Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

Study into the architectures of next-generation computer memories by making use of emerging non-volatile memories and modelling techniques.

Xiong Qiangqiang 1st , Muhammad Ezanuddin Abdul Aziz 2nd

Cite this paper as: Xiong Qiangqiang, Muhammad Ezanuddin Abdul Aziz (2024) Study into the architectures of next-generation computer memories by making use of emerging non-volatile memories and modelling techniques.". *Frontiers in Health Informatics,* (8), 5325-5333

ABSTRACT

A contemporary approach to the design of computer systems puts a priority on energy economy. It is a widely held belief that leakage increases exponentially with smaller CMOS technology. This is due to the fact that standard CMOS scaling theory dictates that threshold and supply voltages are lowered in proportion to device sizes. Because of this, contemporary methods consider leaky power to be a rival to dynamic power. Without a wave of revolutionary technology that can completely change the game, the issue of power budget leakage never be able to be overcome. Within the realm of non-volatile memory technology, there have been a number of significant new advancements that have taken place. "ReRAM," "PCRAM," and "Spin-Torque-Transfer Random Access Memory" (MRAM, STTRAM) are some examples of popular examples of modern non-volatile memories that have desired properties such as low access energy, high cell compactness, and great access performance. These memories consist of a combination of these qualities. Therefore, it is excellent that these new technologies for non-volatile memory are used in the construction of future computers that are not only powerful but also efficient in terms of energy consumption. Due to the fact that these new non-volatile memory technologies are still in the research and development phase, further academic research is required in order to demonstrate their utility. Due to this, this study investigates three different approaches that may be used to facilitate the development of these new types of non-volatile memory. The first step is to create models of a few different types of nonvolatile memory, which include their space requirements, power consumption, and performance at the circuit level.

KEYWORDS: *Non-volatile memory, Computer memory architectures, Aandom-access memory, energy economy.*

1. Introduction

The construction of computer systems, energy efficiency is now emerging as one of the most crucial aspects to take into consideration. As the process node becomes smaller, the influence of the leakage problem becomes more obvious in the cutting-edge CMOS technologies that are now available. In order to enable the construction of the following generation of exascale computing systems that are both cost-effective and efficient, new technology is necessary to provide processing power that is either high in performance or low in power consumption. The primary objective should be to enhance the power and performance characteristics of the traditional memory hierarchy. This is because CPU cores spend orders of magnitude more power than memory access and disc access delay, and that disc power

and system memory power account for up to forty percent of a data center's total power consumption. This is the reason why this is the case. On-chip "Static Random Access Memory" (SRAM), off-chip "Dynamic Random-Access Memory" (DRAM), and storage on hard disc drives (HDDs) are the three primary components that are essential to the modern architecture of computer memory. A rise in the use of solid-state drives (SSDs) as a storage cache between dynamic random-access memory (DRAM) and hard disc drives (HDDs) or even as a replacement for HDDs has been brought about by recent advancements in the density, speed, and affordability of NAND flash technology. Due to the fact that it is mechanically constructed, a hard disc drive (HDD) can only support a certain maximum access speed, which is a significant performance constraint. Despite the fact that contemporary solid-state drives (SSDs) have improved their performance, NAND flash devices are not able to easily replace SSDs in the near future. This is due to the fact that NAND flash devices have a low write endurance of 105 and a slow programming speed (Haensch et al., 2023).

Because of the current state of DRAM main memory, which is typified by high power consumption and growing leakage power, it is less likely that SRAM off-chip primary memories and DRAM on-chip caches able to be dropped to the level of technology that used in the future generation. There is an acute need for innovative technologies that can improve the performance of memory hierarchies while simultaneously reducing the amount of power that is used (Fakhry et al., 2023).

