Architectural Evaluation of 3D Stacked RRAM Caches

Dean L. Lewis                        Hsien-Hsin S. Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
{dean, leehs}@ece.gatech.edu

ABSTRACT

The first memristor, originally theorized by Dr. Leon Chua in 1971, was identified by a team at HP Labs in 2008. This new fundamental circuit element is unique in that its resistance changes as current passes through it, giving the device a memory of the past system state. The immediately obvious application of such a device is in a non-volatile memory, wherein high- and low-resistance states are used to store binary values. A memory array of memristors forms what is called a resistive RAM or RRAM.

In this paper, we survey the memristors that have been produced by a number of different research teams and present a point-by-point comparison between DRAM and this new RRAM, based on both existent and expected near-term memristor devices. In particular, we consider the case of a die-stacked 3D memory that is integrated onto a logic die and evaluate which memory is best suited for the job. While still suffering a few shortcomings, RRAM proves itself a very interesting design alternative to well-established DRAM technologies.

1. INTRODUCTION

Die stacking is an exciting new manufacturing technology that allows multiple layers of silicon to be stacked one on top of the other and tightly integrated with short, fast through silicon vias (TSVs). In the near-term, the simplest and most logical application of 3D integration is to continue the trend of bringing more and more functionality on-chip. For 3D processors, a logical choice is to integrate memory into the stack. Most work in 3D memory integration has focused on the addition of a large last-level cache [2, 7]. The memory technologies under consideration run the gambit from well-studied SRAM and DRAM to non-volatile technology like Flash to promising new memories like PRAM and MRAM.

While a large last-level cache is a nice feature, what we would really like is to be able to integrate the entire system memory into the 3D stack. Unfortunately, none of the above memory technologies is quite capable of realizing such a design. The main deficiency is simply density. Even DRAM, with just a capacitor and transistor in each cell, still requires too many silicon layers—sixteen or more—to even come close to providing the several gigabytes of memory required for modern systems. This of course ignores the very important concerns of power consumption and heat dissipation.

In order to move the system memory into the stack, a new memory technology is required. A very exciting recent development in memory technology is the discovery of the memristor, a fundamental circuit element that promises an order-of-magnitude reduction in cell size compared to DRAM.

This research is supported in part by the C2S2 center of the SRC’s Focus Center Research Program and an NSF grant CCF-0811738.

2. MEMRISTORS

The existence of the memristor was first theorized by Dr. Leon Chua nearly four decades ago [3]. Chua noted the four electrical circuit properties—voltage, current, magnetic flux, and charge—should allow for six different relationships (Figure 1). The first two are simple time-derivatives: voltage to flux and current to charge. Three more relations are covered by the traditional passive circuit elements: resistors (voltage and current), inductors (current and flux), and capacitors (voltage and charge). But what about the fourth, flux and charge? If the relationship is linear, we just have a resistor. But if it is non-linear, we get a very interesting behavior. Chua named this behavior memory resistance or memristance. Simply described, the resistance of the device changes in response to an applied voltage or current.

Figure 1: The four circuit properties and their relations.
Figure 2: A generic memristor structure. A memristive layer is sandwiched between two electrodes. The boundary between high and low resistance regions moves up and down as current passes through the device, changing the total effective resistance.

atoms in the film, changing the memristor’s resistance. A positive voltage reduced the resistance, while a negative voltage increased it. The resistance swing ranged across several orders of magnitude, a significant and easily-observable change. A sufficiently small voltage potential was used to probe the state of the memristor without changing its resistance.

Figure 3: An example hysteresis loop. Note the four regions: the high and low resistance states, and the two transition regions. As shown, some threshold voltage $V_{\text{trans}}$ must be reached to change the memristor’s resistance. Smaller voltages can be used to probe the resistance of the device.

and we are looking at a storage density several orders of magnitude greater than is feasible with DRAM.

Taking a different approach, a team from the University of Michigan created a memristor in a CMOS-compatible process [11, 10]. This memristor has the same structure from Figure 2, but, rather than metal, the bottom electrode was made from the p-doped silicon of a traditional CMOS process, allowing these memristors to be built directly on a silicon substrate. The top electrode was Ag and the switch material amorphous Si, neither material exotic by modern fabrication standards. The downside of CMOS compatibility is that the device is limited to CMOS dimensions, and the transistors in this experiment had a half-pitch of approximately 120nm, much larger than DRAM.

And in a fairly bizarre design, the National Institute of Standards and Technology created a memristor that is 2.5cm on each side [6]. Their target market is inexpensive portal electronics like disposable sensors, so they sought out a durable, low-manufacturing-cost memristor instead of the seemingly more logical high-capacity devices of the previous two teams. However, the actual memristive layer is still only 60nm thick, which is fairly consistent with the other two designs as well as with equation 1.

