STUDY OF NEW MEMORY CIRCUIT TECHNOLOGIES BY SIMULATION

Zoltán Szandtner; assistant; Dennis Gabor College; szandtner@gdf.hu

István Vári-Kakas; professor; Dennis Gabor College; vari@gdf.hu

Keywords: memory simulation, memory modelling, cache, SRAM, non-volatile memory (NVM or NVRAM), STT-RAM, Destiny

1. ABSTRACT

The increasing parallelization in general purpose computing has led to ever greater demand from the memory system. The memory bandwidth and latency are both important parameters for high performance computing. While bandwidth was steadily increased by architectural means, latency depends primarily on technology instead. In this article we take an overview of mature, promising memory technologies and compare representative specimens with simulation tools. The comparative results show the current performance of production ready devices and pinpoints the most favorable application area of each technology.

2. INTRODUCTION

The development of semiconductor type circuits for the operational memory of computers has been ongoing for more than fifty years. Allegedly Bob Norman already proposed a purely semiconductor memory in 1961 [1] — the first such *static* memories were produced at *Fairchild Semiconductor* — while *dynamic* memory was invented by Robert H. Dennard at IBM in 1970.

The working principle of the two memory types, SRAM (*Static Random-Access-Memory*) and DRAM (*Dynamic Random-Access-Memory*) has remained unchanged ever since, with SRAM using a pair of inverters in a feedback loop and DRAM using a capacitor to store a single bit of information (Figure 1).

The circuits making up the cell are sometimes also called the *microstructure* of the memory. Along with the parameters of the discrete elements — storage and access transistors, trench capacitors or novel elements in case of new memory types — contained within, it dictates several key figures of merit: cell *switching-time, dissipation,* schedule of the operation cycle. It's important to note that *switching time only* refers to the time necessary for the cell to switch states and does not account for the necessary operations to retrieve the data stored within.

