Complimentary Memory Technologies to Reform the Homogeneity, Energy Landscape, and Volatility of Future Memory Hierarchies

Brian Lichtman
Department of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL, USA
Bri353@knights.ucf.edu

Abstract—With the limits of transistor size rapidly approaching and the memory wall approaching us at unfathomable speeds, a need to revisit our existing memory technologies has arisen. Fortunately, there are new non-volatile memory (NVM) technologies on the horizon showing great promise. In this paper we will examine some of the new NVM technologies being developed. We will present the basic characteristics of these new technologies as well as the properties that make them unique and the possibilities for the future they offer.

Keywords—non-volatile; volatile; memory; memristor; RRAM; MRAM; SRAM; STT-RAM; PCM; PRAM

I. INTRODUCTION

In 1965, Gordon Moore, then Director of R&D at Fairchild Semiconductors and later co-founder of Intel, published an internal paper with a prediction that the number of components per chip would double every year [1]. In 1975, after taking into account new data, Moore announced a modification to his perdition at the IEEE International Electron Devices Meeting, lowering his expectation from doubling every year to every other year [1] [2]. Carver Mead later coined his perdition as Moore’s Law which has been the standard industry has strived to achieve ever since [1].

While this perdition has stayed true since its formulation and may continue to stay as such for few more years to come, several limiting factors have begun to prevent us from continuing to scale memory at this pace. The first of which is physical [3] [4] [5]. Manufacturing components with transistors at the 14nm scale has proven to be quite difficult. Manufacturing such small size components and such high densities requires new manufacturing processes to be able to provide near absolute precision.

Additionally, transistors can only be made so small before quantum mechanical effects begin to have noticeably large effects. Quantum tunneling is one of these quantum mechanical phenomena. In quantum tunneling, electrons are able to pass through an electric field barrier which would normally not be possible. [6] states that quantum tunneling occurs when barriers reach a thickness of around 1 – 3nm or smaller. Device manufactures have been seeing these effects for a while already and they have only gotten worse.

Quantum tunneling is a source leakage current that plagues the shrinking transistor causing not just a significant drain in power but also the heating issues experienced by modern semiconductor technologies [7].

Nonetheless, quantum effects aren’t the only hindrance. Memory technologies that make use of capacitors for bit storage suffer from non-quantum based leakage current plights. Specifically, they deal with dielectric leakage current caused by manufacturing defects and/or material imperfections in the dielectric material insulator [8]. This in turn also causes these memory technologies to rely on a refresh signal resulting in even more power consumption and heat accumulation. As transistors become smaller, allowing for denser concentrations, the problem only increases [9] [10] [11] [12]. Intel’s 14nm Broadwell release suffered from these obstacles causing its release to be delayed from 2014 to 2015 [13].

Samsung and Intel’s latest CPUs, the Broadwell and Exynos 7 Octa, bring us to a 14nm achievement through the use of FinFET technologies [14] [15]. Intel’s CFO, Stacy Smith announced that they expect 10nm CPUs sometime in the near future. Smith also announced that Intel has begun taking what he calls “an early look” into 7nm technologies and that “…our early look tells us that we can continue to bring down the cost of transistors with historical rates of improvement”, after which is the five nanometer milestone. Smith wasn’t able to comment directly, but was able to defer to Intel CEO Brian Krzanich with the comment “Brian would say that the difference between seven and five is that it’s only going to get harder, and there will be fewer and fewer people able to do it [4].”

Based on current semiconductor manufacturing trends, Hesseldahl states we can probably expect 5nm technologies around late 2020. Then assuming a two-year product cycle, we will probably be able to make it to 2022. But after that, “some three years short of its 60th anniversary, [it is predicted] Moore’s Law will reach its logical end [4].”

As stated by Mushiva in [5], “At 5nm the laws of physics turn the chip into a frying pan and quantum mechanics at that size scrambles the atom and disrupts information flow (ability for signals to travel through a logic gate on a silicon wafer in a coordinated fashion). So Moore’s law falls short at postulating
leaps in computational power primarily because the axiom is untenable at a certain size and that limit is fast approaching.”