The equation for memristance as derived at HP Labs for their device is as follows:

$$ M(q) = R_{\text{OFF}}(1 - \frac{\mu V R_{\text{ON}}}{D^2} q(t)) $$

(1)

Here, $R_{\text{ON}}$ and $R_{\text{OFF}}$ are the low and high resistances, $\mu V$ is the dopant mobility, and $D$ is the thickness of the memristive material. The $1/D^2$ term is key; to achieve memristance of an observable magnitude, very small device dimensions are required, dimensions that manufacturing technology has reached in only the past few years. This is why it took researchers so long to discover this first device.

Perhaps what makes memristance so exciting is that memristor devices are only starting to be discovered. As stated previously, HP Labs created a metal-oxide-metal memristive device. This device has a half-pitch of just 30nm, compared to DRAM’s half-pitch of 59nm that same year [1]. But that is just the beginning. As shown in Figure 4, starting from a single substrate, multiple arrays of these metal-oxide-metal devices can be stacked one on top of the other, separated by insulating material—a sort of 3D-on-wafer technology similar to buried devices [8]. Such a technique would increase the per-wafer density several fold. Pairing this processing with 3D die-stacking, and we are looking at a storage density several orders of magnitude greater than is feasible with DRAM.

Most interesting, though, is that because these are brand-new devices, we are only just beginning to understand what makes a good memristor and what the limits of memristor scaling might be. Just a few years down the road, the HP Labs team is predicting 5nm half-pitch devices from their process while the Michigan team expects 20nm devices from theirs. On the other hand, we cannot expect 20nm DRAM cells until 2017 [1], and many manufacturing challenges have yet to be solved to realize such devices.

While density is the poster child for memristors, they have other important characteristics. A big plus is that memristors are non-volatile like Flash but obviously with much more density. The HP memristors have already demonstrated retention up to a couple years, and retention near a decade is predicted, more than sufficient for any
memory application. On the downside, writes are a little slow, on the order of nanoseconds, and, sad to say, the extreme density will add to addressing delays. More significantly, endurance has so far been a major problem, with devices only surviving around $10^5$ to $10^7$ write cycles. Even as a last level cache or main memory, these devices would not last longer than a minute or so. Fortunately, the device inventors are optimistic and expect to add several orders of magnitude to the endurance in just a few years.

This combination of extreme density, moderate speed, and moderate endurance makes near-future memristors an excellent candidate for a memory-on-chip system architecture. We call such a memory *resistive random access memory* or RRAM.

3. **RRAM VS. DRAM**

So RRAM sounds very promising, but how does it stack up against DRAM, which is the definitive king of on-stack memory. A summary comparison is given in Table 1. Each point of comparison will be discussed below.

First and most importantly is density. When we talk about stacked memory, capacity is king. The architectural innovations of the past 30 or so years have largely insulated the processor from the latency of main memory, so simply providing as much data as possible is the goal. On this front, RRAM is the clear winner. Looking out to the end of the ITRS roadmap, we expect to have 16nm half-pitch DRAM capacities in the range of seven to ten years. Once again, this is a pretty standard 16ms before requiring a refresh operation. RRAM, in stark contrast, has definitively demonstrated its retention time over very long periods. Some groups have shown successful retention time for years, with anticipated retention times in the range of seven to ten years. Once again, this is not really a fair comparison, but the non-volatility of RRAM is quite big advantage.

Closely related to read speed is readability, the ability of the memory cell to report its state. DRAM is a notoriously weak signaling memory because each DRAM cell holds very little charge with which to change the potential of the bitline. As a result, sense amplifiers likely would not change too much with cell size, as they need to be kept large in order to respond quickly to signals on the bit lines. So overall, we can anticipate that the area efficiency increases with shrinking cell size, another advantage for RRAM.

Now let us consider speed. FaStack 3D Memory from Tezzaron has a round-trip access time of 10ns, including bus latency, addressing, transaction, and data return. Memristors by comparison require approximately 10ns just to write to a single cell, to which bus and addressing latencies must be added. This means RRAM will be much slower than DRAM for writes. Unfortunately, the authors have been unable to locate numbers for RRAM read speed in the literature, but it is reasonable to expect that a read operation will be much faster than a write operation. Unlike DRAM where the small, weak memory cell must drive the read operation, RRAM can rely on large and relatively high-current read drivers to produce a strong probe signal.

Next up is yield. DRAM memories have a very high quality on the order of one defect per thousands or even millions. But to achieve this incredible quality, manufacturers have to play a lot of tricks with ECC, bad row and column replacement, and so on. The bare yield of the DRAM cells themselves is only around 90% [5]. The memristors surveyed vary a bit in yield, but all fall within a range of 80-95%, which is obviously quite similar to DRAM. So in this metric, both technologies perform equally well.

