RRAM

Ethan Tola
School of Computer Engineering
Rochester Institute of Technology
Rochester, New York 14623

Pamineo Richards
School of Computer Engineering
Rochester Institute of Technology
Rochester, New York 14623

Amrit Gautam
School of Computer Engineering
Rochester Institute of Technology
Rochester, New York 14623

***Abstract*—When discussing emerging engineering technologies, it is important to understand the pivotal role that non-volatile memory (NVM) plays. Several emerging NVM technologies like STT-MRAM and PCRAM have been tracked in parallel, constantly comparing the ideal characteristics that such memory devices should have. In the past decade, RRAM technology has shown much promise in being able to deliver NVM applications and other potential novel applications as well. This paper will give much insight into the background of metal-oxide based technology from device fabrication to array architecture. This paper will examine several performance factors such as programming speed, variability control and multi-level operations. A review of issues such as uniformity, endurance and retention will also be discussed.**

I. Introduction

Technology is progressing to the next generation and emerging memory technology is one of the biggest future computer memory hierarchy architecture designs we have today. The power of computing systems, their performance and functionality vastly depend on the characteristics of the memory subsystem. RRAM is one of the memory subsystems inside the layers of some well known hierarchy such as static random access memory(SRAM), dynamic random access memory (DRAM), FLASH(NOR and NAND), Spin-transfer Torque Magnetoresistive random access memory (STT-MRAM) and phase change random access memory (PCRAM). SRAM, DRAM, FLASH fall under mainstream memory technologies and SST-MRAM PCRAM, RRAM fall under emerging technologies [1].

Device characteristics of mainstream and emerging memory technologies are based on various factors such as cell area, multi-bit, voltage, read and write time, retention and write energy. RRAM is attractive due to its lower programming voltage and faster write/read speed, thus the primary target of RRAM is to replace the NOR FLASH for code storage and more ambitiously to replace NAND FLASH as data storage [1].

Unlike other forms of memory, RRAM is a form of non volatile memory (NVM), which maintains its data even when powered off. On the other hand, volatile forms of memory lose its data when powered off. Each form of memory will be discussed in detail in other sections [1].

The rest of the paper is organized as follows. After the introduction, Section 2 overviews the background and the theory for the RRAM. Section 3 introduces the result with pros of the RRAM and its application. Section 4 presents the discussion with cons of the RRAM and its application. Finally, Section 5 contains concluding remarks followed by references in Section 6.

II. Background

RAM is a form of memory that can be both volatile and non-volatile. It typically resides on the second and third levels of the memory hierarchy. The volatile form of this memory loses all of the data stored in the system when powered off. The most common form of this type of memory is DRAM. On the other hand, non-volatile memory (NVM) maintains its data when powered off. The most common form of this memory is harddisks or solid state drives (SSD). Unlike those forms of memory, non-volatile RAM is much faster but holds less data.

RRAM is a form of NVM which uses materials that change resistances when a voltage is applied to it. There are two primary forms of RRAM. The first form uses conductive filaments known as oxygen vacancies while the other form uses metal atoms to create conductive filaments [1]. Oxygen based RRAM has been taking the lead in many of the RRAM applications. The internal components or memory cells of this NVM are made up of mem-ristors. The first of this kind of technology has been reported to be back in the 1960s [1]. Because of the usage of mem-ristors, RRAM typically has faster switching speeds than most other NVMs and far less power draw than other mass storage media like NAND Flash or SSDs [2]. With these clear advantages and the dropping prices of RRAM manufacturing, RRAM has the potential to become a highly valuable form of memory as its usage can be found from edge devices and sensors to handling high performance neural net computations.

With any technology, RRAM also comes with its fair share of drawbacks. Unintended filaments inside the mem-ristors, called sneak paths, can be problematic when trying to produce this memory on a large scale. As technology advances and many engineers push forward with this new technology, these drawbacks may soon be resolved.

III. Theory

How memory stores its data is what gives the memory its various characteristics and capabilities. RRAM is a class of NVM technologies that uses resistances to determine whether or not the device is on. As stated previously, RRAM creates filaments between two terminals to store a ‘1’ and destroys these filaments to store a ‘0’.

