Critical Issues in Advanced ReRAM Development

Martin Peckerar[1], Po-Chun Huang[1], Rachid Ahmad Jamil[1], Bruce Jacob[2], and Donald Yeung[1]

[1] *Department of Electrical and Computer Engineering*
University of Maryland
College Park, MD 20742
[2] *United States Naval Academy*
Annapolis, MD 21402

Abstract

For many years, Resistive RAM (ReRAM) technology has been pursued as a potentially high yield 3D memory. Recent improvements include the addition of diode-select devices that reduce sneak path leakages. These memory structures can be made using only back-end-of-the-line processing steps. ReRAM materials are fully compatible with backend processes and the resulting memories are planar, stackable units. No active devices are present in these stacks. These devices are highly suitable for new memory architectures, such as edge computing or compute-in-memory. We have developed interlayer interconnect architectures that minimize individual cell sizes, which we disclose in this paper. Persistent problems remain. These include poor device yield and poor cycle endurance. These issues can be traced to the basic mechanism of ReRAM operation: the formation (and destruction) of conducting filaments creating the set and reset states. The tips of these filaments develop exceptionally high electric fields due to field-line compression (lightning rod effect). The filament tips will undergo field-forming rearrangement, leading to arc-over and ultimate device failure. In this paper, we describe alternative methods of conduction bridge formation in which these high fields are not necessary for realizing the set/reset cycle. In conventional ReRAM devices, current flow during read and write is perpendicular to the chip surface. In the structures we propose, current flow is horizontal with respect to this surface. We refer to these devices as HReRAMs. Process flows and characterization results for these structures will be prescribed in this paper.

1 Introduction

As Moore's Law comes to an end, computer designers can no longer rely on device scaling to improve the energy efficiency and performance of computer systems. This is occurring at an extremely inopportune time. Enabled by deep learning, artificial intelligence is beginning to demonstrate capabilities rivaling humans across many important tasks. For example, convolutional neural networks (CNNs) can now achieve better-than-human performance on object recognition. Also, the emergence of Transformer models has enabled unprecedented natural language processing capabilities in systems like ChatGPT. In order to continue making advances in AI, computer system designers will need to provide even higher levels of performance and efficiency on key AI kernels in the future, but without the benefit of Moore's Law scaling.

Modern deep neural networks (DNNs) like CNNs and Transformers require performing linear algebra operations with massive amounts of low-precision floating point. On general-purpose systems, these computations are the bottleneck, but they are easily accelerated by special-purpose hardware. After using accelerators to address this computational bottleneck, the bottleneck in DNNs shifts to the memory system . Memory is a major limitation for machine learning workloads because the models they employ are so large, a problem that is getting worse over time. For most machine learning (ML) models, access to off-chip "global" memory remains the primary bottleneck.

To address the memory bottleneck problem in the post-Moore era, we believe it will be necessary to leverage new memory technologies, rather than try to scale or re-package old ones. In recent years, there has been significant interest in non-volatile memories?such as resistive RAM (ReRAM), magnetic RAM (MRAM), ferroelectric RAM (FeRAM), or phase change memory (PCM) either as a supplement to or replacement for DRAM. Many researchers have tried to design new memory system architectures around these non-volatile memory technologies. Despite all of this research, prior work has overlooked what we believe to be a significant opportunity for emerging non-volatile memories: their compatibility with CMOS logic. Whereas fabricating DRAM requires special VLSI processes tuned for implementing DRAM memory cells, fabricating non-volatile memory can be done within the context of a standard CMOS logic process. This implies that we can integrate non-volatile memory directly into logic dies, including DNN accelerators.

Most importantly, ReRAM is a true 3D technology. The memory element itself is not an "active" element, requiring access to high-quality semiconductor material layers. It is fabricated as part of the back-end-of-line processing steps within a standard VLSI process flow, making use of both metals and oxides (or other insulators). This is inherently much simpler and less costly than die

stacking, and has the potential to scale to many more layers. Also, the single-die solution that we seek further reduces cost, and is well suited not only for high-performance systems, but also for edge and IoT devices.

Unfortunately, as with any new technology, significant problems are introduced by what is a radically new technology. These problems and our approach to their solution are discussed below. We use resistive random access memories (ReRAMs) as an example. But many of the issues encountered are common to 3D technology as mentioned above. We describe the weakness in current vertical-transport ReRAM (VReRAM) technology and how to overcome these issues with a horizontal transport (HReRAM) approach. relevant references supporting the work documented here are: [1] [2] [3].

2 The Evolution of the VReRAM to the HReRAM Device

In this section, we describe the operating principles of our new HReRAM memory cell. We first outline the weaknesses of existing ReRAM technology and then demonstrate how these weaknesses can be eliminated with changes to the basic non-volatile memory cell.

