100	0	2.5	2	10	0

investigated the dependence of resistance levels in HRS with the applied reset condition. In GeS_2 and HfO_2(1 nm)+GeS_2 we found that for applied pulse on the gate higher than 2.2 V R_{off} saturates at $6 \times 10^5\,\Omega$ and 2×10^6 respectively. On the contrary, in HfO_2(2 nm)+GeS_2, R_{off} achieves a resistance of $2 \times 10^8\,\Omega$ when 2.6 V is applied. We performed Eldo simulations using the model of the integrated 130 nm MOSFET, to evaluate, for different values of V_{gate}, the effective voltage drop on the CBRAM cell before the reset event. The CBRAM cell resistance value used in the simulations was 10 kΩ. Results are shown on the top axis of Fig. 3.9 (left down). A voltage drop of about 1.3 V on HfO_2(2 nm)+GeS_2 based CBRAM is required to obtain an R_{off} of $2 \times 10^8\,\Omega$. Fig. 3.9 (right up) shows that R_{off} saturates for bitline voltages higher than 1.8 V. This is related to the saturation of the voltage drop on the CBRAM when the bitline voltage lies between 1.8 V and 2.4 V for a fixed V_{gate} of 2.5 V (Fig. 3.9 top axis). From the above considerations the HfO_2(2 nm)+GeS_2 stacks was selected as the most promising candidate for the 1T-2R NVE application. In Sec. 4.4 we will discuss in more details the architecture and the foreseen advantages with respect to existing solutions.

3.5 Conclusions

In this chapter some of the electrical performances of CBRAM cells were studied through DC I-V quasi static measurements and dynamic measurements. Interface engineering effects in W-GeS_2 or Ta-GeS_2 based CBRAM were analyzed. Among other measured CBRAMs technologies, our dual-layer electrolyte stack (2 nm HfO_2−30 nm GeS_2) leads to a resistance ratio ($R_{\text{off}}/R_{\text{on}}$) higher than 10^6, reset current of 100 µA

using a compliance current of 240 µA without forming step. We explained the improved memory resistance ratio by means of physical modeling. Expected benefits of this technology on 1T-2R NVE architecture will be discussed in the next chapter. It is worth to note that our HfO_2 based CBRAM paves also the way for multi-level storage in high density applications.

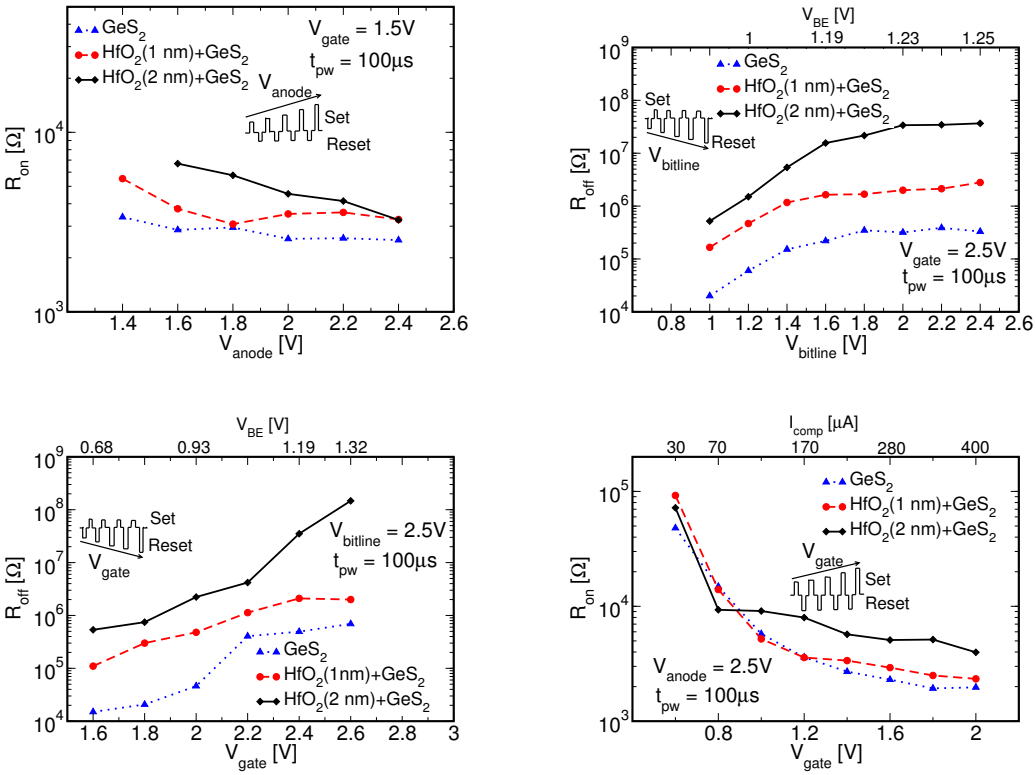


Figure 3.9: (Left up) Measured LRS for the three different studied CBRAM stacks as a function of the voltage applied at the anode during the set operation. (Right up) Measured HRS for three different studied CBRAM stacks as a function of the voltage applied at the bitline during the reset operation. Effective voltage drop calculated for HfO_2(2 nm)+GeS_2 before the reset event is shown on the top axis. (To calculate V_{BE} we used an R_{on} =10 kΩ). (Left down) Measured HRS for three different studied CBRAM stacks as a function of the voltage applied at the gate during the reset operation. Effective voltage drop calculated for HfO_2(2 nm)+GeS_2 before the reset event is shown on the top axis. (Right down) Measured LRS for the three different studied CBRAM stacks as a function of the voltage applied at the gate during the set operation. R_{on} can be reduced by increasing the voltage on V_{gate}. Compliance current flowing in the CBRAM just after the set event and corresponding to V_{gate} is shown on the top axis.

4

Nonvolatile hybrid (logic and ReRAM) architectures

Back-End-of-the-Line integration, CMOS compatible voltage, fast programming time, low power consumption and scalability are interesting features of ReRAM that are pushing the development of new hybrid architectures. These architectures are also expected to boost specific markets such as low power (mobile) embedded (Sec. 1.5), reconfigurable logic (Sec. 1.9) or even in neuromorphic computing (Sec. 1.10). For this reason, in this chapter, we will propose and discuss three architectures that belongs to the three categories cited above. The first design that we will present is a nonvolatile SRAM based on OxRRAM devices. We are going to show a methodology to verify the robustness of this architecture in the case of variability of the logic. We will use an advanced design kit (22 nm FDSOI) to set up worst case and Monte-Carlo simulations in order to understand the critical points for reliable operations in this architecture. This analysis will give us some indications on the most suitable ReRAM technology for such application and the constraints on the logic. Hence, experimental results that we obtained in Chap. 2 and 3 helped us in suggesting the most appropriate technology. Based on these results, a new NV-SRAM based on CBRAM has been designed, simulated using our compact model in CADENCE environment workflow, and finally fabricated due to a bilateral project with ALTIS semiconductor. Next, we will discuss a Nonvolatile element (1T-2R NVE) that could be used to control a pass gate in an FPGA switching block or to store a data in a Look-Up Table (LUT). In this case, advantages in using our optimized HfO_2/GeS_2 CBRAM will be highlighted. Finally, a circuit will be proposed

that exploits the unavoidable intrinsic variability occurring in CBRAM devices. This architecture, which can implement stochastic firing and might be useful for stochastic hardware neural network, will be simulated and discussed.

4.1 OxRRAM based nonvolatile SRAM (8T2R NV-SRAM)

The structure of the nonvolatile 8T2R NV-SRAM cell is presented in Fig. 4.1. The cell is designed with a 6T-SRAM cell (M1-M6) along with two additional p-type control transistors (CM1 and CM2) connected between the data nodes (D,DN) of the SRAM cell and the OxRRAMs (R1, R2). R1 and R2 are made accessible to the SRAM cell by CM1 and CM2. TE and BE represent the top and bottom electrodes of the OxRRAMs. Conventionally, a store (set) operation on the OxRRAM corresponds to a logical 1 and the reset corresponds to a logical 0. All the simulations were performed using Eldo simulator. The results are obtained using our 22 nm CMOS–FDSOI process design kit. A behavioral bipolar OxRRAM model, calibrated on the experimental results obtained on TiN/HfO_2/Ti based OxRRAM devices was used (Fig. 4.2). Accordingly to the model the threshold voltage for the OxRRAM store and reset operations are 0.7 V and −0.7 V respectively. This is because a set voltage of 0.7 V has been obtained applying a voltage ramp of $10^7\,\mathrm{V\,s^{-1}}$ (Fig. 4.2 right). No pulse measurements were available at the time of developing the model, thus the set operation is determined by the rising part of the applied pulse (typically around $10^8\,\mathrm{V\,s^{-1}}$). Moreover, the model predicts a continuos change of the resistance from the HRS to the LRS and the final LRS is determined by the current flowing into the OxRRAM. On the contrary, the switching from LRS to HRS is abrupt and the HRS has been fixed to 88 kΩ. This value was obtained as a mean of low-field measurements during quasi-static DC measurements. In the proposed 6T-SRAM the ratio W/L of pull-up (M2 and M4), pull-down (M1 and M3), transfer transistor (M5 and M6) and control transistors (CM1 and CM2) of the NV-SRAM are set to: 100 nm/25 nm, 260 nm/25 nm, 100 nm/25 nm, and 220 nm/25 nm, respectively (Fig. 4.1). The transistors are operated at a voltage of 1.1 V accordingly to process design kit. The NV-SRAM cell operation follows the sequence: normal (read/write), reset (switching from LRS to HRS), store (switching from HRS to LRS), power-down, power-up and restore.

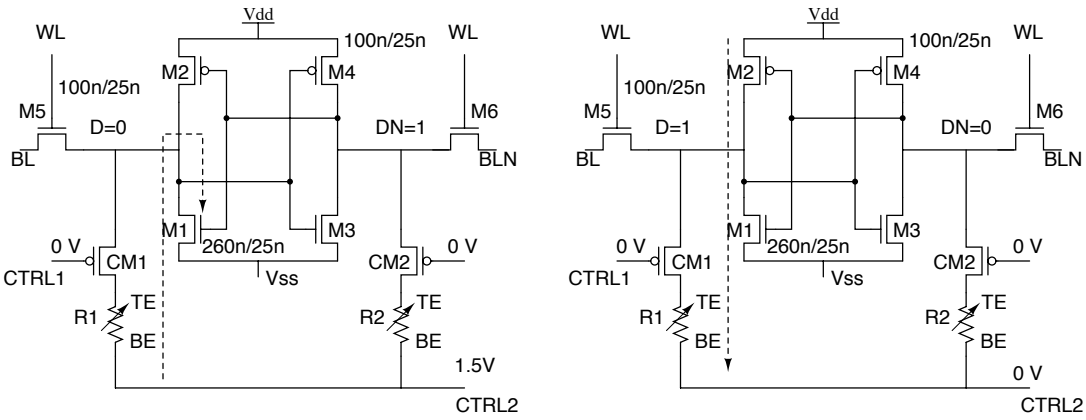


Figure 4.1: Schematic representation of Nonvolatile 8T2R SRAM (NV-SRAM) cell. Dashed lines show the current path during RESET operation (left) and store operation (right).

Before storing the information in the OxRRAMs, the reset operation was simulated to ensure the HRS for the corresponding OxRRAM device. When the nodes D and DN are equals to 0 and 1, respectively, R1 can be resetted by turning on the control transistors CM1 and CM2 and putting CTRL2 line to 1.5 V. At this point, the resistance of R2 remains unchanged due to the low voltage drop on it. At the same time, R1 faces a negative voltage drop (0.8 V) that, if required, switches the OxRRAM into the HRS. Similar to the R1 case, R2 can be reset if the SRAM nodes are flipped and the sequence is repeated.

During the store operation, the logic state of the SRAM cell is stored in the OxRRAMs R1 and R2. Let us assume that the node D has 1 and DN has 0. To backup the information in the OxRRAM, the control transistors are turned on by lowering CTRL1 to 0 V and grounding CTRL2. The potential difference at the top and bottom electrode of R1 results in a positive voltage drop across R1 which sets the resistance to a LRS thereby, storing the information in the OxRRAM (Fig. 4.3). At the same time, since there is no voltage drop across R2, its value remains unchanged in high resistive state thereby storing the value 0. We used a pmos instead of a nmos to build a sufficient voltage on the TE net.

Normally, a power-down in a circuit is performed by putting all the control lines to ground. In our simulations, however, the power-down operation is performed by raising the VSS pin of the cell to VDD and hence, the nodes D and DN are pulled up to VDD.

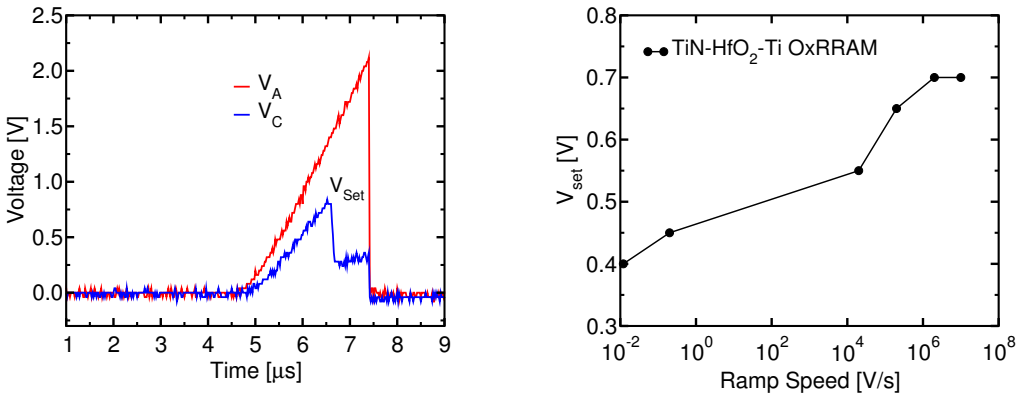


Figure 4.2: (Left) Voltage ramp of $10^6\,\mathrm{V\,s^{-1}}$ applied to 1R-OxRRAM devices. The Set Voltage is 0.7 V. The experimental setup used has been defined in Fig. 2.5. (Right) Set voltage V_{Set} as a function of voltage ramp speed.

This is done to avoid resetting the state of the OxRRAM during the restore operation that would cancel the resistance asymmetry between R1 and R2 and could compromise the correct level restoration. The control lines BL/BLN and CTRL2 are also raised to VDD to reduce the leakage through the transfer transistors and control transistors, respectively.

The logic state of the SRAM has to be restored at power-up according to the following sequence (Fig. 4.3): CTRL1 is lowered to ground turning on the control transistors, VSS is pulled down to 0 V with a delay of 5 ns with respect to CTRL1, and CTRL2 is kept at VDD as it was during power-down. As VSS is pulled down to 0 V, the resistance difference results in different discharging currents and difference in voltage between D and DN, which is amplified by the SRAM latch. The low resistance of R keeps D node at 1 while DN in pulled down to 0 through M3 due to the high resistance of R2. In this way the logic levels in NV-SRAM are restored.

Before a new store operation the OxRRAM in the LRS must be switched to HRS. This is possible forcing a logical 0 in the latch side where the OxRRAM is in LRS and then raising CTRL2 with the control transistor in on state.

4.1.1 8T2R NV-SRAM cell static noise margin

The 6T-SRAM cell core of the 8T2R NV-SRAM has to be optimized such that even under the worst-case conditions, it still functions properly. The sizing of the 6T core of

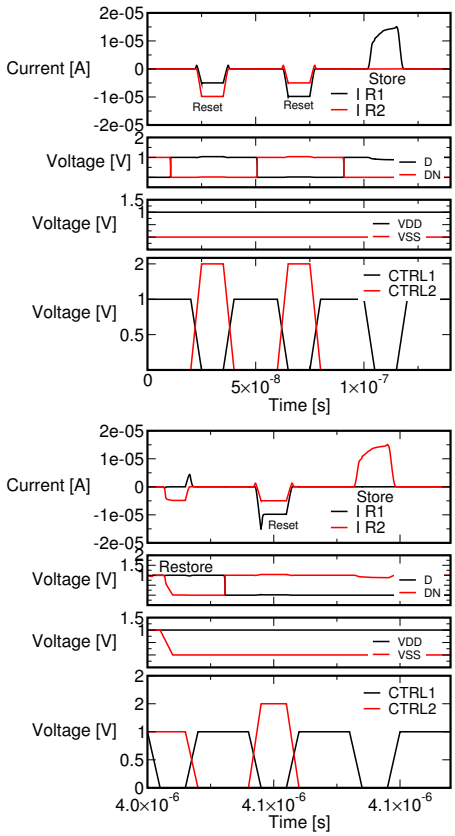


Figure 4.3: (Up) Transient analysis of the OxRRAM based NV-SRAM to demonstrate reset and store operations. The Store operation starts at 100 ns. The maximum current flowing into the OxRRAM is 16 µA. (Down) Restore of the data after power-down. The RESET operation at 4.04 µs shows that in the HRS state a current of 12 µA still flows in the OxRRAM (R1) before the end of the pulse on CTRL1. A new store operation is simulated on R2.

the 8T2R cell does not comply with the typical sizing of the pull-up pmos in 6T-SRAM, which is usually kept at minimum due to area and stability constraints. The reason for having a wider pmos transistor in the 8T2R NV-SRAM comes from the requirement of a current of a few µA in order to write the OxRRAM during the store operation as predicted by the model used. Since in the 22 nm FDSOI technology used in this work, the pmos has a high threshold voltage of approximately 500 mV, the store current could only be met by increasing the pmos width. However, a lower threshold for pmos would lead to a decrease of the sizing of these transistors. The pull-up and the transfer transistors are considered to have the same strength to enable a reasonable value for

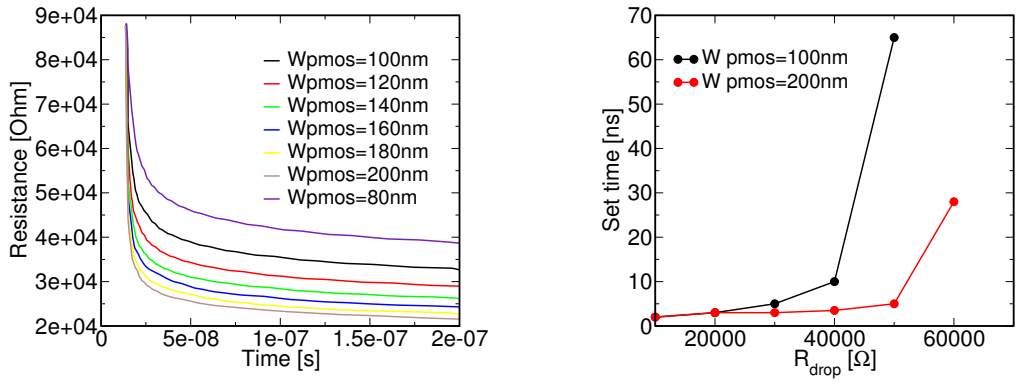


Figure 4.4: (Left) R_{on} time evolution during store operation. The width of the pmos M2 is sampled in a range between 80 nm and 200 nm. (Right) Store time as a function of R_{drop}.

the Read Static Noise Margin (RSNM) and Write Static Noise Margin (WSNM). The Static Noise Margins (SNMs) are key parameters in the SRAM analysis and can be defined as the highest value of noise between the two cell inverters for which the proper functionality in each operation mode is maintained. The SNMs can be defined as the largest (for retention and read) or the smallest (for write) square that can be fitted between the butterfly curves consisting of the inverted and non-inverted curves corresponding to respective operation mode. With the above mentioned sizing, the NV-SRAM cell has a pull-up ratio of 1 (defined as the ratio of W/L between the pull-up and transfer transistors and a cell ratio of 2.6 (defined as the ratio of W/L between the pull down and transfer transistors) ensuring a RSNM of 180 mV and WSNM of 320 mV. The addition of the control transistors and the OxRRAM does not affect the normal operation of the SRAM cell as well as the SNMs. This is because of the use of CM1 and CM2 which block the current leakage path through the OxRRAMs during the normal operation.

4.1.2 The influence of V_T variability on the stability of 8T2R NV-SRAM

The impact of V_T variability on the NV-SRAM operation is expected to be a major constraint for obtaining a high yield. This is caused both by the potential difficulty in ensuring reliable NV-SRAM specific operation modes, if mismatch is considered and by the unorthodox 6T SRAM core sizing, with a larger than minimum pmos which

degrades the write stability in the typical SRAM operation. The key yield-limiting factor of the 8T2R OxRRAM based NV-SRAM is the reliability of the restore operation as it strongly depends on the $R_{\text{off}}/R_{\text{on}}$ ratio. Simulated store operation never fails as the positive voltage applied on an OxRRAM device in HRS is always going to lower its resistance by some value in a continuos model after the set threshold is reached. However, under the random V_T variations, the obtained R_{on} for a fixed operation time will vary and therefore, the restore reliability may be compromised. The optimization of store operation from the perspective of maximizing the cell $R_{\text{off}}/R_{\text{on}}$ under timing and transistor sizing (and hence stability) constraints becomes therefore the starting point for the 8T2R NV-SRAM reliability analysis.

In particular, the R_{on} value that can be obtained in the store operation is determined by two key parameters: the sizing of the pull-up transistor in the SRAM cell, here analyzed only through its width, and by the duration of a store operation (i.e the time in which the CM pmos is on). Fig. 4.4 (left) presents the dependence of the OxRRAM resistance as a function of time during store for different values of the width of the pmos in the latch. Accordingly to the behavioral model the decrease of resistance can be separated into two different parts: the abrupt resistance drop in the first 20 ns and the saturation of the resistance value as the store time increases.

The extracted R_{drop} as a function of the store time shows clearly a threshold-like behavior for both widths of the pull-up transistor. Once the decrease of the resistance enters the saturation region the time that is necessary for obtaining a significant R_{drop} exponentially increases (Fig. 4.4 right). A consequence of the results presented in Fig. 4.4 is that if a specific, high, R_{drop} from the R_{off} is targeted, either the store time or the width of the pmos have to be increased. The former solution however, would cause a significant increase in the power consumption during store operation. The latter solution would lead to the increase of the cell pull-up ratio, decreasing the SRAM cell reliability during write operation and to the increase of the current flowing through the OxRRAM. As depicted in Fig. 4.4 (right), if a significant R_{drop} is targeted, the use of a large width for the pmos can still lead to a more energy-efficient store operation due to a much shorter store time. In order to illustrate this, let us assume that a 35 kΩ R_{on} should be obtained and that the increase of the width linearly affects the current during store operation. Decrease of the width from 200 nm to 100 nm leads therefore to a 2× reduction in store current. At the same time though, such change of the width

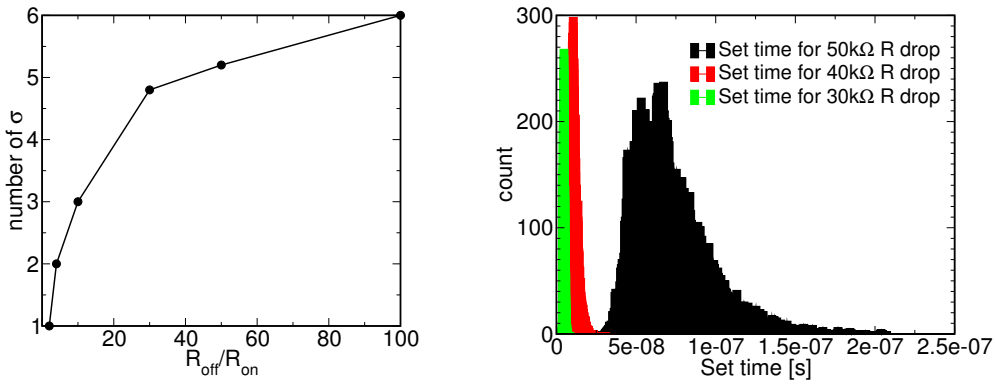


Figure 4.5: (Left) Variation of $\pm\sigma V_{\mathrm{T}}$ a V_{T} fail point as a function of the resistance ratio in the worst case variation analysis. (Right) Distributions of store time required to obtain various $\mathrm{R_{drop}}$ obtained from a 10k samples Monte Carlo simulation for pmos width of 100 nm.

increases the required store operation time from 16 ns to 110 ns. As a consequence, despite the lower magnitude of the current flowing through the OxRRAM, the energy of store operation increased by approximately 5× for pmos width = 100 nm as compared to the 200 nm. This behavior indicates that the store time is the dominant factor in power optimization and should be minimized, as was described in Sec. 1.5.

