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

In order to gain an efficient and more refined understanding of the architectures of next-generation computer memories, making use of emerging non-volatile memories and modeling techniques are necessary.

Xiong Qiangqiang 1st , Muhammad Ezanuddin Abdul Aziz 2nd

Cite this paper as: Xiong Qiangqiang, Muhammad Ezanuddin Abdul Aziz (2024) In order to gain an efficient and more refined understanding of the architectures of next-generation computer memories, making use of emerging non-volatile memories and modeling techniques are necessary.". *Frontiers in Health Informatics,* (8), 5289-5298

ABSTRACT

Energy efficiency is given a high level of importance in the design of computer systems that are based on modern methodologies. There is a widespread idea that the amount of leakage rises exponentially with the size of the CMOS processing technology. The reason for this is because the traditional CMOS scaling theory stipulates that the threshold and supply voltages should be decreased in proportion to the size of the device. The reason for this is because modern approaches regard leaky power to be a competitor to dynamic power. In the absence of a wave of new technology that has the potential to radically alter the game, the problem of power budget leakage never be able to be entirely resolved. There have been a number of notable new advances that have taken place within the field of non-volatile memory technology. These improvements are discussed more below. Examples of popular examples of current non-volatile memories include "ReRAM," "PCRAM," and "Spin-Torque-Transfer Random Access Memory" (MRAM, STTRAM). These types of memories have desirable characteristics such as low access energy, high cell compactness, and excellent access performance. These recollections are made up of combinations of all of these characteristics. Therefore, it is wonderful that these new technologies for non-volatile memory are being used in the creation of future computers that are not only powerful but also efficient in terms of the amount of energy that they consume. Because these fresh non-volatile memory technologies are still in the research and development phase, further academic study is necessary in order to establish their applicability. This is because the aforementioned technologies are continually evolving.

KEYWORDS: *Non-volatile memory, Random-access memory, Energy economy. Modern methodologies.*

1. Introduction

The architecture of computer systems, energy efficiency is quickly becoming one of the most important factors to consider. The effect of the leakage issue becomes increasingly visible in the cutting-edge CMOS technologies that are now accessible. This is because the process node is becoming smaller. To allow the building of the next generation of exascale computing systems that are both cost-effective and efficient, new technology is required to supply processing power that is either high in performance

or low in power consumption. This is essential in order to enable the creation of these systems. Enhancing the power and performance characteristics of the conventional memory hierarchy needs to be the major goal of this endeavour. The reason for this is that central processing unit cores use orders of magnitude more power than memory access and disc access latency, and that disc power and system memory power account for up to forty percent of the overall power consumption of a data centre. This is the reason why things are the way they are. There are three basic components that are crucial to the present design of computer memory. These components include storage on hard disc drives (HDDs), on-chip "Static Random Access Memory" (SRAM), and off-chip "Dynamic Random-Access Memory" (DRAM). Recent developments in the density, speed, and affordability of NAND flash technology have led to an increase in the utilisation of solid-state drives (SSDs) as a storage cache between dynamic random-access memory (DRAM) and hard disc drives (HDDs), or even as a substitute for HDDs. This trend has been brought about by the fact that SSDs are now more affordable than HDDs. Hard disc drives (HDDs) are physically manufactured, which means that they can only support a specific maximum access speed. This is a major performance restriction since it limits the amount of data that can be stored on the unit (Taheri et al., 2024).

In spite of the fact that modern solid-state drives (SSDs) have undergone performance enhancements, NAND flash devices are not going to be able to readily replace SSDs in the near future. This is because NAND flash chips have a sluggish programming speed and a low write endurance of 105. Other factors that contribute to this are also at play. Because of the present state of DRAM main memory, which is characterised by high power consumption and increased leakage power, it is less probable that SRAM off-chip primary memories and DRAM on-chip caches able to be reduced to the level of technology that employed in the generation that come after this one (Syed et al., 2024).