However, the end of Moore’s Law is not as bad as it may seem. There are new non-volatile memory (NVM) technologies currently being developed which show a lot of promise for the future of memory. In this paper we will take a look at some of the new NVM technologies.

II. MEMORY TECHNOLOGIES – AN OVERVIEW

Over the years, a multitude of memory technologies have come about. These technologies come in two forms, volatile, and non-volatile. Modern commercial PCs use Static Random Access Memory (SRAM) for their caches and Dynamic Random Access Memory (DRAM) for their main memory. These are both volatile memory technologies which are discussed further in Section A. Additionally, the modern PC make use of non-volatile memory technologies for its long term storage. Such devices as Hard Disk Drives and Flash based storage drives will be explored in Section B.

Nonetheless, these technologies are not the end solution. There are many complications with these technologies causing research to lead us towards a new direction. Lastly, Section D, Section E, and Section F we will examine some of the emerging NVM technologies that are being considered as potential replacements to the current memory standard. These technologies include memristors, phase change memory (PCM) sometimes referred to as PRAM, and spin-transfer torque magnetoresistive random-access memory (STT-RAM). A taxonomy of the current as well as emerging memory technologies can be seen in Fig. 1. Additionally, while reading the next sections, Fig. 2 provides a contextual refer to where the current as well as emerging technologies map to today’s memory hierarchy.

A. Volatile Memory

The use of volatile memory has become quite popular as it offers high speeds with an almost unlimited endurance [16]. See TABLE I for more details. DRAM is the most common form of volatile memory. This is mainly because of its size, $6 \times 10^2$, and its relatively fast read/write rates of $10 - 60\text{ns}$.

Fig. 1 Memory Taxonomy [11]

Fig. 2 Memory Technologies [19]
However, one of the limitations we see in DRAM is that it not getting much faster anymore. At least when it comes to the energy cost needed to increase the RAMs speed and capacity [11].

But the biggest problem yet is the amount of power wasted by these chips. A cell of DRAM is made up of a NMOS transistor and a capacitor [17]. See Fig. 3. Because the data is stored in the capacitor, in order for a DRAM cell to maintain its state, a refresh signal must constantly be applied. This is because conductors suffer from dielectric leakage current caused by manufacturing defects and/or material imperfections in the dielectric material insulator [8]. This creates an imperfect capacitor resulting in this requirement.

Additionally, transistors themselves suffer from a different leakage current. Specifically, they suffer from the Quantum Mechanical effect of quantum tunneling. In quantum tunneling, electrons are able to pass through a normally unpassable electric field barrier. This is because in quantum mechanics, movement of particles is probabilistic. In quantum physics, particles are thought of to move with wave-like probabilities. So, as the wave approaches the barrier, although the probability of a particle crossing an electric barrier is small, it is possible even if not probable. Fig. 4 provides a visual representation of this phenomenon.

As transistors become smaller, the electric field barrier becomes smaller and the probability of a particle crossing the electric field becomes larger. At a thickness of around 1 – 3nm or smaller, the effects of quantum tunneling become noticeable [6]. Then, allowing for denser concentrations, the problem only increases. As more and more transistors are put on to a component, the amount of leakage current rises as well as the amount of heat dissipated [9] [10] [11] [12] [18] [19].

In addition to DRAM, we all have SRAM, the second most widely used volatile memory and provides for even faster speeds than DRAM. This is because SRAM is normally made up of a pair of cross-coupled inverters acting as a flip-flop. Because SRAM doesn’t rely on a refresh single to maintain its state, can process data with read/write rates of about 1 – 2ns.

SRAM is usually implemented in the form of 4 to 6 transistors. This can be seen in Fig. 3. As just mentioned though, for DRAM, small transistors suffer from leakage current from quantum tunneling effects. Well, with SRAM, 4 – 6 times the leakage current from quantum effects is realized because they are manufactured with 4 – 6 times the number of transistors per cell of memory. This also leads to the other downfall of SRAM in that it is quite large with a size of 50 – 120F^2 [3] [10] [19] [20] [21] [11] [22] [23] [21]. See TABLE I for more details about the specifications of SRAM.