A. Low Level RRAM Basics

In general, there are two ways to accomplish this: using either metal oxides or fast-diffusive gold or copper ions. These two types do not vary when it comes to higher level characteristics and array design choices. The major differences between different RRAM technology comes from the materials used for the cell creation. For oxide based RRAM, there are a multitude of various elements that are used to create a cell. For CBRAM, primarily either gold (Ag) or copper (Cu) are used as intermediate layers between the two electrodes. The electrode elements also vary and can be summarized in Fig. 1.

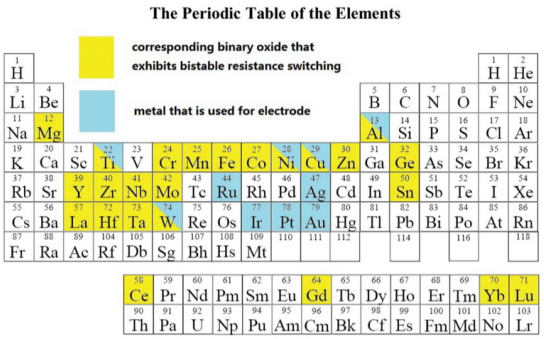

Fig. 1. Periodic Table of Used Elements [1]

Aside from the elements used, the oxide or Ag/Cu layer are sandwiched between two electrodes. To create either the filament or conductive bridge requires a voltage to be applied to one of the terminals. Fig. 2 helps illustrate this concept.

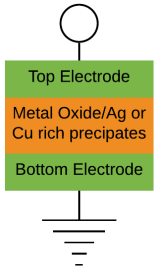

Fig. 2. RRAM Cell Breakdown

For oxide based RRAM, when the same voltage is applied to the top terminal that creates or destroys the internal filaments, this is known as unipolar switching. This switching mode only cares about the amplitude of the voltage rather than the polarity. Bipolar switching mode takes into account the polarity of the voltage where one polarity creates a filament and only the other polarity can destroy it. For both switching modes, there is a compliance current that is used to avoid permanent dielectric breakdown in the forming/set process. In general, bipolar switching mode is used as unipolar switching requires a higher compliance current and has larger variability as well. Using bipolar switching mode, there is an additional layer known as the interfacial oxygen barrier which is why the opposite voltage is needed. In general, a typical RRAM cell consists of a top electrode, metal capping layer for oxygen gettering, the oxide, and a bottom electrode. Both the top and bottom electrodes are typically the same material.

Overall, these lower level considerations typically affect the write speeds or power consumption of the cells as a whole. One of the main advantages of using these metals depicted in Fig. 1 is that they work well with in-place CMOS fabrication processes. Aside from how the cell is made, a majority of its applications are derived by how its used. The next section explains some various array architectures that RRAM cells are used in.

B. Array Architecture

Before diving into the world of the many common RRAM array architectures, it is important to realize both the cell design and process technology. For cell design, because of the abrupt current increases that can occur in filamentary RRAM devices during the forming/setting process, it is necessary to optimally clamp such current to avoid possible degradation. A transistor is almost always used here to limit this behavior due to the fast response times and large resistance in the saturation region. It is also important to note that during the forming/setting process the resistance changes instantly while the voltage drop across the RRAM cannot. It is to no surprise that many of these characteristics are possible due to the presence of parasitic capacitance. It is commonly believed that the peak value of the overshoot current is the most important factor affecting filament characteristics [3]. This peak is linked to the parasitic capacitance which is why we must try to avoid some degree of this. Special consideration must be taken into account when looking to the design of the memory cell.

For process technology, no preferred material for RRAM has yet to be determined. RRAM device fabrication is known to use conventional semiconductor fabrication tools and is compatible with the silicon CMOS (BEOL) process. The fabrication process of the memory cell should be carefully controlled to prevent the oxide layers from suffering any kind of damage. Furthermore, there must be a compromise between the forming voltage and memory window. The forming voltage undesirably increases when increasing the device area. Because the forming process leads to oxide breakdown in the RRAM device, this linear correlation was also found to be between the forming voltage and the device size and can be used to monitor the fabrication process of the memory cell across a range of cell sizes [4].

