Prestieesci Research Review

RESEARCH TO COMPREHEND NEXT-GENERATION COMPUTER MEMORY ARCHITECTURES BY USING MODELLING TECHNIQUES AND UTILISING 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

Modern computer system design places a premium on energy efficiency. Since conventional CMOS scaling theory states that threshold and supply voltages are reduced in relation to device sizes, the widespread assumption is that leakage would grow exponentially with shrinking CMOS technology. Therefore, modern techniques count leaky power as a competitor to dynamic power. The power budget leakage problem can't be solved unless there's a surge of innovative, game-changing technology. A number of noteworthy new developments have occurred in the area of non-volatile memory technology. Popular examples of contemporary non-volatile memories with desirable characteristics such as low access energy, high cell compactness, and outstanding access performance include "ReRAM," "PCRAM," and "Spin-Torque-Transfer Random Access Memory" (MRAM, STTRAM). So, it's fantastic that these new non-volatile memory technologies will be used to construct future computers that are both powerful and energy efficient. To prove their value, further academic study is needed, since these new non-volatile memory technologies are still in the research and development phase. Because of this, three methods for facilitating these new forms of non-volatile memory are explored in this research. We begin with models of several forms of nonvolatile memory, including their space requirements, power consumption, and performance at the circuit level. Second, they assess the effects of write operations on non-volatile memory and provide several techniques at the architectural level to reduce such effects. Finally, they look at actual uses of this cutting-edge innovation in case studies.

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

INTRODUCTION

Nowadays, energy efficiency is one of the most important factors to consider while constructing computer systems. The effect of the leakage issue becomes more noticeable in current CMOS technologies as the process node shrinks. To construct the subsequent generation of cost-effective exascale computing systems, new technology is required to provide processing power with either high performance or low power consumption. Improving the power and performance characteristics of conventional memory hierarchy should be the first priority (Esatu, 2023). This is due to the fact that processing cores consume orders of magnitude more power than memory access and disc access latency, and that disc power and system memory power account for up to 40% of a data center's overall power consumption. Modern computer memory architecture relies on three main components: on-chip "Static Random Access Memory" (SRAM), off-chip "Dynamic Random-Access Memory" (DRAM), and storage on hard disc drives (HDDs). Recent developments in the density, speed, and affordability of NAND flash technology have led to the increased use of Solid-State Drives (SSDs) as a storage cache between DRAM and HDDs, or even as a substitute for HDDs. Because of its mechanical design, an HDD can only support a certain maximum access speed, which is a major performance bottleneck. Despite the improved performance of modern solid-state drives (SSDs), NAND flash devices will not be able to supplant SSDs any time soon because to their poor write endurance of 105 and sluggish programming speed. The present state of DRAM main memory is characterised by high power consumption and increasing leakage power, making it less probable that SRAM off-chip primary memories and DRAM on-chip caches can be decreased to the next generation technology level. Improving memory hierarchy performance with minimal increase in power consumption is an immediate demand for new technologies (Wang et al., 2021).

BACKGROUND OF THE STUDY

Changing the threshold voltage of the gate and storing bits in a drifting gate is how flash memory works. Due to its inexpensive cost, large variety of uses, and compact cell size, NAND flash has surpassed conventional non-volatile memory. Changing the number of electrons in the isolated floating gate allows one to alter the threshold voltage of the flash memory cell. Either hot carrier injection (HCI) or Fowler-Nordheim (FN) tunnelling is used by NAND to power or discharge the eddy current gate. During programming, the floating gate experiences tunnelling charges, which cause the threshold voltage to become negative. The voltage becomes positive after an erase technique removes charges. Despite its widespread usage, NAND flash isn't the most efficient non-volatile memory technology (Cojocar et al., 2020). Programming becomes more complicated when dealing with NAND flash memory since it can only be erased in "block" sizes. There is a major issue with write endurance in NAND memory, which is well-known. This indicates that the number of program-erase cycles that can be performed by a single flash storage cell could be restricted. A "Flash Translation Layer" (FTL) is practically necessary for wear-leveling in order to simplify the access procedure and enable efficient wear-leveling. Scaling NAND flash memory beyond the 22nm technology node is challenging due to its inherent physical limits and dependency on declining lithographic accuracy. Some of these limitations include a low drifting gate electron charge, a short channel impact, a strong interference from the floating gate, and a poor coupling ratio (Sandor et al., 2019).