By moving away from these volatile memories which dissipate so much unutilized power, we can use devices will be able to make much better use of the power they are provided as well as dissipate less heat. This would significantly impact the consumer electronics market. Specifically this will have an impact on battery powered mobile devices such as cellular phones, laptops and tablets. These devices would see a drastic improvement in their battery life.

B. Non-volatile Memory

Today, this is the type of memory tends only to be used for long term storage. Being that the memory is non-volatile means that when you remove power, the memory persists [10] [16]. This is very advantageous considering that offline storage is the purpose of this technology.

Hard disk drives are probably the most common form. This is specifically so because of their reliably and high endurance. However, their speeds can be quite slow with access latencies in the millisecond range [3] [24]. This is why recently, many consumer electronics have begun replacing disks with flash based equivalents.
Flash technologies are the second most popular NVM technologies in production today. They come in the form of NAND and NOR memory. The reason for its popularity is its ability for high capacities, small size footprint ($2 - 10^2$), and its relatively low cost per byte. However, it suffers from speeds much slower than DRAM.

However, being that Flash memory is slower than DRAM doesn’t mean the memory technology itself is inherently bad. This memory is still quite an improvement over older technologies such as disk. Compare the speed a traditional disk drive to that of a solid state drive. NAND technology read/writes are in the microsecond as compared to disk’s millisecond access time [3]. Actually, the real concern with Flash based technology is its endurance. As Flash based memories become denser, their endurance decreases. Current high density Flash memories only have an average write endurance of $10^4 - 10^5$ [3] [10] [22] [23]. A few authors have cited endurances for NAND as high as $10^6$ [10] [21]. [10] cited NOR write endurances as high as $10^7$. These write endurances are unacceptable. The write demand needed is not fulfilled by the provided endurance of these technologies.

C. Emerging Technologies

Although currently used NVM technologies have some fallbacks, there is still quite a lot of research being done on it. The findings that have come about are quite intriguing. New NVMs have been discovered recently. These include resistive RAM memory such PCM and memristors and magnetoresistive RAM such as STT-RAM. These technologies show a lot of promise although most are still in the early stages of development. Table I shows a comparison of some of these new technologies.

As can be seen, some of these new memory technologies provide similar if not sometimes better memory densities over current DRAM and Flash based technologies while consuming much lower amounts of power to read, write as well as to maintain data. This is due to the fact of them being NVM technologies which don’t require power to maintain state.

Viking Technology has actually realized the benefits of using some of these NV memory technologies and actually has already begun production of non-volatile DIMMs called the ArxCis-NVTM. Their goal was to create a memory replacement that would allow for almost instant recovery time on their clients’ computers after a power outage. With NVM, they realized they could create data centers that didn’t have to have servers that backed up memory upon power failure. They expressed how using volatile memory could cause hours of downtime while servers restore their data into RAM upon a reboot. But, with NVM, the memory doesn’t need to be restored. This also brings about massive power savings for the duration of time that the system is spending restoring instead of actually working [16].

The goals of emerging NVM research are as follows. First, in order for these technologies to be deemed advantageous, they

