RRAMSpec: A Design Space Exploration Framework for High Density Resistive RAM

Deepak M. Mathew[1], André Lucas Chinazzo[1], Christian Weis[1], Matthias Jung[2], Bastien Giraud[3], Pascal Vivet[3], Alexandre Levisse[4], and Norbert Wehn[1]

[1] Technische Universität Kaiserslautern, Germany
{deepak,chinazzo,weis,wehn}@eit.uni-kl.de
[2] Fraunhofer Institute for Experimental Software Engineering (IESE), Germany
matthias.jung@iese.fraunhofer.de
[3] Univ. Grenoble Alpes, CEA-LETI, MINATEC Campus, Grenoble, France
{bastien.giraud,pascal.vivet}@cea.fr
[4] Embedded System Laboratory (ESL), EPFL, Switzerland
alexandre.levisse@epfl.ch

Abstract. Resistive RAM (RRAM) is a promising emerging Non-Volatile Memory candidate due to its scalability and CMOS compatibility, which enables the fabrication of high density RRAM crossbar arrays in Back-End-Of-Line CMOS processes. Fast and accurate architectural models of RRAM crossbar devices are required to perform system level design space explorations of new Storage Class Memory (SCM) architectures using RRAM e.g. Non-Volatile-DIMM-P (NVDIMM-P). The major challenge in architectural modeling is the trade-off between accuracy and computing intensity. In this paper we present RRAMSpec, an architecture design space exploration framework, which enables fast exploration of various architectural trade-offs in designing high density RRAM devices, at accuracy levels close to circuit level simulators. The framework estimates silicon area, timings, and energy for RRAM devices. It outperforms state-of-the-art RRAM modeling tools by conducting architectural explorations at very high accuracy levels within few seconds of execution time. Our evaluations show various trade-offs in designing RRAM crossbar arrays with respect to array sizes, write time and write energy. Finally we present the influence of technology scaling on different RRAM design trade-offs.

Keywords: RRAM · ReRAM · Crossbar · NVM.

1 Introduction

In present day off-chip memory hierarchy, there exists a large gap in bandwidth between main memory (DRAM) and storage memory (NAND Flash/HDD). Some of the new emerging *Non-Volatile Memories* (NVMs), such as *Resistive RAM* (RRAM), *Spin-Transfer Torque Magnetic RAM* (STT-MRAM) and *Phase Change Memory* (PCM), exhibit the potential to bridge this gap since they have the performance and cost per bit in between DRAM and Flash [1–3]. Therefore, in recent years, an additional layer of memory hierarchy called *Storage-Class*

Memory (SCM) [4, 5] is under discussion in order to integrate these emerging NVMs into the existing memory hierarchy.

Metal oxide based RRAM is a promising emerging NVM candidate for SCM due to its properties such as fast switching (~ 100 ns for writes), good scalability, and CMOS compatibility. The integration of RRAM in the CMOS process is done in the *Back-End-Of-Line* (BEOL). Despite the above mentioned advantages, adoption of RRAM as an SCM in embedded and high performance computer architectures is still facing challenges due to variability issues [6–8] and sneak currents [9, 10] in high density crossbar arrays. RRAM crossbar memories will have shorter read and write latencies than flash, with lower leakage and higher density than DRAM, making it an ideal candidate for SCM [1]. Therefore, researchers and industry consortia are exploring novel hybrid memory architectures such as NVDIMM-P [11], where RRAM and DRAM share the main memory address space. In order to conduct early design space explorations for such novel memory architectures using RRAM, fast and accurate architectural models of high density RRAM crossbar memories are required. Such a model has to provide timings, energy, and area of the high density RRAM from low level parameters of the RRAM device at high accuracy and fast execution speed.

A RRAM cell model with a voltage dependent write time, and an array model with an accurate voltage drop analysis are the essential components required for developing such an architectural modeling framework. The existing RRAM modelling frameworks [12–14] are either less accurate due to the approximate array voltage drop analysis or very slow since they depend on SPICE simulations [9]. This paper makes the following new contributions:

1. We present an architectural modeling approach to evaluate timings, energy consumption, and silicon area of high density RRAM crossbar memories.
2. We prove that the conventional modelling approach (assuming constant sneak currents) fails at lower technology nodes due to increased voltage drop in the crossbar array, and due to the non-linearity of the selector.
3. We show the trends in RRAM read/write times and energies with the increase in crossbar array size. Contrary to the popular belief, we show that the write time can decrease with the increase in array size.
4. Finally, we compare the output of our framework with a state-of-the-art NVM modelling framework [12].

RRAMSpec gets a technology input file and an architectural input file. The technology input file includes RRAM cell, selector, and CMOS technology related parameters. The architectural input file contains the required density, the optimization target (fixed, performance, energy) etc.

