EXPLORATION OF SOT-MRAM BASED DEVICES IN MEMORY HIERARCHY FOR NEXT-GENERATION COMPUTING

Thesis

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

KALLINATHA H D

Reg. No. 187083CO002

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA (NITK)
SURATHKAL, MANGALORE - 575 025
October 30, 2024

DECLARATION

By the Ph.D. Research Scholar

I hereby *declare* that the research thesis entitled **"Exploration of SOT-MRAM Based Devices in Memory Hierarchy for Next-Generation Computing"**, which is being submitted to the ***National Institute of Technology Karnataka, Surathkal*** in partial fulfilment of the requirements for the award of the Degree of ***Doctor of Philosophy*** in ***Computer Science and Engineering*** is a *bonafide report of the research work carried out by me.* The material in this thesis has not been submitted to any University or Institution for the award of any degree.

KALLINATHA H D
(187083CO002)
Department of Computer Science & Engineering

Place: NITK, Surathkal

Date: October 30, 2024

CERTIFICATE

This is to *certify* that the Research Thesis entitled **"Exploration of SOT-MRAM Based Devices in Memory Hierarchy for Next-Generation Computing"**, submitted by **Kallinatha H D** (Register Number:187083/187CO002) as the record of the research work carried out by him, is *accepted* as the *Research Thesis submission* in partial fulfilment of the requirements for the award of degree of ***Doctor of Philosophy***.

30/oct/24

Dr. Basavaraj Talawar
Research Supervisor
Associate Professor
Department of CSE
NITK Surathkal - 575025

30/10/24

Dr. Manu Basavaraju
Chairperson - DRPC
(Signature with Date and Seal)

Department of CSE

NITK Surathkal - 575025

DUGC / DPGC / DRPC
Dept. of Computer Science & Engineering
NITK- Surathkal
Srinivasnagar - 575 025

ACKNOWLEDGEMENTS

This thesis is a culmination of hard work, support and patience. During this journey, numerous individuals contributed to the completion of my research. I owe profound thanks to my supervisor, Dr Basavaraj Talawar, Associate Professor in the Department of Computer Science and Engineering at NITK, Surathkal, who provided me with the opportunity to pursue my PhD. His guidance, support, and encouragement have been invaluable.

I submit my reverential pranamas at the lotus feet of His Holiness *Dr Sri Sri Sri Sivakumara Mahaswamyjigalu*, Founder President and the current President, His Holiness *Sri Sri Siddalinga Mahaswamyjigalu*, Siddaganga Math, expressing my deepest gratitude for their blessings and support throughout this journey. I am deeply grateful to the Sree Siddaganga Education Society, Tumkur, for providing me with the opportunity to pursue my PhD at NITK, Surathkal. I extend my heartfelt appreciation to the management and principal of the Siddaganga Institute of Technology (SIT), Tumkur. My colleagues at the Dept. of CSE, SIT Tumkur, have been supportive I express my appreciation for them.

I am grateful to my Research Progress Assessment Committee (RPAC) members, Dr Jeny Rajan and Dr A.V. Narasimhadhan, for their suggestions and continued support in refining my research.A special thank you to Prof. Manu Basavaraj, Head of the Department of CSE, and all former Heads of the Department. I also appreciate the support provided by Prof. Ravi B., Director of NITK Surathkal, and past directors, department heads, and members of the DRPC.

My thanks extend to the teaching, technical, and administrative staff, friends, and fellow research scholars in the Department of CSE at NITK. I am thankful for the assistance from the library and the academic section of NITK during my tenure as a research scholar. My research lab, SPARK, has been a supportive and enjoyable environment, enhanced by the camaraderie among colleagues and friends.

I appreciate all current and former research group members, including Mrs Sadhna Rai, Dr Bhemmappa Halavar, Dr Khyamling Parane, Dr Prabhu Prasad, Dr Pramod, and Dr Anil Kumar, for their stimulating discussion throughout my stay at NITK. I extend my heartfelt thanks to my friends who have been an integral part of this journey, including Dr PrurshottamT.L. and family, Dr Somesh, Dr.Druva Kumar and

family,Sachin D.N., Dr Praveen Ramteke, Sarswathi, Manjunath, Sneha, Dr Vishal Rathod, Dr. Rashmi, Dr Shubham, Mrs.Spoorthi, Basavaraju K. S., Sandeep M., Aarabhi Putty, Ramu S., Moulya D M., Anusha Hegde.,Sushma,Keerthan Kumar and many other friends. Your companionship, encouragement, and unwavering support have been invaluable throughout my stay at NITK.

I must mention Mr. Karunakar, the Guest House in charge, for his exceptional support during the final stages of my stay at NITK. The NITK Health Care Center deserves thanks for providing necessary medical care during my stay.

I am profoundly thankful for the blessings of my late father, Mr Doddaiah B., a teacher par excellence. My heartfelt thanks go to my mother, Smt. Nagarathnamma K.G. for always motivating me to get an education, and my siblings, whose support has been a constant in my life.

Finally, my wife, Mrs. Sahana T.S., and my daughter, Ms.Manogna D.K., have been my pillars of strength. Their love and support have been constant in my endeavours. My wife's PhD journey at NITK did not deter her from supporting my work extensively. I am also grateful to my father-in-law, Mr. Sadashivaiah T.S., and my mother-in-law, Smt. Umadevi and my brother-in-law, Mr Girish T.S., thank you for your incredible support during my time at NITK, Surathkal.

This thesis could not have been completed without the collective support of my family, friends, research supervisor, and well-wishers, who have contributed directly and indirectly to my research work.

Kallinatha H D

Place: Surathkal

Date:October 30, 2024

ABSTRACT

The "Memory Wall" problem, which refers to the increasing gap between CPU processing speed and memory access time, is a significant challenge in modern computing systems. Traditional Static Random Access Memory (SRAM) caches have limitations such as high area, leakage power and scalability issues, especially as Complementary Metal-Oxide-Semiconductor (CMOS) technology scales. This necessitates exploring alternative memory technologies.

Spin-Orbit-Torque Magnetic RAM (SOT-MRAM) stands out due to its separate read and write paths, lower leakage power, and higher endurance than Spin-Transfer-Torque Magnetic RAM (STT-MRAM) and other non-volatile memory technologies.

The first part of the thesis work proposes a Multi-Factor Scaling (MFS) framework for utilizing SOT-MRAM as an alternative to SRAM caches to address the absence of a scaling road map and structural influence of memory cells. The focus is on applications such as Artificial Intelligence (AI), Natural Language Processing (NLP), and general-purpose workloads. This work introduces an advanced MFS framework for evaluating the impact of SOT-MRAM density replacement in cache memory design for power performance improvement. The framework includes Design Space Exploration (DSE) across various scaling scenarios, comparing SRAM and SOT-MRAM configurations.

Second, the work proposes and investigates the achievable Relative Lifetime Improvement (RLI) through the use of a Physical Split Cache (PSC) with Virtual Reordering (VRO), which dynamically manages Write Variation (WVAR) to extend cache lifetime and reliability. The PSC design leverages the advantages of both technologies: the high-speed access of SRAM and the energy efficiency and non-volatility of SOT-MRAM. The VRO algorithm optimizes cache performance by dynamically reordering cache lines to balance write distribution and enhance endurance.

Finally, this research work explores the integration of SOT-MRAM as an alternative to DRAM in main memory systems, targeting embedded systems and multi-core environments. The lack of publicly available parameters for SOT-MRAM poses a challenge in evaluating its performance with reliable simulations. Therefore, micro-architectural DSE and comprehensive full-system simulations are employed to derive

and validate the necessary parameters for robust analysis of SOT-MRAM-based memory systems under various configurations and capacities.

The research methodology involves evaluating benchmark programs representing diverse application domains to assess the performance, energy efficiency, and reliability of SOT-MRAM compared to traditional memory technologies. The findings show that SOT-MRAM significantly improves power efficiency, reduces latency, and enhances overall system performance, making it a compelling alternative for modern computing systems. In summary, this thesis presents an end-to-end framework for rapidly evaluating and adopting SOT-MRAM in cache and main memory designs. The comprehensive analysis and simulation results emphasize SOT-MRAM's potential to overcome the limitations of SRAM and DRAM, providing scalable and efficient memory solutions for advanced applications. The insights gained from this research lay the groundwork for future developments in the memory hierarchy, ensuring that computing systems can meet the increasing demands of AI, NLP, and other data-intensive workloads.

Keywords: Memory, Cache, NVM, MRAM, Hybrid, Main Memory.

Dedication

I dedicate and submit this thesis to Bhagwan Sri Krishna and Sri Kalleshwara Swamy for divine guidance and support.

I also extend my heartfelt gratitude to all my teachers, from school through to my PhD, for their invaluable guidance and wisdom.

Contents

List of Figures

List of Tables

Abbrevations

AIoT Artificial intelligence Internet of Things

NVM Non Volatile Memory

PSC-VRO Physically Split Cache with Virtual Reordering

MRAM Magnetic Random Access Memory

STT-MRAM Spin-Transfer-Torque Magnetic RAM

SOT-MRAM Spin-Orbit-Torque Magnetic RAM

PCM Phase Change Memory

ReRAM Resistive Random Access Memory

AI Artificial Intelligence

NLP Natural Language Processing

ANN Artificial Neural Network

SNA Social Network Analytics

MFS Multi-Factor Scaling

LI Lifetime Improvement

PSC Physical Split Cache

VRO Virtual Reordering

WVAR Write Variation

DSE Design Space Exploration

SRAM Static Random Access Memory

DRAM Dynamic Random Access Memory

MTJ Magnetic Tunnel Junctions

IoT Internet of Things

LLC Last-Level Cache

PPA Power, Performance, and Area

EDP Energy-Delay Product

PL Pinned Layer

FL Free Layer

TB Tunnelling Barrier

RBL Read Bit Line

RWL Read Word Line

SL Source Line

BL Bit Line

HM Heavy Metal

WBL Write Bit Line

WL Word Line

WWL Write Word Line

SHE Spin Hall Effect

VLSI Very Large Scale Integration

CMOS Complementary Metal-Oxide-Semiconductor

MSHR Miss Status Handling Registers

PoLF Probabilistic Set Line Flush

SWWR Static Window Write Restriction

DWAWR Dynamic Way Aware Write Restriction

HP High Performance

LOP Low Power Performance

CA Cell Area

AR Aspect Ratio

RLI Relative Lifetime Improvement

HPC High Performance Computing

Chapter 1

Introduction

The "Memory Wall" problem is the increasing disparity between the processing speed of CPU cores and memory access latency(McKee, 2004). The problem was addressed by implementing larger multilevel cache memories (Gholami et al., 2024). Static Random Access Memory (SRAM) has been the dominant form of cache memory technology because of its superior performance compared to other computing memory technologies (Alhalabi et al., 2018). Today's Chip Multiprocessor systems heavily rely on extensive on-chip cache hierarchies to enhance system performance by minimizing memory access delay and the need to access off-chip memory with high latency(Smith, 1982). The cache sizes(Fig.1.1) of Intel core have grown from 1MB to 32, AMD core cache sizes have grown from 2MB to 512MB and NVIDIA from less than 1MB to 64MB from generation to generation(wik, b,a,c; Inci et al., 2022).

1.1 Cache

Last Level Cache(LLC) sizes have been consistently increasing over the past few CPU/GPU generations(Fig.1.1). The down-scaling or size reduction of CMOS devices has made it possible to increase the size of SRAM while keeping the proportion of total chip area to on-chip cache memory area and static power leakage in line with Moore's law(Schwierz and Liou, 2020; Liao et al., 2020). The increasing leakage power at feature sizes below 45nm makes up around 60% of total energy consumption due to quantum tunnelling effects(Cargnini et al., 2014). Meeting the need for greater capacity has become more difficult as the SRAM is reaching lithography scaling lim-

its, rising leakage currents, and having process variation problems with down-scaling in CMOS devices. The increased leakage power limited the processor's performance to just a few gigahertz, presenting significant challenges when using SRAM as the primary technology for on-chip cache design (Mittal et al., 2015; Inci et al., 2022).

The growth of Artificial Intelligence (AI), Natural Language Processing (NLP), Social Network Analytics(SNA), Graph and general computing applications have led to demand for larger cache capacities to handle increasing data size, workloads, and computational complexities(Inci et al., 2022). SRAM alone cannot fulfil this demand due to its larger area and leakage power problem.

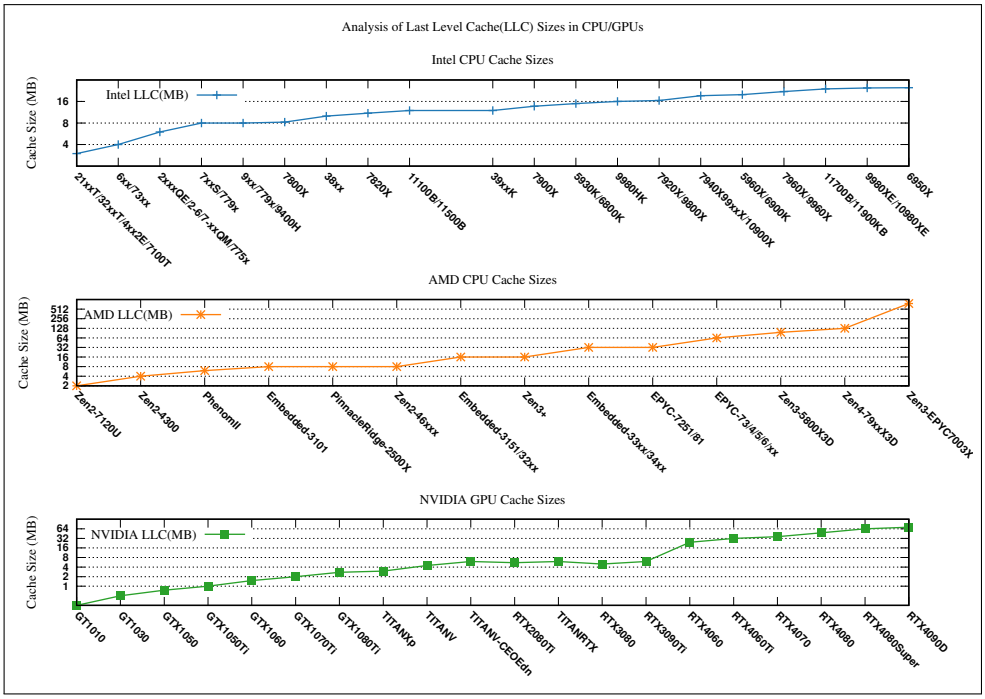


Figure 1.1: LLC Cache Capacity in CPUs/GPUs(wik, b,a,c; Inci et al., 2022).

The attributes of the SRAM cache change depending on its size. A larger cache increases the likelihood of finding a requested item, improving hit rates. However, larger caches also lead to higher access latency and longer processing times. Additionally, they require more power and occupy extra space on the chip. Therefore, when designing a system with an SRAM cache, it is important to consider the trade-offs between cache size, hit rate, access latency, power consumption and chip area(Oboril et al., 2015; Evenblij et al., 2019). On the other hand, increasing the LLC capacity can reduce off-chip memory access by up to 24%(Fig.-1.2). The SRAM cache problem is exasperated by the slow growth of main memory speed compared to server technology

speed. All these factors led to the exploration of SRAM alternatives in cache memory.

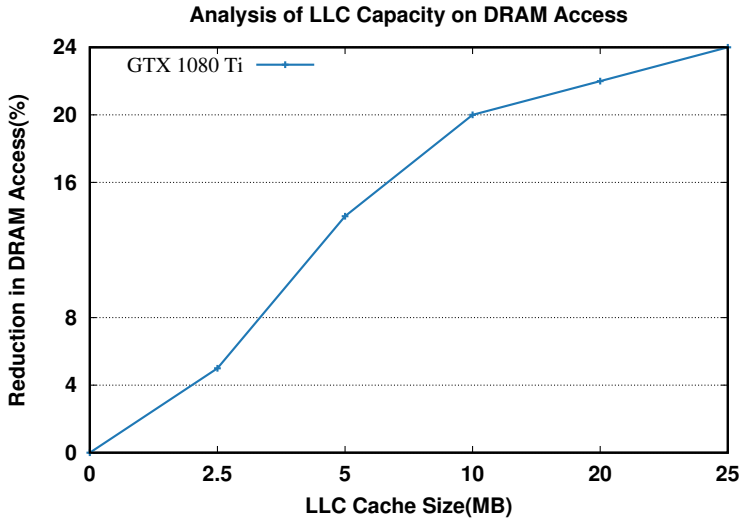


Figure 1.2: Reduction in DRAM Access with LLC Cache Capacity (Inci et al., 2022).

Researchers have been exploring alternative memory technologies to address the limitations of CMOS-based SRAM. One promising alternative is the use of Non-Volatile-Memory(NVM). MRAM has recently gained popularity because of its non-volatility, high speed and density(Wang et al., 2019). MRAM uses Magnetic Tunnel Junctions(MTJ) to store data. STT-MRAM is an NVM-MRAM that uses current to change tunnel junctions' magnetic/spin polarization. It uses the STT effect to switch the orientation of MTJs. There are two variants of MRAM, electron spin-based STT-MRAM and SOT-MRAM. They have gained attention because they are byte-addressable, non-volatile, highly dense and ultra-low leakage power characteristics (Singh et al., 2020). These emerging NVM technologies can potentially replace SRAM in lower levels of the computer systems memory hierarchy (Inci et al., 2022). STT-MRAM has higher write energy consumption and longer latency than SRAM, especially for write-intensive workloads (Singh et al., 2020). Also, due to the common read/write mechanism used in STT-MRAM, it suffers from read disturbance and limited endurance problems, which can further impact its performance. To address these concerns and explore the potential of SOT-MRAM in on-chip caches, this work proposes an alternative framework to cache design and lifetime extension of cache. SOT-MRAM cells vary in area, aspect ratio, read/write mechanism, cache access method, optimization target, device type, and memory array arrangement. The choice of SOT-MRAM cells and configuration significantly impacts cache power and performance.

Choosing, designing and evaluating MRAM cache performance for modern applications with multiple influencing factors requires an integrated framework solution.

1.2 LLC lifetime extension.

Endurance is a limiting factor when adopting the NVM in the memory hierarchy. The need for cache management and endurance extension mechanisms and their effect on the cache performance and lifetime improvement are unknown at design time. The Design Space Exploration(DSE) of the new lifetime extension mechanism for fault-tolerant cache design for various systems like IoT requiring a small LLC cache to CPU/GPU requiring several MB of LLC cache in one framework is time-consuming (Han and Jiang, 2023; Escuin et al., 2023). So, hybrid design is used to overcome these issues. In the hybrid cache design, the hierarchy is partitioned with SRAM as the smaller part of the cache for low-latency and high-endurance access. SOT-MRAM is used for the larger part of the cache for higher density and lower leakage power. The hybrid cache design approach aims to balance the performance, energy efficiency, and reliability of on-chip caches, leveraging the advantages of both SRAM and SOT-MRAM technologies in LLC.

1.3 SOT-MRAM as DRAM Alternative

DRAM, the popular memory technology, faces issues with power and scalability when scaled to large sizes. A significant limitation of DRAM is that when scaled, the capacitors are more vulnerable to errors due to their reduced size (Asifuzzaman et al., 2022). The primary bottleneck for system performance is the memory bandwidth. Over the past 20 years, server hardware FLOPS have been increasing at 3.0×/2 years, outpacing the growth rates of DRAM and interconnect bandwidth, which have advanced by only 1.6 and 1.4 times every two years, respectively. This difference has made memory the primary obstacle in AI applications(Gholami et al., 2024). Traditional cache memory technologies such as SRAM and DRAM face restrictions regarding packaging density and power leakage(Jia et al., 2017). Numerous alternatives have been suggested as substitutes for the current DRAM memories based on CMOS technology for

multiprocessor systems. Replacing DRAM with Non-Volatile Memory (NVM) is the most promising option. NVM technologies that could be used as a replacement for traditional DRAM memory are Phase Change Memory (PCM), Spin-Transfer-Torque Magnetic RAM (STT-MRAM), and Resistive RAM (ReRAM). These technologies offer byte-addressability, non-volatility, and fast read times. STT magnetization-based switching memory was first introduced in (Hosomi et al., 2005), leading to the genesis of several memory technologies based on this method. STT-MRAM technology has a similar capacity, frequency, and device size as DRAM(Asifuzzaman et al., 2022), making it a potential replacement for DRAM. However, due to the common path shared by read as well as write operations, STT-MRAM suffers from the read-write disturbance problem (Alhalabi et al., 2018) (Wang et al., 2019).

SOT-MRAM is being widely explored as a potential technology for cache memory applications. SOT-MRAM boasts separate paths for the read and write process, rendering it an attractive and efficient memory device. The potential of SOT-MRAM as a low-power, high-speed spintronic device for on-chip memory and main memory in High Performance Computing (HPC) and AI applications is highlighted by Zheng et al. (Zheng et al., 2021). Reliable evaluation of SOT-MRAM-based main memory is challenging due to the public unavailability of parameters related to timing and current. The significant advantage of MRAM devices is scalability, and they do not require any refresh, reducing static power consumption. STT-MRAM has been explored as the main memory in recent research (Komalan et al., 2018; Asifuzzaman et al., 2022) providing promising results. The average speedups of Open and Close Page Techniques with a 1.2x configuration, as mentioned in (Asifuzzaman et al., 2022), are 2.4% and 2.7%, respectively. Additionally, there was 4.6 times more power consumption than DRAM. One setback for STT-MRAM is its write latency, write energy and reliability (Komalan et al., 2018). SOT-MRAM provides better read/write latency, retention time, reliability and endurance than STT-MRAM (Komalan et al., 2022) (Zheng et al., 2021).

SOT-MRAM has not yet been evaluated as the primary memory for embedded and multi-core systems using full system simulators and benchmark programs, such as PARSEC (Bienia et al., 2008a). Given its attractive features, SOT-MRAM has the potential to serve as the main memory. This research seeks to propose and explore

its behaviour in a multi-core environment when handling shared memory workloads. The SOT-MRAM main memory's timing and current parameters are derived using the approach in (Asifuzzaman et al., 2022). The authors in (Asifuzzaman, 2019; Asifuzzaman et al., 2022) have validated the method and the parameters from industry estimations.

1.3.1 Problem Statement

This work aims to develop a comprehensive Multi-Factor Scaling (MFS) framework for SOT-MRAM on-chip caches to enhance modern application performance. The thesis focuses on designing and implementing a Physically Split Cache (PSC) that combines SRAM and SOT-MRAM at the Last Level Cache (LLC) with algorithms to reduce write variation and improve lifetime. It explores replacing DRAM with SOT-MRAM in main memory systems for embedded and multi-core environments to improve memory performance and energy efficiency.

1.3.2 Objectives

1. An End-to-end Framework for Multi-Factor Scaling(MFS) Design of Non-Volatile SOT-MRAM On-chip Caches for Modern Applications.

2. Design and Implementation of Physically Split Cache(PSC) with SRAM and SOT-MRAM for LLC with Write Variation and Lifetime Improvement(LI) algorithm.

3. Integration and Investigation of SOT-MRAM as an alternative to DRAM in Main Memory for Embedded Systems and Multi-Core Environment.

1.4 Contributions

SOT-MRAM memories bring a paradigm shift in the existing memory hierarchy. Hence, it is necessary to perform a scaling roadmap analysis, design and evaluate a LI mechanism and Integrate SOT-MRAM in the main memory.

1.4.1 Solution Approaches

The need for a new memory hierarchy is studied, and experiments are conducted to study the behaviour of SRAM, DRAM, and SOT-MRAM at the circuit, architectural, and application levels. Cache and main memory simulation analyse the area, power, and performance trade-offs in the memory hierarchy.

The proposed approach leverages a multi-factor scaling strategy that combines various techniques to enhance the performance of SOT-MRAM caches and improve lifetime by utilizing cache management policies and hybrid cache designs that reduce the number of writes to SOT-MRAM cells.

DRAM replacement by SOT-MRAM is designed by considering the parameters of the DDRx protocol. Deriving parameters and designing SOT-MRAM main memory involved DSE of circuit-level main memory analysis tuning and then sensitivity analysis of timing and current parameters for reliable simulations.

1.4.2 Thesis Contribution

This work is dedicated to improvising the LLC and main memory, and the following are the contributions:

1. This thesis work evaluates the scaling roadmap for SOT-MRAM within the context of Last-Level Cache (LLC).

2. The proposed MFS framework offers a comprehensive integration, providing an end-to-end evaluation flow from device-level survey to micro-architecture, application, and cache management policy evaluation.

3. We propose a reliable PSC design featuring a Life Improvement (LI) and Write Variation algorithm for enhanced cache lifetime management.

4. Extensive multi-level integrated simulations with relevant workloads, we deliver a thorough evaluation encompassing Power, Performance, and Area (PPA), to-

tal latency, application power consumption, bandwidth utilisation, Energy-Delay Product (EDP), Write Variation (WVAR) reduction, and RLI.

5. Integration of SOT-MRAM into hybrid or stand-alone DDRx standard for embedded and multi-core systems, extending its application scope.

6. A comprehensive survey and micro-architectural exploration to address SOT-MRAM main memory parameter absence.

7. To overcome the challenge of parameter absence, the work estimates and scales parameters, ensuring a robust analysis of three memory structures and making the parameters available through publication.

8. Thorough system-level simulations provide a holistic evaluation, covering power consumption, bandwidth utilization, EDP, and total latency.

1.5 Organisation of the Thesis

The rest of the thesis is organised as follows:

Chapter 2: summarizes the research related to three major issues addressed by the thesis , i.e. MFS with density replacement study, hybrid LLC design and implementation of LI algorithm, and SOT-MRAM main memory.This chapter explores the fundamentals and evolution of memory technology. It discusses the core principles and various NVM technologies. The chapter concludes with a literature review & research gaps, setting the stage for the thesis's contributions.

Chapter 3: Gives details about MFS design framework for end-to-end evaluation of the SOT-MRAM cache with a density replacement study for modern applications.

Chapter 4: Presents the Physically Split Cache with Virtual Reordering (PSC-VRO), a hybrid cache design for reliability and a WVAR algorithm for lifetime extension. It provides insights about how the inter/intra set WVAR can be used to improve lifetime of SOT-MRAM cache segment.

Chapter 5: Provides details about the need for SOT-MTAM main memory design, integration and investigation of the DSE of SOT-MRAM, deriving SOT-MRAM main memory parameters for embedded and multi-core systems. A thorough full system simulations using multi-threaded shared memory workloads.I also give insignts about the factors affecting main memory design.

Chapter 6: Conclusions, the contributions of this thesis, along with some important conclusions, have been summarized with future research directions.

Chapter 2

Background and Literature Review

Chapter 1 of this thesis highlighted that the conventional SRAM LLC in current Chip Multi-Processors(CMPs) are generally shared and large but suffer from scalability, density, and leakage power consumption issues. This thesis aims to leverage emerging MRAM technologies for next-generation workloads . The thesis addresses their main challenges, such as limited write endurance and costly write operations. This chapter begins by outlining the basics of the current memory systems and the emerging NVM, discussing the challenges associated with integrating NVMs into the cache hierarchy. Subsequently, the latest advancements in SOT-MRAM technology, including scaling strategies, life extension techniques through wear levelling, and the potential integration of SOT-MRAM in main memory systems, are reviewed.

2.1 Conventional Memory Technology Based on Charge

An ideal memory technology should be fast, dense, reliable, energy-efficient, and affordable. However, only some types meet all of these criteria. For instance, SRAM is quick but expensive and power-intensive, whereas DRAM is cheaper and denser but slower and less reliable. Similarly, flash memory offers reliability and density but struggles with write endurance and latency. By strategically employing various memory types across the cache hierarchy, systems can maximize speed, energy efficiency, and cost-effectiveness. This method effectively harnesses the unique strengths of each memory type based on the principle of locality of reference, ultimately achieving balanced performance across the system. In the following subsections, a concise

explanation of conventional charge-based memories is given.

2.1.1 SRAM Cache Memory

The standard design of an SRAM cell for cache memory consists of six transistors, as shown in Fig.2.1. This schematic depicts an SRAM cell composed of four transistors (T1 to T4) that form two cross-coupled inverters responsible for storing binary data. These stored states remain stable as long as power (Vdd) is provided. Additional transistors (T5 and T6) are used for read and write operations. During a read operation, activating the word line (WL) allows the stored states to be read from the bit lines ($\overline{\text{BL}}$ and BL). The bit lines are set to the desired state for writing before the word line (WL) is activated to update the cell.

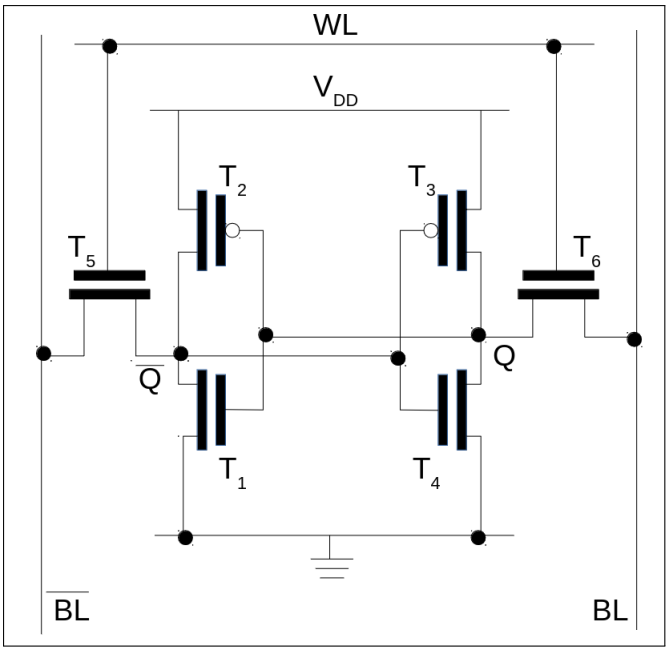


Figure 2.1: Bit Cell of SRAM(Agarwal and Kapoor, 2020)

In addition to its design specifics, SRAM possesses certain advantages and drawbacks worth mentioning:

- SRAM offers high-speed access; data becomes available immediately upon activation of the word line (WL).
- A conventional six-transistor SRAM design consumes more chip area, resulting in lower density than other memory technologies.
- Continuous power supply is necessary for SRAM to maintain data.

- On a cost-per-bit basis, SRAM is more expensive.

2.1.2 DRAM Main Memory

In Fig.2.2, a depiction of a DRAM cell, which consists of a capacitor and a transistor, is shown. The capacitor *C* is responsible for maintaining the cell's state as a charge, and access to this charge is facilitated through the transistor *T*.

Read Operation: Applying voltage to the access line *AL* enables current flow through the data line *DL* based on the charge present in the capacitor. If no charge is detected, there will be no current flow.

Write Operation: The data line *DL* is set to the desired state, followed by the application of voltage to *AL* to charge or discharge the capacitor accordingly. While DRAM offers the advantage of smaller feature sizes and lower costs than SRAM due to its simplistic structure, it is important to note that DRAM capacitors have limited charge retention. This necessitates frequent refresh operations, which consume additional time and energy.

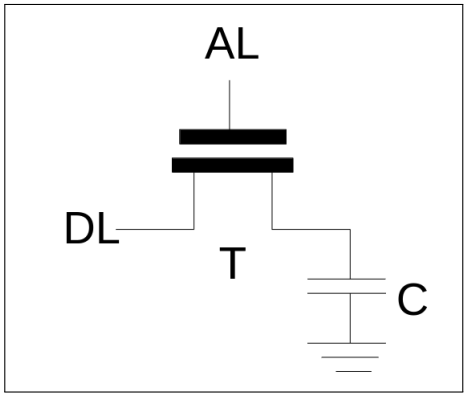


Figure 2.2: Bit Cell of DRAM(Agarwal and Kapoor, 2020)

The key features of DRAM are as follows:

- DRAM's simplified cell structure allows for high density and reduced cost per bit.
- DRAM access speeds are slower compared to SRAM.
- The need for frequent refresh operations to uphold data integrity leads to increased power consumption in DRAM.

2.2 Emerging NVM Technology

Currently, there is a significant amount of interest in various emerging NVM technologies. A detailed review paper (Kryder and Kim, 2009) provides an overview of twelve NVM technologies, such as STT-MRAM, Phase Change RAM (PCRAM), Resistive RAM (ReRAM), Ferroelectric RAM (FeRAM), Nano RAM (NRAM), Conductive-Bridging RAM (CBRAM), Single Electronic Memory (SEM), Polymer, Molecular, Racetrack, Holographic, and Probe memories. While some of these technologies have reached advanced stages of development and are commercially accessible, others are still in the early phases of exploration. For the reasons discussed in Chapter 1 section 1.1, this dissertation specifically focuses on SOT-MRAM in the memory hierarchy(LLC and Main memory), which has been extensively researched and is considered a promising solution for cache memory applications.

2.2.1 MRAM Background

- **STT-MRAM:** The cell structure of STT-MRAM is depicted in Fig.2.3(Lu et al., 2024), consisting of a Magnetic Tunnel Junction (MTJ) connected to an access transistor. The MTJ comprises two ferromagnetic layers separated by a thin insulating barrier made of magnesium oxide (MgO). One of these layers, called the Free Layer (FL) can change its magnetization direction using a spin-polarized current, while the other layer, known as the Pinned Layer (PL) remains fixed. The orientation of the layer's magnetization encodes the data: a parallel orientation represents a '0' state (low resistance), while an anti-parallel orientation signifies a '1' state (high resistance). To read from the STT-RAM cell, the access transistor is activated, creating a small voltage difference across the source and bit line. This voltage difference generates a current, which is then compared against a reference current to determine the stored state. During writing operations, a large voltage difference is applied; a positive voltage writes a '0', and a negative voltage writes a '1'.

- **SOT-MRAM:** SOT-MRAM is an NVM technology that works on the orbital spin torque of nanomagnets for high-speed data reads and writes. The SOT-MRAM bit cell, depicted in Fig.2.4(Lu et al., 2024), consists of an Magnetic

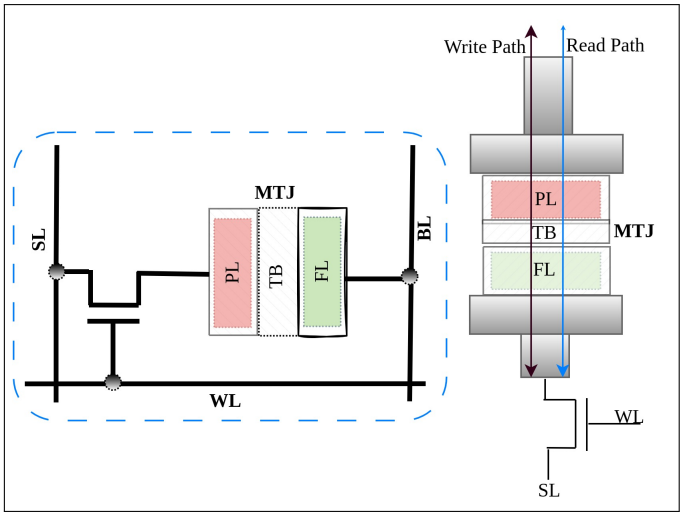


Figure 2.3: Schematic of STT-MRAM cell

Tunnel Junctions (MTJ) and a perpendicular magnetic anisotropy layer, allowing efficient magnetization switching and reliable data storage(Oboril et al., 2015)(Alhalabi et al., 2018).

- Heavy Metal (HM): The HM/SOT metal layer is composed of elements like platinum (Pt) or tantalum (Ta). This layer generates the spin current needed to switch the free layer.
- FL: The orientation of the magnetization in this layer, composed of magnetic elements such as cobalt-iron-boron(CoFeB), determines the stored bit value (0 or 1).
- Tunnelling Barrier (TB): The TB layer commonly consists of magnesium oxide (MgO); this insulating layer enables Tunnelling Magneto Resistance (TMR), a crucial component for data retrieval.
- PL: This layer is made of CoFeB and determines the magnetization orientation, serving as a reference point for the free layer.

SOT-MRAM and STT-MRAM differ in the free layer's magnetisation mechanism. In STT-MRAM, the spin-polarized current is directly passed through the magnetic layers to induce torque(change the direction of nanomagnets) as shown in Fig.2.3. In contrast, SOT-MRAM utilizes the charge current in an HM layer to generate a spin current via the Spin Hall Effect(SHE) (Shao et al., 2021). The current does not pass through the insulating barrier, thus reducing the risk

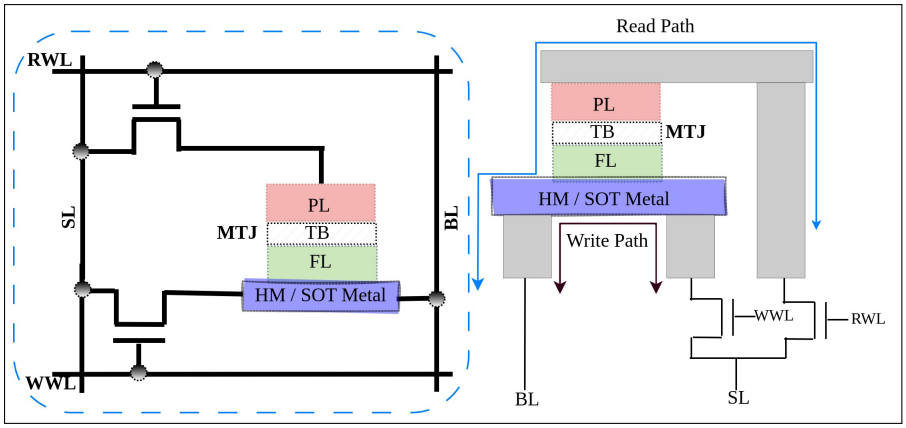


Figure 2.4: Schematic of SOT-MRAM cell

of damage to the SOT-MRAM bit cell while writing. This changed mechanism helps in faster write speeds for SOT-MARM than STT-MRAM. This approach reduces the direct disturbance to the magnetic layers, potentially enhancing endurance and reducing power consumption(Van Beek et al., 2023). As shown in Fig.2.4, separating read and write paths in SOT-MRAM can optimize read and write operations independently. This enhances its endurance, making it suitable for applications requiring frequent write operations.SOT-MRAM can achieve lower write power due to the efficient use of spin current for switching the magnetization(nanomagnets direction)(Seo and Kwon, 2020a; Gupta et al., 2020).