The magnitude stability drop caused by the larger width of the pmos will also depend on the initial $\mathrm{V_T}$ ratio between the nmos and pmos transistors, but should the width increase from the 100 nm to 200 nm, it will always be non-negligible. In our case, with the $\mathrm{V_{TP}}$ approximately 50% larger than the $\mathrm{V_{TN}}$ and the transfer nmos width of 100 nm, changing the width between 100 nm and 200 nm leads to the stability factor drop (represented as the μ/σ extracted from the statistical distributions of write static noise margin in such a way, that the tail is properly evaluated) from 12.3 to 9.22. In order to meet the typical 6σ yield target, the stability factor (μ/σ) should be higher than 6. In this particular case the SRAM yield is therefore still maintained, but the magnitude of this stability factor drop indicates a high importance of this analysis. The only way to avoid an excessively high pmos width and a long store time is therefore to reduce the targeted $R_{\mathrm{off}}/R_{\mathrm{on}}$. This is in turn limited by the minimum $R_{\mathrm{off}}/R_{\mathrm{on}}$ required to ensure the reliability of restore operation.

In order to investigate the stability in restore mode, a worst case ($\pm n\sigma V_{\mathrm{T}}$) analysis on V_{T} was applied to each transistor in the 6T-SRAM cell, thus creating a situation

where each device is working against the restore of the correct values in the nodes of the SRAM cell. Increasing the value of n leads to a larger worst case mismatch and allows analyzing the $\pm n\sigma V_{\mathrm{T}}$ point where the proposed restore operation fails. The data restore operation in our proposed scheme does not affect the resistance value of the OxRRAM, so the use of passive elements should correspond exactly to the case with OxRRAM devices. Moreover, substituting the OxRRAM devices with resistors allows including $R_{\mathrm{off}}/R_{\mathrm{on}}$ cases not covered by our OxRRAM model. Fig. 4.5 (left) depicts the values of $\pm n\sigma V_{\mathrm{T}}$ fail points for different $R_{\mathrm{off}}/R_{\mathrm{on}}$ ratios. The reference resistance value in this analysis was the R_{off}. Decrease of R_{on} aids the restore operation therefore in two ways: lower R_{on} means larger current and stronger impact on data restore and larger difference between R_{off} and R_{on} increases the current difference on both sides of the cell. Both phenomena increase the stability of restore operation, exemplified by a larger n. Accordingly to the model used, even for a pmos width of 200 nm and the delay of 200 ns, R_{on} during store reaches only 21 kΩ, corresponding to a ratio $R_{\mathrm{off}}/R_{\mathrm{on}}$ of four. The worst case mismatch analysis reveals that for this resistance ratio the restore operation fails for n= 2 (Fig. 4.5 left) which is low from the stability point of view.

In order to correlate the worst-case analysis result with a typical yield evaluation, a Monte Carlo (MC) simulation with random Gaussian variation applied to all transistors of the SRAM cell was performed. This MC simulation was performed for a $R_{\mathrm{off}}/R_{\mathrm{on}}$ of four corresponding to n=2, as this is the maximum resistance ratio obtainable in our 8T2R NV-SRAM. In this analysis, 41 failures in 10k samples were observed, corresponding approximately to a 3σ yield. Since in the typical SRAM design, a yield of 5 to 6σ is targeted, it can be expected that in the worst-case analysis the value of n should be equal to at least 3 or 4. However, as depicted in Fig. 4.5 (left), the value of n equal to 3 and 4 occurs only for a $R_{\mathrm{off}}/R_{\mathrm{on}}$ of 10 and 20, respectively.

Yet another constraint for a reliable OxRRAM-based SRAM operation becomes evident in the Monte Carlo analysis of the store mode. Fig. 4.5 (right) depicts a set of histograms from a 10k sample Monte Carlo analysis for a pmos width of 100 nm demonstrating the spread of store time required to obtain a $\mathrm{R}_{\mathrm{drop}}$ between 30 kΩ and 50 kΩ. Clearly, if a low R_{on} is targeted to satisfy restore conditions, the spread of the required store time increases significantly, especially as the saturation region (above 20 ns in Fig. 4.4) of the resistance curve is reached.

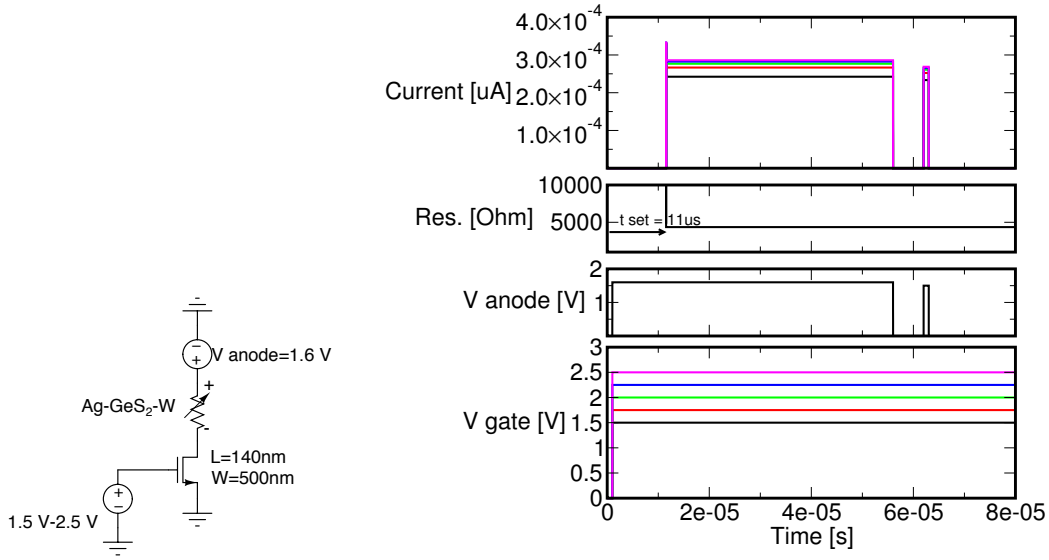


Figure 4.6: Transient parametric analysis of 1T-1R structure that shows both the time required to set the CBRAM and the final resistance value. No modulation of R_{on} is obtained increasing the voltage on the gate.

The stability of the restore operation, as demonstrated before, strongly depends on the minimization of the R_{on} because the R_{off} was fixed by the model and by the selected OxRRAM technology. In order to ensure that the expected, low R_{on} is obtained, the length of the CTRL1 signal should be extended accordingly up to the worst case store time in the far tail of statistical distribution (Fig. 4.5 right). The minimization of the difference between the mean and the worst case far tail store times is therefore important to reduce the power overhead coming from the operation time extension required for reliable restore. As a result, the only efficient method to obtain a high ratio and hence a high restore stability, is either by increasing the width of the pmos, or improving the OxRRAM or use another type of resistive memory with higher R_{off} as will be described in Sec. 4.3.

4.2 Simulations of 1T-1CBRAM structure

Development of behavioral compact models for a specific CBRAM technology usually require a large variety of logic and CBRAM structures to achieve a good comprehension of the device response to programming conditions or to size and/or type of the in series

transistor. For example in [98] a specialized integrated circuit was used to program the CBRAM devices. In this work we dispose of 1R devices, 1T-1R devices and 8×8 memory array. The model described in Sec. 2.1.4 was refined with electrical results on 8×8 memory array to include the cell to cell variability. The time required to set the device was assumed to be the pulse width applied at the gate of the transistor in series (Sec. 2.2). This is a conservative assumption that does not correspond to the real set time, which can be precisely measured as described in Sec. 2.1.3. The efficiency curves were used to establish the conditions that give more than 70% set/reset efficiency on 64 devices (Sec. 2.2). In implementing the model we defined a pulse width of 60 µs and a voltage on the anode of 1.5 V as the programming condition that gives 70% set efficiency on 64 devices (Fig. 2.14 left). Fig. 4.6, obtained through a parametric transient analysis, shows a t_{set} of 11 µs when the voltage on the anode is raised to 1.6 V. Thus the model predicts a lower t_{set} when the voltage on the anode is raised and the set efficiency is around 80%. Modulation of R_{on} with respect to V_{gate} has not been taken into account so far in the proposed model. The R_{on} value (CBRAM switched in the LRS) is 4300 Ω (Fig. 4.6). In a similar way, to establish reset conditions, we defined a pulse width of 10 µs and a voltage on the bitline of 2.5 V. This condition gives 95% reset efficiency on 64 devices (Fig. 2.16 left). Fig. 4.7 shows that the time required to reset the CBRAM is 50 µs when the voltage on the bitline is 1.5 V. The model still predicts a t_{reset} higher when the voltage on the bitline is decreased. The R_{off} value (CBRAM switched in the HRS) is 8.8 MΩ.

4.3 CBRAM based 8T2R NV-SRAM

In Sec. 4.1.2 we analyzed the critical points in designing a nonvolatile SRAM with respect to stability of the normal operations (read/write margins) and reliability of specific operations such as store operation and data restore after power up. One important conclusion was done on the ratio R_{off}/R_{on} that should be greater than 100 for a 6σ yield as we predicted using the FDSOI 22 nm technological node design kit. Hence, using our W-GeS_2-Ag based CBRAM device we re-designed a 8T2R NV-SRAM. This CBRAM stack is promising for this application because of several reasons: i) no forming step is required. ii) CMOS compatible voltages are required to program the device. iii) an R_{off}/R_{on} ratio of 100 can be obtained thus making the restore operation highly reliable.

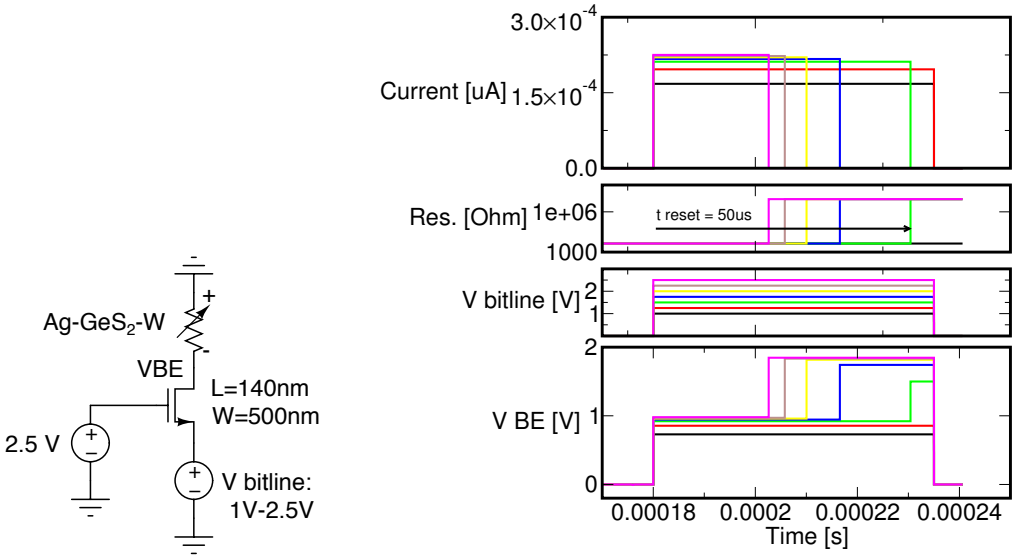


Figure 4.7: Transient parametric analysis of 1T-1R structure that shows both the time required to reset the CBRAM and the final resistance value. Note the changing of VBE before and after the switching.

Accordingly to our previous analysis this resistance ratio is sufficient and no stack engineering of the CBRAM is required. Unfortunately, the higher t_{set} with respect to OxRRAM devices lead to a power consumption during the store and restore operations that can be higher with respect to the consumption during a classical stand-by operation as we discussed in Sec. 1.5.

The entire CADENCE workflow from schematic to layout verification (Layout Versus Schematic) and Design Rule Checking was performed in the framework of the Design Kit provided by ALTIS Semiconductor. The Design Kit is based on 130 nm technological node and include CBRAM behavioral compact model for electrical simulation based on our experimental results. The 6T-SRAM core was reoptimized because of the different technological node. The ratio W/L of pull-up (M2 and M4), pull-down (M1 and M3), transfer transistor (M5 and M6) and control transistors (CM1 and CM2) of the NV-SRAM are set to: 480 nm/120 nm, 1250 nm/120 nm, 480 nm/120 nm, and 800 nm/120 nm, respectively. The transistors are operated at a voltage of 1.5 V. The WSNM and the RSNM are 0.518 V and 0.318 V respectively. As described in Sec. 4.1 the NV-SRAM cell operation follows the sequence: normal (read/write), reset (switching from LRS to HRS), store (switching from HRS to LRS), power-down, power-up and

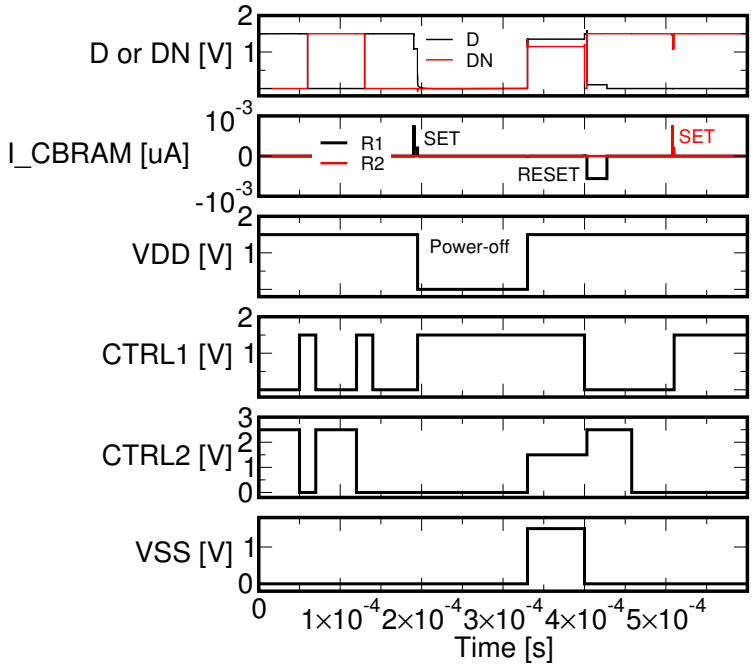


Figure 4.8: Transient analysis of the CBRAM based NV-SRAM to demonstrate the main operations. The Store operation starts at 130 µs. After the set operation a current of 100 µA flows in the CBRAM. The restore operation starts at 400 µs. The reset operation of the CBRAM requires a logical 0 in the latch to establish the voltage drop between the bottom electrode (CTRL2) and the top electrode and ends at 430 µs. A new set operation is applied at 510 µs.

restore. Fig. 4.8 shows the transient simulation of the designed NV-SRAM. Fig. 4.9 (left) shows the final layout of the fabricated structure. The two CBRAMs consist of the orange square connected to the CTRL2 line and the control pmos. Fig. 4.9 (right) is a magnified picture of the fabricated designs. The final scribe includes 24 pads for electrical testing.

4.4 1T-2CBRAM as nonvolatile memory element in FPGA

In this section we discuss a possible integration of CBRAM devices in a voltage divider configuration to control a pass gate or to store the data in a Look-Up Table (LUT) (Fig. 4.10). For this design the stack engineering and hence the maximization of R_{off} is extremely important to reduce the power consumption. In fact, although the 1T-2R NVE solution eliminates stand-by power consumption, the leakage current through the

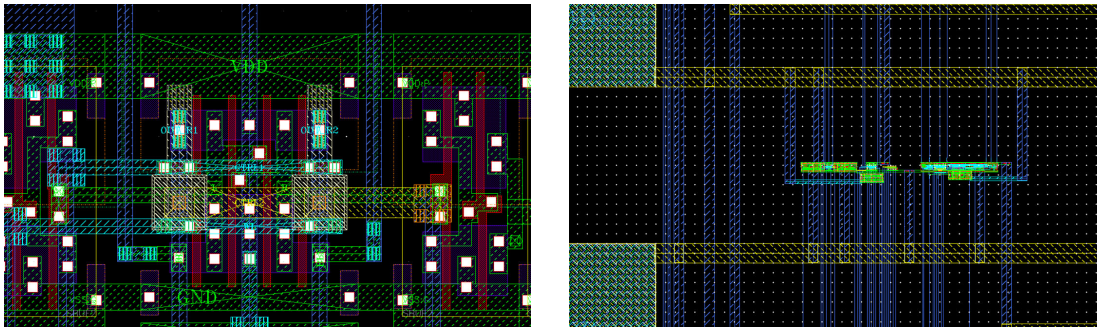


Figure 4.9: (Left) Layout of the designed and fabricated 8T2R NV-SRAM. Accordingly to the design rules the instantiation of the CBRAM takes an area of 0.25 µm in the back-end (square orange connected to the CTRL2 line). (Right) Layout of the designed and fabricated structures with the integrated CBRAM. Two pads for electrical measurements appear on the left corners.

ReRAM during run time (i.e. in continuous read operation) depends on the resistance of the high resistive state. Maximizing the high resistive state is essential to reduce the static power consumption during FPGA run time. In Sec. 3.4 we discussed set/reset conditions and their effect on the resistance levels (Fig. 3.7) claiming that HfO_2(2 nm)+GeS_2 is a promising solution to obtain a very high R_{off}. To demonstrate the concept, we report Eldo transient simulation of the programming scheme to achieve complementary resistance levels in CBRAMs (Fig. 4.11). We used the compact model calibrated on our GeS_2 memory devices [82], since the programming scheme is independent of the chosen stack. We simulate the switching from HRS to LRS and vice-versa in the CBRAM-2 (Fig. 4.10), where the top electrode is directly connected to the polarization line V_{anode} (as in the experimental case). Initially both the CBRAMs are in HRS (Fig. 4.10 (a)). When V_{anode} and V_{gate} are raised (t_{pw}=100 µs) the CBRAM cell switches in the LRS in 50 µs. In this case the 1T-2R NVE is configured to a logic zero on V_{net} during FPGA run time. A logical one can be obtained by resetting CBRAM-2 and then applying set pulses on CBRAM-1. After having proposed a suitable scheme to program the two CBRAMs in two different states, we have to clarify which advantages can be obtained by using the HfO_2(2 nm)+GeS_2 stack with respect to the reference stack.

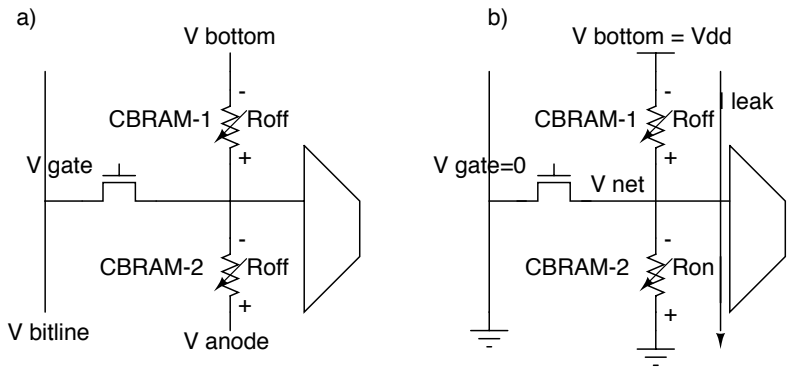


Figure 4.10: a) 1T-2R NVE in the initial state, when both CBRAM-1 and CBRAM-2 are in HRS (or in the pristine state). b) 1T- 2R NVE programmed to keep a logical zero on V_{net} in continuous read operation, after switching the CBRAM-2 from HRS to LRS.

4.4.1 Pulsed-tests and read disturb analysis

For reconfigurable logic architecture a cycling endurance of about 10^3 is required [99]. Hence, a cycling test was performed on GeS_2 and HfO_2(2 nm)+GeS_2 1T-1R devices with programming conditions that allow to maximize the $R_{\text{off}}/R_{\text{on}}$ ratio (Tab. 3.1). Experimentally obtained resistance pairs (R_{off} and R_{on}) were used to calculate the voltage between the two CBRAMs connected in series (V_{net} in Fig. 4.10) to estimate the deviations from a logical zero. The worst case corresponds to the voltage calculated with the lowest resistance ratio in the cycling test. In GeS_2 and HfO_2(2 nm)+GeS_2 the maximum voltage on V_{net} are 60 mV and 18 mV respectively (Fig. 4.12). It is worth to note that in continuous read operation at V_{dd}, the voltage on V_{net} may lead to an unwanted switching from LRS to HRS, thus potentially erasing the logical content in the 1T-2R NVE. Unwanted switching may occur because a non-zero DC voltage is always present on the bottom electrode of the CBRAM-2. To further investigate this point, we performed DC read disturb tests on 8×8 array. The cells were programmed in the LRS state. The switching time (t_{switch}) was defined as the time required to increase the resistance by a factor of 10. The mean t_{switch} for a given applied stress was extracted at 50% of the t_{switch} distribution obtained on the 64 cells of the memory array. In Fig. 4.13 (left) we extrapolated 10 years read disturb immunity at 40 mV and 6 mV for HfO_2(2 nm)+GeS_2 and GeS_2 respectively. The projected 10 years read disturb in HfO_2(2 nm)+GeS_2 is higher than the calculated worst case voltage. Thus, with the optimized HfO_2 barrier, not only the devices can be programmed more than

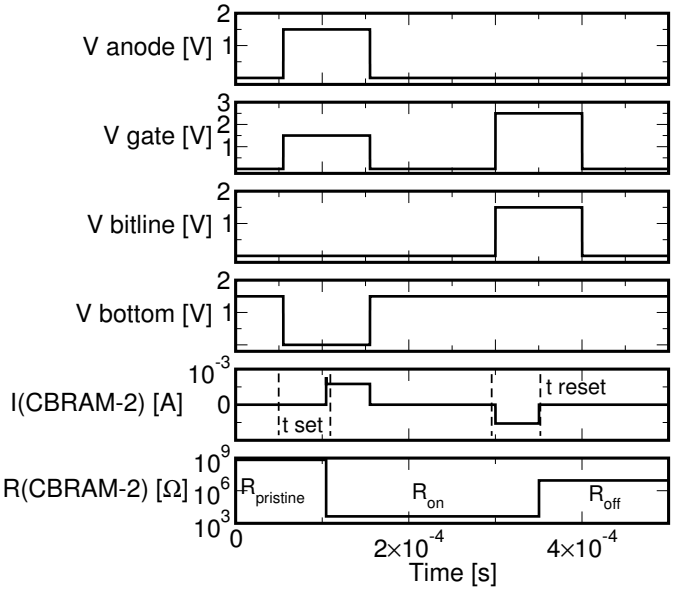


Figure 4.11: Transient simulation of the GeS_2 based CBRAM complementary programming scheme in 1T-2R NVE. The time required to switch the CBRAM-2 (Fig. 4.10) from HRS to LRS is 50 µs with V_{anode}=1.5 V. In the LRS, R_{on} is 3 kΩ. Reset operation requires 50 µs with V_{bitline}=1.5 V. In the HRS, R_{off} is 2×10^6 Ω.