2.1 Issues with existing ReRAM Non-Volatile cells and Proposed Solutions

Non-volatile semiconductor memory (NVSM) has been an essential part of the electronics tool kit for almost half a century [6]. Its attractions are obvious. It stores information almost indefinitely with no attendant power consumption. In addition, 3D instantiations have been realized (albeit with massive process complexity and related yield problems.

The main material system used in current NVSMs is either the metal-nitride-oxide-semiconductor (MNOS) or the silicon oxide nitride oxide silicon (SONOS) layer structure. Both of these systems rely on fixed charge stored beneath a relatively familiar field effect transistor (FET) gate. These transistors require source-drain terminals with an active channel region between.

The source-drain regions take up space and makes 3D stacking difficult. A number of alternate materials solutions have been offered. Magnetic (MRAM), ferroelectric (FeRAM), phase change (PCM) and resistive (ReRAM) cells come to mind. These cells do not require diffused junction contacts. Rather, they work using extensions of the already present sense and control lines of the

RAM cell. One of the easiest of these advanced cells to manufacture is the ReRAM. our focus is on that technology.

A typical VReRAM is shown in Figure 1. The material components of the VReRAM are described in this figure. Note, there are many design and material selection paths for ReRAM

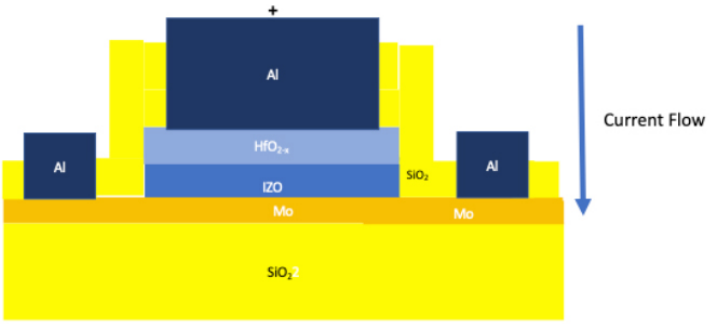

Figure 1: A cross section of a "normal" vertical transport resistive memory (VReRAM). The top aluminum electrode sits over the lower layer structure comprised of a hafnium oxide (HfO_{2-x}) film on top of an indium zinc oxide film. This is referred to as the ReRAM stack. The "stack", in turn, lays flat on a molybdenum film. The molybdenum film (and the aluminum lines that contact it) form the bottom contact. The device is, essentially, blanketed in oxide and thus each cell is fully isolated from other cells or other devices.

architecture. We demonstrate a single one here (emphasizing HfO_2 as the primary component of the memory cell stack), but others are possible. The cell functions in the following way. A positive bias on the aluminum top electrode will pull negatively charged oxygen ions off of the hafnium oxide network. This will lead to a filament of contiguous oxygen vacancies (V+), as shown in Figure 2.

What results is basically a "bed of nails" field emitting surface. The problem with this structure is that the tip electric field is very high, and current density through the necked-down tips is very high as well. There is no clear end to the "field forming" process. Some tips "light off" early, and others later. This leads to current hogging, premature tip blow out, a variable memory window and poor endurance. To overcome this problem, we propose a variant of this basic design, which we call a horizontal ReRAM, or HReRAM.

The basic structure of our HReRAM is similar to that of the VReRAM, but current flow during the read cycle is horizontal with respect to the surface of the underlying substrate. The vacancy forming region (tv) is thicker, as

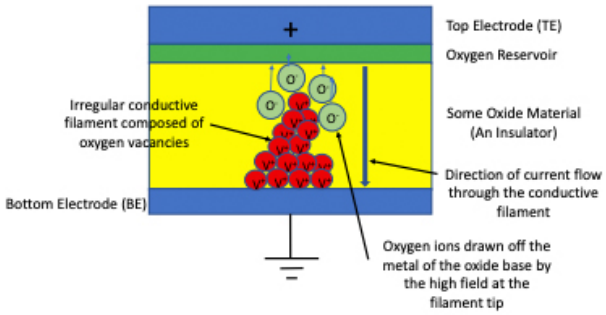

Figure 2: Filament formation in a hafnium oxide resistive RAM cell.