Fig.2.5 depicts the memory array organization of the MRAM cache, comprising n memory locations, each with m bits. Like SRAM arrays, they are structured into n-c rows and c columns. The word line decoder activates the addressed row, and the column decoder activates the bit cells(Shao et al., 2021; Singh et al., 2020). Despite SOT-MRAM cells having a larger footprint than STT-MRAM, they still occupy less area than SRAM cells (Shao et al., 2021). As depicted in the Fig.2.4 schematic, read/write operations on the SOT-MRAM bit and cache array work in the following ways:

1. To read the stored data, a small voltage has to be applied connecting the Bit Line (BL) and the Read Word Line (RWL). The resistance across the tunnelling barrier(i.e.TMR) is measured. If the FL magnetization is parallel to the PL, the resistance is low (indicating a stored '0'). The high resistance indicates that the

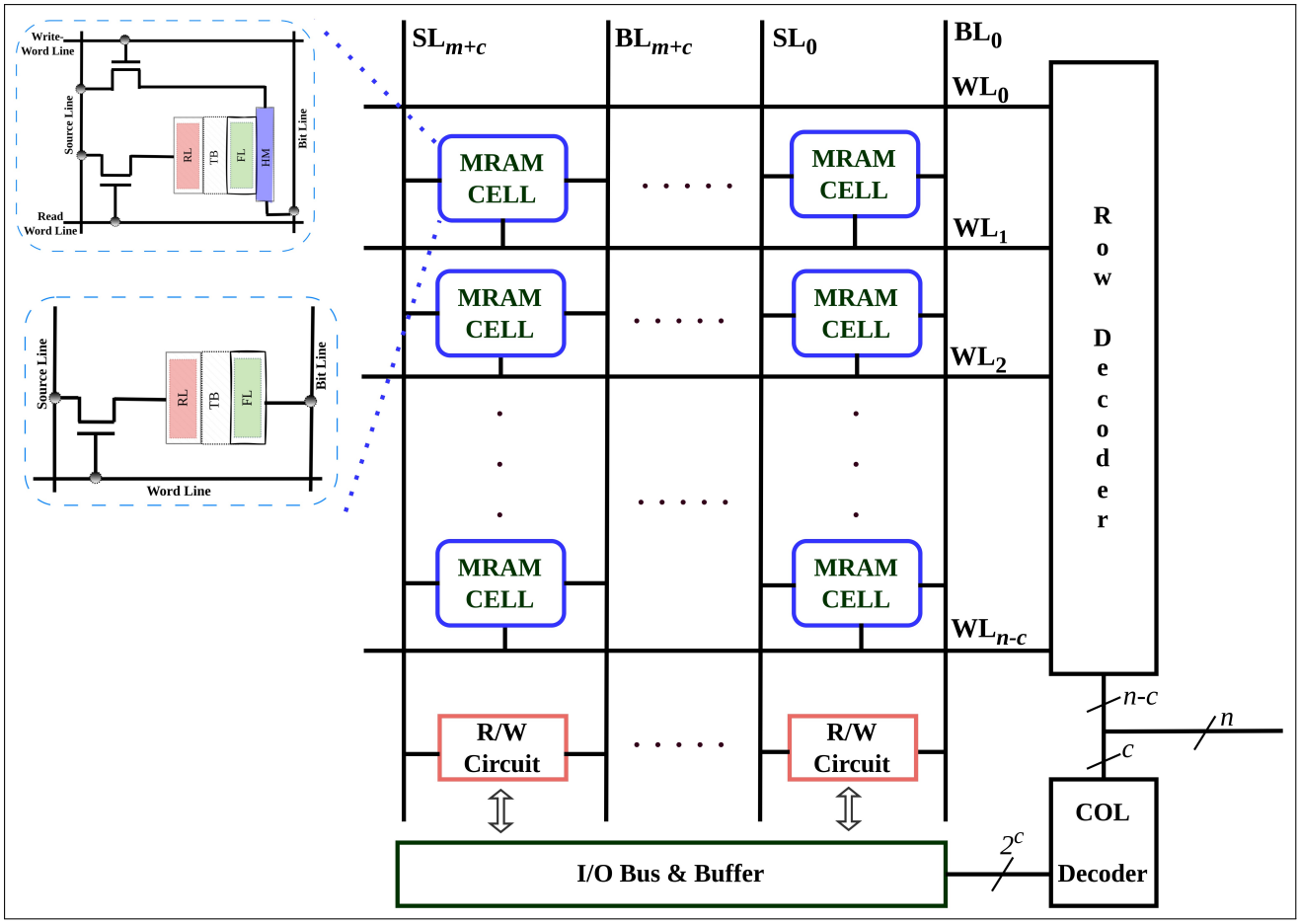


Figure 2.5: MRAM cache memory array organization.

anti-parallel magnetization storing '1'(Van Beek et al., 2023; Shao et al., 2021).

2. For writing data, the current drives through the HM layer via the Write Word Line (WWL) and the Source Line (SL). The current in the HM layer generates an orbital spin current due to the Spin Hall Effect (SHE) in HM. This spin current applies an orbital torque (SOT) to the FL. The direction of the spin current decides whether the magnetization of the FL switches parallel or antiparallel to the PL, storing a '0' or a '1'. This switching mechanism enables data storage using magnetic alignment to represent binary values(Van Beek et al., 2023; Shao et al., 2021).

The performance of both MRAMs depends on the retention time, which is determined by the energy required for magnetization stability. The thermal attempt frequency (taf_0) and magnetization stability energy height (Δ) influence retention time. Stability energy height is influenced by cell size but has an inverse relationship with temperature. Increasing cell size can extend retention time and lead to larger cell

area and capacitance, negatively impacting density and write energy. Using magnetic materials with higher magnetic anisotropy can enhance the magnetization stability energy height and prolong retention time without necessitating a larger cell size(Shao et al., 2021; Liao et al., 2020).

$$T_{\text{retention}} = \frac{1}{taf_0} e^{\Delta} \tag{2.1}$$

The parameter Δ is determined by factors such as volume (Vo), in-plane anisotropic field (A_k), saturation magnetization strength, and absolute temperature (Temp). MRAM acts as an NVM with near-permanent data retention abilities. According to production specifications from Intel and Samsung, MRAM can handle up to 100 million erasures at 105 °C and up to ten billion erasures at 85 °C. This is a significant improvement over SRAM, which can only manage up to one million cycles. Therefore, MRAM has a clear advantage over SRAM regarding retention time(Han and Jiang, 2023).

$$\Delta \propto \frac{Vo \times A_k \times SM_s}{Temp} \tag{2.2}$$

Integrating MRAM technologies into cache memory presents challenges due to reported variations in bit cell values in VLSI literature. These variations can impact parameters such as area, power consumption, and speed. Accurate modelling and prediction of these parameters are crucial for optimizing cache memory design. Operational parameters are crucial for understanding the dynamics of read and write operations when employing SOT-MRAM. The delay in reading, as adapted from (Liao et al., 2020), is expressed as:

$$t_{\text{READ}} = t_{\text{WL}} + t_{\text{sense}} \tag{2.3}$$

The WL delay time (t_{WL}) can be calculated using this equation:

$$t_{\text{WL}} = 0.7 R_{\text{drive}} C_{\text{WL}} + 0.4 R_{\text{WL}} C_{\text{WL}} \tag{2.4}$$

In this formula, R_{drive} refers to a driver resistance, which is defined as an inverter of 5× minimum size. R_{WL} denotes the resistance of the interconnect, while C_{WL} stands for the capacitance of interconnect .

The energy consumption during the read operation, denoted as E_{READ}, comprises several components and is expressed as:

$$E_{\mathrm{READ}} = 2V_{\mathrm{READ}} I_{\mathrm{bias}} t_{\mathrm{READ}} + E_{\mathrm{WL}} + E_{\mathrm{SA}} \tag{2.5}$$

In this context, the first term refers to the Joule heating produced by the currents flowing through the controlled and reference MTJs. E_{WL} denotes the energy required to charge the word line, which is computed as $(C_{\mathrm{WL}}/N_{\mathrm{bit}} + C_{\mathrm{tran}})V^2_{\mathrm{READ}}$, while E_{SA} refers to the sense amplifier energy dissipation, determined by $P_{\mathrm{SA}}.t_{\mathrm{READ}}$. The power of sense amplifier, P_{SA}, is derived from previous SPICE simulation results in(Shao et al., 2021; Liao et al., 2020).

For a write operation, the access time t_{WRITE} can be calculated by combining the word line delay, bit line delay, and magnetic switching time t_{mag}. The validity of t_{mag} has been confirmed through micromagnetic simulations in (Liao et al., 2020).

$$t_{\mathrm{WRITE}} = t_{\mathrm{WL}} + t_{\mathrm{BL}} + t_{\mathrm{mag}} \tag{2.6}$$

The energy required for writing, denoted as E_{WRITE}, is determined by taking into account the resistive effects and capacitance of the circuit.

$$\begin{aligned} E_{\mathrm{WRITE}} = {} & I^2_{\mathrm{WRITE}} \left(R_{\mathrm{BL}} + R_{\mathrm{SL}} + R_{\mathrm{tran}} + R_{\mathrm{SOT}}\right) \\ & \cdot t_{\mathrm{mag}} + \left(C_{\mathrm{WL}}/N_{\mathrm{bit}} + C_{\mathrm{tran}}\right) V^2_{\mathrm{dd}} + \left(C_{\mathrm{BL}} + C_{\mathrm{tran}}\right) V^2_{\mathrm{WRITE}} \end{aligned} \tag{2.7}$$

The comprehensive analysis highlights the operational intricacies of SOT-MRAM and underscores the significance of accurate parameter estimation in evaluating the memory's performance within contemporary technological paradigms.

The proportional cell area equation(Shao et al., 2021) is a standardized metric for memory cell area relative to the square of the process size, which is crucial for assessing area efficiency across different memory technologies or process nodes.

$$Ar[F^2] = \frac{len_{\mathrm{cell}} \times wdt_{\mathrm{cell}}}{sz^2_{\mathrm{proc}}} \tag{2.8}$$

Here, len_{cell} and wdt_{cell} represent the length and width of the cell, while sz_{proc} denotes the process size, for example, 22 nm.

These equations emphasize the importance of utilizing specific formulations to assess the structural and electrical properties of SOT-MRAM. These formulations play a crucial role in conducting simulations to refine memory design, enabling the exploration of how various physical and electrical factors influence performance and energy efficiency, ultimately aiding in optimizing memory design.

2.3 Exploring the Challenges in Employing SOT-MRAM Memory Hierarchy

Designing efficient cache and main memory using MRAM by balancing capacity, performance, energy, endurance, and lifetime has opened up several challenges, such as

1. **The cell structural influence**: The design of the cell significantly affects system performance by influencing parameters such as area, power consumption, and speed. Creating an optimal cell structure for SOT-MRAM technology involves balancing density, speed, and power efficiency while addressing disturbance and endurance issues. This is made more difficult by the need to manage trade-offs between cell miniaturization and reliable magnetic switching properties. Developing a cell structure that integrates well into existing memory hierarchies while meeting modern application requirements presents a substantial challenge for the widespread adoption of SOT-MRAM technology. So, there is a need to develop a framework for assessing cell structure, array, device type, technology node and density scaling for SOT-MRAM adoption.

2. **Enhancing Capacity Without Expanding Chip Real Estate**: Advanced fabrication techniques are needed to increase bit density within the existing chip layouts while preserving performance standards such as speed and energy efficiency. Achieving this balance requires meticulous memory cell and array design, making the integration of SOT-MRAM a complex task that involves various aspects of semiconductor engineering and design architecture.

3. **Difficulties with Write Operations and Weak Write Endurance**: The endurance of SOT-MRAM is better than STT-MRAM, but SOT-MRAM struggle with weak write endurance compared to SRAM, limiting their ability to handle

repeated write cycles without degradation, posing significant reliability concerns in intensive computing environments.

4. **Overhead in integration and evaluation**: Current tools do not provide an integrated framework to achieve device level, architecture level, policy level and application level rapid idea integration and evaluation.

5. **Design, Integration and Reliable simulation**: Another major challenge in building SOT-MRAM main memory is the public availability of parameters, DDRx protocol compatible SOT-MRAM main memory design, integration and reliable simulation. This is due to the fact that different cells have varying PPA factors. In the subsequent sections, the above issues and the possible solutions proposed in the past are discussed.

Table 2.1,memory technologies are sourced from these publications(Liao et al., 2020; Santhalia and Dahiya, 2021; Kallinatha et al., 2024; Saha et al., 2022; Agarwal and Kapoor, 2020; Wu et al., 2020a; Kryder and Kim, 2009).

Table 2.1: Memory Technology Comparative Analysis

PARAMETER	SRAM	DRAM	PCM	STT-MRAM	SOT-MRAM
Nonvolatility	No	No	Yes	Yes	Yes
Cell-Size(F^2)	50-120	6.0-10	4.0-19	6.0-20.0	6.0 -24.0
ReadTime(ns)	< 10	10-60	48	< 10	< 5
Write/EraseTime(ns)	< 10	10-60	40-150	2-20	< 5
Endurance(Cycles)	$> 10^{16}$	$> 10^{16}$	10^{10}	$> 10^{12}$	$> (10^{12} \rightarrow 10^{15})$
WritePower	low	Low	High	High	Low
Future-Scalability	Good	Limited	Limited	Very Good	Very Good
Leakage Power	Very High	Very High	Low	Very Low	Very Low
Maturity	Product	Product	Advance Development	Advance Development	Early Development
Retention	While voltage is applied	$<< second$	$> 10yr$	$> 10yr$	$> 10yr$

2.4 Related work

Implementation of NVM in cache systems, including STT-MRAM and SOT-MRAM technologies, have been assessed for their impact(Han and Jiang, 2023; Lu et al., 2024; Inci et al., 2022; Han and Jiang, 2024) on performance, power, and reliability. This section examines current literature in three main areas. The first part focuses on the latest developments in NVM (STT/SOT) MFS design and density replacement analysis, highlighting research gaps and evaluation. The second part explores reliability

designs and lifetime extension mechanisms for SOT-MRAM, as well as strategies to improve memory technology. The third part discusses SOT-MRAM as a main memory device alternative to DRAM or STT-MRAM for embedded and multi-core systems. Table 2.1 presents a comparative analysis of the existing memory technologies—SRAM and DRAM—against the emerging NVM technologies.

2.4.1 NVM-based Cache MFS Design for Modern Application

NVMs are under consideration for use in different cache levels as a hybrid or complete substitute for SRAM in on-chip caches due to their superior characteristics compared to CMOS memory. This section provides an overview of the literature on NVM caches. The study by Carnigij et al.(Cargnini et al., 2014) involved simulating devices, architecture, and full systems using SRAM and STT-MRAM at 45nm with low-power devices. The research centred on exploring a multi-level device, array, and architecture for the embedded system L1 and L2 cache hierarchy. It was observed that conventional STT-MRAM required more latency and energy for read/write operations. This work experienced 5-10% CPI penalties in both best and average cases while attempting to achieve the desired performance by utilising 4 times more MRAM capacity to replace SRAM. An area-efficient SOT-MRAM design at 45nm was proposed in(Wu et al., 2020a). Performance was evaluated against SRAM up to 8MB. The proposed SOT-MRAM showed a 38.9% improvement in read speed over SRAM, while the write power doubled when using optimized latency mode values in the experiments.

The study in(Wang et al., 2019) proposed an area-efficient SOT-MRAM with an access circuit at 45nm. Prenat et al.(Prenat et al., 2016) compared STT, SOT-MRAM, and SRAM for latency and area scaling up to 4MB, finding significant reductions in area and power consumption for SOT-MRAM. In (Oboril et al., 2015), a hybrid memory hierarchy was evaluated across different sizes, demonstrating favourable capacity scaling for SOT-MRAM over SRAM with an area reduction of 30% and energy savings of 60%. The work concludes that the SOT-MRAM is a suitable replacement for SRAM based on capacity scaling.

In (Singh et al., 2020), a comparison between SRAM, STT-MRAM and SOT-MRAM cache memory capacity of up to 4MB at a technology node of 22nm, utilizing a page size of 256 bytes was performed. Typically, 64B is in the Intel architecture.

The analysis found that for a capacity of 256KB, SOT-MRAM exhibited an area reduction of 27.74%, 2.97 times faster, and demonstrated significantly lower leakage by around 76.05% compared to SRAM. These results were obtained using latency-optimised configurations.

Zhang et al. (Zitong Zhang and Jiang, 2022) investigated the roadmap of scale effect on STT-MRAM with SRAM up to 32MB without comparing SOT-MRAM. The authors utilized a latency-optimized fixed cell layout at 45nm and observed a 5.4% reduction in write latency and a total of 42.1% leakage power reduction. The work concludes that STT-MRAM outperforms SRAM cache above 32MB, suggesting the need for further exploration of SOT-MRAM alongside SRAM.

In (Liao et al., 2020), simulations at material and circuit levels are performed to enhance the density and reduce energy consumption in different MRAM cells. Four materials are employed for SOT-MRAM, including HMs, alloys, Weyl semi-metals, and topological insulators. The performance of the SOT memory cell/array is evaluated based on read/write time, energy efficiency, and reliability. Alloys with high conductivity and significant spin hall angles show promise as SOT channel materials. Comparative analysis suggests that SOT-MRAM provides faster operation with lower energy usage than STT-MRAM, but it may necessitate approximately 25% larger cell area. This work concludes that the SOT-MRAM can perform better than STT-RAM/SRAM, but no scaling roadmap study was conducted. The SOT-MRAM scaling roadmap study needs to be conducted.

The study(Marinelli et al., 2022), aims to explore different micro-architectural elements to overcome challenges in using STT-MRAM as the LLC in embedded systems. These elements include the number of cache banks, management of Miss Status Handling Registers(MSHR) write buffer entries and integration of hardware prefetchers. Optimizing these parameters could minimize performance degradation while achieving over 60% average energy efficiency improvements for the LLC. Furthermore, comparing energy results from validated technology models with publicly available tools highlights the vital role of accurate models in architectural analysis.

Table 2.2: Summary of scaling techniques in the design of MRAM caches with modern applications.

S	T	Tech	App	Fi	MFS Parameters					
					CAS	DS	TNS	CLT	MAE	TiOv
(Cargnini et al., 2014)	Hardware, Architecture and Full System Simulation	SRAM and STT-MRAM only at 45nm with LOP(Low Power) devices.	Multi-level Device, Array and Architecture for Embedded system L1 and L2 cache hierarchy.	Conventional STT-MRAM consumes more latency and energy for read/write operations. By using 4x more MRAM capacity in place of SRAM, the authors are trying to achieve the required performance, but the penalty is high.	✔	✗	✗	✗	✗	✗
(Oboril et al., 2015; Prenat et al., 2016)	Hardware and Simulation	SRAM, STT and SOT	Multi-level (from Device/Bitcell to system-level)	Analysis of 3 memories for capacity (up-to 4MB) scaling at the circuit and architectural level. SOT-MRAM area reduction was 50%, and power reduction by 60% than SRAM.	✔	✗	✗	✔	✗	✗
(Wang et al., 2019)	Hardware and Simulation	SRAM, SOT-MRAM	Two-level (Device and Architecture)	Area efficient SOT-MRAM cell and cache array capacity scaling performance were analyzed up to 8MB. In this SOT-MRAM, read speed is enhanced by 38.9% to SRAM and write power by two times.	✔	✗	✗	✔	✗	✗
(Wu et al., 2020a)	Hardware and Simulation	SRAM, STT, SOT-MRAM	Device and Architecture for the proposed 3T2SOT area-efficient SOT-MRAM cell.	Hybrid CMOS/MTJ, memory cell designs, include 3: the standard 1T1MTJ, 2T2MTJ STT-MRAM, and traditional 2T1SOT MRAM at 45nm. SOT-MRAM read speed is reduced by 38.9% to SRAM and write power by two times.	✔	✗	✗	✔	✗	✗
(Singh et al., 2020; Saha et al., 2022)	Hardware and Simulation	SRAM, STT and SOT-MRAM	Device, Circuit and Architecture level large LLC	For 256KB, SOT-MRAM is 27.74% area efficient, 2.97 times faster and 76.05% less leakage than SRAM. In (Saha et al., 2022), SOT-MRAM was found to be 4x faster than STT-MRAM. In read/write energy and latency, SOT-MRAM outperforms STT-MRAM.	✔	✗	✗	✗	✗	✗
(Zitong Zhang and Jiang, 2022)	Simulation	SRAM and STT-MRAM only	Device and Bank level at 45nm technology	Proposed a study on the scale effect roadmap of STT-MRAM with SRAM up to 32MB. Achieved a write latency reduction of 5.4% and a total leakage power reduction of up to 42.1%.	✔	✔	✔	✗	✗	✔
(Liao et al., 2020)	Material and Circuit level simulation	Different cells of STT-MRAM and SOT-MRAM for improving density and decreasing energy	Four SOT materials are used. The SOT memory cell/array is optimized and benchmarked for read/write time, energy, and reliability.	SOT-MRAM has the potential for fast operation and reduction in energy usage with around 25% more cell area than STT-MRAM.	✗	✔	✗	✔	✗	✗
(Marinelli et al., 2022)	Full system simulation	Typical and worst-case STT-MRAM with LOP/LSTP devices	Micro-architecture exploration of in-house tuned cell and array models with architecture and system-level analysis	Optimizing the number of cache banks, MSHR, and writing buffer entries. Integration of hardware prefetchers minimizes performance degradation with 60% average energy efficiency improvements for the LLC.	✗	✔	✗	✔	✗	✗
(Han and Jiang, 2023; Lu et al., 2024; Inci et al., 2022; Han and Jiang, 2024)	Simulation	SRAM/STT/SOT	Full system simulation for application power-performance analysis	In all these works, SOT always outperforms the STT cache at various levels except the chip area. Only one set of device values is used.	✔	✗	✗	✗	✔	✗
This work	Simulation	SRAM/SOT-MRAM	None of the aforementioned studies have focused on SOT-MRAM MFS. This work investigates all MFS parameters using novel SOT-MRAM cell designs	The SOT-MRAM density increases the capacity by twice/thrice the SRAM cache and can reduce power consumption by almost half while maintaining similar performance to SRAM.	✔	✔	✔	✔	✔	✔

Notes:- S: Study; T: Type; Tech: Technology; App: Approach; Fi: Finding; CAS: Capacity Scaling; DS: Device Scaling; TNS: Tech Node Scaling; CLT: Circuit-Level Techniques; MAE: Modern Apps Evaluation; TiOv: Time Overhead.

In (Han and Jiang, 2023; Lu et al., 2024; Inci et al., 2022; Han and Jiang, 2024), simulations were performed for SRAM, STT-MRAM, and SOT-MRAM memories simulation for application power-performance analysis. These studies consistently found that SOT outperforms STT caches at various levels except for chip area, using only one set of device values. These studies did not address the scale-effect roadmap for SOT-MRAM cell arrangement. We leverage and extend the idea in(Inci et al., 2022) for the MFS framework.

Past work thoroughly compared the characteristics of SRAM, STT, and SOT-MRAM(Kalllinatha and Talawar, 2023; Kallinatha et al., 2024). Comparisons between STT and SOT-MRAMs have established SOT-MRAM as superior in all aspects except for the memory array/cell area (Singh et al., 2020; Saha et al., 2022; Han and Jiang, 2023). The literature has emphasized the necessity of conducting a scale effect study of SOT-MRAM with structural parameters and modern applications. We noticed that there is a gap in the existing research regarding the integration of SOT-MRAM MFS scale effect and density replacement studies to enhance modern application performance(Liao et al., 2020; Zitong Zhang and Jiang, 2022; Han and Jiang, 2023; Wu et al., 2020a).

Existing works often prioritize improving the functionality of the SOT-MRAM system architecture without considering how the structural arrangement of an individual SOT-MRAM cell affects the overall system. This work proposes a framework and examines the impact of the structural configuration of individual SOT-MRAM cells on the system's architecture, an aspect often overlooked in previous works. The focus is exploring SOT-MRAM as an LLC cache in CPUs, evaluating its performance with different cache capacities, areas, and aspect ratios. This comprehensive evaluation aims to improve power consumption, reduce write latency, and enhance overall cache efficiency. The next sub-section summarizes the literature on NVM cache write variation and lifetime extension.

Contributions

The existing literature emphasizes the substantial impact of memory cell design characteristics on the PPA parameters of LLC. In response, a comprehensive framework has been developed, enabling expedited prototyping and detailed evaluations focused

on scaling roadmaps and density enhancements—crucial for optimizing LLC power performance. The framework, enhanced by the MFS algorithm, has yielded superior outcomes. Additional insights into this technique are elaborated in Chapter 3.

2.4.2 NVM-based Cache Write Variation, Lifetime Improvement(LI) of Physical Split Cache(PSC)

Several works have devised multiple strategies to circumvent the issue of NVM's limited lifetime, including methods to distribute writes more evenly, integrate hybrid caching systems(Agarwal and Kapoor, 2020; Mittal and Vetter, 2014a; Sarkar et al., 2021; Mittal et al., 2014), and apply data compression techniques(Escuin et al., 2022). This literature review will discuss pioneering methods to improve the lifetime of NVM-based Last-Level Caches (LLCs).

J. Wang et al. i^2wap(Wang et al., 2013) introduced an innovative approach utilizing two global counters and registers for enhancing NVM endurance in LLCs. This method includes Swap-Shift for inter-set wear leveling and Probabilistic Set Line Flush (PoLF) to tackle intra-set write variance.PoLF is a cache management technique randomly determining which cache lines to flush to reduce write variation and enhance endurance. It selectively flushes lines based on a set probability, balancing performance and wear levelling by distributing writes more evenly across the cache.

Mittal et al. suggested methods named EqualWrites (Mittal and Vetter, 2015) and EqualChance (Mittal and Vetter, 2014b), targeting intra-set write variance reduction, thereby aiding in the extension of NVM cache life. The EqualWrites concept relies on the idea that significant variance within a cache set indicates a threshold-exceeding difference in writes between two blocks. By swapping data between these blocks, writes are spread more uniformly. EqualChance periodically changes the physical location of write-intensive data, using set-specific write counts to guide write-redirection operations, offering robust wear-levelling, particularly for sets with significant write disparities.Other noteworthy methods include Static Window Write Restriction (SWWR) (Agarwal and Kapoor, 2017) and its dynamic counterparts DWWR and Dynamic Way Aware Write Restriction (DWAWR) (Agarwal and Kapoor, 2019).

Table 2.3: Summary on hybrid/split cache with Lifetime Improvement(LI) techniques.

Study	Type	Mem.Tech	Approach	Findings	CPR	VRO	Overhead	RD
(Wang et al., 2013)	i^2WAP	STT-MRAM	PoLF and Swap-shift	Assumes the hot block is the reason for wear. Uses two counters	✗	✗	Area, power	✗
(Mittal and Vetter, 2015)	Intra-Set	NVM	EqualWrite	Uses SRAM per-set counters which grow with cache size	✗	✗	Area, power	✗
(Mittal and Vetter, 2014b)	Intra-Set	NVM	EqualChance	Physical cache-block location of a write-intensive data item is modified periodically	✗	✗	Area, power	✗
(Agarwal and Kapoor, 2017)	Intraset	Hybrid	SWWR	Uses round-robin fashion static window	✗	✗	Area, power	✗
(Agarwal and Kapoor, 2019)	IntraSet	Hybrid	DWWR, DWAWR	Dynamic write restricted counter and no inter-set WAR	✗	✗	Some part of cache unavailable for write operations, degrading performance	✗
(Sivakumar et al., 2023)	IntraSet	Any NVM	WALL-NVC and LRU-Coldblock	Uses replacement policy with write distribution	✔	✗	✗	✗
(Mittal et al., 2014)	IntraSet	STT-MRAM	Write Smoothing	Cache unavailable for write operations, degrading performance	✗	✗	✗	✗
(Sivakumar et al., 2023)	Both	NVM	Partitions Cache into I/D cache	Uses fewer counters and does not restrict any portion of cache memory	✗	✔	✔	✗
(Agarwal and Kapoor, 2020)	Both	NVM/Hybrid	Various policies	Performance degradation and restricted	✗	✔	✔	✗
(Escuin et al., 2022, 2023)	Both	Hybrid	Frame/Byte fault handling policies	Fault-tolerant and compression based insertion policy for endurance extension.	✗	✗	✗	✔
This Work	Both	Hybrid/SOT-MRAM	PSC with VRO with FT	Easy DSE, efficient cache with lifetime extension	✔	✔	✔	✔

Notes:CPR:Cache-Part Restriction;VRO:Virtual ReOrganization;RD:Reliable Design.

These techniques segment the cache into windows, selectively restricting writes to specific windows to distribute the write load. SWWR operates on a fixed schedule, cycling through the cache windows, while DWWR and DWAWR use counters to track write frequencies, dynamically adjusting write restrictions for optimal wear distribution. Traditional cache replacement policies often exacerbate wear, with frequently accessed blocks wearing out faster. The WALL-NVC (Sivakumar and Jose, 2023) technique employs an NVM-friendly policy, the Least Recently Used Cold Block (LRU-CB) (Sivakumar and Jose, 2023), which considers both write frequency and recent usage in block eviction, ensuring a more balanced write distribution.

While these wear-levelling and write distribution methods show promise, they also introduce new challenges. The reliance on SRAM-based counters for methods like EqualWrites, EqualChance, and PoLF increases area and power overhead, a concern amplified as cache sizes grow. Similarly, write restriction strategies can compromise performance by making cache segments intermittently unavailable for writes. The proposed work seeks to overcome these limitations by employing minimal counters and ensuring that no portion of the cache is off-limits, potentially leading to performance improvements without significant power or area penalties. This work develops a cache management policy to leverage the MFS framework for comprehensive end-to-end evaluation. The cache management policy optimizes the proposed algorithm to address write variation in the NVM cache segment.

2.4.3 Contributions

The existing literature suggests that the endurance and WVAR have a significant impact on the reliability, LI, and performance of hybrid caches. As a result, a cache management policy has been devised to take into account the importance of WVAR and LI in the design of the last-level cache (LLC). This approach involves writing to the SRAM segment, reducing and distributing writes evenly across the SOT-MRAM segment, and migrating loop blocks. These actions are carried out through a replacement policy and the use of an additional bit to identify the loop blocks. For additional details and outcomes of this approach, please see Chapter 3.

2.4.4 SOT-MRAM Main Memory for Embedded and Multi-core Systems

Though SOT-MRAM is investigated as a caching technology, it has not been explored yet as a main memory device alternative to DRAM or STT-MRAM for multi-core systems. This section reviews the integration of NVM as primary memory in horizontal structure, vertical structure, and hybrid memory systems for various workloads. Although the feasibility of using SOT-MRAM technology as the primary system memory in embedded, HPC, or general-purpose computers is not yet thoroughly studied, some researchers have investigated the potential of using STT-MRAM or PCM as a hybrid or full main memory in various computing systems.

We first list (Table-2.4) all the works using the commercially available hardware in the experiments. In(Peng et al., 2020), Intel's first NVM-based Optane DC persistent memory module was tested on a 24-core processor for high-performance computing (HPC) applications, known as the "seven dwarfs". The results show that DRAM-cached NVM enhances HPC application performance and allows for handling more significant problems than DRAM alone. HPC applications' performance on uncached-NVM was categorized into three levels: insensitive, scaled, and bottlenecked. They introduced two optimization methods: a predictive model and a write-aware data placement, which doubled performance and cut DRAM use by 60%. However, these findings on HPC applications may differ for other applications or workloads using NVM-based memory.

Liu et al. (Liu et al., 2022) have devised a hybrid DRAM/STT-MRAM memory for IoT systems to enhance STT-MRAM's write speed with a fast data migration method. They built this system using Micron DDR3 SDRAM (Inc., 2006) and Everspin DDR3 STT-MRAM (Inc., Inc.), focusing on reducing power use on standby mode. Their findings have shown only a minor drop in STT-MRAM performance. Importantly, they have shared specific timing and power details, offering valuable insights for further research.

Table 2.4: Summary of studies on NVM main memory systems

Study	Type	Technology	Approach	Findings	Parameter Disclosures	Micro-arch. Parameters	Various Mem. Org. and Capacities
(Peng et al., 2020)	Hardware	NVDIMM (DRAM-cached-NVM)	NVDIMM on 24-core processor using DRAM-cached-NVM	Enhanced HPC application performance twofold while reducing DRAM usage by 60%	✗	✗	✗
(Liu et al., 2022)	Hardware	Hybrid DRAM/STT-MRAM	Hybrid DRAM/STT-MRAM for IoT systems	Minimal impact on STT-MRAM performance with fast data migration	Limited disclosure	✗	✗
(Li et al., 2023)	Hardware	STT-MRAM	STT-MRAM and DRAM in IoT systems	15% power reduction, 714x faster data restoration time	Limited disclosure	✗	✗
(Fu et al., 2022)	Simulator	DRAM-PCM (CAHRAM)	Content Aware Hybrid DRAM-PCM system	Outperformed existing solutions in I/O and PCM life	✗	✗	✗
(Jing and Li, 2022)	Simulator	STT-MRAM	STT-MRAM architectures	Minimal impact, benchmark evaluation needed	✗	✗	✗
(Mahdavi et al., 2022)	Simulator	STT-MRAM	Techniques for STT-MRAM issues	Reduced read disturbance, write failure, performance	✗	✗	✗
(Asifuzzaman, 2019)	Simulator	STT-MRAM	STT-MRAM impact on HPC and real-time systems	2-10% performance degradation in HPC, potential for real-time systems	✔	✗	✗
(Ma et al., 2021)	Simulator	STT-MRAM	STT-MRAM reliability and performance	Up to 5.996% latency, 20.65% energy increase	✗	✗	✗
(Oh et al., 2023)	Simulator	Hybrid DRAM/STT-MRAM	SMART architecture	Lower energy, higher performance than conventional STT-MRAM and DRAM	✔	✗	✗
This Work	Full System Simulator	SOT-MRAM	Stand-alone SOT-MRAM main memory and Hybrid memory with real-time workloads	Substantial Power and Performance gains at both circuit and system level	✔	✔	✔

Work in(Li et al., 2023) examines the effects of commercially available STT-MRAM and DRAM on energy harvesting in IoT systems. The study evaluates different scenarios and finds that using STT-MRAM results in a 15% decrease in power consumption and a 714x faster data restoration time. However, the study is limited by a lack of published timing and power parameters for 32MB STT-MRAM, which may affect its accuracy in representing applications with higher memory requirements.

The rows (Table-2.4) with type simulator are open-source simulators with limited publicly available configuration, timing, and power parameters. In (Fu et al., 2022), a new system called CAHRAM is introduced to improve PCM in a hybrid setup. CAHRAM uses deduplication to reduce access overheads. Results using Gem5 (Binkert et al., 2011) and NVMain (Poremba and Xie, 2012) show the system outperforms existing methods in speed, PCM lifetime, and efficiency. This work's effectiveness may be limited for workloads with low-redundancy data. The study lacks detailed timing and current parameters. Jing et al. (Jing and Li, 2022) analyze advancements in heterogeneous memory with STT-MRAM, introducing three optimized schemes: hierarchical, parallel, and hybrid architectures. However, the study lacks full system simulation analysis with benchmark programs, highlighting the need for further research to test these schemes under diverse workloads.

In (Mahdavi et al., 2022), Mahdavi et al. review STT-MRAM issues like read disturbances and write failures. They propose reducing write operation errors and optimising encoding to decrease read disturbances. Their simulations on a quad-core processor using Gem5 (Binkert et al., 2011)-NVMain (Poremba and Xie, 2012) show a 92% reduction in read disturbances and a 22% decrease in write failures. Despite minimal increases in area, power, and performance(0.1%, 1.52%, and 1.19%) overheads, the study lacks detailed timing and current parameters.

Focusing on NVM-based main memory systems, we review research with publicly available timing and current parameters for open-source simulator verification.

In the (Asifuzzaman, 2019) thesis, the author evaluates STT-MRAM in HPC systems with trace-based simulations, assuming STT-MRAM is 50% and 100% slower than DRAM. Performance degradation ranges from 2 to 10%. Another study (Asifuzzaman et al., 2022) assesses STT-MRAM as primary memory in high-performance computing compared to DRAM. Using SPEC2006 benchmarks on ZSim and DRAM-

Sim2 with timing parameter scaling of 1.2x, 1.5x, and 2.0x, they find an average performance drop of 5.4% and 11.3% in integer and floating-point benchmarks, respectively, when timing parameters are twice that of DRAM. Lastly, considering STT-MRAM's benefits like radiation resistance and zero leakage power, its suitability for real-time systems, including space, automotive, and avionics applications, is evaluated (Asifuzzaman et al., 2019). Using benchmarks tailored for these sectors, the study suggests STT-MRAM could be a strong candidate for this sector.

Haoyuan Ma et al.(Ma et al., 2021) introduced a framework to assess STT-MRAM's reliability and performance at the computer architecture level using GEM5+NVMain. Their results show that STT-MRAM's average latency and energy use could increase by up to 5.996% and 20.65% compared to standard models. These results indicate that reliability and performance issues at the system level might limit STT-MRAM's use in high-performance and real-time systems.

Another issue with STT-MRAM is its modified sensing methods. The SMART system in (Oh et al., 2023), uses a modified approach for STT-MRAM to address the sensing challenges more efficiently than DRAM. SMART's sense amplifier design reduces row buffer size, leading to higher activation energy and lower performance. However, it improves by delaying bit-line sensing until a column access command is received. This approach offers numerous benefits, including larger page sizes, less reliance on sense amplifiers, reduced activation power, better parallelism, lower latency, and more effective repair of faulty columns. SMART surpasses traditional STT-MRAM and DRAM in energy efficiency and performance while being more compact. This study provides publicly available timing and current parameters.

Research on STT-MRAM has been extensive, exploring its use in various settings, from IoT to real-time systems, using real hardware and simulations. Despite some performance drawbacks, studies demonstrate its potential to replace DRAM in many computing systems. However, high write energy, slow write speeds, and reliability issues still hinder STT-MRAM's adoption as the main memory. The following section will discuss the potential of SOT-MRAM as an alternative. Table 2.4 summarizes this research, noting gaps and challenges for future studies. "✗" marks in the table indicate unresolved issues, while "✔" marks show where studies have provided solutions or key parameters. The final row outlines the contribution of our study in this work.

2.4.5 NVM Based Embedded System Main memory

Several alternatives have been proposed to replace the existing CMOS-based DRAM memories. The most favourable choice is to use NVM devices as an alternative to DRAM. STT-MRAM, one of the NVM devices, has significant advantages regarding high data density and low energy consumption. Still, it suffers from a read-write disturbance problem because of the shared path for read and write operations. Apart from this, STT-MRAM also suffers from large current consumption and reliability issues (Alhalabi et al., 2018). Many researchers have analyzed the impact of using STT-MRAM-based main memory as an alternative to DRAM. In (Kültürsay et al., 2013), an experiment was carried out to find the suitability of STT-MRAM as a memory device. They have suggested that STT-MRAM can be used as an alternative to DRAM with modifications to save energy and time. After adapting techniques like partial write and write bypass, they were able to achieve an average power reduction of 60% when compared to DRAM. In (Kim et al., 2020), STT-MRAM is used as a part of hybrid memory and the DRAM by making energy-efficient data placement methods to optimize energy consumption. Chen et al. proposed a new swapping mechanism for hybrid memory systems comprising NVM devices such as STT-MRAM and PCM. This work focused on reducing swap operations for embedded systems that combine DRAM and NVM as hybrid memory (Chen et al., 2017). However, they have used STT-MRAM as a main memory component. In (Mahdavi et al., 2022), STT-MRAM is studied as the main memory. Methods have been proposed to solve the read-write disturbance problems it faces before adopting it as the main memory. This work proves that STT-MRAM-based devices can be used as main memory by handling issues they face. A detailed analysis of STT-MRAM was performed in (Asifuzzaman et al., 2022), which details the timing and energy parameters to use STT-MRAM as the main memory; this is used as a base to derive the timing and energy parameters for SOT-MRAM in our work. All these works show strong evidence that STT-MRAM is used as the main memory. Nevertheless, STT-MRAM faces issues concerning read-write disturbance, reliability, and high write latency.

As mentioned, SOT-MRAM resolves the read-write disturbance issue of STT-MRAM by using separate read-write paths. There is evidence to show that SOT-MRAM can be an alternative to SRAM. A study in(Garello et al., 2018) demonstrated

using SOT-MRAM in the CMOS-compatible integration process to handle cache replacement. The analysis in(Prenat et al., 2016) to study the impact of SOT-MRAM-based devices has suggested that SOT-MRAM can perform better than SRAM and be considered a cache candidate due to their low power consumption and good performance. SOT-MRAM is often considered a successor of STT-MRAM, but the drawback is that it takes up a lot of cell space. In (Seo and Kwon, 2020b), an area-optimized SOT-MRAM is proposed. In this, SOT-MRAM achieved an area efficiency of 42% compared to conventional SOT-MRAM. Recent work in (Lu et al., 2024) proposes a new three-terminal-based SOT-RAM which can perform better than conventional two terminals, one MTJ-based SOT-MRAM. They have developed methods that can overcome the drawbacks of conventional technologies. All these past works give promising results for STT-MRAM-based main memory and SOT-MRAM-based cache. Since SOT-MRAM and STT-MRAM have common features and only STT-MRAM is used as the main memory, we have designed DDRx-compatible and integrated the SOT-MRAM as the main memory.