Probably the most popular and common RRAM array architecture to date is the one-transistor and one-resistor array (1T1R). In this design there are several RRAM cells which are all in series with a cell selection transistor. The addition of these transistors allows for individual cells to be turned on or off as needed. The word line (WL) controls the gate of the transistor. The bit line (BL) connects to the RRAM cell's top electrode. The source line (SL) connects to the source of the transistor. 1T1R is the desirable array architecture for

embedded applications. This is primarily due to the fact that performance and reliability are top priority.

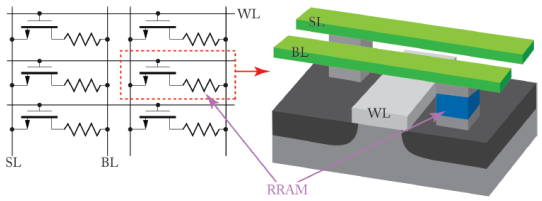

Fig. 3. 1T1R Array (Schematic). Cell Area: $6F^2 to 12F^2$

1T1R is again a desirable choice when it comes to the typical write/read scheme For set operations, word line (WL) voltage is turned on so that the transistor state is ON for the desired cell, and a write voltage applied to the bit line (BL) while grounding the source line (SL). For reset operations, word line (WL) voltage is turned on and a write voltage applied to the source line (SL) while the bit line (BL) is grounded thus reversing the current flow. For unselected rows and columns, their lines are grounded. For reading data from the array, the write line (WL) voltage is turned on and a read voltage applied to the bit line (BL) while SL is grounded.

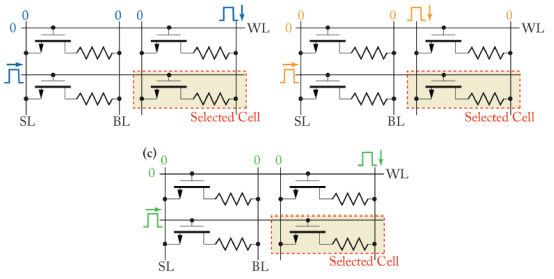

Fig. 4. 1T1R Array (Write/Read scheme). Set (Left) / Reset (Right) / Read (Bottom)

Another common RRAM array architecture is the cross-point array. In this design the rows and columns need to be perpendicular to each other with RRAM cells in the middle. It is important to realize here that the cross-point array can achieve a cell area of 4 F^2. This gives it a competitive advantage over 1T1R because of its higher density, which makes it better for the standalone NVM applications.

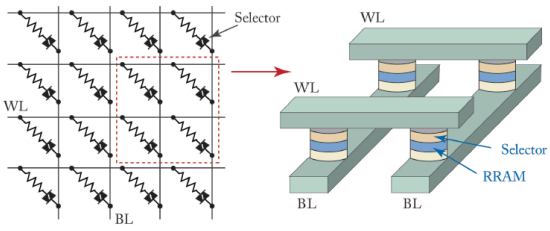

Fig. 5. Cross-point Array (Schematic). Cell Area: $4F^2$

The write/read scheme for the cross-point array differs greatly from the 1T1R. Because there are no selection transistors, cross talk is possible between cells in the architecture. Two write schemes can be applied here, V/2 scheme and V/3 scheme. Only the V/2 scheme will be discussed here because it typically uses less power and consumes less energy. For set operations, the desired cell's word line (WL) and bit line (BL) are biased at the write voltage and ground, respectively. For reset operations, the bias conditions for the word line (WL) and bit line (BL) are reversed. For reading the data from the array, all columns are biased at the read voltage. Meanwhile the selected row is biased at ground and the unselected rows are biased at the read voltage. This results in only the cells in the selected row being the only one to see the read voltage.

