Prestieesci Research Review

A STUDY TO UNDERSTAND NEXT-GENERATION COMPUTER MEMORY ARCHITECTURES THROUGH THE USE OF MODELING AND LEVERAGING EMERGING NON-VOLATILE MEMORIES

Xiong Qiangqiang*, Muhammad Ezanuddin Abdul Aziz

Lincoln University College, 47301 Petaling Jaya, Selangor D. E., Malaysia.

Corresponding author: XIONG QIANGQIANG, Lincoln University College, 47301 Petaling Jaya, Selangor D. E., Malaysia, Email: xqqxqqxqq4228@126.com

ABSTRACT

Energy efficiency in computer system architecture is of the utmost importance nowadays. The common belief is that as CMOS technology becomes smaller, the leakage would increase exponentially since traditional CMOS scaling theory predicts that threshold and supply voltages were decrease in response to device sizes. As a result, methods of the present generation see leaky power as rival to dynamic power. Prior to power budget leakage being an important issue, there has to be a boom of groundbreaking, industry-altering technologies. In the field of non-volatile memory technology, there have been several exciting new advancements. “Resistive Random Access Memory” (ReRAM), “Phase-Change Random Access Memory” (PCRAM), and “Spin-Torque-Transfer Random Access Memory” (MRAM, STTRAM) are all examples of modern non-volatile memories that have appealing properties including low access energy, high cell compactness, and excellent access performance. So, it's great to see these new non-volatile memory technologies being used to build low-power, high-performance computers in the future. Since these novel non-volatile memory technologies are still in their early stages of development, further academic research is required to demonstrate their worth. In light of this, this dissertation investigates three strategies for assisting these novel types of non-volatile memory. Space, power consumption, and circuit-level performance models of several nonvolatile memory types serve as the starting points. Secondly, they propose and evaluate several architecture-level strategies to mitigate write operations' detrimental impacts on non-volatile memory. Lastly, they conduct case studies of real-world applications for this state-of-the-art technology.

KEYWORDS: Non-volatile memory, Computer memory architectures, Resistive random-access memory, Phase-change random access memory.

INTRODUCTION

When it comes to building computer systems these days, energy efficiency is a major consideration. As the process node becomes smaller, modern CMOS processes begin to experience the impacts of the leakage problem. Innovative technologies that provide high-performance or low-power processing power are needed to build the next generation of economical exascale computing systems. Because processing cores consume orders of magnitude more power than memory access and disc access latency, and because disc power and system memory power make up as much as 40% of a data center's overall power consumption, improving the power and performance characteristics of conventional memory hierarchy should be the first priority (Esatu, 2023). The three pillars of contemporary computer memory design are storage on hard disc drives (HDDs), off-chip “Dynamic Random-Access Memory” (DRAM), and on-chip “Static Random Access Memory” (SRAM). With the recent advancements in density, speed, and affordability of NAND flash technology, Solid-State Drives (SSDs) have become more popular as a storage cache between DRAM and HDDs or even as a replacement for HDDs. A substantial performance hurdle is introduced by the mechanical architecture of the HDD, which establishes a maximum limit on the access speed. While new SSDs provide better performance, NAND flash devices won't replace them anytime soon because of their slow programming speed and low write endurance of 105. It is less likely that SRAM off-chip primary memories and DRAM on-chip caches can be reduced down to the next generation process level due to the current DRAM main memory's high power consumption and rising leakage power. There is an urgent need for innovative technologies that can improve memory hierarchy performance without appreciably increasing power consumption (Wang et al., 2021).

BACKGROUND OF THE STUDY

Flash memory works by altering the gate voltage at the threshold and storing bits in a drifting gate. NAND flash has overtaken traditional non-volatile memory because of its small cell size, low cost, and wide range of application requirements. The isolated floating gate's electron count may be adjusted to modify the flash memory cell's threshold voltage. To power or discharge the eddy current gate, NAND employs either Fowler-Nordheim (FN) tunnelling or hot carrier injection (HCI). When the floating gate is subjected to tunnelling charges during a programming process, the threshold voltage grows negative. Charges are eliminated using an erase procedure, and the voltage then flips back to positive. Although NAND flash is now the most widely used non-volatile memory technology, it isn't the most efficient (Cojocar et al., 2020). The fact that flash memory made of NAND can only be erased in "block" sizes complicates programming. It is also widely recognised that NAND memory suffers from a serious write endurance problem. This suggests that a single flash storage cell's capacity to do program-erase cycles may be limited. In practical terms, wear-leveling requires a "Flash Translation

Layer" (FTL) to provide effective wear-leveling and simplify the access process. NAND flash memory's fundamental physical limitations and dependence on deteriorating lithographic precision make scaling it beyond the 22nm technology node difficult. Among these restrictions include low coupling ratio, severe floating gate interference, short channel impacts, and low drifting gate electron charge (Sandor et al., 2019).