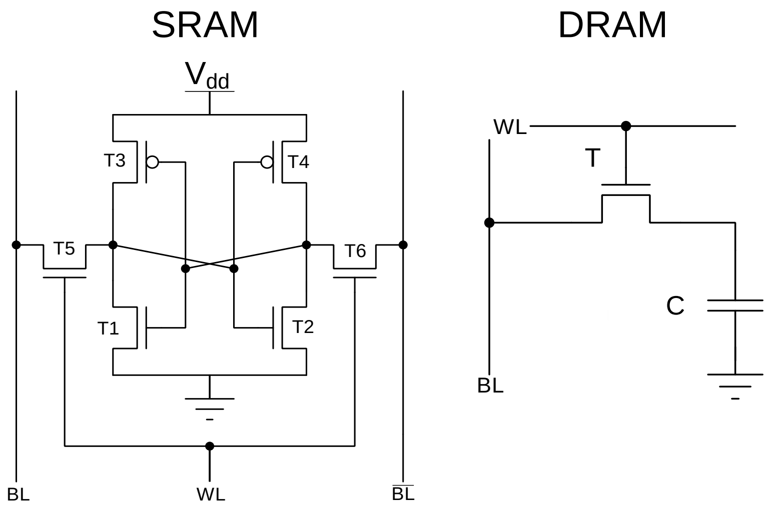

Figure 1.
SRAM and DRAM cells storing 1bit

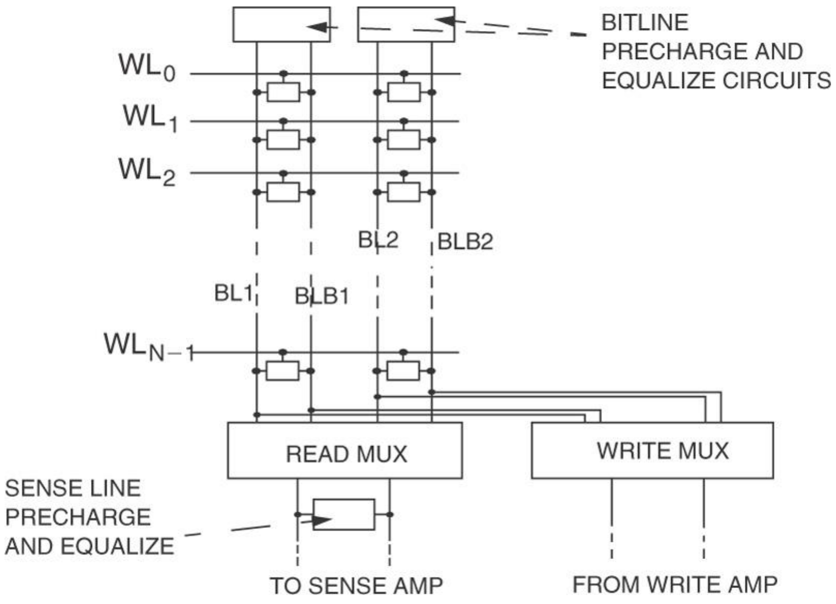
BITLINE PRECHARGE AND EQUALIZE CIRCUITS
WL0
WL1
WL2
BL1
BLB1
BL2
BLB2
WLN−1
READ MUX
WRITE MUX
SENSE LINE PRECHARGE AND EQUALIZE
TO SENSE AMP
FROM WRITE AMP

Figure 2.
SRAM macrostructure [2]

For SRAM the big number of elements – four storage and two access transistors — leads to a bigger cell size on the wafer and therefore lower capacity compared to DRAM which only needs a single transistor and a trench capacitor for every cell. This also applies to dissipation, since SRAM has more active elements than DRAM. On the flipside, the switching time of the two technologies is drastically different with SRAM switching times on the order of a few picoseconds, whereas DRAM cells usually take several nanoseconds to achieve this. Furthermore, DRAM needs to be periodically refreshed due the leakage currents slowly discharging its capacitor.

While elements of the cells have undergone developments of their own as dictated by new manufacturing technologies, the microstructure itself has remained the same. Structural developments instead happened in what is called the *macrostructure* of the memories.

Except for very small caches, the most effective arrangement of memory cells is grid array of wires with each intersection containing a single cell (Figure 2). These arrays form the basic building block for the memory macrostructure and its qualities are the root cause of the difficulty of memory architecture optimisation.

The bigger the array the more effectively wafer area is used, however this leads to longer wires with worse resistance and capacitance values. Address decoding and output multiplexing are other areas where similar trade-off optimisations can take place. In his work, Jacob proved that memory systems parameters can’t be fine-tuned in isolation and instead a holistic approach is necessary [3]. The sum of micro- and macro-structure qualities are the critical figures of interest for the memory system designer. In our study we focused on three such critical parameters: *access time, bulk dissipation* and *capacity.*

The rest of the article is structured as follows: we overview the makeup and principles of operation of new memory types at the cell level, we explain the rationale behind focusing on macrostucture simulations and present the setup of the simulation environment and comparatively discuss the obtained results.

3. NEW TECHNOLOGIES UNDER DEVELOPMENT

Without exception the major memory technologies currently under research or development are based on variable resistance elements. As a consequence dummy cells are typically used in the memory array to create a reference voltage for the sense amplifier’s comparison, similar to how modern DRAM operates. The persistent nature of these technologies means that they can be used for both operational and storage class memory applications. Due to their better write endurance, they might even displace *flash* when the later ceases to scale to smaller technology nodes.

3.1. Magnetic Memories: MRAM, STT-RAM, SOT-RAM, and MeRAM

The new MRAM (*Magnetoresistive RAM*) memories use the same magnetic resistance effects that are used by modern HDD (*Hard Disk Drive*) read heads [4]. Their storage element is the *Magnetic Tunnelling Junction* (MTJ), a nanolayer structure consisting of two magnetic layers with an insulating tunnel barrier sandwiched between (Figure 3). Of the ferromagnetic layers, the strongly magnetised is called the fixed or reference one, the other, with lower, changeable magnetisation, the free layer.