2.4.6 Contributions

The literature suggests a comprehensive evaluation of SOT-MRAM as the main memory for embedded and multi-core systems using a full system simulator and benchmark programs like PARSEC(Bienia et al., 2008a). Designing and integrating SOT-MRAM into embedded and multi-core systems is crucial, extending its potential applications across three memory structures. The challenge of publicly available parameters and reliable simulations is addressed by micro-architectural and system-level evaluation of the three memory structures. More information about this work can be found in chapter 5.

Chapter 3

MFS Design of SOT-MRAM On-chip Caches for Modern Applications

This chapter will address two primary topics:(a) the scaling roadmap for SOT-MRAM caches using the framework and (b) density replacement to improve LLC capacity and performance within the same chip real estate for modern applications. The chapter proposes an MFS framework for exploring the SOT-MRAM on-chip cache design space. MFS framework aims to analyze the design of various solutions for specific cache sizes and assess their performance across applications, including AI, NLP, Social Network Analytics (SNA), and generic computing applications. This work also includes evaluating potential power performance improvements by replacing equivalent capacity SRAM with SOT-MRAM and integrating larger SOT-MRAM caches within the same chip area occupied by SRAM. The MFS framework facilitates evaluating and comparing different SOT-MRAM cell configurations and optimization targets to identify the most suitable solution for a targeted configuration design. The proposed approach leverages a multi-factor scaling strategy that combines various techniques to enhance the performance of SOT-MRAM caches:

- Device scaling: Scaling down the SOT-MRAM cell dimensions to improve write energy efficiency and endurance (Gupta et al., 2020).

- Technology scaling: Utilizing advanced technology nodes to reduce the write

current and energy consumption further(Van Beek et al., 2023).

- Circuit-level techniques: Employing efficient write circuits and schemes to reduce the overall write energy and latency, thereby improving endurance (Seo and Kwon, 2020a; Shao et al., 2021).
- Architecture-level optimizations: Utilizing cache management policies and hybrid cache designs that reduce the number of writes to SOT-MRAM cells (Escuin et al., 2023; Agarwal and Kapoor, 2020; Sivakumar and Jose, 2023).

3.1 MFS for On-chip Cache Design

One of the challenges in designing on-chip caches with SOT-MRAM is the scaling factors of bit cells (cell area, aspect ratio, and I/V parameters), which are required to achieve the desired cache capacity with an efficient technology node. This scaling approach enables the design of smaller, faster caches and consumes less power than traditional SRAM-based caches. The cell area and aspect ratio scaling can increase the number of memory cells in a given chip area, resulting in a higher cache density. The advanced technology nodes with smaller feature sizes enable a higher density of memory cells, further increasing the cache capacity.

The MFS approach is an alternative to SOT-MRAM on-chip cache design. MFS scale various parameters to enable the design of smaller, faster, and more energy-efficient caches. The MFS approach optimises the cache design for various workloads and applications. This work investigates the impact of MFS on SOT-MRAM on-chip cache design. It evaluates the performance of caches under different scaling parameters such as cell area, aspect ratio, and technology node. This approach also provides an assessment of density replacement for modern applications.

The following steps are followed in the MFS roadmap:

- Identify the key factors affecting the performance of the cache.
- The range of values for each factor used in the scaling process is established.
- The simulation evaluates the performance of the cache under each combination of values. The developed algorithm is used with DESTINY(Mittal et al., 2017) simulator in this step.

- Analyze the results to identify the optimal combination of factor values to provide the best overall performance for the cache. Identify the SOT-MRAM cells and targeted configurations which are providing balanced results.

- The identified cells and configurations for each capacity are used in density replacement evaluation studies for modern applications. For example, iso-capacity uses 1MB:1MB of SRAM and SOT-MRAM, and iso-area uses 4MB:8MB and 4MB:16MB of SRAM and SOT-MRAM for evaluation. The NVMExplorer(Pentecost et al., 2022) simulator is used for this step.

The simulations use workloads from AI(ResNet, Facebook, Wikipedia), NLP(ALBERT), SPEC2006, SPEC2017(Leskovec and Krevl, 2014; SPEC, SPEC) and corresponding input data to model the cache's behaviour for modern applications accurately. In density analysis, the SRAM is replaced with a cache of equal size and 2x, 3x, and 4x the size using SOT-MRAM.

Fig.3.1(a) presents the MFS-DSE framework flowchart for scaling and optimization of cache memory design using SOT-MRAM technology. SOT-MRAM cells' structural and electrical characteristics feed the targeted cache configuration. Array layout and operational aspects are considered to explore scaling roadmap parameters. These inputs are then fed into a microarchitectural MFS Algorithm 1, which explores the design space of cache memory using the DESTINY tool. The algorithm identifies balanced and optimal cache configurations for each capacity and corresponding SOT-MRAM cell, such as 12/1 or 24/1. The results are subsequently analyzed.

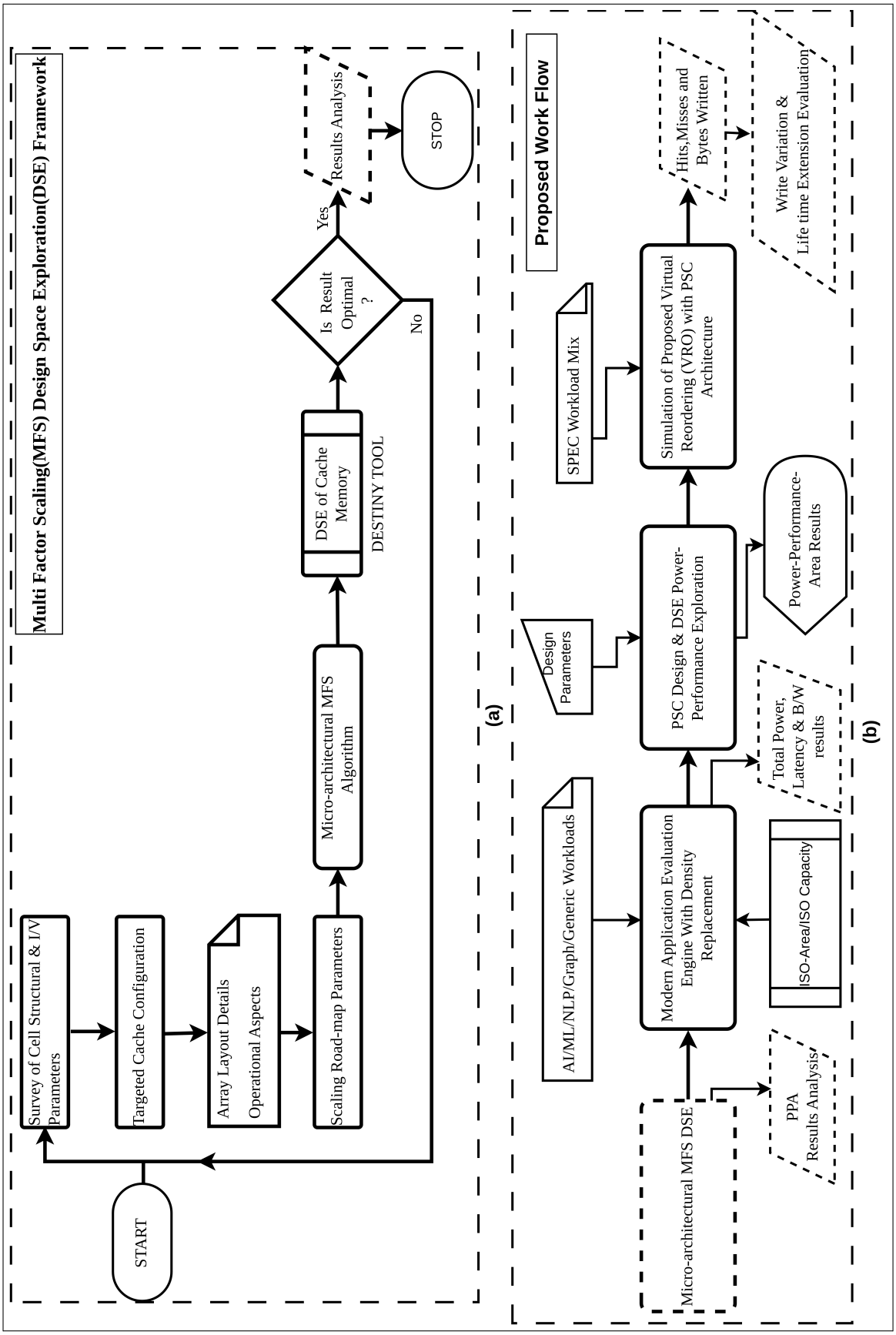


Figure 3.1: The end-to-end MFS framework overview and integration of Lifetime Improvement (LI) in the design process.

Fig.3.1(b) shows the end-to-end workflow of the proposed framework. MFS-DSE gives optimal results when choosing the balanced combination parameters for the next step of modern application density replacement. In the iso-area/capacity, the work evaluates the performance of a mix of workloads utilizing the MFS-DSE framework results. The next step is a hybrid PSC design for reliability and LI extension. The same MFS-DSE is used to obtain PSC power and performance results, and the optimal results are presented. Finally, PSC is evaluated with LI extension algorithm dynamic VRO for WVAR distribution and minimization. Design parameters and the SPEC workload mix are used in the LLC design and power-performance exploration. The work then simulates the proposed dynamic virtual reordering(VRO) with a PSC cache. In this work, VRO or dynamic reordering(DRO) are used interchangeably. The unified LLC is split into SRAM and SOT-MRAM by PSC architecture. The SOT-MRAM part is logically split into read-and-write ways by *VRO or DRO*, which will only store the corresponding blocks. Based on the number of writes and predefined reordering thresholds, PSC modifies the logical mapping, causing the ways exclusively reserved for heavily written data to function as read ways and vice versa. This results in a uniform distribution of heavily written ways across the cache memory. PSC-VRO evaluations include the results of power, performance, and area from DESTINY and write variation and lifetime extension from HyCSim(Escuin et al., 2022). This comprehensive approach ensures thorough analysis and optimization of SOT-MRAM performance in the cache. Further details of PSC-VRO with LI is discussed in the next Chapter 4.

The Algorithm 1 systematically explores the design space of SOT-MRAM cache configurations to optimise various performance metrics. It commences by initialising parameters such as memory types (SRAM, SOT-MRAM), size range (16KB to 32MB), cell parameters (area, aspect ratio, current, voltage, CMOS access length), and target optimisation parameters (read/write latency, read/write energy, read/write EDP, area, leakage)(lines1 to 7). From line 8, the algorithm iterates through technology nodes (22nm to 65nm), memory types, sizes, and cell parameters to find the optimal configuration that meets the desired performance criteria(lines 8 to 26). It evaluates each combination using the DESTINY tool within nested loops(lines 15 to 17), which provides power, performance, and area (PPA) estimates. The algorithm dy-

Algorithm 1: MFS SOT-MRAM Cache Design Tuning

Input: Memory *M*, Size *S*, Target Optimization *TO*, Cells Parameters *CP*
Output: Meeting all the Optimal results of *TO*

```
1  Memo ← {SRAM, SOT − MRAM};
2  Size ← {16KB, 32KB, ..., 512KB, 1MB, 2MB, ..., 32MB};
3  Cells_Para ← CeP ;          /* Area, Aspect ratio, Current, Voltage
     parameters, CMOS access length */
4  Opt_Paras ← OPt ;  /* ReadLatency, WriteLatency, ReadEnergy, Write
     Energy, Read EDP, Write EDP, Area, Leakage */
5  Accs ← {Normal, Fast, Sequential};
6  Dev ← {HP, LOP};
7  Tech_node ← {22nm, ..., 65nm};
8  while Tech_node ∈ Tn do
9      while Memo ∈ Mm do
10         for Sizes ∈ Si do
11             for Cell_Para ∈ CeP do
12                 CirLev ← ∞;
13                 for Opt_Paras ∈ OPt do
14                     for Accs ∈ Ac do
15                         for Devi ∈ De do
16                             CirLevP ← DESTINY(PPA);
17                             Process the results;
18                             if CirLevP+ ≥ CirLevP then
19                                 CirLevP+ ← CirLevP;
20                             end
21                         end
22                     end
23                 end
24                 MFSResult.append(argv(CirLevP));
25             end
26         end
27     end
28     return CirLevP
29 end
```

namically updates the best configuration based on these estimates, ensuring that the optimal parameters are identified and recorded(line 21). This comprehensive exploration facilitates a detailed analysis of SOT-MRAM cache designs, balancing power and performance considerations for advanced memory technologies. SOT-MRAM device parameters such as access transistor width, set/reset current, read/write time, cell area, and aspect ratio from relevant studies were derived (Liao et al., 2020; Wang et al., 2019; Seo and Kwon, 2020a; Van Beek et al., 2023; Lu et al., 2024; Wu et al., 2020a). This work integrated models from reliable and concise SOT-MRAM design

cell-level parameters. Also, it utilised information from sources detailing highly-dense, area-optimal, performance-improved SOT-MRAM devices and other related key parameters for scaling road-map design from(Liao et al., 2020; Wang et al., 2019; Seo and Kwon, 2020a; Van Beek et al., 2023; Lu et al., 2024; Wu et al., 2020a). This work uses the DESTINY(Mittal et al., 2017) simulator in the MFS approach to improve the micro-architectural performance of caches at various capacities. Depending on the optimisation target chosen, the cache memory power, performance, and area results are extracted.

3.2 Scaling Framework

The scaling methodology begins with initialising the Algorithm 2 to optimize cache memory configurations for modern computing environments. This algorithm selects the optimal cache configuration and associated cell parameters, tailoring the memory system to specific performance and efficiency goals. Following this, an evaluation engine NVMExplorer(Pentecost et al., 2022) is integrated, providing a robust platform for testing and validating the chosen configurations against a diverse set of modern application workloads. This step ensures that the cache designs are versatile and effective across various real-world scenarios. Subsequently, the methodology employs an *iso-area/capacity* evaluation to assess how different cache configurations perform under fixed area constraints, focusing on maximizing storage capacity without increasing the chip area by density replacement. This evaluation is crucial for determining the practical applicability of the configurations in chip real estate-constrained systems.

Finally, the methodology refined results based on a detailed analysis of latency, power consumption, and the Energy-Delay Product (EDP) of the applications are obtained. This cyclical approach allows for continuous improvement of the cache configurations, ensuring they meet the evolving demands of modern applications with limited hardware real estate.

3.2.1 Scaling DSE Algorithm

The Algorithm 2 commences by defining input and output parameters, focusing on structural, operational, and targeted cache configurations (Lines 1-2). The initial

Algorithm 2: Scaling Road-map DSE Framework

```
   Input: Parameters including cell structural, operational aspects, and
          targeted cache configurations
   Output: Optimized cache configuration and performance results
   /* Initial Setup                                                    */
1  Survey of Cell structural and Key parameters
2  Determine targeted cache configuration details
3  Note array layout details and operational aspects
   /* Scaling Road-map Definition                                      */
4  Define scaling road-map parameters based on initial data
   /* Algorithm Application                                            */
5  Apply Micro-architectural MFS Tuning Algorithm to develop a scaled model
   /* Iterative Design Space Exploration                               */
6  while not optimal and iterations <= max_iterations do
       /* Design Space Exploration with DESTINY Tool                   */
7      DSE of cache memory using DESTINY tool
       /* Decision Block                                               */
8      if Result is optimal then
9          Proceed to results analysis
10         return Optimized cache configuration and detailed results
11     end
12     else
13         Adjust parameters or model based on feedback
14         Reevaluate structural parameters and update design criteria
15         Refine operational aspects based on performance metrics
           /* Reapplication of MFS Algorithm                           */
16         Apply Micro-architectural MFS Algorithm with updated parameters
           /* Increment iteration counter                              */
17         iterations ← iterations + 1
18     end
19 endw
   /* Final check if the optimal solution was not found                */
20 if not optimal then
       /* Further Optimizations                                        */
21     Suggest further Optimizations
22 end
```

setup involves surveying cell structural and current-voltage parameters, determining cache configuration details, and noting array layout and operational aspects (Lines 3-5). The scaling roadmap parameters are established based on initial data (Line 6).

Applying the micro-architectural scaling mechanism develops a scaled model through iterative DSE. This process continues until optimal results or maximum iterations are reached(Lines 9-15). Utilizing the DESTINY(Mittal et al., 2017) tool, the MFS tuning algorithm performs a detailed DSE of cache memory. If the results are optimal, the

MFS algorithm concludes by returning the optimized cache configuration; otherwise, it adjusts parameters and iteratively reapplies the scaling algorithm with updated data until optimal configurations are realized (Lines 16-19).

3.3 Experimental Setup

The simulation tools, setup, structural device parameters, constraints, and workloads involved in the MFS exploration, density replacement study and PSC-VRO are described here. Tables 3.1,3.2, and 3.3 list all the detailed parameters used in the study. The memory array parameters are employed in density replacement to estimate the overall effect of the SOT-MRAM LLC by considering the application workload. The framework integrates DESTINY(Mittal et al., 2017) and NVMExplorer(Pentecost et al., 2022) for the study.

3.3.1 Simulation Setup for MFS and Density Replacement Evaluation

Data from compact SOT-MRAM models (Mittal et al., 2017; Liao et al., 2020; Sura and Nehra, 2021; Wang et al., 2019; Wu et al., 2020a; Mondal et al., 2023; Jang and Park, 2022) and the repository in (Mittal et al., 2017) were used to establish the bit-cell device parameters. Key parameters, including cell area, aspect ratio, set/reset current, read/write time, and access transistor width, were obtained from these sources(Liao et al., 2020; Sura and Nehra, 2021; Wang et al., 2019; Wu et al., 2020a; Mondal et al., 2023; Jang and Park, 2022; Saha et al., 2022; Singh et al., 2020; Han and Jiang, 2023, 2024; Lu et al., 2024; Kalllinatha and Talawar, 2023). Algorithm-1 is designed to integrate the modified DESTINY(Mittal et al., 2017) simulator core. This was used for MFS scale effect experiments.

NVMExplorer(Pentecost et al., 2022) is a cross-stack framework for comprehensive DSE and application evaluation. It evaluates various on-chip memory options by considering system limitations and their effects at the application level. The framework uses optimal cell configurations, array parameters, and a wide range of modern workloads detailed in Table 3.2 to analyse both *iso-area* and *iso-capacity* configurations for LLC replacements. The included application workloads cover diverse domains such as

Feature	**SRAM**	**SOT-MRAM**
Cell area	146 F^2	12 to 24 F^2
Aspect Ratio	1.46 F	0.5 to 2.0 F
Access CMOS width	1.31 F	2 to 6F
Resistance Parallel and Anti-parallel	–	3K to 30K
Write pulse	–	$< 0.5ns$
Capacity Range	16 KB to 32MB	
Device type	HP/LOP	HP/LOP
Technology node	—	22nm-65nm
Simulators	Modified DESTINY and NVMExplorer with Proposed MFS method	

Table 3.1: Simulators and Parameters used for MFS roadmap study(Mittal et al., 2017; Liao et al., 2020; Sura and Nehra, 2021; Wang et al., 2019; Wu et al., 2020a; Mondal et al., 2023; Jang and Park, 2022)

graph processing, natural language processing, image processing, and general computing, with multiple instances for each, ensuring a thorough and relevant performance assessment across different scenarios.

Table 3.2: Workload details of modern applications

No. of Instance	**Workload Name**	**Application Domain**
3	Facebook and Wikipedia	Graph processing /Social network
3	ALBERT	Language Model(NLP)
3	ResNet	Image Processing DNN
14	SPEC2017	Generic Computing

Application workloads are outlined in Table 3.2(Leskovec and Krevl, 2014; Lan et al., 2019; SPEC, SPEC; Pentecost et al., 2022; Inci et al., 2022). These workloads cover a range of application domains, including graph processing and social networks (Facebook and Wikipedia), natural language processing (ALBERT), image processing (ResNet), and general computing (SPEC2017). Each workload has multiple instances: three for Facebook/Wikipedia, ALBERT, and ResNet, and fourteen for SPEC2017. This variety ensures a comprehensive evaluation across different application types, yielding realistic and pertinent performance measurements.

3.4 Results and Discussion

The results and findings of the work are presented in two sections. First, the work examines the scaling road map for SOT-MRAM structural influences on cache memory.

Table 3.3: Bit Cell Device Parameters(Liao et al., 2020; Sura and Nehra, 2021; Wang et al., 2019; Wu et al., 2020a; Mondal et al., 2023; Jang and Park, 2022).

Parameter Description	**SOT1**	**SOT2**	**SOT3**	**SOT4**
MTJ area (nm^2)	30×30	40×40	$\Pi \times 20 \times 20$	40×35
Heavy-Metal dimension (nm^3)	$50 \times 40 \times 3$	$40 \times 60 \times 2$	$100 \times 50 \times 3$	$60 \times 35 \times 2.5$
Free layer thickness (nm)	0.8	1	2	0.6
Oxide layer height (nm)	1.1	0.85	1.1	1
Spin Hall angle (θ)	0.3	0.3	0.3	0.4
Magnetic anisotropy (A/m)	1.2×10^5	1.33×10^5	8×10^4	8.9×10^6
Saturation Magnetization (A/m)	1×10^5	1×10^6	9×10^5	1.080×10^6
Tunnel Magnetoresistance (TMR)	130%	120%	135%	150%
Heavy-Metal resistivity ($\mu\Omega \cdot$ cm)	190	200	200	180
Damping constant(")	0.5	0.02	0.015	0.01

The second part focuses on the density effect on modern applications. The results of the experiments carried out based on the proposed work in Fig.3.1 with different parameter combinations like cache capacity, cell area, aspect ratio, High Performance (HP), Low Power Performance (LOP) models, and 45nm process nodes unless specified otherwise are as follows.

3.4.1 SOT-MRAM MFS Scale Effect on Device type

The device type scaling, specifically for HP and LOP devices, plays a significant role in the design of cache memories. In this study, we evaluated the impact of device-type scaling on the performance and power consumption of SOT-MRAM-based cache memories. The cache memories were designed using HP and LOP devices at various technology nodes with different cell areas and aspect ratios.

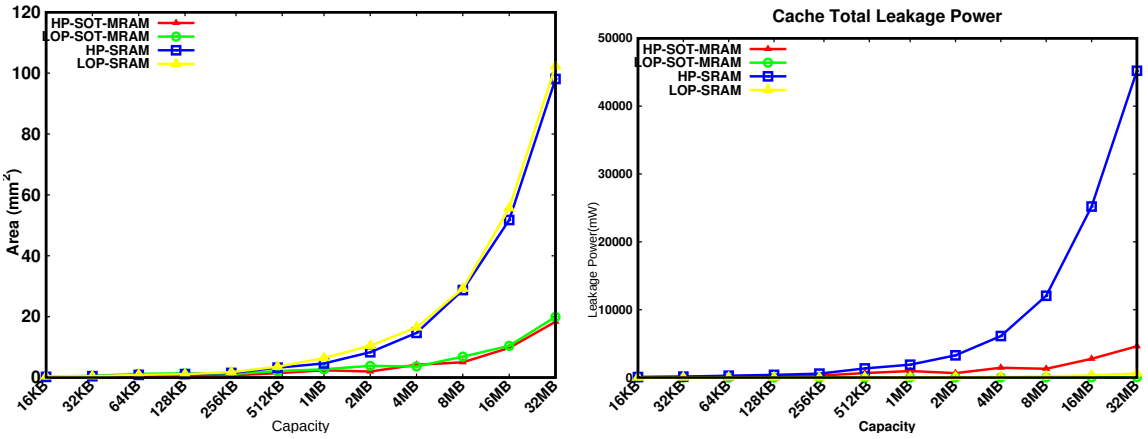
LOP and HP Devices for an SRAM Memory Bank

Fig. 3.2 shows the analysis of parameters like cache area, leakage power, write latency, hit latency, write energy, and hit energy with the characteristics of SRAM caches across various capacities. As depicted in Fig.3.2a, the area of the SRAM caches using LOP and HP devices are similar for memory capacities ranging from 0 to 32 MB. However, at the 32 MB capacity, there is a difference, with the LOP device occupying 4.3% more area than the HP device SRAM cache. The results presented in Fig.3.2b show that the LOP device type exhibits significantly lower leakage power than the HP device type,

with the LOP device demonstrating a substantial reduction, on an average 71× times less leakage power than the HP device. At 16KB, LOP devices have 55× less leakage power than HP devices. This gap widens to 75× at 32MB capacity. This considerable difference in leakage power is due to the design optimization of LOP devices, which prioritize energy efficiency over performance. In contrast, HP devices are engineered for high-speed operations, producing higher leakage power. The increased area of the LOP device at 32 MB can be attributed to the additional circuitry required to achieve lower power operation. These findings highlight the trade-offs between energy efficiency and performance in SRAM cache designs.

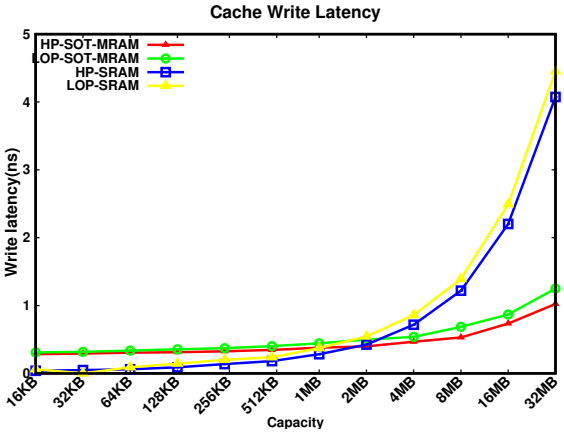
Fig.3.2c illustrates that the HP SRAM device's average write latency is 24% lower than that of the LOP SRAM device. Specifically, at a cache capacity of 16 KB to 256KB, the difference in latency between the two types is insignificant. Nevertheless, as the cache capacity increases, the disparity in write latency between the two types diminishes to less than 10%. When the cache capacity reaches 32 MB, the LOP SRAM device exhibits 9.3% more write latency than its HP counterpart.

According to the results presented in Fig.3.2d, the read/hit latency of the LOP device type is higher than that of the HP device type. At a capacity of 16 KB, the difference between the two device types is 2×. However, the difference is observed to decrease with an increase in capacity. At 32MB, LOP SRAM has 6.08% more read latency than HP SRAM. As depicted in Fig.3.2e and 3.2f, when you need to access data quickly or write data, LOP SRAM uses less energy than HP SRAM. The energy reduction is about *half*. This means that LOP SRAM is more energy efficient. This difference in energy efficiency remains the same even when the capacity of SRAM increases. On average, LOP SRAM uses 1.96× less energy for reading data and 1.97× less energy for writing data than HP SRAM. The disparity in area and speed between the SRAM cache utilising LOP and HP devices is negligible when considering the application level performance(Fig.3.2). But, when it comes to power leakage, LOP devices are much better than HP devices, with an average of 71× less leakage. This difference becomes even more significant as the capacity of the memory increases. Therefore, LOP devices are a better choice for larger-capacity memory for embedded systems or computing systems requiring energy-efficient caches.

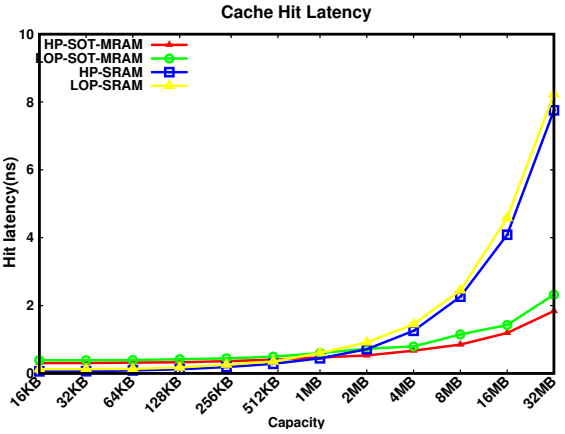


(a) Area

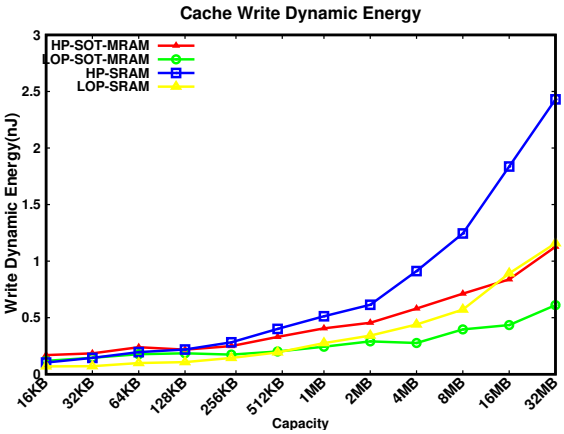
(b) Leakage Power



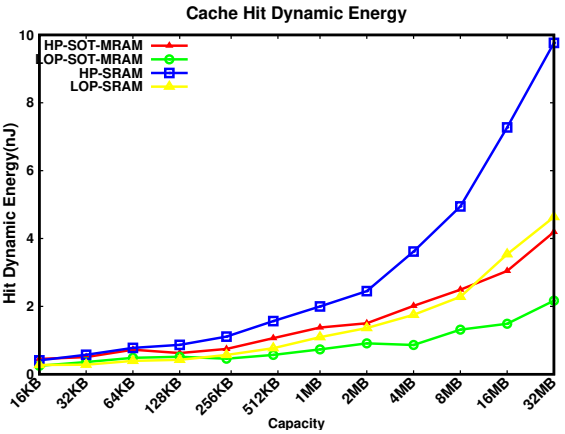
(c) Write Latency



(d) Hit Latency



(e) Write Energy



(f) Hit Energy

Figure 3.2: SRAM and SOT-MRAM with HP/LOP device cache capacities

LOP and HP Devices for a SOT-MRAM Memory Bank

This section evaluates the memory bank of SOT-MRAM using LOP and HP devices in a $45nm$ process. The parameters used for the evaluation are listed in Table-3.1 and Table 3.3. Our findings(Fig 3.2) demonstrate minimal differences in SOT-MRAM memory banks with LOP and HP devices, particularly for capacities ranging from

0 to 32 MB. However, at 32 MB(Fig.3.2a), the area of LOP SOT-MRAM is 8.8% larger than its HP counterpart due to the influence of cell density on the overall area. In terms of latency, the HP device type exhibits lower read latency than the LOP device type. As observed in Fig.3.2c and 3.2d, the read and write latencies are higher for LOP SOT-MRAM by an average of 26.25% and 16.55%, respectively. These performance differences between LOP and HP devices are noticeable, and the primary factors determining SOT-MRAM write latency include the duration of the write pulse, H-tree latency, and row decoder latency (Zitong Zhang and Jiang, 2022; Cargnini et al., 2014; Evenblij et al., 2019). Regarding power consumption, the LOP device type demonstrates significantly lower leakage power than the HP device type(Fig.3.2b). At 32 KB, the leakage power of the LOP device is 53 times lower than that of the HP device, and this difference increases to approximately 75 times at 32 MB. Additionally, the LOP device type has roughly half the dynamic energy and write dynamic energy of the HP device type(Fig.3.2e, 3.2f due to the lower MOS switching current required by LOP devices(Shao et al., 2021).

Based on the analysis, it can be inferred that the SOT-MRAM cache that employs LOP devices outperforms the cache that uses HP devices, particularly for larger capacities. This finding is similar to the performance characteristics exhibited by SRAM. LOP devices are typically utilized in applications where energy efficiency and power conservation are crucial, such as battery-operated devices, portable electronics, and energy-sensitive computing environments. On the other hand, HP devices are preferred in HPC applications, such as data centres, servers, and high-speed processors, where maximizing computational speed and throughput is the primary goal, even at the cost of higher power consumption (Gholami et al., 2024; Kalllinatha and Talawar, 2023).

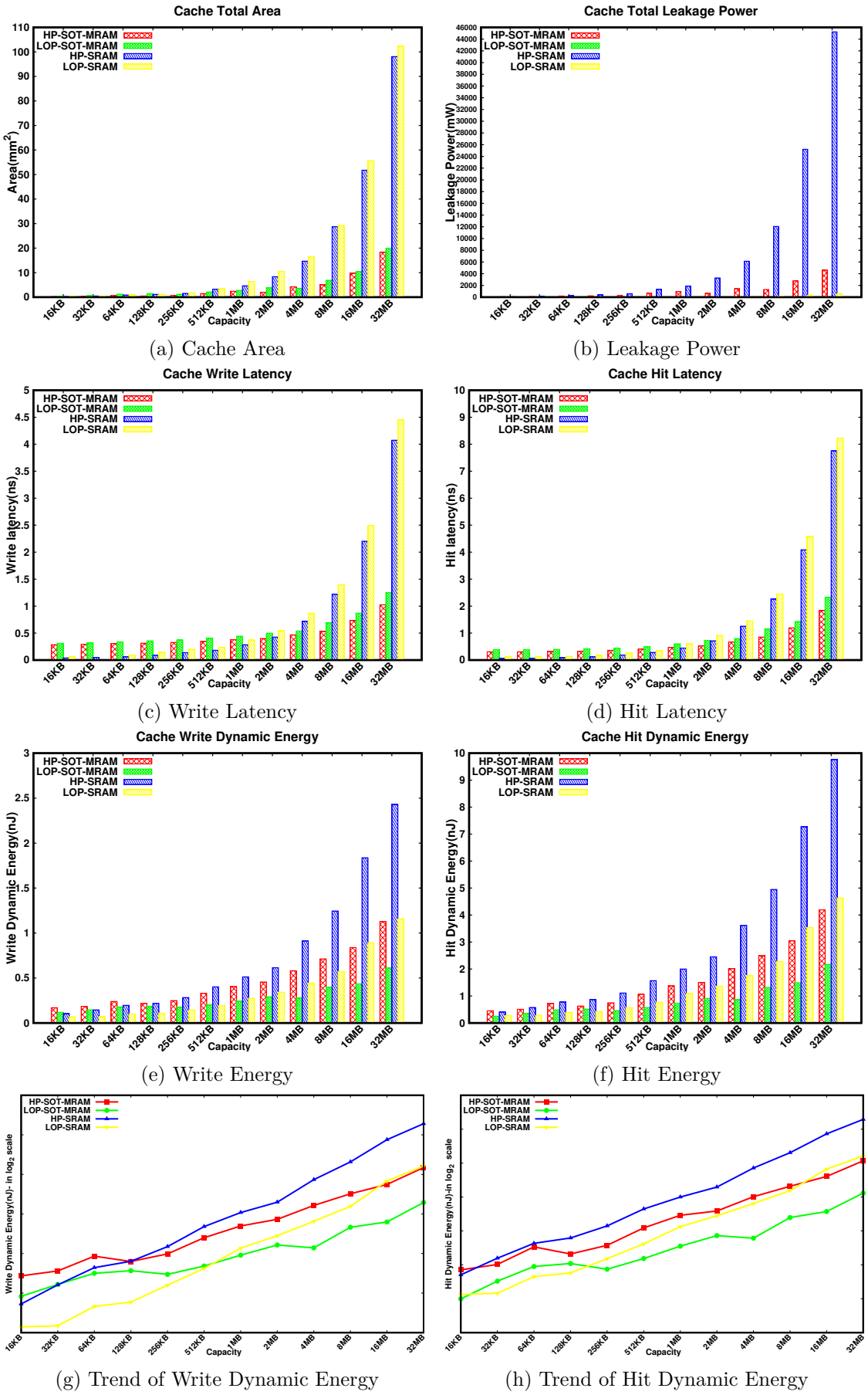
Comparison of SRAM and SOT-MRAM

The analysis presented in Fig.3.3 compares the performances of the HP and LOP devices in the SRAM and SOT-MRAM memory banks. At the capacity of 16 KB, the SOT-MRAM memory banks have a larger area than the SRAM memory bank. However, as the capacity increases, the area of the SOT-MRAM memory bank increases slowly, while the area of the SRAM memory bank increases exponentially. At 32

MB, the area of the SRAM memory bank is approximately $4\times$ larger than that of the SOT-MRAM memory bank, as shown in Fig.3.3a.

At smaller capacities, the area of 512KB HP-SOT-MRAM is 55.60% less than HP-SRAM. In contrast, the same for LOP counterparts is a 41.07% reduction in chip area for LOP-SOT-MRAM.So, there is a difference of 14.53% between the HP and LOP-SOT-MRAM cache chip areas. From 16KB to 256KB, the difference is relatively comparable, suggesting that the specific technology has a limited impact on the cache area up to 256KB because the peripheral circuitry dominates the chip area in either type of caches(Han and Jiang, 2024). However, significant differences become apparent as the cache capacity increases beyond 1MB. HP and LOP SOT-MRAM technologies exhibit moderate increases in the area(18.3 and 19.9mm^2), indicating their efficient design and higher cell densities. In contrast, HP-SRAM and LOP-SRAM show substantial area increases, with LOP-SRAM having the largest area of 102.28mm^2 at 32 MB. The increase in chip area in SRAM is attributed to the inherently larger footprint of SRAM cells and the additional circuitry required for managing larger capacities (Liao et al., 2020). Therefore, in a capacity below 1MB, the density replacement is not feasible. For the cache size between 2MB to 8MB, we can increase density by $2\times$, whereas for larger capacity caches, the density can be increased up to $4\times$. On the other hand, SRAM technologies, particularly LOP-SRAM, are optimized for lower power consumption, resulting in a larger area due to additional power management circuitry. With its smaller area footprint and high performance, HP-SOT-MRAM is well-suited for high-speed processors and performance-critical systems. LOP-SOT-MRAM balances area efficiency and power consumption, making it suitable for portable electronics and battery-operated devices. In contrast, LOP-SRAM is suited for energy-sensitive systems, but its chip area is the largest. Therefore, HP-SOT-MRAM can replace both HP and LOP-SRAM in terms of area.

Fig.3.3b illustrates the total leakage power of HP and LOP memory banks of SOT-MRAM and SRAM. At a small capacity of 16 KB, the leakage power of the HP-SOT-MRAM memory bank is 2.23% more than the SRAM memory bank. However, as the capacity increases, the leakage power of SRAM device-based memory banks rises exponentially. At 512KB, HP-SOT-MRAM can save 51% leakage power than HP-SRAM, whereas LOP-SOT-MRAM can reduce 41% compared to LOP-SRAM. It is



(a) Cache Area

(b) Leakage Power

(c) Write Latency

(d) Hit Latency

(e) Write Energy

(f) Hit Energy

(g) Trend of Write Dynamic Energy

(h) Trend of Hit Dynamic Energy

Figure 3.3: Characteristics of SRAM and SOT-MRAM with HP/LOP device cache at varying capacities

worth noting that both SRAM memory banks exhibit a more rapid increase in leakage power than the SOT-MRAM memory banks beyond 1MB. At the highest capacity of 32 MB, the leakage power of SRAM is *10*× higher than that of SOT-MRAM for both LOP and HP devices. The total reduction in leakage power of both HP and LOP counterparts is 89.80% for SOT-MRAM. This difference can be attributed to the inherent properties of MTJ devices in SOT-MRAM, which are less susceptible to leakage. Most leakage in SOT-MRAM comes from peripheral circuits such as MOS devices and sense amplifiers rather than the MTJ cells themselves(Singh et al., 2020; Shao et al., 2021). SOT-MRAM offers superior leakage power performance compared to SRAM, particularly at larger capacities, making it the preferred option for high-capacity, energy-efficient memory designs. Integration of SOT-MRAM into the LLC could yield significant power savings and enhanced overall performance.