2. Background

There are two components that are responsible for the functioning of flash memory. These components are the modification of the voltage at the threshold of the gate and the storage of bits in a drifting gate. The old non-volatile memory has been surpassed by NAND flash in terms of its cheap cost, broad variety of applications, and tiny cell size. NAND flash has also surpassed it in terms of its small cell size. Adjusting the quantity of electrons that are contained inside the isolated floating gate gate is one method that may be used to bring about a change in the threshold voltage of the flash memory cell. For the purpose of either powering or discharging the eddy current gate that it is accountable for, NAND makes use of either hot carrier injection (HCI) or Fowler-Nordheim (FN) tunnelling. In the course of the programming process, the floating gate is exposed to tunnelling charges, which ultimately leads to a threshold voltage that is negative. As a consequence of the elimination of charges by the use of an erase technique, the voltage becomes positive (Zhang & Pazos, 2024).

In spite of the fact that it is extensively used, the study that was conducted in the year 2020 by Cojocar and colleagues concluded that NAND flash is not the most efficient non-volatile memory technology. Because NAND flash memory can only be erased in "block" sizes, the process of programming becomes more complicated when dealing with this kind of memory. This is because of the physical

limitations of memory. The fact that NAND memory has a big difficulty with its capacity to tolerate writing is something that is commonly known. The fact that this is the case gives rise to the possibility that the number of program-erase cycles that a single flash storage cell is able to perform may be restricted. With regard to the process of wear-leveling, a "Flash Translation Layer" (FTL) is basically necessary in order to simplify the access method and make it possible for wear-leveling to be carried out in an efficient way (Harabi, 2023).

3. The purpose of the research

The purpose of the research is to develop an efficient and more refined understanding of the architectures of next-generation computer memories. Research into new non-volatile memory (NVM) technologies and the application of sophisticated modelling methods are essential steps in developing and deploying memory systems that overcome the drawbacks of older memory architectures in areas like speed, scalability, energy efficiency, and data retention. Innovative applications in domains demanding high-performance and dependable memory solutions are the target of the study, which also seeks to improve computing performance while reducing power consumption.

4. Literature Review

In order to assess the memory and cache architecture of the system that is based on SRAM and DRAM architecture, a number of modelling tools have been created over the course of the last decade. These tools have been developed in order to analyse the system. To measure the efficiency, power consumption, and capacity of caches that are made up of dynamic random-access memory (DRAM) and static random-access memory (SRAM), it is standard practice for computer architects to apply the CACTI technique. This is done in order to determine the efficacy of caches. There are a great number of models that fall into this category. Some examples of these models are those that consider large-capacity caches, energy models for SRAMs, leaky power, and organisations that are interconnect-centric. The key reason for the disparities that exist between the NVM chips that are manufactured and the NVM circuit implementations that are utilised in real life is because CACTI is unable to align its core assumptions. This is the primary reason for the variances. There are a number of architectural techniques that have been presented as possible solutions to the issues that are connected with eNVM write operations. Some of these ideas include write pause, data relocation, early write termination, and dynamically duplicated memory. It is possible for attackers to take advantage of the limited writing endurance of eNVM to their advantage and install malicious apps, which eventually result in the memory being destroyed. Non-Volatile System Laboratory at the University of California, San Diego has been working on the creation of storage prototypes in order to get a better knowledge of the possibilities of non-volatile memory as a form of long-term data storage. This is being done in order to better grasp the potential of non-volatile memory. The device that they provide is called Moneta, and it is a storage array that mimics PCRAM and has a capacity of 64 gigabytes. It is attached to PCIe and has similar characteristics (Khan et al., 2024).