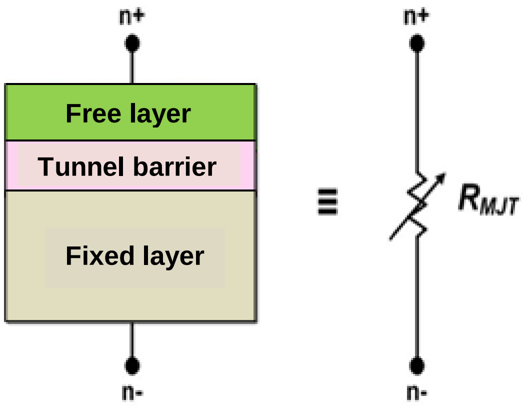

Figure 3.
MTJ structure

Resistance depends on the relative magnetisation direction of the free and fixed layers, low in parallel, high in anti-parallel (Figure 5). The memory cell is practically the same as DRAM's with the distinction that refresh is not necessary since the storage element is persistent (Figure 5).

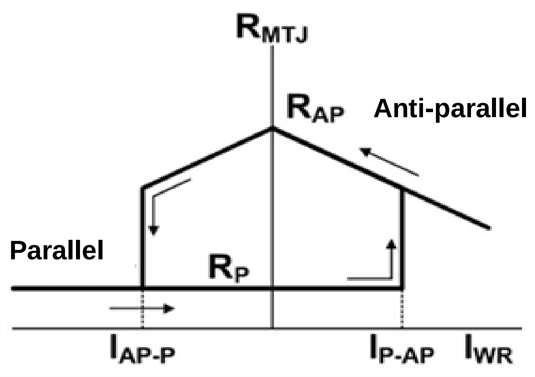

Figure 4.
MTJ resistance states depending on write current

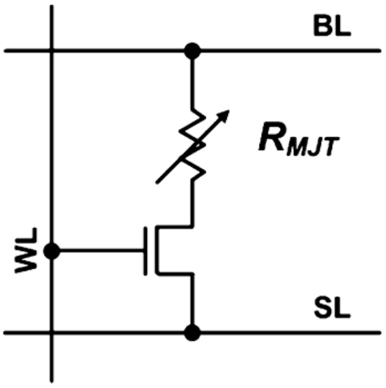

Figure 5.
MRAM cell (WL = Write Line, BL = Byte Line, SL = Source Line)

STT-MRAMs use *Spin Transfer Torque* (STT) to toggle free layer magnetisation. The phenomenon was independently discovered by Slonczewski and Berger in 1996 [5] [6]. The electrons passing through the fixed layer become spin polarised which can be retained by a properly chosen barrier so it exerts torque on the free layer's magnetisation. A sufficiently large current will switch magnetisation. While highly scalable, the writing method still uses high currents, and subsequent barrier breakdown limits write endurance. The technology found a small niche in special non-volatile memory applications where its higher write endurance and low latency allowed it to displace *flash* memory. At current production pitch widths, its capacity is still too low to compete with *flash* in storage class and with DRAM in main memory applications.

There are several improved magnetic memory technologies under development that can be considered direct descendants of STT-RAM. Their common feature is the use of further novel magnetic phenomenon to alleviate the high write energy problems of STT-RAM. *SOT-RAM* utilizes *Spin Orbit Torque* effects in addition to STT and separates the write and read paths, leading to a three terminal device [7]. *Magnetoelectric RAM* (MeRAM) uses an electric field to lower the anisotropy of the free layer, thereby lowering switching energy [8]. These new technologies are in their early experimental phase and promise switching times comparable to SRAM.

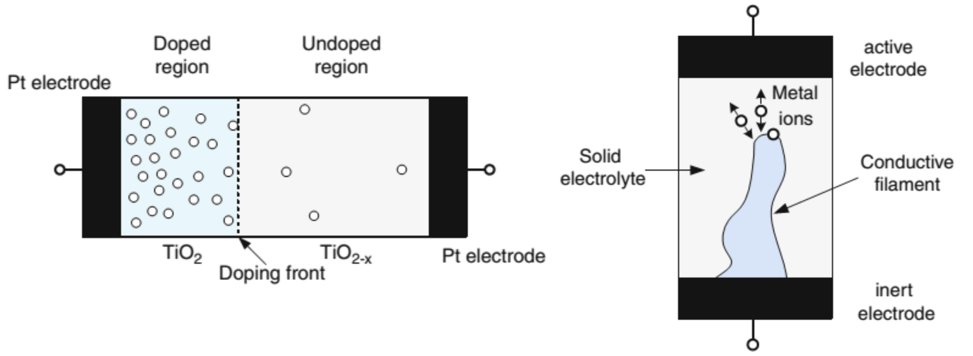

Figure 6.
RRAM & CBRAM storage elements [9]