The write and read latency of the SOT-MRAM memory bank in Fig.3.3c,3.3d is initially *2*× higher than that of the SRAM memory bank at a capacity of 16 KB. However, as the capacity increases beyond 1MB, the read and write latency of the SOT-MRAM grows slowly. In contrast, both latencies of the SRAMs increase significantly to around *fourfold* that of the corresponding SOT-MRAM cahe capacity. The main factor contributing to the write delay of SOT-MRAM is the time needed for the MTJ device switching, which remains constant irrespective of the capacity. Conversely, as the capacity increases, the H-tree latency increases for the SRAM, leading to a longer write path and a rapid increase in write latency(Oboril et al., 2015; Saha et al., 2022). The above analysis underscores the advantages of SOT-MRAM, especially in LLC applications, where its latency remains stable and relatively low across a wide range of cache sizes.

The comparison between the dynamic energy of SRAM is *2*× less than SOT-MRAM at 16KB capacity. However, as the capacity increases beyond 1MB, both SRAM memories rapidly increase read and write energy. At 32MB capacity, the dynamic energy of both forms of SRAM rapidly increases to *twofold* than SOT-MRAM in HP/LOP form. This disparity is depicted in Figure 3.3e and 3.3f. The trend in dynamic energy consumption is further illustrated in Figure 3.3g and 3.3h, demonstrating that SRAM's energy is more sensitive to size than SOT-MRAM.Moreover, beyond 512KB, both read/write energy of SRAM in the HP/LOP version increases more rapidly with

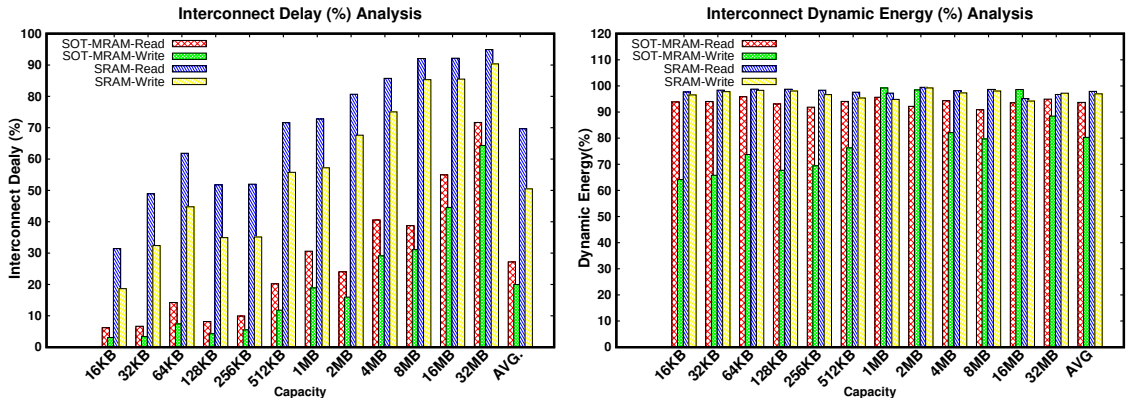
cache size. Therefore, for future memory architectures that prioritize high performance without compromising on speed or energy efficiency, HP/LOP-SOT-MRAM of 1MB capacity and beyond is more suitable for LLCs than SRAM.

Interconnect delay Analysis

The interconnect delay is a critical factor affecting access latency in SRAM and SOT-MRAM memory cells. In Fig.3.4a, for smaller capacities, such as 16 KB, the interconnect delay contributes 31.43% and 18.65% of the total read and write latency, respectively, for SRAM and less than 6.20% and 3.08% of the total read and write latency respectively for SOT-MRAM. However, as the capacity increases, this delay becomes a more significant portion of the latency. For example, at 32 MB, the interconnect delay for SRAM can make up to 90% of the total access latency, while for SOT-MRAM, it accounts for up to 70%. The average interconnect delay of SRAM read 70% and write is 50% of the total latency. The average SOT-MRAM interconnect delay is 27% of the read and 20% of the write delay. This difference highlights the superior scalability of SOT-MRAM compared to SRAM. The smaller interconnect delay in SOT-MRAM is primarily due to its smaller cell size and a 3D vertical structure, which enable shorter and more efficient signal paths. It is crucial to understand that the analysis shows the delay as a proportion of the total access latency rather than the actual delay values. In reality, the delay values for SRAM are approximately three times higher than those for SOT-MRAM at every capacity level.

The increasing size of SRAM results in longer wire lengths and more intricate routing, leading to heightened interconnect delay due to increased resistance and capacitance(Lu et al., 2024). On the other hand, the 3D structure of SOT-MRAM allows for more condensed cell arrangements, reducing interconnect lengths and minimizing delays. This characteristic makes SOT-MRAM ideal for high-density LLC designs and modern memory architectures(Gholami et al., 2024; Van Beek et al., 2023).

Fig.3.4b delineates the relationship between dynamic energy involved in interconnection as a fraction of the overall energy expended for access operations within SRAM and SOT-MRAM at varied memory sizes ranging from 16 KB to 32 MB. At the commencement point of 16 KB, the energy consumed by SRAM during read and write operations constitutes approximately 97% and 96%, respectively, of the aggregate



(a) Delay (b) Energy

Figure 3.4: Interconnect Delay and Energy Analysis

access energy. Conversely, SOT-MRAM showcases notably reduced energy consumption, with the energy for reading and writing activities around 93% and below 64%, respectively. This difference persists with increasing memory capacities, emphasising SOT-MRAM's enhanced efficiency.

As the memory capacity grows to considerable sizes, such as 1 MB, 2 MB, and larger, the variances in energy consumption profiles between SRAM and SOT-MRAM become increasingly disparate. Specifically, at a 4 MB capacity, SRAM's interconnect energy per access for reading almost touches 98.17% of the total access energy, while writing energy is around 97%. In contrast, SOT-MRAM demonstrates a lower propensity for interconnect dynamic energy consumption, sustaining read energy at around 94% and write energy at approximately 82%. This tendency persists across memory sizes; on average, the read and write interconnect energy of SRAM is 98% and 97% of total latency where, whereas SOT-MRAM energy is 94% and 82%, respectively. This distinct energy efficiency underscores SOT-MRAM's potential as a preferable option for applications demanding high capacity and performance.

3.4.2 Scale effect of bit cell structural influences on cache memory

After analysing the results of the previous section for cache capacities smaller than 1MB, peripheral circuits are the main influencing factor on the circuit's performance. As a result, the remainder of the section will focus on analyzing cache capacities ranging from 1MB to 32MB for LLC applications.

The structural characteristics of bit cells, such as cell area and aspect ratio, play a crucial role in determining the performance metrics of SOT-MRAM cache memory. The structural parameters effects on write latency and leakage power are summarized in Tables 3.4, 3.5, and 3.6.

The MFS framework uses three bit-cell configurations ranging from a 12 to a 24 Cell Area(CA) with a fixed Aspect Ratio(AR). The results are analysed to discuss their trade-offs. A 12/1 configuration offers lower write latency but higher leakage power, prioritizing speed over power conservation. The 24/1 setup is more energy-efficient but with longer write latency. The analysis emphasizes the need for designers to balance these factors when tailoring cache memory for different applications. Table-3.4 shows

Cell Area(F^2)			
Size	**12**	**18**	**24**
1MB	0.37875	0.384754	0.391686
2MB	0.398096	0.41184	0.425901
4MB	0.466465	0.48176	0.505281
8MB	0.52961	0.636898	0.674353
16MB	0.73418	0.828511	0.919948
32MB	1.0236	1.1636	1.3395

Table 3.4: compares the write latency (in ns) of SOT-MRAM caches with varying CAs, keeping the AR=1.

the correlation between the SOT-MRAM bit-cell CA size and the cache write latency. Analysis indicates that as the cell area increases, the write latency also increases. When the CA increases from 12 to 18, there is a 9% increase in average write latency. This tendency continues as the CA grows from 18 to 24, with a 7% increase in latency. When directly comparing the smallest and largest cell sizes, 12 and 24, a significant 17% increase in write latency is observed.

Furthermore, this correlation between cell size and write latency persists across different cache sizes. Within the same cache size, write latency consistently rises alongside increasing CA. For example, a cache size of 1MB experiences a modest 2.2% increase in write latency with CA growth, with this percentage incrementally rising for larger cache sizes - reaching 4.6% for 2MB, 5.4% for 4MB and peaking at 19.8% for a 32MB cache. The 8MB and 16MB caches deviate from this gradual increase pattern,

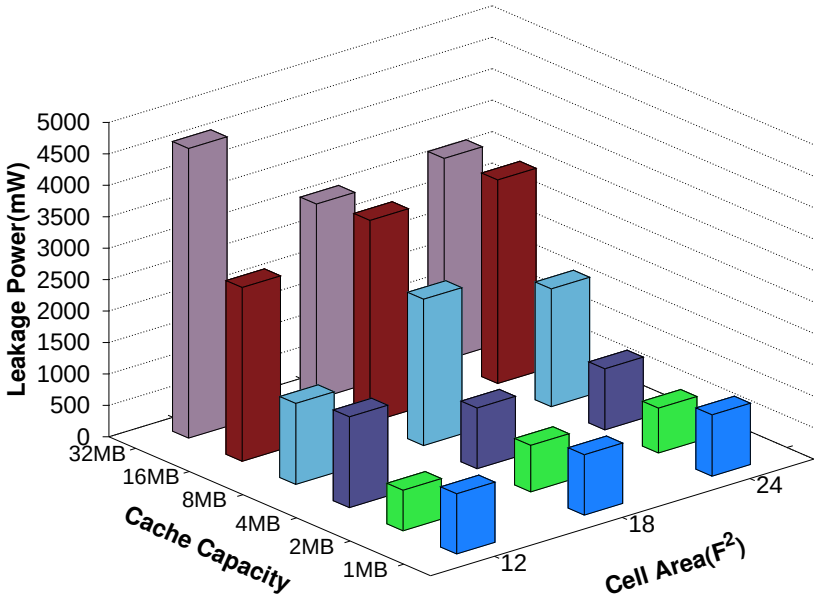
showing substantive increases of 17.8% and 16.3% in their respective latency. This trend indicates that larger cell areas, while potentially improving other characteristics such as operational stability, inherently lead to higher write latencies.

SIZE	**Aspect Ratio**			
	0.5	**1**	**1.5**	2
1MB	0.403772	0.391683	0.373032	0.367461
2MB	0.449925	0.425901	0.444489	0.428493
4MB	0.499503	0.505281	0.511825	0.53881
8MB	0.622124	0.674353	0.636376	0.579635
16MB	0.93308	0.910948	0.913579	0.845725
32MB	1.25094	1.3395	1.21809	1.19985

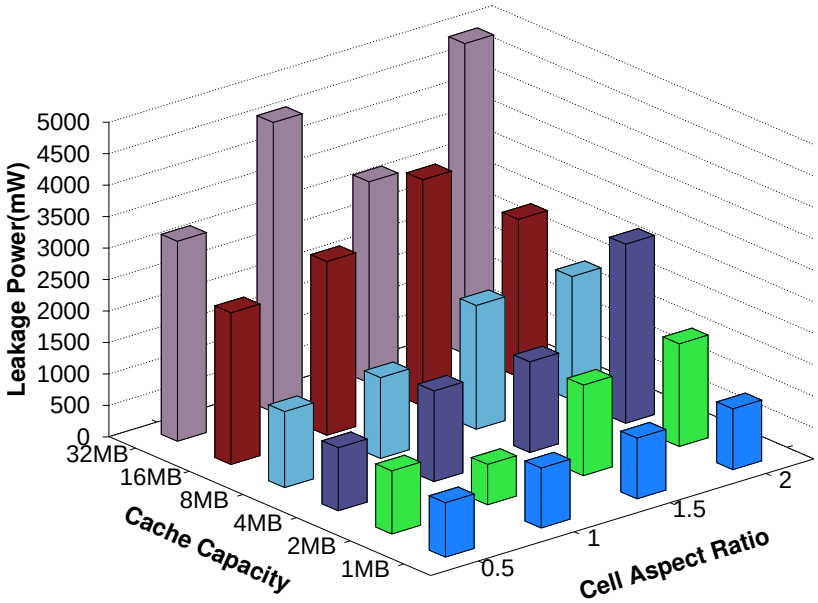
Table 3.5: To study the effect of Write Latency(ns) with different aspect ratios and for 24 F^2 constant Cell area.

Table 3.5 shows the effect of different aspect ratios on write latency for a constant cell area of 24 F^2. The results indicate that aspect ratios also significantly influence write latency. For example, at 1MB, the write latency decreases from 0.403772 ns for an aspect ratio of 0.5 to 0.367461 ns for an aspect ratio of 2.0. This reduction suggests that higher aspect ratios can help mitigate the increase in write latency associated with larger cell areas, potentially due to more efficient packing and signal propagation. The accompanying Fig.3.5b illustrates how the performance of the MTJ device changes with varying aspect ratios. For a device cell area of $24F^2$, the following observations can be made: 1) (4.89898Fx4.89898F) is the device length and width with an aspect ratio of 1.0. 2) 6.9282F is the device length, and the width is about 3.4641F for the aspect ratio of 0.5. A larger aspect ratio increases leakage power for smaller capacity caches, such as 2MB and 4MB. Aspect ratios of 0.5 and 1.5 reduce the leakage power for all cache sizes compared to aspect ratios 1 and 2.

In Table 3.6, you can find information about the impact of transitioning from a 12/1.0 to a 24/2.0 configuration on write latency and leakage power. The results indicate that different cache sizes experience varied effects. For example, at 32MB, there is a 17.2% increase in write latency but a significant 50.6% decrease in leakage power. This highlights the trade-off in optimizing these parameters and emphasizes the need for balanced design considerations.



(a) Cell Area Scaling



(b) Aspect ratio Scaling

Figure 3.5: Cell area and Aspect ratio Scaling Analysis

SIZE	Configuration	Write Latency(ns)	Leakage Power(mW)
1MB	12/1 to 24/2	-2.9%	0.6%
2MB	12/1 to 24/2	7.6%	-35.6%
4MB	12/1 to 24/2	15.5%	0.7%
8MB	12/1 to 24/2	9.4%	-26.3%
16MB	12/1 to 24/2	15.1%	-29.6%
32MB	12/1 to 24/2	17.2%	-50.6%

Table 3.6: The effect on Write Latency(ns) and Leakage Power for 12/1.0 to 24/2 targeted configuration.

Fig.3.5b depicts the variation in leakage power across different cache capacities and AR with a fixed CA. Generally, higher aspect ratios correspond to increased leakage power, particularly noticeable in larger caches such as 4MB and 32MB. For example, the leakage power of the 4MB cache jumped from 1443.83 mW at an aspect ratio of 1 to 2851.39 mW at an aspect ratio of 2. Likewise, the leakage power of the 32MB cache rapidly increased from 3183.76 mW at an aspect ratio of 0.5 to 4934.74 mW at an aspect ratio of 2. The average leakage power across all capacities also rises with higher aspect ratios, going from 1615.04 mW at an aspect ratio of 0.5 to 1951.38 mW at an aspect ratio of 1 to 2478.41 mW at an aspect ratio of 2. This indicates a trade-off between aspect ratio and power efficiency, highlighting the need for careful optimization in cache design.

Practical implications of the stable and lower latencies with reduced leakage power and interconnect dynamic energy position SOT-MRAM as a highly scalable and energy-efficient solution for LLC designs.

The leakage power analysis in Fig.3.5a examines the leakage power (in mW) for various cache capacities and cell areas with a fixed AR of 1. For a 1MB cache, the leakage power remains relatively consistent across different cell areas, ranging from 956.832 mW to 975.777 mW. In the case of a 32MB cache, the leakage power is consistently high across all cell areas, peaking at 4608.14 mW for 12 but demonstrating lower values for 18 and 24 at 3105.21 mW and 3212.78 mW, respectively. On average, the leakage power decreases as the cell area increases, with an average of 1951.3836 mW for 12, 1885.605 mW for 18, and 1831.1445 mW for 24. This analysis underscores that larger cell areas generally result in lower average leakage power, although this

trend is not uniform across all cache capacities.

The current density of the MTJ device, which is influenced by the aspect ratio (AR) determined by the switching current and the width and length of the MTJ, is related to its switching characteristics. The access CMOS transistor width is also related to the switching current. It can be observed from Fig.3.5a that the cache size has a more significant impact on the leakage power than the cell area. The analysis indicates that optimizing cache capacity, with cell area, is paramount in minimizing leakage power in SOT-MRAM memory. It is essential to balance both parameters to ensure optimal performance. Choosing the 12/1 and 24/2 CA/AR configurations provides balanced parameters for conducting comprehensive MFS studies on the technology node. The 12/1 configuration allows exploration of the impact of a moderate aspect ratio, finding a balance between power and performance. The 24/2 configuration represents a higher aspect ratio, offering insights into the upper limit of the AR on area, leakage power and latency. Analysing these configurations allows for thoroughly examining power-performance trade-offs, scalability, and optimisation strategies in crafting efficient cache systems, ensuring high performance while minimising power consumption. The next subsection analyses the technology node MFS results.

3.4.3 MFS on Technology Node

In this section, the trade-off analyses results of using technology nodes from 65nm to 22nm on SOT-MRAM cache performance from 1MB to 32MB.In Fig.3.6a, the impact of shrinking technology nodes on the total area of SOT-MRAM caches is demonstrated. As the technology node scales down from 65nm to 45nm, the chip area reduces by 2.6×, similarly from 45nm to 32nm and from 32nm to 22nm, the chip area reduces by 2×. The total chip area of the 32MB cache at 65nm decreases 10× at 22nm node. Also, the CA/AR increase of 12/1 to 24/2 increases chip area by 34%. This exemplifies the efficiency gains that can be realized with advanced process technologies with smaller CA/AR configurations. These are essential for enhancing the density and performance of memory systems for modern applications.

Fig.3.6b shows the trends in leakage power scaling for SOT-MRAM across technology nodes. As the technology node scales from the old to the new generation, the chip leakage power is reduced by *half*. The leakage power for a 32MB cache decreases from

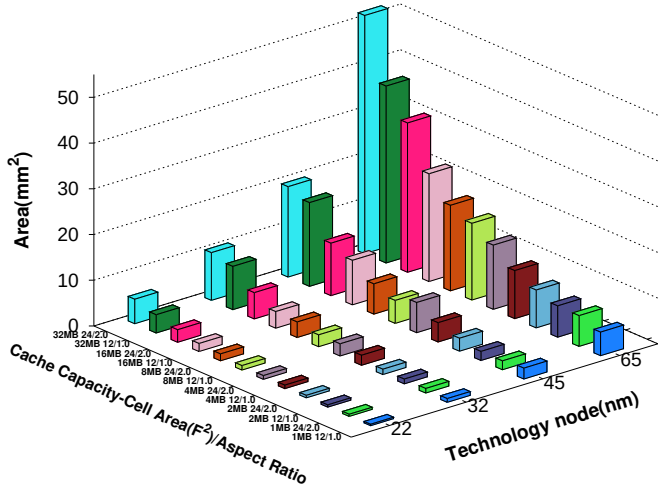
approximately 10000 mW at 65nm to around 2250 mW at 22nm, a 4× reduction in leakage power. This reduction in leakage power directly contributes to overall power consumption reduction and enhanced energy efficiency in the cache. This improvement is advantageous for applications where power efficiency is a consideration.

Fig.3.6c depicts the relationship between write latency, technology nodes, and cache capacity with 12/1 and 24/2 CA/AR cells. While smaller technology nodes lead to improvements in area and leakage power, they also result in increased write latency. As the technology node scales from the old to the new generation, the write latency increases by 10%. For instance, the write latency for a 32MB cache grows from around 1.2ns at 65nm to approximately 1.5ns at 22nm. This increase underscores the need for careful optimization in memory design, as there is a trade-off between achieving lower area and power consumption and maintaining write latency.

In summary, the analysis provides valuable insights into the trade-offs and benefits of scaling technology nodes for the SOT-MRAM cache MFS. It highlights the need to balance area, leakage power, and write latency when designing efficient memory architectures for modern computing environments. The 12/1 CA/AR for density replacement studies balance latency, leakage power, and area efficiency at the 45nm technology node. Further exploration of SOT-MRAM with contemporary applications evaluation continues in the next subsection.

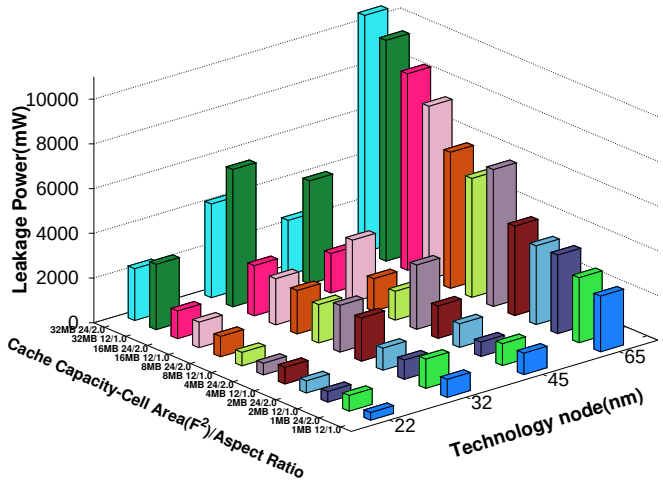
3.5 Density replacement studies to enhance modern application performance

In this study, we investigate the potential performance impact of replacing SRAM with SOT-MRAM in modern computing applications, such as social network graph processing, NLP language models, image processing DNN, and generic traffic applications. The detailed evaluation of these modern applications can be found in Table 3.2. This research is crucial as it examines the potential for replacing SRAM with SOT-MRAM, which could significantly enhance the efficiency and performance of contemporary computing systems. By analyzing the performance of SOT-MRAM at equal, 2x, and 4x the size of SRAM, we aim to determine the most effective configuration for various applications. The insights from the MFS study presented in the previous



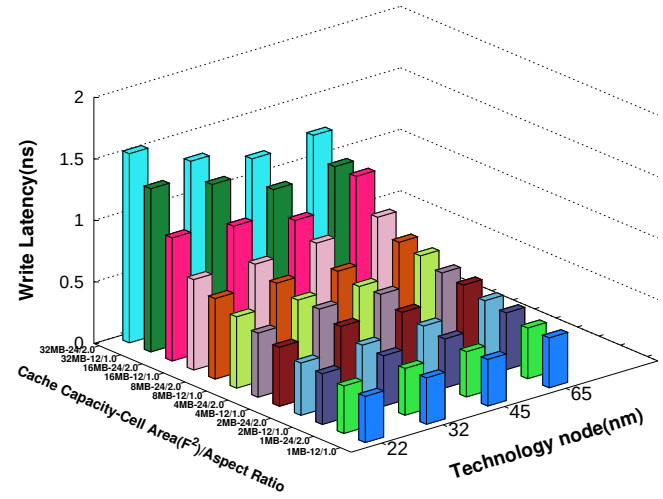
(a) Area Scaling road-map

Leakage Power Scaling



(b) Leakage Power Scaling

SOT-MRAM Scaling Roadmap of Write Latency



(c) Write Latency

Figure 3.6: Technology Road-map scaling

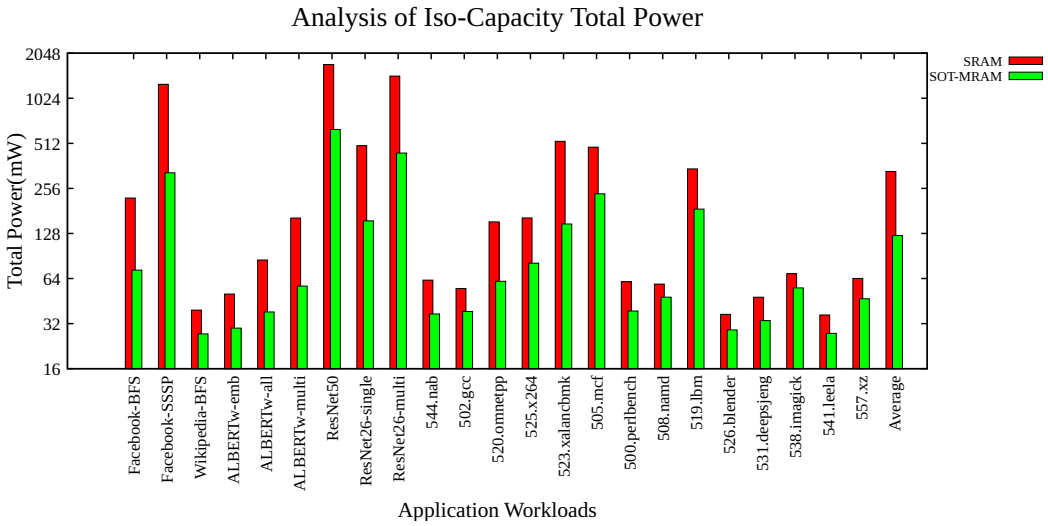
sections provide a strong foundation for selecting optimal parameters, such as cell area and aspect ratio, ensuring that our performance evaluations are based on optimized configurations.

Our density analysis focuses on two scenarios: iso-capacity and iso-area replacements. This section compares the performance of a 1MB SRAM with a 1MB SOT-MRAM in terms of iso-capacity. This work also evaluates replacing a 4MB SRAM cache with an 8MB SOT-MRAM cache regarding iso-area. This work examines how SOT-MRAM can fit within the same chip area as SRAM while offering higher capacity. These comparisons are based on the insights derived from the analysis in sections 3.4.2 and 3.4.3. This comprehensive analysis aids in assessing the feasibility and advantages of replacing SRAM with SOT-MRAM in high-density, high-performance computing environments, ultimately contributing to more efficient and scalable memory architectures.

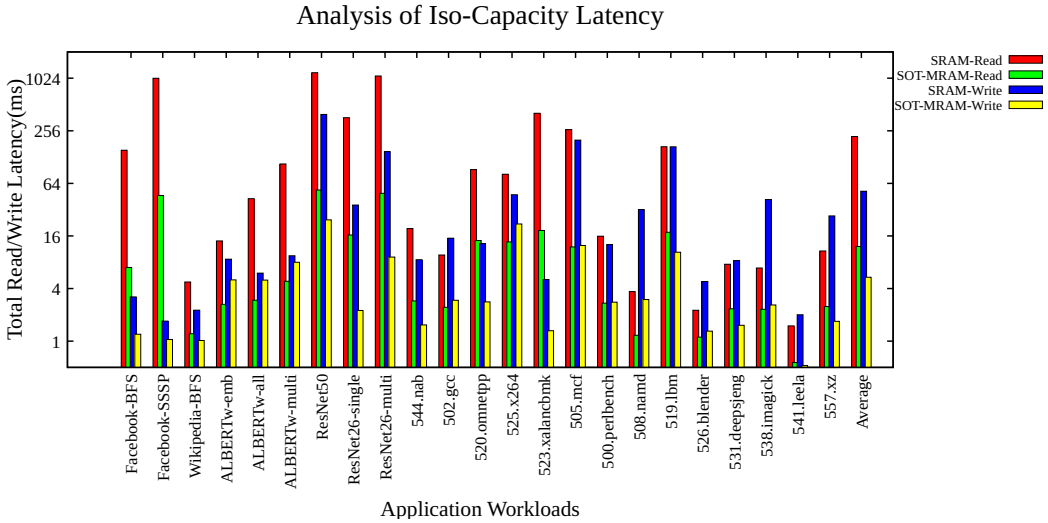
3.5.1 Iso-Capacity Analysis

In the iso-capacity density analysis, we are assessing the performance of 1MB SRAM compared to 1MB SOT-MRAM based on a 12/1 cell area/aspect ratio at 45nm. This design was identified in sections 3.4.2 and 3.4.3 as an optimal configuration for our study, striking a balance between area, leakage power, and latency. The analysis of 1:1 MB iso-capacity as in Fig.3.7a illustrates the power efficiency advantage of SOT-MRAM over SRAM across diverse application workloads. SOT-MRAM consistently consumes less power than SRAM, with the magnitude of the difference varying based on the workload. For instance, in the Facebook-BFS workload, SOT-MRAM consumes roughly 75% less power than SRAM. Similarly, for the ALBERTw-multi workload, the power savings with SOT-MRAM are approximately 65%. Even in demanding scenarios like the SPEC benchmark 525.x264, SOT-MRAM demonstrates a reduction in power consumption of nearly 50%. On average, the total power consumption across all workloads is significantly lower for SOT-MRAM, with an average power reduction of about 60% compared to SRAM.

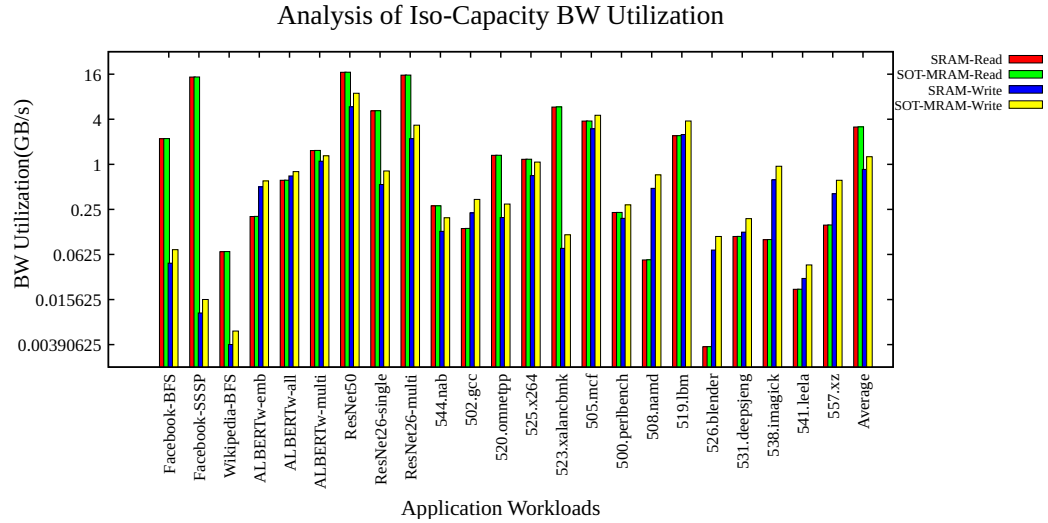
This points to the energy efficiency benefits of SOT-MRAM in contemporary AI and NLP applications, underscoring its lower power consumption, cost-saving potential, and extended battery life in mobile devices. The scalability of SOT-MRAM makes



(a) Application Total Power

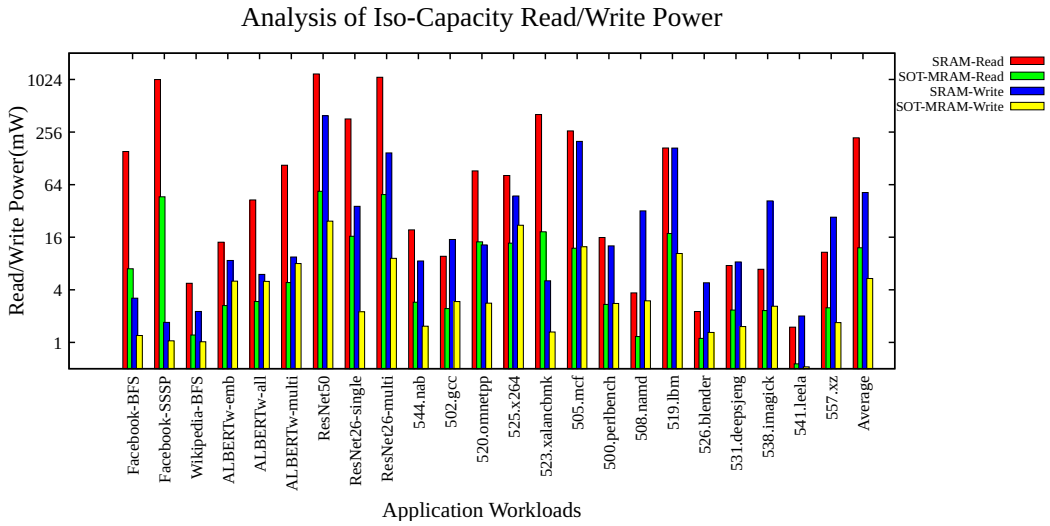


(b) Application Latency

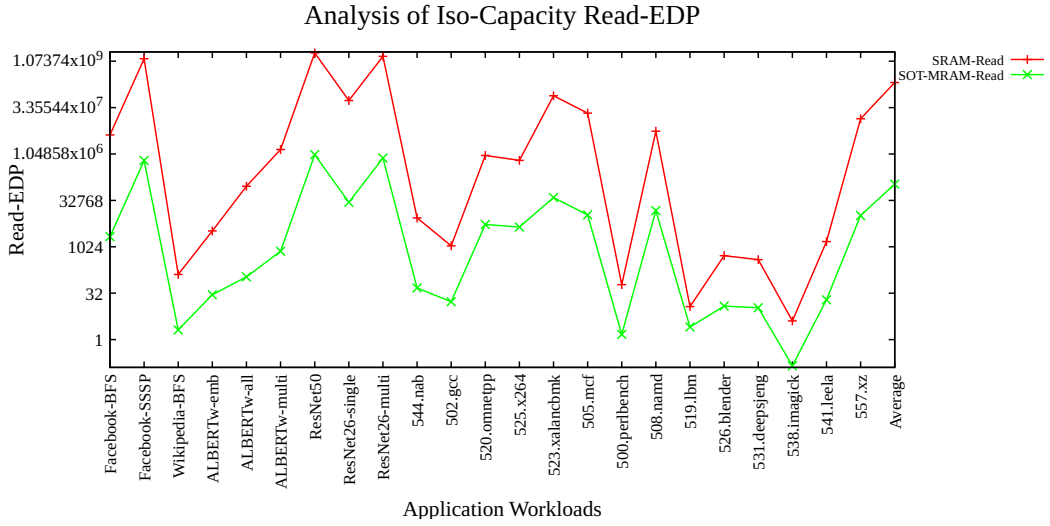


(c) Application Bandwidth Utilization

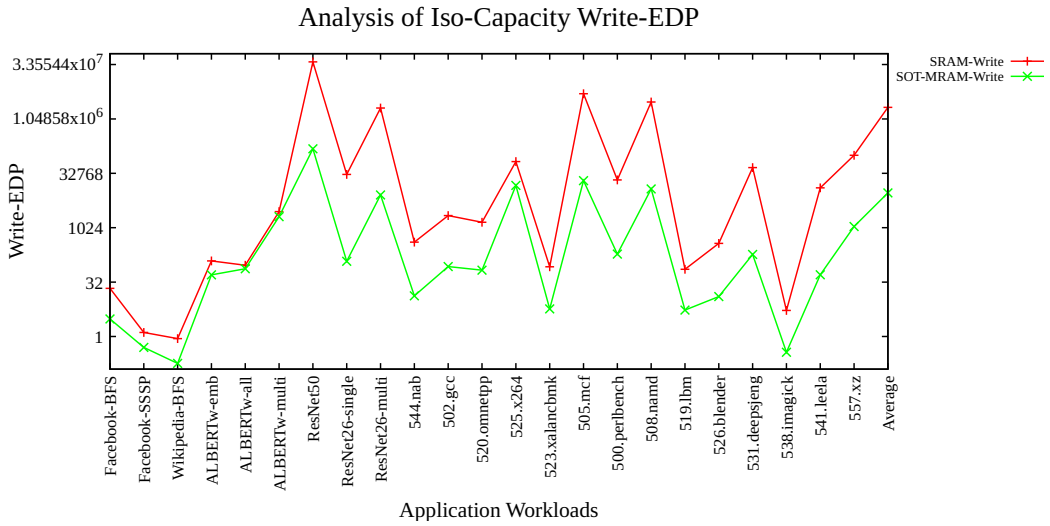
Figure 3.7: Modern Application Performance Analysis for *iso-capacity(1MB)*



(a) Application Read/Write Power



(b) Read EDP



(c) Write EDP

Figure 3.8: Modern Application Read/Write Power and EDP *iso-capacity(1MB)*

it well-suited for a variety of applications, positioning it as a compelling substitute for SRAM in high-performance and energy-efficient memory designs. The analysis presented in Fig.3.7b illustrates a significant contrast in read and write latencies between SRAM and SOT-MRAM across various application workloads. SOT-MRAM consistently demonstrates notably lower latency when compared to SRAM for most workloads. For example, in the Facebook-BFS workload, the total read latency for SRAM is approximately 256 ms, whereas for SOT-MRAM, it is approximately 64 ms, signifying a 75% decrease in latency. Similarly, for the ALBERTw-multi workload, SRAM's total read latency is about 512 ms, while SOT-MRAM's is around 128 ms, displaying a 75% improvement. Write latencies follow a similar pattern, with SOT-MRAM consistently outperforming SRAM. For the SPEC benchmark 525.x264, SRAM's write latency is roughly 1024 ms, while SOT-MRAM's is around 256 ms, indicating a 75% reduction.

The iso-capacity bandwidth utilization analysis presented in Fig.3.7c shows that SOT-MRAM consistently demonstrates significantly higher bandwidth utilization across all workloads than SRAM. This is due to SRAM's lower utilization because of longer H-tree paths and more complex routing in larger capacities. The average read bandwidth utilization for SOT-MRAM is approximately 40% higher than SRAM, and the write bandwidth utilization is about 50% higher. These performance advantages highlight the potential of SOT-MRAM to deliver faster data access and improved overall system performance, positioning it as an excellent replacement for SRAM in high-capacity cache implementations.

The analysis of iso-capacity read/write power in Fig.3.8a indicates that SOT-MRAM consistently consumes less power than SRAM across a range of application workloads, with approximately 75% lower read and write power on average. This demonstrates the superior energy efficiency of SOT-MRAM, making it a compelling alternative to SRAM in large cache implementations. Fig.3.8b and 3.8c compare the Read Energy-Delay Product (EDP) and Write Energy-Delay Product (EDP) for iso-capacity of SOT-MRAM and SRAM. SOT-MRAM shows an 85% decrease in Read EDP and an 80% decrease in Write EDP compared to SRAM, on average, across various application workloads. SOT-MRAM's efficiency is attributed to its lower read and write latency and energy consumption, making it a superior choice for read and write-intensive applications and enhancing overall system performance.

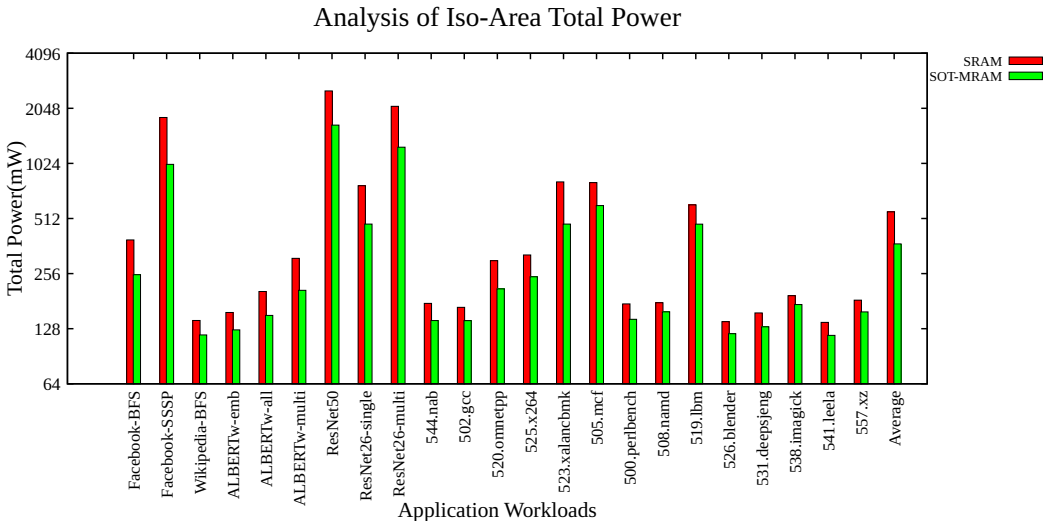
In conclusion, the iso-capacity analysis indicates that SOT-MRAM demonstrates significantly superior power efficiency to SRAM across different workloads, consistently using approximately 60% less power on average. Moreover, SOT-MRAM outperforms SRAM in read and write operations, showcasing average latency reductions of 75%. Additionally, SOT-MRAM achieves approximately 40% higher read bandwidth and 50% higher write bandwidth than SRAM. When considering Read EDP and Write EDP for iso-capacity, SOT-MRAM displays considerable reductions compared to SRAM, making it the optimal choice for read- and write-intensive applications that enhance overall system performance.