5. Question

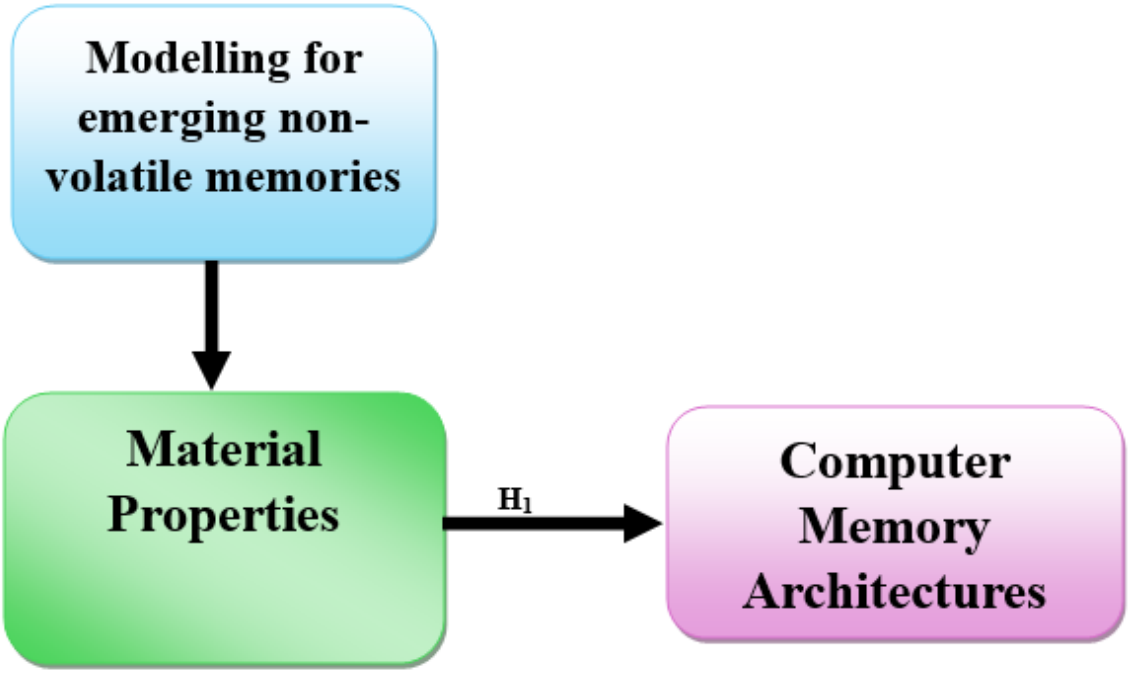
- What role do emerging non-volatile memories and advanced modeling techniques play in achieving an efficient and refined understanding of next-generation computer memory architectures?

6. Methodology

6.1 Research design

Multiple bits of digital data may be stored by a Multi-Level Cell (MLC), which makes it a promising option for electronic non-volatile memory (eNVM). This change has elevated eNVM to the level of a formidable market rival. This is because there isn't a straightforward way to increase the density of NAND flash arrays, even if the technology already has MLC capability. PCRAM and ReRAM are examples of eNVMs with MLC capabilities; in comparison to SLC eNVMs, they often have a longer programming time and worse cell longevity. This is due to the fact that MLC's capacities are comparable to SLC's. So, it suggests an eNVM design that can be easily changed to work with either MLCs or SLCs. In order to maximise the big MLC capacity and the rapid SLC access speed, this design should consider the workload details and the required lifetime. Using MLC PCRAM as an example helps researchers keep their results generalisable.