2. Background

Changing the voltage at the threshold of the gated and storing bits in a drifting gate are the two components that make up the operation of flash memory. In comparison to traditional non-volatile memory, NAND flash has overtaken it in terms of its low cost, wide range of applications, and small cell size. Altering the threshold voltage of the flash memory cell may be accomplished by adjusting the number of electrons that are contained inside the isolated floating gate gate. NAND employs either hot carrier injection (HCI) or Fowler-Nordheim (FN) tunnelling in order to either power or discharge the eddy current gate that it is responsible for. During the programming process, the floating gate is subjected to tunnelling charges, which result in the threshold voltage being negative. As a result of the removal of charges by an erase procedure, the voltage turns positive. According to Cojocar et al.'s research from 2020, NAND flash is not the most effective non-volatile memory technology, despite the fact that it is widely used. For the reason that NAND flash memory can only be erased in "block" sizes, the programming process becomes more difficult when working with this kind of memory. The fact that NAND memory has a significant problem with its ability to withstand writing is well knowledge (Taheri et al., 2024).

The fact that this is the case suggests that the number of program-erase cycles that a single flash storage cell is capable of doing could be limited. Within the context of wear-leveling, a "Flash Translation Layer" (FTL) is essentially required in order to simplify the access method and allow wear-leveling to be carried out in an effective manner. The problem of scaling NAND flash memory beyond the 22nm technology node is complicated by the intrinsic physical limitations of the technology as well as its dependence on a diminishing lithographic precision. These restrictions include a low drifting gate

electron charge, a short channel impact, a high interference from the floating gate, and a poor coupling ratio. These are only some of the drawbacks (Inglese, 2023).

3. The purpose of the research

The development of a memory structure that is applicable to any and all circumstances is the major goal of this electronic neural network model (eNVM) research. A broad variety of design alternatives need to be made available by each of these eNVM techniques. These design options should vary from CPU caches that give priority to delay optimisation to extra storage that attempts to optimise density on the other hand. A distinct collection of auxiliary circuits is required for each and every optimisation aim. Because there are so few eNVM technologies that have been completed, there are not a lot of prototype chips that are available. In every respect, the amount of space that is available for design is quite restricted. In an attempt to reduce the amount of time and effort required to construct prototypes, researchers are searching for circuit-level prediction models that can estimate the performance of eNVMs, as well as their energy consumption and chip size. Therefore, they began their investigation on eNVM by developing NVSim, a model that allows for assessments of device-level functions, power consumption, and space. In comparison to CACTI, NVSim is now able to handle a greater variety of memory variations as a result of the inclusion of support for additional memory types such as NAND flash, PCRAM, ReRAM, and STTRAM.

4. Literature Review

Over the course of the last decade, a number of modelling tools have been developed in order to analyse the memory and cache architecture of the system that is based on SRAM and DRAM architecture. It is common practice for computer architects to use the CACTI approach in order to determine the effectiveness, power consumption, and capacity of caches that are composed of dynamic random-access memory (DRAM) and static random-access memory (SRAM). There are a lot of models that fit into this category, such as those that take into consideration large-capacity caches, energy models for SRAMs, leaky power, and organisations that are interconnect-centric. The inability of CACTI to align its fundamental assumptions is the primary reason for the differences that exist between the NVM chips that are made and the NVM circuit implementations that are used in real reality. Write pause, data relocation, early write termination, and dynamically duplicated memory are some of the architectural approaches that have been proposed as potential solutions to the problems that are associated with eNVM write operations (Ahmed et al., 2021). Attackers may make use of the limited writing endurance of eNVM to their advantage and install harmful applications, which ultimately result in the memory being destroyed. The Non-Volatile System Laboratory at the University of California, San Diego has been working on the development of storage prototypes in order to get a better understanding of the possibilities of non-volatile memory as a form of long-term data storage. Moneta is their product, which is a storage array that attaches to PCIe and has a capacity of 64 gigabytes and mimicked PCRAM (Mishra et al., 2023).

The relationship between the program and the hardware was designed with a great deal of consideration. In addition, the revolutionary solid-state drive (SSD) Onyx, which is based on PCRAM, was created. A method to the creation of a morphable memory system that is based on memory

monitoring; this technique includes a main memory space that is constructed on PCRAM and has MLC and SLC sections (Pan et al., 2024).