3.5.2 Iso-Area Analysis

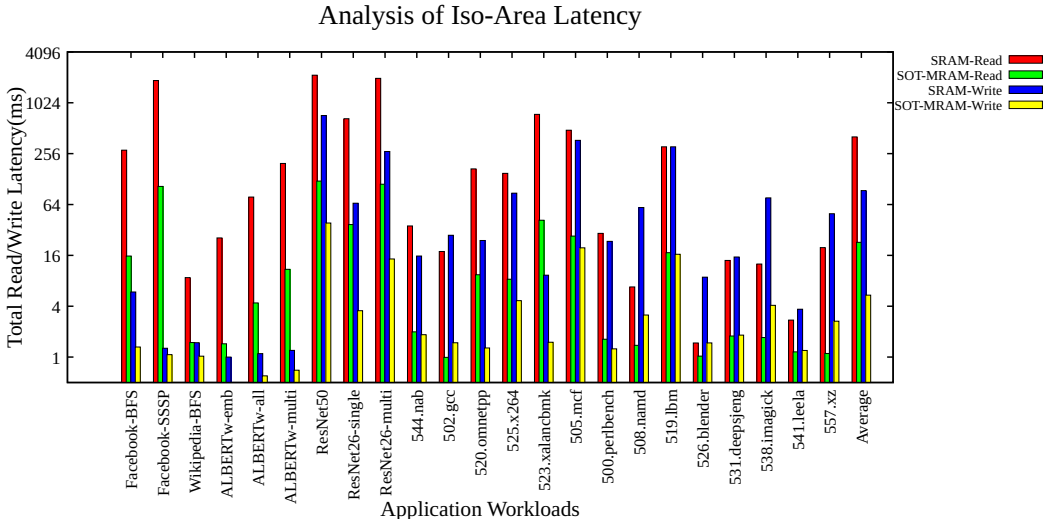
In the iso-area density analysis, we assess the performance of 4MB SRAM compared to 8MB SOT-MRAM based on a 12/1 cell area/aspect ratio at 45nm. This design was identified in sections 3.4.2 and 3.4.3 as an optimal configuration for our study, striking a balance between area, leakage power, and latency.

In an iso-area analysis(Fig.-3.9), replacing a 4MB SRAM cache with an 8MB SOT-MRAM cache significantly decreases overall power consumption across different application workloads. Fig.-3.9a shows that applications like Facebook-BFS and ResNet50 consume up to 2048mW with SRAM. At the same time, SOT-MRAM demonstrates considerably lower power consumption at around 1024mW, marking a 50% reduction. This pattern is consistent across workloads such as ALBERTw-all and 523.xalancbmk, with SOT-MRAM consistently exhibiting lower power consumption. On average, SOT-MRAM showcases a 45% reduction in power consumption compared to SRAM, emphasizing its superior energy efficiency.

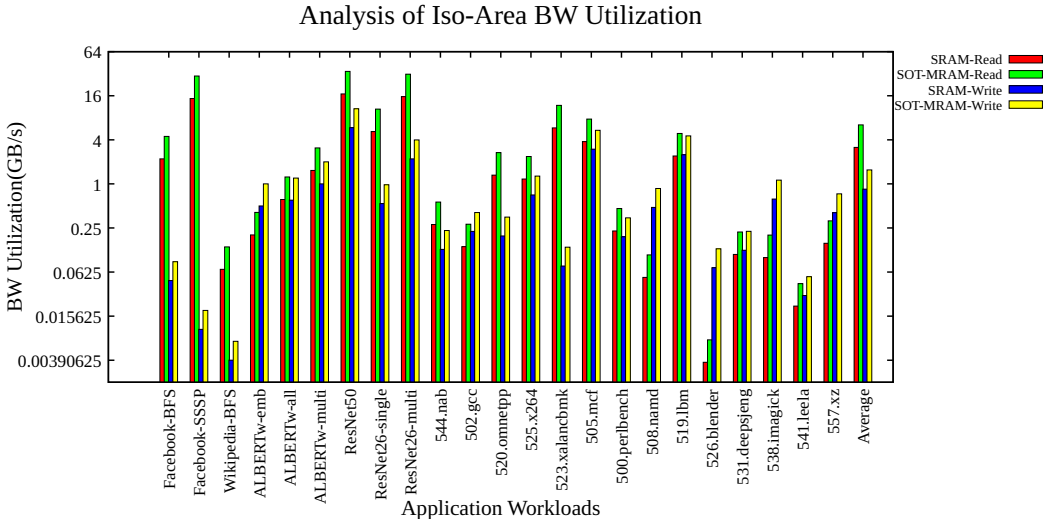
Fig.-3.9b shows that for applications like Facebook-BFS and ALBERTw-all, the read latency for SRAM can go up to 1024ms, whereas for SOT-MRAM, it is significantly lower at around 64ms. Similarly, the write latency for SRAM is consistently higher across all workloads compared to SOT-MRAM. On average, SOT-MRAM achieves a latency reduction of about 70% and 65% for read and write operations, respectively, because of the faster switching times of SOT-MRAM cells and their efficient handling of read and write operations. This makes SOT-MRAM a highly efficient alternative to SRAM in high-capacity memory applications.



(a) Applications Total Power

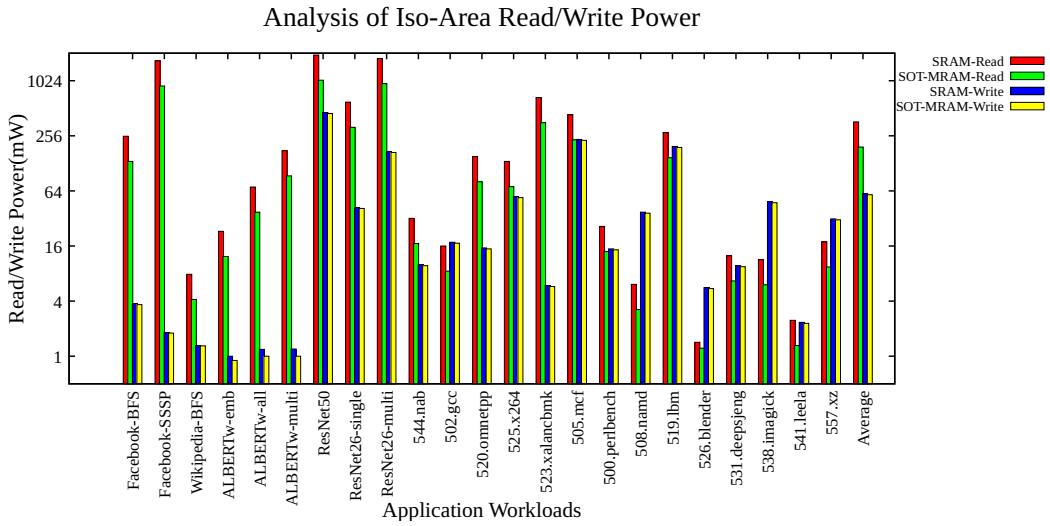


(b) Applications Total Latency

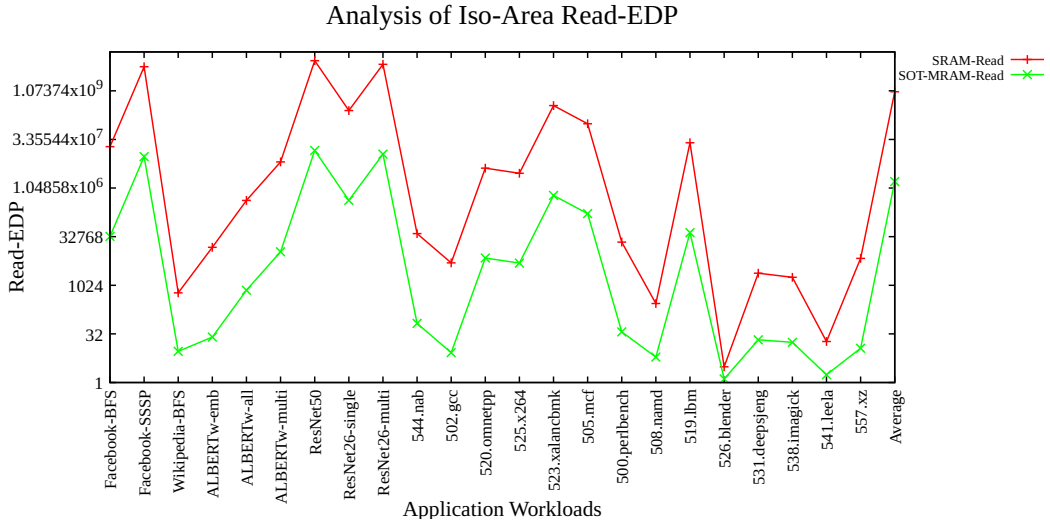


(c) Bandwidth Utilization

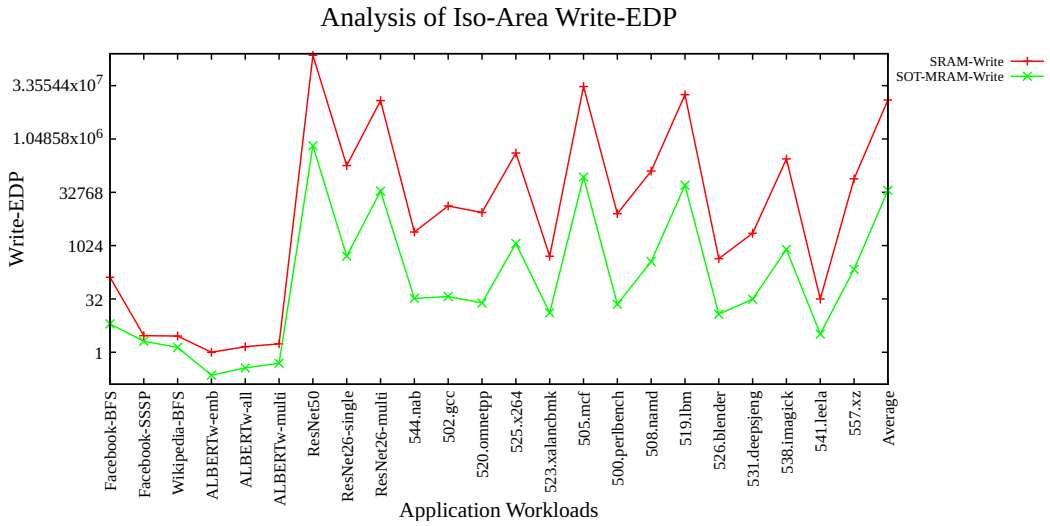
Figure 3.9: Modern Application Performance Analysis for *iso-area* of 4MB and 8MB



(a) Read/Write Power



(b) Read EDP



(c) Write EDP

Figure 3.10: Modern Application EDP and Read/Write Power for *iso-area* of 4MB and 8MB

Furthermore, in Fig.3.9c, SOT-MRAM demonstrates about 30% higher bandwidth utilization on average compared to SRAM. This increased bandwidth utilization allows SOT-MRAM to effectively manage more data transfers per unit of time. It is well-suited for applications requiring high data throughput, such as AI and NLP workloads. In Fig.3.10, iso-area analysis of read-write power and EDP across workloads is presented. The comparison of read/write power in an iso-area configuration in Fig.3.10a shows that an 8MB SOT-MRAM cache is more power-efficient than a 4MB SRAM cache. SOT-MRAM demonstrates lower power consumption in read and write operations across various application workloads. It significantly reduces read power, staying below 256mW compared to SRAM's 1024mW in certain workloads. Additionally, SOT-MRAM exhibits significantly reduced write power consumption, frequently exceeding 50% savings across all workloads. Overall, SOT-MRAM outperforms SRAM in reducing read/write power due to its lower leakage power and efficient switching characteristics, making it an attractive option for power-sensitive applications requiring high density and energy efficiency.

A comparison of the iso-area EDP(Fig.3.10b,3.10c) between 4MB SRAM and 8MB SOT-MRAM across different application workloads reveals significant performance enhancements for SOT-MRAM in both read and write operations. When it comes to read-intensive tasks, SOT-MRAM exhibits notable reductions in EDP, with an average read-EDP reduction of approximately 80% compared to SRAM. Similarly, for write-intensive tasks, SOT-MRAM maintains its advantage with an average Write-EDP reduction of around 80%. These improvements can be attributed to the efficient switching characteristics and lower power consumption of SOT-MRAM cells, making it a superior choice for modern, high-performance memory applications.

In conclusion, substituting 4MB SRAM with 8MB SOT-MRAM results in performance enhancements and efficiency across various contemporary application workloads. SOT-MRAM reduces total power consumption by approximately 45% and exhibits lower read and write latencies, with reductions of around 70% and 65%, respectively, compared to SRAM. Furthermore, SOT-MRAM demonstrates improved bandwidth utilization, particularly in read operations, with an increase of approximately 30%. The EDP metrics show an average reduction of around 80% for both read and write operations, underscoring the potential of SOT-MRAM to replace SRAM in

high-capacity cache designs for modern high-performance computing applications.

Density replacement of 16MB SOT-MRAM in place of 4MB SRAM

Upon rigorous analysis comparing the 4MB SRAM with the 16MB SOT-MRAM, it is advised against transitioning to the 16MB SOT-MRAM for high-density memory applications. This recommendation is based on the marginal performance enhancements, characterized by only an 18% improvement in read/write latency and a 20% to 30% enhancement in Energy-Delay Product (EDP). Furthermore, the SOT-MRAM demonstrates higher read power consumption than SRAM. When 4MB SRAM is replaced by 16MB SOT-MRAM, the area advantage of SOT-MRAM diminishes, thereby negating the area advantage. Alternatively, an 8MB SOT-MRAM configuration is recommended instead of a 4MB SRAM. It presents a balanced blend of performance and efficiency, positioning it as a more suitable replacement for the 4MB SRAM.

3.6 Summary

This chapter comprehensively evaluates the performance enhancements and efficiency gains offered by SOT-MRAM over SRAM. The analysis uses the parameters of power consumption, latency, area efficiency, and density replacement studies. SOT-MRAM outperforms SRAM in HP/LOP devices, significantly reducing power consumption with acceptable performance loss.LOP-SOT-MRAM is a good choice for energy-sensitive systems, but HP-SOT-MRAM is also a good choice wherever HP/LOP-SRAM is used. Iso-capacity evaluation over AI,NLP,SNA, and generic computing applications show a 60% average reduction in power consumption and a 75% reduction in latency for 1MB SOT-MRAM. Similarly, in iso-area comparisons, replacing 4MB of SRAM with 8MB of SOT-MRAM retains the same physical chip area and *halves* the power consumption, leading to a 45% overall reduction. This configuration also demonstrates a significant decrease in both read and write latencies by approximately 65-70%. These findings underscore the potential of SOT-MRAM to substantially enhance performance and energy efficiency within existing chip real estate, supporting more robust and capable modern computing environments. Smaller technology nodes (from 65 to 22nm) effectively reduce chip area while managing power consumption and write latency. The

12/1 area/aspect ratio at 45nm for density replacement studies optimizes the balance between area, power, and latency. Overall, the analysis underscores the potential of SOT-MRAM to enhance performance and energy efficiency within existing chip real estate, positioning it as a potential technology for future high-density, high-performance memory architectures.

Chapter 4

PSC-VRO for LLC LI

4.1 Introduction

The relentless advancement in computing technologies continuously demands enhancements in memory architectures, especially at the LLC level, which is critical for overall system performance and efficiency. NVM technologies, particularly SOT-MRAM, have emerged as promising candidates to meet these demands due to their superior density and energy efficiency than prevalent SRAM caches. However, the integration of NVM into LLC poses significant challenges, primarily related to write endurance and the lifetime of cache due to write operations. These challenges necessitate innovative approaches to cache design that can leverage the benefits of NVM while mitigating its drawbacks.

4.2 Motivation, Parameters and Background

Fig.4.1 shows maximum and average write counts of various workload mixes from Table-4.3. The graph shows considerable variance in write operations, leading to uneven wear in cache memory cells, particularly in SOT-MRAM. This variability results in the degradation of memory cell endurance over time, causing reduced cache lifetime and impacting overall system reliability. The variation in writes results from uneven write distribution across the cache sets. This uneven distribution of cache writes leads to accelerated degradation(wear out) of heavily written sets compared to less frequently accessed ones. The term $Write_{avg}$ represents the average number of write

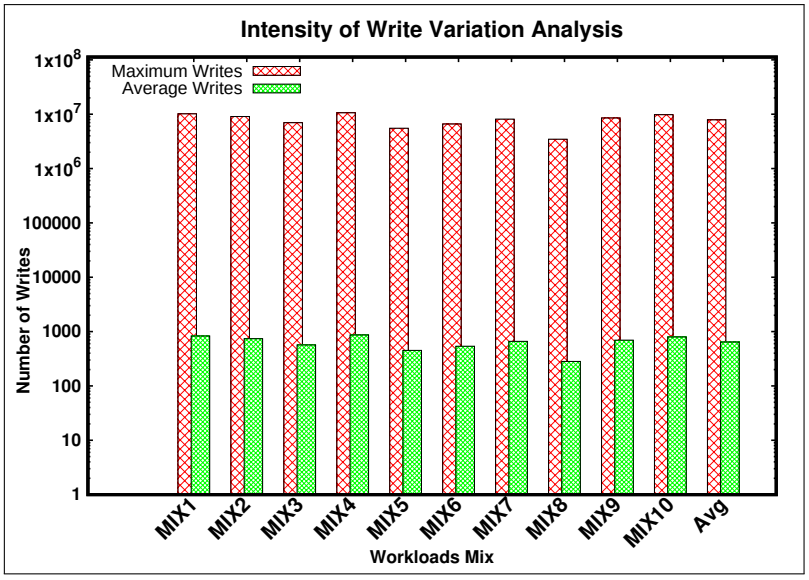


Figure 4.1: Analysis of Intensity of Write variation

Table 4.1: Parameters used in the equations

Parameter	**Meaning**
Aso	Cache associativity
Set	Number of cache sets
$Write_{avg}$	Average write count
$W_{k,l}$	Number of writes in cache set k and way l

operations per block in a cache bank. The write average is determined by summing all the write operations across the blocks and dividing this total by the number of blocks. This measure is crucial for understanding the overall write activity in the cache. It is used to analyze the write variations within individual sets and across different sets.

$$Write_{avg} = \frac{\sum_{i=1}^{Aso}\sum_{j=1}^{Set} w_{i,j}}{Aso \cdot Set} \tag{1}$$

The inter-set write variation coefficient(interv) is the variation in write counts across different cache sets. It measures uneven write distribution among the sets, with higher values indicating significant WVAR. The cache performance and lifespan depend on interv, as it aids in identifying and addressing hotspots that could cause premature wear-out of heavily written sets.

$$interv = \frac{1}{Write_{avg}} \sqrt{\frac{\sum_{k=1}^{Set} (\sum_{l=1}^{Aso} \frac{W_{k,l}}{Aso} - Write_{avg})^2}{Set - 1}} \tag{2}$$

The coefficient of intra-set write variation(IntraV) is the variation in write counts across blocks within a cache set. It measures uneven write distribution among the blocks in a set. The higher the values of *intrav*, the higher the disparity in write counts between blocks, resulting in certain blocks wearing out faster than others.

$$IntraV = \frac{1}{Set.Write_{avg}} \sum_{k=1}^{Set} \sqrt{\frac{\sum_{l=1}^{Aso} (W_{k,l} - \sum_{m=1}^{Aso} \frac{W_{k,m}}{Aso})^2}{Aso - 1}} \tag{3}$$

The Table 4.1 lists the variables used in the equations.

Cache lifetime can be categorized into two distinct types: raw and error-tolerant. Raw lifetime is characterized by the first instance of failure in a cache line. Error-tolerant lifetime accounts for the raw lifetime and the employed error recovery techniques(Wang et al., 2013). This work addresses the raw lifetime.

The lifetime of a cache block is the inverse of its maximum write count. $\forall$ in LI represents the individual write count for each cache block(Mittal and Vetter, 2014b).

$$LI = \frac{1}{\forall_{k=1}^{Set} \forall_{l=1}^{Aso} \max(W_{k,l})} \tag{4}$$

In terms of write variations, the lifetime is influenced by three important factors: intra-set write variation (IntraV), inter-set write variation (interv), and average write count in a cache bank $Write_{avg}$(Wang et al., 2013).

$$LI = \frac{Write_{avg\_base} \cdot (1 + interv_{base} + IntraV_{base})}{Write_{avg\_pt} \cdot (1 + interv_{pt} + IntraV_{pt})} - 1 \tag{5}$$

$Write_{avg}$ represents a cache bank's average number of writes. The subscripts "base- and "pt"in each term represent the values for the baseline and the proposed technique, respectively. PSC can segment the cache into different endurance zones; SRAM is considered high endurance, and the SOT-MRAM part is considered low endurance. Allowing frequently written data to be placed in the high-endurance sector(SRAM). This ensures that write operations are not concentrated in certain regions, thus mitigating premature wear-out and extending the lifetime of the cache.

VRO within the PSC framework can dynamically adjust the data placement strategy based on write patterns, virtually reordering the cache's logical view to balance the write load evenly across all cells. VRO can significantly reduce inter-set variance (differences in write activity among different sets) and intra-set variance (differences within the same set).

4.3 Design of PSC Architecture with Virtual Re-Ordering(VRO)

PSC presents a novel architectural solution designed to enhance the functionality of LLCs with NVM technologies. This design aims to segregate the cache into distinct regions physically, each optimized for different usage patterns and access frequencies. Such a configuration reduces write variation across the cache where certain regions endure fewer writes than others, thereby significantly improving the overall endurance and lifetime of the cache memory. Implementing PSC in LLCs involves balancing the trade-offs between access latency, power efficiency, and endurance. PSC ensures that high-frequency write areas are isolated and specifically tailored to handle higher endurance, sustaining performance over extended periods and reducing the overall power consumption. This architectural design depicts(Fig.4.2) a multi-core processor system incorporating hierarchical cache structures using SRAM and SOT-MRAM technologies to optimize cache management and reliability. Each processing core (Core_1, Core_2, Core_3, ..., Core_n) is equipped with a private L1 cache, while larger L2 caches serve as an intermediary between the fast L1 caches and the main memory. The PSC is the LLC. The PSC cache controller dynamically manages cache configurations to enhance performance and extend cache lifespan, utilizing components such as the Trace Reader, Config/Disabling Manager, and the proposed VRO LI algorithm. The PSC is divided into tag and data arrays, with the data array further segmented into *'M'* SOT-MRAM and *'N'* SRAM ways, combining the speed of SRAM with the non-volatility and energy efficiency of SOT-MRAM. Trace Reader reads memory access patterns to monitor the cache usage patterns. Config / Disabling Manager configures and potentially disables parts of the cache based on the workload requirements and the fault map. The Tag Array stores the metadata (tags) for the cache lines, indicating the address of

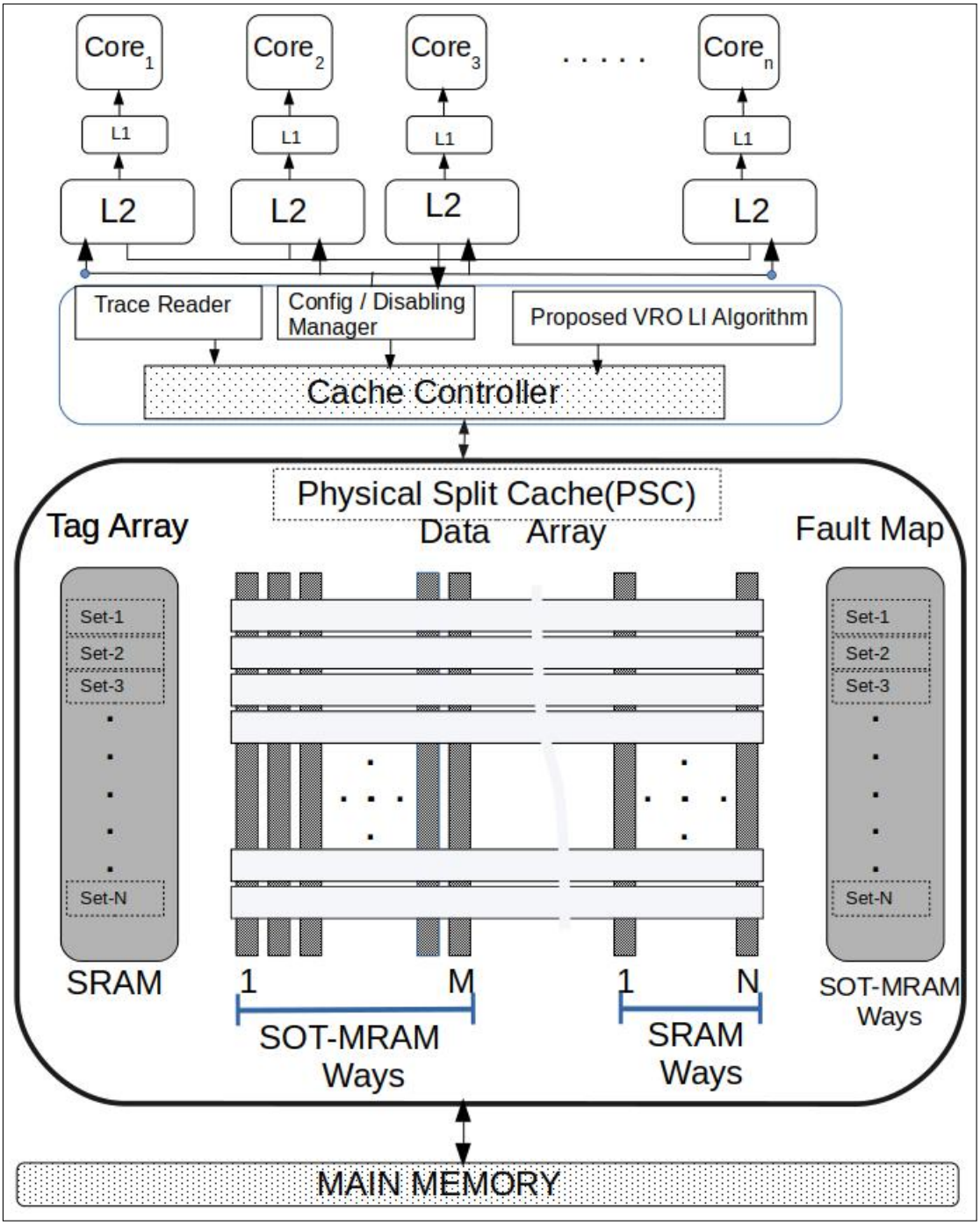


Figure 4.2: PSC Architecture with VRO

the data stored in the data array. It is typically implemented with SRAM for speed. Main memory, typically DRAM, interacts with the PSC to manage data misses in the LLC. This architecture leverages the dynamic reordering capabilities of the VRO LI algorithm to maintain write variation(WVAR) across SOT-MRAM ways, creating a stable and efficient cache system and ensuring a balance between performance, energy efficiency, and reliability.VRO further segments SOT-MRAM ways into separate read

and write ways. It divides a group of *M*-way LLC into *P*-ways for reading and *(M-P)* for writing.

Integrating a Design Space Exploration (DSE) tool with PSC facilitates swift and efficient iterations through various design configurations, significantly expediting the cache design cycle. The DSE tool capitalizes on the strengths of HyCSim(Escuin et al., 2022), a trace-driven simulator, to conduct rapid comparisons and optimizations of diverse hybrid LLC configurations. By leveraging HyCSim's capability to swiftly simulate and assess the impact of different insertion and replacement policies. This study reaches optimal SOT-MRAM cache configurations.

Our approach incorporates one global counter to reduce hardware overhead(Mittal and Vetter, 2014b). This singular counter simplifies the wear-levelling process across the cache, unlike other methods that may rely on multiple counters, thus leading to a leaner, more efficient hardware design. By simplifying the wear-levelling mechanism, our approach minimizes the additional area and power requirements typically associated with NVM cache designs, making it a more viable solution for modern applications.

Utilizing DESTINY(Mittal et al., 2017) for micro-architectural performance exploration, our approach provides a detailed and accurate assessment of various cache configurations. The synergy between HyCSim and DESTINY enables designers to explore various potential designs and validate their performance implications with high fidelity(Escuin et al., 2022).

4.3.1 Design of Cache VRO for WVAR and LI

Motivated by the use of instruction and data ways in (Sivakumar and Jose, 2023), we have used a similar approach and modified suitably to adapt it to the PSC architecture. The major goals of VRO are:

1. Distribute writes in SOT-MRAM.
2. Reduce SOT-MRAM writes
3. Balance LLC hits
4. Improve LLC Read hits

Algorithm 3 initializes and continuously processes cache requests(lines 1 to 14), distinguishing between read and write requests and handling each according to the cache's state. It migrates blocks between SRAM and SOT-MRAM as necessary(lines 4 to 5). The process involves dynamically adjusting the *sram_threshold* and utilizing *PSC-VRO* to optimize cache operations(lines 10 to 12).*PSC-VRO* is designed for lifetime extension of the SOT-MRAM part of the PSC hybrid architecture through VRO.

Algorithm 3: Proposed PSC-VRO

input : LLC cache-request, cache block, SRAM-set-threshold
output: Decision on read/write operation

```
1  repeat
2      for each LLC cache request do
3          if Read request and block in PSC-LLC then
4              if block is Read-loop block and in SRAM then
5                  Migrate to read ways of SOT-MRAM-cache-part;
6              Perform regular read operation;
7          else if Write request and block in SRAM then
8              if Write < SRAM-set-threshold then
9                  Write to SRAM;
10                 Update SRAM-set-threshold;
11         else if Read or Write request and block in SOT-MRAM part then
12             Call Dynamic-VRO() ;
13                                                  ▷ Algorithm ??
14         else
15             Regular Miss Operation ;
16         endif
17     end
18 until until end of requests;
```

Algorithm 4 of the PSC architecture initializes various parameters(lines 1 to 11) and functions essential for the Dynamic-VRO algorithm(lines 1 to 26). The function *exchange()* iterates through the cached read ways, swapping each read way with the corresponding write way to balance the cache's read/write operations(lines 1 to 19). This process allows frequently accessed data to be dynamically moved between read and write ways, reducing WVAR performance. The function *set_reorder()* recalculates(lines 20 to 26) the starting points for read and write ways after each reordering process. This ensures that the cache dynamically adapts to varying read/write patterns, optimising performance and extending the cache's lifetime. The algorithm also

includes regular read and write operations, incrementing or decrementing SOT_write_count based on the type of operation performed.

Algorithm 3 takes the type of request and block as input. Its results are read/write operations and decisions about reordering NVM ways in PSC. The algorithm dynamically manages read and write operations in the SOT-MRAM part of the PSC architecture by adjusting the partition ratio based on the SOT_write_count(lines 1 to 29). It begins by checking if the SOT_write_count meets specific thresholds, and if so, it calls the *exchange()* and *set_reorder()* functions to reorder the cache ways(lines-6 and 7). Suppose the SOT_write_count is less than or equal to zero(line 8), indicating fewer writes than expected. In that case, the algorithm increments the number of read ways using *inc_read_way()* after verifying that the *read_ways_count* is greater than the *write_ways_count*, then enforcing other conditions(lines 8 to 15). Conversely, if the SOT_write_count exceeds a maximum threshold T, it calls *dec_read_way()* to increment the write ways(lines 16 to 23). The functioning of VRO is optimised for SOT-MRAM LLC segment LI with WVAR distribution by relaxing the write restriction. Re-ordering the partition ratio using the global counter SOT_write_count effectively reduces the overhead of estimating heavy writes at the way level(Mittal and Vetter, 2014b). VRO in PSC architecture dynamically adjusts its read-to-write partition ratio from at least one read and remaining writes(1:11) to equal read and write partition(6:6) based on the observed write patterns to NVM. This flexibility enables PSC to manage a higher volume of data writes by increasing the number of write blocks.

The VRO utilises a flexible management system for partition ratios based on the application *"SOT_write_count"*. This counter increases with each write and decreases with each read in the LLC(Sivakumar and Jose, 2023). Depending on the updated value of SOT_write_count, PSC-VRO decides whether the LLC access primarily deals with read blocks or write blocks. Initially, the read-to-write way ratio is fixed at 8:4, and the re-ordering interval can be either a fixed value, a set of predetermined values, a counter, or a combination of the two. This work uses the *roi* counter. At the end of each reordering interval, the SOT_write_count is checked to ensure it stays between 0 and 1,023. The usual writing function continued without changing the split ratio if it remained stable within this range. If it isn't stable, the way reordering is initiated.

Algorithm 4: Generalized Dynamic Reordering of Cache Ways

Input : Cache associativity, initial read and write ways, starting indices for read and write ways

Output: Updated sequence of cache ways for reads and writes

```
Initialize the count of read and write ways
Set initial positions for read and write cache ways
Define a counter for write operations and set it to zero
Specify a threshold for the number of write operations
while system is operational do
    if there is a need to adjust the configuration then
        Swap the designated roles of selected read and write ways
        Update the starting points for read and write ways based on recent changes
        if write operations exceed the set threshold then
            Reallocate ways from writes to reads to balance the workload
            Reset the write operation counter
        endif
        else if read operations demand more resources then
            Shift resources from reads to writes to improve efficiency
            Maintain an updated count of operations to guide further adjustments
        endif
    endif
    Maintain a list of read and write ways, adjusting as necessary to optimize cache performance
end
```

Algorithm 5: Dynamic Virtual Reordering in Cache Systems

Input : Cache requests redirected from primary control algorithm, type of cache block

Output: Decision on read/write operations and adjustment of cache ways

```
1  while cache requests continue do
2      if request is for reading then
3          Execute standard read operation
4          Decrease write operation counter
5      else
6          if write operation counter is beyond threshold then
7              Exchange read and write cache ways
8              Reorder cache ways for optimal performance
9              Perform standard write operation
10             if write operation counter is too low then
11                 Increment read focus
12                 Reset counter adjustment parameters
13         else if write operation counter is very high then
14             Decrement read focus
15             Reset counter adjustment parameters
16         else
17             Perform standard write operation
18             Adjust write operation counter based on the operation type
19         endif
20     endif
21 end
22 Function IncreaseReadWay():
23     Decrease the number of write ways
24     Increase the number of read ways
25     Reset the write operation counter
26 Function DecreaseReadWay():
27     Increase the number of write ways
28     Decrease the number of read ways
29     Reset the write operation counter
```

A SOT_write_count of '0' indicates that there have been more reads to read blocks than expected. This triggers a flag. Suppose this flagging occurs for *'roi'* consecutive reordering intervals. The number of ways for reading increases by converting a writing way into a reading way to handle more read blocks in the subsequent reordering intervals. Conversely, when the SOT_write_count reaches 1,023, a reading way is changed into a writing way to accommodate more write blocks, following the same process for 'roi' consecutive reordering intervals. After reordering, the cache operates with the new partition ratio, which is repeated periodically. The requirement of *'roi'* consecutive reordering intervals is adhered to for changing the partition ratio, as the 8:4 ratio is considered the most stable configuration. Any deviation from this ratio should only occur if an application consistently indicates the need for a different partition. At the end of each reordering interval, the SOT_write_count is reset. However, if the current partition ratio deviates from 8:4, reordering is triggered by either SOT_write_count is '0', or SOT_write_count is '1,023' without the need for *'roi'* consecutive intervals. This ensures a preference for maintaining the 8:4 partition ratio. PSC is configured to always reserve at least one read way.

Implementing the PSC and integrating it into the proposed framework, coupled with dynamic virtual reordering and DSE tools, presents a substantial leap in designing hybrid caches. It allows for a more streamlined and effortless design process, reduces complexity, and ensures a more cost-effective cache solution in less time.

4.4 Evaluation

4.4.1 Experimental Setup

The simulation tools, setup, and workloads involved in the PSC-VRO are described here. Table 4.2 and 4.3 list all the tools and detailed parameters used in the work. The memory array parameters are employed in density replacement to estimate the overall effect of the SOT-MRAM LLC by considering the application workload. The proposed PSC-VRO uses Table 4.3 workload mix to evaluate the impact of the cache management policy on the hybrid LLC. The framework integrates DESTINY(Mittal et al., 2017) and HyCSim(Escuin et al., 2022) for the study. After establishing the micro-architectural feasibility, the LLC memory traces are generated. The traces were

collected from gem5(Binkert et al., 2011). Table 4.3 presents the SPEC CPU 2006 and 2017 mix. These traces were then utilized for LLC DSE and evaluation of the VRO WVAR algorithm. HyCSim(Escuin et al., 2022) is modified to include the proposed PSC-VRO, enhancing its capability to simulate the WVAR behaviour of LLC configurations.

Feature	**SRAM**	**SOT-MRAM**
Cell area	146 F^2	12 F^2
Aspect Ratio	1.46 F	1 F
Associativity	4	12
Write pulse	–	$< 0.5ns$
Timing and current	Standard model	As per compact model selection
Temperature	350K	
Capacity Range	1MB	
Device type	HP	
Technology node	45nm	
Simulators	Modified DESTINY, gem5 and HyCSIM	

Table 4.2: Parameters used for PSC power and performance experiments

In our experimental setup, the initial phase involved utilizing the DESTINY tool to perform a detailed DSE of the PSC configurations. DESTINY provided micro-architectural performance results such as power, and timing estimates for 1MB to 32MB PSC design. In this section we only present 1MB results. This phase determines the parameters for PSC and ensures that the proposed cache configurations are feasible under micro-architectural constraints.

Application workloads are outlined in Table 3.2(Leskovec and Krevl, 2014; Lan et al., 2019; SPEC, SPEC; Pentecost et al., 2022; Inci et al., 2022). These workloads cover a range of application domains, including graph processing and social networks (Facebook and Wikipedia), natural language processing (ALBERT), image processing (ResNet), and general computing (SPEC2017). Each workload has multiple instances: three for Facebook/Wikipedia, ALBERT, and ResNet, and fourteen for SPEC2017. This variety ensures a comprehensive evaluation across different application types, yielding realistic and pertinent performance measurements.

Table 4.3: Mixes of SPEC CPU 2006 and 2017(SPEC, SPEC; Escuin et al., 2023).

Mix	**Applications**
Mix 1	bzip206 ,gobmk06,zeusmp06,dealII06
Mix 2	bzip206,roms17, wrf06 , hmmer06
Mix 3	soplex06, hmmer06,cactuBSSN17 zeusmp06
Mix 4	astar06,omnetpp06, milc06, libquantum06
Mix 5	leslie3d06,mcf17,xalancbmk06, bwaves17
Mix 6	xz17,wrf06, GemsFDTD06,lbm17
Mix 7	dealII06,xalancbmk06, libquantum06 ,cactuBSSN17
Mix 8	mct17, milc06 ,lbm17, gobmk06
Mix 9	astar06,soplex06, bwaves17 xz17
Mix 10	leslie3d06,omnetpp06, roms17 ,GemsFDTD06

4.5 Results and Analysis of PSC-VRO WVAR and LI

The hybrid design provides a better balance between density, power performance, and LLC lifetime extension. First, this section does the micro-architectural performance DSE, including power and timing estimates for various PSC designs, ranging from a smaller scale of 1MB to a larger cache size of 32 MB. This phase was critical in determining the foundational parameters for PSC and ensuring that the proposed cache configurations were feasible regarding micro-architectural constraints. This section only presents the results for the 1MB LLC design.