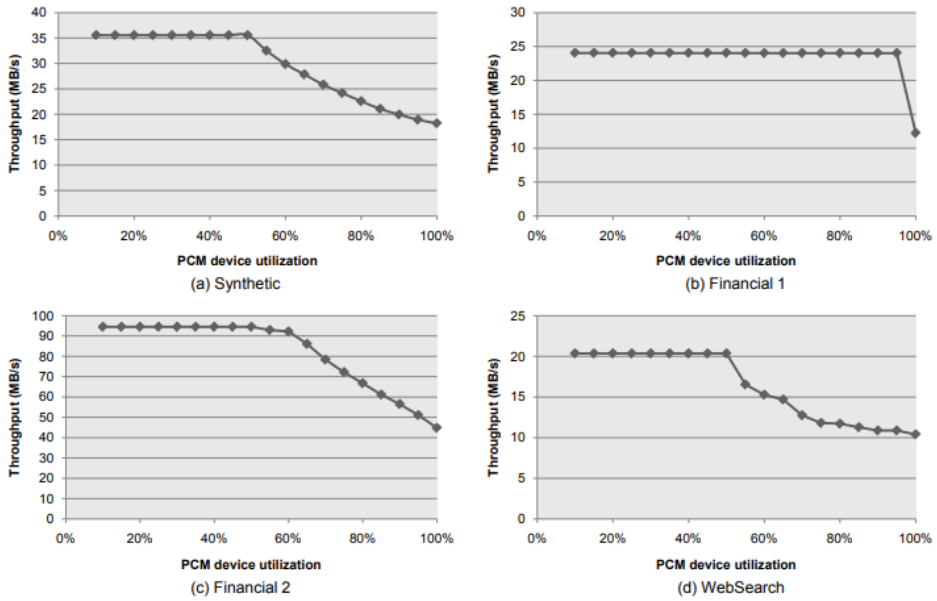
6.2 Conceptual framework



7. Results

This section's goal is to assess how the adaptive MLC/SLC method improves PCRAM devices' performance and duration while they're not under heavy load. The real I/O trace on the Linux 2.6.32-23 kernel was captured using a 2GB RAMDISK with a block size of 4KB and set as Ext2 memory. This was done to ensure that the suggested method could be tested in a realistic setting. Beginning with 1,500,000 randomly written documents read from RAMDISK, they constructed a false file system trace. From 5 KB to 10 MB, that's how much data these papers included.

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



Extrapolating disc activity from data maintained by the Storage Management Council allowed for more accurate replication. Web servers, database servers, and webpages—applications used by enterprises—could now compete with them thanks to this. Individual traces that originate from SPC are referred to as Financial 1, Financial 2, or Web Search, whereas the synthetic trace is called Synthetic.

7.1 Modelling the Timing of PCRAM MLC/SLC

Read and resultant latencies in SLC and MLC modes were calculated using a prototype of NVSim. In order to simplify the computation, the latency of the MLC type was assumed to be four P&V steps, while the write delay of the SLC was found by adding the delays of the SET and RESET operations. Table 1 provides a detailed numerical overview of the computation results and may be accessed here. The SET and RESET operations, which need a significant current, limit the reading width to 64 cells and the writing width to 16 cells, respectively. In contrast, the latter use 128 bits, while the former was reserved for SLC, which makes use of 64 bits. The I/O bandwidth in SLC mode is almost double that in MLC mode due to these assumptions.