PURPOSE OF THE STUDY

Establishing a universal memory structure is the main purpose of this eNVM research. All of these eNVM techniques need to provide a variety of design alternatives, ranging from CPU caches with an emphasis on delay optimisation to extra storage with an aim on density optimisation. This implies that a distinct set of auxiliary circuits is required for every optimisation aim. Sadly, not many eNVM technologies are completely developed, and as a result, not many prototype chips are available. What is available only encompasses a small portion of the whole space for design. Researchers want circuit-level predicting models to predict eNVM performance, usage of energy, and chip size, which will save labour and time while constructing prototypes. Thus, the initial phase in their eNVM research was to build NVSim, a model for evaluating eNVM performance, power, and area at the device level. Now that NAND flash, PCRAM, ReRAM, STTRAM, and other memory types have been added, NVSim can handle more memory varieties than CACTI.

LITERATURE REVIEW

In the last decade, a number of modelling tools have been developed to study the system-level architecture of memory and cache that are based on SRAM or DRAM. In computer architecture, the CACTI approach is often used to determine the capacity, power consumption, and efficiency of DRAM and SRAM caches. Examples of such models include leaky power models, energy models for SRAMs, large-capacity caches, and interconnect-centric organisations. However, discrepancies between produced NVM chips and real-world NVM circuit implementations arise from the fact that CACTI's fundamental assumptions do not align. Some architectural options that have been proposed to address eNVM write difficulties include write pause, data relocation, early write termination, and dynamically duplicated memory (Ahmed et al., 2021). Several types of data redundancy are part of these strategies to fix access failures caused by insufficient eNVM write endurance. Even worse, by capitalising on eNVM's poor writing endurance, attackers may introduce malicious applications that destroy the memory. In an effort to learn more about the possibilities of non-volatile memory as a sort of long-term data storage, the Non-Volatile System Laboratory at UC San Diego has been building storage prototypes. Their product was Moneta, a storage array that is PCIe-

attached and has 64 GB of emulated PCRAM storage. Its software-hardware interface was carefully designed. Onyx, an innovative solid-state drive (SSD) based on PCRAM, was also produced. Methods for developing a morphable memory system that makes use of memory monitoring; the system's main memory space is built on PCRAM and has MLC and SLC regions. It is possible that writing performance might be enhanced using on-chip caches constructed from non-volatile memory technologies like ReRAM and STTRAM. Using STTRAM, they constructed last-level caches to save writing energy and eliminate writing errors. Finally, they looked at the hybrid cache hierarchy, which uses non-volatile and volatile memory technology (Nagarajan et al., 2021).

RESEARCH QUESTION

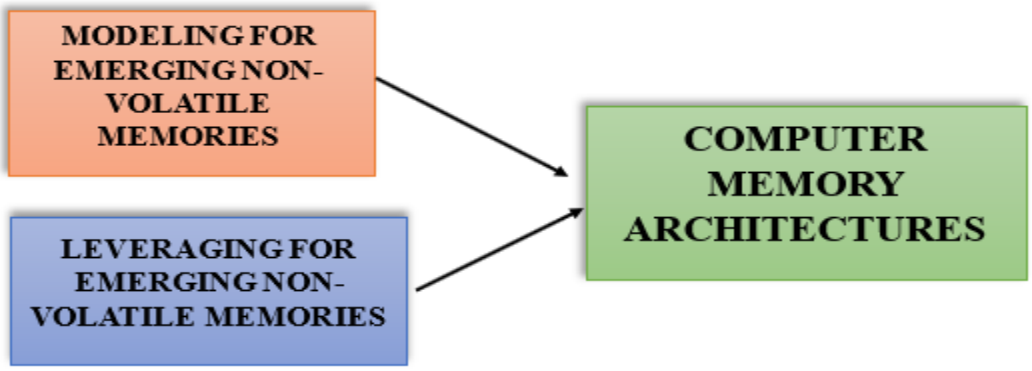
What are the emerging non-volatile memory technologies?

What is a non-volatile memory in computer architecture?