4.5.1 PSC Micro-architecture results

The LLC cache parameters presented in Table-4.4 were obtained using DESTINY(Mittal et al., 2017), assuming sequential cache access, a 45nm process, and a cache design optimized for minimizing the energy-delay product for write operations. In the case of a hybrid cache, it is assumed that the energy and delay for cache misses resemble those of a SOT-MRAM cache. The energy and delay for cache hits and writes vary based on whether the operation involves reading from an SRAM or a SOT-MRAM part of the PSC. Additionally, the leakage power is presumed to increase linearly following

the number of ways designed using SRAM and SOT-MRAM. This section analyses the results for 1MB LLC design for PSC cache power-performance results listed in Table-4.4. The comparison in Table-4.4 demonstrates the efficiency of SOT-MRAM over SRAM in power and performance parameters for caches. SOT-MRAM has a lower read latency of 0.42 ns, 16% less than SRAM's 0.5 ns, and a write latency of 0.35 ns, 20% lower than SRAM's 0.44 ns. These reductions lead to faster data access and improved system performance.

Regarding energy consumption, SOT-MRAM shows substantial improvements. Its read energy is 1.1 nJ, approximately 45% less than SRAM's 1.99 nJ, and its write energy is 0.34 nJ, 33% lower than SRAM's 0.51 nJ. These reductions enhance the energy efficiency of memory operations, making SOT-MRAM a more sustainable option for large-scale cache implementations. These results improve as the cache size increases.SOT-MRAM's leakage power is significantly lower than SRAM's, measuring 970 mW compared to SRAM's 1893.86 mW, marking a nearly 50% reduction. This reduction is important for decreasing the overall power consumption of the system, especially in standby modes.

The Hybrid PSC architecture demonstrates superior performance across various parameters, including read and write latency and energy efficiency. It offers a balanced trade-off between performance, energy consumption, and chip area(50% PSC), making it a promising choice for memory-intensive applications.

Table 4.4: Power and Performance

Parameters	SRAM	SOT-MRAM
Cache Read Latency (ns)	0.5	0.42
Cache Write Latency (ns)	0.44	0.35
Cache Read Energy(nJ)	1.99	1.1
Cache Write Energy(nJ)	0.51	0.34
Leakage(mW)	1893.86	970

4.5.2 LI extension with WVAR results

Analyzing the intra-set write variation results(Fig.4.3) for different workload mixes, it's evident that the PSC-VRO method yields significant improvements over the baseline and other strategies (SA-1 and SA-2 from (Agarwal and Kapoor, 2020)). The intra-set

write variation(Fig.4.3) is instrumental in evaluating the wear-levelling effectiveness of cache memory; a lower value indicates a more uniform distribution of write operations, which is desirable for prolonging NVM lifetime.

The intra-set write variation(IntraV) analysis across different workloads (Fig.4.3) reveals significant differences between the baseline, SA-1, SA-2, and PSC-VRO. In the MIX1 workload, the baseline approach demonstrated an IntraV of 6.67%, while the PSC-VRO approach achieved a reduction to 3.84%, representing a 42.4% improvement. In the MIX2 workload, the baseline showed an IntraV of 124.91%, reduced to 27.78% with the PSC-VRO method, marking a significant 77.8% reduction. These trends are consistent across other workloads, indicating the effectiveness of the PSC-VRO approach in minimising intra-set write variation.

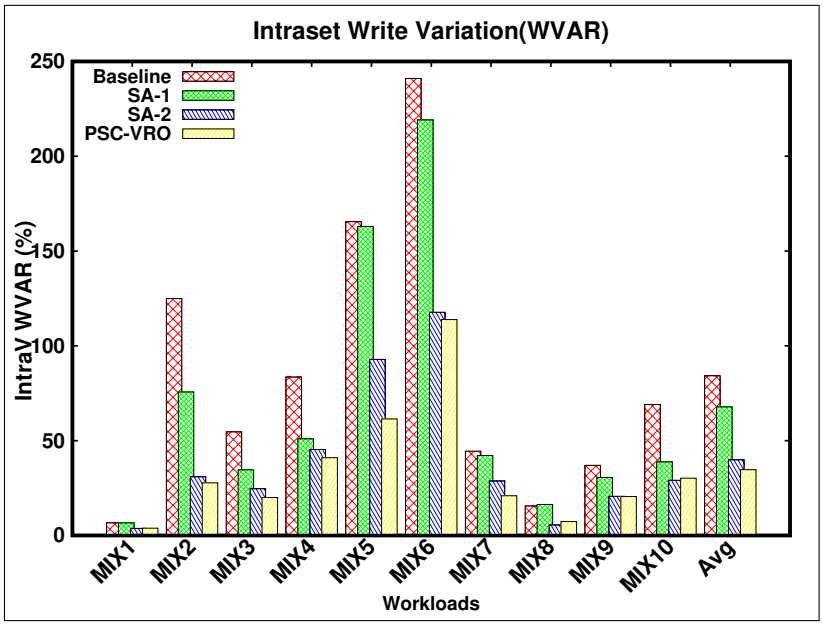


Figure 4.3: Analysis of Intra Set Write Variation

For instance, in the MIX3 workload, the *intrav* decreased from 54.63% in the baseline to 20.01% with the PSC-VRO method, signifying a 63.4% reduction. In the more challenging MIX6 workload, the baseline *intrav* of 241.04% was reduced to 113.86% with the PSC-VRO method, representing a significant 52.7% improvement. This consistent performance enhancement is vital for enhancing the reliability and longevity of memory systems, as lower write variation ensures more uniform wear across the memory cells, reducing the likelihood of premature failures and enhancing overall system stability.

The average *intrav* values across all workloads further highlight the effectiveness of the PSC-VRO approach. The baseline methodology resulted in an average *intrav* of 84.24%, whereas the PSC-VRO method significantly decreased this to 34.70%, marking a 58.8% reduction.PSC-VRO consistent reduction in interv continued with 48% and 13% than SA-1 and SA-2, respectively. This consistent reduction in *intrav* underscores the potential of the PSC-VRO approach to deliver substantial improvements in LI, especially for modern HPC applications where consistent and reliable memory performance is essential. By effectively distributing writes and minimising hotspots, the proposed PSC-VRO mitigates the adverse effects of uneven write patterns, making it a valuable strategy for future memory architecture designs. Such results strongly advocate implementing the PSC-VRO in LLCs to achieve balanced write distribution and enhance cache memory lifetime.

The inter-set write variation results (Fig.-4.3) from the various workload mixes provide valuable insights into the efficacy of the proposed cache management strategy compared to the baseline and other schemes (SA-1 and SA-2(Agarwal and Kapoor, 2020)). Inter-set write variation(Fig.4.4) measures the uniformity of write distribution across different cache sets, and lower values indicate a more even write spread, which is beneficial for the overall endurance of the cache.

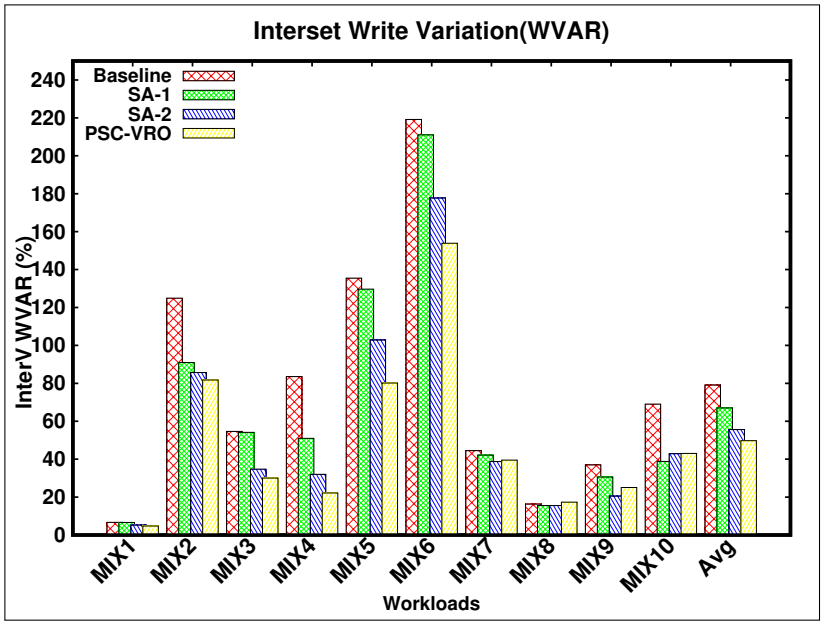


Figure 4.4: Analysis of Inter Set Write Variation

The inter-set write variation (interv) analysis has shown significant differences

across workloads across the baseline, SA-1, SA-2, and PSC-VRO. In the MIX1 workload, the baseline approach showed an *interv* of 6.67%, while the PSC-VRO method achieved a slight reduction to 4.75%, marking a 28.8% improvement. Similarly, in the MIX2 workload, the baseline exhibited a high *interv* of 124.91%, significantly reduced to 81.77% with the Proposed PSC-VRO, indicating a 34.53% decrease. These trends are consistent across other workloads, demonstrating the effectiveness of the PSC-VRO approach in minimising inter-set write variation.

For instance, in the MIX3 workload, the *interv* decreased from 54.63% in the baseline to 30.05% with the PSC-VRO method, showing a 45% reduction. In the more challenging MIX6 workload, where the baseline interv was considerably high at 219.19%, the PSC-VRO method reduced this to 153.86%, representing a 29.80% improvement. This consistent performance improvement is crucial for enhancing the reliability and longevity of memory systems, as lower write variation ensures more uniform wear across the memory cells, reducing the likelihood of premature failures and improving overall system stability.

The average *interv* values across all workloads further highlight the efficacy of the PSC-VRO approach. The baseline methodology resulted in an average *interv* of 79.12%, while the PSC-VRO method lowered this to 49.76%, reflecting a 37.10% reduction. PSC-VRO consistent reduction in interv continued with 25.78% and 10.50% than SA-1 and SA-2, respectively. Compared to 58.8% *intrav* reduction, a mere 37.10% interv reduction underscores the potential of the PSC-VRO approach to deliver substantial improvements in LI despite higher inter-set write variation. The proposed methodology mitigates the adverse effects of uneven write patterns of intra-set writes distributing and minimising hotspots. Due to *intrav* minimization indirectly, *interv* is minimized to some degree. This makes PSC-VRO a valuable strategy for LLC architecture designs.

In summary, the proposed cache management strategy significantly enhances the uniformity of write distribution across cache sets. This improvement is crucial in the NVM part of the PSC caches, where the balance of write operations can directly impact the longevity and reliability of the memory. These results validate the effectiveness of the proposed PSC and dynamic virtual reordering strategy in reducing inter-set write variation, contributing to extended cache endurance and potentially greater overall

system performance.

The Relative LI(RLI) analysis in Fig.4.5 for the PSC-VRO compared to the baseline and other state-of-the-art methods are significantly higher across all workload mixes. A higher percentage indicates better performance in terms of extending the lifetime of the cache.

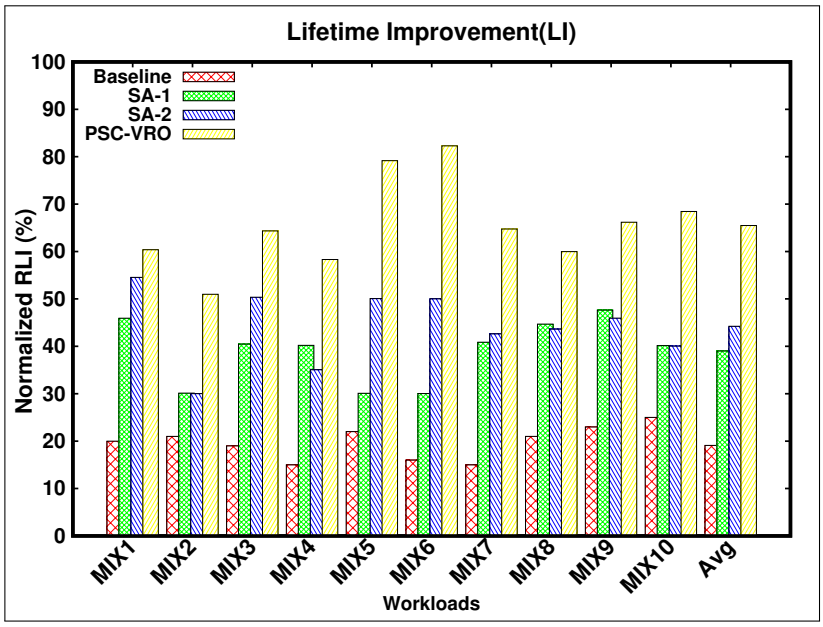


Figure 4.5: Analysis of Relative Lifetime Improvement

The following analysis of each workload is as follows:

- **MIX1:** The PSC-VRO method shows a significant improvement in lifetime, achieving 60.38%, which is a 10% increase compared to SA-2 and a 25.10% increase compared to SA-1.

- **MIX2:** The PSC-VRO method demonstrates a notable enhancement with a 50.97% improvement, which is 41.14% more LI improvement than both SA-1 and SA-2, which are around 30.12% and 30.02%, respectively.

- **MIX3:** The PSC-VRO approach leads with a 64.35% improvement, far exceeding SA-1 (40.48%) and SA-2 (50.36%). This reflects a 31.88% and 21.87% increase compared to SA-1 and SA-2, respectively.

- **MIX4:** Here, the PSC-VRO method improves lifetime by 58.31%, compared to 40.20% for SA-1 and 35.06% for SA-2, showing a 23.21% and 18.13% increase over SA-1 and SA-2.

- **MIX5:** The PSC-VRO method achieves a 79.16% improvement, surpassing SA-1's 30.09% and SA-2's 50.06% by 62.02% and 36.70%, respectively.
- **MIX6:** The PSC-VRO method marks a significant improvement with a 82.30% increase, significantly higher than SA-1's 30.03% and SA-2's 50.02%, showcasing an improvement of 63.41% and 39.02% respectively.
- **MIX7:** With a 64.75% improvement, the proposed method outperforms SA-1 (40.86%) and SA-2 (42.64%) by 23.89% and 22.11%.
- **MIX8:** The PSC-VRO method indicates a 59.99% improvement in lifetime, significantly higher than SA-1's 44.69% and SA-2's 43.64%, with differences of 15.31% and 16.35%.
- **MIX9:** The PSC-VRO method leads with a 66.17% improvement, greatly surpassing SA-1's 47.66% and SA-2's 45.95%, indicating improvements of 28.51% and 31.22%.
- **MIX10:** The PSC-VRO method achieves a 68.45% improvement, compared to 40.14% for SA-1 and 40.06% for SA-2, with a notable increase of 28.31% and 28.39%.
- **Average Analysis:** On average, the proposed PSC-VRO method results in a 65.48% improvement in relative lifetime, which is substantially higher than SA-1's 39.02% and SA-2's 44.24%. The PSC-VRO method consistently shows an average improvement of 40.43% over SA-1 and 32.80% over SA-2, highlighting its superior performance in enhancing the cache's relative lifetime across various workloads.

The improved lifetime reliability achieved by the PSC-VRO method can be attributed to its ability to optimize cache ways allocation and access strategies, reducing wear and prolonging the lifespan of LLC. The analysis demonstrates that the PSC-VRO method consistently outperforms SA-1 and SA-2 regarding Relative Lifetime Improvement (RLI).

The summary of the analysis is that PSC-VRO addresses directly intra-set and indirectly inter-set write variation, resulting in a 65.48% improvement in RLI. The

PSC-VRO ensures balanced wear levelling and minimizes wear-out of memory cells, enhancing memory reliability and prolonged lifespan. We conclude that the findings underscore the effectiveness of the PSC-VRO method in improving RLI and reliability, making it a promising solution for modern computing systems.

4.6 Summary

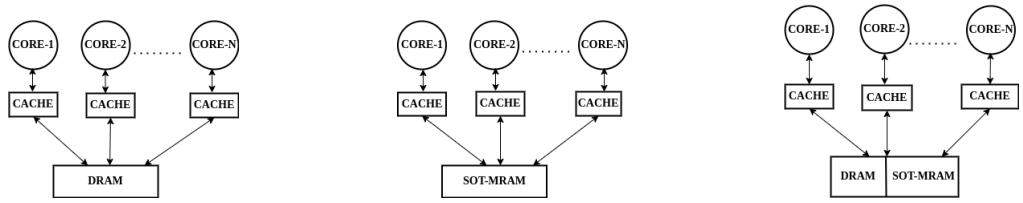
The results of our experiments show that the coefficient of WVAR can be a useful metric for prioritizing write minimization and distribution in the SOT-MRAM segment of PSC architecture. When developing cache management policies for hybrid caches, it is essential to consider the differential read-write operations of NVM. By taking into account that the number of writes affects endurance and reliability, we can significantly extend the overall lifetime of the system, which is crucial for modern computing systems. VRO, a cache WVAR policy, is specifically designed for SRAM-SOT-MRAM hybrid caches. VRO operates based on the type of operation that influences block read and write decisions for higher or lower endurance segments. Since SRAM writes do not impact the endurance of the written block, PSC-VRO manages the writes with SRAM blocks while controlling the costly NVM writes. VRO outperforms existing methods in terms of both intervention and LI. It is more efficient in reducing writes and enhancing LI, making PSC-VRO the top choice when lower energy consumption and improved performance with LI are important.

Chapter 5

SOT-MRAM as an alternative to DRAM in Main Memory for Embedded Systems and Multi-Core Environment

5.1 Introduction

As the current primary memory technology is reaching its limits, it is essential to explore alternative memory technologies to accommodate modern applications and use cases. However, using new memory technology poses the challenge of deriving accurately estimated parameters for designing and integrating new memory technology and performing reliable simulations. This chapter proposes a new approach integrating SOT-MRAM into hybrid and full main memory architectures for embedded and multi-core systems, encompassing various memory configurations and capacities. The work addresses the challenge of evaluating SOT-MRAM-based memory systems when specific SOT-MRAM memory parameters are not publicly available. The research methodology includes micro-architectural (circuit-level) design space exploration and comprehensive full system simulations, which evaluate benchmark programs representing diverse application domains. The evaluation includes three memory structures with varying memory organizations and capacities. Figure 5.1 depicts the different memory architectures used in this work.



(a) DRAM based memory (b) NVM based memory (c) Hybrid memory

Figure 5.1: Different memory architectures

5.2 SOT-MRAM Main Memory

To address the challenges and retain the benefits of STT-MRAM, researchers have investigated other memory technologies, such as SOT-MRAM. Recent studies have demonstrated that SOT-MRAM could serve as a promising alternative to SRAM, offering similar benefits while overcoming some of the limitations of STT-MRAM. F. Oboril et al. provided compelling evidence of SOT-MRAM as a viable alternative to SRAM in their pioneering work, as noted in (Oboril et al., 2015) for multi-core systems. The research has demonstrated that SOT-MRAM offers a 60% reduction in energy consumption, a 1% performance improvement, and an outstanding 27-fold reduction in retention failure probability. The results provided in their work make a strong case for using SOT-MRAM as a cache memory (Oboril et al., 2015).

The integration of SOT-MRAM with the CMOS manufacturing process is another crucial aspect that needs to be addressed. As demonstrated in (Zheng et al., 2021), SOT-MRAM has been successfully integrated into CMOS technology for cache replacement.

According to a study by Garello et al., (Garello et al., 2018), SOT-MRAM has shown superior power efficiency and performance compared to SRAM, which makes it a promising choice for cache applications. However, the larger bit cell area is still considered a major disadvantage.

To solve the problem of larger bit cell area, a potential solution has been presented in (Seo and Kwon, 2020b) with an area-optimized cell design that achieved a remarkable area optimization of 42%. Also, a similar pioneering area-optimized SOT-MRAM bit cell and memory organization was proposed in (Liu et al., 2023). The result shows an area reduction of 32%. However, the requirement of two access transistors per bit has made SOT memories area-intensive. Similar to the previous works, the problem of

area efficiency was solved by introducing a multiple-bit SOT-MRAM cell in (Mishra et al., 2021). A multi-bit SOT cell with a shared write channel among multiple bits is proposed to address this challenge, enabling an area-efficient memory design with improved device density. The authors conclude that SOT-MRAM may become a preferred choice for a wide range of memory hierarchies(Mishra et al., 2021) including main memory.

All these above works prove that SOT-MRAM can be a potential candidate for main memory. Previous studies have shown favourable outcomes for the main memory implemented using STT-MRAM, and SOT-MRAM is adopted as the cache memory technology in(Oboril et al., 2015). Since SOT-MRAM and STT-MRAM exhibit similarities, and STT-MRAM is employed as the primary memory, we have evaluated the effects of employing SOT-MRAM as the primary memory in a multi-core environment.

In the absence of reliable current and timing parameters from manufacturers, the proposed SOT-MRAM-based main memory for the multicore environment uses approaches demonstrated in (Asifuzzaman, 2019; Asifuzzaman et al., 2022; Oh et al., 2023). These methods use simulators and commercially available hardware to validate their timing and current parameter scaling methodology. The results strongly suggest that read and write operations have identical costs after the data is loaded into the row buffer regardless of the memory technology(Asifuzzaman et al., 2022).

Given the compelling evidence presented by prior research, this work explores the viability of SOT-MRAM as a main memory technology. The evaluation involves deriving and using separate scaled parameters for row buffer-associated and non-associated parameters. We estimated and scaled the timing and power parameters using established and validated methods outlined in previous studies such as (Asifuzzaman et al., 2022) and (Oh et al., 2023). Additionally, we conducted circuit-level memory exploration experiments to further validate the scaling factors employed. The details of this analysis can be found in the results section. The following section lists and provides details of the timing parameters.

Our decision to use DDR3-1600 for comparisons is based on the accessible and validated parameters specific to the DDR3 standard, as per their industry partner Everspin Technologies under a co-operation agreement (Asifuzzaman et al., 2022). However, the parameters used do not correspond to their commercial product param-

eters. Moreover, the DDR3-1600 parameters also apply to other DDRx standards, as highlighted in recent studies (Asifuzzaman, 2019; Liu et al., 2022). This approach ensures that our findings are relevant across various DDR technologies.

5.2.1 Parameters for Estimation of Timing

The DDRx protocol compatibility of SOT-MRAM memory and its similarity in organization and CPU interface to STT-MRAM provides valuable insights into the timing parameters of SOT-MRAM memory. In the case of DRAM and STT-MRAM primary memory devices, a row buffer serves as an interface between the memory bus and cell arrays (Asifuzzaman et al., 2022). The same holds true for SOT-MRAM. Only timing parameters associated with the row buffer differ between DRAM and SOT-MRAM. This is because, following loading a row of data into the buffer, the timing parameters for subsequent operations are identical for both types of memory(Asifuzzaman et al., 2019). The circuitry beyond the row buffer for DRAM and SOT-MRAM is essentially the same(Asifuzzaman, 2019). The timing parameter tCWD, which denotes the delay between issuing a column write command and placing the data on the bus, remains the same for DRAM and SOT-MRAM. Other timing parameters unrelated to row operations, such as tBURST, tCAS, and tWTR, are identical. Table-5.1 presents these timings, expressed in cycles of DDR3-1600.

The primary distinction between SOT-MRAM and DRAM's main memory is the technology used in their storage cells, namely MTJ and capacitor. The cell access mechanism differs between these two memory technologies, leading to differences in the timing parameters related to SOT-MRAM row operations compared to DRAM. The access operation of DRAM is voltage-based, whereas SOT-MRAM's access operation is current-based (Asifuzzaman et al., 2019).

The timing parameter related to *precharging* a bit line to a reference voltage before the cell access is *Row pre-charge* (tRP)(Asifuzzaman, 2019). In DRAM, to access a cell, a bit line is pre-charged, and then the word line is activated, enabling the sensing circuit to sense the data.SOT-MRAM cell array access is different from DRAM, as it uses a current operation mode to read data stored in MTJ by activating a word line and applying a small amount of current to sense the data through the bit-line(Asifuzzaman et al., 2019).tRCD is the timing parameter that represents the time required to access

Table 5.1: The SOT-MRAM timing parameters unrelated to row operation(scaled from DDR3-1600 in Cycles).

Timing Parameter	**Description**	**DRAM**	**SOT-MRAM**
tRTP	Read-to-pre-charge-delay	6	6
tCWD	Column-write-delay	10	10
tWTR	Write-to-read-delay-time	6	6
tCCD	Column-to-column-delay	4	4
tAL	Added latency-to-column-access	0	0
tRTRS	Rank-to-rank switching-time	1	1
tBURST	Burst-length	4	4
tCAS/tCL	Column-access-strobe-latency	11	11
tFAW	Four-row-activation-window	24	24
tCKE	Next-power-up for an idle device	4	4
tWR	Write-recovery-time	12	12
tCMD	Command-transport-duration	1	1
tXP	Exit-power-down with DLL on to any valid command	5	5

and retrieve the data from a row and has it ready in the row buffer.tRCD,tRP, and row to row activation delay(tRRD) are three values which differ between DRAM and SOT-MRAM.

The timing parameters specific to SOT-MRAM have not been standardized or publicly disclosed due to the constantly evolving nature of the technology. As a result, memory manufacturers working on STT-MRAM and SOT-MRAM are not disclosing these parameters. Therefore, sensitivity analysis is needed on the parameters that vary from DRAM to SOT-MRAM. In this study, a conservative scaling of SOT-MRAM parameters from DRAM-1600 is adopted using the methodology in (Asifuzzaman et al., 2022; Inci et al., 2022). The scaling methodology was validated in the work(Asifuzzaman et al., 2022) and (Oh et al., 2023). SOT-MRAM tRFC(Refresh cycle time) and tRAS(Activate to pre-charge delay) values are taken as 0. Parameters are listed in Table-5.2.

Unlike DRAM, SOT-MRAM access operation is *non-destructive*, so no row restoration is required. The next row access operations in SOT-MRAM are initiated sooner than in DRAM. The SOT-MRAM Row cycle (tRC) is shorter than DRAM in certain cases despite having a longer tRCD(Wang et al., 2014). In this work, we assume the SOT-MRAM operation to be symmetric.

Table 5.2: The SOT-MRAM timing parameters related to row operation(scaled from DDR3-1600(in Cycles)).

Timing Parameter	Description	DRAM	SOT-MRAM
tRCD	Row-to-column-command-delay	11	12
tRP	Row-pre-charge	11	12
tRRD(R/W)	Row-activation to Row activation delay	5	6
tRFC	Refresh-cycle-time	208	0
tRAS	Min. Row-active-time or Activate to pre-charge-delay	28	0

5.2.2 Parameters for Estimation of Power

The fundamental distinction between the two main memory technologies is the storage cell. The power parameters linked with the access of these cells vary between SOT-MRAM and DRAM. Regarding the power consumption of DRAM and SOT-MRAM, three parameters differ in the present models. These parameters are the Active pre-charge-Current (IDD0), Active Read pre-charge-Current (IDD1), and Operating-Burst-Current (IDD4(R/W))(Asifuzzaman et al., 2022). SOT-MRAM uses current mode to access its cells, unlike DRAM's voltage-mode cell operations. So, we choose the same methodology adopted in timing parameter estimation and scale the values to account for current-based sensing methods of SOT-MRAM as in (Asifuzzaman et al., 2022),(Oh et al., 2023).

Table 5.3: Current parameters(in mA) used in the study(scaled from DDR3-1600).

Current Parameter	Description	DRAM	SOT-MRAM
IDD0	**Active pre-charge Current**	**53**	**63**
IDD1	**Active Read-pre-charge Current**	**66**	**76**
IDD2P	pre-charge Power Down Exit Current	18	18
IDD2N	pre-charge Standby-Current	24	24
IDD3P	Active-Power-Down-Current	15	15
IDD3N	Active-Standby-Current	20	20
IDD4(R/W)	**Operating Burst-Current**	**90**	**104**
IDD5	Refresh-Current	152	0
IDD6	Self-Refresh-Current	15	0

In the case of SOT-MRAM, since it does not require refresh, the Refresh Current (IDD5) and Self Refresh Current (IDD6) are set to 0. However, the current parameters not associated with any operation accessing the cell(pre-charge and active, power down

and stand-by current remain unchanged from DRAM to SOT-MRAM.Table 5.3 lists in bold the row-related parameters, and the rest are the same for both memory types except refresh currents.

5.3 Evaluation

This section details the experimental setups utilized for system-level evaluations in embedded and multi-core systems, as well as the experimental setup for the main memory micro-architecture exploration using the MFS approach.

5.3.1 Experimental Set-up of Main Memory Micro-architecture DSE-MFS

We conducted a study on the latest device parameters for STT-MRAM and SOT-MRAM by analyzing the following works (Liao et al., 2020; Zhang et al., 2012; Wang et al., 2019; Yang et al., 2022; Wu et al., 2020b). We extracted and incorporated information from a reliable, compact STT-MRAM model (Zhang et al., 2012; Wang et al., 2019; Yang et al., 2022) to determine the cell file parameters. Similarly, we used parameters from sources that detail high-density, area-optimized, and performance-enhanced devices to determine the parameters for SOT-MRAM (Wu et al., 2020b; Wang et al., 2019). Additionally, we derived other important parameters, such as cell area, aspect ratio, set/reset current, read/write time, and access transistor width, from the same sources (Mittal et al., 2017; Liao et al., 2020; Zhang et al., 2012; Wang et al., 2019; Yang et al., 2022; Wu et al., 2020b). We have presented these parameters concisely in Table 5.4. For these experiments, we utilized a refined version of Algorithm-6, which we adapted and optimized based on the work in (Inci et al., 2022). To enhance the circuit-level performance of the main memory across different memory capacities, we integrated the DESTINY simulator (Mittal et al., 2017) into our approach to obtain the circuit-level evaluation. Algorithm 6 used for MFS-DSE of the main memory design tuning. The power, performance, and area results of primary memory significantly vary based on the optimization target chosen in DESTINY (Mittal et al., 2017). The optimal configuration for each memory technology is selected using Algorithm 6 to obtain the optimal results.

Table 5.4: Bit Cell Device Parameters(Mittal et al., 2017; Liao et al., 2020; Zhang et al., 2012; Wang et al., 2019; Yang et al., 2022; Wu et al., 2020b).

Parameter Description	STT	SOT
MTJ area	$40 \times 40\,\mathrm{nm}^2$	$40 \times 40\,\mathrm{nm}^2$
Heavy-Metal dimension	–	$40 \times 60 \times 2\,\mathrm{nm}^3$
Free layer thickness	1.3 nm	1 nm
Oxide layer height	0.85 nm	0.85 nm
Spin Hall angle	–	0.3
Magnetic anisotropy	1.3×10^5 A/m	1.33×10^5 A/m
Saturation Magnetization	1.58×10^5 A/m	1×10^6 A/m
Tunnel Magnetoresistance(TMR)	120 %	150 %
Heavy-Metal resistivity	–	$200\,\mu\Omega \cdot \mathrm{cm}$

Algorithm 6: MFS Main Memory Design Tuning Algorithm

Input: Memory M, Size S, Target Optimization *TO*
Output: Meet all the Optimal *TO*

$Mem \in M = \{SRAM, SOT-MRAM\}$;
$Size \in S = \{1, 2, 4, 8, 16, 32, 64, ...\}$;
$Cell\_Para \in CP$; `/* Area, Aspect ratio, Current, Voltage parameters, CMOSaccesslength */`
; $Opt\_Para \in OP$; `/* ReadLatency, WriteLatency, ReadEnergy,...Write Energy, Read EDP, Write EDP, Area, Leakage */`
; $Acc \in A = \{Normal, Fast, Sequential\}$;
$Dev \in D = \{HP, LOP\}$;
while $Mem \in M$ **do**
  while $Size \in S$ **do**
    while $Cell\_Para \in CP$ **do**
      $CirLev \leftarrow \infty$;
      while $Opt - Para \in OP$ **do**
        while $Acc \in A$ **do**
          while $Dev \in D$ **do**
            $CirLev \leftarrow Calculate(EDAP)$;
            if $CirLev^+ \geq CirLev$ **then**
              $CirLev^+ \leftarrow CirLev$;
            end
          end
        end
      end
      TunedResult.append(argv(*CirLev*));
    end
  end
end
return *CirLev*

5.3.2 Experimental Set-Up for Embedded Systems

Experiments were conducted using well-established simulators for the design and analysis of main memory. Initially, we conducted a comparative study of three different technologies, i.e., DRAM, STT-MRAM, and SOT-MRAM, to identify the best candidate regarding area efficiency, access latency, and energy. These results were obtained from the DESTINY (Mittal et al., 2017) simulator. Experiments in the DESTINY simulator were performed by setting the circuit-level parameters using references from the previous work. STT-MRAM parameters were chosen from (Zitong Zhang and Jiang, 2022), while SOT-MRAM parameters were chosen as mentioned in (Saha et al., 2022), (Prenat et al., 2016). The results obtained from these experiments were promising, which excited us to perform system-level analysis as well. System-level analysis is performed using Gem5 & NVMAIN. Gem5 can simulate various architectures at system emulation mode or full system mode (Binkert et al., 2011). In this work, we have used system emulation mode. NVMain is an architectural simulator that can be used for NVM (Poremba and Xie, 2012) simulation. Table 5.5 and 5.6 list the details of the various parameters used in the experiment.

Table 5.5: Embedded System Experimental Set-Up Details

System Configuration	
Simulator	Gem5-NVMain
ISA	ARM
CPU	TimingSimpleCPU
Memory-(SOT-MRAM/DRAM)	4 GB
L1 Cache	64kB (D-Cache),128kB (I-Cache)
L2 Cache	2 MB

The challenge in carrying out the experiments was the availability of SOT-MRAM parameters. Commercial data sheets are not available for SOT-MRAM-based memory. To solve this issue, we applied a similar approach proposed in (Asifuzzaman et al., 2022) by K. Asifuzzaman et al. According to them, new memory devices can

be designed to be compatible with DDRx protocol standards. In such a case, once the data is loaded into the row buffer, the operation would take the same time, irrespective of the memory technology. We have adopted this approach in our experiments. Since SOT-MRAM has better read-write latency than STT-MRAM, we have used 1.1x scaling instead of the 1.25x scaling used for STT-MRAM in Asifuzzaman et al. (2022). Other parameters not associated with row operation are set per the DDR3-1600 standard using the values from the micron data sheet (Micron, 2018). Power and performance analysis is performed using DRAM and SOT-MRAM as main memory. Benchmark programs are chosen from MiBench, an open-source embedded benchmark suite (Guthaus et al., 2001); the list of programs chosen for the experiment is in Table 5.7.

Table 5.6: Memory Configuration Details

Parameter	**DRAM**	**SOT-MRAM**
Channel	4	
Banks	8	
Rank	1	
Clock	800Mhz	
tRCD	11	12
tRP	11	12
tRRD	5	6
Memory Controller	FRFCFS	

5.3.3 Experimental Set-Up for Multi-core Systems

DESTINY (Design Space Exploration for Non-volatile Memory Technology)Mittal et al. (2017) is a simulator for exploring and analyzing non-volatile memory (NVM) technologies. It is a tool used in computer architecture and memory design to evaluate different NVM technologies and configurations' performance, power consumption, and area characteristics. The DESTINY simulator core was employed for conducting

Table 5.7: Benchmark program Details

Category	Programs
Automotive	Bitcount, Qsort, Susan
Network	Patricia
Security	Blowfish, Rijndael
Telecomm	FFT (Fourier)

micro-architectural circuit-level experiments, optimizing the main memory's results for various memory capacities. The system-level experiments integrate two standard

Table 5.8: Workloads from the PARSEC (Bienia et al., 2008a) benchmark suite used in the study.

Program	Application Domain
blackscholes	Financial Analysis
canneal	Engineering
dedup	Enterprise Storage
facesim	Animation
ferret	Similarity Search
fluidanimate	Animation
freqmine	Data Mining
swaptions	Financial Analysis
vips	Media Processing
x264	Media Processing

Gem5(Binkert et al., 2011)- NVMain(Poremba and Xie, 2012) Simulators. Gem5 is a system-level simulator for computer architecture simulation in full system mode. NVMain supports non-volatile main memory technology in hybrid and stand-alone modes. The benchmark programs used in the experiments are summarized in Table 5.8. Benchmark programs for the experiments are taken from the popular benchmark suite Princeton Application Repository for Shared-Memory Computers (PARSEC) benchmark suite(Bienia et al., 2008b). It is a collection of programs used to evaluate multi-core machines. The accuracy of experimental results is crucial in evaluating a

system's performance, which is often achieved through simulation environments. In this study, experiments were conducted using the *full system(FS)* simulation environment of the *gem5-NVMain* simulator. The FS mode provides a more accurate representation of system interactions with the operating system compared to system emulation (SE) mode(Binkert et al., 2011).

Table 5.9: Gem5-NVmain Simulator Set-up

Parameter	**Description**
ISA	X86
CPU	Detailed,2GHz
L1-I & D cache	64KB,64B Block size
L2	1024KB,64B Block size
Main Memory	4 GB (DRAM /SOT-MRAM)
Hybrid Main Memory	1GB DRAM, 3GB SOT-MRAM
The kernel used in Full System Mode	Linux-2.6.22.9

Table 5.10: Memory Configuration Details

Memory Parameters	**Values**
Channels	4
Rank	1
Banks	8
Rows	32768
Cols	64
Memory Scheduling	FRFCFS
Row Buffer Policy	ClosePage

The NVMain configuration files for SOT-MRAM were populated using the approach proposed to design new memory devices compatible with *DDRx* protocol standards (Asifuzzaman et al., 2022). In our experiments, we employed a lower scaling factor for SOT-MRAM, considering its improved performance and modified sensing methods while still adhering to the DDRx protocol, similar to DRAM. This conservative approach was based on our circuit-level analysis results and supported by strong evidence from references (Oh et al., 2023) (Rai et al., 2022) and (Saha et al.,

2022). In contrast, STT-MRAM was scaled at 1.25x, resulting in comparatively inferior read-write latency(Asifuzzaman et al., 2022). Additionally, we set other parameters following the DDR3-1600 standard, utilizing models sourced from the micron data sheet(Micron, 2018). Table 5.9 lists all the multi-core environment-related parameters, viz., clock speed, caches, memory controller and OS kernel used for FS simulation. Table 5.10 provides information on the parameters associated with the memory configuration and scheduling algorithm used in the FS simulation.

5.4 Results and Discussion

This section comprehensively analyses the circuit-level main memory parameters for three memory technologies: DRAM, STT-MRAM and SOT-MRAM. An initial system-level emulation of the embedded system results is discussed. Then, the full system simulation and analysis consider the influence of a multi-core environment with different memory organizations and various memory capacities on memory structures. The work encompasses various metrics across workloads, including latency, power consumption, and bandwidth. Additionally, it examines the impact of various design parameters on the performance of SOT-MRAM. To conduct this analysis, we employed the exploration Algorithm 6, which allowed us to obtain valuable circuit-level results and insights.

5.4.1 Analysis of Micro-architecture Level NVM Main Memory Design Exploration

This section presents the results of micro-architecture level main memory exploration. We will focus on the circuit-level analysis of 1GB to 128GB of main memory. The ensuing analysis is conducted based on our in-depth understanding and by compelling evidence from (Rai et al., 2022) and (Saha et al., 2022).