A very important metric for memory is retention time. This is a bit of an apples to oranges comparison because DRAM is volatile memory, while RRAM is non-volatile. But non-volatility is an important advantage of RRAM and so worth mentioning in this comparison. DRAM, under normal conditions (0 to 85°C) can retain its data for a pretty standard 16ms before requiring a refresh operation. RRAM, in stark contrast, has definitively demonstrated its retention time over very long periods. Some groups have shown successful retention for just a few days or weeks. The HP Labs team, however, have memristors that have held their data for nearly two years, with anticipated retention times in the range of seven to ten years. Once again, this is not really a fair comparison, but the non-volatility of RRAM is quite a big advantage.

RRAM has some other advantages. First, as DRAM scaling continues to reduce the size of the memory cell, the capacity of each cell is limited by the physical size of the cell. For RRAM, this is not really an issue. As RRAM cells get smaller, the bit lines get shorter, reducing load and thus required drive power. Sense amplifiers likely would not change too much with cell size, as they need to be kept large in order to respond quickly to signals on the bit lines. So overall, we can anticipate that the area efficiency increases with shrinking cell size, another advantage for RRAM.
Table 1: This table quickly lists the relevant metrics for each memory system and reports the superior option in each metric. See Section 3 for detailed discussions of each.

<table>
<thead>
<tr>
<th>Metric</th>
<th>DRAM</th>
<th>RRAM</th>
<th>Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity (GB/cm²)</td>
<td>46</td>
<td>466</td>
<td>RRAM</td>
</tr>
<tr>
<td>Area Efficiency</td>
<td>Less</td>
<td>More</td>
<td>RRAM</td>
</tr>
<tr>
<td>Write Speed (ns)</td>
<td>10, round trip</td>
<td>10, just one cell</td>
<td>DRAM</td>
</tr>
<tr>
<td>Read Speed (ns)</td>
<td>10, round trip</td>
<td>Uncertain, likely better</td>
<td>RRAM</td>
</tr>
<tr>
<td>Yield</td>
<td>90%</td>
<td>80-95%</td>
<td>tie</td>
</tr>
<tr>
<td>Retention Time</td>
<td>16ms</td>
<td>2yr</td>
<td>RRAM</td>
</tr>
<tr>
<td>Readability</td>
<td>10⁻¹</td>
<td>10⁹</td>
<td>RRAM</td>
</tr>
<tr>
<td>Endurance</td>
<td>10^{10} write cycles</td>
<td>10^6 write cycles</td>
<td>DRAM</td>
</tr>
</tbody>
</table>

nearly an order of magnitude greater than the highest on resistance ($5 \times 10^5$). Such large switching ranges are obviously much easier to process than the minute current and voltage changes in DRAM.

Lastly, we come to endurance. In this metric, DRAM really shines. DRAM cells can endure so many writes, the authors have found that no one even reports endurance results anymore. But we can still come up with a rough estimate. DRAM must be accessed at least once every 16ms for refresh operations, and DRAM modules are known to operate without fail for several years at a time. This works out to an effective endurance of at least $10^{10}$ write cycles. By comparison, endurance has proven to be the Achilles heel of RRAM. The Michigan team reported an endurance on the order of $10^5$ write cycles, far too few cycles, sadly, for RRAM to serve as memory in any capacity within a computer system. Endurance is presently a major show-stopper for RRAM. Of course, this does not rule out RRAM as an option for mass storage like hard disk and USB drives, but such applications are not very interesting. However, Stanley from HP Labs is quite confident in the team’s ability to quickly improve the endurance of their memristors. After all, those devices were the first memristors ever discovered, and a more thorough exploration of the manufacturing options will almost certainly produce devices with greatly enhanced properties, including endurance.

4. RRAM ARCHITECTURES

A quick survey of Table 1 highlights the many advantages of RRAM over conventional DRAM as last-level 3D stacked memory. But before any of these advantages can be exploited, the problem of endurance has to be addressed. But what can we do with RRAM today, even given this handicap? Obviously, RRAM’s niche is write-a-little-read-a-lot applications. The first thing that comes to mind is FPGAs. The very limited number of write cycles, combined with FPGAs’ demand for lots and lots of bits of programmed data, makes for an obvious first RRAM application. For a more general use in a processor, we can consider a hybrid architecture. Such an architecture was proposed for a DRAM/MRAM system in [9]. The basic idea would be to place stable data (i.e. data that is not constantly changing) in the RRAM memory while placing the dynamic data in the DRAM memory. Full consideration of these and similar architectural designs is left to future work.

5. CONCLUSION

In this paper, we have surveyed the existing memristor-based RRAM technologies and drawn some useful conclusions about the high-levels trends in relevant metrics like capacity, speed, and endurance. Our comparison with conventional DRAM revealed that RRAM is a very good choice for future memory designs, but if and only if the endurance problem can be solved. Hopes are high amongst memristor research teams that viable solutions do exist, but for now we must wait and see. But inspite of these limitations, there are several interesting architectural possibilities to consider that can take advantage of these exciting new devices and all they have to offer.

6. REFERENCES