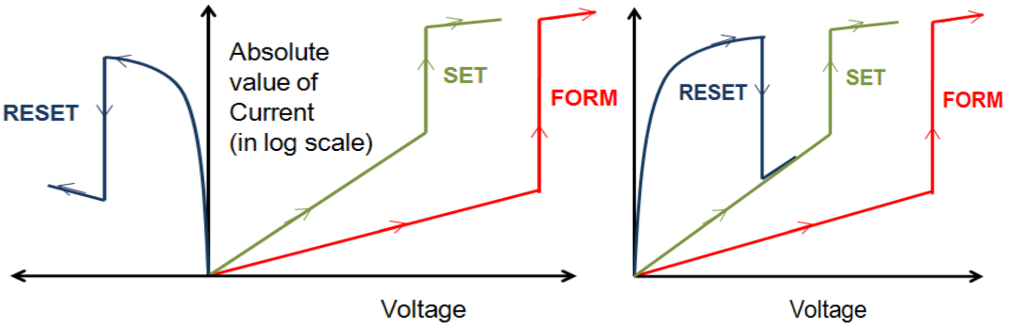

Figure 7.
Memristor write operations for bi- and unipolar devices [11]

3.2. Resistive Memories: RRAM, CBRAM, and PCRAM

The other big family of new devices use different physical phenomenon instead magnetic-resistance to affect resistance [9]. *Resistive RAM* (RRAM) and *Conductive Bridging Memory* (CBRAM) have very similar characteristics, but their switching mechanism is slightly different. Both involve the creation of an induction filament between electrodes, but in RRAM the channel consists of oxygen vacancies in the dielectric body, whereas in CBRAM it is the material of the electrode itself that dissolves into the electrolyte (Figure 6).

After a forming operation creates the initial conductance channel, high-current set and reset operations can increase and decrease conductance (Figure 7). The resistance of the element can be read at a lower voltage. The existence of such devices was predicted by Leon Chua in 1976 [10]. Creating a relation between charge and flux, in essence “remembering” the current that flowed through them he dubbed the elements *memristors.* The term therefore is often used in the resistive memory literature.

Until recently *Phase Change Memory* (PCRAM or PRAM) had little to offer compared to the previous two technologies. Instead filament conduction, resistance change is achieved by changing the crystalisation structure, that is the *phase* of a chalcogenide material, similar to those employed in CD-R and DVD-R disks. Using a heating element, the material is heated, then allowed to recrystalize, heating profile dictating eventual conduction properties (Figure 8).

Compared to competing memory technologies, the melting and crystalisation processes are woefully slow and produce great waste heat. However, recent discoveries of resonant bonding phenomenon in crystalisation might give the technology not only a new lease of life, but allow the creation of elements with similar performance to SRAM [13]. For the moment not even experimental devices have been created, since research is still in the material study phase.

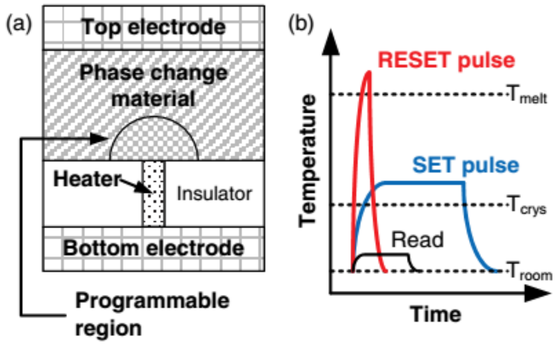

Figure 8.
PCM element cross-section (a) and heating profiles (b) [12]

4. SIMULATION RESULTS

4.1. Preliminaries

At the moment, we consider cell level simulation comparison to be exceedingly difficult as the differing modelling assumptions would need to be matched in a single framework. In a previous work [14], we replicated the cell level measurements of a comparative STT-RAM study by the University of Virginia [15]. We used the built-in models of NVMSpice [9], since reproducing the results in this framework would have allowed not only STT-RAM but cross-technology comparisons. However our findings significantly differed from the reference values (Figure 9). Fast switching only occured at high voltage, close to barrier breakdown (3.5 V). By comparison, the reference model achieved ~10 ns switching time at 1.1 V. We eventually identified the cause of the widely diverging results in the respective modelling assumptions by the authors of the reference study and the measurement environment.