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

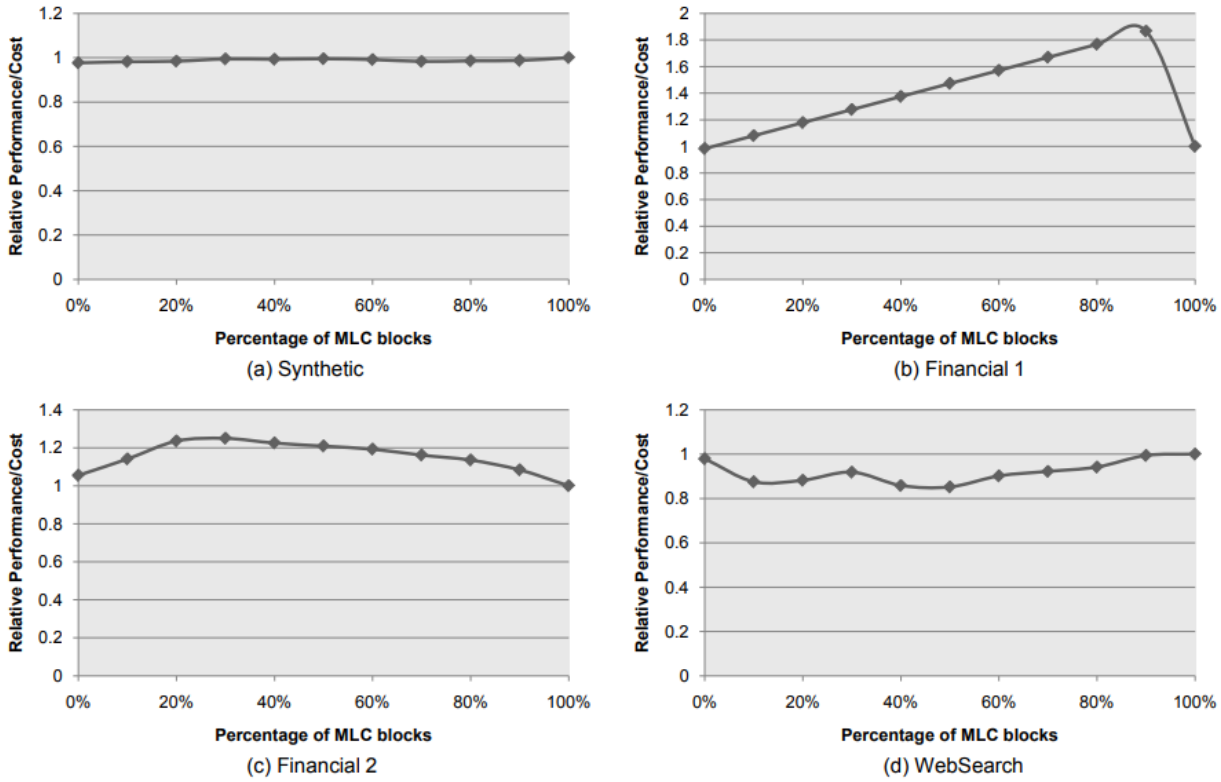


Table 1: PCRAM Cell Timing Model for SLC And MLC Modes

	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 Results of Performance-Aware Management

The first step they take to enhance speed is to demonstrate the performance-aware partitioning strategy in action. The recommended adaptive MLC/SLC PCRAM's performance is greatly affected by the I/O access distribution. It is critical to recover data from the MLC regions to provide consistent access across the file system. Alternatively, files with seldom use might be partitioned using MLC sections, while frequently used data should be stored in SLC parts. It is possible to carry out this operation if

the access pattern shows bias. Research shows that by connecting many PCRAM devices in an array format, a large device capable of storing the set of operations may be built. How many PCRAM chips a gadget has been one determinant of its usefulness. The device enters MLC mode when the utilisation of all PCRAM blocks drops below 100% and SLC mode when it drops below 50%. When this happens, the adaptive MLC/SLC method provides the necessary capacity. A fifty percent to a hundred percent drop in utilisation occurs. Figure 1 depicts the connection between the average throughput and the different workloads as a function of the utilisation of PCRAM devices. As Synthetic and Financial 2 are used more often, their throughput gradually decreases. This outcome is the consequence of the fact that these two tasks rely heavily on data stored locally. Because of the high volume of files needed just once, this event has a magnified impact on Financial 1. The uniform distribution of I/O access patterns brought forth by these workloads significantly reduces WebSearch's performance by 50%.

7.3 Assessment of Cost-Effectiveness

Assuming the price of a PCRAM chip remains constant, their previous work on the link between performance and device utilisation may be utilised to infer the relationship between performance and cost. The workload access pattern of Financial 1 provides two extreme choices. Two variations exist: one uses 94 MB/s bandwidth from SLC-exclusive PCRAM chips, while the other uses 44 MB/s bandwidth from MLC-exclusive PCRAM chips but uses half as many PCRAM chips. Figure 2 shows the throughput-per-cost measure in its optimal form, and Figure 1 shows its rephrased conclusion relevant to the investigations. Typical throughput-per-cost gains are close to 28 percent.

7.4 An Assessment of One's Lifetime

As the amount of RESET operations grows, the RESET/SET resistance margin of an MLC PCRAM cell decreases. Therefore, it is the time when it must be formally acknowledged as an SLC. Before starting the lifetime-aware partitioning procedure, make sure that the LSB bank blocks are empty and that the PCRAM chunks in the MSB bank are initialised with the MLC type. When the OS monitor detects an increase in a block's accessing probability, it activates the matching blocks in the LSB banks to convert that block to SLC mode. According to the lifetime model, the lifespan-aware partitioning strategy may provide 100 lifetime benefits. Lifespan is greatly enhanced due to the decreased device capacity.