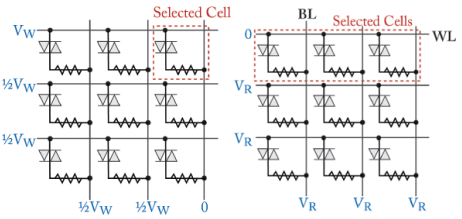

Fig. 6. Cross-point Array (V/2 Write/Read scheme). Write (Left) / Read entire row (Right)

C. RRAM as a Whole

RRAM has a big potential to influence and replace other memory systems inside the layers of some well known hierarchy. RRAM may be the large capacity NVM technology to replace FLASH (NOR or NAND) but it also offers suitable embedded application. One of the main advantages of this is because it provides the low programming voltage that other hierarchy memory like FLASH does not provide. It likewise offers the non-volatility that DRAM doesn't offer but has a speed that is similar to DRAM, which gives RRAM an edge to serve as the storage class memory between FLASH and DRAM in memory hierarchy [1].

CMOS logic process is important for storage class memory. For a storage class memory to be compatible with the CMOS logic process it must have improved cycling endurance and low programming voltage, which is known to have much higher priority than other attributes. The estimated cycling endurance rate for RRAM is estimated to be quite high for RRAM. If the application target is to place the RRAM close to the DRAM based main memory, the endurance of a single cell should be $4.8x10^17$ cycles, which appears to be quite hard for RRAM to accomplish. However there is a way to get around it. With the help of architectural wear-leveling techniques, it is possible to uniquely distribute write events to various sub-blocks which can reduce the endurance of a single cell to $4.8x10^9$ cycles. Similarly for RRAM to target the application level of FLASH memory the cycling endurance needs to be much reduced [1].

RAMM has shown to have radiation- hardened characteristics which is seen to have potential to serve the aerospace or nuclear industry as it can survive in harsh radiation environments. It is suggested that RRAM memory cells are robust against many radiation effects [1].

IV. Results

As seen before, RRAM has a multitude of characteristics that can be applied to various fields. Like most NVMs, RRAM can be a replacement to traditional memory such as DDR and FLASH depending on the application that it's needed for. This section will explore the benefits of RRAM and what the current applications for this emerging technology are.

A. Benefits of RRAM

The scalability to the nanometer regime is one of the key motivations that push the development of metal-oxide RRAM technology [4]. The scalability of RRAM is proved to be exceptional and has been successfully fabricated with sub-10 nm devices. While it was once thought that RRAM would be suited best for standalone NVM applications, research now shows that it may be better suited for Embedded Applications. It doesn't hurt that this technology also has a set of unique physical properties that could allow for new functionality and features in systems. RRAM also offers a lower programming voltage than that of embedded flash and non-volatility that DRAM does not. This makes RRAM an attractive option for storage class memory. If concrete opportunities are to come from this technology, researchers must place extreme emphasis on this being part of the future computing model.

B. RRAM Applications

One of the most major and promising applications of RRAM technology is in neuromorphic computing. Today, deep neural nets (DNN) are pioneering many fields such as image processing, natural language processing, and robotics. The problem with DNNs is not only the high memory usage but the immense amount of power it takes to not only train but operate DNNs. Some of the benefits of having low activation voltage is that RRAM can be utilized with higher efficiency while maintaining the high read/write speeds and high density [5]. With the low power usage of RRAM devices, DNNs can also be put on edge devices for more flexible usages of DNNs. Aside from power, DNNs also require an immense amount of calculations. With new methodologies such as crossbar to increase efficiency in calculating the common multiple-and-accumulate (MAC) operations, RRAM can even compete with mainstream memory technologies [5].

Alongside the usage of RRAM in DNNs, this low power usage of RRAM can also be used in embedded applications, especially when it's tied with fast write speeds. Compared to FLASH which can write to memory at best in 8 μs, RRAM has been measured to successfully write in as little as 5 ns [6]. With this incredible out performance, RRAM may soon be the leading NVM storage cell for embedded devices. Along with that, RRAM has been tested to be fairly resilient to heat up to 150 °C, thus making it an even more promising candidate for embedded applications [7]. As stated previously, this resilience of RRAM technology extends even towards radiation making it a possible candidate for aerospace or nuclear applications.

V. Discussion

Although RRAM is great, there are many NVM and RAM technologies out there that have been and are currently being used now. Even though RRAM brings a lot to the table, there are plenty of drawbacks to RRAM that make it difficult to incorporate in large memory spaces where memories such as DDR2 or FLASH have taken hold.