This paper is organized as follows. RRAM technology, operation, and the sneak current problem with crossbar arrays are discussed in Section 2. In Section 3, we summarize the previous works in RRAM crossbar array modeling. Our modeling approach is detailed in Section 4. The results are discussed in Section 5. Finally, Section 6 concludes this paper.

2 RRAM Background

A basic metal-oxide RRAM cell has a *Metal-Insulator-Metal* (MIM) structure with the insulator layer composed of a binary or ternary transition metal oxide (e.g. HfO2, TaO2, SrTiO3) [15]. The resistance state of the cell, either a *High Resistance State* (HRS) or a *Low Resistance State* (LRS), is used to store logic `0` and logic `1` respectively. When writing a `1` to RRAM, known as SET operation, it switches from HRS to LRS. When writing a `0` to RRAM, known as RESET operation, the device switches from LRS to HRS. For bipolar RRAMs, the switching process (SET or RESET) depends on the polarity of the applied voltage, while for unipolar RRAMs the switching happens irrespectively of the polarity of the applied voltage. In this paper, we focus mainly on bipolar metal-oxide RRAMs, although it is easily extendable for unipolar RRAMs. Further details of the switching mechanism of RRAMs is explained in [15].

The simple MIM structure of RRAM device permits building high density crossbar arrays with minimum cell size ($4F^2$), where F is the minimum feature size of the technology node, which corresponds to half of the minimum metal pitch. Figure 1 shows the schematic of an $m \times n$ crossbar array with m *Wordlines*(WLs) and n *Bitlines*(BLs). RRAM cells are placed at the intersection of each WL and BL. *Analog Multiplexer*s (AMUXes), which are connected to the edge of WLs and BLs connect the selected lines to a voltage V_{SEL} and unselected lines to the voltage V_{USEL}.

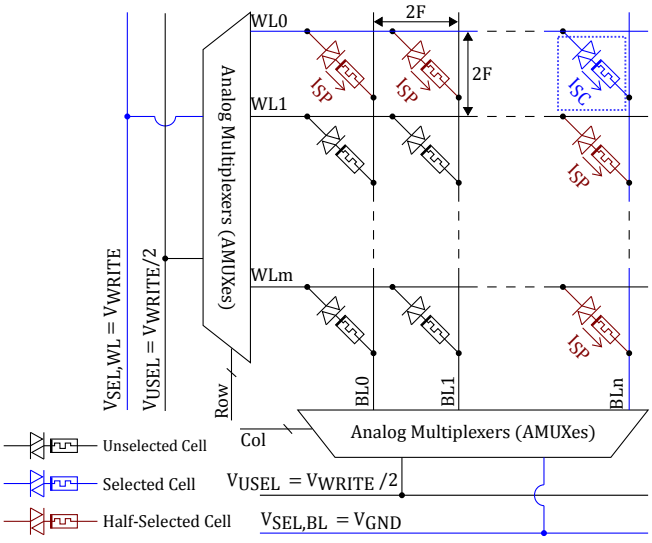


Fig. 1: Basic Biasing Scheme to Perform a SET Operation.

Figure 1 shows the array voltages for performing a write (SET in this case) operation on the cell farthest from AMUXes. A voltage V_{WRITE} is applied to the

selected WL. The selected BL is grounded. This will ensure that the full write voltage is applied to the selected cell. All unselected WLs and BLs are supplied by a voltage $V_{WRITE}/2$ such that the effective voltage across the unselected cells is zero, preventing any unwanted current flowing through these cells. This biasing scheme, known as *Half-Bias Scheme* [10], is commonly used for biasing RRAM crossbar arrays. All cells, which share the WL or BL with the selected cell, known as *Half-Selected Cells*(HSCs), experience sneak path currents (I_{SP}) due to the voltage $V_{WRITE}/2$ across them. These sneak currents cause additional voltage degradation in the selected WL/BL, thereby reducing the effective voltage across the selected cell. Since the time for RRAM SET or RESET operation exponentially increases with the decrease in write voltage [9, 16], the voltage degradation due to sneak currents largely increases the write time, and may even cause write failures.

To mitigate the sneak current problem, a selector device (eg. a bipolar diode) is integrated to the RRAM cell as shown in Figure 1. The selector device has a non-linear switching characteristic similar to a diode, resulting the device significantly reducing current flow at low voltages. This limits the sneak currents flowing through HSCs. However, even using selectors with a non-linearity factor of ~ 1000, the total sneak current in arrays larger than 1 Mb may be comparable to the selected cell current. Therefore, there exist various sneak current compensation schemes [17–19], which increase the reliability of the read and SET operations by externally emulating the sneak currents. But, these schemes do not reduce the effect of the voltage drop across the crossbar array.

There are two main components of the voltage drop: the drop across the AMUXes, and the drop across the metal lines. Due to the relatively high voltages needed for the resistive state switching, thicker oxide transistors, which have higher resistance for the same area, are necessary for designing AMUXes in advanced technology nodes (≤ 65 nm). Furthermore, as the metal wire width shrinks with the technology scaling, its sheet resistance increases. Both effects result in higher voltage drop in the crossbar array.