RESEARCH DESIGN

A Multi-Level Cell (MLC) for electronic non-volatile memory (eNVM) has been shown to be feasible, allowing a cell to store several bits of digital data. Thanks to this upgrade, eNVM is now a serious contender in the market and is seen as the natural next step after NAND flash technology, which has MLC capabilities but lacks a straightforward scaling route to achieve denser arrays. In contrast to their single-level cell (SLC) equivalents, eNVMs with MLC capabilities, including PCRAM and ReRAM, often have a longer programming time and a shorter cell lifespan. Hence, it proposes a reconfigurable eNVM architecture that is flexible between MLCs and SLCs, considering the peculiarities of the workload and the necessary lifespan to make the most of the huge MLC capacity and the quick SLC access speed. Researchers examine MLC PCRAM as an example while maintaining generalizability.

CONCEPTUAL FRAMEWORK



RESULTS

The ways that the adaptive MLC/SLC technique enhances performance and extends the life of PCRAM devices when they are not being used to their maximum capacity are evaluated in this section. Using a 2GB RAMDISK configured as an Ext2 memory with a 4KB block size, they captured the genuine I/O trace on the Linux 2.6.32-23 kernel to assess the efficacy of the recommended approach on a real platform. They generated a fictitious file system trace by first flooding the RAMDISK with independently created files, sized between 5KB and 10MB, and then reading those documents 1,500,000 times at random.

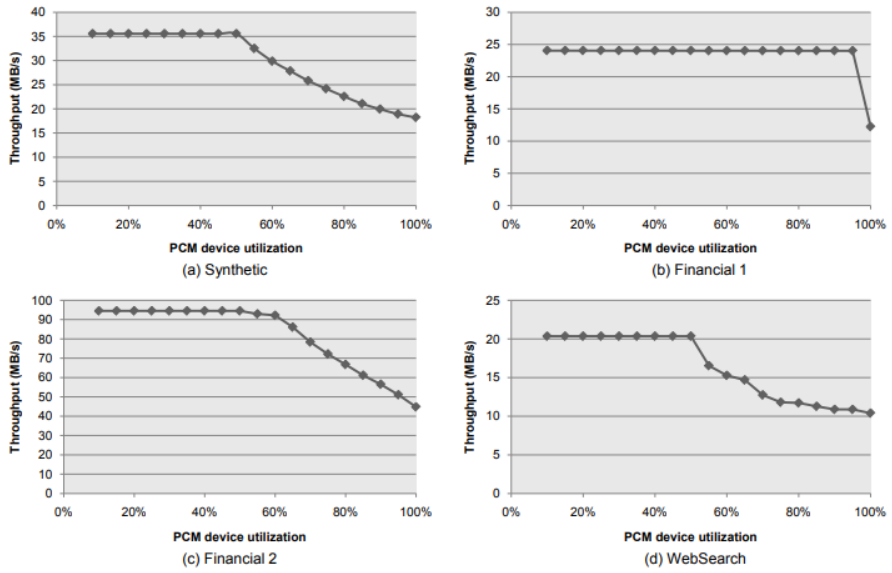


Figure 1. The Adaptive MLC/SLC Solution's Performance in Various Use Cases

They further used disc records from the Storage Management Council to more closely emulate the disc activity on enterprise-level applications such as web servers, database servers, and internet searches. Synthetic is the name given to the synthetic trace, while Financial 1, Financial 2, or WebSearch are the labels given to the traces from SPC.

- **Modelling the Timing of PCRAM MLC/SLC**

The read and results latencies in SLC or MLC modes were calculated using a prototype rendition of NVSim. To keep things simple, the MLC type latency was established by assuming a mean of four P&V steps, while the SLC write delay was determined by combining the SET or RESET delays. Table 1 presents a rounded numerical summary of the calculations' outcomes. The read width was restricted to 64 cells (64 bits for SLC or 128 bits for MLC) and the write width to 16 cells (16 bits for SLC or 32 bits for MLC) due to the high current required for the SET and RESET operations. These presumptions result in an I/O bandwidth for SLC mode that is almost double that of MLC mode.