1k cycles keeping a very high resistance ratio, but also immunity to DC read disturb is demonstrated. On the contrary, in standard GeS_2 memory device a voltage of 60 mV can be sustained for only one day, before switching the CBRAM to HRS.

4.4.2 Comparison and leakage currents estimations

Fig. 4.13 (right) provides a comparison of R_{off} and R_{on} mean values obtained during pulsed cycling tests for our 2 nm HfO_2 devices and other state of the art ReRAM technologies, as reported in: [21], [22], [100, 101, 102, 103, 104, 105, 106, 107, 108], [7], [30]. These data refer to sub-µm devices cycled at least 1k times in pulse mode. The proposed solution offers the best $R_{\text{off}}/R_{\text{on}}$ ratio presented so far to our knowledge. It is worth to note that the OxRRAM based on N-doped AlO_x presented in Fig. 4.13 (down) was used to demonstrate the first nonvolatile 3D-FPGA reported in [49]. Tab. 4.1 reports leakage currents for 6T-SRAM and 1T-2R NVE. In 6T-SRAM leakage current strongly depend on V_{th}, the oxide thickness, and the feature size and vary in the range of pAs [57], [58]. This implies that an R_{off} value higher than 10^{12} Ω at a read voltage of

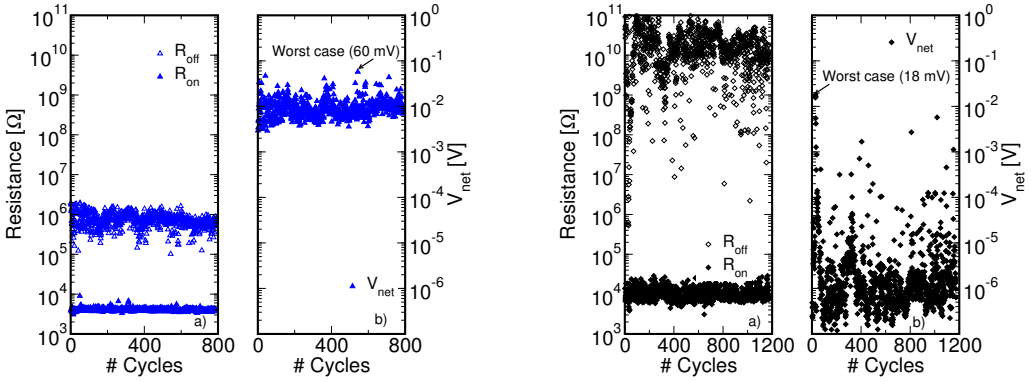


Figure 4.12: (a) Pulse cycling test for GeS_2 (left) and HfO_2(2 nm)+GeS_2 (right) based CBRAM devices. (b) Voltage between two GeS_2 (left) and HfO_2(2 nm)+GeS_2 (right) based CBRAM devices in a 1T-2R NVE (V_{net} in Fig. 4.10) calculated using every resistance pair of the cycling test in (a).

Table 4.1: Comparison of 6T-SRAM and proposed 1T-2R NVE for reconfigurable logic applications.

Ref.	Architecture	Technology	$I_{\text{leak. at Vdd}}$ [pA]	V_{dd} [V]	Volatility	Standby Cons.
[57]	6T-SRAM	-	25	1.2	yes	yes
[58]	6T-SRAM	-	12-50	1.2	yes	yes
[109]	1T-2PCM	GST	5×10^7	1	no	no
[47]	1T-2CBRAM	W−GeS_2−Ag	10^7	1	no	no
[55]	1T-2CBRAM	Pt−ZnCdS−Ag	10^6	1	no	no
[This PhD]	1T-2CBRAM	W−HfO_2−GeS_2−Ag	1300	1	no	no

1 V should be targeted to further improve this specification. In this work, considering the HfO_2(2 nm)+GeS_2 stack, we report a mean leakage current of 1.3 nA at 1 V reverse read operation (Fig. 3.7), the lowest compared to other 1T-2R NVE solutions reported so far [55], [47], [109]. In conclusion, we demonstrated that HfO_2(2 nm)+GeS_2 based CBRAM is a promising candidate to implement the 1T-2R NVE architecture. As we discussed in Sec. 1.7.1 we can obtain a density enhancement because of the integration in the third dimension, hence reduction of dynamic power consumption is expected.

4.5 Stochastic synapses for neuromorphic applications

In this section we discuss how weak programming conditions could help in implement stochastic hardware neural networks using the results presented in Sec. 2.2. An intro-

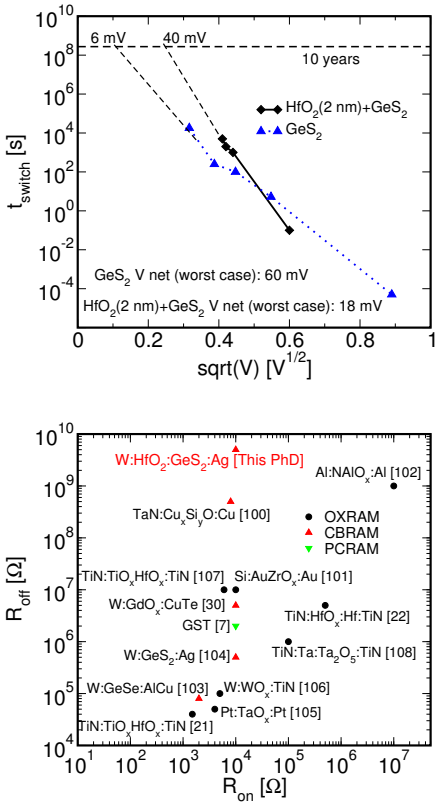


Figure 4.13: (Up) Switching time (t_{switch}) from LRS to HRS during low negative stress bias for GeS_2 and $HfO_2(2\,\text{nm})-GeS_2$ samples. Each point corresponds to the mean of 64 cells. Projected 10 years disturb immunity of 6 mV and 40 mV were extracted for GeS_2 and $HfO_2(2\,\text{nm})+GeS_2$ respectively. (Down) Benchmark of LRS (R_{on}) and HRS (R_{off}) for several ReRAMs reported in the literature obtained with cycling test (more than 1k cycles). Mean values are reported.

duction on neural network was provided in Sec. 1.10. In [110] we propose to implement synapses using CBRAM devices. The synaptic weight is represented by the resistance values of the CBRAM cells. In a deterministic hardware neural network there are some learning rules that define the changing in the synaptic conductance between two neurons and the synaptic programming is physically achieved using voltage pulses. Usually the synaptic programming is not supposed to fail i.e. the CBRAM cell that implements the synapse is going to be (strong) programmed by a signal that leaves the input neuron. Finally, in this way, a part of the CBRAM cell will be in the LRS and the others in HRS (accordingly to the rule ex. LTP or LTD). On the other hand, if the signal that leaves

the neuron is based on weak programming conditions we will obtain a probabilistic switching of the device. In particular for a population of synapses supposed to switch to LRS, only a percentage will effectively switch (Sec. 2.2). In few words: even if the Long Term Potentiation rule was satisfied the CBRAM will not switch to the LRS because of the distribution of t_{set} (i.e. the time required to set the memory). Another solution to introduce the stochasticity is the use of a Pseudo Random Number Generator (PRNG). The PRNG output allows or blocks the input neuron signals according to the defined probability levels randomly generated. In few words: even if the Long Term Potentiation rule was satisfied the CBRAM will not switch to the LRS because of the switching probability enforced by the PRNG that tune the signal entering in the CBRAM [110]. Exploiting the intrinsic CBRAM switching probability avoids the presence of the PRNG circuits, thus saving important silicon footprint. It also reduces the programming power, as the programming pulses are weaker compared with the ones used for deterministic switching. It might, however, be difficult to precisely control the switching probability of individual synapse using weak conditions in a large-scale system. When weak programming conditions are used, both device to device and cycle to cycle variations contribute to probabilistic switching, hence can be difficult to chose the right weak condition that is effective to be weak also after many cycles. Decoupling the effect of the types of variations is not straightforward in CBRAM devices thus also the choice of the programming conditions can be difficult. In Sec. 4.6 we will provide another solution for implementing stochasticity in hardware neural networks.

4.6 Stochastic neurons

In Sec. 4.5 we provide two solutions to implement a stochastic hardware neural network. Even if the conditions for the application of a learning rule are satisfied, the probabilistic switching of the device, determined by extrinsic or intrinsic ways, will impact the final state of the CBRAM. In the extrinsic way a PRNG was used that tune the (strong) signal entering into the synapse, in the intrinsic way a weak signal enters in the synapse and the synapse itself will decide if it will switch to another state or not. We claimed that weak conditions can switch a percentage of devices lower than 100%. In this section, we will discuss another possibility that does not require neither PRNG nor weak conditions. We can foresee an area gain, because PRNG can be avoided and the implementation of the

stochastic neuron is more robust with respect to the application of weak conditions that suffer of cycle to cycle instability. In the literature, different techniques to implement controlled stochasticity in hardware neural networks have been proposed. It is possible to exploit the thermal noise in the CMOS but this may lead to silicon overheads and unwanted correlations [66]. Other techniques exploit CMOS circuits with using noise but have significant area overhead [111], or the noise of photons with photodetectors [112] or even special kinds of 'noisy transistors' [113]. Finally it was proposed to use fundamentally probabilistic nanodevices like single electron transistors [114], but which might suffer from poor CMOS compatibility and room temperature operation. As was reported in Sec. 2.3 by cycling many times CBRAM devices a statistical distribution of the high resistive state (R_{Off}) is usually obtained regardless of the programming conditions used. Dispersion in R_{Off} was interpreted in terms of stochastic breaking of the filament during the reset process, due, for example, to the unavoidable defects close to the filament which act as preferential sites for dissolution. In Sec. 2.3 we showed, with the help of modeling, that a distribution in R_{Off} leads to a spread in others physical quantities like, for example, the left-over filament height (h) and the t_{set}. To validate the operation of a stochastic neuron we characterized the kinetic of the set operation by pulse measurements as was described in Sec. 2.1.3. Fig. 4.14 (left b inset) shows an example of the oscilloscope trace for the evolution of voltage drop across the cell (V_c) during a set pulse. Starting from some of the measured values of R_{Off} (Fig. 4.14 left a) we collected the spread in t_{set} when the applied pulses were V_a=3 V and t_{pulse}=5 µs (Fig. 4.14 left b). The dotted line in Fig. 4.14 (left b), shows the simulated values of t_{set}. To obtain the simulated curve of t_{set}, first the distribution of h was calculated and then the t_{set} was calculated using Eq. 2.1 and Eq. 2.3. It is worth to note that in designing a stochastic neuron the key aspect is the dispersion in t_{set} and not the relation (if any) between R_{Off} and the t_{set}. In fact we simply propose a methodology to obtain a t_{set} distribution starting from the measured R_{Off}. This methodology can be integrated in the compact model to exploit variability when designing and simulating new architectures. Nevertheless, as a proof of concept assuming a distribution of t_{set} is enough to introduce the stochasticity in the neuron circuit and even the CBRAM compact model can be skipped for a proof-of-concept simulation.

The complexity of a neuron circuit depends on the overall functionality of the neural network and of the chosen biological models. For our purpose of concept validation,

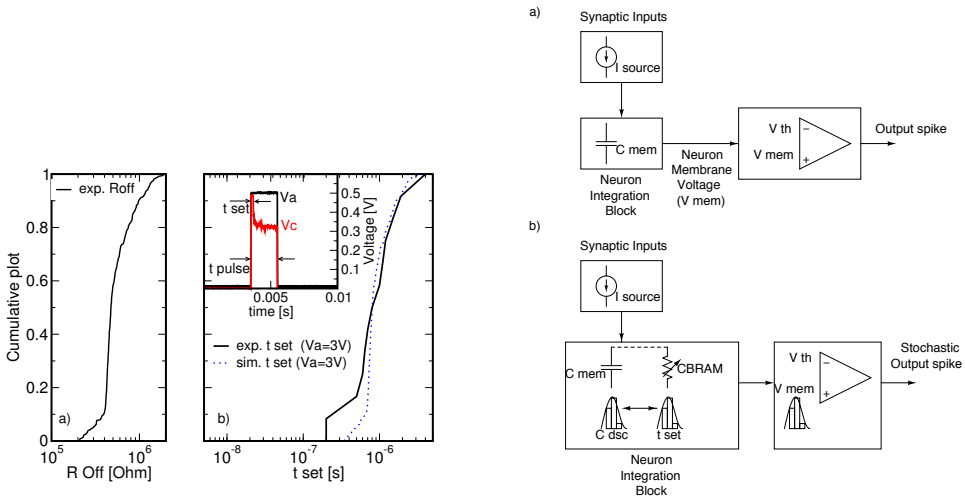


Figure 4.14: (Left) (a) R_{Off} distribution (cycle to cycle) obtained in Ag-GeS_2-W based 1R CBRAM devices.(b) Experimental (line) and simulated (dotted) t_{set} distribution obtained cycling the CBRAM cell with a pulse amplitude V_a=3 V. (b in the inset) Example of a typical oscilloscope trace tracking the voltage on the CBRAM (V_c) and the applied pulse (V_a). Between every set operation a reset operation was performed (not shown). (Right) (a) Schematic image shown the basic concept of a Integrate and Fire neuron [115]. (b) Schematic showing the basic concept of our proposed Stochastic Integrate-Fire neuron (S-IF).

we chose one of the simplest, the Integrate and Fire neuron model. Fig. 4.14 (right a) shows the concept of a simple Integrate and Fire neuron model. It constantly sums (integrates) the incoming synaptic-inputs or currents (excitatory and inhibitory) inside the neuron integration block using a capacitor. More advanced designs also work with this principle [115]. This integration leads to an increase in the membrane potential of the neuron V_{mem}. When the membrane potential reaches a certain threshold value V_{th}, the neuron generates an output spike (electrical signal). After the neuron has fired the membrane potential goes back to a resting value (initial state), through discharging of the capacitor C_{mem}. Usually, the output firing activity of a Integrate and Fire neuron is deterministic because the neuron fires every time the membrane potential reaches a defined threshold value.

To introduce non-deterministic or stochastic behavior in Integrate and Fire neuron, we propose to connect a CBRAM device to the capacitor C_{mem}, such that C_{mem} could only discharge through the CBRAM device by switching it to the low-resistive state

(Fig. 4.14 right b). The anode of the CBRAM and the V_{mem} net of the capacitor should be connected. The duration for which current can flow through the low-resistive CBRAM device can be controlled using a transistor. In such a configuration, the spread on the t_{set} of the CBRAM would translate to a spread on the discharge-time (t_{dsc}) of the capacitor. For consecutive neuron spikes, this would lead to different initial state of C_{mem}, thus making the firing of the neuron stochastic. Fig. 4.15 (left) illustrates conceptually the impact of four different values of t_{set} (keeping constant pre-synaptic weights), on the inter-spike interval. In case (a), t_{set} is very long thus the capacitor has a very weak discharge. As a consequence just few additional incoming pre-neuron spikes are required to charge back the V_{mem} to the level of V_{th}, thus leading to an output pattern with the shortest inter-spike interval. In case (b), t_{set} was the shortest, and hence the capacitor discharged the most. Thus for this case, more incoming pre-neuron spikes are needed to recharge V_{mem}. Case (c) represents a deterministic Integrate and Fire situation with full V_{mem} discharge. Finally, case (d) depicts a situation with different t_{set} durations for consecutive output spikes. It is a possible representation of neuron inter-spike intervals for a random sequence of t_{set} values that can be obtained by cycling the CBRAM device multiple times (note the experimental dispersion of t_{set} in Fig 2.20). The circuit equivalent of the Stochastic-Integrate and Fire neuron concept shown in Fig. 4.14 (right) is presented in Fig. 4.15 (right). It consists of a current-source to simulate input currents coming from synapses and pre-neurons, a capacitor C_{mem} to integrate the current and build up the neuron membrane-voltage V_{mem}, a nMOS transistor M1 to perform set operation, two nMOS transistors M2 and M3 to perform the reset operation, a comparator block, a spike-generation block, a delay-element Δt and a CBRAM device. The delay element is used to perform the reset operation of the CBRAM device at the end of each neuron spike.

In Fig. 4.15, initially the CBRAM is in high-resistive state. As incoming pre-synaptic current is accumulated in C_{mem}, V_{mem} would constantly build up at the anode of the CBRAM. During this time M1, M2 and M3 are off. When the neuron spikes, the spike-generation block will generate an output-spike and two additional pulsed-signals (S1, S2) going to M1 and Δt respectively. S1 acts as a gating signal to turn on M1. V_{mem} build-up and switching on of M1 will enable set-operation of the CBRAM since a positive voltage drop is established between the anode and the cathode. However during the set-operation, M2 and M3 are not turned on, as Δt delays the signal S2.

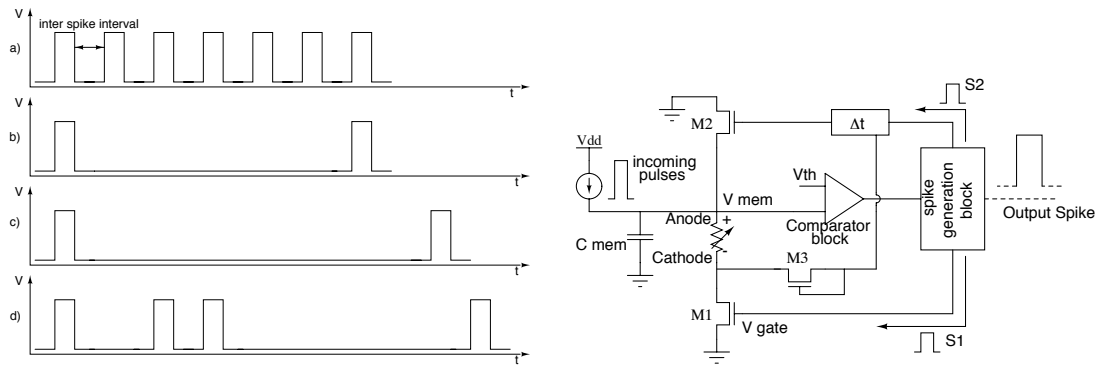


Figure 4.15: (Left) (a)-(d) Schematic of output neuron firing patterns for different example test cases. Proposed circuit-equivalent of the S-IF neuron.

At the end of the set-operation, the signal S2 will turn on M2 and M3 thus building up the voltage at the cathode to switch the CBRAM to the off-state (reset). Thus, before the next consecutive neuron spikes the CBRAM device is automatically reset and reprogrammed to a different initial R_{Off} state. Note that the flow of current through the CBRAM, during the set-operation, leads to a discharge of the capacitor C_{mem} thus decreasing the membrane voltage V_{mem}. The amount of decrease in V_{mem} can be estimated by calculating the total duration (t_{dsc}) for which current flows through the switched CBRAM. t_{dsc} is the difference of the pulse-width of the signal S1 and the t_{set} (inset of Fig. 4.14 left b). Depending on the value of t_{set} every time the neuron spikes, different amount of C_{mem} discharge will occur. Thus, in between any two firing cycles, the neuron may require different amount of incoming current to charge V_{mem} to the level of V_{th}.

We performed Eldo transient simulation to validate the proposed concept using a simplified circuit shown in Fig. 4.16 (left). Transistors and capacitors sizing were not optimized with respect to a real implementation, but to give a simple proof-of-concept. Fig. 4.16 (right a) shows a simulated train of incoming pulses (excitatory currents) and the corresponding evolution of the V_{mem} (Fig. 4.16 right b) between two consecutive neuron spike-cycles. When V_{mem} reaches a threshold voltage V_{th} ($V_{th} \simeq 3.5\,V$ in our simulation), the CBRAM device undergoes set-operation, and C_{mem} begins to discharge. Fig. 4.16(right b) shows the discharging and re-charging of C_{mem} for four different simulated values of t_{set} (in the range 300 ns - 600 ns). Fig. 4.16 (right c), shows the expected output of the neuron. Note that different number of incoming pulses are

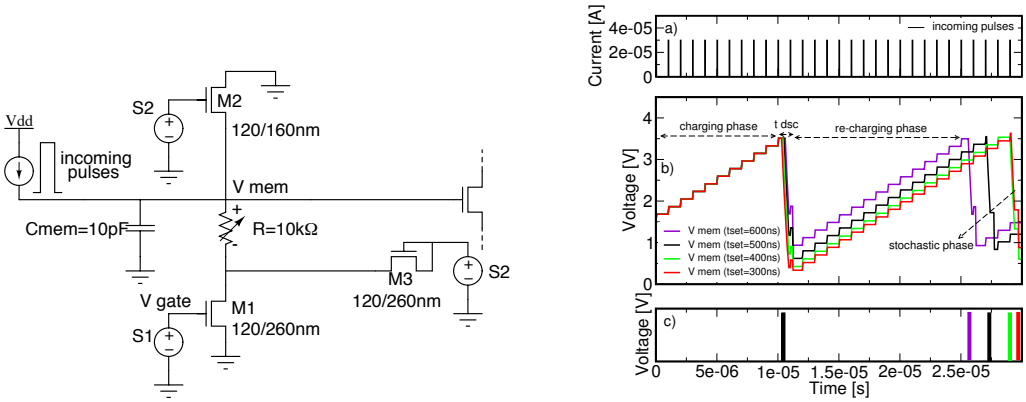


Figure 4.16: (Left) Circuit used to demonstrate the concept of a S-IF effect when the CBRAM is in the set state. (Right) Full evolution of V_{mem} simulating the circuit shown on the left. (a) Pre-neuron incoming pulses are used to build up V_{mem}. (b) Initially V_{mem} builds up as consequence of incoming currents (charging phase). Set operation lead to different discharge of C_{mem} (t_{dsc}). During the recharging phase a different number of incoming pulses will raise V_{mem} till V_{th}. (c) Expected different inter-spike intervals depending on the t_{set}.

required to reach the neuron firing threshold again, since the initial V_{mem} value is dominated by the stochasticity in t_{set}. Five additional incoming pulses are needed to reach the threshold for the shortest value of t_{set} (300 ns). Fig. 4.17 (right) shows the zoomed version of C_{mem} discharging for the different simulations shown in Fig. 4.16. Note that the longest t_{set} (600 ns) corresponds to the least amount of C_{mem} discharge, and vice-versa. To simulate the reset operation, a pulse of 45 ns with an amplitude of 3 V was applied at M2 and M3, while keeping M1 off. Such high voltage on M3 is required to build up a voltage on $V_{cathode}$. Fig. 4.17 (right) shows the time evolution of $V_{cathode}$ and V_{mem} when the initial value of V_{mem} was generated by a t_{set} of 300 ns for two different width of M3. The actual voltage drop on the CBRAM can be increased increasing the size of the nMOS as shown in Fig. 4.17 (right a). Moreover, during the reset, an additional discharge of V_{mem} is possible depending on the size of M3, since M2, that is directly connected to V_{mem}, is turned on by S2.