One of the main benefits of using RRAM is its capabilities to reduce power significantly compared to memories such as FLASH which require high voltages to read and write. Increasing RRAM's resistance to reduce power brings forth dynamic resistance instability (DRI) which are issues such as, retention, short-time instability, and time-dependent variability (TDV) [8]. These are issues that arise after the verification process.This issue becomes a problem with neuromorphic computing applications as TDV can cause large shifts in established weights or even degredad data retention over large periods of time [9]. According to Zhizehn Yu and others paper, increasing the resistance to around $100k\Omega$ will cause an "early-stage" fluctuation (ERF) in RRAM cells that vary weights in neuromorphic computing. This is different from TDV or retention problems as techniques such as relaxation or refreshing cannot be applied as it is too time consuming and occurs too frequently (time span of 0 10000s) for those methods respectively. With ERF moving un-checked in this application can degrade a neural net's accuracy from 98% to 58% after 10000s [8].

Aside from the intrinsic problems of RRAM technology, RRAM also brings challenges to its large scale approach with high varioustions from device-to-device [10]. This can typically be solved within the structure of the circuits but there are some applications where that cannot be done and its performance depends directly on how well the RRAM cells operate. This is specifically applicable for multi-level RRAM [11]. Regardless, like any emerging technology, there are numerous engineering challenges that have to be overcome before the technology can be widely used and RRAM is no exception. As it is right now, RRAM still poses a threat to both mainstream RAM and NVM technologies.

A. RRAM vs Mainstream RAM

All mainstream memory technologies (SRAM, DRAM, FLASH) are based on some sort of charge storage mechanism. What they all have in common is the case that they face many challenges in being scaled down to sub-10 nm nodes or beyond. This is a common problem because of physical properties which make it quite difficult to retain a charge on the nanoscale. This is one motivating factor for finding a non-charge based counterpart, exploring emerging technologies like RRAM, STT-MRAM, and PCRAM.

Mainstream RAM technologies have numerous disadvantages compared to emerging technology like RRAM. One of the disadvantages is being not able to use proper modification and optimization techniques that RRAM uses. RRAM also has the benefit of having the characteristics to adopt a hybrid design. RRAM also has the potential to add and modify any

new memory architecture design which is not possible for mainstream RAM [12].

Other big problems mainstream technologies deal with are low density, which limits increase of on-chip memory capacity as well as high leakage power consumption. High leakage power consumption delay or some time stop the progress of memory hierarchy design. Another problem they have is they are more vulnerable to radiation based errors which cases performance improvement and more power uses. This is where RRAM solves most of these problems because of its advantages with high density, zero standby power, fast access speed, and non volatility [12].

B. RRAM vs Emerging NVM

Traditional memory has been suggested to be slower than computer logic technology. Other mainstream memory technologies, such as SRAM and DRAM have low density, high standby power and are unreliable. The idea here is to replace these memory technologies with other various emerging nonvolatile memory (NVM) technologies such as STT-RAM, RRAM, and PCRAM. NVMs offer high density, zero standby power, fast access speed, and non volatility.

For emerging technologies, STT-MRAM and PCRAM share some common features like being two terminal devices, nonvolatile, and differentiate state using HRS and LRS. The device characteristics amongst the emerging technologies are very different due to the underlying physics. STT-MRAM relies on the difference in resistance between the parallel and antiparallel configuration of two ferromagnetic layers separated by a thin insulator layer. PCRAM relies on chalcogenide material to switch between the crystalline phase and the amorphous phase. RRAM appears to have a lower programming voltage and faster read/write speeds. STT-MRAM has an advantage over SRAM, having a smaller cell area [1]. The other interesting aspect here is to compare RRAM with other nonvolatile memory technologies with various factors such as cell area, multi-bit, voltage, read and write power, data retention and write energy.

	PCM	STT-RAM	RRAM
Data retention	Y	Y	Y
Cell factor (F^2)	6–12	4–20	<1
Read latency (ns)	20–50	2–20	<50
Write latency (ns)	50–120	2–20	<100
Write numbers	10^{10}	10^{15}	10^{15}
Read/write power	High	Low	Low
Other power	None	None	None

TABLE I
COMPARISON OF THE DIFFERENT NVMS TECHNOLOGIES [12].