PURPOSE OF THE STUDY

The primary objective of this eNVM study is to develop a memory structure that is applicable to all situations. A wide range of design options should be provided by all of these eNVM approaches, from CPU caches that prioritise delay optimisation to additional storage that aims to optimise density. All optimisation goals need a separate set of auxiliary circuits. There aren't a plenty of prototype chips accessible since so few eNVM technologies are finished. The available room for design is severely limited in its entirety. In order to save time and effort when building prototypes, researchers are looking for circuit-level prediction models that forecast eNVM performance, energy consumption, and chip size. So, they started their eNVM study by creating NVSim, a model for device-level performance, power, and area evaluations. With the addition of support for new memory kinds such NAND flash, PCRAM, ReRAM, and STTRAM, NVSim is now able to manage more memory variants compared to CACTI.

LITERATURE REVIEW

Several modelling tools have been created in the last ten years to investigate the SRAM and DRAM-based memory and cache architecture at the system level. Computer architects often use the CACTI method to ascertain the efficiency, power consumption, and capacity of dynamic random access memory (DRAM) and static random access memory (SRAM) caches. Many models fall into this category, including those that account for large-capacity caches, energy models for SRAMs, leaky power, and interconnect-centric organisations. The failure to align CACTI's basic assumptions is the root cause of the disparities between manufactured NVM chips and NVM circuit implementations in the actual world. Write pause, data relocation, early write termination, and dynamically duplicated memory are some of the architectural alternatives that have been suggested to solve eNVM write issues (Ahmed et al., 2021). These techniques include various forms of data redundancy to address access failures caused by inadequate eNVM write endurance. Worse, attackers may use eNVM's short writing endurance to their advantage and install destructive apps, thus destroying the memory. Researchers at the University of California, San Diego's Non-Volatile System Laboratory have been developing storage prototypes to better understand the potential of non-volatile memory as a form of long-term data storage. Moneta, their product, is a 64 GB storage array with simulated PCRAM and is PCIe-attached. A lot of thought went into designing its software-hardware interaction. Additionally, the groundbreaking

PCRAM-based solid-state drive (SSD) Onyx was manufactured. A memory monitoring-based approach to creating a morphable memory system; including a primary memory space with MLC and SLC sections built on PCRAM. Using on-chip caches built from non-volatile memory technologies such as ReRAM and STTRAM might potentially improve writing speed. In order to reduce writing energy consumption and writing mistakes, they built last-level caches using STTRAM. The authors concluded by discussing the hybrid cache hierarchy, a memory architecture that combines non-volatile and volatile components (Nagarajan et al., 2021).