Due to the intrinsic physics of CBRAM device, some constraints in implementing the proposed circuit should be considered. In particular, V_{th} has to be greater than the minimum value of the voltage-drop required to set the CBRAM device for a given pulse-width. The amplitude of S1 should be sufficient to turn on the gate of M1, while

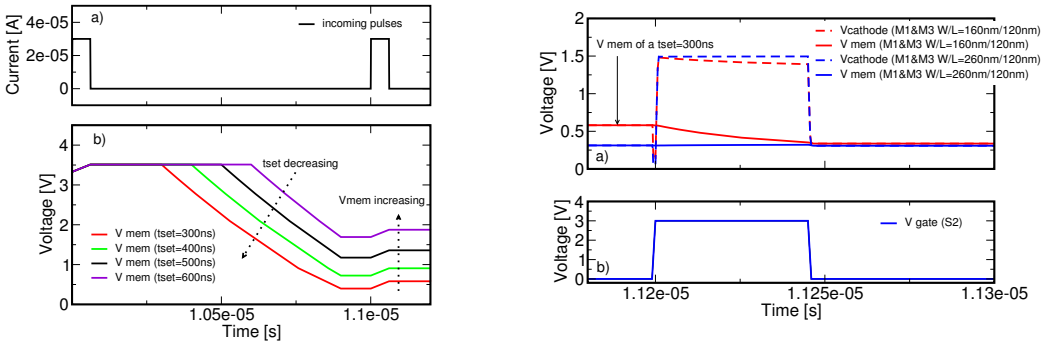


Figure 4.17: (Left) (a) Pre-neuron incoming pulses are used to build up V_{mem}. (b) Zoom on V_{mem} during the discharging phase for different t_{set} in the range 300 ns-600 ns. Lower t_{set} leads to lower residual membrane voltage V_{mem}. (Right) (a) Time-evolution of V_{mem} and $V_{cathode}$ that establish a voltage drop on the CBRAM to enable reset operation. Larger M3 increase the voltage drop, since $V_{cathode}$ builds up more. V_{mem} corresponding to a t_{set} of 300 ns is considered. (b) Pulse applied to M3.

the pulse-width of S1 depends on the V_{th} and the spread on t_{set}. If S1 pulse-width is very long it would always lead to a complete discharge of C_{mem} and the t_{set} stochasticity cannot be exploited. However S1 cannot be arbitrarily small, it has to be greater than the minimum t_{set} value at a given voltage applied on the anode of the CBRAM device. In [104], we have shown the dependence of applied pulse-width and the amplitude of V_a for the CBRAM set-operation. Thus, by tuning the characteristics of S1, the stochastic response of the neuron can be controlled. The amplitude of S1 would determine the amount of current flowing through M1 (compliance current) and thus the final value of the CBRAM resistance in the set state. The set state resistance would determine the programming conditions for the consecutive reset-operation. Thus, the characteristics of S2 can be tuned based on the final CBRAM resistance obtained after the set-operation. For the proposed S-IF, additional energy consumption per spiking cycling of the neuron will be devoted to perform set and reset operation. The extra-energy consumption is dependent on the ratio R_{Off}/R_{On}; in particular on R_{On} since hundreds of µA can flow before M1 would be turned off, if the low resistance state is $\simeq 10^4\,\Omega$, thus raising the power consumption. We estimated the energy consumption during the set operation using: $E_{set}=V_{set}\ I_{set}\ t_{set}$. In our simulations we used V_{set}=3.5 V (i.e. V_{th}), I_{set}=350 µA, t_{set} in a range between 300 ns and 600 ns that gives a E_{set} energy mean value of 55 n J.

The energy devoted to reset the CBRAM is negligible. For a real system, E_{set} can be strongly reduced increasing the resistance of the low resistive value thus reducing I_{set}, since for the proposed application the ratio R_{Off}/R_{On} is not a major constraint.

4.7 Conclusions

In this chapter we discussed three hybrid architectures. In Sec. 4.1 a NV-SRAM based on OxRRAM memories was presented. In this case, the constraints on the logic (size of the pmos in the latch) to achieve a reliable store operation and the minimum resistance ratio to achieve a reliable restore operation were identified by parametric transient simulations, worst case analysis and Monte-Carlo simulations. In Sec. 4.4 a 1T-2R NVE was analyzed. The architecture requires very high R_{Off} to be competitive with state of the art FPGA based on volatile 6T-SRAM, in which leakage current is minimized (Sec. 1.7.1). Using our experimental results we suggested that HfO_2(2 nm)+GeS_2 is, at our knowledge, the most promising technology to renovate some parts of FPGA embodiments. Finally, we exploited the variability in t_{set} to design a stochastic neuron that can be integrated in stochastic harware neural networks (Sec. 4.6). The proposed circuit offers several advantages with respect to the other ways to implement stochasticity such as PRNG or weak programming conditions discussed previously. In fact, it is extremely compact (1R-3T), and the energy consumption is dominated by the CBRAM set process, hence can be minimized providing an high R_{On}.

5

Conclusions

This manuscript addresses CBRAM technology aspects in view of developing hybrid architectures. CBRAM stacks were electrically characterized to understand the programming conditions and to reveal the main switching parameters, the memory ratio and the reset current. A thermally activated hopping model was used to describe the ion migration and the consequent filament growth/dissolution during set/reset processes. The model parameters extracted by both DC quasi-static and pulse measurements were used to implement a compact model for electrical simulations of the main operations of the circuits. Hence, using compact modeling and circuit level simulations, we show that CBRAM devices can be integrated in hybrid architectures such as nonvolatile SRAM, routing switches that consists of 1T-2CBRAM architectures and even in neurons of hardware neural network. We showed performance and reliability improvement of Ag-GeS_2 based CBRAM devices by addition of a 2 nm thick HfO_2 layer between the electrolyte and the W bottom electrode and the foreseen advantages in using this stack in 1T-2CBRAMs hybrid routing switches were elucidated. We also exploited the intrinsic variability of CBRAM devices in designing fault tolerant ultra-scalable architectures for neuromorphic computing, highlighting the benefits of novel non memory technologies, whose impact may go beyond traditional memory markets.

In Chap. 1 we introduced the state of the art in memory research to introduce non volatility at several levels of the memory hierarchy with the integration of ultra-scalable (sub 20 nm), low power, byte addressable and supporting in-place writing emerging memories. We focused on nonvolatile SRAM that integrates 6T-SRAM cells and NVM devices forming a direct bit-to-bit connection in a 3D or vertical arrangement to achieve

fast parallel data transfer and fast power on/off speed. By stacking emerging memories with CMOS devices, new routing switches can be achieved to reduce FPGA area, remove external Flash, thus reducing board area and obtain instantly-on capability. As a consequence we expect to reduce the power consumption in SRAM based FPGA related to the in-rush and SRAM reconfiguration after power up, and dynamic power consumption due to the area reduction. Although hybrid routing solutions eliminate stand-by power consumption, the leakage current through the ReRAM during run time depends on the resistance of the high resistive state. We have concluded the chapter with a remark on the high resistive state that should be maximized to reduce the static power consumption during FPGA run time.

Chap. 2 introduces the equations of the empirical model used to explain and predict the main switching parameters in CBRAM devices measured by electrical characterization both in quasi-static and pulse configuration. We performed a statistical analysis on the correlation between the programming conditions and the percentage of CBRAM devices in a 8×8 NOR memory array that reversibly switch from high resistive state (HRS) to low resistive state (LRS). Some of the possible causes of the cell to cell and cycle to cycle variability were discussed. We exploited the cell to cell variability using CBRAM as stochastic synapse in neuromorphic applications.

In Chap. 3 four different CBRAM stacks: i) W-GeS_2-Ag, ii) Ta-TaO_x-GeS_2-Ag, iii) W-SiO_x-GeS_2-Ag, iv) W-HfO_2-GeS_2-Ag were investigated in view of developing hybrid architectures. This was mainly done to improve the 1T-2R NVE architecture developed for FPGA embodiments. Among the characterized CBRAMs technologies, our dual-layer electrolyte stack (2 nm HfO_2−30 nm GeS_2) leads to a resistance ratio ($R_{\mathrm{off}}/R_{\mathrm{on}}$) higher than 10^6, reset current below 100 µA enforcing a compliance current (I_{comp}) of 240 µA. Moreover, no forming step is required. We have also explained the improved memory resistance ratio by means of physical modeling.

In Chap. 4 three hybrid architectures were described and the circuits operation was demonstrated by electrical simulations. Critical points (both related to logic and memory characteristics) to obtain reliable operations in NVSRAM were elucidated through worst case and Monte-Carlo analysis. Using our optimized 2 nm HfO_2−30 nm GeS_2 stack, we defined foreseen advantages in implementing a hybrid 1T-2R routing switch for an FPGA with respect to SRAM based solutions. Finally, we exploited the variability in switching parameters of the CBRAM to design a stochastic neuron

that can be useful in hardware neural networks. The proposed circuit offers several advantages with respect to the other ways to implement stochasticity such as PRNG or weak programming conditions of CBRAM devices used as synapse.

5.1 Perspectives

Monolithically stacked 3-D FPGA [49] and Nonvolatile SRAM [33] have demonstrated the integration of logic and emerging memories. Behind the performance of the fabricated circuits a huge research at the memory device level was mandatory and it is expected to play a fundamental role also in the next years. Process and material optimization, effects of scaling, reduction of variability and physical understanding are key aspects that need to be addressed more. During this thesis we were able to fabricate and measure CBRAM stacks to optimize specific applications or to use non optimized CBRAM stacks to build fault tolerant designs and, even more, exploiting the variability to enhance artificial neural networks performances. We designed and taped-out both nonvolatile 6T-SRAM and fundamental blocks of an FPGA to collect new experimental data that could increase the knowledge of the memory devices and the cross-effects between memories and logic. Electrical measurements on these prototypes should be the focus of the research in hybrid architectures in our group in the next months, providing new feedbacks for the optimization of the memory stacks.

Appendix A

List of Patents and Publications

A.1 Patents

1. Manan Suri, **Giorgio Palma**, *Neurone artificiel comprenant une mémoire résistive*, DD. 14473.

2. Elisa Vianello, Olivier Thomas, Gabriel Molas, **Giorgio Palma**, *Conception d'une mémoire résistive de type CBRAM pour une application logique reconfigurable*, DD. 14750.

A.2 Conference and Journal Papers

1. **G. Palma**, E. Vianello, O. Thomas, M. Suri, S. Onkaraiah, A. Toffoli, C. Carabasse, M. Bernard, A. Roule, O. Pirrotta, G. Molas and B. De Salvo, *Interface engineering of Ag−GeS_2 based Conductive Bridge RAM for reconfigurable logic applications*, submitted to Transactions on Electron Devices.

2. **G. Palma**, E. Vianello, O. Thomas, H. Oucheikh, S. Onkaraiah, A. Toffoli, C. Carabasse, G. Molas and B. De Salvo, *A novel HfO_2−GeS_2−Ag based Conductive Bridge RAM for reconfigurable logic applications*, accepted for oral presentation at IEEE ESSDERC 2013.

3. **Giorgio Palma**, Manan Suri, Damien Querlioz, Elisa Vianello, Barbara De Salvo, *Stochastic neuron design using Conductive Bridge RAM*, accepted for lecture presentation at IEEE/ACM NANOARCH 2013.

4. F. Longnos, E. Vianello, G. Molas, **G. Palma** , E. Souchier, C. Carabasse, M. Bernard, B. De Salvo, D. Bretegnier, J. Liebault, *On disturb immunity and P/E kinetics of Sb-Doped GeS_2/Ag Conductive Bridge memories*, to be published in proceedings of IEEE IMW 2013.

5. M. Suri, D. Querlioz, O. Bichler, **G. Palma** , E. Vianello, D. Vuillaume, C. Gamrat, B. De Salvo, Bio inspired stochastic computing using binary CBRAM synapses. IEEE Transactions on Electron Devices, vol. 60, no. 7, pp. 2402-2409, 2013.

6. **Giorgio Palma**, Elisa Vianello, Gabriel Molas, Carlo Cagli, Florian Longnos, Jérémy Guy, Marina Reyboz, Catherine Carabasse, Mathieu Bernard, Faiz Dahmani, Damien Bretegnier, Jacques Liebault, and Barbara De Salvo *Effect of the Active Layer Thickness and Temperature on the Switching Kinetics of GeS_2-Based Conductive Bridge Memories*, Japanese Journal Of Applied Physics, vol. 52 pp. 04CD02 (2013).

7. M. Suri, O. Bichler, D. Querlioz, **G. Palma**, E. Vianello, D. Vuillaume, C. Gamrat, B. De Salvo *CBRAM devices as binary synapses for low-power stochastic neuromorphic systems. Auditory (Cochlea) and visual (retina) cognitive processing applications*, IEEE IEDM Tech .Dig., 2012, pp. 10.3.1-10.3.4.

8. E. Vianello, G. Molas, F. Longnos, P. Blaise, E. Souchier, C. Cagli, **G. Palma**, J. Guy, M. Reyboz, C. Carabasse, M. Bernard, F. Dahmani, D. Bretegnier, J. Liebault, *Sb-doped GeS_2 as performance and reliability booster in Conductive Bridge RAM*, IEEE IEDM Tech. Dig. 2012, pp. 31.5.1-31.5.4.

9. **Giorgio Palma**, Elisa Vianello, Gabriel Molas, Carlo Cagli, Florian Longnos, Jérémy Guy, Marina Reyboz, Catherine Carabasse, Mathieu Bernard, Faiz Dahmani, Damien Bretegnier, Jacques Liebault, and Barbara De Salvo *Effect of the Active Layer Thickness and Temperature on the Switching Kinetics of GeS_2-Based Conductive Bridge Memories*, International conference on Solid-State-Devices and Materials, SSDM 2013.

10. Hraziia, A. Makosiej, **G. Palma**, J.-M. Portal, M. Bocquet, O. Thomas, F. Clermidy, M. Reyboz, S. Onkaraiah, C. Muller, D. Deleruyelle, A. Vladimirescu,

A. Amara, C. Anghel, *Operation and stability analysis of bipolar OxRRAM based nonvolatile 8T2R NVSRAM as solution for information back-up*, in press on Solid-State Electronics (2013).

11. **G. Palma**, E. Vianello, C. Cagli, G. Molas, M. Reyboz, P. Blaise, B. De Salvo F. Longnos, F. Dahmani *Experimental investigation and empirical modeling of the set and reset kinetics of Ag-GeS_2 Conductive Bridging Memories*, IEEE IMW pp.1-4 (2012).

A. LIST OF PATENTS AND PUBLICATIONS

Appendix B

Résumé en Français

B.1 Introduction

L'intégration lors des dernières étapes de réalisation du circuit VLSI, la tension d'alimentation compatible CMOS, le temps de programmation rapide, ainsi que la miniaturisation sont les caractéristiques recherchées pour la conception de nouvelles architectures hybrides de mémoires résistives (ReRAM). Ces architectures pourront etre utilisées dans différentes applications telles que la faible puissance embarquée, la logique reconfigurable, ou encore les circuits neuromorphiques.

B.1.1 Mémoire ReRAM

Les matériaux actifs utilisés dans les mémoire ReRAM sont généralement des oxydes de métaux de transition incorporés entre deux électrodes métalliques. L'effet mémoire s'appuie sur la transition réversible entre un état fortement résistif (reset) et un état faiblement résistif (set) suite à l'application d' un champ électrique sur la structure. Ce mécanisme de commutation est l'effet de réactions thermochimiques et/ou électrochimiques; il peut être déclenché soit par l'amplitude du champ électrique appliqué (commutation unipolaire) soit par la polarité du champ appliqué (commutation bipolaire). En particulier, en mode bipolaire, le processus dit du set ne peut se produire qu'avec une seule polarité alors que la réinitialisation se produit uniquement avec l'inversion de polarité (Fig. B.1). Par ailleurs, le comportement de commutation ne dépend pas seulement des oxydes utilisés mais dépend également du choix des électrodes métalliques et de la physique au niveau des interfaces. Les Oxydes de métaux de

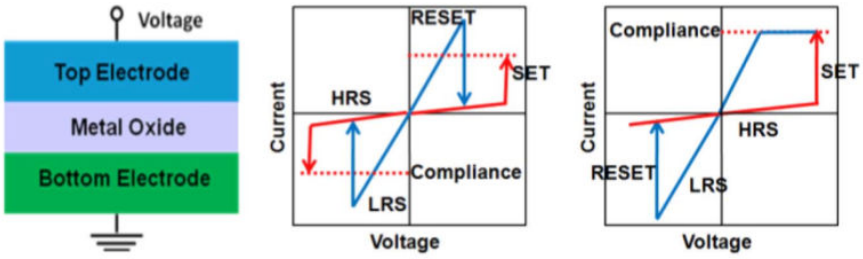


Figure B.1: Représentation schématique de la structure MIM pour l'oxyde métallique ReRAM et caractéristiques typiques I-V en régime DC présentant un comportement unipolaire et bipolaire [20].

transition les plus utilisés sont ceux qui exploitent le réseau sous-stoechiométrique en atomes d' oxygène, tels que TaO_x, HfO_x ou AlO_x. En particulier, il a été observé que l' état résistif faible peut être attribué à une phase pauvre en oxygène dans l'isolant qui mène à la conduction métallique. En outre, il a été démontré récemment que la concentration des lacunes d'oxygène dans l'isolant peut être contrôlée par une couche d'interface tels que le Ti ou le Zr, ce qui améliore sensiblement les performances de la cellule ReRAM. Une récente étude rapportée dans [21] sur une mémoire résistive utilisant l'empilement Ti/HfO_x. Hf/HfO_x a montré une vitesse de commutation 10 ns et une endurance de 10^{10}. De plus, elle a pu etre miniaturisé jusqu'à 10 nm [22]. Un autre type de ReRAM avec une densité de 32 Gb a été fabriqué en technologie 24 nm [23].

Dans cette thèse, nous nous intéressons en particulier aux mémoires ReRAM bipolaires de type CBRAM (Conductive Bridge RAM). Les structures CBRAM étudiées dans ce cadre utilisent l'argent (Ag) comme métaux électrochimiquement actifs et de sulfure amorphe, agissat comme électrolyte.

B.1.2 SRAM non volatile pour applications embarquées à faible consommationes

Dans les applications mobiles à faible puissance on trouve souvent des mémoires Flash et des mémoires SRAM créées par deux différent macros, qui utilise une interface en série pour transférer l'information entre eux. Cette approche consomme de la puissance soit pour écrire la mémoire Flash avant extinction du dispositif soit pour restaurer les données dans la mémoire SRAM après avoir allumé le dispositif. De cette manière, on va limiter la durée de vie de la batterie. Néanmoins, cette opération implique un

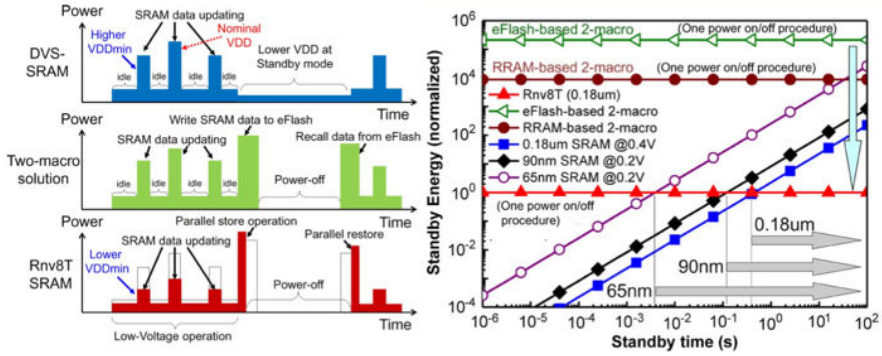


Figure B.2: (Gauche) Consommation de potence pour SRAM, la solution avec deux macro et nonvolatile SRAM pendant la phase active et du stand-by. (Droite) Comparaison du consommation pendant la phase du standby [33].

gain de puissance si le stand-by fonctionnant à 0.2 V prendre un temps plus grand que 10^3 s comme indiqué dans la Fig. B.2 (à droite) pour la technologie 65 nm CMOS. Pour ces raisons, on trouve dans [33] une mémoire SRAM de type non volatile qui intègre les cellules SRAM et le cellules ReRAM. Celles ci sont integrées dans le niveau des interconnections de la logique et vont permettre un transfert rapide de données de façon parallèle. En particulier, une SRAM non volatile basée sur des cellules ReRAMs à démontré que la procédure d' extention et d' allumage du dispositif vont contribuer au gain de puissance, si le correspondant stand-by prendre un temps plus grand que 2 ms. Cette fonction dépend principalement de l'énergie de commutation pour écrire et effacer la mémoire non volatile. Les operations de programmation de la mémoire SRAM non volatile et aussi la tolérance aux fautes seront expliquées dans la Sec. B.4.1.

B.1.3 Architecture hybride de type FPGA et ReRAM

L'amélioration des performances du FPGA en utilisant des dispositifs non volatiles de type ReRAM gagne en importance ces dernières années et peut être considérée comme un des sujets principaux de cette thèse avec le developpement de mémoire de type SRAM non volatile. Dans les FPGA, la possibilité d'intégration de la mémoire et la logique de manière distribuée sans surcoût en surface silicium donnera lieu à des architectures hybrides à l'intérieur des éléments de logique (comme non volatile Flip-Flop [47] [18] ou non volatile Look-up-Table) et des ressources de routage. Les nouveaux commutateurs de routage peuvent être réalisés en empilant mémoires émergents et dispositifs CMOS:

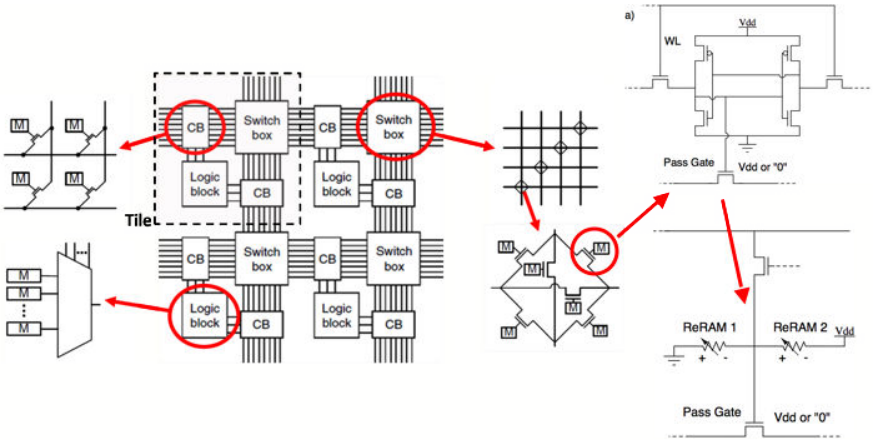


Figure B.3: Structure des blocs dans un FPGA. Les cellules 6T-SRAM (M) commandont les portes de routage dans les blocs de connexion (CB) et des blocs de commutation pour définir le routage général et pourront etre substitue par des éléments des commutations 1T-2CBRAM avec une configuration à pont diviseur.

pour réduire la consommation dans les FPGA, eliminer la mémoire flash à l'exterieur de la puce, pour obtenir un allumage instantané. En conséquence, nous nous attendons à réduire la consommation d'énergie liée à la reconfiguration des SRAM après la mise sous tension, et la consommation d'énergie dynamique en raison de la surface inférieure.