3 Related Work

In this section, we briefly describe state-of-the-art RRAM/NVM modeling frameworks, their drawbacks, and the advantages of RRAMSpec in comparison to the existing modelling frameworks.

NVsim [13] is the first architectural exploration framework for NVMs, which models a variety of NVMs including RRAM crossbar memories. Later, Poremba et.al. presented an advanced version of this framework called DESTINY[12], which permits 3D-modelling of NVMs. Both NVSim and DESTINY have two major drawbacks. Firstly, they assume a constant I_{SP} across all HSCs in the crossbar array, which does not hold for RRAM crossbar arrays using selector devices, especially in advanced technology nodes. In Section 5, we show that the error using this approach can go even higher than 100% for large arrays in advanced technology nodes. Secondly, they do not consider the voltage drop on the selected WL and BL of the crossbar array, which has a large impact

on RRAM cell write time as explained in Section 2. Instead, they use a cell write voltage and write time provided by the designer as an input parameter. In contrast, RRAMSpec calculates the voltages and sneak currents at each HSC. Its RRAM cell model calculates the write time based on the effective write voltage on the selected cell considering the crossbar array size. Recently, Levisse et.al. [14] proposed a methodology calculating analytically the voltage drop evolution across the array (assuming constant I_{SP} across all HSCs) while not considering the effect on the programming time. Another RRAM modeling framework proposed by [9] invokes HSPICE to simulate the complete crossbar array. Therefore, it is not suitable for fast design space explorations with multiple array sizes due to its long simulation time.

None of these frameworks consider the fact that periphery circuitries can be partially placed below the crossbar array in high density RRAM crossbar memories [20]. In RRAMSpec, we consider the placement of the periphery below the crossbar array, and model the influence of crossbar array scaling on the voltage drop across the AMUXes and the crossbar array metal lines.

4 Modelling of RRAM

In this section, we present the various models used in the RRAMSpec framework.

4.1 RRAM Cell and Selector

The RRAM cell is modelled as two resistance states: HRS and LRS, provided by the designer. The cell switching time for an applied write voltage on the selected cell (V_{SC}) is calculated based on the following equation [9, 16]:

$$t_{SET/RESET} = C_{S/R} \cdot e^{-K_{S/R} \cdot V_{SC,S/R}} \tag{1}$$

The parameters $C_{S/R}$ and $K_{S/R}$ are constants which depends on RRAM cell properties. The selector is modelled as a look-up table with the voltages and respective currents (see Table 1) extracted from measured data of a state-of-the-art selector device [21]. The designer can modify the input file to add another selector.

4.2 Crossbar Array

In the center of our modelling approach is the crossbar array model, which accurately calculates the currents and voltages at each node inside the crossbar array. In a crossbar array as shown in Figure 1, there is no current flowing through unselected cells due to the half-bias scheme. Therefore, we can neglect the voltage drop on the unselected cells, and use a reduced array as depicted in Figure 2 for performing steady state analysis. This reduced model largely improves the simulation time without affecting the accuracy[1]. In Figure 2, R_{AMUX} is the

[1] The number of components in simulation are reduced from $\mathcal{O}(\#WLs \times \#BLs)$ to $\mathcal{O}(\#WLs + \#BLs)$)

driver resistance of the AMUX. In state-of-the art RRAM chips [20] drivers are placed underneath the crossbar array. Therefore, larger array sizes provide more space to fit the drivers below the array. This allows larger driver transistors, and lowers R_{AMUX}. V_{ED} in Figure 2 is the maximum voltage applied to AMUXes. This voltage is limited by the breakdown voltage of the transistors in AMUX. Therefore, it is a CMOS technology related parameter provided by the designer. The existing modelling approaches [12, 13] assume a constant sneak path current along the HSCs. We calculate the sneak path current (I_{SP}) at each HSC using an iterative approach, since the error in array voltage drop calculation using constant sneakpath current approach is very high in lower technology nodes due to the increased WL/BL resistance. In Section 5, we perform a quantitative analysis of the error in array voltage drop calculation using constant sneakpath current approach in comparison to our new approach and SPICE simulations using the reduced array model.

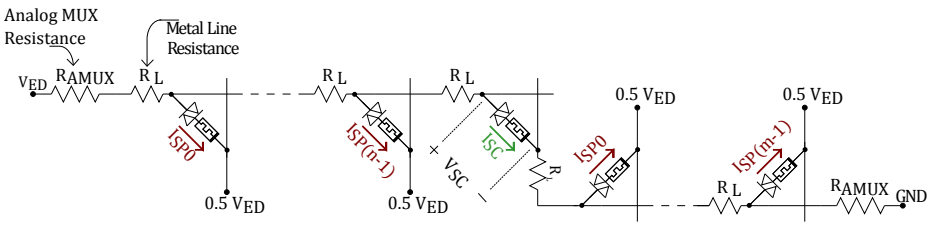


Fig. 2: Reduced Array Model for Steady State Analysis of an $m \times n$ array.