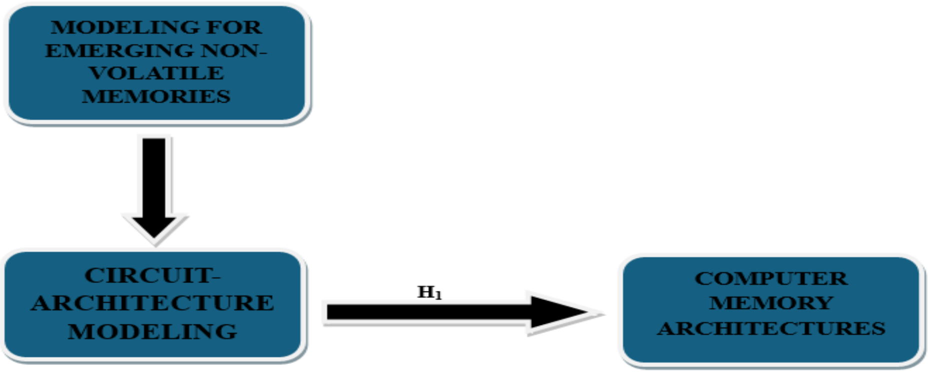
RESEARCH QUESTION

1. What constitutes non-volatile memory in computer architecture?

RESEARCH DESIGN

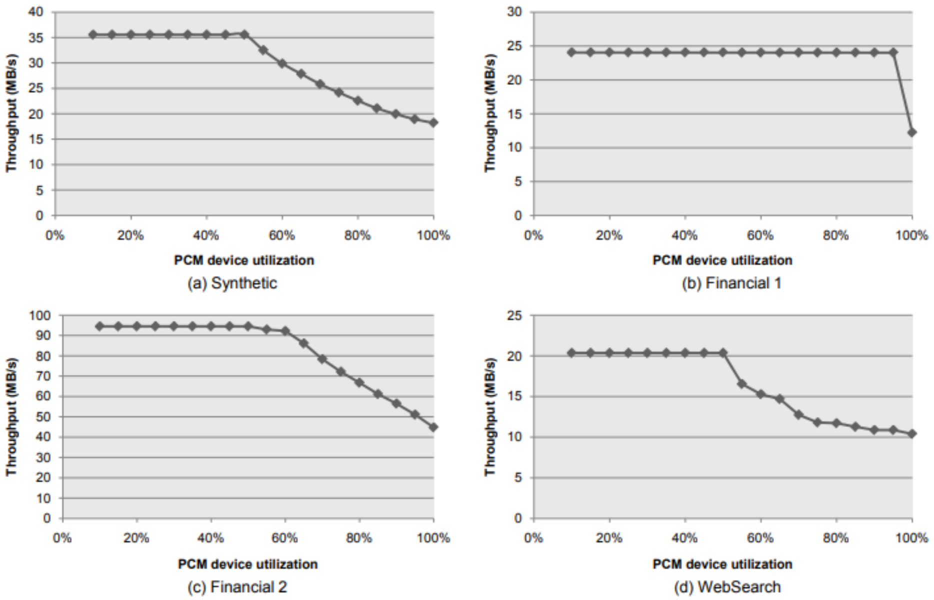
There is evidence that a Multi-Level Cell (MLC) can store multiple bits of digital data, making it a viable option for electronic non-volatile memory (eNVM). Since NAND flash technology already has MLC capabilities but no easy way to scale to denser arrays, this update has made eNVM a formidable competitor in the market. In comparison to their SLC counterparts, eNVMs that include MLC capabilities, such as PCRAM and ReRAM, often exhibit a longer programming time and a lower cell lifetime. Therefore, it suggests a reconfigurable eNVM design that can switch between MLCs and SLCs with ease, taking into account the specifics of the workload and the required lifetime to maximise the massive MLC capacity and the fast SLC access speed. In order to keep things generalisable, researchers look upon MLC PCRAM as an example.

CONCEPTUAL FRAMEWORK


MODELING FOR EMERGING NON-VOLATILE MEMORIES
CIRCUIT-ARCHITECTURE MODELING
H_1
COMPUTER MEMORY ARCHITECTURES

RESULTS

In this part, we assess how the adaptive MLC/SLC method improves performance and prolongs the life of PCRAM devices when they are not in use at full capacity. In order to evaluate the suggested method on a real-world platform, they recorded the actual I/O trace on the Linux 2.6.32-23 kernel using a 2GB RAMDISK set up as Ext2 memory with a 4KB block size. First, they read 1,500,000 randomly produced documents ranging in size from 5 KB to 10 MB from RAMDISK, creating an artificial file system trail.