Dans des FPGA basés sur des cellules SRAM, le transistor de routage (pass gate) est gerée par une cellule 6T-SRAM. En utilisant deux mémoires de type CBRAM avec une configuration à pont diviseur au-dessus de la partie logique on peut développer des éléments de commutation non volatile avec une amélioration de la densité par rapport à la solution 2D. On appellera ces mémoire de configurations 1T-2R NVE. Dans [49] il a été démontré que le 3D-FPGA basé sur la technologie ReRAM, peut atteindre un gain de surface de 40% et un gain de 28% du produit délai énergie par rapport à une référence 2D-FPGA. Bien que la solution de NVE 1T-2R élimine la consommation d'énergie en mode stand-by, le courant de fuite à travers la ReRAM lors de l'exécution (c'est à dire en fonctionnement continue de lecture) dépend de la résistance de l'état résistif élevé. Maximiser l'état résistif élevé est essentiel pour réduire la consommation d'énergie statique lors de l'exécution FPGA. Par conséquent, l'ingénierie des matériaux et des conditions spécifiques de programmation, les quels devont etre compatible avec la logique, sont tenus de fournir une solution compétitive par rapport aux FPGA basées sur la technologie SRAM. Les résultats éléctriques obtenus sur different empilement

de type CBRAM pour augmenter le valeur du R_{off} seront discuté dans la Sec. B.3.3. Néanmoins, la courant de fuite dans les 6T-SRAM dépend fortement du V_{th}, l'épaisseur d'oxyde, et la taille, et va varier dans la gamme du picoampere [57], [58]. Cela implique qu'un R_{off} de plus que $10^{12}\,\Omega$ à une tension de lecture de 1 V devrait être ciblé afin de réduire la consommation d'énergie lors de l'exécution du FPGA.

B.1.4 Systeme neuromorphiques

La recherche dans le domaine neuromorphique a gagné beaucoup d'importance au cours des dernières années en raison de la faible puissance requise, la tolérance au pannes, et algorithmes ultra-adaptatifs [60],[61],[62],[63],[64]. Les réseaux de neurones sont utilisés pour classer des modèles basés sur l'apprentissage à partir d'exemples. Règles d'apprentissage définissent et ainsi permettent une modification de la conductance synaptique sur lequel repose l'effet de mémoire dans le cerveau biologiques. Règles de potentialisation à long terme (LTP) ou dépression de longue durée (LTD) définissent une amélioration ou dépression dans la transmission du signal entre deux neurones. Différents paradigmes de réseaux de neurones utilisent différentes règles d'apprentissage, mais dans la majorite de cas les règles vont déterminer les statistiques de motif à partir d'un ensemble d'échantillons de formation et ensuite classer de nouveaux modèles sur la base de ces statistiques. Les méthodes actuelles comme arrière propagation utilisent des approches heuristiques pour découvrir les statistiques. L'approche heuristique impliquent généralement beaucoup de petites modifications aux paramètres du système qui améliorent progressivement les performances du système. De plus, l'approche de l'adaptation progressive des back-propagation est susceptible de faux minima [65]. Pour améliorer cette approche, de nombreux algorithmes exploite nombres aléatoires pour améliorer l'apprentissage. En particulier, la littérature dans les domaines des réseaux de neurones [66], [67] et de la biologie [68] suggère que dans de nombreuses situations, en fait fournir un certain degré de stochastique, bruyant ou le comportement probabiliste dans leurs blocs de construction peut améliorer la capacité et la stabilité des systèmes neuromorphique. Certains types de réseaux neuronaux même fondamentalement s'appuyer sur les neurones stochastiques, comme les machines de Boltzmann. Pour cette raison on proposera dans la Sec. B.5 un neurone qui utilise la variabilite dans le temps de programmation pour montrer des characteristiques stochastiques.

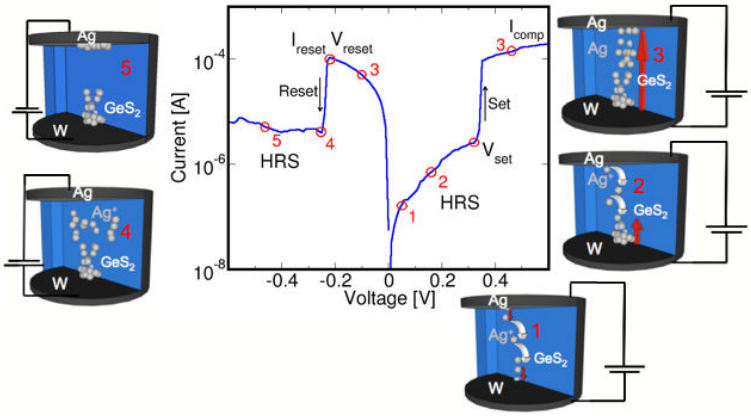


Figure B.4: Étapes du processus de commutation pour des cellules CBRAM à base de W-GeS_2-Ag et la caractéristique courant-tension correspondante (les paramètres principaux de commutation sont indiqués). Première étape: l'oxydation de l'électrode supérieure d'argent (Ag) et diffusion dans l'électrolyte GeS_2. Deuxième étape: réduction des ions Ag^+ dans l'électrode inférieure et nucléation de la nouvelle phase. Troisième étape: formation du filament conducteur (FC) riche en Ag (set) avec la commutation de l'état hautement résistif (HRS) vers l'état faiblement résistif (LRS). Quatrième étape: dissolution du FC avec ré-oxydation de l'Ag pendant le reset. Cinquième étape: réduction des ions Ag^+ dans l'électrode supérieure.

B.2 Caractérisation et modélisation de mémoire CBRAM basée sur Ag-GeS_2

Fig. B.4 affiche une caractéristique courant-tension pour une mémoire CBRAM à base de Ag-GeS_2 obtenu en régime continu. Au départ la cellule se trouve dans l'état hautement résistif (HRS). Pour faire passer la cellule de l'état HRS à l'état faiblement résistif (LRS), une tension positive est appliquée à l'anode d'argent qui s'oxyde, générant de l'Ag^+ ions (étape 1). Ces cations sous l'influence du champ électrique, migrent vers la cathode W où ils sont réduits pendant la formation du Filament Conducteur (FC) riche en Ag (étape 2). Au moment où le FC est assez grand pour créer un contact métallique avec l'électrode opposée, la cellule passe à l'état LRS à la tension de set (V_{set}). La conductance est limitée par le courant I_{comp}, qui peut être appliqué via le semi-conducteurs (SPA) ou un MOSFET intégré. Pour passer de l'etat LRS à HRS, une tension négative est appliquée. Pendant le processus de reset (étape 3), un courant électrochimique introduit des ions Ag $^+$. Ces derniers ne contribuent pas à la conduction

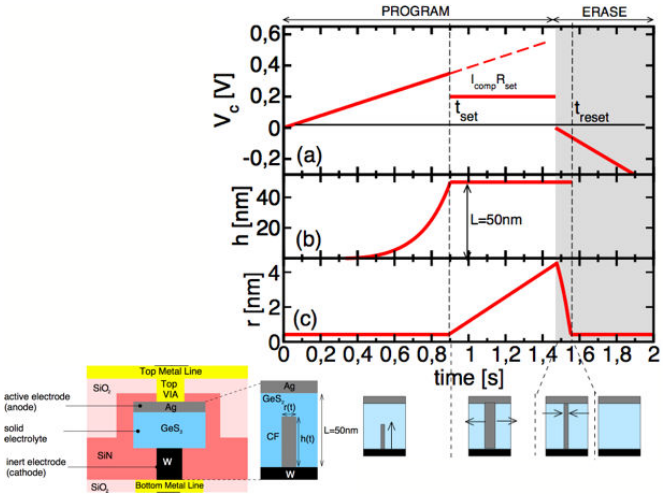


Figure B.5: Représentation schématique de la cellule CBRAM à base de W-GeS_2-Ag. On utilise un plug de tungstène (W) comme électrode inferieur. L'électrolyte est constitué de 50 nm de GeS_2 déposé par RF-PVD. Une couche d'argent est déposée en tant qu'électrode supérieure. Le FC est supposé cylindrique avec une hauteur $h(t)$ et un rayon $r(t)$. Une séquence simulée des opérations de set et de reset: (a) Tension DC appliquée à la cellule CBRAM; (b) évolution verticale et (c) évolution latérale du FC. Le set se produit lorsque le FC atteint l'électrode supérieure $h(t) = L$. Etant donné que le courant appliqué dans le dispositif est à sa valeur de saturation, la tension appliquée, V_c, diminue brusquement à une valeur constante: V_C=$R_{set}I_{comp}$ et le rayon du FC peut croître latéralement [79]. Au début du processus de reset, le FC commence à se dissoudre latéralement.

métallique dans les FC et sont réduits dans l'électrode d'Ag. Durant ce processus, la cellule passe de l'état LRS à HRS. Lors du reset, le FC peut être partiellement ou complètement dissous dans le GeS_2 dépendant de plusieurs facteurs (étape 4). Dans ce manuscrit, les cellules 1CBRAM isolées (1R) et les cellules 1T-1CBRAM (1T-1R) ont été mesurées électriquement.

La formation des FC est déterminée par la redistribution de masse associé au courant ionique $J(t)$. Dans ce manuscrit, nous considérons le FC cylindrique, avec un rayon $r(t)$ et une hauteur $h(t)$ (Fig. B.5), nous supposons que l'évolution verticale et latérale du FC sont proportionnelles à la densité de courant ionique [72]. Pour reproduire les conditions expérimentales, nous introduisons un paramètre (Δ) dans l'expression de Mott-Gurney (Eq. 2.1). En particulier si $V(t) < \Delta$, certains processus comme l'oxydation de l'argent, la migration des ions à travers le chalcogénure, le transfert d'électrons entre les ions et la cathode ou la nucléation d'une nouvelle phase à travers la

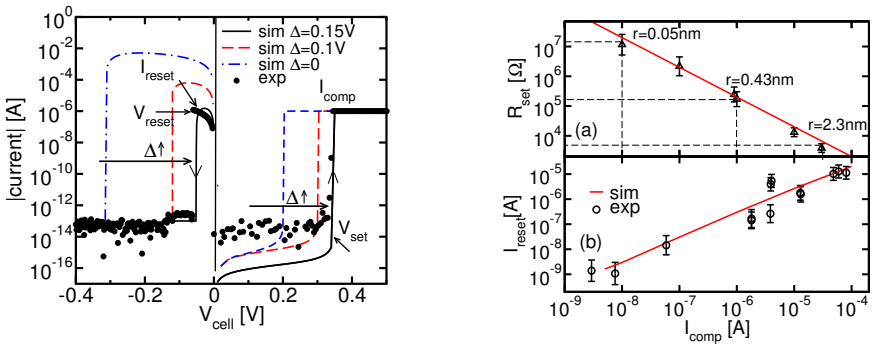


Figure B.6: (Gauche) Courbes expérimentales (symboles) et simulées (lignes) courant-tension obtenues en appliquant une rampe de tension. Les données expérimentales montrent une forte asymétrie pour la tension de set et du reset. Différents Δ changent V_{set}, V_{reset} et I_{reset}. (Droit a) Dépendance du R_{set} avec I_{comp}. (Droit b) I_{reset} en fonction de I_{comp} forcée lors de l'opération de set précédent.

cathode ne sont pas suffisamment activées pour permettre le passage de l'état HRS vers l'état LRS. Fig. B.5 illustre la procédure adoptée pour simuler une transition set/reset en DC (ou quasi-statique). La tension appliquée aux CBRAM à base de GeS_2 (50 nm) est illustrée en (a) ainsi que la hauteur verticale $h(t)$ correspondante (b) et latérale $r(t)$ (c) évolution des FC selon les équations Eq. 2.1 et Eq. 2.2.

Fig. B.6 (gauche) montre la caractéristique IV quasi-statique expérimentale et simulée. Une asymétrie dans le set et le reset apparaît. Des simulations avec différents Δ ont été effectuées pour s'adapter aux résultats expérimentaux V_{set} et I_{reset}. Une augmentation de Δ réduit la chute de tension efficace à l'intérieur de l'électrolyte solide, la décélération ainsi que la croissance verticale et latérale du FC. Fig. B.6 (droite) montre la résistance d'ensemble obtenue en programmant des CBRAM avec différentes valeurs de I_{comp}. Pour étudier la dépendance de V_{set} avec la cinétique du processus de commutation, on a fait varier la pente γ du signal de tension sur sept ordres de grandeur tout en enregistrant les tensions de set obtenues (Fig. B.7). Fig. B.7 (droite) affiche le temps de commutation en fonction de la tension appliquée à la cellule. L'inverse du temps de commutation est une fonction exponentielle de la tension appliquée. Cependant, à des tensions inférieures, t_{set} augmente beaucoup plus rapidement. Le paramètre Δ peut expliquer la saturation à faible tension pour les deux courbes.

L'effet de la température sur les caractéristiques IV a été aussi étudié. Lorsque la température augmente, V_{set} diminue tandis que V_{reset} reste à peu près constante. Les

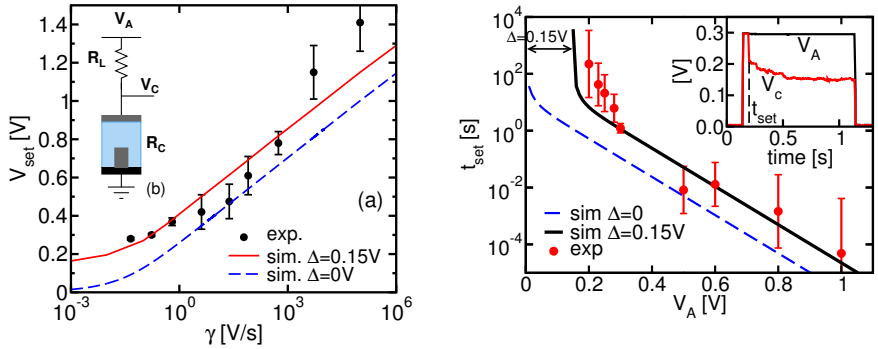


Figure B.7: (Gauche a) Tensions de commutation expérimentales (symboles) et simulées (lignes) V_{set} en fonction de la pente de la rampe. L'introduction du paramètre Δ dans le modèle permet de reproduire la saturation de V_{set} pour les pentes très faibles. (B gauche). Représentation schématique du protocole expérimental pour effectuer un test pulsé. (Droite) Temps de commutation t_{set} expérimentaux (symboles) et simulés (lignes) en fonction de la valeur de la tension appliquée. Les simulations réalisées avec $\Delta = 0.15$ V reproduisent l'augmentation brusque de t_{set} lorsque $V_{\text{A}} \simeq 0.2$. L'inset est un signal d'oscilloscope typique de V_{c} et V_{A} lors d'une opération de set.

résultats expérimentaux sont bien reproduits par les simulations. Le modèle est capable de prédire une augmentation de V_{set} pour les basses températures.

Nous avons caractérisé le comportement de la commutation de 64 cellules organisées dans une matrice mémoire 1T-1R NOR 8×8. Le transistor en série a une longueur L=140 nm et une largeur W=500 nm. Le but de ce paragraphe est de fournir une analyse statistique de la corrélation entre les conditions de programmation et le pourcentage de cellule CBRAM dans le réseau de mémoire qui peuvent basculer entre l'état HRS et LRS (et vice-versa). L'efficacité du set est rapporté dans la Fig. B.10. L'efficacité du set est défini comme la moyenne du pourcentage de cellules qui commutent pendant le premier (ou le second cycle) et le pourcentage de cellules qui commutent dans les deux cycles. Par exemple, sur 64 cellules, un total de 42 cellules sont activées dans le premier set ou dans la seconde opération du set, tandis que 36 commutent dans les deux cycles. L'efficacité de l'ensemble est donc de 60%. Par convention, les cellules sont prises en compte dans l'état LRS si la résistance est inférieure à 20 kΩ, ou dans l'état HRS si la résistance est supérieure à 200 kΩ. Fig. B.10 (gauche) affiche l'efficacité du set en fonction de la tension sur l'anode pour une tension de wordline constante de 1.5 V (à gauche). L'efficacité augmente avec la tension sur l'anode et aussi avec la largeur

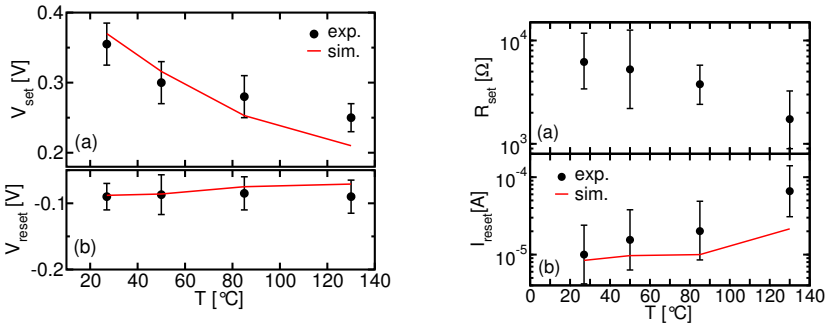


Figure B.8: (Gauche) V_{set} (a) et V_{reset} (b) expérimentaux (symboles) et simulées (lignes) en fonction de la température. (Droite) Courbe expérimentale (symboles) et simulée (ligne) de R_{set} (a) et I_{reset} (b) en fonction de la température.

d'impulsion, même si l'effet de la tension est plus fort. Des impulsions plus longues augmentent l'efficacité de set, lorsque la même tension V_a est appliquée. Fig. B.10 (droit) affiche l'efficacité du reset.

B.3 Ingénierie de l'empilement pour augmenter R_{off}

Dans les systèmes hybrides (mémoires non volatiles et logiques) la fonctionnalité du circuit est fortement dépendante des caractéristiques de commutation des cellules ReRAM intégrés. Ces dernières dépendent elles-mêmes du circuit logique (tel que les MOSFET d'accès) utilisé pour programmer la mémoire. La performance du circuit peut être améliorée en utilisant différents matériaux actifs pour les CBRAMs, afin de satisfaire les spécifications de l'application visée. Par exemple, il a été proposé pour les circuits FPGA de gérer les éléments de routage avec deux CBRAM en séries en utilisant une configuration type pont diviseur de tension. Dans ce cas-là, l'optimisation du R_{off} est très importante. Cette architecture nécessite des recherches au niveau de la mémoire résistive pour augmenter la valeur de l'état HRS, autrement dit elle ne peut pas être compétitive avec les solutions existantes. Le but de ce chapitre est de trouver des solutions au niveau des empilements pour augmenter la résistance du off en vue de développer des commutateurs de routage.

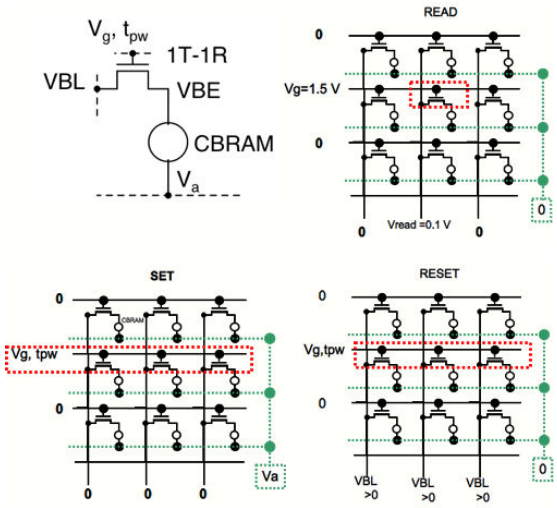


Figure B.9: (Gauche haut) Schéma de la structure 1T-1R. (Droit haut) Schéma du réseau mémoire 8×8 NOR (seulement trois lignes représentées). Lors de la lecture (droit haut), l'anode (V_a) est à la masse, la bitline (VBL) est polarisée à 0.1 V et la wordline à 1.5 V. Pendant le set (gauche bas), la bitline est à la masse, la grille est en pulsée (V_g, t_{pw}) et l' anode est polarisée à 1.5 V. Pendant le reset (droite bas), l'anode est à la masse, la grille est en pulsée et la bitline est polarisée.

B.3.1 Ta-TaO$_x$-GeS$_2$-Ag CBRAM

Nous proposons une CBRAM constituée de Ta-TaO$_x$-GeS$_2$-Ag comme solution possible pour augmenter R_{off}. Les caractéristiques IV en statique sont présentées dans la Fig. B.11 pour des structures 1T-1R. Il est intéressant de noter que pendant les premiers cycles la tension de set apparaît aux alentours de 0.5 V et un second V_{set} autour de 2 V. Les cycles suivants montrent une tension de set équivalente à la première (V_{set1}). Seulement quelques cycles montre un I_{off} très faible. De ce fait, le rapport $R_{\text{off}}/R_{\text{on}}$ est fortement réduit dû à la diminution de R_{off}. La double tension de set pourrait s'expliquer par une légère rupture du GeS$_2$ suivie par la rupture de la couche TaO$_x$ à tension plus élevée. Nous supposons que les ions Ag diffusent dans le TaO$_x$ et dégradent l'oxyde après quelques cycles Set/Reset, de ce fait le TaO$_x$, censé agir comme un obstacle à la conduction à l'état bloqué devient inefficace.

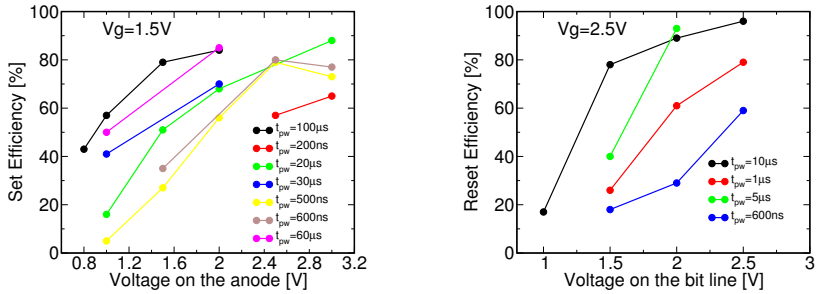


Figure B.10: (Gauche) Efficacité du set (probabilité de commutation) pour 64 cellules de la matrice tout en variant la tension appliquée sur l'anode (tension de grille 1.5 V. Efficacité du reset (probabilité de commutation) pour 64 cellules de la matrice tout en faisant varier la tension sur la bitline (tension de grille 2.5 V). Une commutation est considérée comme réussie si R_{on} inferieur à 20 kΩ et R_{off} supérieur à 200 kΩ.