<table>
<thead>
<tr>
<th>Attributes</th>
<th>SRAM</th>
<th>DRAM</th>
<th>Flash NOR</th>
<th>Flash NAND</th>
<th>Non-STT MRAM</th>
<th>STT-RAM</th>
<th>PCM</th>
<th>R-RAM (Memristors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory Cell Factor ($F^2$)</td>
<td>50-120</td>
<td>6-10</td>
<td>10</td>
<td>2-5</td>
<td>16-40</td>
<td>4-20</td>
<td>6-12</td>
<td>&lt;4</td>
</tr>
<tr>
<td>Read Time (ns)</td>
<td>1 - 2</td>
<td>10 - 60</td>
<td>10</td>
<td>50</td>
<td>3-20</td>
<td>2-20</td>
<td>20-60</td>
<td>10 – 50</td>
</tr>
<tr>
<td>Write/Erase Time (ns)</td>
<td>1 - 2</td>
<td>10 - 60</td>
<td>$10^4$-$10^7$</td>
<td>$10^4$-$10^8$</td>
<td>3-20</td>
<td>2-20</td>
<td>50-120</td>
<td>10 - 100</td>
</tr>
<tr>
<td>Endurance</td>
<td>$10^{16}$</td>
<td>$10^{16}$</td>
<td>$10^3$</td>
<td>$10^4$</td>
<td>$10^{15}$</td>
<td>$10^{10}$–$10^{12}$</td>
<td>$10^{15}$</td>
<td></td>
</tr>
<tr>
<td>Power Consumption – Read/Write</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Med/High</td>
<td>Read Low / Write High</td>
<td>Med</td>
<td>Low</td>
</tr>
<tr>
<td>Power Consumption – Other than R/W</td>
<td>&gt;1pJ</td>
<td>2pJ</td>
<td>Unspecified</td>
<td>10nJ</td>
<td>Unspecified</td>
<td>0.02pJ</td>
<td>1nJ</td>
<td>2pJ</td>
</tr>
<tr>
<td>Embedded/SoC Friendly</td>
<td>Y</td>
<td>N (Thermal)</td>
<td>N (Thermal)</td>
<td>N (Thermal)</td>
<td>Y</td>
<td>Y</td>
<td>N (Bi-polar)</td>
<td>Y</td>
</tr>
<tr>
<td>Data Retention</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Data Obtained From [3] [10] [19] [20] [21] [11] [22] [23].
must meet the following criteria. Second, they must have equal or lower power requirements compared to current conventional technologies. Third, they must have the potential for higher speed reads and writes. See Fig. 5. Fourth, they need to allow for higher densities of data storage. Last, they need to provide a smaller form factor than the current memory technologies.

### D. Memristor Technology

HP was the first to discover how to actually make a memristor in the lab and recognize what it was. Although it was hypothesized a long time before in 1971 by Leon Chua nobody could produce a useful physical model or an actual memristor. This is specifically because nobody conceived on a basic element that would change state and hold a memory based on an applied voltage. [9] [12] Where it fits in can be seen in Fig. 6.

The key significance of the memristor is its ability to remember state. By applying a large electric field on the memristor, you can change the resistivity of the memristor and use this knowledge to represent bits. However, even more important that that is the fact that memristors change state very fast. Current averages are from 10 – 100ns. It is hypothesized that they will be able to change of a bit at speeds of about 1 nanosecond based on experiments done at HP Labs [9] [12]. Because not a lot of power is required to change the state of a memristor, about 2pJ, this could prove to be an invaluable new element that could potentially replace not just memory but the transistor as well. [9]

Lastly, memristors interesting property of letting them maintain state also allows them to be able to store not just data but logic as well. Memristors can represent implication logic. With memristors, HP sees the potential of replacing logic circuits with memristors allowing for faster microprocessors than were previously thought possible. This is because implication logic can represent NOT logic in fewer instructions [9] [25]. Fig. 7 provides an example of using memristors to provide implication logic.

---

### E. Spin-Transfer Torque RAM

One of the most promising technologies coming about is STT-RAM. This memory takes advantage of magnetic tunneling, through the use of a Magnetic Tunnel Junction (MTJ) to store the state of bits instead of the energy produced from by electric field. In STT, there are two layers, a free layer and a fixed layer. The fixed layer is fixed with a magnetic polarity in one direction. The free layer can have its polarity switch by applying a field through the MTJ. Based on the direction of the free layer, the STT cell will experience different levels of
resistivity providing a means to represent a logic 0 or 1 [18] [19]. Fig. 8 provides a model of a STT cell.

Because the memory is non-volatile and therefore power is not required to maintain the state of the cell tremendous power savings can be realized.

Some of the key benefits of STT memory technologies include low power reads and a small size of 4 – 20 F². Additionally, reads have been recorded to be fast, with speeds of 2 – 20ns [11] [18] [19].