Therefore we focused on using cell-level data from existing, tested devices. We conducted our research on a higher, macrostructural scale to investigate how the technologies fare when the effects of the interconnect circuitry take effect. To get as close to life results as possible we focused our simulation on cache memory architectures.

4.2. The Simulation

For main memory simulation the use of more complex simulation tools that take dynamic system level behaviour into account are necessary. Existing such tools are not yet broad enough in scope for cross-technology comparisons [16]. There are promising future candidates in development though [17].

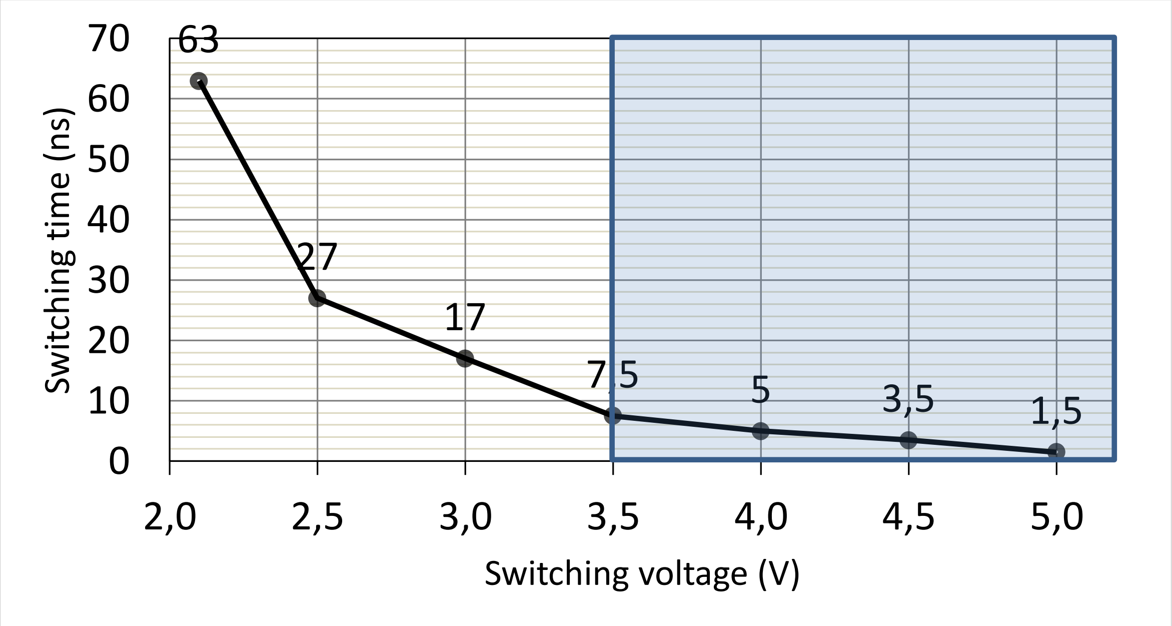

Figure 9.
NVMSpice simulation of an in-plane STT-RAM cell

The traditional simulation tools for macro-scale memory design and evaluation are descendants of *CACTI* developed by HP [18]. For our measurements we chose *Destiny* [19], a newer software jointly developed by Oak Ridge National Lab, Penn State University, and UCSB. It unifies and extends the abilities of *CACTI 6.5* [20], *CACTI-3DD* [21], and *NVSim* [22]. For our purposes the important feature is that the tool comes with *built in discrete models* for STT-RAM, memristors and PCRAM cells in addition to the traditional SRAM and eDRAM (*embedded DRAM*) ones.

Inputs consist of basic *memory cell parameters* — cell *area* expressed in pitch width (F) and *aspect*, *switching times* and *energies* — and *macro parameters,* — total *capacity*, *word width*, *associativity* — that define the overall chip level architecture.