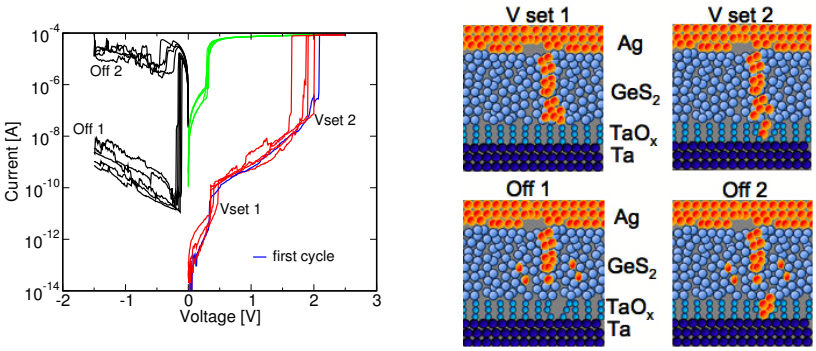


Figure B.11: (Gauche) La caractéristique I-V en statique pour une CBRAM à base de 1T-1R Ta-TaO_x-GeS_2-Ag. (Droite) Schéma des étapes pendant le set et le reset.

B.3.2 W-SiO_x-GeS_2-Ag CBRAM

Nous avons aussi développé une CBRAM composée de W-SiO_x-GeS_2-Ag. Les caractéristiques IV en statique sont présentées dans la Fig. B.12 pour le dispositif 1T-1R. La cellule CBRAM nécessite une opération de forming à une tension de 1.3 V. Les 40 premiers cycles montrent un rapport R_{off}/R_{on} très élevé de 10^6, après que la fenêtre de mémoire soit réduite, le R_{off} stabilise autour de 100 kΩ.

B.3.3 W-HfO_2-GeS_2-Ag CBRAM

Nous avons étudié trois empilements de matériaux actifs différents dans les CBRAM. Le premier empilement est le W-GeS_2-Ag utilisé comme référence. Le deuxième et le

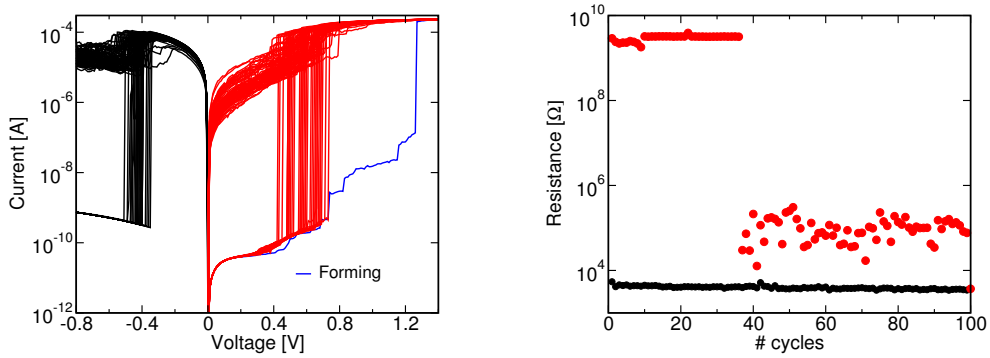


Figure B.12: (Gauche) Caractéristiques I-V en statique pour une CBRAM à base de W-SiO$_x$-GeS$_2$-Ag (100 cycles). (Droit) Evolution du R_{on} et R_{off} lors du cyclage.

troisième empilement ont des couche de HfO_2 supplémentaires de 1 nm et SI2nm respectivement insérées entre le W et le GeS_2 de 30 nm. La couche de HfO_2 a été déposée par atomic layer deposition (ALD). Nous avons effectué des mesures électriques à la fois sur les dispositifs 1T-1R isolés et les 8×8 matrice de mémoire. Les caractéristiques typiques I-V en statique sont présentées dans la Fig. B.14. Les moyennes SI0.35V/−0.2 V (GeS_2), 0.4 V/−0.3 V(HfO_2(1 nm)+GeS_2), 0.5 V/−0.4 V (HfO_2(2 nm)+GeS_2) sont montrées. Il est intéressant de souligner qu'aucune étape de formation est nécessaire pour le cas HfO_2(2 nm)+GeS_2. De ce fait R_{on} est d'environ 5 kΩ pour les trois empilements et R_{off} augmente de manière significative avec l' épaisseur de la barrière HfO_2. Un rapport de résistance de deux, cinq voir six ordres de grandeur a été obtenu dans le GeS_2, HfO_2(1 nm)+GeS_2 et HfO_2(2 nm)+GeS_2 .

En conclusion, les effets d'ingénierie de l'interface dans des mémoires W-GeS_2 ou Ta-GeS_2 CBRAM ont été analysés. Parmi les autres technologies CBRAMs mesurée, notre (2 nm HfO_2−30 nm GeS_2) a un taux de résistance (R_{off}/R_{on}) de plus que 10^6, une courant de reset current de 100 µA sans étape de forming. De ce qui précède l' empilement HfO_2(2 nm)+GeS_2 a été choisi comme le candidat le plus prometteur pour l'application comme commutateur de routage. Nous allons discuter plus en détail l'architecture et les avantages prévus par rapport aux solutions existantes.

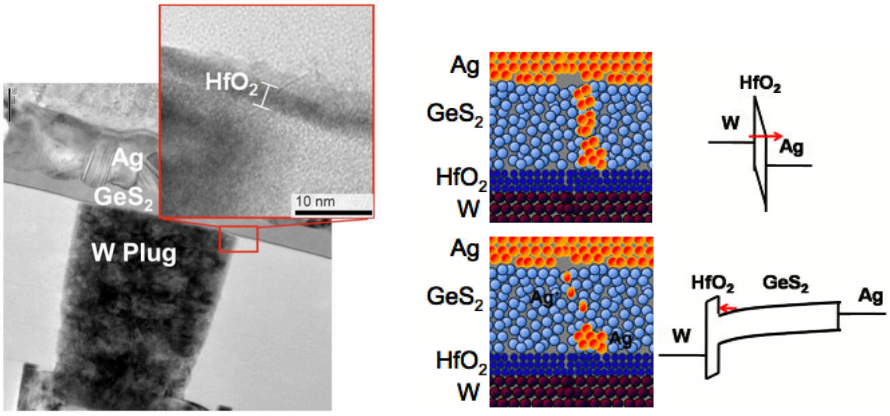


Figure B.13: (Gauche)Image TEM d'une HfO_2(2 nm)+GeS_2 CBRAM. (Droit) Schéma des set et reset et diagramme de bande (semi-conducteur GeS_2 de type p et Φ(W)>Φ(Ag).

B.4 Nouvelles architectures hybrides: Logique et Mémoire ReRAM

B.4.1 NVSRAM basée sur des elements OxRRAM

Le schéma électrique d'une 8T2R NV-SRAM est présenté sur la Fig. B.15. Cette structure est assemblée avec une cellule 6T-SRAM (M1-M6) et avec deux transistors PMOS de contrôle (CM1 et CM2) connectés entre les nœuds de donnée (D, DN) des cellules SRAM et OxRRAM (R1, R2). Lors de l'opération d'écriture, l'état logique de la structure SRAM est stocké dans les cellules OxRRAM. On suppose que les nœuds D et DN sont respectivement dans les états 1 et 0. Afin de retrouver l'information enregistrée dans les cellules OxRRAM, les transistors de contrôle sont enclenchés en abaissant la tension CTRL1 jusqu'à 0V et en mettant CTRL2 à la masse. En conséquence, la différence de potentiel positive entre les électrodes supérieures et inférieures de R1 fixe la résistance à une valeur très faible (LRS) permettant ainsi de stocker l'information dans la structure OxRRAM. Comme dans le même temps aucune chute de tension ne se produit à travers R2, la valeur de sa résistance reste inchangée dans le cas d'un état résistif élevé, ce qui dans ce cas, permet de stocker l'état 0. Afin d'alimenter suffisamment le nœud TE en tension, un transistor PMOS est préféré à un transistor NMOS.

Dans un second temps, l'état logique de la cellule SRAM doit être restauré durant la remise en marche du circuit. Cette séquence se décompose selon l'enchaînement suivant:

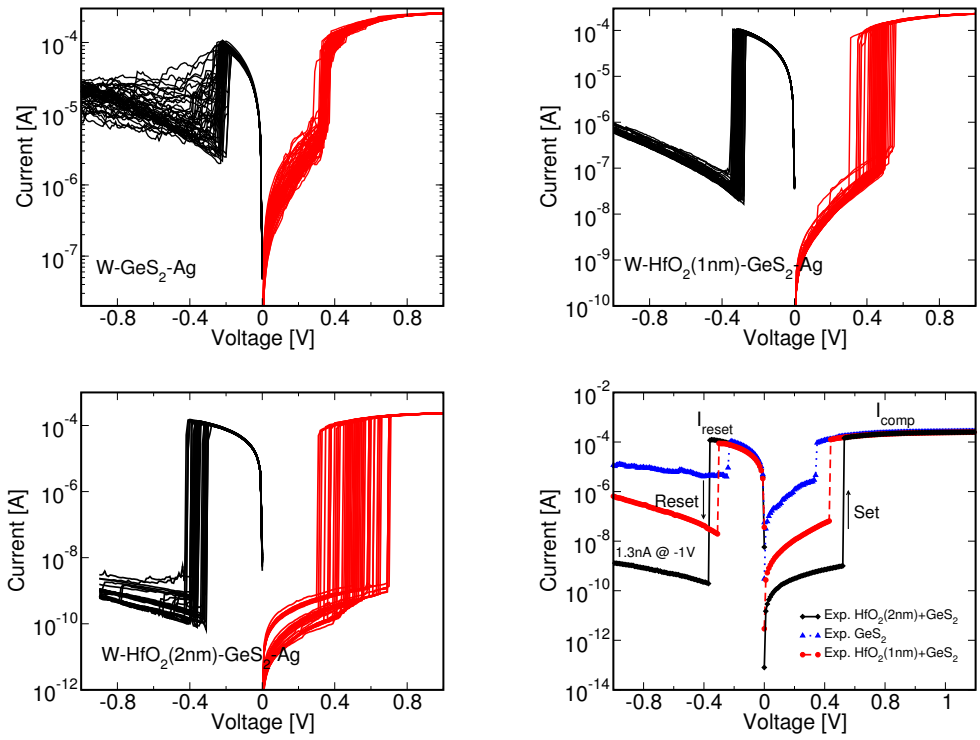


Figure B.14: Caractéristique I-V en régime pour GeS_2 (gauche haut), HfO_2(1 nm)+GeS_2 (droit haut) et HfO_2(2 nm)+GeS_2 (gauche bas) CBRAM. Comparaison pour les trois types (droit bas).

CTRL1 est mis à la masse enclenchant ainsi les transistors de contrôle, puis la tension VSS est mise à 0V avec un retard de 5ns par rapport à CTRL1, et CTRL2 est maintenu à la tension VDD comme durant l'arrêt de l'alimentation du circuit. Lorsque VSS est mis à 0V, la différence de résistance entre les deux états résistifs entraîne non seulement deux courants de décharge différents, mais aussi une différence de tension entre les nœuds D et DN amplifiée par la cellule SRAM. De plus, la faible valeur résistive de R1 permet de maintenir le nœud D à la valeur 1 tandis que le nœud DN est mis à 0 en raison de la forte valeur résistive de R2. De cette manière, l'information contenue dans la cellule NVSRAM est restituée.

Afin d'étudier la stabilité du processus de restauration de l'information, le système est soumis à un cas particulier dans lequel la tension de seuil varie selon $\pm n\sigma V_{\mathrm{T}}$ pour chaque transistor de la structure SRAM. Dans cette situation, chaque transistor fonctionne dans un cas défavorable à la restauration de l'information de la cellule SRAM. L'augmentation

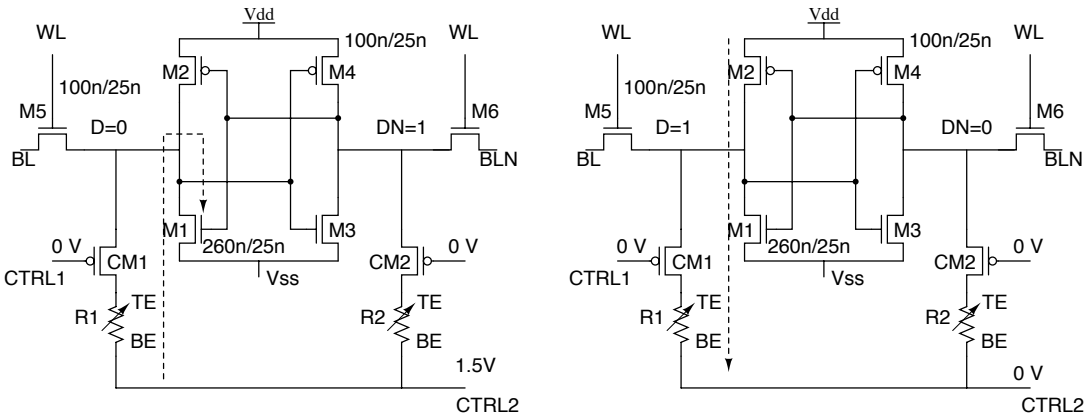


Figure B.15: Représentation schématique des 8T2R SRAM cellules non volatiles (NVS-RAM). Les lignes pointillées indiquent le chemin du courant pendant l'opération de RESET (à gauche) et la mémorisation (à droite).

de n entraîne un cas plus défavorable et permet d'extraire la valeur seuil de $\pm n\sigma V_{\mathrm{T}}$ pour laquelle la restauration de l'information échoue. Dans le schéma présenté en Fig. 4.3, la restauration de l'information n'affecte pas la valeur de la résistance de la cellule OxRRAM. De ce fait, l'utilisation de simples résistances substituant R1 et R2 est équivalente à l'utilisation d'éléments OxRRAM. D'autre part, cette substitution permet d'ajuster les valeurs de R1 et R2. La Fig. B.16 montre les valeurs maximales de n pour lesquels la restauration de l'information échoue. Dans cette simulation, la valeur référence de la résistance R_{off}. La diminution de R_{on} renforce l'opération de restauration de l'information suivant deux manières. D'une part, le fait d'abaisser la valeur de permet d'avoir un courant plus fort et une restauration de l'information plus importante. D'autre part, une plus grande différence de résistance entre R_{off} et R_{on} augmente la différence de courant de chaque coté de la cellule. Ces deux phénomènes améliorent la stabilité de la restauration de l'information. D'après notre modèle, utilisé également pour un transistor PMOS d'une largeur de 200nm et un temps de programmation de 200ns, R_{on} durant le stockage ne dépasse pas 21 kΩ, correspondant à un rapport de 4. L'analyse du pire cas montre que pour un même rapport de résistance l'opération de restauration de l'information échoue pour n=2 correspondant à une trop faible valeur.

D'autre part, la fiabilité du système OxRRAM-SRAM durant la restauration de l'information est analysée via la méthode Monte Carlo. La Fig. B.16 illustre les résultats des analyses Monte Carlo faites sur 10 000 échantillons pour une structure PMOS de

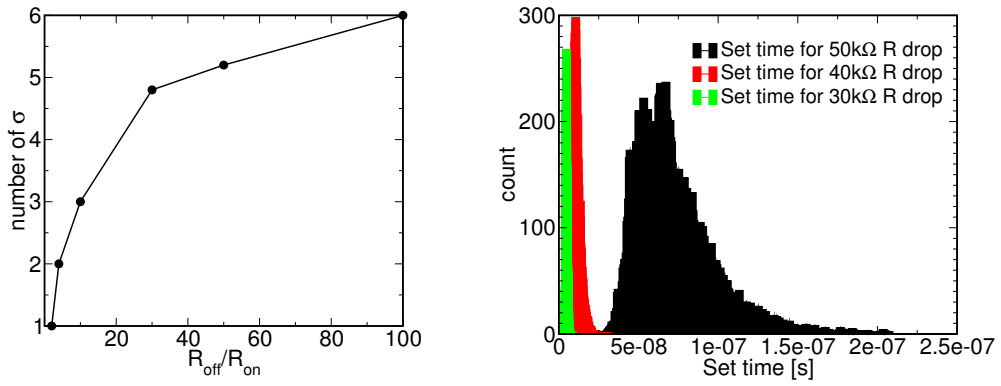


Figure B.16: (Gauche) Variation du $\pm\sigma V_{\mathrm{T}}$ a V_{T} fail point en fonction du rapport de la résistance à l'analyse de worst case. (Droite) Distributions of store time requi pour obtenir different $\mathrm{R_{drop}}$ (10k samples Monte Carlo simulation pour un largeur du pmos egal à 100 nm.

100nm de largeur. Ce résultat met en avant l'étendue du temps de restauration pour obtenir une difference entre le valeurs de résistances comprise entre 30 kΩ et 50 kΩ. Par conséquent, la seule méthode permettant d'avoir un fort rapport de résistance et ainsi d'avoir un phénomène de restauration d'information plus stable consiste soit à augmenter la largeur du transistor PMOS, soit d'utiliser une autre structure de mémoire résistive avec une plus forte valeur de R_{off}.

B.4.2 1T-2CBRAM en tant qu' élément non volatile dans les FPGA

Dans cette section, nous proposons une nouvelle intégration de dispositifs CBRAM utilisés dans une architecture à diviseur de tension, afin non seulement de commander un transistor mais aussi de stocker des données dans un tableau de référence (Look-up Table ou LUT). Cette architecture nécessite une très grande attention dans l'ingénierie de l'interface afin de maximiser R_{off} et donc de réduire la consommation d'énergie. En effet, lors de la mise en marche de l'alimentation du circuit associé, le courant de fuite à travers la ReRAM dépend de la résistance de l'état hautement résistif. Maximiser cet état est donc essentiel afin de réduire la consommation d'énergie statique lors de l'exécution de la cellule FPGA. Dans les sections précédentes, nous avons discuté des conditions du set/reset ainsi que de leurs effets sur les niveaux de résistance. Il en a été conclu que $\mathrm{HfO_2}$(2 nm)+$\mathrm{GeS_2}$ est une solution prometteuse pour obtenir une très haute R_{off}. Dans ce chapitre, nous présentons des simulations Eldo en régime transitoire pour

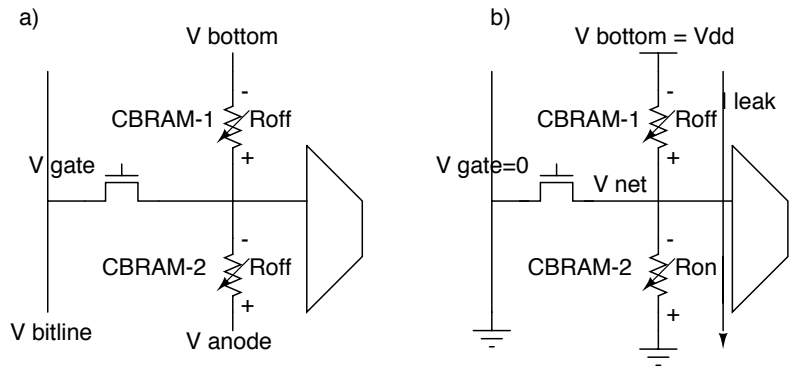


Figure B.17: a) 1T-2R NVE dans l'état initial: les deux CBRAM (1 et 2) sont dans l' HRS. b) 1T- 2R NVE programmée pour avoir un zero logique sur V_{net} pendant une opération de lecture, après le set de la CBRAM 2 à partir de HRS vers LRS.

atteindre les niveaux de résistance complémentaires dans les deux CBRAM. Pour ce faire, nous avons utilisé notre modèle compact. Initialement, les deux CBRAM sont en état de résistance élevée (High Resistive State ou HRS). Lorsque V_{anode} et V_{gate} sont élevés, les cellules CBRAM passe à l'état de basse résistance (Low Resistive State ou LRS). Dans ce cas, le NVE 1T- 2R est configuré pour un zéro logique appliqué à V_{net} au moment de l'exécution du FPGA.

Pour les architectures logique reconfigurable, une grande endurance du procédé de restauration de l'information pendant les tests de cyclages d'environ 10^3 est nécessaire. Ainsi, un test d'endurance a été effectué sur des dispositifs 1T-1R à des conditions de programmation qui permettent de maximiser le rapport de résistance entre les différents états. Les paires de résistances obtenues expérimentalement ont été utilisées pour calculer la tension entre les deux CBRAM connectés en série afin d'estimer les écarts par rapport à un zéro logique. Le pire cas déterminé correspond à la tension calculée avec le ratio de résistance plus faible dans le test de cyclage. Il est intéressant de noter qu'en fonctionnement continu à lire la tension sur V_{net} peut conduire à une commutation indésirables de LRS à HRS pouvant entraînement la perte du contenu logique dans le NVE 1T- 2R . La commutation non désirée peut se produire parce que la tension continue non nulle est toujours présent sur l'électrode inférieure de la CBRAM -2. Pour approfondir ce point, nous avons effectué une analyse en régime continu (DC) sur un réseau de mémoires CBRAM. Les cellules ont été programmées dans l'état LRS. Le temps de commutation (t_{switch}) a été défini comme le temps nécessaire

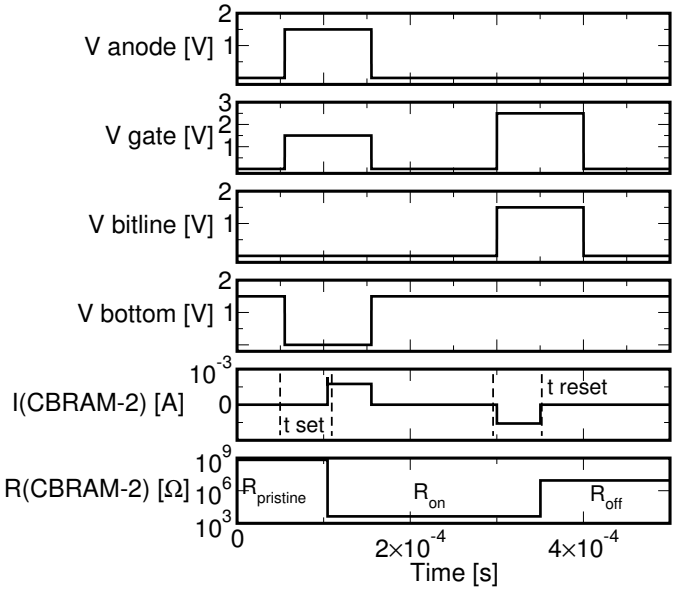


Figure B.18: Simulation transitoire sur la programmation complémentaire de la structure 1T-2R NVE dans GeS_2 CBRAM. Le temps nécessaire pour changer l'état résistif de la CBRAM-2 (Fig. 4.10) à partir du HRS vers LRS est 50 µs avec V_{anode}=1.5 V. Dans l'état LRS, R_{on} est 3 kΩ. L'opération de reset nécessite un temps de 50 µs avec V_{bitline}=1.5 V.

pour augmenter la résistance d'un facteur 10. Dans la figure Fig. B.20, nous avons extrapolé une stabilité sur dix ans de l'information pour des tensions allant de 40 mV à 6 mV pour HfO_2(2 nm)+GeS_2 et GeS_2 respectivement. Ainsi, avec la couche de HfO_2 optimisée, non seulement les mémoires peuvent être programmés plus de 1000 cycles avec un rapport de résistance très élevé, mais acquièrent une très grande stabilité du point du vue du stockage de l'information durant le processus de lecture.