However, they face one issue, writing memory can be much more power consuming than traditional DRAM technologies. Additionally, stacking these memories has proven to be difficult. Nonetheless, recent research has found possibilities for improvement. [18] found new ways to use STT-RAM to be able to read and write data somewhere between 3 - 10ns. This is faster than current DRAM.

F. Phase change memory

The memory is another possible technology. PCM uses the melting and crystallization of a semiconductor to represent the states of a bit. By applying heat at a temperature below the melting point (~600° C) but above the crystallization threshold (~300° C) to the chalcogenide glass, these states can be attained [19].

When the material is in a crystalline state, it has a low resistivity and when in an amorphous state has a high resistivity. When in the crystalline state, the PCM is said to be set with a represent logic of 1. When in the amorphous state, PCM is said to be reset with a represent logic of 0 [3] [19] [26]. See Fig. 9.

Some of the advantages of PCM includes its ability for relatively fast reads for around 20 – 50ns. Additionally, PCM has analog state holding properties allowing once cell to represent multiple bits just as memristors can. The last benefit is that its size is also quite small with being 6 – 12F². However, PCM faces some large challenges compared to the previously mentioned NVM technologies [19] [21].

The immense challenges for this technology are the fact that writing has an overwhelmingly large write latency compared to the others. As long as 120ns. Additionally, writes have been recorded to take as much as 1nJ per cell. Additionally, the heat produced to change states can drift between cells causing memory to become corrupted. Lastly, write times are inconsistent and cannot be predicted [19] [26].

III. Future of NVM Memories

The movement of these memory technologies into consumer products is not too far away. Of these technologies, we see the memristor and STT based memories as having the largest potential. Everspin Technologies has already begun production of STT MRAM modules [27].

Once memristors are able to consistently hit 1ns read/write times with memory sizes as large as current SRAM, we see this as a viable L1 cache replacement. However, new materials for memristors need to be found that can provide endurance of 10¹⁵ in high density applications and not just in small cell combinations in labs. However, if only endurance is able to be realized in the near term, a hybrid cache architecture making use of SRAM for the L1 cache and memristors for L2 and L3 cache would still make for significant improvements to power usage, heat output and speeds.

Next we also see the potential for STT memories to replace current main memory. Given its ability to perform high speed reads in conjunction with its longer write times, we believe that STT memories will fit perfectly here into the memory hierarchy. This is because although STT has larger write times and higher write power consumption than memristors, main memory does not require the same write times as a cache and will tend to have many less writes as well over your caches. Therefore, we believe that the ultra-low power reads provided by STT will work exceptionally well at this stage of the memory hierarchy. However, should research be able to improve write times as well as write power consumption, we also see the potential for STT memory technologies to be used for off chip caches as well as L3 caches.

When it comes to long term storage, we believe that either STT memory or memristor technology would make a great alternative to the current NAND based hard drives. This also assumes that the endurance of these technologies has been improved to consistently achieve the maximums listed in Table I. Just like with main memory, off chip storage tends to be more read intensive than memories closer to the top of the hierarchy. For this reason, STT is a great solution when it comes to power savings. However, when it comes to density, we see memristors being a clear winner. Because of memristors ability to store multiple bits per cell at the expensive of time, memristors provide for the ability to have ultra-high density memories as compared to other technologies. However, the power

![Fig. 9 PCM State Transitions [19]](image-url)
implications of these designed has not thoroughly been researched and will therefore need to be looked into.

Lastly, we see the potential for memristor based technologies to replace current NOT based logic designs. Again, assuming that 1ns read/write rates are fully attained, implication based logic could provide for a much faster logic devices such as CPUs and GPUs. We believe if CPUs and GPUs move forward into Heterogeneous System Architectures as described in [28] by AMD, migration to an implication based system would be quite easy.

IV. CONCLUSION

In conclusion, we see great potential for both STT and memristor based technologies. Although PCM is an interesting technology, it does not appear to be to meet the requirements of the consumer market. Specifically, cost reduction inabilities has been a large factor [29].

In the end, it will all come down to speed and cost per bit. The largest concern that seems to be on consumer markets mind is how costs of new memory technologies will be affected once we finally reach the limits of Moore’s Law [30].

REFERENCES