(a) Synthetic

(b) Financial 1

(c) Financial 2

(d) WebSearch

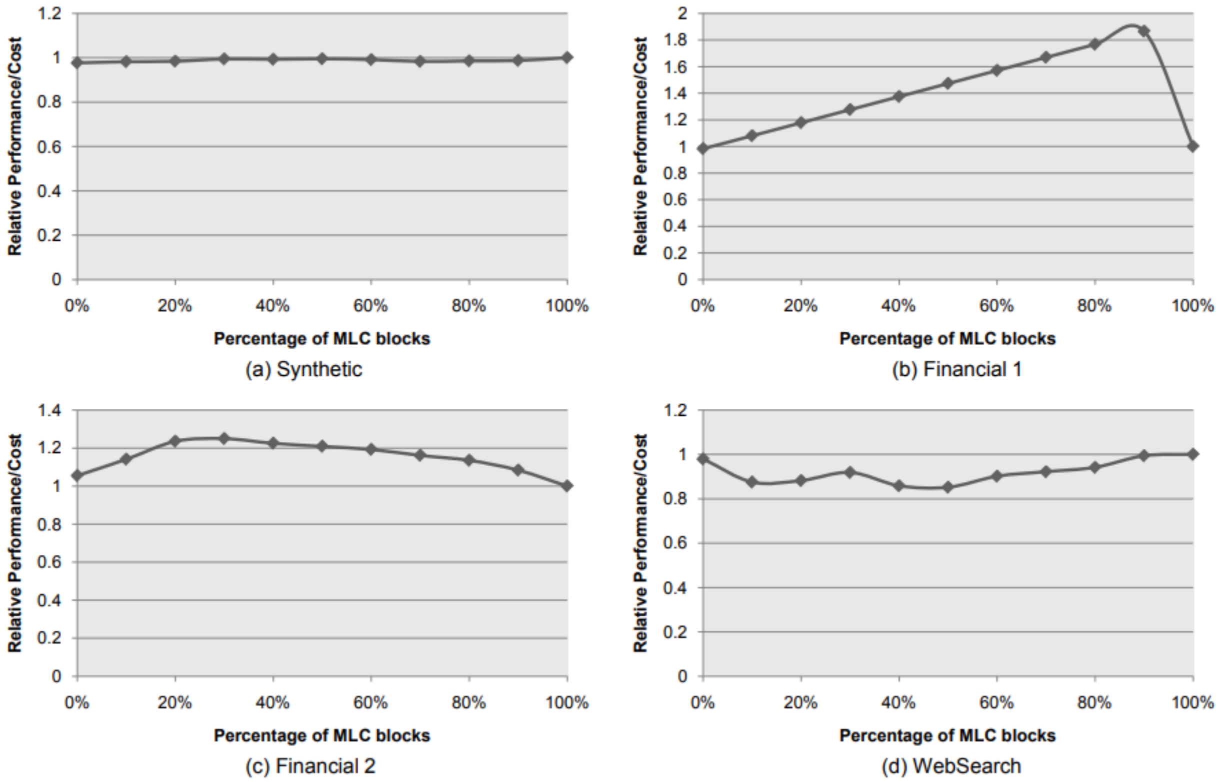
Figure 1. The Adaptive MLC/SLC Solution's Performance in Various Use Cases

In addition, they drew on disc records maintained by the Storage Management Council to simulate disc activity more accurately, allowing them to compete with enterprise-level applications like web servers, database servers, and websites. The synthetic trace is known as Synthetic, while the traces from SPC are referred to as Financial 1, Financial 2, or WebSearch.

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

In order to determine the read and results latencies in SLC or MLC modes, a prototype version of NVSim was used. The SLC write delay was calculated by adding the SET and RESET delays, whereas the MLC type latency was assumed to be four P&V steps to make things simple. A rounded numerical overview of the results of the computations is shown in Table 1. The reading and writing widths were both limited to 64 and 16 cells, respectively, because of the high current needed for the SET and RESET procedures.

The former was for SLC, which uses 64 bits, while the latter uses 128 bits. In SLC mode, the I/O bandwidth is almost twice as high as in MLC mode due to these assumptions.