8. Discussion

In the adaptive MLC/SLC strategy that has been shown, the vast capacity of the MLC and the quick access speed of the SLC are used. This technique considers the characteristics of the workload as well as the needs for the lifetime. The MLC/SLC mode management approach is concluded in this section of the dissertation, which brings the development of the technique to an end. The designs of an adaptive MLC/SLC eNVM array at the circuit level are the foundation upon which this technique is built. For the aim of establishing whether or not the adaptive MLC/SLC technique that has been described is appropriate, a case study using a storage device that is based on PCRAM is used. According to the

findings of Khan et al. (2019), the adaptive MLC/SLC strategy has the potential to enhance the throughput-per-cost of PCRAM devices by a median of 28%, and by a factor of 100% when the device utilisation goes below 50%. This was found out by a study of four I/O traces that were taken from the actual world (Schulman et al., 2024).

9. Conclusion

For a number of reasons, including their high density, incredible scalability, rapid access, and lack of volatility, the non-volatile memory technologies that are on the horizon, such as STTRAM, PCRAM, and ReRAM, are gaining the attention of the community of computer designers. Despite the fact that it has been more than thirty years since these technologies were first developed, they have already produced a disruption in the conventional memory hierarchy and offered a threat to both SRAM and DRAM. At the architectural level, this dissertation presents models for improved design, and at the circuit level, it presents models for evaluating area, energy, and performance. Both of these models are offered in this dissertation. All of the various levels of application are covered by the case examples that are offered. This is in spite of the fact that a great number of various kinds of non-volatile memory are still at the prototype stage. The limits that are connected with write operations in non-volatile memory were explored in the second part, which focused on the development of solutions at the architectural level in order to find a solution to these restrictions. Through the execution of evaluations at the architectural level, they were able to provide evidence that these tactics were successful. Checkpointing, memory hierarchy, and secondary storage are just a few examples of the various applications that these case studies illustrate how non-volatile memory technologies may increase power economy or performance. Other examples include the numerous applications that are demonstrated. These applications are only a handful of the many that are available. In light of the outcomes of these case studies, it is reasonable to anticipate that non-volatile memory technologies would eventually replace previous memory and disc technologies. This is a reasonable expectation to have. It is possible that this may speed the development of these technologies and contribute to the revolution in non-volatile memory that is presently taking place. This revolution is currently taking place (Harabi, 2023).

References

Haensch, W. (2024). A Simple Packing Algorithm for Optimized Mapping of Artificial Neural Networks onto Non-Volatile Memory Cross-Bar Arrays. *arXiv preprint arXiv:2411.04814*.

Harabi, K. E. (2023). *Energy-Efficient Memristor-Based Artificial Intelligence Accelerators using In/Near Memory Computing* (Doctoral dissertation, Université Paris-Saclay).

Khan, A. A., De Lima, J. P. C., Farzaneh, H., & Castrillon, J. (2024). The Landscape of Compute-near-memory and Compute-in-memory: A Research and Commercial Overview. *arXiv preprint arXiv:2401.14428*.

Khan, M.N.I.; Nagarajan, K.; Ghosh, S. Hardware Trojans in Emerging Non-Volatile Memories. In Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence,

Italy, 25–29 March 2019.
Laleni, N., Müller, F., Cuñarro, G., Kämpfe, T., & Jang, T. (2024). A High Efficiency Charge Domain Compute-In-Memory 1F1C Macro Using 2-bit FeFET Cells for DNN Processing. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*.
Schulman, A., Huhtinen, H., & Paturi, P. (2024). Manganite Memristive Devices: Recent Progress and Emerging Opportunities. *Journal of Physics D: Applied Physics*.
Syed, G. S., Le Gallo, M., & Sebastian, A. (2024). Non von neumann computing concepts. In *Phase Change Materials-Based Photonic Computing* (pp. 11-35). Elsevier.
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.
Zhang, X., & Pazos, S. (2024). Roadmap to neuromorphic computing with emerging technologies.