5. Question

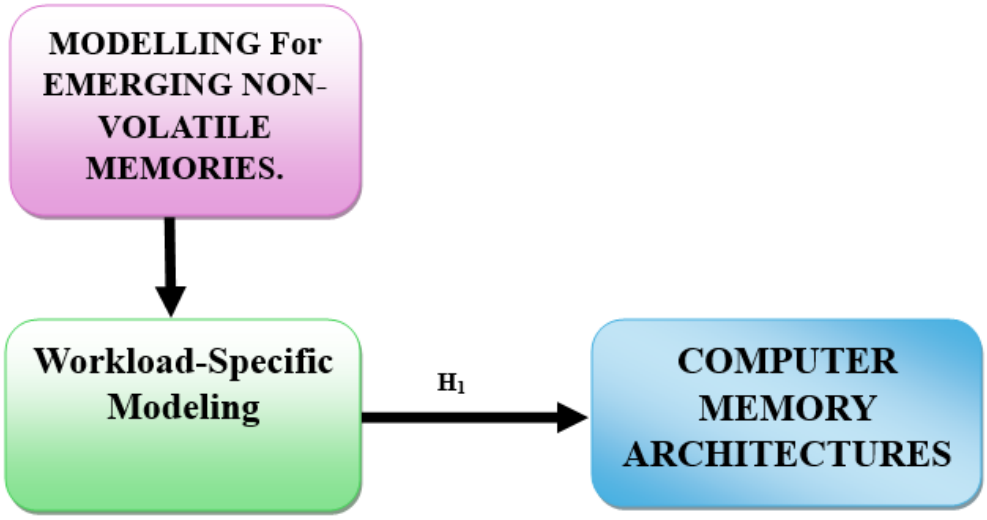
- What are the most effective computational modeling methods to simulate the behavior of non-volatile memories like RRAM, PCM, and STT-MRAM in diverse computing environments?

6. Methodology

6.1 Research design

There's evidence that a Multi-Level Cell (MLC) is capable of storing multiple bits of digital data, which makes it a potential candidate for electronic non-volatile memory (eNVM). eNVM has become a serious competitor in the market as a result of this development. This is due to the fact that NAND flash technology currently has MLC capabilities, but there is no simple solution to expand to denser arrays. eNVMs that have MLC capabilities, such as PCRAM and ReRAM, often display a longer programming time and a poorer cell lifespan in contrast to their SLC counterparts. This is because MLC capabilities are similar to those of SLC. As a result, it proposes a reconfigurable eNVM architecture that is capable of switching between MLCs and SLCs with ease. This design should take into consideration the particulars of the workload as well as the needed lifespan in order to optimise the large MLC capacity and the quick SLC access speed. For the sake of maintaining the generalisability of their findings, researchers turn to MLC PCRAM as an example.

6.2 Conceptual framework

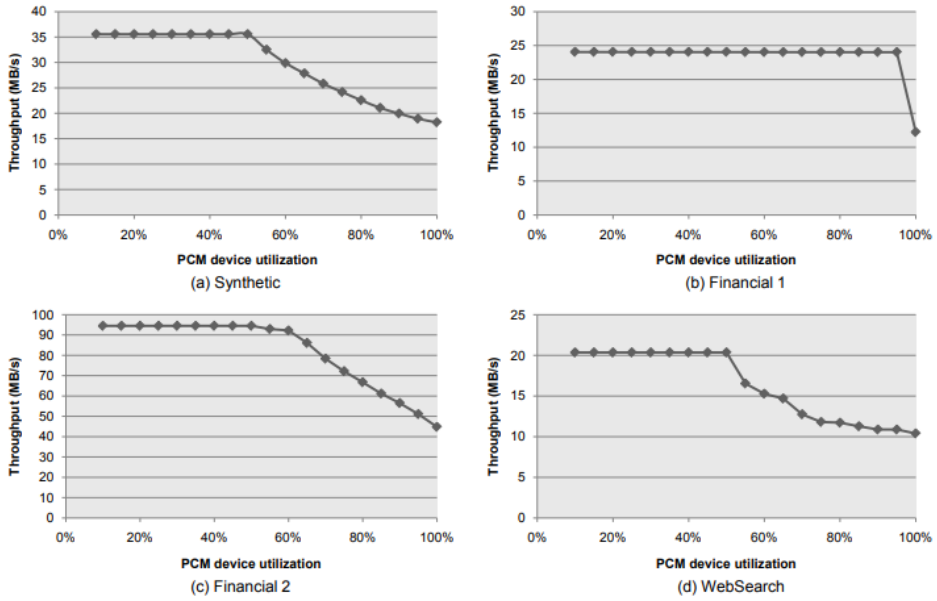


7. Results

The purpose of this section is to evaluate the ways in which the adaptive MLC/SLC approach enhances the performance of PCRAM devices and extends their lifespan when they are not being