(a) Synthetic

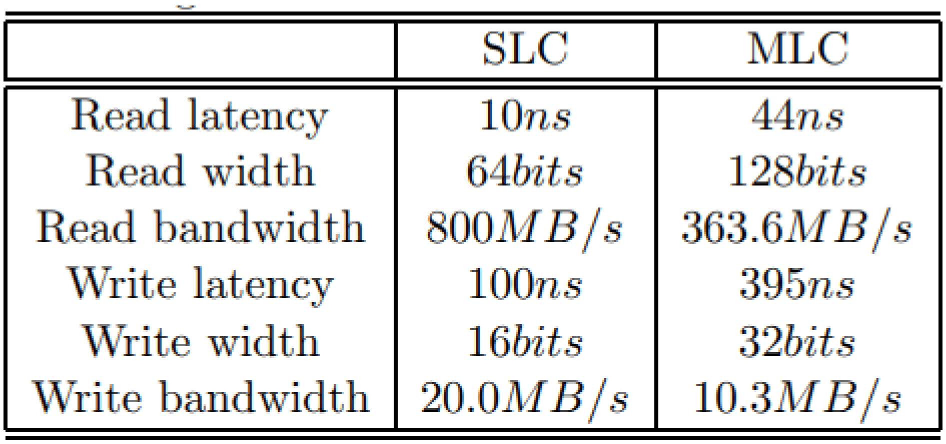
(b) Financial 1

(c) Financial 2

(d) WebSearch

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**

In order to increase speed, they start by demonstrating the performance-aware partitioning strategy in action. The I/O access distribution greatly affects the efficiency of the recommended adaptive MLC/SLC PCRAM. Data retrieval from the MLC regions is necessary for consistent access across the file system. Conversely, most often used data should be stored in SLC sections, while less frequently used files may be partitioned using MLC regions. Assuming the access pattern exhibits bias, this can be accomplished. The study found that by connecting many PCRAM devices in an array, a huge device capable of storing the set of operations may be formed. The amount of PCRAM chips is one component that affects the device's use. When all PCRAM blocks' device usage drops below 50% and 100%, respectively, a state transition between SLC and MLC modes happens. Utilisation decreases between 50% and 100%, at which point the adaptive MLC/SLC method supplies the required capacity. The relationship between different workloads and average throughput as a function of PCRAM device usage is shown in Figure 1. The throughput of Synthetic and Financial 2 gradually decreases as the usage rises. This is the outcome of the fact that these two tasks depend significantly on data stored locally. Additionally, this impact is magnified in Financial 1 because of the high number of files that are needed just once. These workloads uniformly distributed I/O access pattern cause WebSearch's performance to decrease precipitously at 50%.

- **Cost-Efficiency Evaluation**

Assuming the price of a PCRAM chip remains constant, they may use their previous finding on the correlation between performance and device utilisation to infer the relationship between performance and cost. Two extreme configurations are available under Financial 1's workload access pattern. One utilises 94 MB/s bandwidth from SLC-only PCRAM chips, while the other uses 44 MB/s bandwidth from MLC-only PCRAM chips, while using half the amount of PCRAM chips. In Figure 2, we can see the throughput-per-cost metric at its most advantageous, and in Figure 1, we can see the conclusion rephrased for our research of this measure. The usual gain in throughput-per-cost is around 28 percent.

- **A Lifetime Evaluation**

When an MLC PCRAM cell has too many RESET operations, its RESET/SET resistance margin decreases. Therefore, it must be established as an SLC at that time. Starting the lifetime-aware partitioning approach entails initialising the PCRAM chunks in the

MSB bank to MLC type and leaving the blocks in the LSB bank empty. Upon detecting that the accessing probability of a certain block is increased, the operating system monitor activates the matching blocks in the LSB banks to convert that block to SLC mode. According to the lifetime model, the lifetime-aware partitioning method has the potential to provide 100 lifetime benefits. That greatest gain in longevity is due to the lower device capacity.

DISCUSSION

The proposed adaptive MLC/SLC method takes workload characteristics and lifetime requirements into account, making advantage of the massive MLC capacity and fast SLC access speed. This section of the dissertation concludes the development of the MLC/SLC mode management approach, which follows the designs of an adaptive MLC/SLC eNVM array at the circuit level. A case study with a PCRAM-based storage device is used to evaluate the appropriateness of the proposed adaptive MLC/SLC method. An analysis of four real-world I/O traces revealed that the adaptive MLC/SLC method may potentially enhance 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).

CONCLUSION

The computer design community is intrigued by forthcoming non-volatile memory technologies such as STTRAM, PCRAM, and ReRAM due to its high density, great scalability, rapid access, and lack of volatility. These technologies have been present for almost 30 years, but they have finally disrupted the conventional memory hierarchy and threatened SRAM and DRAM. This dissertation presents case examples at several levels of application, models for improved design at the architectural level, and models for measuring area, energy, and performance at the circuit level. Although many types of non-volatile memory are still in the prototype stage, the first part of the dissertation constructs and specifies models for energy, area, performance, and circuit level (Oliveir et al., 2023). The limitations of write operations in non-volatile memory were addressed in the second part by creating techniques at the architectural level. To prove these techniques worked, they conducted evaluations at the architectural level. Checkpointing, memory hierarchy, and secondary storage are just a few of the many applications that these case studies show how non-volatile memory technologies may enhance power economy or performance. In light of the results of these case studies, it is reasonable to assume that non-volatile memory technologies will eventually replace older memory and disc technologies. This could speed up the development of these technologies and add to the revolution in non-volatile memory that is now taking place (Huang & Wang, 2023).

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.