For memory technologies with small cell sizes, the space and energy expenditure of the interconnect network and its various repeaters can dominate the design. In some optimisation cases, bigger capacity memories might have smaller footprints as the ideal interconnect to cell area design might be more compact than that of a smaller capacity memory. Therefore for our simulation we chose a big-cache scenario where interconnect circuitry delays and energies starts to dominate, and the lower cell size of the new technologies could grant them an advantage over SRAM (Table 1).

Since the necessary models for advanced magnetic memory types are not yet available, we chose the mature STT-RAM technology for our comparison. From the memristive memories we picked the PCRAM technology. A conventional DRAM-like array was used instead crossbar for PCRAM to ensure an even playing field with conventional memory types as we focused more on operation speed instead capactiy.The 45 nm process node was chosen as all technologies were successfully produced at the pitch width.

We used conservative cell parameters (Table 2). For STT-RAM we used data on existing devices compiled by Yiren Chan [23]. For PCRAM we chose a typical device whose values were already supplied with Destiny [24].

Capacity	Word width	Associativity	Wire model	Optimisation
1 MB– 32 MB	512 bit	16	Global Conservative	Write EDP

Table 1.
Macro parameters

Cell	SRAM	STT-RAM	PCRAM
Size [F^2]	146	15	4
SET time [ns]	*auto*	20	150
RESET time [ns]	*auto*	20	40
SET current [µA]	*auto*	75	150
RESET current [µA]	*auto*	75	300
Read energy [pJ]	*auto*	-	20
Read power [µW]	*auto*	30	-
Process node [nm]	45	45	45

Table 2.
Cell parameters

Unlike the non-volatile technologies, the parameters of an SRAM cell can not be assumed to be static, as it's primarily driven by bitline and wordline length, themselves dictated by the arrangement of peripheral circuitry and mat size. The complex optimization of this multi-parameter setup was the original intent of CACTI tools. SRAM performance is therefore recalculated for each design from underlying CMOS technology parameters. We used a standard CACTI values with the fast switching HP model [25].

The more economic use of wafer space by the new technologies is apparent at first glance (Figure 10). If we replaced a 2 MByte SRAM cache with an STT-RAM one using close to equal area, a bigger 16 Mbyte capacity cache could be achieved with a presumed better hit-miss ratio.