used to their maximum capacity. A 2GB RAMDISK that was configured as Ext2 memory and had a block size of 4KB was used to capture the actual I/O trace on the Linux 2.6.32-23 kernel. This was done so that the recommended approach could be evaluated on a platform that was representative of the real world. In the beginning, they created a fake file system trail by reading 1,500,000 documents that were written at random from RAMDISK. These documents ranged in size from 5 KB to 10 MB of data.

Figure 1: The Adaptive MLC/SLC Solution's Performance in Various Use Cases



Additionally, in order to replicate disc activity with greater precision, they relied on disc records that were kept by the Storage Management Council. This enabled them to compete with enterprise-level applications such as web servers, database servers, and webpages. Synthetic is the name given to the synthetic trace, while Financial 1, Financial 2, or Web Search are the names given to the individual traces that originate from SPC.

7.1 Modelling the Timing of PCRAM MLC/SLC

For the purpose of determining the read and resulting latencies in SLC or MLC modes, a prototype version of NVSim was used. The write delay of the SLC was determined by summing the delays of the SET and RESET operations, while the latency of the MLC type was assumed to be four P&V steps in order to simplify the calculation. A comprehensive numerical summary of the outcomes of the calculations is presented in Table 1, which may be found here. As a result of the large current that was required for the SET and RESET processes, the reading and writing widths were both restricted to 64 and 16 cells, respectively. On the other hand, the latter makes use of 128 bits, whereas the former was

for SLC, which utilises 64 bits. As a result of these assumptions, the I/O bandwidth in SLC mode is almost twice as high as it is in MLC mode.