Assuming V_{ED} at the AMUX for a write operation, RRAMSpec iterates inwards calculating the voltages and currents at each HSC. Finally, the voltage V_{SC} and current I_{SC} on the worst-case cell in the array (the cell farthest from WL and BL AMUXes) is calculated via multiple iterations. The cell-switching time is computed with Equation 1 using the calculated V_{SC}. The sum of cell switching time and the RC delay of the WL/BL is used to calculate the internal write time of the crossbar array. This process is repeated for different array sizes. Our modelling approach also checks that the voltage on the cells near to AMUX is not causing unwanted resistance switching due to write disturbance while writing to the farthest cell.

Reading data from the selected RRAM cell is performed by applying a voltage (V_{READ}) on the AMUXes and sensing the current flow. We assume the *Primary Sense Amplifiers* (PSAs) to be placed underneath the array along with AMUXes. For a reliable read operation, it is important that the ratio of the cell currents in LRS and HRS is high enough for the sense-amplifier to sense. Therefore, to find the optimal read voltage, V_{READ}, the designer provides a *Design Current Ratio*, $I_{RATIO} = I_{LRS}/I_{HRS}$ according to the sense-amplifier specifications. RRAMSpec performs a binary search between 0 V and the previously defined V_{WRITE} to find the minimum voltage that provides the required I_{RATIO}. This voltage is selected as the V_{READ}. RRAMSpec also ensures that the selected V_{READ} is low enough to

avoid any disturbance on the resistance state of the read cell (read-disturbance). Reading an RRAM cell in a crossbar array typically involves an additional sneak current estimation step before the actual read process itself [17–19]. Therefore, the internal read time is calculated by summing up the individual delays: the sneak current estimation delay, the sensing delay, and the RC delay of WL/BL.

4.3 RRAM Architecture

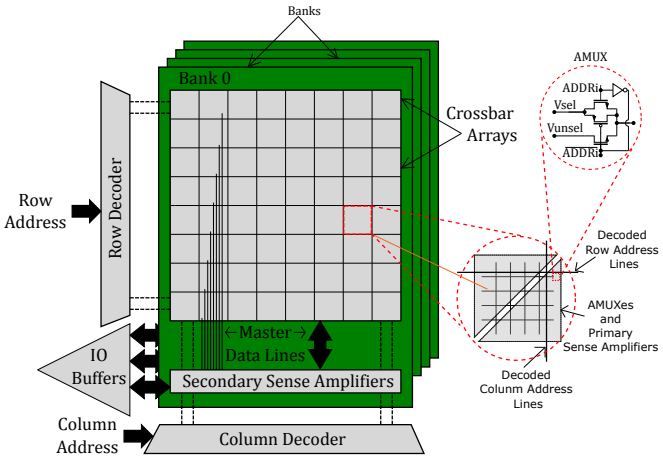


Fig. 3: Architecture of the RRAM.

Figure 3 shows the internal architecture of a complete RRAM memory modelled in RRAMSpec. It consists of several banks. Each bank has its own row/column decoders and secondary sense-amplifiers. The global circuitries such as data IO lines and command lines are shared between multiple banks. A bank is organized as a 2-D matrix of crossbar arrays. AMUXes and PSAs are placed under the crossbar array. Row and column addresses are decoded at the edge of each bank, similar to the decoding scheme in DRAMs. A single bit is accessed from each crossbar array during a read or write operation. The number of activated crossbar arrays at each access depends on the data bus width and the prefetch size. For a x4 chip with a prefetch size of 8, 32 crossbar arrays are activated in parallel, and the 32 bits are transferred to/from secondary sense amplifier. Various decoding delays, command delays and data transfer delays are added to the internal write/read times calculated in Section 4.2 to estimate the total write/read times of the RRAM chip.

4.4 Area Model

In a memory chip, the peripheral circuitries such as row and column decoders, sense amplifiers etc. occupy a considerable fraction of the total silicon area. But,

the BEOL integration of RRAM enables fabrication of crossbar arrays' internal control circuitries (AMUXes and PSAs) underneath the memory array itself [20]. This improves the area efficiency (bits/area), but limits the total area of the AMUXes and the control circuitries to the area of the crossbar array itself.