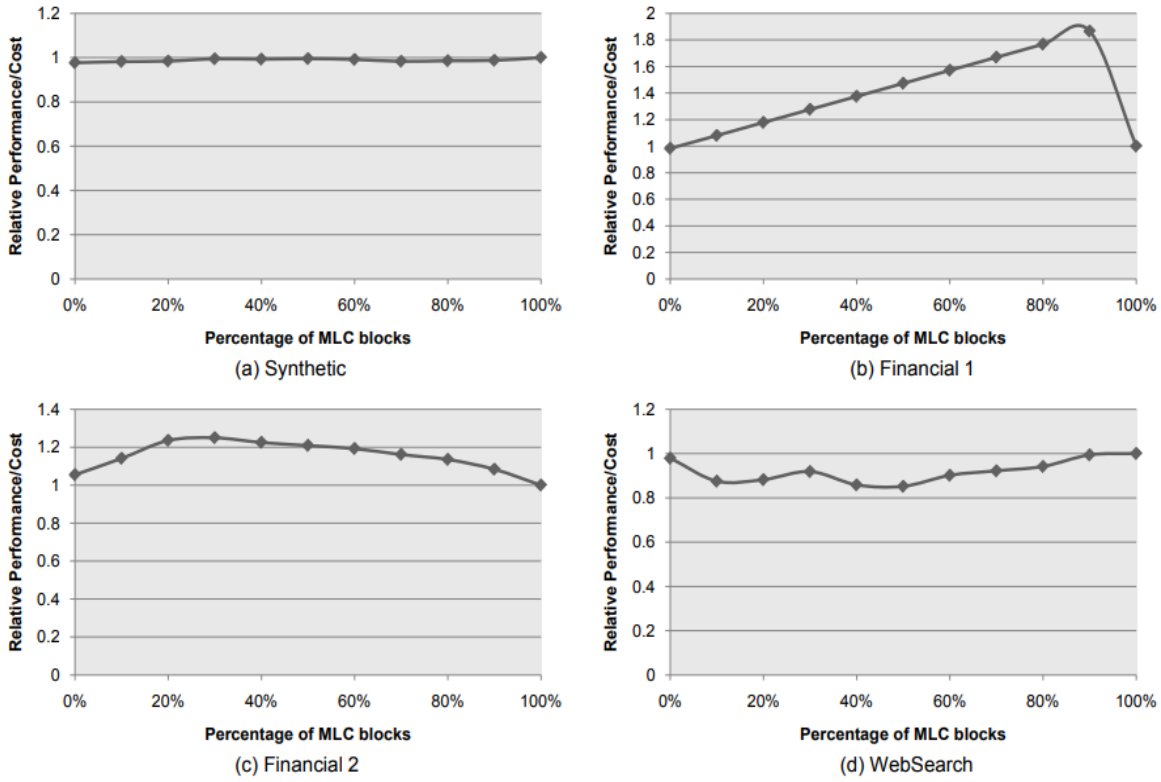


Figure 2. A Cost-Benefit Study of The Adaptive MLC/SLC System

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	$20.0MB/s$	$10.3MB/s$

- **The Outcome of Performance-Aware Management**

They begin by showing how the performance-aware partitioning method works to boost speed. The suggested adaptive MLC/SLC PCRAM's efficiency is heavily dependent on the I/O access distribution. For the file system as a whole to have uniform access, a certain quantity of data must be retrieved from the MLC regions. The flip side is that SLC areas are where most of the data that is accessed often may be accessible, while MLC regions can be used to split files that are not regularly visited. This can be done if the access pattern is biassed. According to the research, a large PCRAM device that can store the set of operations is created by joining several PCRAM devices in an array. One factor influencing how the device is used is the quantity of PCRAM chips. There is a state transition between SLC and MLC modes that occurs when the device utilisation is less than 50% and 100%, respectively, for all PCRAM blocks. When the utilisation falls anywhere between 50% and 100%, the necessary capacity is supplied using the adaptive MLC/SLC approach. Figure 1 shows the correlation between various workloads and the average throughput as a function of PCRAM device utilisation. As the utilisation grows, it becomes apparent that Synthetic and Financial 2 exhibit a progressive decrease in throughput. Because these two workloads rely heavily on locally stored data, this is the result. Also, since a lot of files in Financial 1 are only used once, this effect is amplified there. Because of the evenly distributed I/O access pattern of this workload, WebSearch's performance abruptly drops at 50%.

- **Cost-Efficiency Evaluation**

They may extrapolate the link between performance and cost from the earlier conclusion on the relationship between performance and device utilisation, supposing that the cost of each PCRAM chip is constant. One extreme configuration under Financial 1's workload access pattern uses SLC-only PCRAM chips with a bandwidth of 94 MB/s, while the other uses MLC-only PCRAM chips with a bandwidth of 44 MB/s and half the number of PCRAM chips needed. Figure 2 illustrates the throughput-per-cost metric at its greatest value, and Figure 1 rephrases the result for investigation of the throughput-per-cost measure. An increase of around 28% in throughput-per-cost is typical.