Figure 2: A Cost-Benefit Study of The Adaptive MLC/SLC System

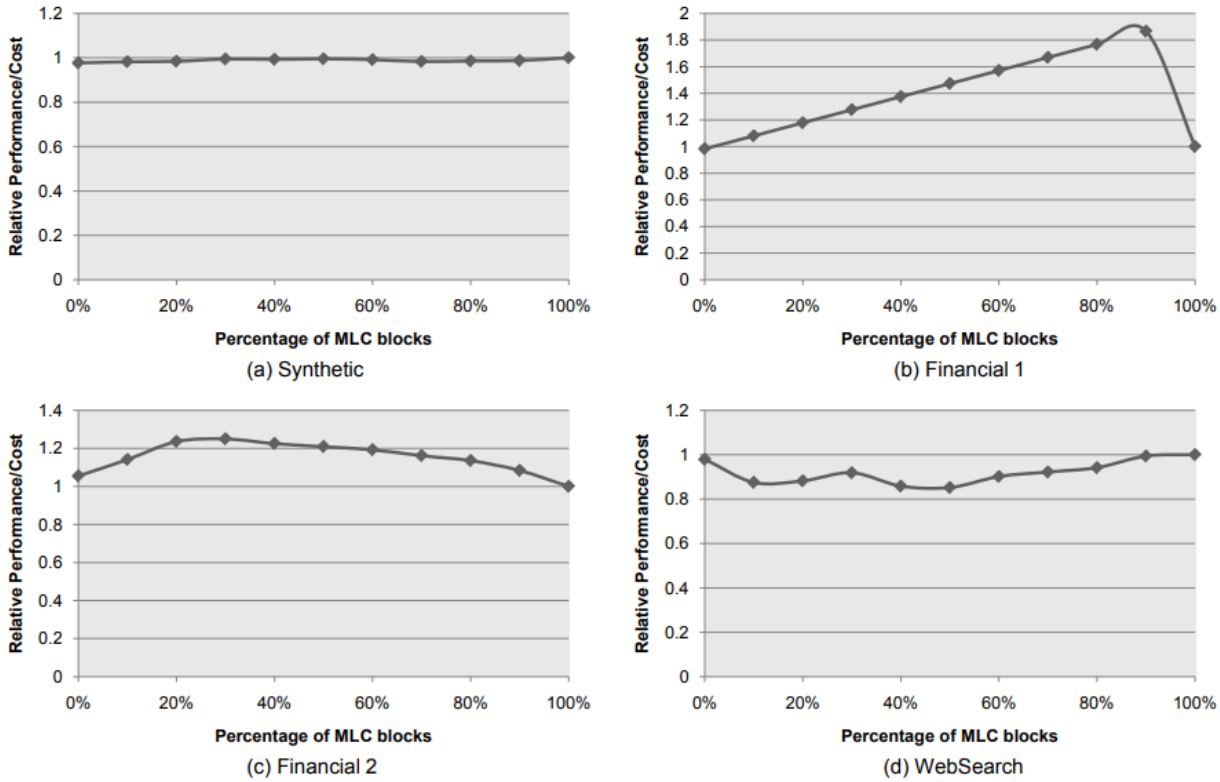


Table 1: PCRAM Cell Timing Model for SLC And MLC Modes

	SLC	MLC
Read latency	$10ns$	$44ns$
Read width	$64bits$	$128bits$
Read bandwidth	$800MB/s$	$363.6MB/s$
Write latency	$100ns$	$395ns$
Write width	$16bits$	$32bits$
Write bandwidth	$20.0MB/s$	$10.3MB/s$

7.2 The Outcome of Performance-Aware Management

Starting with a demonstration of the performance-aware partitioning approach in action is the first step

that they take in order to boost speed. There is a significant impact on the effectiveness of the adaptive MLC/SLC PCRAM that is advised by the I/O access distribution. For the purpose of ensuring consistent access across the file system, it is important to retrieve data from the MLC regions. In the other direction, the data that is utilised most often needs to be kept in SLC portions, and files that are accessed less frequently could be partitioned utilising MLC areas. This is something that can be performed, supposing that the access pattern displays bias. According to the findings of the research, a massive device that is able to store the set of operations may be created by joining a number of PCRAM devices in their array configuration. One factor that influences how the device is used is the quantity of PCRAM chips that it contains. There is a transition between SLC and MLC modes that takes place when the device use of all PCRAM blocks falls below 50% and 100%, respectively. This is the point at which the adaptive MLC/SLC technique delivers the needed capacity. Utilisation reduces between fifty percent and one hundred percent. A representation of the relationship between the various workloads and the average throughput as a function of the utilisation of PCRAM devices is shown in Figure 1. The use of Synthetic and Financial 2 causes a progressive decline in the throughput of both of these applications. It is because of the fact that these two jobs are highly dependent on data that is kept locally that this result has occurred. In addition, the effect of this occurrence is amplified in Financial 1 because of the large number of files that are required just once. As a result of these workloads, the I/O access pattern is evenly distributed, which causes WebSearch's performance to drop dramatically by fifty percent.

7.3 Cost-Efficiency Evaluation

Their prior discovery on the association between a performance and device utilisation may be used to deduce the relationship between performance and cost, assuming the price of a PCRAM chip stays constant. Financial 1's workload access pattern offers two extreme options. One makes use of 94 MB/s bandwidth from PCRAM chips that are exclusive to SLC, while the other utilises 44 MB/s bandwidth from PCRAM chips that are exclusive to MLC but consumes half as many PCRAM chips. We can see the throughput-per-cost metric in its most favourable state in Figure 2, and its rephrased conclusion for the study can be found in Figure 1. Gains of around 28% in throughput-per-cost are typical.

7.4 A Lifetime Evaluation

The RESET/SET resistance margin of an MLC PCRAM cell shrinks as the number of RESET operations increases. Consequently, it is when it has to be officially recognised as an SLC. To begin the lifetime-aware partitioning method, set the MLC type initialisation for the PCRAM chunks in the MSB bank and leave the LSB bank blocks unfilled. When the OS monitor notices that a block's accessing probability has gone up, it switches that block to SLC mode by activating the corresponding blocks in the LSB banks. The lifetime model predicts that there might be one hundred lifetime benefits from the lifespan-aware partitioning approach. Because of the reduced device capacity, there is a significant improvement in lifetime.