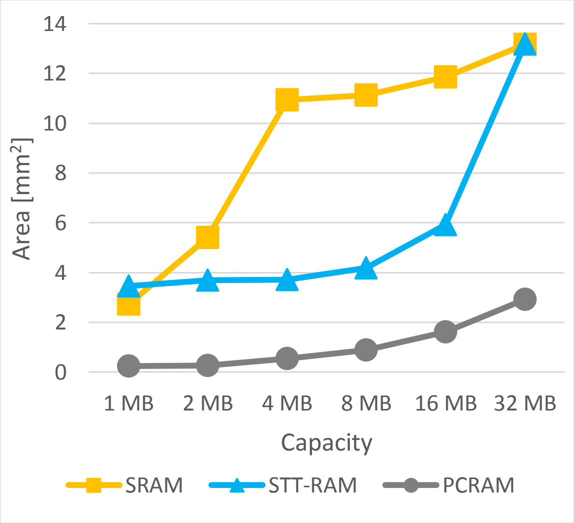

Figure 10.
Chip area vs capacity

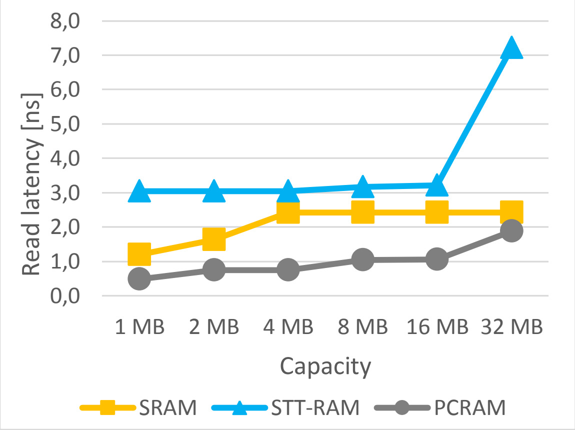

Figure 11.
Read latency vs capacity

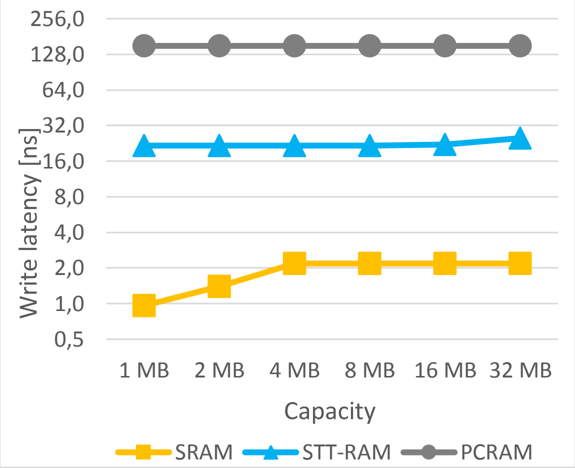

Figure 12.
Write latency vs capacity

However to unequivocally perform better, the new cache would also have to have at least comparable latencies. Therefore, when evaluating the results, the most important figures are the read (Figure 11) and write latency (Figure 12) ones.

4.3. Discussion

Unfortunately neither of the new technologies is yet fast enough for this, due their lesser write latencies. While the latency of SRAM initially steadily increased by interconnect related latencies, the trend settles down. This is because interconnect complexity does not rise at the same rate as capacity does, and for SRAM latency is primarily dictated by the later. Although interconnect latency also influences the other technologies, in the case of STT-RAM and PCRAM their switching time dominates as it's almost a magnitude greater than interconnect introduced latencies. This contradicts the finding of Smullen et al. [15] that STT-RAM should be able to replace SRAM in large caches. However, in their study they used parameters from an STT-RAM cell expressly optimized for cache memory instead existing devices under production like we did. The very good read performance of PCRAM was surprising as most of the literature focuses on its failings in write-performance.

5. CONCLUSIONS

In this article we first overviewed the cell level operation of new memory technologies and highlighted their particularities based on existing literature. Our previous work showed cell level comparison to be unviable, therefore we used the *Destiny* simulation software to compare

SRAM, STT-RAM and PCRAM at the macro-structure level. Running a big-cache scenario we could draw several conclusions. We focused on cache applications since memory cell size and switching characteristics are comparable to SRAM.

In terms of raw switching performance — read and write latency — current STT-RAM devices don't yet measure up to SRAM, so even with its lesser cell footprint STT-RAM can't displace SRAM as the primary cache technology. As expected PCRAM couldn't compete with either technology in terms of writing speed, however if used as a read-only memory it not only competes, but its superior cell size makes it a superb firmware store.

As passive dissipation is a critical design parameter, its further study should also be investigated. The examination of memristive memories for using them in storage devices is also promising because of their great cell density.

6. REFERENCES

[1] G. Moore, Interviewee, *The Fairchild Chronicles.* [Interview]. 1995-2004.

[2] B. Jacob, S. W. Ng and D. T. Wang, Memory Systems: Cache, DRAM, Disk, First ed., San Francisco, CA, USA: Morgan Kaufmann, 2008.

[3] B. Jacob, "The Memory System: You Can't Avoid It, You Can't Ignore It, You Can't Fake It," *Synthesis Lectures on Computer Architecture,* vol. 4, no. 1, pp. 34-39, 2009.

[4] C. P. Lacaze and J.-C. Lacroix, Non-volatile Memories, London SW19 4EU: John Wiley & Sons, Inc., 2014.

[5] J. Slonczewski, "Current-driven excitation of magnetic multilayers", *Journal of Magnetism and Magnetic Materials,* vol. 8853, no. 96, p. 53–58, 1996.

[6] L. Berger, *"Emission of spin waves by a magnetic multilayer traversed by a current",* Physical Review B-Condensed Matter, vol. 54, no. 13, p. 9353–9358, 1996.

[7] S. Fukami, H. Sato, M. Yamanouchi, S. Ikeda, F. Matsukura and H. Ohno, "Advances in spintronics devices for microelectronics — From spin-transfer torque to spin-orbit torque" in *Design Automation Conference (ASP-DAC),* Suntec, Singapore, 2014.