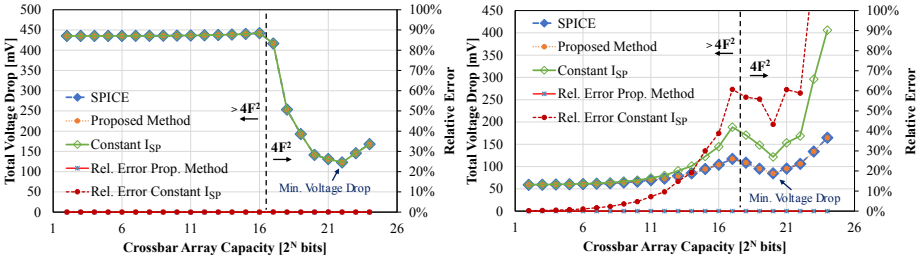
The schematic of an AMUX modelled in our framework is shown in Figure 3. Our area model calculates the maximum width of each transistor in the AMUX such that the complete periphery fits under the memory array, i.e. the total area of AMUXes should not exceed the area of the crossbar array. The resistance of AMUX transistors is then calculated using their width. For square arrays, the number of AMUXes increases linearly with the array size, i.e. number of rows and columns, while the crossbar array area increases quadratically. This provides more space for placing the AMUXes underneath the crossbar array, permitting to increase the width of driver transistors in AMUXes, thereby decreasing their resistances. However, for very small arrays, due to both the constant area occupied by the control circuitry and the minimal width of each transistor, the set of AMUXes might not fit underneath the memory array. In those scenarios, the area occupied by AMUXes and PSAs is used for the total chip area calculation. In addition to the crossbar array area, the total chip area calculation also includes area occupied row/column decoders, secondary sense amplifiers, I/O drivers etc.

4.5 Energy Model

RRAMSpec calculates the operational energy for reads and writes, and the leakage energy when the RRAM crossbar memory is in the idle state. The following sources of operational energy are accounted: crossbar array, global circuitries, and global interconnects. Inside the crossbar array, static current flow during reads and writes (both I_{SC} and I_{SP}), and the capacitive charging currents of metal lines during voltage transitions ($V_{WRITE}/2$ to V_{WRITE}, $V_{WRITE}/2$ to V_{GND}, and $V_{WRITE}/2$ to V_{READ}) are considered. Among these two sources, the power originated from high static currents ($\sim 50 - 150\,\mu$A) is dominant. Energy model uses the static currents which are already calculated by the crossbar array model described in Subsection 4.2. The energy due to capacitive charging of metal lines during read or write operation is calculated using the following equation.

$$E = C \cdot (V_1^2 - V_{HALF} \cdot V_1) \tag{2}$$

In this equation, C represents the line capacitance, and V_1 indicates the operational voltage, which is either V_{WRITE} or V_{READ}. V_{HALF} is the half-bias voltage, which is fixed to $V_{WRITE}/2$. In global circuitries and interconnects, majority of the energy is consumed during voltage transitions due to the capacitances of metal lines. Global wires are usually wider and thicker than the local wires. Moreover, they are also much longer, resulting in an appreciable amount of energy spent in each transition. The idle state leakage currents are negligible inside crossbar arrays and their periphery circuitries due to the half-bias scheme. Therefore, the major sources of leakage energy are the voltage level translators [22] used in global address decoder circuitries. RRAMSpec calculates the leakage in those voltage level translators based on the number of address lines of the device.



(a) 65 nm technology (b) 28 nm technology

Fig. 4: Comparison of the Accuracy of Proposed Method with SPICE and Constant Sneakpath Approach for Different Crossbar Array Sizes (e.g. $2^{20} = 1K$ x $1K$, $2^{21} = 1K$ x $2K$, $2^{22} = 2K$ x $2K$, and so on).

5 Results and Discussion

In this section, we first compare the results of the array voltage drop calculations using our new method to the constant sneak path current (I_{SP}) approach and to SPICE simulations. Our SPICE simulations are based on the RRAM memory device and selector models from [16] and [21], respectively. Table 1 lists the CMOS and RRAM technology parameters used for our analysis. Figure 4 depicts the comparison results of voltage drop calculations using different methods for crossbar arrays at 65 nm and 28 nm CMOS technology nodes. The periphery is designed using thick oxide transistors that support high voltages (5 V for the 65 nm and 3.3 V for the 28 nm) needed for performing writes. The total voltage drop in Figure 4 is the sum of voltage drops in crossbar array metal lines and the AMUX while performing a SET operation. We consider a single bit access per crossbar array. For crossbar arrays at 65 nm (Figure 4a), both, the total voltage drop calculated using our approach and the drop calculated using constant I_{SP} approach matches with the voltage drop calculation using SPICE simulations. Therefore, the relative error in our modelling approach and in the constant I_{SP} approach compared to SPICE simulations is very low. But, for crossbar arrays at 28 nm (Figure 4b), the total voltage drop calculated using the constant I_{SP} approach deviates much from the voltage drop calculated using SPICE. The relative error is very high (>100%) for high density crossbar arrays. The huge error in voltage drop calculation using the constant I_{SP} approach is due to the increased metal line resistance at 28 nm compared to 65 nm. The high metal line resistance ($R_{\square}$ in Table 1) results in a large difference in the voltages at the first cell and the last cell of the selected WL and BL, resulting in large difference in the sneak currents. The constant I_{SP} approach fails here because it assumes the same sneak current across all HSCs. On the other hand, as demonstrated in Figure 4, our approach calculates the array voltage drop at negligible relative error in comparison with SPICE even for larger arrays. This is because it calculates the voltages and currents at each node in an iterative way as explained in Section 4.2.