we wish to form a contiguous layer of vacancies at the interface between the bottom electrode and the vacancy-forming layer. We do not wish to "short" the filament to the top electrode. Note the thin nitride layer between the top electrode and the oxygen reservoir layer. This layer (and the thick vacancy forming layer) will have lower parasitic capacitance than the VReRAM counterpart. This, in turn will speed the cell response time. The write and read operations for HReRAM are accomplished in a fashion similar to conventional ReRAM memory. A bias is applied to the top contact drawing oxygen ions off of their network sites, leaving a Filament Forming Layer

of vacancies behind. These vacancies coalesce into a conductive filament. But the vacancies also agglomerate at the base of the filament and eventually form a conducting sheet along the back contact. This "active channel" (the contiguous conducting layer) of the HReRAM is contacted by the source and drain metallization, as shown in the figure. The field-formed filament and active channel are both formed by the vertical field and the low resistance channels create a logic zero state. So the operational procedure goes like this. The source and drain are grounded and a bias is applied between the bottom electrode (BE) and the top electrode (TE). This will set the logic state of the cell. Next, BE and TE are grounded and bias is applied between the source and drain contacts, clearly labelled in Figure 3. The net result is a current flow in the horizontal direction shown in the figure.

One very significant benefit of the HReRAM structure is minimization of

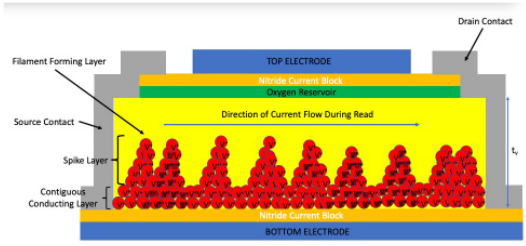

Figure 3: The base structure and direction of read current flow in an HReRAM. Here, a positive bias is applied to the source contact. This gives us a current in the direction shown by the arrow in the figure.

power dissipation during the write part of the memory storage cycle. Even though the write current in any ReRAM is low, write voltages are high, leading to a relatively large VI (power) product. In our design, the nitride insulating films blanketing the top and bottom electrodes completely cuts off current to external circuitry. This, in turn, lowers dissipation during this power-consumptive part of the cell operation.

The main difference between the VReRAM and the HReRAM is this. Vertical fields form the logic state in both cases. But for HReRAM, the cell is read using a "horizontal" field created by bias between the source and drain. In the HReRAM case, these fields are not enhanced by the curvature of a filament tip. The low field established in the read part of the cycle enhances cycle life and improves endurance. It is the goal of our research to utilize this effect to create high reliability, long lasting memory cells.

2.2 Materials Issues and Experimental Results in HReRAM Cell Fabrication

A number of oxide layer structures have been studied for use in resistive RAM applications. Two are described here.As described in this reference, a number of mechanisms can come into play in a resistive memory. Vacancy filaments, metallic filaments, phase change and other bias-dependent charging mechanisms can give rise to bias dependent resistance changes. First, HfO_2 films are the most common hi-K dielectrics used in deeply scaled CMOS memories and microprocessors. In addition, $HfO_2$is one of the most studied ReRAM materials. Han, et al.[4], have successfully fabricated a horizontal device, as

shown in figure 4.

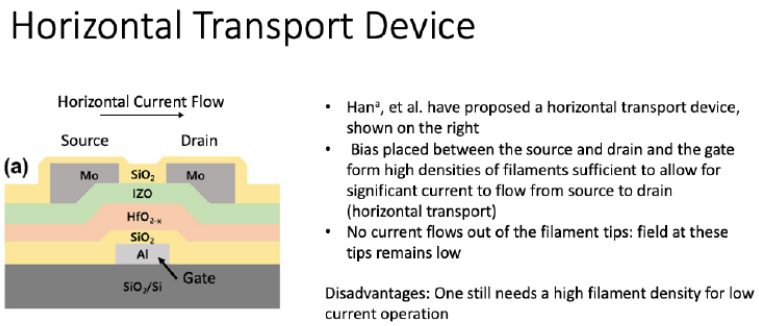

Figure 4: A version of an HReRAM fabricated and described in [4]

The material system itself is stable over time and over a range of temperatures. The metal "stack" at the heart of the memory element is cheap and easy to fabricate. The resulting devices (shown in the figures above) is nearly planar and can be stacked to form a 3D memory without the use of exotic structures (such as transistors built vertically on pillar sidewalls). Deep, Through the Silicon Vias (TSVs) are unnecessary. The HfO_2 structure, as shown in figure 4) develops a memory window in the milliamp range (easily detected by conventional latch technology), albeit at relatively high voltages (20V).

More recently, we have explored a new approach: a metal-doped oxide approach which utilizes field assisted positive metal migration to open and close the RAM switching element, as shown in figure 5.