The comparison between different NVMs technologies are shown in Table I. STT-RAM memory is more suitable for on-chip memory design. PCM are usually used for main memory design or as the cache of NAND flash [12]. RRAM have shown various features, in respect of performance, energy and density with low read/write power.

VI. CONCLUSION

In summary, this paper covered the basics of RRAM technology, ranging from how a single cell is constructed to how various arrays are constructed and utilized in the larger memory hierarchy. With that basis, key advantages of RRAM were discussed as well as current applications of this technology. To bring it together, challenges of RRAM technology were discussed alongside how RRAM compares to its competitors, highlighting where the advantages and disadvantages are seen.

In conclusion, RRAM is a highly promising technology that could revolutionize the memory industry. With its low power, high read/write speeds, dense architectures, and compatibility with current CMOS manufacturing processes, RRAM can be utilized in nearly every field. Some of the greatest advantages of RRAM stem in the fields of embedded and especially, neuromorphic computing. But before RRAM can be used at large scale, the various engineering challenges have to be overcome. If they can be, RRAM will certainly be seen in nearly every piece of technology, replacing all sorts of memory such DDR2 and FLASH as discussed in this paper.

REFERENCES

[1] S. Yu, *Resistive random access memory (RRAM): from devices to array architectures*. Morgan & Claypool Publishers, 2016.

[2] M. Rouse, "What is rram or reram (resistive ram)? - definition from whatis.com," Jun 2017. [Online]. Available: https://searchstorage.techtarget.com/definition/RRAM-or-ReRAM-resistive-RAM

[3] P. R. Shrestha, D. M. Nminibapiel, J. P. Campbell, J. T. Ryan, D. Veksler, H. Baumgart, and K. P. Cheung, "Analysis and control of rram overshoot current," *IEEE Transactions on Electron Devices*, vol. 65, no. 1, pp. 108–114, 2018.

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

[5] G. Charan, A. Mohanty, X. Du, G. Krishnan, R. V. Joshi, and Y. Cao, "Accurate inference with inaccurate rram devices: A joint algorithm-design solution," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, pp. 1–1, 2020.

[6] S. Sheu, K. Cheng, M. Chang, P. Chiang, W. Lin, H. Lee, P. Chen, Y. Chen, T. Wu, F. T. Chen, K. Su, M. Kao, and M. Tsai, "Fast-write resistive ram (rram) for embedded applications," *IEEE Design Test of Computers*, vol. 28, no. 1, pp. 64–71, 2011.

[7] J. Liu, H. Lv, X. Xu, D. Dong, P. Yuan, Z. Yu, and M. Liu, "Resistive switching memory towards embedded application in 28 nm node and beyond," in *2018 IEEE International Conference on Electron Devices and Solid State Circuits (EDSSC)*, 2018, pp. 1–2.

[8] Z. Yu, Z. Wang, J. Kang, Y. Fang, Y. Chen, Y. Cai, and R. Huang, "Early-stage fluctuation in low-power analog resistive memory: Impacts on neural network and mitigation approach," *IEEE Electron Device Letters*, pp. 1–1, 2020.

[9] J. Kang, Z. Yu, L. Wu, Y. Fang, Z. Wang, Y. Cai, Z. Ji, J. Zhang, R. Wang, Y. Yang, and R. Huang, "Time-dependent variability in rram-based analog neuromorphic system for pattern recognition," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 6.4.1–6.4.4.

[10] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, p. 075201, jan 2012. [Online]. Available: https://doi.org/10.1088%2F0957-4484%2F23%2F7%2F075201

[11] S. R. Lee, Y. Kim, M. Chang, K. M. Kim, C. B. Lee, J. H. Hur, G. Park, D. Lee, M. Lee, C. J. Kim, U. Chung, I. Yoo, and K. Kim, "Multi-level switching of triple-layered taox rram with excellent reliability for storage class memory," in *2012 Symposium on VLSI Technology (VLSIT)*, 2012, pp. 71–72.

[12] G. Sun, J. Zhao, M. Poremba, C. Xu, and Y.Xie, "Memory that never forgets: emerging nonvolatile memory and the implication for architecture design," pp. 1–5, 2018.