This clearly shows a major drawback of the existing modelling approach in state-of-the-art RRAM modelling frameworks [12, 13]. Besides that, our new methodology provides a very good speed versus accuracy trade-off. On an Intel Xenon CPU (X5680) it performs the exploration for 2 × 2 to 8K × 8K array sizes in less than two seconds, while SPICE takes around four minutes to complete using the reduced array.

Table 1: CMOS and RRAM Technology Parameters.

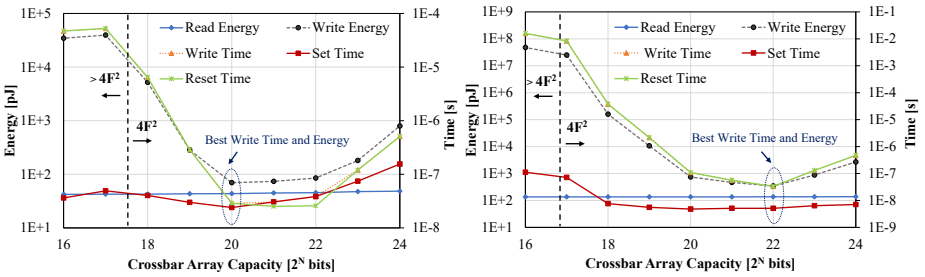
	65 nm	28 nm
R_{HRS} [KΩ]	30	300
R_{LRS} [KΩ]	4	7
I_{CC}[2] [μA]	150	50
V_{PP} [V]	5	3.3
I_{SP} [nA]	69	21
$R_{\square}$[3] [mΩ]	150	450

Selector I-V Table						
V [V]	0	1.5	1.7	1.9	2.1	2.3
$I_{65\,nm}$ [A]	0	462 p	523 p	585 p	647 p	708 p
$I_{28\,nm}$ [A]	0	24 n	538 n	8.1 μ	63 μ	205 μ
V [V]	2.5	2.55	2.6	2.65	2.7	2.75
$I_{65\,nm}$ [A]	4.2 n	699 n	26 μ	121 μ	240 μ	368 μ
$I_{28\,nm}$ [A]	406 μ	462 μ	518 μ	576 μ	635 μ	695 μ

Another interesting trend in Figure 4 is the decrease in total voltage drop with the increase in crossbar array capacity, especially for 65 nm technology, which is explained as follows. The AMUXes and other periphery circuitry are constructed underneath the crossbar array [20]. For any crossbar array capacity, there is a lower limit on the minimum area occupied by the complete periphery due to the required minimum dimensions (forced by the design rules of CMOS technology) of the thick oxide transistors (minimum width, minimum length etc.) used in designing the AMUXes. If the area of a crossbar array is smaller than the minimum required area by the periphery, then the spacing between cells in the crossbar array is increased (i.e $> 2F$ in Figure 1) such that it occupies the same area as the periphery. Those crossbar array capacities that do not permit a $4F^2$ RRAM cell size are indicated in Figure 4 as the $> 4F^2$ region. In this region, the periphery and AMUXes are designed using minimum size transistors, resulting in a huge voltage drop in them. Thus, the total voltage drop, which is the sum of voltage drops in crossbar array metal lines and the AMUX is also very high, $\sim$ 425 mV for 65 nm technology. The small value of the total voltage drop ($\sim$ 50 mV) for 28 nm in Figure 4b is attributed higher HRS of the 28 nm cell compared to the 65 nm cell, resulting in a ($\sim 10\times$) lower SET current. Expanding the array capacity in the $>4F^2$ region slightly increases the total voltage drop due to the increase in the length of metal lines, and the increase in sneak currents due to the rise in number of cells. This is more prominent in Figure 4b due to the higher metal line resistance in 28 nm.

If the area of a crossbar array is larger than the minimum required area by the periphery (indicated as $4F^2$ region), then the periphery can be expanded such that its area is matched with the area of the crossbar array. This permits to

[2] Compliance current: the limiting current for *SET* operation.
[3] $R_{\square}$ is the sheet resistance of metal lines.