8. Discussion

Taking into consideration the features of the workload as well as the requirements for the lifespan, the adaptive MLC/SLC approach that has been presented takes use of the enormous capacity of the MLC and the rapid access speed of the SLC. The development of the MLC/SLC mode management technique is brought to a close in this portion of the dissertation. This strategy is based on the designs of an adaptable MLC/SLC eNVM array at the circuit level. For the purpose of determining whether or not the adaptive MLC/SLC approach that has been described is suitable, a case study with a PCRAM-based storage device is used. Khan et al. found that the adaptive MLC/SLC approach has the ability to improve the throughput-per-cost of PCRAM devices by a median of 28%, and by 100% when the device utilisation falls below 50% (Khan et al., 2019). This was discovered via an investigation of four real-world I/O traces (Taheri et al., 2024).

9. Conclusion

The non-volatile memory technologies that are on the horizon, such as STTRAM, PCRAM, and ReRAM, are attracting the attention of the computer design community because of their high density, tremendous scalability, quick access, and absence of volatility. It has been over thirty years since these technologies were first introduced, but they have now caused a disruption in the traditional memory hierarchy and posed a challenge to both SRAM and DRAM. Models for enhanced design at the architectural level, as well as models for assessing area, energy, and performance at the circuit level, are presented in this dissertation. Case examples are presented at many different levels of application. This is despite the fact that many different forms of non-volatile memory are still in the prototype stage. In the second section, strategies were developed at the architectural level in order to solve the restrictions that are associated with write operations in non-volatile memory. By conducting assessments at the architectural level, they were able to demonstrate that these strategies were effective. A few examples of the numerous applications that these case studies demonstrate how non-volatile memory technologies may improve power economy or performance include checkpointing, memory hierarchy, and secondary storage. These are only a few of the many applications. Taking into consideration the findings of these case studies, it is plausible to expect that non-volatile memory technologies ultimately take the place of older memory and disc technologies. This might hasten the development of these technologies and contribute to the revolution in non-volatile memory that is now going place that is currently taking place (Pan et al., 2024).

References

Haensch, W., Raghunathan, A., Roy, K., Chakrabarti, B., Phatak, C. M., Wang, C., & Guha, S. (2023). Compute in-memory with non-volatile elements for neural networks: A review from a co-design perspective. *Advanced Materials*, *35*(37), 2204944.

Fakhry, D., Abdelsalam, M., El-Kharashi, M. W., & Safar, M. (2023). A review on computational storage devices and near memory computing for high performance applications. *Memories-Materials, Devices, Circuits and Systems*, *4*, 100051.

Taheri, N., Tabrizchi, S., & Roohi, A. (2024). Intermittent-Aware Design Exploration of Systolic Array Using Various Non-Volatile Memory: A Comparative Study. *Micromachines*, *15*(3), 343.

Inglese, P. (2023). *Exploration of security threats in In-Memory Computing Paradigms* (Doctoral dissertation, Université Grenoble Alpes [2020-....]).

Mishra, V., Kumar, A., & Akashe, S. (2023, July). New Non-Volatile Memory Technologies and Neuromorphic Computing. In *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 857-862).

Pan, J., Wang, Z., Zhao, B., Yin, J., Guo, P., Yang, Y., & Ren, T. L. (2024). Recent Progress of Non-Volatile Memory Devices Based on Two-Dimensional Materials. *Chips*, *3*(4), 271-295.

Sarkar, M. R. (2024). *The Future of Computing: An Energy-Efficient In-Memory Computing Architectures with Emerging VGSOT MRAM Technology* (Doctoral dissertation, Virginia Tech).

Ajayan, J., Mohankumar, P., Nirmal, D., Joseph, L. L., Bhattacharya, S., Sreejith, S., ... & Mounika, B. (2023). Ferroelectric field effect transistors (FeFETs): advancements, challenges and exciting prospects for next generation non-volatile memory (NVM) applications. *Materials Today Communications*, *35*, 105591.