Total Area

In Figure 5.2, we observe the impact of memory size on the area occupied by different main memory technologies, viz DRAM, STT-MRAM, and SOT-MRAM. As we vary the memory size from 1GB to 128GB, it is evident that the overall memory area also

increases by approximately two times across all memory technologies. The results show that the area occupied by the DRAM to STT-MRAM main memory chip undergoes a significant reduction approximately four times when STT-MRAM replaces DRAM. In parallel, our analysis also shows a substantial reduction in the total chip area occupied by SOT-MRAM to DRAM, which amounts to three times. These results underscore the efficiency and area-saving advantages of employing SOT-MRAM as an alternative memory technology. The same is listed in Table 5.11. Positive values indicate an increase in the area while negative indicates a decrease in the area (percentage-wise).

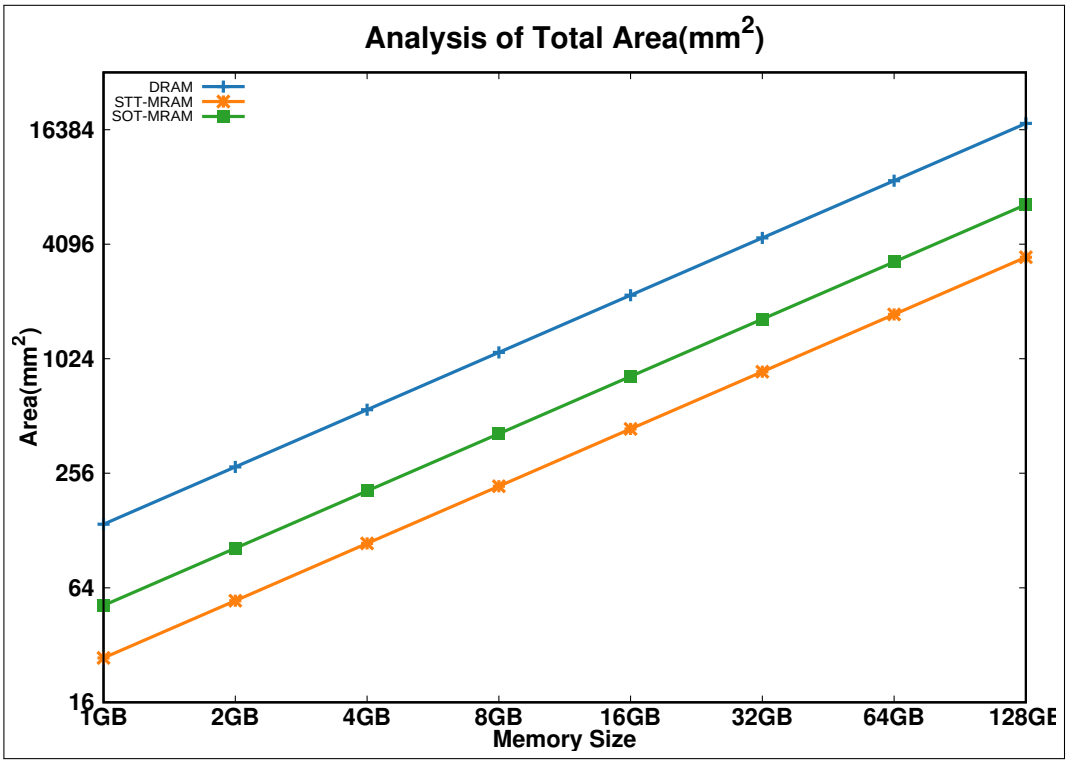


Figure 5.2: Analysis of Area(mm^2)

Table 5.11: Average Percentage change of Circuit Level Parameters for Main memory.

Parameters(%)	**DRAM to STT-MRAM**	**STT to SOT**	**SOT to DRAM**
Total Area(mm^2)	-80.19	88.88	-62.59
Read Latency(ns)	-75.22	-13.91	-78.72
Write Latency(ns)	-70.68	-29.89	-79.98
Read Dynamic Energy(pJ)	-72.17	-16.62	-75.98
Write Dynamic Energy(pJ)	-78.84	-65.48	-92.70
Leakage Power	187.64	-15.90	139.38
EDP-Read	-92.73	-27.95	-94.54
EDP-Write	-93.44	-77.36	-98.56

The reduction in the area presents two significant advantages. Firstly, it allows for higher memory density in a given physical space, making it possible to accommodate larger memory sizes. Secondly, it provides the opportunity to maintain the same memory capacity as DRAM while utilizing MRAM technology with a more compact memory area. STT-MRAM enables a 4x increase in memory size, while SOT-MRAM offers 3x more memory capacity. SOT-MRAM is preferred over STT-MRAM as it overcomes the limitations associated with the latter.

Read and Write Latency

The analysis of main memory read latency (Fig.5.3) reveals significant latency reductions that can be achieved by adopting STT-MRAM and SOT-MRAM over traditional DRAM. At 1GB memory size, STT-MRAM offers approximately 4.56 times faster read access than DRAM, while SOT-MRAM showcases an even more impressive 5.61 times improvement. These reductions continue across various memory sizes, with STT-MRAM and SOT-MRAM consistently outperforming DRAM in read access times.

Similarly, regarding write latency, STT-MRAM and SOT-MRAM demonstrate substantial improvements over DRAM. At 1GB memory size, STT-MRAM provides around 2.68 times faster write access, while SOT-MRAM achieves a remarkable 6.34 times improvement. These gains continue at larger memory sizes, with STT-MRAM offering approximately 3.56x faster write access on average compared to DRAM and SOT-MRAM, delivering an impressive 5.16x improvement. Also, STT-MRAM takes 3x to 2x more write access time than SOT-MRAM at various memory capacities.

In conclusion, adopting MRAM-based technologies, especially SOT-MRAM presents a compelling opportunity to significantly reduce both read and write access times compared to traditional DRAM. These latency reductions have the potential to enhance overall system performance and make MRAM-based main memory systems a promising choice for future memory architectures. SOT-MRAM's ability to overcome the long and unreliable write times of STT-MRAM makes it a superior replacement for DRAM regarding access latencies.

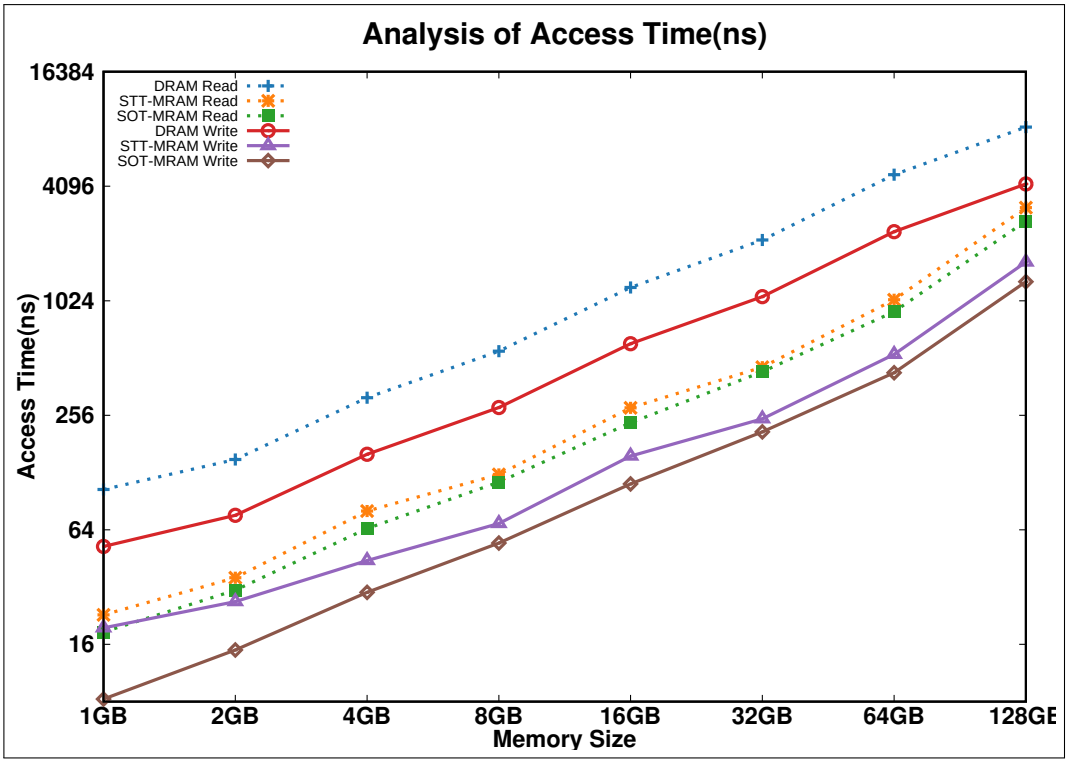


Figure 5.3: Analysis of Access Latency(ns)

Dynamic Energy

The analysis of access energy consumption in main memory reveals notable differences among memory technologies (Fig.5.4). When DRAM main memory is compared to STT-MRAM main memory, significant reductions in read dynamic energy are observed, with an average decrease of approximately 72.18%. Similarly, SOT-MRAM has 75.98% read energy reduction compared to DRAM. Further, SOT-MRAM reduces the read dynamic energy by 16.62% on average compared to STT-MRAM.

The write dynamic energy (Fig.5.4) reduction from STT-MRAM in place of DRAM results in an average decrease of about 78.84%. In contrast, the STT-MRAM to SOT-MRAM main memory comparison yields a reduction of approximately 65.48%. Replacing DRAM from SOT-MRAM leads to the most significant energy reduction, with an average decrease of around 92.70% for write dynamic energy.

These findings highlight the potential benefits of adopting MRAM-based technologies, especially SOT-MRAM, in reducing the energy consumption of main memory systems. The considerable energy reductions achieved through these transitions can improve energy efficiency and lower power consumption in memory-intensive applications, making SOT-MRAM-based main memory systems a promising choice for

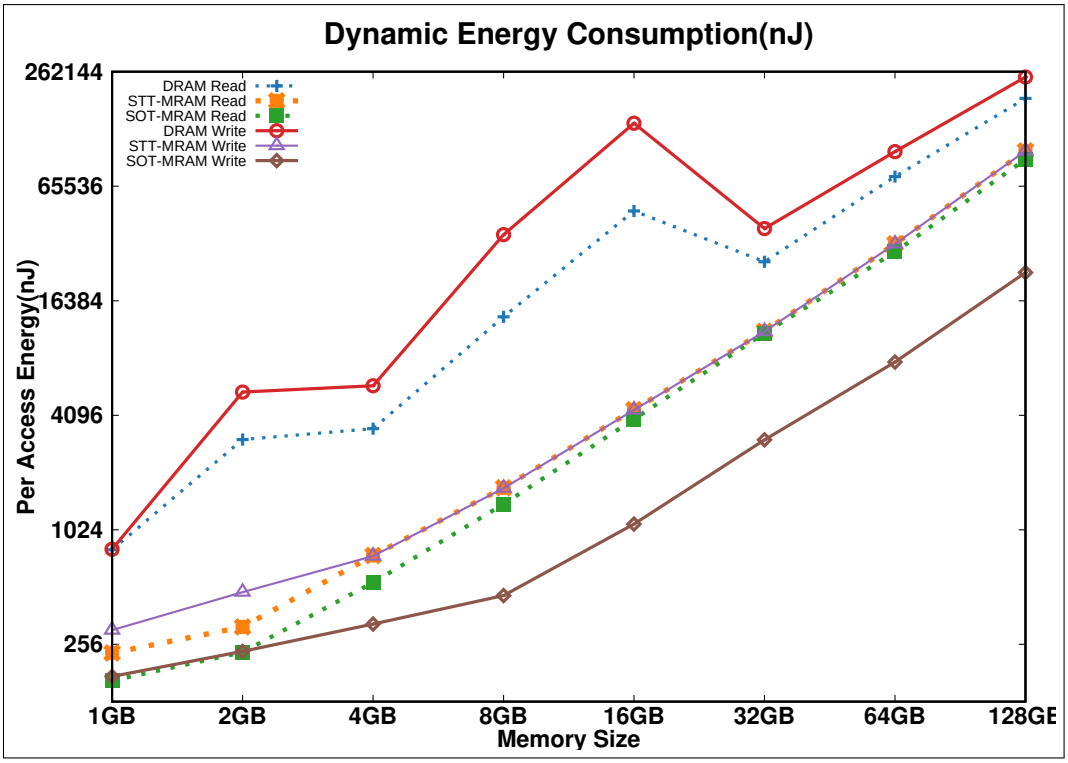


Figure 5.4: Analysis of Access Energy consumption(nJ)

energy-conscious memory architectures.

Energy-Dealy Product(EDP)

Fig.5.5 presents the Energy-Delay Product (EDP) analysis in main memory micro-architecture, revealing significant advantages of transitioning from DRAM to MRAM-based technologies, particularly SOT-MRAM. At 1GB memory size, STT-MRAM offers approximately 93.71% lower EDP for read operations and around 85.99% lower EDP for write operations than DRAM. However, SOT-MRAM outperforms DRAM and STT-MRAM, with approximately 96.35% and 96.62% lower EDP for read and write operations, respectively.

At 8GB memory size, STT-MRAM shows around 97.16% and 98.86% EDP reduction for read and write operations, respectively, compared to DRAM. Yet, SOT-MRAM exhibits even greater improvements, with approximately 97.88% and 99.75% lower EDP for read and write operations, respectively.

For the 64GB memory size, STT-MRAM achieves approximately 90.23% lower EDP for read operations and around 92.59% lower EDP for write operations compared to DRAM. On the other hand, SOT-MRAM again surpasses DRAM and STT-MRAM,

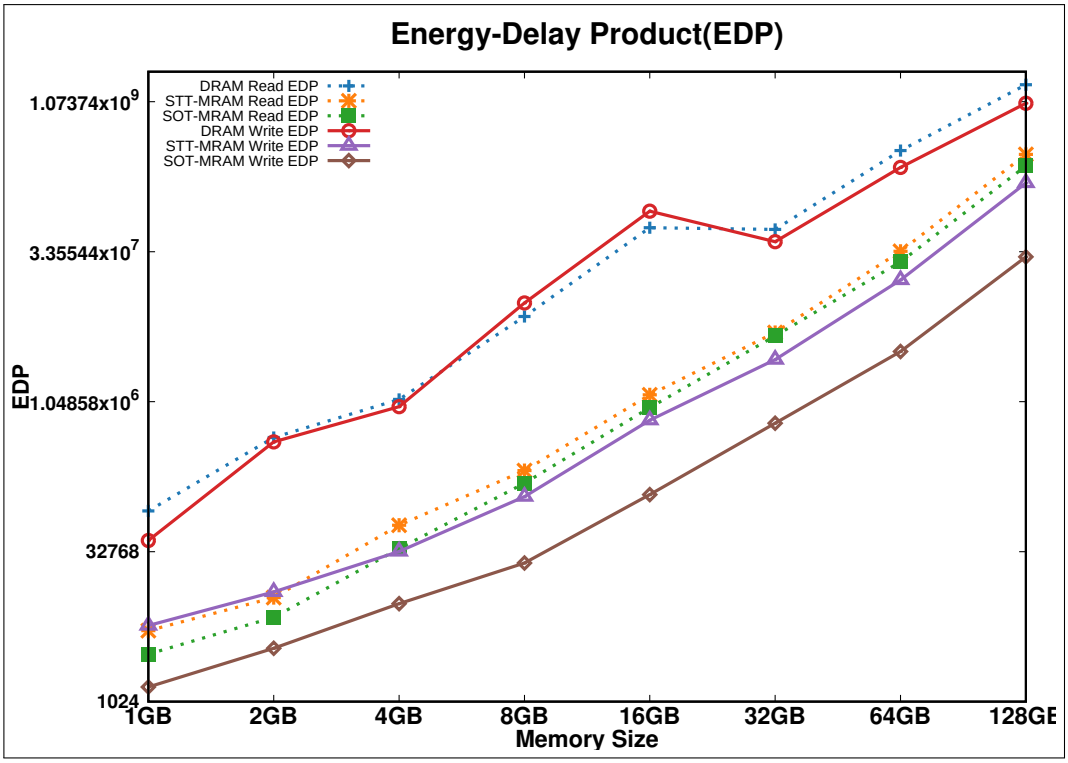


Figure 5.5: Analysis of micro-architecture level EDP

showing approximately 92.35% and 98.58% lower EDP for read and write operations, respectively.

Finally, at 128GB memory size, STT-MRAM exhibits around 80.04% lower EDP for read operations and approximately 84.14% lower EDP for write operations compared to DRAM. Nevertheless, SOT-MRAM remains the superior choice, showcasing approximately 84.68% and 97.12% lower EDP for read and write operations, respectively.

On average, STT-MRAM provides an EDP reduction of approximately 92.73% for read operations and around 77.37% for write operations compared to DRAM. However, SOT-MRAM continues to exhibit the most significant improvements, with an average EDP reduction of approximately 94.55% for read operations and 98.57% for write operations compared to DRAM.

The substantial EDP reductions observed in SOT-MRAM indicate its superiority over DRAM and STT-MRAM. The technology's ability to deliver remarkably lower energy consumption and access delays makes it a highly favourable alternative for energy-efficient memory architectures, offering the potential for enhanced system performance and reduced power consumption.

Leakage Power

In Fig.5.6, the main memory analysis reveals interesting insights regarding leakage power. STT-MRAM consumes approximately 1.8 times more than DRAM, while SOT-MRAM consumes around 1.3 times more than DRAM and has 0.5 times less leakage power than STT-MRAM. Despite this higher leakage power, other essential performance parameters strongly favour SOT-MRAM.

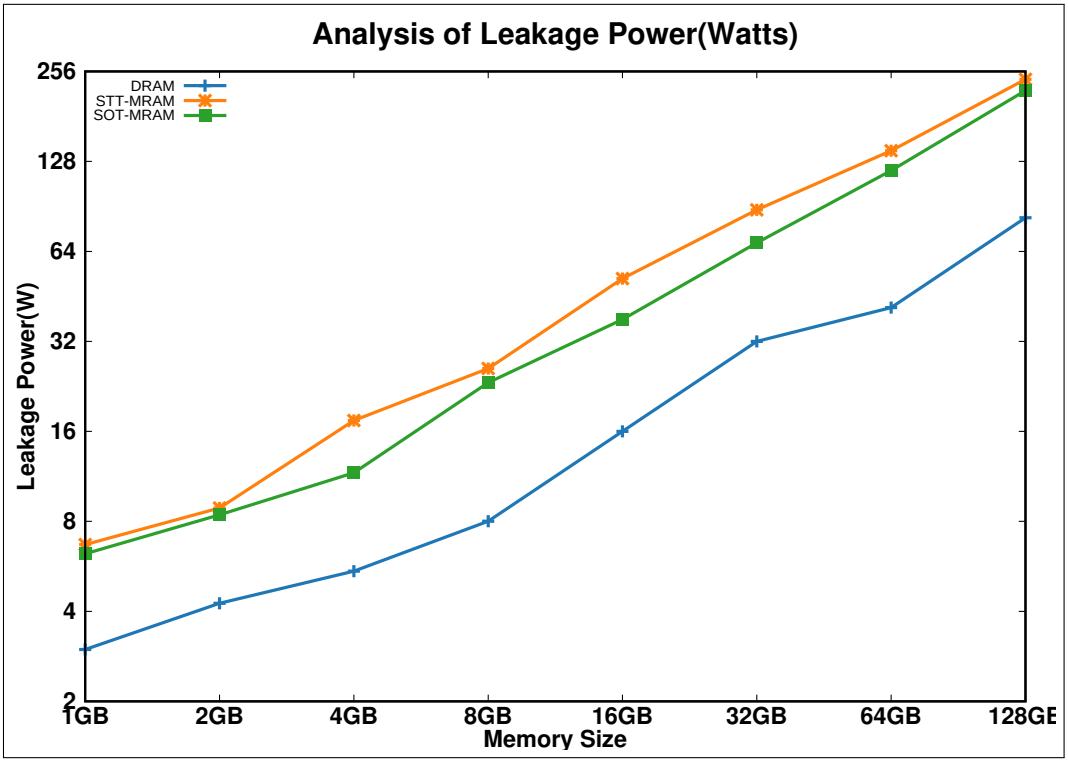


Figure 5.6: Analysis of Leakage Power(W)

In conclusion, the analysis reveals that STT-MRAM and SOT-MRAM consistently outperform DRAM regarding access time. SOT-MRAM showcasing the most impressive performance in Table 5.11. On average, SOT-MRAM provides approximately 4.83 times faster read access and 5.16 times faster write access than DRAM, making it a superior option for memory-intensive tasks. Furthermore, SOT-MRAM leads in energy efficiency, achieving remarkable reductions in dynamic energy consumption compared to DRAM. On average, SOT-MRAM demonstrates a 75.98% reduction in read dynamic energy and a significant 92.70% reduction in write dynamic energy. While STT-MRAM also exhibits energy savings, SOT-MRAM's improvements are more substantial.

The Energy-Delay Product (EDP) further highlights the dominance of SOT-MRAM,

offering approximately 94.54% and 98.56% reductions in read and write EDP compared to DRAM, respectively, outperforming both DRAM and STT-MRAM. Regarding chip area, SOT-MRAM is more efficient, occupying approximately three times less space than DRAM. Although STT-MRAM can provide higher memory density, it has significant disadvantages, such as read disturbance, high write time, and energy consumption.

Despite having slightly higher leakage power than DRAM, SOT-MRAM exhibits nearly 16% less leakage than STT-MRAM. Combined with its superior access time, dynamic energy savings, lower EDP, and better memory density, as shown in Table 5.11 (fourth column), SOT-MRAM proves to be a compelling alternative to replace DRAM in primary memory systems. Its potential for improved memory performance and energy efficiency, along with overcoming the drawbacks of STT-MRAM Table 5.11 (third column), solidifies SOT-MRAM as the more promising choice for next-generation memory solutions.

Table 5.11 summarizes three memory technologies' average micro-architectural parameter values. Positive values denote an increase, whereas negative values indicate a reduction in the corresponding parameters. The first column lists the parameters, while the subsequent columns compare values between DRAM and STT-MRAM, STT-MRAM and SOT-MRAM, and DRAM and SOT-MRAM, respectively.

The analysis shows that at the circuit level, SOT-MRAM performs better than DRAM. However, when adapting an NVM cell for main memory using the same DDRrx protocol, we took a cautious approach. In the following section, we will delve into system-level simulations and their results, where we intentionally scaled the current and timing parameters for SOT-MRAM beyond what DRAM typically uses as in (Asifuzzaman et al., 2022; Asifuzzaman, 2019). Even with these conservative adjustments, SOT-MRAM not only exhibited similar performance to DRAM but even outperformed it in some aspects.

5.4.2 Embedded System Level Analysis

Circuit level analysis gives us strong evidence that in terms of leakage power and write energy, SOT-MRAM has better qualities than DRAM and STT-RAM. In this subsection, we discuss the results obtained while performing system-level analysis using

MiBench(Guthaus et al., 2001), an open-source embedded benchmark. Though we included several programs from the benchmark, we included only a few in our results, which had comparative results. This work could be used as evidence to establish SOT-MRAM as a memory in embedded devices. These devices have strict energy constraints, and hence, power consumption is an important factor; hence, a reduction in power consumption can enhance the battery life of such devices.

Analysis of Power

Power contribution is an important factor to be considered in embedded systems. We analyzed the power consumed for burst operation and the total power consumed by different programs. The total power consumed is the sum of power consumed for a refresh as well as other operations. Figure 5.7 and 5.8 give the details of these power consumption. For all these benchmark programs, the total power consumed by SOT-MRAM is much lower than that of DRAM. However, in the case of burst power, there is only a difference of 6.1%. On average, SOT-MRAM consumes 46.09% less total power than DRAM; for the benchmark programs considered in our experiments, this reduction is mainly because there is no refresh power in SOT-MRAM. In most cases, SOT-MRAM has nearly a 50% power consumption reduction. The total power consumed may depend on write operations as well because write consumes more energy in DRAM.

Performance Analysis

Even though power is an essential factor in embedded systems, we cannot compromise performance entirely. Next, we analyzed the memory behaviour regarding time spent; figure 5.9 shows the time spent by different benchmark programs using DRAM and SOT-MRAM as main memory. Even though the parameters for SOT-MRAM were set at 1.1x times higher than DRAM, the total time taken for the benchmark program is less when compared to DRAM. On an average, there is a performance improvement of nearly 30% when we compared the time spent in cycles, this is because as we see in latency comparison graph, read latencies of SOT-MRAM is lower than DRAM when the size chosen is 4 GB. In our analysis, we have considered the benchmark programs which have nearly similar time spent, Patricia, Qsort have large memory footprints,

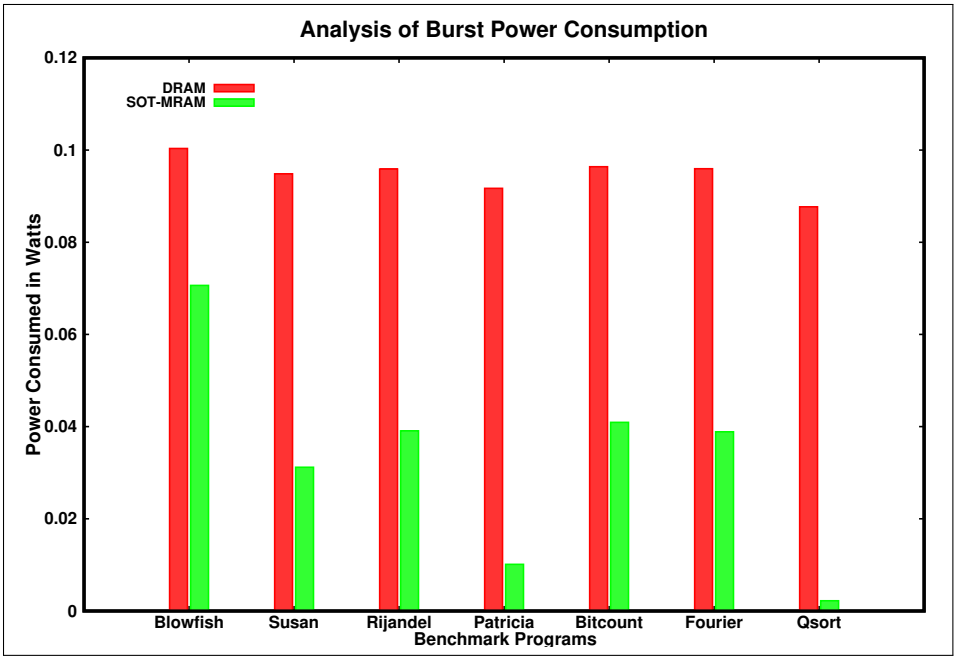


Figure 5.7: Analysis of Burst Power Consumption

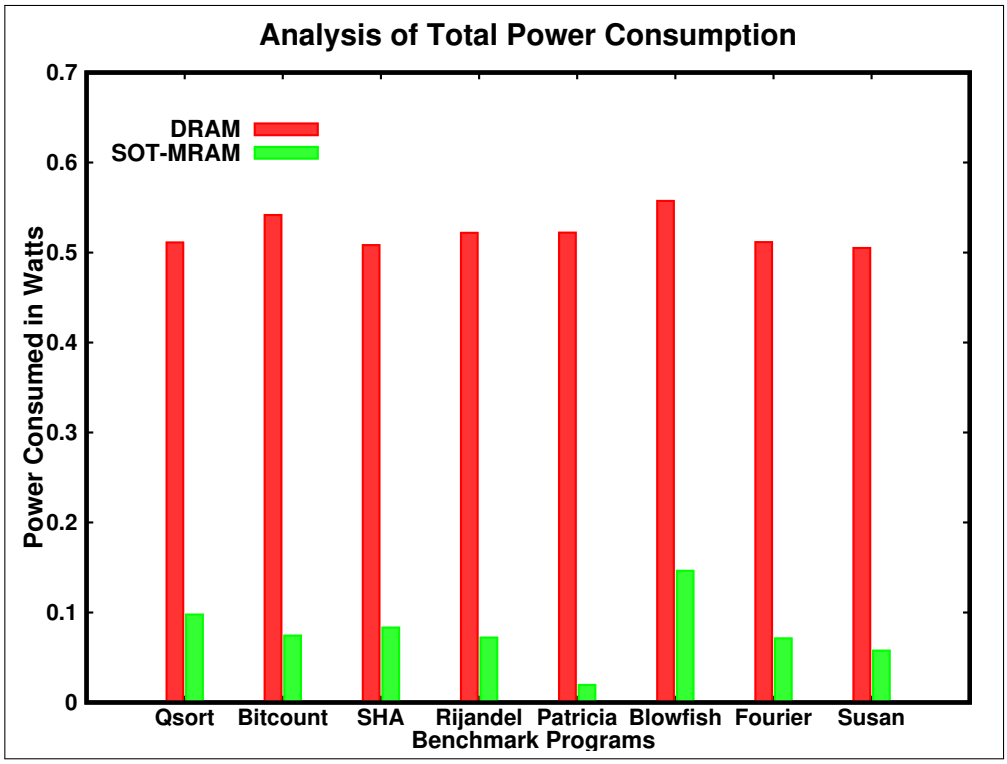


Figure 5.8: Analysis of Total Power Consumption

and they have too much deviation from other programs; hence we have not included them in the graph. In this work, we have proposed and analyzed the impact of using SOT-MRAM as the embedded system's main memory technology. Our experiments show that SOT-MRAM has a power reduction of up to 46.09 % in comparison to DRAM on average. In the case of performance, SOT-MRAM is 30% better than DRAM as a potential main memory candidate.

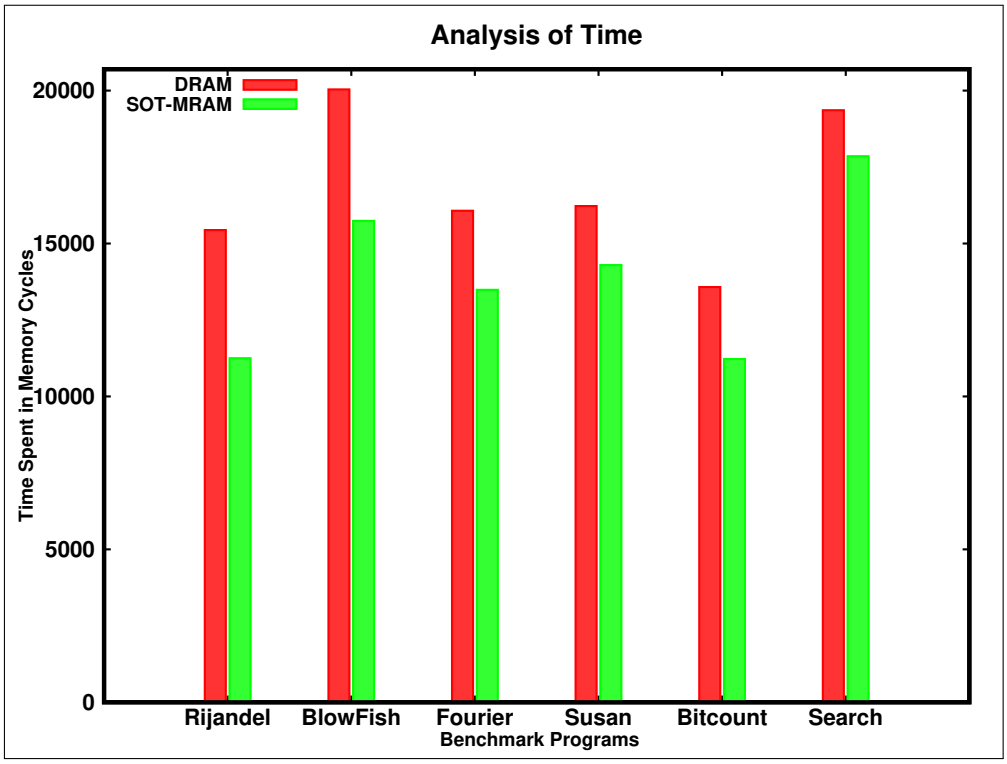


Figure 5.9: Analysis of Time Spent for Execution

5.4.3 Full System Analysis of Multi-core Environment

In this section, we examine how different applications in a multi-core environment perform in terms of power consumption, bandwidth utilization, EDP (Energy-Delay Product), and total latency when using DRAM, hybrid memory, and SOT-MRAM as the main memory structures.

Analysis of Total Power Consumption

This section analyses the total power consumed by different memory structures. Figure 5.10a,5.11a,5.12a and 5.13a depicts the total power consumed by different benchmark programs. Analysis was performed in single-core, dual-core, quad-core, and octa-core environments. The values in Table 5.12 shed light on the power consumption of each memory technology across different core configurations. When values are positive, there's an increase in total power consumption, while negative values indicate a reduction. This analysis helps us understand how core counts influence the total power consumption of different memory technologies under various application workloads.

We consistently observe reductions when comparing DRAM and Hybrid memory power consumption across various core configurations. In a single-core setup, Hybrid memory consumes 62.16% less power than DRAM. This reduction remains consistent as the number of cores increases, with power reductions of 61.32%, 60.99%, and 61.59% in 2-core, 4-core, and 8-core environments, respectively.

It is interesting to note that full SOT-MRAM main memory outperforms DRAM in terms of power consumption. Specifically, in a 1-core environment, full SOT-MRAM exhibits 74.78% less power consumption, which significantly improves over DRAM. This trend continues as the number of cores increases, with reductions of 73.73%, 73.97%, and 73.73% in 2-core, 4-core, and 8-core environments, respectively.

The analysis shows that SOT-MRAM consistently has lower power consumption than Hybrid memory. In a 1-core environment, the SOT-MRAM configuration shows a 33.36% decrease in power consumption compared to Hybrid memory. This reduction remains consistent in 2-core, 4-core, and 8-core environments, with 32.09%, 33.26%, and 31.61% less power consumption, respectively. These findings suggest that SOT-MRAM is more efficient and cost-effective for power-conscious applications than DRAM.

Table 5.12: Comparison of the memory structures Total power consumption(%)

Total Power Consumption			
No.of Cores	**DRAM and Hybrid**	**DRAM and SOT-MRAM**	**Hybrid and SOT-MRAM**
1-core	-62.16	-74.78	-33.36
2-Core	-61.32	-73.73	-32.09
4-Core	-60.99	-73.97	-33.26
8-Core	-61.59	-73.73	-31.61
Average	**-61.51**	**-74.05**	**-32.58**

In conclusion, regardless of the number of cores, on average, full SOT-MRAM demonstrates superior power reduction(74.05%) compared to both DRAM and Hybrid memory configurations. Its consistent performance across various core configurations makes it a promising and energy-efficient memory technology for multi-core environments.

Analysis of Bandwidth

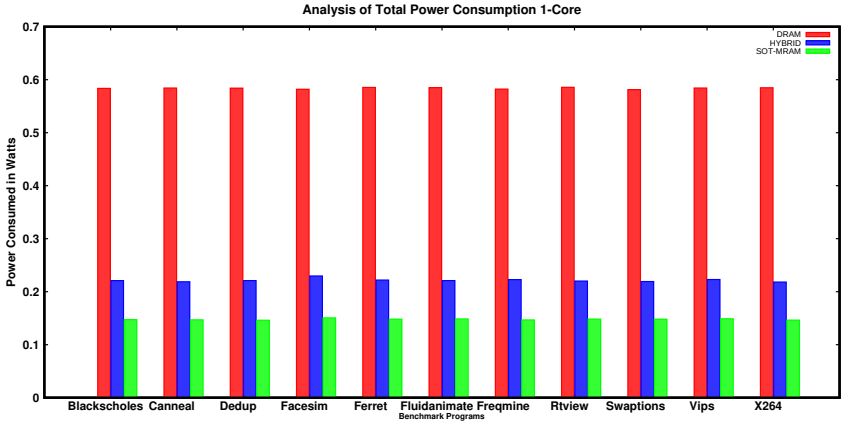
The exploration of bandwidth utilization across different core configurations and memory technologies yields insights into three memory structures. The values presented in Table 5.13 provide insights into the performance of each memory technology across varying core configurations. Positive values indicate an increase in bandwidth utilization, while negative values signify a reduction. This analysis elucidates the impact of core counts on the bandwidth utilization of different memory technologies.

Table 5.13: Comparison of the memory structures Bandwidth Utilization(%)

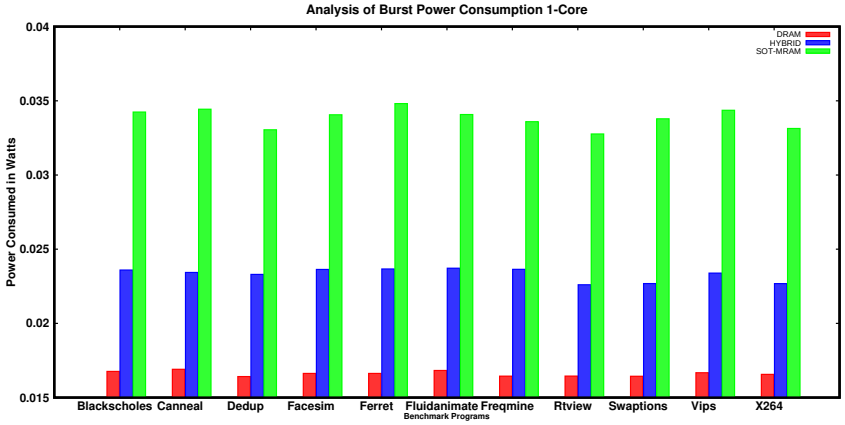
Average Bandwidth(%)			
No.of Cores	**DRAM and Hybrid**	**DRAM and SOT-MRAM**	**Hybrid and SOT-MRAM**
1-core	29.24	40.36	-8.60
2-Core	29.18	41.24	-9.34
4-Core	27.98	40.89	-10.09
8-Core	26.98	37.90	-8.60
Average	**28.34**	**40.10**	**-9.16**

Figure 5.10d,5.11d,5.12d, and 5.13d presents the details of the bandwidth utilization for different memory technologies. In bandwidth utilization higher the value better the result.