(a) 28 nm technology (b) 65 nm technology

Fig. 5: Comparison of Energies and Write Times for Different Crossbar Array Sizes during the Architectural Exploration of a 1 Gb RRAM Bank.

increase the width of transistors (i.e. decreasing their resistance) used in designing AMUXes, resulting in a decrease in the voltage drop in AMUXes. This explains the sudden decrease in the total voltage drop in Figure 4 with the increase in crossbar array capacity at the starting of $4F^2$ region. Even a small decrease in AMUX resistance will result in a large change in the total voltage drop for 65 nm due to the high current flow in the crossbar array. Further increasing the array capacity will result in a point where the voltage drop in the metal lines exceeds the voltage drop in the AMUXes. The total voltage drop increases beyond this point with increasing array size since the metal line resistance is dominant in this region. This transition point in Figure 4 corresponds to an array capacity of 2^{22} bits and 2^{20} bits for 65 nm and 28 nm technology nodes respectively.

Next, we show the influence of the total voltage drop on the read and write timings (SET and RESET), and energies of various crossbar array capacities. Figure 5 plots the results obtained during the architectural exploration of a complete 1 Gb RRAM bank at 65 nm and 28 nm technology nodes using RRAMSpec. Figure 5a plots the variation of SET and RESET times for different crossbar array capacities in 28 nm technology node. There is a direct correlation between the voltage drop in Figure 4b and the SET time in Figure 5a due to the exponential voltage-time dependency of RRAM cell, which is modelled using Equation 1. Therefore, the crossbar array capacity, which results in the minimum SET time in Figure 5a is also the one with minimum voltage drop in Figure 4b. This results in an optimal array capacity of 2^{20} bits for SET. For RESET operation, the minimum RESET time is achieved for an array capacity of 2^{22}. Therefore, the optimal array size with respect to write time (the maximum of SET and RESET times) is 2^{20}. When RESET starts, the selected cell is in LRS. Therefore, it withdraws high current from the driver, resulting in a large voltage drop in the AMUX, and consequently very high RESET times. Thus, allowing wider AMUX transistors ($4F^2$ region) will result in a large decrease in the RESET time as shown in Figure 5a. On the other hand, SET time is not much reduced by the increase in AMUX transistor width. The reason for this is the small current flow through the device under SET due to its HRS, causing less voltage drop in

AMUX. A similar behavior can be observed in Figure 5b for the 65 nm technology node.

In both Figures 5a and 5b, the write energy follows the same trend of write time. It is worth to note that increasing the array size can reduce the total energy spend during a write operation, even though more cells are leaking (HSCs). This is because larger arrays allow wider AMUXes, which drastically reduces the voltage drop in the AMUX due to lower resistance. This results in a larger voltage across the selected cell, decreasing write time and write energy. However, the read energy remains nearly constant while increasing the array size due to very low sneak currents at read voltages. Read voltages are usually much lower than the write voltages, thus, the non-linear selector blocks the sneak currents.

Table 2: Comparison of RRAMSpec with DESTINY

1 Gb bank	DESTINY	RRAMSpec
Area (mm^2)	17.97	10.30
Crossbar array size	2K by 512	1K by 1K
Read latency (ns)	1.37	20.93
Write latency (ns)	53.35	41.38
Read energy (pJ)	42.95	43.64
Write energy (pJ)	72.33	69.76
Leakage power (mW)	7278	0.68

The validation of our modelling approach against the manufactured prototype testchip of high density RRAM crossbar [20] is not possible since the manufacturer does not disclose their proprietary RRAM cell/selector technology details. Instead, we compare the results of RRAMSpec with a state-of-the-art NVM modeling framework, DESTINY [12]. We performed the architectural exploration of a 1 Gb RRAM bank with 4 bits per access using RRAMSpec and DESTINY, targeting a write latency optimized solution. It is important to note that DESTINY does not compute the write latency of the RRAM cell, instead, this value has to be provided by the designer. RRAMSpec calculates the cell write latency based on the available voltage at the selected cell, which is accurately computed considering the crossbar array size, and sneak currents through each HSC as explained in Section 4.2. Therefore, in order to ensure a fair comparison, DESTINY is provided with the RRAM cell write latency, which is already calculated by RRAMSpec. Table 2 shows the comparison results. The chip area estimated by DESTINY is 74% larger than the value computed by RRAMSpec. This is possibly due to the fact that DESTINY does not consider the placement of peripheries underneath the crossbar array. The higher read latency in RRAMSpec compared to DESTINY is mainly originated from the current sensing delay, and the additional sneak current compensation delay. DESTINY outputs a very high leakage power of 7.2 W. We assume this overestimation of the leakage power is due to the assumption that row decoders are present in each cross-point array, and the absence of *Phase* signals [22] that can be activated only during operation.

6 Conclusion and Future Work

In this paper, we presented RRAMSpec: an architectural modeling approach, and a design space exploration framework to evaluate timings, silicon area, and energy consumption of high density RRAM devices. We validated the modeling against SPICE simulations using physics based RRAM models. Sample explorations using RRAMSpec showed optimum array sizes in terms of write/read energies and timings for different technology nodes. Finally, we compared the obtained evaluation results with a state-of-the-art NVM modelling framework. We mainly focused here on bipolar metal-oxide RRAMs, however our framework and modeling approach is extendable to crossbar arrays in general. More advanced features such as multi-bit access per subarray, multi-level cells, and modelling of 3D vertical RRAMs will be included in the future version of the framework. We are not considering the influence of programming voltage on the endurance and retention time of RRAM devices. This will be modelled in the future version of our exploration framework. Later, the tool will be published as open source.