- **A Lifetime Evaluation**

The RESET/SET resistance margin of an MLC PCRAM cell diminishes after an excessive number of RESET operations. Consequently, it is required to be set up as an SLC during that period. To begin the lifetime-aware partitioning method, the MSB bank's PCRAM chunks are initialised to MLC type while the LSB bank's blocks remain empty. The operating system monitor activates the corresponding blocks in the LSB banks to switch particular blocks to SLC mode when it detects that their accessing probability is greater. The lifetime model suggests that the lifetime-aware partitioning approach may provide lifetime improvements of up to 100. The reduced device capacity is responsible for this maximum amount of lifespan improvement.

DISCUSSION

The enormous MLC capacity and quick SLC access speed are used by the suggested adaptive MLC/SLC technique, which considers workload characteristics and lifespan demands. Following the circuit-level design of an adaptive MLC/SLC eNVM array, this portion of the dissertation finishes the creation of the MLC/SLC mode management strategy. The suitability of the suggested adaptive MLC/SLC approach is assessed using a case study with a PCRAM-based storage device. According to a simulation conducted on four real-world I/O traces, the adaptive MLC/SLC approach may increase PCRAM gadget throughput-per-cost by a median of 28%, and 100% if the device use is below 50% (Khan et al., 2019).

CONCLUSION

Because of their non-volatility, high density, quick access, and strong scalability, upcoming non-volatile memory technologies like STTRAM, PCRAM, and ReRAM have piqued the attention of the computer design community. These technologies, which

have been around for more than 30 years, have ultimately upset the traditional memory hierarchy and put SRAM and DRAM in danger. This dissertation includes case studies at the applicability level, architectural level models for better design, and circuit level models for area, energy, and performance measurement. The first part of the dissertation builds and defines models for energy, area, performance, and circuit level for many kinds of non-volatile memory, even if these technologies are still in the prototype stage (Oliveir et al., 2023). By developing architecture-level methods, they addressed the restrictions of write operations in non-volatile memory in the second portion. They carried out assessments at the architectural level to demonstrate the efficacy of these methods. These case studies demonstrate how non-volatile memory technologies may improve power economy or performance in a range of applications, including checkpointing, memory hierarchy, and secondary storage. These case studies provide support to the ideas that future non-volatile memory technologies should take the place of outdated memory/disk technologies. This might hasten the invention of these technologies and further the non-volatile memory revolution now underway (Huang & Wang, 2023).

1. References
2. Ahmed, F.U.; Sandhie, Z.T.; Ali, L.; Chowdhury, M.H. A Brief Overview of On-Chip Voltage Regulation in High-Performance and High-Density Integrated Circuits. IEEE Access 2021, 9, 813–826.
3. Cojocar, L.; Kim, J.; Patel, M.; Tsai, L.; Saroiu, S.; Wolman, A.; Mutlu, O. Are We Susceptible to Rowhammer? An End-to-End Methodology for Cloud Providers. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 18–21 May 2020.
4. Esatu, T. (2023). Novel Non-Volatile Memory Devices and Applications. University of California, Berkeley.
5. Huang, K., & Wang, T. (2023). Indexing on Non-Volatile Memory: Techniques, Lessons Learned and Outlook. Springer Nature.
6. 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.
7. Nagarajan, K.; Ahmed, F.U.; Khan, M.N.I.; De, A.; Chowdhury, M.H.; Ghosh, S. SecNVM: Power Side-Channel Elimination Using On-Chip Capacitors for Highly Secure Emerging NVM. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 2021, 29, 1518–1528.
8. Oliveira, G. F., Ghose, S., Gómez-Luna, J., Boroumand, A., Savery, A., Rao, S., ... & Mutlu, O. (2023). Extending Memory Capacity in Modern Consumer Systems With Emerging Non-Volatile Memory: Experimental Analysis and Characterization Using the Intel Optane SSD. IEEE Access.

9. Sandor, V.K.A.; Lin, Y.; Li, X.; Lin, F.; Zhang, S. Efficient decentralized multi-authority attribute based encryption for mobile cloud data storage. J. Netw. Comput. Appl. 2019, 129, 25–36.
10. Wang, S.; Lee, H.; Grezes, C.; Amiri, P.K.; Wang, K.L.; Gupta, P. Adaptive MRAM Write and Read with MTJ Variation Monitor. IEEE Trans. Emerg. Top. Comput. 2021, 9, 402–413.