Fig. B.20 présente une comparaison des valeurs R_{off} et R_{on} obtenues lors des tests de cyclage pulsés pour notre HfO_2(2 nm)+GeS_2 et d'autre technologies ReRAM. La solution proposée offre le meilleur $R_{\text{off}}/R_{\text{on}}$ ratio présenté.

B.5 Neurones stochastiques

Dans cette section, nous présentons une méthodologie originale pour concevoir des circuits neuronaux hybrides (CMOS + mémoire résistive non volatile) avec un comportement stochastique par rapport à l'opération de spiking. Pour ce faire, nous utilisons la variabilité intrinsèque des CBRAM, en particulier la variabilité sur le temps du set et

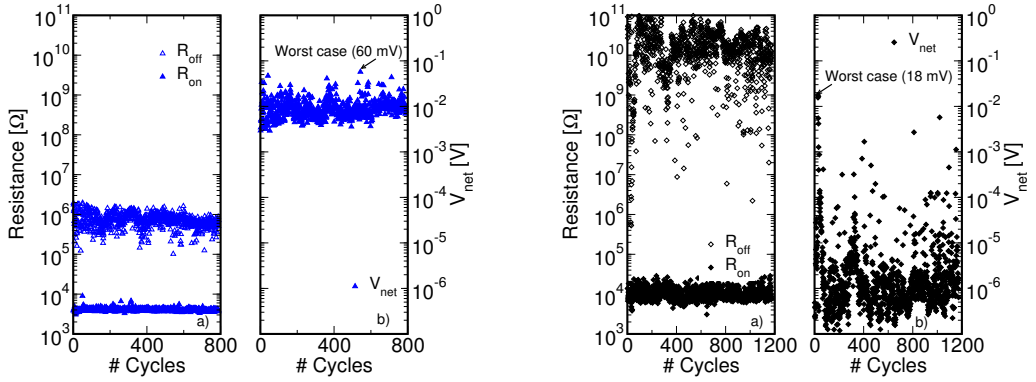


Figure B.19: (a) Test pulsée sur GeS_2 (gauche) et HfO_2(2 nm)+GeS_2 (droit) CBRAMs. (b) Tension entre le point diviseur de tension (V_{net} in Fig. B.17) calculé en utilisant chaque paire de résistance de le test du cyclage (a).

sur l'état de haute résistivité pour les structures faites en Ag-GeS_2. Nous proposons ici un circuit et une technique d'auto-programmation pour l'utilisation de dispositifs CBRAM dans des neurones de type Integrate and Fire.

Pour introduire un comportement stochastique dans un neurone Integrate and Fire, nous connectons un dispostif CBRAM au condensateur C_{mem}, tel que C_{mem} puisse se décharger à travers le dispositif CBRAM en faisant basculer son état à l'état de basse résistivité. Ainsi, l'anode de la CBRAM et le point V_{mem} doivent être connectés. La durée pendant laquelle le courant circule à travers la cellule CBRAM dans l'état ON (i.e de basse résistivité) peut être contrôlé avec un transistor. Dans une telle configuration, la distribution sur le t_{set} de la CBRAM se traduirait par une distribution sur le temps de décharge (t_{DSC}) du condensateur, ce qui conduirait à une distribution de tension sur C_{mem}, et en conséquence à une reponse stochastique.

B.6 Conclusion

Cette thèse aborde les différents aspects des technologies CBRAM dans le but de développer des architectures hybrides. Des structures CBRAM ont été testées électriquement afin de comprendre les conditions de programmation ainsi que de déterminer les paramètres de commutation, le ratio de la mémoire et le courant de reset. Un modèle se basant sur les sauts activés en température a été développé afin de décrire la migration ionique et les phénomènes de croissance/dissolution des filaments qui en résultent durant

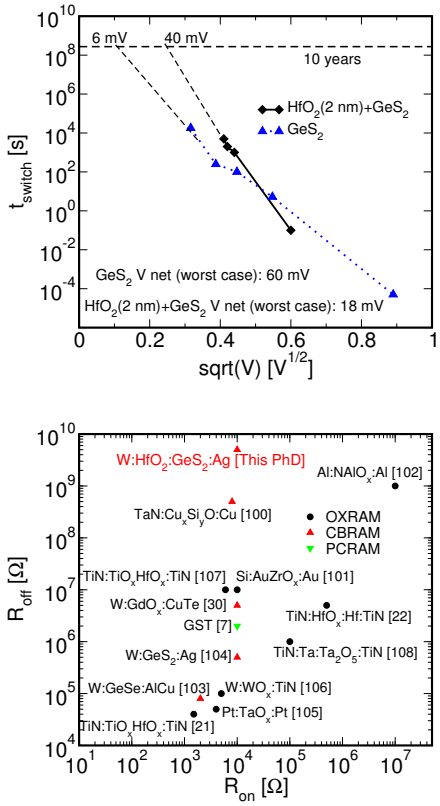


Figure B.20: (Haut) Temps du switch (t_{switch}) entre l'état LRS et l'état HRS lors d'un stress négatif en tension pour les structures en GeS_2 and $HfO_2(2\,\text{nm})-GeS_2$. Chaque point correspond à la moyenne calculée sur 64 cellules. Projections sur 10 ans pour la perturbation en lecture 6 mV et 40 mV sont extraites. (Bas) Résumé pour les états LRS (R_{on}) et HRS (R_{off}) pour certaines ReRAMs obtenu avec des tests de cyclage.

les opérations set/ reset. Les paramètres de ce modèle sont extraits de mesures DC quasi-statiques et de mesures pulsées. Ils sont utilisés pour implémenter le modèle compact utilisés pour les simulations électriques des principales opérations ayant lieu au sein du circuit. Dans le chapitre 1, nous avons présenté l'état de l'art de la recherche dans le domaine des mémoires. Pour cela, nous avons présenté des mémoires non volatiles compatibles avec une intégration à très petite échelle (le nœud sub-20nm), et utilisant de faibles puissances. Nous nous sommes focalisés sur des SRAM qui intègrent des cellules 6T-SRAM et des dispositifs mémoires non volatiles utilisant des connexions bit-to-bit en 3D ou respectant un arrangement vertical dans le but d'atteindre un transfert rapide de données parallèles ainsi qu'une vitesse rapide. En empilant des mémoires émergentes

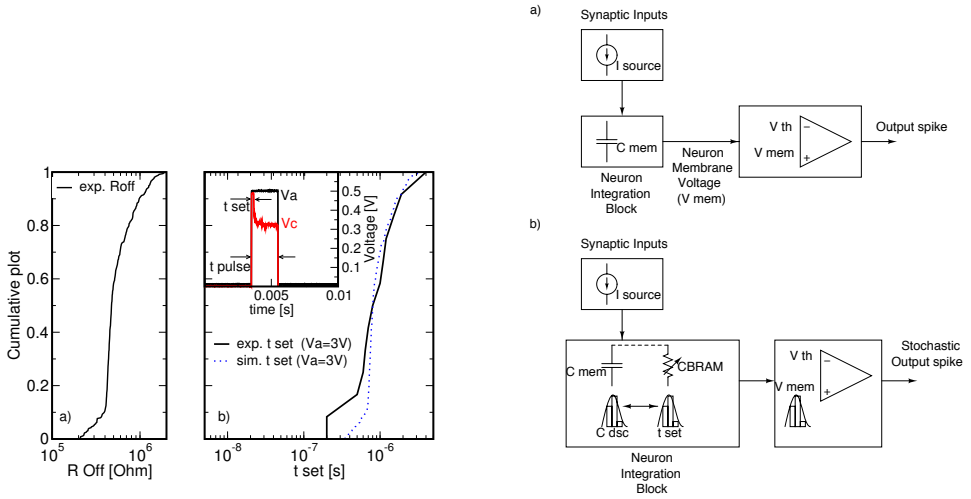


Figure B.21: (Gauche) (a) Distribution de R_{Off} (cyclage) dans Ag-GeS_2-W 1R CBRAM.(b) Résultat expérimental (ligne) et simulation (point) de la distribution de t_{set} obtenue avec un cyclage utilisant une amplitude de pulse V_a=3 V. (b dans l'encart) Exemple d'une trace d'oscilloscope pour connaitre la tension sur la CBRAM (V_c). (Droit) (a) Schéma électrique pour montrer le concept de neurone de type Integrate and Fire ou bien le concept d'un neurone stochastique Integrate-Fire (S-IF) (b).

avec des dispositifs CMOS, on peut utiliser de nouvelles procédures de commutation permettant de réduire la surface du FPGA, de s'affranchir des mémoires Flash externes, et ainsi réduire la surface du circuit et obtenir l'état ON de manière instantanée. Par conséquent, on peut réduire la consommation en puissance du FPGA basée sur des SRAM. Ceci est relié à la reconfiguration de la SRAM, ainsi qu'à la consommation de puissance dynamique due à la réduction de surface. Malgré que les solutions de routage pour mémoires hybrides éliminent la veille en terme de consommation, le courant de fuite à travers la ReRAM durant le temps de marche dépend de la résistance de l'état hautement résistif. Nous avons conclu ce chapitre sur une réflexion autour de l'état hautement résistif: celui-ci doit être maximisé afin de réduire la consommation statique durant le temps de fonctionnement du FPGA. Le chapitre 2 introduit les équations du modèle empirique utilisées pour expliquer et prédire les principaux paramètres de commutation des dispositifs CBRAM mesurés à l'aide des caractérisations électriques dans les configurations quasi-statique et pulsée. Nous avons effectué une analyse statistique de la corrélation entre les conditions de programmation et le pourcentage de dispositifs CBRAM intégrés dans un réseau de mémoire 8×8 NOR commutant de manière réversible

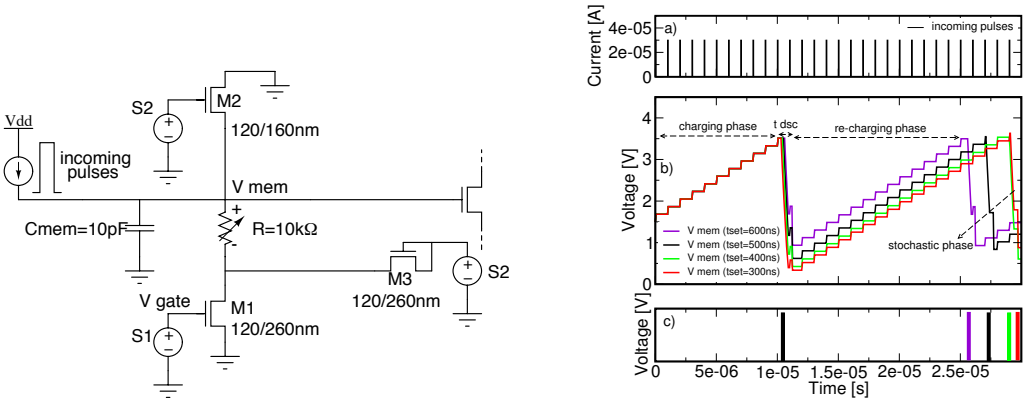


Figure B.22: (Gauche) Circuit utilisé pour démontrer le concept pour un neurone stochastique lorsque le CBRAM est dans l'état LRS. (Droite) Evolution de V_{mem} simulant le circuit représenté à gauche. (a) les impulsions entrantes sont utilisées pour augmenter la tension sur V_{mem}. (b) Au départ, V_{mem} s'accumule en raison des courants entrants (phase de charge). L' opération de set conduit à différentes décharges de C_{mem}. Pendant la phase de recharge un nombre différent d'impulsions permettra de réhausser V_{mem} jusqu'à V_{th}. (c) différents intervalles inter-spike affichés en fonction du t_{set}.

entre un état hautement résistif (HRS) et un état faiblement résistif (LRS). Nous avons développé une réflexion autour des causes possibles de cette variabilité des cellules que nous avons exploitée par la suite dans des applications neuromorphiques utilisant les CBRAM comme synapses stochastiques. Dans le chapitre 3, quatre différents stacks de CBRAM ont été étudiés en vue d'une utilisation dans les architectures hybrides: i) W-GeS_2-Ag, ii) Ta-TaO_x-GeS_2-Ag, iii) W-SiO_x-GeS_2-Ag, iv) W-HfO_2-GeS_2-Ag L'objectif principal de cette partie est d'améliorer l'architecture 1T-2R NVE. Parmi les technologies de CBRAM caractérisées, notre empilement bicouche électrolyte (2 nm HfO_2−30 nm GeS_2) a permis d'obtenir un ratio de résistance (R_{off}/R_{on}) plus élevé que 10^6, ainsi qu'un courant de reset inférieur à 100 µ A sans l'étape de forming. Nous avons expliqué également l'amélioration du ratio de résistances de la mémoire à l'aide de la modélisation physique. Enfin, nous avons exploité la variabilité des paramètres de commutation des CBRAM pour dessiner un neurone stochastique pouvant être utile dans un réseau de neurone. Les circuits proposés présentent différents avantages par rapport aux autres méthodes utilisées pour implémenter la stochasticité telles que le PRNG ou les faibles conditions de programmation utilisés pour les synapses.

List of Figures

List of Tables

Bibliography

[1] K. Zhang, "Isscc 2013 trends," 2013. 2, 3, 129

[2] M. Corporation 2013. 2

[3] "Kingston corporation," 2013. 2

[4] K. Zhang, *Embedded Memories for Nano-Scale VLSIs.* Springer Publishing Company, Incorporated, 1st ed., 2009. 2, 11, 12, 15, 130

[5] M. K. Qureshi, S. Gurumurthi, and B. Rajendran, *Phase Change Memory: From Devices to Systems.* Morgan and Claypool Publishers, 1st ed., 2011. 2, 3, 129

[6] H.-S. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010. 4

[7] G. Servalli, "A 45nm generation phase change memory technology," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, 2009. 5, 9, 86

[8] Y. Choi, I. Song, M.-H. Park, H. Chung, S. Chang, B. Cho, J. Kim, Y. Oh, D. Kwon, J. Sunwoo, J. Shin, Y. Rho, C. Lee, M.-G. Kang, J. Lee, Y. Kwon, S. Kim, J. Kim, Y.-J. Lee, Q. Wang, S. Cha, S. Ahn, H. Horii, J. Lee, K. Kim, H. Joo, K. Lee, Y.-T. Lee, J. Yoo, and G. Jeong, "A 20nm 1.8v 8gb pram with 40mb/s program bandwidth," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 46–48, 2012. 5

[9] M. Takata, K. Nakayama, T. Izumi, T. Shinmura, J. Akita, and A. Kitagawa, "Nonvolatile sram based on phase change," in *Non-Volatile Semiconductor Memory Workshop, 2006. IEEE NVSMW 2006. 21st*, pp. 95–96, 2006. 5

[10] D. Takashima, "Overview of ferams: Trends and perspectives," in *Non-Volatile Memory Technology Symposium (NVMTS), 2011 11th Annual*, pp. 1–6, 2011. 6, 129

[11] H. Shiga, D. Takashima, S. Shiratake, K. Hoya, T. Miyakawa, R. Ogiwara, R. Fukuda, R. Takizawa, K. Hatsuda, F. Matsuoka, Y. Nagadomi, D. Hashimoto, H. Nishimura, T. Hioka, S. Doumae, S. Shimizu, M. Kawano, T. Taguchi, Y. Watanabe, S. Fujii, T. Ozaki, H. Kanaya, Y. Kumura, Y. Shimojo, Y. Yamada, Y. Minami, S. Shuto, K. Yamakawa, S. Yamazaki, I. Kunishima, T. Hamamoto, A. Nitayama, and T. Furuyama, "A 1.6gb/s ddr2 128mb chain feram with scalable octal bitline and sensing schemes," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pp. 464–465,465a, 2009. 6, 9

[12] T. Miwa, J. Yamada, H. Koike, H. Toyoshima, K. Amanuma, S. Kobayashi, T. Tatsumi, Y. Maejima, H. Hada, and T. Kunio, "Nv-sram: a nonvolatile sram with backup ferroelectric capacitors," *Solid-State Circuits, IEEE Journal of*, vol. 36, no. 3, pp. 522–527, 2001. 6, 15

[13] Y. Shimojo, A. Konno, J. Nishimura, T. Okada, Y. Yamada, S. Kitazaki, H. Furuhashi, S. Yamazaki, K. Yahashi, K. Tomioka, Y. Minami, H. Kanaya, S. Shuto, K. Yamakawa, T. Ozaki, H. Shiga, T. Miyakawa, S. Shiratake, D. Takashima, I. Kunishima, T. Hamamoto, and A. Nitayama, "High-density and high-speed 128mb chain feram; with sdram-compatible ddr2 interface," in *VLSI Technology, 2009 Symposium on*, pp. 218–219, 2009. 6, 129

[14] J.-G. Zhu, "Magnetoresistive random access memory: The path to competitiveness and scalability," *Proceedings of the IEEE*, vol. 96, no. 11, pp. 1786–1798, 2008. 7

[15] T. Kawahara, K. Ito, R. Takemura, and H. Ohno, "Spin-transfer torque ram technology: Review and prospect," *Microelectronics Reliability*, vol. 52, no. 4, pp. 613 – 627, 2012. Advances in non-volatile memory technology. 7, 129, 130

[16] K. L. Wang, J. G. Alzate, and P. K. Amiri, "Low-power non-volatile spintronic memory: Stt-ram and beyond," *Journal of Physics D: Applied Physics*, vol. 46, no. 7, p. 074003, 2013. 7, 130

[17] S. Chung, K.-M. Rho, S.-D. Kim, H.-J. Suh, D.-J. Kim, H. Kim, S. Lee, J.-H. Park, H.-M. Hwang, S.-M. Hwang, J. Y. Lee, Y.-B. An, J.-U. Yi, Y.-H. Seo, D.-H. Jung, M.-S. Lee, S.-H. Cho, J.-N. Kim, G.-J. Park, G. Jin, A. Driskill-Smith, V. Nikitin, A. Ong, X. Tang, Y. Kim, J.-S. Rho, S.-K. Park, S.-W. Chung, J.-G. Jeong, and S. J. Hong, "Fully integrated 54nm stt-ram with the smallest bit cell dimension for high density memory application," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 12.7.1–12.7.4, 2010. 7, 9

[18] Y. Guillemenet, L. Torres, and G. Sassatelli, "Non-volatile run-time field-programmable gate arrays structures using thermally assisted switching magnetic random access memories," *Computers Digital Techniques, IET*, vol. 4, no. 3, pp. 211–226, 2010. 7, 24, 107

[19] W. Zhao, E. Belhaire, and C. Chappert, "Spin-mtj based non-volatile flip-flop," in *Nanotechnology, 2007. IEEE-NANO 2007. 7th IEEE Conference on*, pp. 399–402, 2007. 7

[20] H.-S. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. Chen, and M.-J. Tsai, "Metal oxide rram," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012. 8, 106, 130, 139

[21] Y. S. Chen, H. Lee, P. Chen, P. Gu, C. Chen, W. Lin, W. H. Liu, Y. Y. Hsu, S. S. Sheu, P.-C. Chiang, W.-S. Chen, F. Chen, C. Lien, and M.-J. Tsai, "Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, 2009. 9, 86, 106

[22] B. Govoreanu, G. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. P. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, O. Richard, T. Vandeweyer, K. Seo, P. Hendrickx, G. Pourtois, H. Bender, L. Altimime, D. Wouters, J. Kittl, and M. Jurczak, "10x10nm2 hf/hfox crossbar resistive ram with excellent performance, reliability and low-energy operation," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 31.6.1–31.6.4, 2011. 9, 86, 106

[23] T.-Y. Liu, T. H. Yan, R. Scheuerlein, Y. Chen, J. Lee, G. Balakrishnan, G. Yee, H. Zhang, A. Yap, J. Ouyang, T. Sasaki, S. Addepalli, A. Al-Shamma, C.-Y. Chen,

M. Gupta, G. Hilton, S. Joshi, A. Kathuria, V. Lai, D. Masiwal, M. Matsumoto, A. Nigam, A. Pai, J. Pakhale, C. H. Siau, X. Wu, R. Yin, L. Peng, J. Y. Kang, S. Huynh, H. Wang, N. Nagel, Y. Tanaka, M. Higashitani, T. Minvielle, C. Gorla, T. Tsukamoto, T. Yamaguchi, M. Okajima, T. Okamura, S. Takase, T. Hara, H. Inoue, L. Fasoli, M. Mofidi, R. Shrivastava, and K. Quader, "A 130.7mm2 2-layer 32gb reram memory device in 24nm technology," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, pp. 210–211, 2013. 9, 106

[24] M. Kozicki, C. Gopalan, M. Balakrishnan, and M. Mitkova, "A low-power non-volatile switching element based on copper-tungsten oxide solid electrolyte," *Nanotechnology, IEEE Transactions on*, vol. 5, no. 5, pp. 535–544, 2006. 9

[25] W. Wang, A. Gibby, Z. Wang, T. W. Chen, S. Fujita, P. Griffin, Y. Nishi, and S. Wong, "Nonvolatile sram cell," in *Electron Devices Meeting, 2006. IEDM '06. International*, pp. 1–4, 2006. 9

[26] C. Chevallier, C. H. Siau, S. Lim, S. Namala, M. Matsuoka, B. Bateman, and D. Rinerson, "A 0.13 µm 64mb multi-layered conductive metal-oxide memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pp. 260–261, 2010. 9

[27] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, K. Tanabe, T. Nakamura, Y. Sumimoto, N. Yamada, N. Nakai, S. Sakamoto, Y. Hayakawa, K. Tsuji, S. Yoneda, A. Himeno, K. Origasa, K. Shimakawa, T. Takagi, T. Mikawa, and K. Aono, "An 8mb multi-layered cross-point reram macro with 443mb/s write throughput," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 432–434, 2012. 9

[28] S. Kaeriyama, T. Sakamoto, H. Sunamura, M. Mizuno, H. Kawaura, T. Hasegawa, K. Terabe, T. Nakayama, and M. Aono, "A nonvolatile programmable solid-electrolyte nanometer switch," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 1, pp. 168–176, 2005. 10, 25, 26, 131

[29] S. Dietrich, M. Angerbauer, M. Ivanov, D. Gogl, H. Hoenigschmid, M. Kund, C. Liaw, M. Markert, R. Symanczyk, L. Altimime, S. Bournat, and G. Mueller,

"A nonvolatile 2-mbit cbram memory core featuring advanced read and program control," *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 4, pp. 839–845, 2007. 10

[30] K. Aratani, K. Ohba, T. Mizuguchi, S. Yasuda, T. Shiimoto, T. Tsushima, T. Sone, K. Endo, A. Kouchiyama, S. Sasaki, A. Maesaka, N. Yamada, and H. Narisawa, "A novel resistance memory with high scalability and nanosecond switching," in *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 783–786, 2007. 10, 86