Acknowledgment

This work was funded by the Carl-Zeiss Stiftung under the Nachwuchsförderprogram 2015 and the EU OPRECOMP project (http://oprecomp.eu) under grant agreement No. 732631. This work was also supported by the the Fraunhofer High Performance Center for Simulation- and Software-based Innovation and ERC Consolidator Grant COMPUSAPIEN (Grant No. 725657). The authors thank the Electronic Materials Research Lab (EMRL) at the RWTH Aachen for their great support.

References

[1] Y. Chen and C. Petti. "ReRAM technology evolution for storage class memory application". In: *46th European Solid-State Device Research Conference (ESSDERC)*. 2016, 432–435.

[2] S. W. Fong, C. M. Neumann, and H. . P. Wong. "Phase-Change Memory Towards a Storage-Class Memory". In: *IEEE Transactions on Electron Devices* 64.11 (2017), 4374–4385.

[3] P. Cappelletti. "Non volatile memory evolution and revolution". In: *2015 IEEE International Electron Devices Meeting (IEDM)*. 2015, 10.1.1–10.1.4.

[4] R. F. Freitas and W. W. Wilcke. "Storage-class memory: The next storage system technology". In: *IBM Journal of Research and Development* 52.4.5 (2008), 439–447.

[5] C. H. Lam. "Storage Class Memory". In: *10th IEEE International Conference on Solid-State and Integrated Circuit Technology*. 2010, 1080–1083.

[6] A. Chen and M. Lin. "Variability of resistive switching memories and its impact on crossbar array performance". In: *2011 International Reliability Physics Symposium*. 2011, MY.7.1–MY.7.4.

[7] A. Fantini et al. “Intrinsic switching variability in HfO2RRAM”. In: *5th IEEE International Memory Workshop*. 2013, 30–33.

[8] Yun-Feng Kao et al. “A Study of the Variability in Contact Resistive Random Access Memory by Stochastic Vacancy Model”. In: *Nanoscale Research Letters* 13.1 (2018), 213.

[9] C. Xu et al. “Overcoming the challenges of crossbar resistive memory architectures”. In: *IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*. 2015, 476–488.

[10] A. Ghofrani, M. A. Lastras-Montao, and K. Cheng. “Toward large-scale access-transistor-free memristive crossbars”. In: *The 20th Asia and South Pacific Design Automation Conference*. 2015, 563–568.

[11] JEDEC. *DDR5 & NVDIMM-P Standards Under Development*. URL: `https://www.jedec.org/news/pressreleases/jedec-ddr5-nvdimm-p-standards-under-development`.

[12] M. Poremba et al. “DESTINY: A tool for modeling emerging 3D NVM and eDRAM caches”. In: *Design, Automation Test in Europe Conference Exhibition (DATE)*. 2015, 1543–1546.

[13] X. Dong et al. “NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31.7 (2012), 994–1007.

[14] A. Levisse et al. “Architecture, design and technology guidelines for cross-point memories”. In: *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*. 2017, 55–60.

[15] H. . P. Wong et al. “MetalOxide RRAM”. In: *Proceedings of the IEEE* 100.6 (2012), 1951–1970.

[16] Karsten Fleck et al. “Uniting Gradual and Abrupt set Processes in Resistive Switching Oxides”. In: *Phys. Rev. Applied* 6 (6 2016), 064015.

[17] A. Levisse et al. “Capacitor based SneakPath compensation circuit for transistor-less ReRAM architectures”. In: *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)* (2016), 7–12.

[18] A. Levisse et al. “SneakPath compensation circuit for programming and read operations in RRAM-based CrossPoint architectures”. In: *15th Non-Volatile Memory Technology Symposium (NVMTS)*. 2015, 1–4.

[19] J. Baek et al. “A Reliable Cross-Point MLC ReRAM with Sneak Current Compensation”. In: *2015 IEEE International Memory Workshop (IMW)*. 2015, 1–4.

[20] T. Liu et al. “A 130.7mm^2 2-layer 32Gb ReRAM memory device in 24nm technology”. In: *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*. 2013, 210–211.

[21] S. Kim, W. Lee, and H. Hwang. “Selector devices for cross-point ReRAM”. In: *2012 13th International Workshop on Cellular Nanoscale Networks and their Applications*. 2012, 1–2.

[22] Brent Keeth et al. *DRAM Circuit Design: Fundamental and High-Speed Topics*. 2nd. Wiley-IEEE Press, 2007.