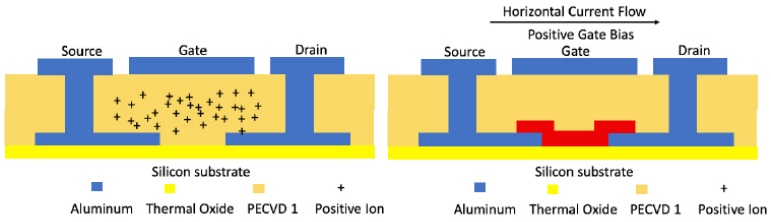

Figure 5: Field Compression Sof Positive Ions into a Conducting Film for Horizontal Transport

Here, the mobile ion introduced is copper - a standard back-end-of-the-line additive. We have fabricated these devices by "sandwiching" an ultra-thin copper film between two plasma-enhanced chemical vapor deposited (PECVD) oxide layers, as shown in figure 6. Figure 7 shows the current memory window

forming after cycling for 3 times with 10V drain bias. the window opening is approximately 3μA.

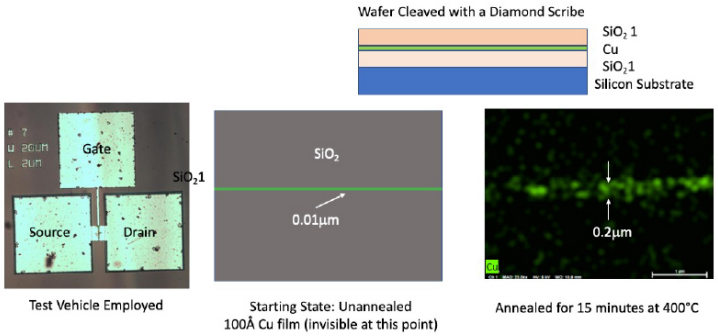

Figure 6: Illustration of the process used to "dope" the PECVD oxide with mobile copper ions using a furnace anneal at 400C for 15 minutes

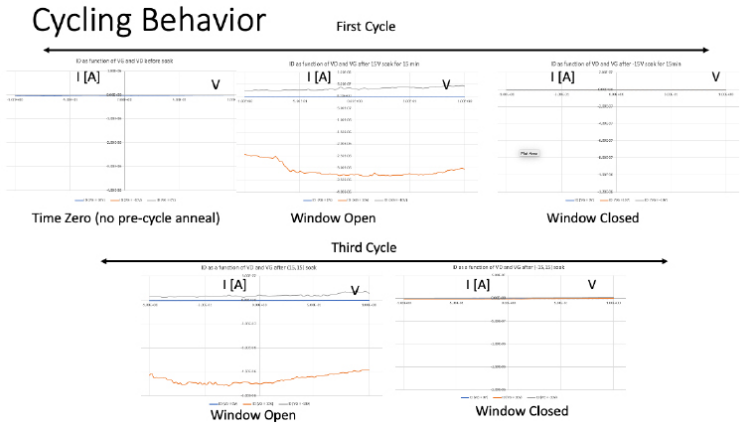

Figure 7: Current memory window openned by annealing copper sandwich for 15 minutes at 400C.

In addition, the HReRAM stack can include other metals and materials such as amorphous silicon.

These can be used to create relatively high performance diodes that can be integrated to form a selector diode. Such selectors are necessary to prevent "sneak path" leakage paths creating read (or write) disturbs and prevent individual device access.

3 Conclusions

Horizontal transport ReRAMs are possible and may solve major problems in 3D non-planar design. The metal ion approach described in this talk shows promise in alleviating major problems in ReRAM application: a. field-stress induced failure during the read/write cycle; and b. yield loss due to random rupture of individual emitters. The process is planar and fully stackable, enabling 3D fabrication. While a read/write window has been demonstrated, much work is necessary in defining optimized thermal cycles, layer thicknesses, and operating voltages.

The presence of copper should not be an issue, as copper is a standard back-end ingredient in VLSI processing today. Diffusion barriers will, of course, be necessary to prevent front-end contamination.

References

[1] G. Campardo, M. Micheloni, D. Novosel, **VLSI Design of Non-Volatile Memories**, Springer, (2005).

[2] J.Brewer, M. Gill, **Nonvolatile Memory Technologies with Emphasis on Flash: A Comprehensive Guide to Understanding and Using Flash Memory Devices**, IEEE Press Series on Microelectronic Systems Book 8, (2011).

[3] B. Prince, **3D Memory Technologies**, Wiley and Sons, Ltd, (2011).

[4] Jimin Han, Boyoung Jeong, Yuri Kim, Joonki Suh, Hongsik Jeong Hyun-Mi Kim, Tae-Sik Yoon, "Nonvolatile memory characteristics associated with oxygen ion exchange in thin-film transistors with indium-zinc oxide channel and HfO2-x gate oxide", Materials Today Advances **15** (2022) 100264.