[8] K. Wang, H. Lee and P. Amiri, "Magnetoelectric Random Access Memory-Based Circuit Design by Using Voltage-Controlled Magnetic Anisotropy in Magnetic Tunnel Junctions", *Nanotechnology, IEEE Transactions on,* vol. 14, no. 6, pp. 992-997, 2015.

[9] H. Yu and Y. Wang, Design Exploration of Emerging Nano-scale Nonvolatile Memory, New York: Springer Science+Business Media, 2014.

[10] C. O. Leon, "Memristor-the missing circuit element", *IEEE Trans Circuit Theory,* vol. 18, no. 5, pp. 507-519, 1971.

[11] S. C. Deepak, "Resistive RAM: Technology and Market Opportunites", NuPGATM Corporation, 16 11 2010. [Online]. Available: http://www.monolithic3d.com/uploads/6/0/5/5/6055488/deepak_ieee_scv_society_talk.ppt. [Accessed 05 12 2014].

[12] H.-S. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi and K. E. Goodson, "Phase Change Memory", *Proceedings of the IEEE,* vol. 98, no. 12, pp. 2201-2227, 2010.

[13] P. C. Lacaze and J.-C. Lacroix, "Recent techniques for improvement of amorphization and crystallization rates of phase-change materials" in *Non-volatile Memories,* London/Hoboken, Wiley-ISTE, 2014, pp. 156-160.

[14] Z. Szandtner, "Új kutatások a félvezető memóriák technológiája és architektúrája területén" in *Óbudai Egyetem,* XXXII. OTDK Műszaki Tudományok Szekció, 2015.

[15] C. Smullen, A. Nigam, S. Gurumurthi and M. Stan, "The STeTSiMS STT-RAM simulation and modeling system" in *Computer-Aided Design (ICCAD),* 2011 IEEE/ACM International Conference on, San Jose, CA, 2001.

[16] D. Wang, B. Ganesh, N. Tuaycharoen, K. Baynes, A. Jaleel and J. Bruce, "DRAMsim: A Memory System Simulator", *SIGARCH Comput. Archit. News,* vol. 33, no. 4, pp. 100-107, 2005.

[17] M. Poremba and Y. Xie, "NVMain: An Architectural-Level Main Memory Simulator for Emerging Non-volatile Memories" in *VLSI (ISVLSI), 2012 IEEE Computer Society Annual Symposium on,* Amherst, MA, 2012.

[18] S. Wilton and N. Jouppi, "CACTI: an enhanced cache access and cycle time model", *Solid-State Circuits, IEEE Journal of,* vol. 31, no. 5, pp. 677-688, 1996.

[19] M. Poremba, S. Mittal, D. Li, J. Vetter and Y. Xie, "DESTINY: A tool for modeling emerging 3D NVM and eDRAM caches" in *Design, Automation & Test in Europe Conference & Exhibition (DATE),* Grenoble, 2015.

[20] N. Muralimanohar, R. Balasubramonian and N. P. Jouppi, "CACTI 6.0: A Tool to Model Large Caches" in *International Symposium on Microarchitecture,* Chicago, 2007.

[21] K. Chen, S. Li, N. Muralimanohar, J. H. Ahn, J. Brockman and N. Jouppi, “CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory” in *Design, Automation & Test in Europe Conference & Exhibition (DATE),* Dresden, 2012.

[22] X. Dong, C. Xu, Y. Xie and N. Jouppi, “NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory”, *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 31, no. 7, pp. 994,1007, 07 2012.

[23] C. Yiran, “Prediction of STT-RAM Parameters for 2012-2025” in *Cloud·Storage·Big Data Summit és 2nd Asian Nonvolatile Memory Workshop (ANVMW),* Shanghai, China, 2013.

[24] B. C. Lee, E. Ipek, O. Mutlu and D. Burger, “Architecting Phase Change Memory As a Scalable Dram Alternative”, *SIGARCH Comput. Archit. News,* vol. 37, no. 3, pp. 2-13, 2009.

[25] S. Thoziyoor, N. Muralimanohar, J. H. Ahn and N. P. Jouppi, “CACTI 5.1”, HP Laboratories, Palo Alto, 2008.