[31] M.-F. Chang and P. cheng Chen, "Embedded non-volatile memory circuit design technologies for mobile low-voltage soc and 3d-ic," in *Solid-State and Integrated Circuit Technology (ICSICT), 2010 10th IEEE International Conference on*, pp. 13–16, 2010. 11, 15

[32] A. Makosiej, O. Thomas, A. Amara, and A. Vladimirescu, "Cmos sram scaling limits under optimum stability constraints," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pp. 1460–1463, 2013. 13, 14, 130

[33] P.-F. Chiu, M.-F. Chang, C.-W. Wu, C.-H. Chuang, S.-S. Sheu, Y.-S. Chen, and M.-J. Tsai, "Low store energy, low vddmin, 8t2r nonvolatile latch and sram with vertical-stacked resistive memory (memristor) devices for low power mobile applications," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 6, pp. 1483–1496, 2012. 15, 16, 99, 107, 130, 139

[34] M. Fliesler, D. Still, and J.-M. Hwang, "A 15ns 4mb nvsram in 0.13u sonos technology," in *Non-Volatile Semiconductor Memory Workshop, 2008 and 2008 International Conference on Memory Technology and Design. NVSMW/ICMTD 2008. Joint*, pp. 83–86, 2008. 15

[35] N. Sakimura, T. Sugibayashi, R. Nebashi, and N. Kasai, "Nonvolatile magnetic flip-flop for standby-power-free socs," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, pp. 355–358, 2008. 15

[36] H. Hassan and M. Anis, *Low-Power Design of Nanometer FPGAs: Architecture and EDA*. Burlington, MA: Morgan Kaufmann, 2010. 16, 19, 20

[37] A. Corporation in *Using FPGA-Based Parallel Flash Loader with the Quartus II Software*, 2007. 18

[38] S. Tapp, "Bpi fast configuration and impact flash programming with 7 series fpgas," 2013. 19

[39] T. Tuan and B. Lai, "Leakage power analysis of a 90nm fpga," in *Custom Integrated Circuits Conference, 2003. Proceedings of the IEEE 2003*, pp. 57–60, 2003. 20

[40] Y. Chen, J. Zhao, and Y. Xie, "3d-nonfar: Three-dimensional non-volatile fpga architecture using phase change memory," in *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*, pp. 55–60, 2010. 21

[41] Lattice, "ice40 lp series ultra low-power mobile fpga family," 2012. 22

[42] L. Hong, "Comparison of embedded nonvolatile memory technologies and their applications," 2009. 22

[43] Microsemi, "Igloo nano low power flash fpgas with flash freeze technology," 2013. 22

[44] Lattice, "Latticexp2family: Instant-on, secure, single-chip fpga with complete development platform," 2012. 23

[45] Eureka, "Nand flash faq," 2012. 23

[46] L. Torres, R. M. Brum, L. V. Cargnini, and G. Sassatelli, "Trends on the application of emerging nonvolatile memory to processors and programmable devices," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pp. 101–104, 2013. 23, 130

[47] S. Onkaraiah, O. Turkylmaz, M. Reyboz, F. Clermidy, J. Portal, and C. Muller, "An hybrid cbram/cmos look-up-table structure for improving performance efficiency of field programmable gate array," in *To be published in Proceedings of Symposium on Circuits and Systems (ISCAS), 2013 IEEE International*, 2013. 24, 25, 87, 107

[48] M. Lin, A. El Gamal, Y.-C. Lu, and S. Wong, "Performance benefits of monolithically stacked 3-d fpga," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 2, pp. 216–229, 2007. 24

[49] Y. Y. Liauw, Z. Zhang, W. Kim, A. Gamal, and S. Wong, "Nonvolatile 3d-fpga with monolithically stacked rram-based configuration memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, pp. 406–408, 2012. 24, 25, 26, 86, 99, 108

[50] K. Ikegami, K. Abe, K. Nomura, S. Yasuda, M. Oda, and S. Fujita, "Designing nonvolatile reconfigurable switch-based fpga through overall circuit performance evaluation," in *Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW), 2012 IEEE 26th International*, pp. 213–220, 2012. 24

[51] M. Miyamura, S. Nakaya, M. Tada, T. Sakamoto, K. Okamoto, N. Banno, S. Ishida, K. Ito, H. Hada, N. Sakimura, T. Sugibayashi, and M. Motomura, "Programmable cell array using rewritable solid-electrolyte switch integrated in 90nm cmos," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pp. 228–229, 2011. 25

[52] S. Tanachutiwat, M. Liu, and W. Wang, "Fpga based on integration of cmos and rram," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 11, pp. 2023–2032, 2011. 25

[53] J. Liang and H.-S. Wong, "Cross-point memory array without cell selectors device characteristics and data storage pattern dependencies," *Electron Devices, IEEE Transactions on*, vol. 57, no. 10, pp. 2531–2538, 2010. 25

[54] H.-Y. Chen, S. Yu, B. Gao, P. Huang, J. Kang, and H.-S. Wong, "Hfox based vertical resistive random access memory for cost-effective 3d cross-point architecture without cell selector," in *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 20.7.1–20.7.4, 2012. 25

[55] S. Yasuda, K. Ikegami, T. Tanamoto, A. Kinoshita, K. Abe, and S. Fujita, "Nonvolatile configuration memory cell for low power field programmable gate array," in *Memory Workshop (IMW), 2011 3rd IEEE International*, pp. 1–4, 2011. 25, 87

[56] M. Tada, N. Inoue, and Y. Hayashi, "Performance modeling of low- k /cu interconnects for 32-nm-node and beyond," *Electron Devices, IEEE Transactions on*, vol. 56, no. 9, pp. 1852–1861, 2009. 26

[57] M. Yamaoka, Y. Shinozaki, N. Maeda, Y. Shimazaki, K. Kato, S. Shimada, K. Yanagisawa, and K. Osada, "A 300-mhz 25- mu;a/mb-leakage on-chip sram module featuring process-variation immunity and low-leakage-active mode for mobile-phone application processor," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 1, pp. 186–194, 2005. 26, 86, 87, 109

[58] K. Nii, Y. Tsukamoto, T. Yoshizawa, S. Imaoka, Y. Yamagami, T. Suzuki, A. Shibayama, H. Makino, and S. Iwade, "A 90-nm low-power 32-kb embedded sram with gate leakage suppression circuit for mobile applications," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 4, pp. 684–693, 2004. 26, 86, 87, 109

[59] A. Saarinen, M.-L. Linne, and O. Yli-Harja, "Modeling single neuron behavior using stochastic differential equations," *Neurocomput.*, vol. 69, pp. 1091–1096, June 2006. 28

[60] D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, A. J. Sherbondy, and R. Singh, "Cognitive computing," *Commun. ACM*, vol. 54, pp. 62–71, Aug. 2011. 28, 109

[61] G. Snider, R. Amerson, D. Carter, H. Abdalla, M. Qureshi, J. Leveille, M. Versace, H. Ames, S. Patrick, B. Chandler, A. Gorchetchnikov, and E. Mingolla, "From synapses to circuitry: Using memristive memory to explore the electronic brain," *Computer*, vol. 44, no. 2, pp. 21–28, 2011. 28, 109

[62] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. Modha, "A digital neurosynaptic core using embedded crossbar memory with 45pj per spike in 45nm," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*, pp. 1–4, 2011. 28, 109

[63] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," in

Electron Devices Meeting (IEDM), 2011 IEEE International, pp. 4.4.1–4.4.4, 2011. 28, 109

[64] M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, "Spin-based neuron model with domain-wall magnets as synapse," *Nanotechnology, IEEE Transactions on*, vol. 11, no. 4, pp. 843–853, 2012. 28, 109

[65] D. F. Specht, "Probabilistic neural networks for classification, mapping, or associative memory," in *Neural Networks, 1988., IEEE International Conference on*, pp. 525–532 vol.1, 1988. 28, 109

[66] J. Alspector, B. Gupta, and R. B. Allen, "Artificial neural networks," ch. Performance of a stochastic learning microchip, pp. 66–78, Piscataway, NJ, USA: IEEE Press, 1990. 29, 90, 109

[67] M. van Daalen, P. Jeavons, and J. Shawe-Taylor, "A stochastic neural architecture that exploits dynamically reconfigurable fpgas," in *FPGAs for Custom Computing Machines, 1993. Proceedings. IEEE Workshop on*, pp. 202–211, 1993. 29, 109

[68] S. Fusi, P. J. Drew, and L. F. Abbott, "Cascade Models of Synaptically Stored Memories," *Neuron*, vol. 45, pp. 599–611, 2005. 29, 109

[69] R. Salakhutdinov and G. Hinton, "An efficient learning procedure for deep boltzmann machines," *Neural Comput.*, vol. 24, pp. 1967–2006, Aug. 2012. 29

[70] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-based resistive switching memories ñ nanoionic mechanisms, prospects, and challenges," *Advanced Materials*, vol. 21, no. 25-26, pp. 2632–2663, 2009. 33, 41, 58

[71] C. Schindler, G. Staikov, and R. Waser, "Electrode kinetics of cu–sio[sub 2]-based resistive switching cells: Overcoming the voltage-time dilemma of electrochemical metallization memories," *Applied Physics Letters*, vol. 94, no. 7, p. 072109, 2009. 33

[72] S. Yu and H.-S. Wong, "Compact modeling of conducting-bridge random-access memory (cbram)," *Electron Devices, IEEE Transactions on*, vol. 58, no. 5, pp. 1352–1360, 2011. 33, 34, 66, 111

[73] U. Russo, D. Kamalanathan, D. Ielmini, A. Lacaita, and M. Kozicki, "Study of multilevel programming in programmable metallization cell (pmc) memory," *Electron Devices, IEEE Transactions on*, vol. 56, no. 5, pp. 1040–1047, 2009. 33, 34, 36, 39

[74] J. R. Jameson, N. Gilbert, F. Koushan, J. Saenz, J. Wang, S. Hollmer, and M. N. Kozicki, "One-dimensional model of the programming kinetics of conductive-bridge memory cells," *Applied Physics Letters*, vol. 99, no. 6, p. 063506, 2011. 33, 39, 59

[75] M. Mitkova and M. Kozicki, "Silver incorporation in gese glasses used in programmable metallization cell devices," *Journal of Non-Crystalline Solids*, vol. 299-302, Part 2, no. 0, pp. 1023 – 1027, 2002. 19th International Conference on Amorphous and Microcrystalline Semiconductors. 33, 131

[76] A. Bid, A. Bora, and A. K. Raychaudhuri, "Temperature dependence of the resistance of metallic nanowires (diameter $\geq$ 15 nm): Applicability of bloch-gruneisen theorem," *Phys. Rev. B*, vol. 74, p. 035426, Jul 2006. 35

[77] S. Z. Rahaman, S. Maikap, W. S. Chen, H. Y. Lee, F. T. Chen, T. C. Tien, and M. J. Tsai, "Impact of tao nanolayer at the gese-w interface on resistive switching memory performance and investigation of cu nanofilament," *Journal of Applied Physics*, vol. 111, no. 6, p. 063710, 2012. 35, 60

[78] S. Choi, S. Ambrogio, S. Balatti, F. Nardi, and D. Ielmini, "Resistance drift model for conductive-bridge (cb) ram by filament surface relaxation," in *Memory Workshop (IMW), 2012 4th IEEE International*, pp. 1–4, 2012. 35

[79] C. Pi, Y. Ren, and W. K. Chim, "Investigation of bipolar resistive switching and the time-dependent set process in silver sulfide/silver thin films and nanowire array structures," *Nanotechnology*, vol. 21, no. 8, p. 085709, 2010. 35, 111, 132, 140

[80] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories: fundamentals, applications, prospects," *Nanotechnology*, vol. 22, no. 25, p. 254003, 2011. 37, 39, 41

[81] C. Schindler, M. Meier, R. Waser, and M. Kozicki, "Resistive switching in age-se with extremely low write currents," in *Non-Volatile Memory Technology Symposium, 2007. NVMTS '07*, pp. 82–85, 2007. 37, 60

[82] M. Reyboz, S. Onkaraiah, G. Palma, and E. Vianello, "Physical compact model of a cbram cell," in *MOS AK Workshop IEEE*, 2012. 40, 53, 84

[83] D. Ielmini, "Modeling the universal set/reset characteristics of bipolar rram by field- and temperature-driven filament growth," *Electron Devices, IEEE Transactions on*, vol. 58, no. 12, pp. 4309–4317, 2011. 41

[84] C. Gopalan, Y. Ma, T. Gallo, J. Wang, E. Runnion, J. Saenz, F. Koushan, P. Blanchard, and S. Hollmer, "Demonstration of conductive bridging random access memory (cbram) in logic cmos process," *Solid-State Electronics*, vol. 58, no. 1, pp. 54 – 61, 2011. Special Issue devoted to the 2nd International Memory Workshop (IMW 2010). 41

[85] K. Okamoto, M. Tada, T. Sakamoto, M. Miyamura, N. Banno, N. Iguchi, and H. Hada, "Conducting mechanism of atom switch with polymer solid-electrolyte," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 12.2.1–12.2.4, 2011. 41, 43

[86] N. Banno, T. Sakamoto, N. Iguchi, H. Sunamura, K. Terabe, T. Hasegawa, and M. Aono, "Diffusivity of cu ions in solid electrolyte and its effect on the performance of nanometer-scale switch," *Electron Devices, IEEE Transactions on*, vol. 55, no. 11, pp. 3283–3287, 2008. 42

[87] D. Ielmini, F. Nardi, and C. Cagli, "Physical models of size-dependent nanofilament formation and rupture in nio resistive switching memories," *Nanotechnology*, vol. 22, no. 25, p. 254022, 2011. 43

[88] C. Cagli, J. Buckley, V. Jousseaume, T. Cabout, A. Salaun, H. Grampeix, J.-F. Nodin, H. Feldis, A. Persico, J. Cluzel, P. Lorenzi, L. Massari, R. Rao, F. Irrera, F. Aussenac, C. Carabasse, M. Coue, P. Calka, E. Martinez, L. Perniola, P. Blaise, Z. Fang, Y. H. Yu, G. Ghibaudo, D. Deleruyelle, M. Bocquet, C. Muller, A. Padovani, O. Pirrotta, L. Vandelli, L. Larcher, G. Reimbold, and B. De Salvo, "Experimental and theoretical study of electrode effects in hfo2 based rram," in

Electron Devices Meeting (IEDM), 2011 IEEE International, pp. 28.7.1–28.7.4, 2011. 43

[89] J. Yi, S.-W. Kim, Y. Nishi, Y.-T. Hwang, S.-W. Chung, S.-J. Hong, and S.-W. Park, "Research on switching property of an oxide/copper sulfide hybrid memory," in *Non-Volatile Memory Technology Symposium, 2008. NVMTS 2008. 9th Annual*, pp. 1–4, 2008. 57, 60

[90] T. Tsuruoka, K. Terabe, T. Hasegawa, and M. Aono, "Forming and switching mechanisms of a cation-migration-based oxide resistive memory," *Nanotechnology*, vol. 21, no. 42, p. 425205, 2010. 58

[91] S. Z. Rahaman and S. Maikap, "Improved resistive switching memory characteristics using novel bi-layered ge0.2se0.8/ta2o5 solid-electrolytes," in *Memory Workshop (IMW), 2010 IEEE International*, pp. 1–4, 2010. 60

[92] R. Soni, M. Meier, A. Rudiger, B. Hollander, C. Kugeler, and R. Waser, "Integration of gexse1-x in crossbar arrays for non-volatile memory applications," *Microelectronic Engineering*, vol. 86, no. 4 - 6, pp. 1054 – 1056, 2009. The 34th International Conference on Micro- and Nano-Engineering (MNE). 60, 62

[93] M. Tada, T. Sakamoto, Y. Tsuji, N. Banno, Y. Saito, Y. Yabe, S. Ishida, M. Terai, S. Kotsuji, N. Iguchi, M. Aono, H. Hada, and N. Kasai, "Highly scalable nonvolatile tiox/tasioy solid-electrolyte crossbar switch integrated in local interconnect for low power reconfigurable logic," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, 2009. 60

[94] L. Vandelli, A. Padovani, L. Larcher, R. Southwick, W. Knowlton, and G. Bersuker, "A physical model of the temperature dependence of the current through sio2-hfo2 stacks," *Electron Devices, IEEE Transactions on*, vol. 58, no. 9, pp. 2878–2887, 2011. 64, 65, 66, 136

[95] L. Larcher, "Statistical simulation of leakage currents in mos and flash memory devices with a new multiphonon trap-assisted tunneling model," *Electron Devices, IEEE Transactions on*, vol. 50, no. 5, pp. 1246–1253, 2003. 64

[96] N. Yang, W. Henson, J. R. Hauser, and J. J. Wortman, "Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in mos devices," *Electron Devices, IEEE Transactions on*, vol. 46, no. 7, pp. 1464–1471, 1999. 64

[97] A. Padovani, L. Larcher, G. Bersuker, and P. Pavan, "Charge transport and degradation in hfo2 and hfox dielectrics," *Electron Device Letters, IEEE*, vol. 34, no. 5, pp. 680–682, 2013. 65, 66, 136

[98] J. Jameson, N. Gilbert, F. Koushan, J. Saenz, J. Wang, S. Hollmer, M. Kozicki, and N. Derhacobian, "Quantized conductance in ag-ges-w conductive-bridge memory cells," *Electron Device Letters, IEEE*, vol. 33, no. 2, pp. 257–259, 2012. 81

[99] T. Sakamoto, N. Banno, N. Iguchi, H. Kawaura, H. Sunamura, S. Fujieda, K. Terabe, T. Hasegawa, and M. Aono, "A ta2o5 solid-electrolyte switch with improved reliability," in *VLSI Technology, 2007 IEEE Symposium on*, pp. 38–39, 2007. 85

[100] M. Wang, W. Luo, Y. Wang, L. Yang, W. Zhu, P. Zhou, J. H. Yang, X. Gong, Y. Lin, R. Huang, S. Song, Q. T. Zhou, H. Wu, J. Wu, and M. H. Chi, "A novel cuxsiyo resistive memory in logic technology with excellent data retention and resistance distribution for embedded applications," in *VLSI Technology (VLSIT), 2010 Symposium on*, pp. 89–90, 2010. 86

[101] W. Guan, S. Long, R. Jia, and M. Liu, "Nonvolatile resistive switching memory utilizing gold nanocrystals embedded in zirconium oxide," *Applied Physics Letters*, vol. 91, no. 6, p. 062111, 2007. 86

[102] W. Kim, S. I. Park, Z. Zhang, Y. Yang-Liauw, D. Sekar, H.-S. Wong, and S. Wong, "Forming-free nitrogen-doped alox rram with sub-µa programming current," in *VLSI Technology (VLSIT), 2011 Symposium on*, pp. 22–23, 2011. 86

[103] S. Z. Rahaman, S. Maikap, C. H. Lin, P. J. Tzeng, H. Lee, T. Y. Wu, Y. S. Chen, F. Chen, M.-J. Kao, and M.-J. Tsai, "Low current bipolar resistive switching memory using cu metallic filament in ge0.2se0.8 solid-electrolyte," in *VLSI Technology Systems and Applications (VLSI-TSA), 2010 International Symposium on*, pp. 134–135, 2010. 86

[104] E. Vianello, G. Molas, F. Longnos, P. Blaise, E. Souchier, C. Cagli, G. Palma, J. Guy, M. Bernard, M. Reyboz, G. Rodriguez, A. Roule, C. Carabasse, V. Delaye, V. Jousseaume, S. Maitrejean, G. Reimbold, B. De Salvo, F. Dahmani, P. Verrier, D. Bretegnier, and J. Liebault, "Sb-doped ges2 as performance and reliability booster in conductive bridge ram," in *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 31.5.1–31.5.4, 2012. 86, 95

[105] Z. Wei, Y. Kanzawa, K. Arita, Y. Katoh, K. Kawai, S. Muraoka, S. Mitani, S. Fujii, K. Katayama, M. Iijima, T. Mikawa, T. Ninomiya, R. Miyanaga, Y. Kawashima, K. Tsuji, A. Himeno, T. Okada, R. Azuma, K. Shimakawa, H. Sugaya, T. Takagi, R. Yasuhara, K. Horiba, H. Kumigashira, and M. Oshima, "Highly reliable taox reram and direct evidence of redox reaction mechanism," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, 2008. 86

[106] W. Chien, Y. R. Chen, Y. Chen, A. Chuang, F. Lee, Y. Lin, E. Lai, Y.-H. Shih, K. Hsieh, and C.-Y. Lu, "A forming-free wox resistive memory using a novel self-aligned field enhancement feature with excellent reliability and scalability," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, pp. 19.2.1–19.2.4, 2010. 86

[107] H. Lee, P. Chen, T. Y. Wu, Y. Chen, C. Wang, P. Tzeng, C. H. Lin, F. Chen, C. Lien, and M.-J. Tsai, "Low power and high speed bipolar switching with a thin reactive ti buffer layer in robust hfo2 based rram," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–4, 2008. 86

[108] I. Baek, C. Park, H. Ju, D. J. Seong, H. S. Ahn, J. Kim, M. K. Yang, S. Song, E. Kim, S. Park, C. Park, C. Song, G. Jeong, S. Choi, H. K. Kang, and C. Chung, "Realization of vertical resistive memory (vrram) using cost effective 3d process," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 31.8.1–31.8.4, 2011. 86

[109] P.-E. Gaillardon, H. Ben Jamaa, G. Beneventi, F. Clermidy, and L. Perniola, "Emerging memory technologies for reconfigurable routing in fpga architecture," in *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, pp. 62–65, 2010. 87

[110] M. Suri, O. Bichler, D. Querlioz, G. Palma, E. Vianello, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Cbram devices as binary synapses for low-power stochastic neuromorphic systems: Auditory (cochlea) and visual (retina) cognitive processing applications," in *Electron Devices Meeting (IEDM), 2012 IEEE International*, pp. 10.3.1–10.3.4, 2012. 88, 89

[111] J. Alspector, J. Gannett, S. Haber, M. Parker, and R. Chu, "A vlsi-efficient technique for generating multiple uncorrelated noise sources and its application to stochastic neural networks," *Circuits and Systems, IEEE Transactions on*, vol. 38, no. 1, pp. 109–123, 1991. 90

[112] K. Cameron, T. Clayton, B. Rae, A. Murray, R. Henderson, and E. Charbon, "Poisson distributed noise generation for spiking neural applications," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 365–368, 2010. 90

[113] T.-J. Chiu, J. Gong, Y.-C. King, C.-C. Lu, and H. Chen, "An octagonal dual-gate transistor with enhanced and adaptable low-frequency noise," *Electron Device Letters, IEEE*, vol. 32, no. 1, pp. 9–11, 2011. 90

[114] T. Oya, T. Asai, and Y. Amemiya, "Stochastic resonance in an ensemble of single-electron neuromorphic devices and its application to competitive neural networks," *Chaos, Solitons and Fractals*, vol. 32, no. 2, pp. 855 – 861, 2007. 90

[115] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. SAœGHI, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, and K. Boahen, "Neuromorphic silicon neuron circuits," *Frontiers in Neuroscience*, vol. 5, no. 73, 2011. 91, 139