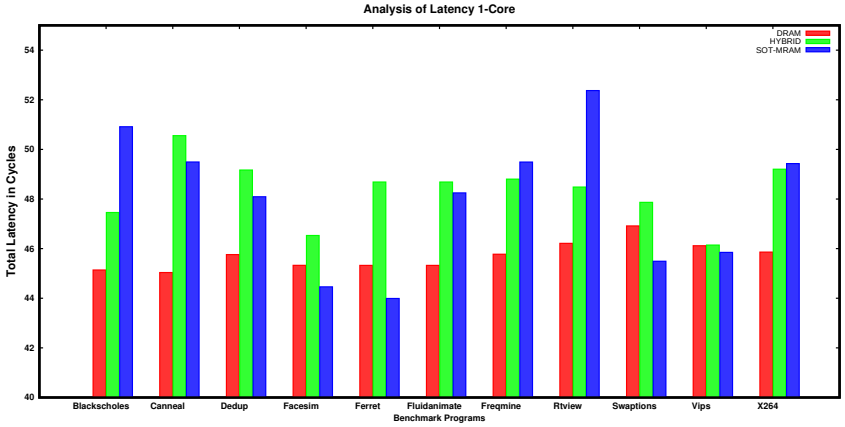
SOT-MRAM consistently demonstrates the highest bandwidth utilization across all core configurations. In a single-core setup, SOT-MRAM achieves an impressive 40.35% increase in bandwidth utilization compared to DRAM. Moreover, Hybrid memory



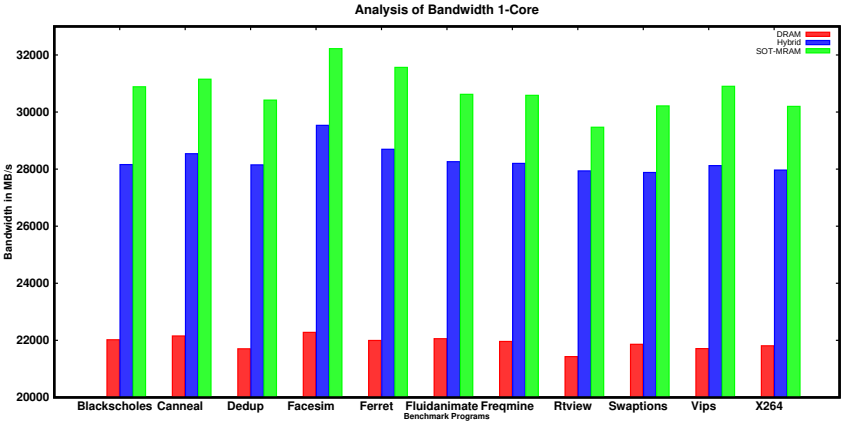
(a) Total Power



(b) Burst Power

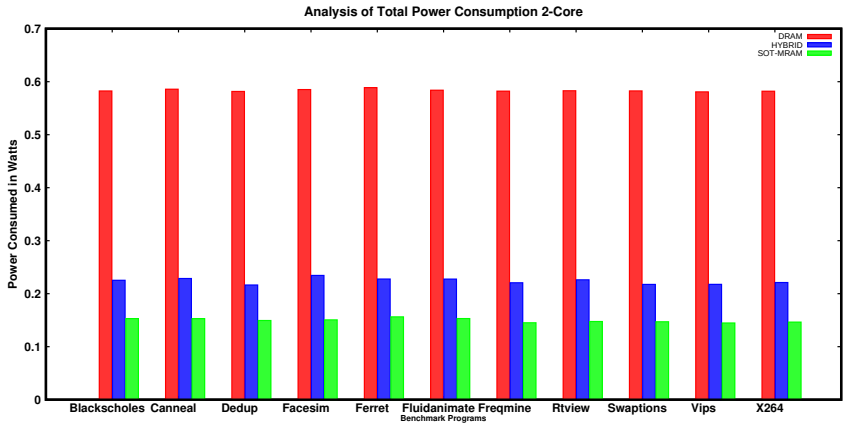


(c) Latency

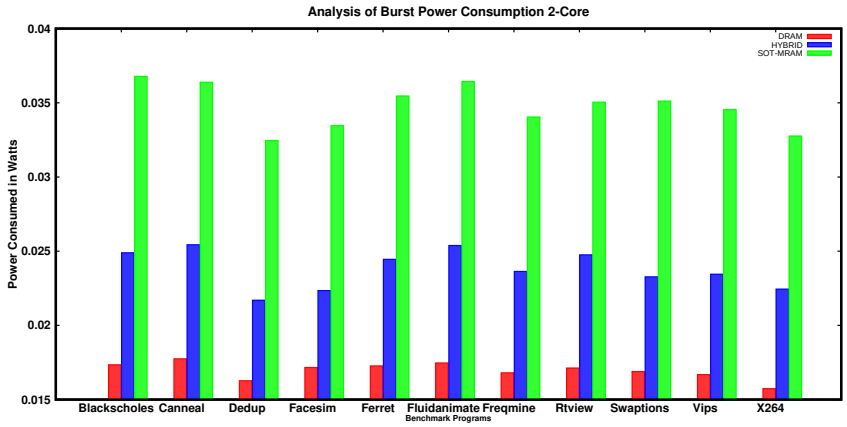


(d) Bandwidth

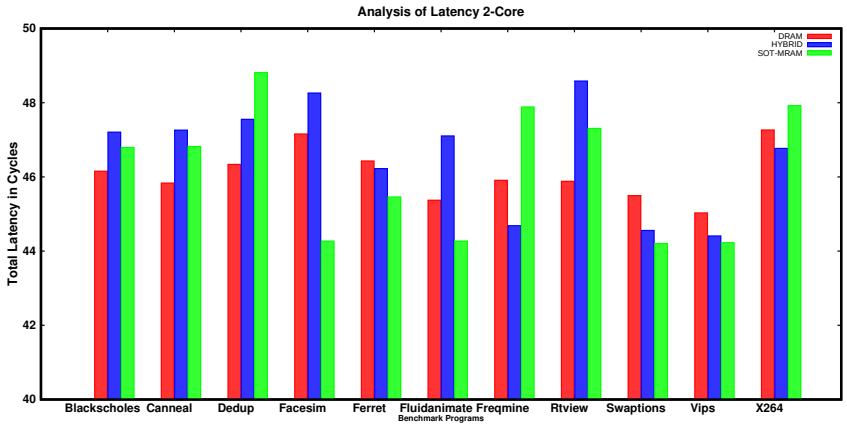
Figure 5.10: Power and Performance analysis of 1- Core



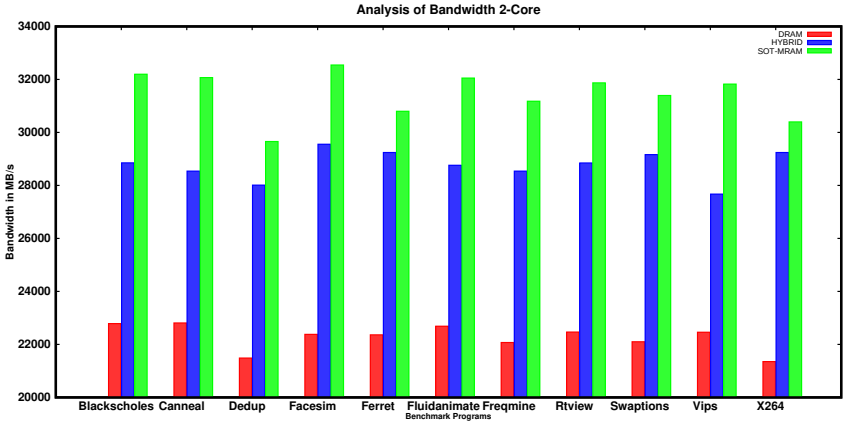
(a) Total Power



(b) Burst Power



(c) Latency



(d) Bandwidth

Figure 5.11: Power and Performance analysis of 2- Core

attains a 29.24% enhancement over DRAM, while its performance is slightly diminished by 8.60% when compared to SOT-MRAM.

As the core counts increase, this trend perseveres. SOT-MRAM maintains its bandwidth utilization superiority in dual-core, quad-core, and octa-core environments, sustaining an average increase of 40.10%. In contrast, DRAM and Hybrid memory showcase comparatively lower values. This consistent pattern underscores the exceptional efficiency of SOT-MRAM in managing data-intensive tasks across diverse computational workloads.

Furthermore, a closer examination of the percentage change values accentuates this prevailing trend. On average, SOT-MRAM exhibits an impressive 40.10% higher bandwidth utilization than DRAM, with Hybrid memory having a noteworthy 28.34% advantage over DRAM. Notably, Hybrid memory's edge over SOT-MRAM diminishes significantly to -9.16%, underscoring the consistent and superior performance of SOT-MRAM in optimizing bandwidth utilization.

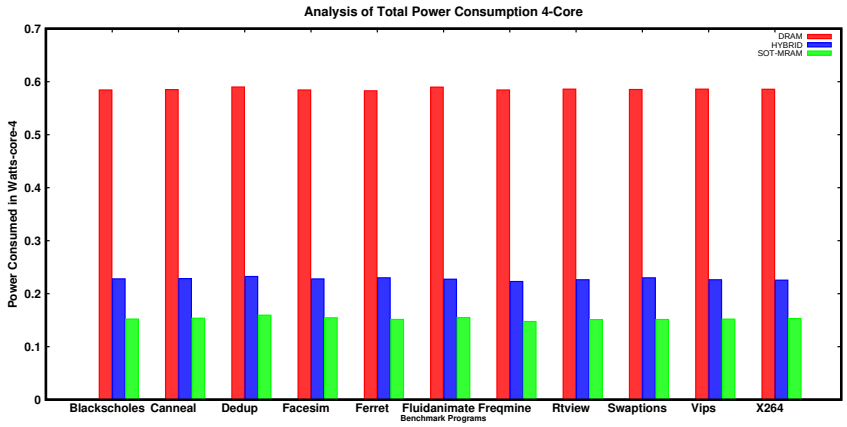
In conclusion, this in-depth analysis highlights the significant influence of memory technology on bandwidth utilization, with SOT-MRAM standing out as the preferred option across a range of core configurations. Its consistently maintaining high bandwidth utilization under varying workloads underscores its potential to enhance overall system performance and efficiency.

Analysis of Total Latency

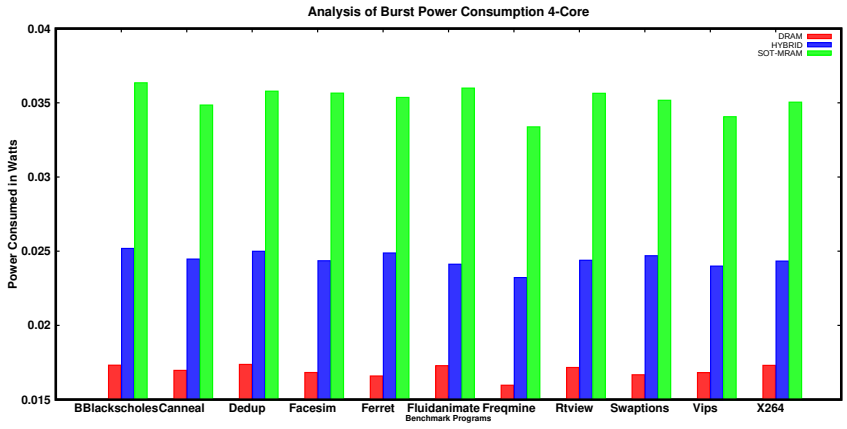
Table 5.14: Comparison of the memory structures Total Latency(%)

Average Latency of Main memories			
No.of Cores	**DRAM and Hybrid**	**DRAM and SOT-MRAM**	**Hybrid and SOT-MRAM**
1-core	5.72	4.98	0.71
2-Core	1.13	0.22	0.91
4-Core	4.37	-0.68	4.84
8-Core	3.61	-4.81	8.12
Average	**3.71**	**-0.07**	**3.64**

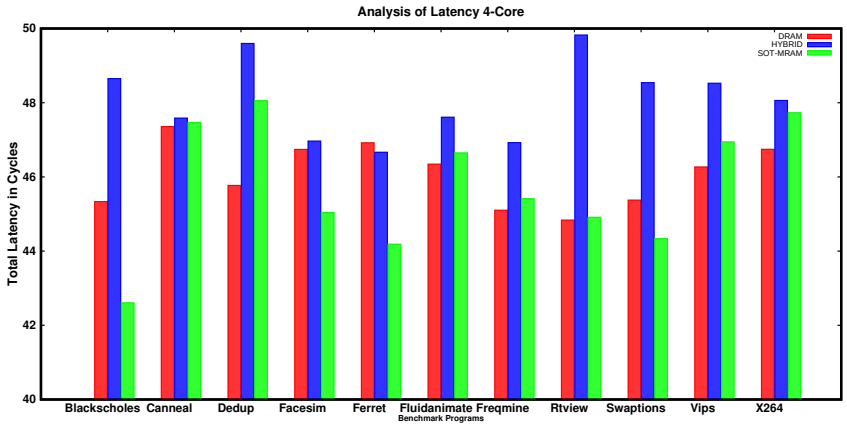
Examining total latency across various core configurations and memory technologies yields significant insights into their performance dynamics. The comparison of DRAM, Hybrid memory, and SOT-MRAM structures, as showcased in Table 5.14, each entry in the table represents the percentage change in total latency when transitioning between memory technologies for different core counts. The interpretation



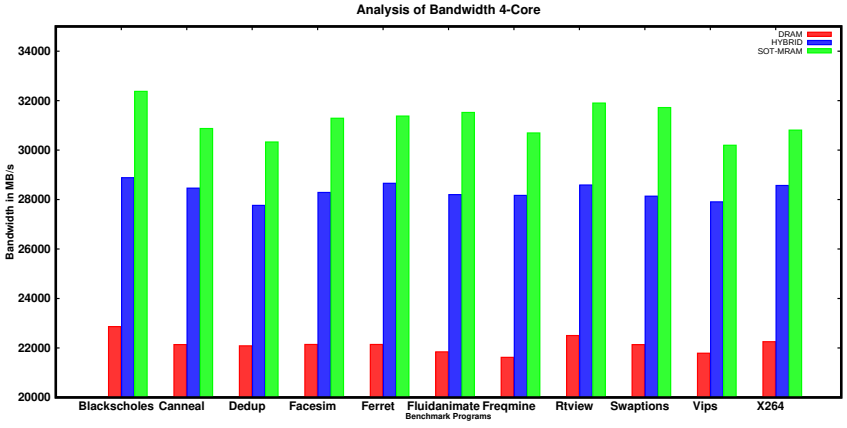
(a) Total Power



(b) Burst Power

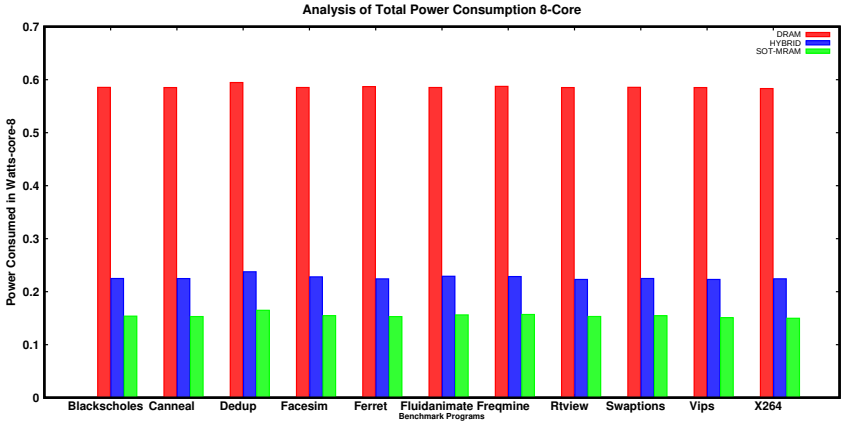


(c) Latency

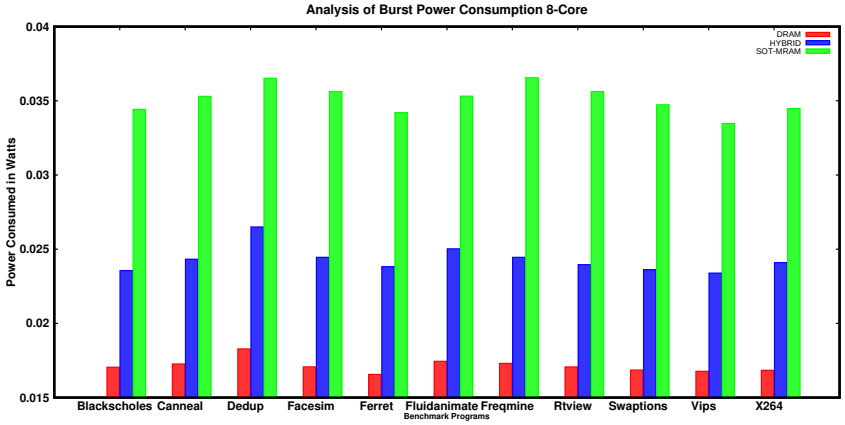


(d) Bandwidth

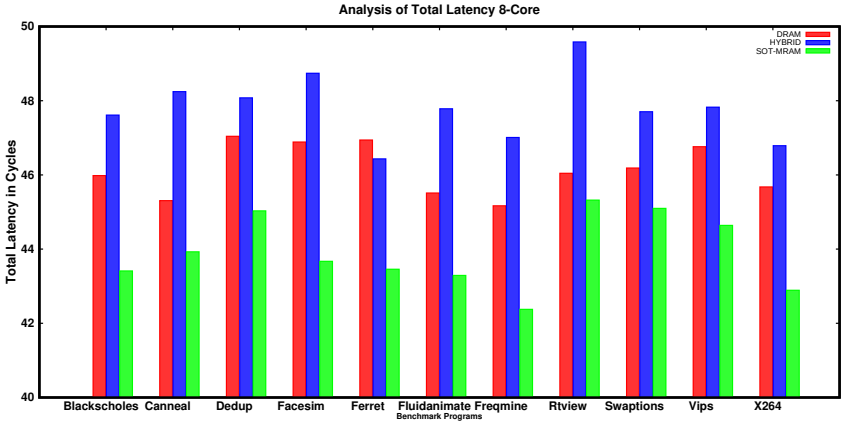
Figure 5.12: Power and Performance analysis of 4- Core



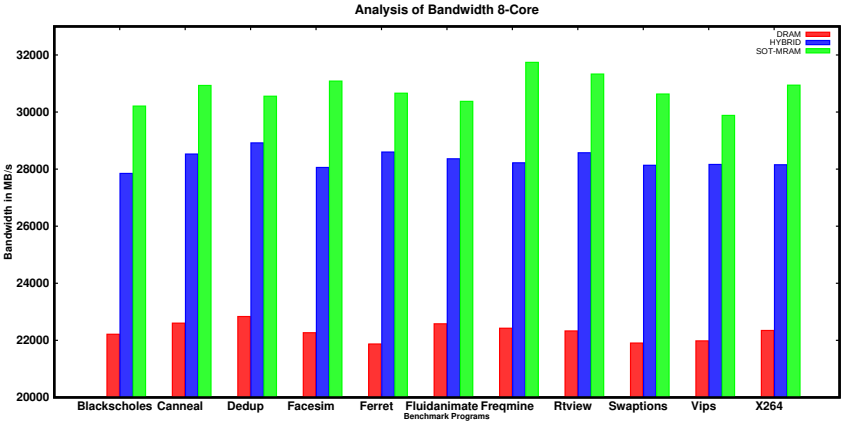
(a) Total Power



(b) Burst Power



(c) Latency



(d) Bandwidth

Figure 5.13: Power and Performance analysis of 8- Core

of positive and negative values reveals the percentage increase or decrease in latency, effectively highlighting the impact of memory technologies. To further enhance clarity, Figures 5.10c, 5.11c, 5.12c, and 5.13c visually represent these trends, elucidating the intricate interplay between core counts and memory technologies in influencing latency outcomes. Notably, a lower latency value indicates better performance, underscoring the critical role of memory technology and core configuration in shaping system responsiveness.

Analyzing the average latency across all core configurations provides a more comprehensive view. On average, Hybrid memory showcases a latency of approximately 47.80 cycles, while DRAM records a latency of about 46.14 cycles, and SOT-MRAM demonstrates the lowest latency at around 43.92 cycles. This reveals that SOT-MRAM consistently offers the lowest latency values regardless of the core count, while Hybrid memory and DRAM follow closely.

In a single-core environment, replacing DRAM with Hybrid memory leads to an increase of 5.72% in total latency, indicating that Hybrid memory technology takes slightly longer to execute tasks in this configuration. Similarly, comparing DRAM to SOT-MRAM, a 4.98% increase in latency is observed for core-1, underlining the potential efficiency of SOT-MRAM in single-core scenarios. As core counts rise to 2, the trend continues. For DRAM to Hybrid memory, the increase in latency drops to 1.13%, showing that the difference in latency between the two memory technologies diminishes with more cores. Comparing DRAM to SOT-MRAM in a dual-core setup yields a mere 0.22% increase in latency for Hybrid memory, implying a minimal impact on performance. Interestingly, core-4 showcases a varying impact. For DRAM to Hybrid memory, the latency increase becomes more significant at 4.37%, suggesting that Hybrid memory may become less favourable in quad-core environments. On the other hand, comparing replacing the DRAM to SOT-MRAM results in a -0.68% latency reduction for core-4, signifying SOT-MRAM's suitability for specific multi-core workloads. Core-8 reveals further intriguing outcomes. DRAM to Hybrid memory results in a latency increase of 3.61%, while switching from DRAM to SOT-MRAM leads to a substantial latency reduction of -4.80%. This divergence highlights the potential of SOT-MRAM to excel in highly parallel computational scenarios.

One notable trend is the consistent latency increase when transitioning from DRAM

to Hybrid memory. Across various core counts, the latency values for Hybrid memory consistently exceed DRAM's. On average, Hybrid memory exhibits around 3.70% higher latency than traditional DRAM. This increase in latency aligns with the characteristics of Hybrid memory, which typically introduces some overhead due to its complex architecture. However, when assessing the shift from DRAM to SOT-MRAM, a different pattern emerges. The average latency change is almost negligible, with SOT-MRAM showcasing a mere 0.07% variation from DRAM. This outcome indicates that SOT-MRAM performs on par with, if not better, DRAM in terms of latency. This holds across different core counts, suggesting SOT-MRAM's consistently delivering efficient memory access, reinforcing its suitability for diverse computational loads.

In summary, the analysis underscores memory technology's pivotal role in influencing total latency. Hybrid memory consistently introduces a modest latency increase compared to DRAM, with variations influenced by core counts. On the other hand, SOT-MRAM maintains latency levels comparable to or better than DRAM across different core configurations. This study advocates for SOT-MRAM's adoption, given its potential to enhance system responsiveness, particularly in multi-core environments. These findings contribute to the broader discourse on memory technology's impact on system performance and highlight SOT-MRAM as a compelling choice for memory system optimization.

Burst Power

Figure 5.10b,5.11b,5.12b and 5.13b shows the burst power analysis conducted across various core configurations (ranging from single-core to octa-core) and memory technologies (DRAM, Hybrid memory, and SOT-MRAM) provides valuable insights into the dynamic interaction between core count and memory technology. As we examine the burst power values, it becomes evident that core count and memory technology play pivotal roles in determining power consumption patterns. When focusing on the average burst power values, we find that the lowest value is associated with DRAM at 0.016 watts, Hybrid memory at 0.023 watts, and SOT-MRAM with the highest burst power consumption at 0.034 watts. Considering the impact of core count, a consistent trend emerges: as the number of cores increases, the burst power consumption converges around the average values mentioned above across all memory technologies.

Analysis of EDP

The analysis conducted on the Energy-Delay Product (EDP) across diverse core configurations and memory technologies offers valuable insights. The comparison of substituting DRAM with Hybrid or SOT-MRAM memory technologies across varying core counts, along with the average EDP values, is presented in Table 5.15. Additionally, the findings illustrated in Fig.5.14 underscore a notable trend: EDP values tend to converge around the memory technology employed rather than the number of cores. This analysis comprehensively explains the intricate relationship between core configurations, memory technologies, and EDP.

The comparative analysis between DRAM and Hybrid Memory Structures demonstrates a consistent average reduction of approximately 57.08% in the Energy-Delay Product (EDP) across varying core configurations. This underscores the inherent energy efficiency improvements that Hybrid memory consistently provides over conventional DRAM, regardless of the number of cores in the system.

Similarly, the comparison between DRAM and SOT-MRAM Memory Structures reveals an average EDP reduction of approximately 72.85% across diverse core counts. This underscores the substantial energy-saving potential inherent in SOT-MRAM when juxtaposed with DRAM, irrespective of the system's core configuration.

Table 5.15: Comparison of Percentage Change in EDP of Main memories(%)

Average Energy-Delay Product(EDP) of Main memories			
No.of Cores	**DRAM and Hybrid**	**DRAM and SOT-MRAM**	**Hybrid and SOT-MRAM**
1-core	-52.82	-69.62	-35.61
2-Core	-58.50	-72.78	-34.41
4-Core	-60.56	-73.60	-33.06
8-Core	-56.45	-75.40	-43.53
Average	**-57.08**	**-72.85**	**-36.65**

Further, examining SOT-MRAM and Hybrid Memory Structures elucidates an average EDP reduction of approximately 36.65% across varying core counts. This finding underscores the synergy between Hybrid memory and SOT-MRAM, showcasing their collective capacity to enhance energy efficiency in memory systems, irrespective of core count fluctuations. The findings conclusively establish that the choice of memory technology exerts a substantial influence on Energy-Delay Product (EDP), whereas the impact of core count remains marginal. Hybrid memory and SOT-MRAM consistently

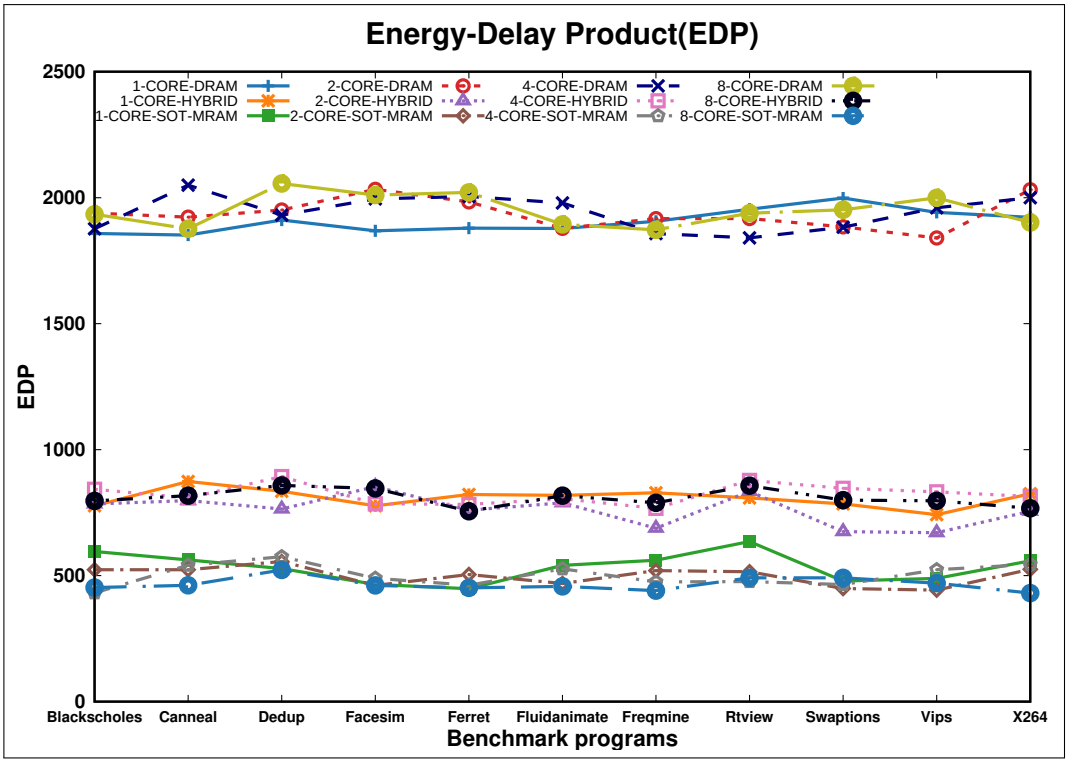


Figure 5.14: EDP of Multi-cores.

emerge as superior alternatives to conventional DRAM in terms of energy efficiency, emphasizing their potential to optimize energy consumption and performance within memory systems.

5.4.4 Memory Organization Analysis

In this section, three memory organizations for DRAM, Hybrid and SOT-MRAM main memory are evaluated. Table 5.17 presents a comparison of system-level parameters for three memory organizations: DRAM-1, DRAM-2, DRAM-3, Hybrid-1, Hybrid-2, Hybrid-3, SOT-MRAM-1, SOT-MRAM-2, and SOT-MRAM-3. The metrics include total power consumption (Watts), burst power (W), average latency (in cycles), bandwidth (MB/s), and Energy-Delay Product (EDP). Positive values indicate an increase, while negative values represent a reduction in the respective parameter.

Fig. 5.15 and 5.16 present EDP, power, and performance analysis of the three memory organizations evaluated.

Looking at the total power consumption, SOT-MRAM-1 stands out as the most power-efficient, followed closely by Hybrid-1 and DRAM-3. Burst power favors Hybrid-1, with SOT-MRAM-1 and SOT-MRAM-2 showing comparable results. Regarding

Table 5.16: Different memory organizations for evaluation

Memory Parameters	DRAM/SOT MRAM	Hybrid	DRAM/SOT MRAM	Hybrid	DRAM/SOT MRAM	Hybrid
	Memory Organization 1		Memory Organization 2		Memory Organization 3	
Channels	4	1+3	2	1+1	4	2+2
Rank	1	1	2	2	1	1
Banks	8	8	8	8	4	4
Rows	32768	32768	32768	32768	8192	8192
Columns	64	64	64	64	512	512
Main Memory Size	4 GB	1GB DRAM + 3GB SOT	4 GB	2 GB + 2GB	4 GB	2 GB + 2GB
Memory Scheduling	FRFCFS					
Row Buffer Policy	ClosePage					
MAT Height	32768				2048	
Row Buffer Size	1				2	

Table 5.17: Comparison of Three Memory Organizations

Comparision of Memory Organizations of Full System Simulation									
System Level Parameters(Avg.)	Dram-1 &2	DRAM-2 & 3	Dram-3 &1	Hybrid-1 &2	Hybrid-2&3	Hybrid-1 &3	SOT-MRAM-1 &2	SOT-MRAM-2 &3	SOT-MRAM-1 &3
Total power(Watts)	-1.55	-36.56	-37.54	51.37	-35.36	-2.16	4.23	10.10	14.75
Burst Power(Watts)	1.03	109.90	112.06	-7.03	29.33	20.24	1.52	-0.74	0.78
Average Latency(in Cycles)	-0.91	-4.05	-4.92	-0.18	103.78	103.41	3.45	6.46	10.14
Bandwidth (MB/s)	5.24	-60.58	-58.51	3.44	-52.87	-51.25	-4.75	-6.30	-10.75
EDP	-3.32	-41.57	-43.51	50.82	170.32	307.70	11.50	24.90	39.26

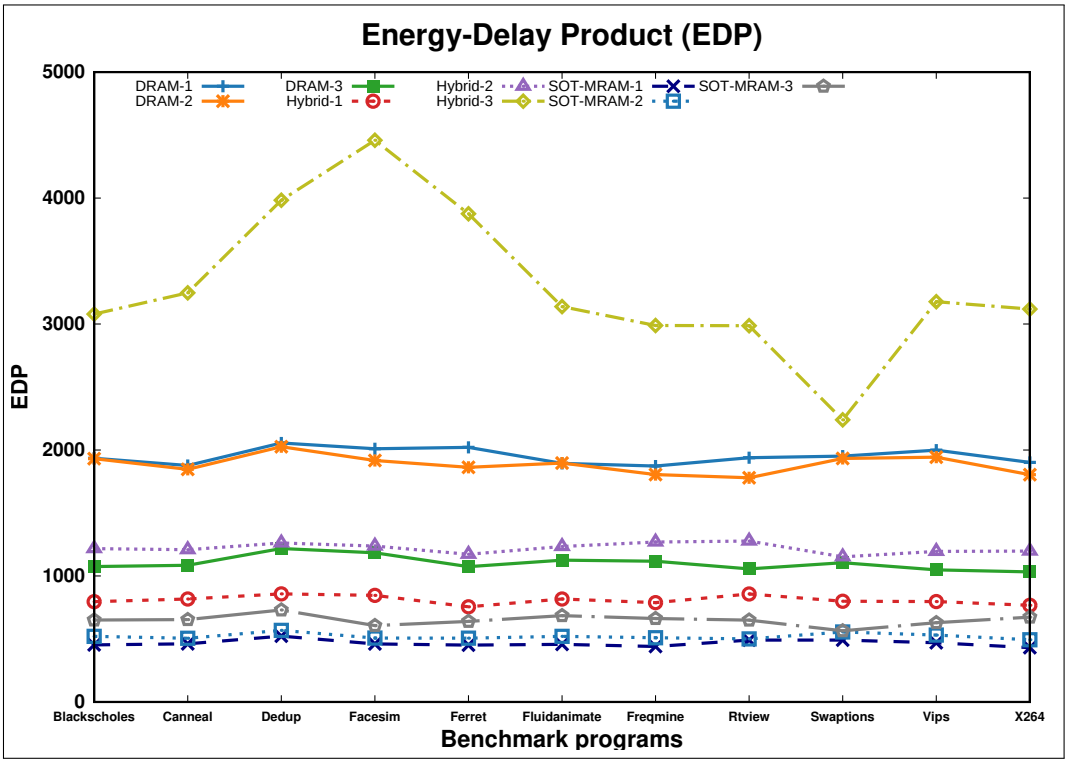
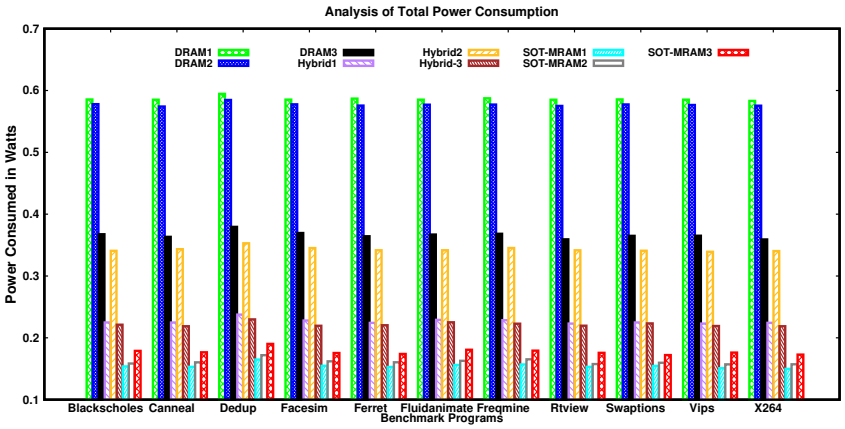
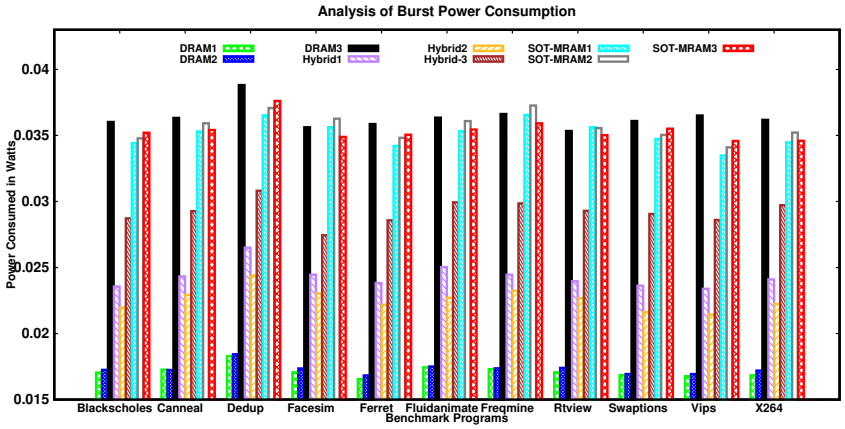


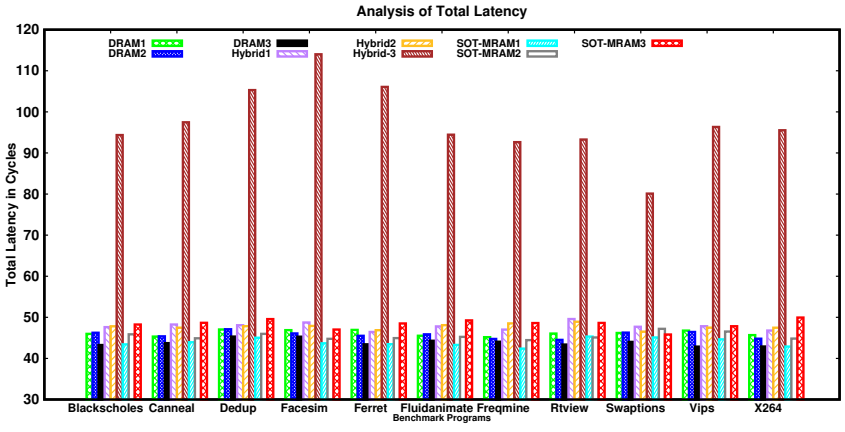
Figure 5.15: Analysis of EDP for memory organizations



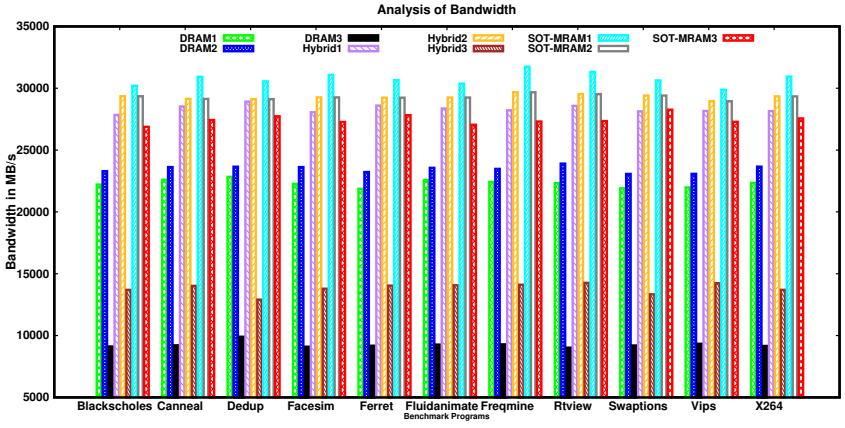
(a) Total power



(b) Analysis of burst power



(c) Analysis of latency



(d) Analysis of bandwidth

Figure 5.16: Analysis of Power and Performance for Memory Organizations

bandwidth, SOT-MRAM configurations exhibit higher values, with SOT-MRAM-1 leading the way. Regarding average latency, DRAM-3 has the lowest latency, while Hybrid-3 shows the highest. As for EDP, DRAM-3 demonstrates the lowest energy-delay product, indicating better overall performance. Comparing the different DRAM configurations, DRAM-2 and DRAM-3 show similar results in most parameters. Hybrid-1 offers the best performance among the Hybrid configurations, while Hybrid-2 and Hybrid-3 have higher latency values.

In the case of SOT-MRAM, SOT-MRAM-1 and SOT-MRAM-2 are generally more power-efficient and have lower latency compared to SOT-MRAM-3. When comparing the best-performing memory configurations among DRAM, Hybrid, and SOT-MRAM, Hybrid-1 and SOT-MRAM-1 stand out with lower power consumption, latency, and bandwidth. In terms of EDP, SOT-MRAM-1 exhibits the most favourable results.

In conclusion, Hybrid-1 and SOT-MRAM-1 emerge as the best-performing memory organizations among the tested configurations, offering a balance of power efficiency, low latency, high bandwidth, and favourable EDP. These findings highlight the potential benefits of adopting Hybrid and SOT-MRAM technologies for next-generation memory solutions, showcasing their superior performance over traditional DRAM configurations.

Regarding total power consumption, Hybrid-1 shows a significant advantage with a 60.74% reduction compared to DRAM-2. However, SOT-MRAM-1 outperforms DRAM-2 and Hybrid-1 with a remarkable 73.21% reduction in total power. Regarding burst power, SOT-MRAM-1 consumes the highest, with a 102.81% increase compared to DRAM-2 and a 44.55% increase compared to Hybrid-1. Average latency is slightly improved in DRAM-2 and SOT-MRAM-1 configuration by 4.55% and 3.94%, respectively, compared to Hybrid-1. For bandwidth, both DRAM-2 and SOT-MRAM-1 show improvements compared to Hybrid-1, with percentage increases of 20.66% and 31.03%, respectively. EDP reduction is most significant in SOT-MRAM-1, showing a 75.27% reduction compared to DRAM-2 and a 42.36% reduction compared to Hybrid-1.

Overall, SOT-MRAM-1 exhibits the best balance of power, performance, and energy efficiency among the three memory configurations, making it the most promising choice for next-generation memory solutions.

5.4.5 Memory Capacity Based Analysis

This section analyses how different main memory technologies with varying capacities, organized into three memory structures, affect system-level parameters when employing an optimal memory organization. The study investigates how these memory configurations impact performance and power consumption in the overall system. The results shed light on the relationship between memory capacity, organization, and key system-level metrics, helping to make informed decisions on memory architecture for improved overall system performance and efficiency. Fig.5.17 presents the results for 4GB to 128GB main memory as the average of values across all workloads from PARSEC. Table-5.18 shows the various capacity hybrid memory compositions from 4GB to 128GB used in the analysis.

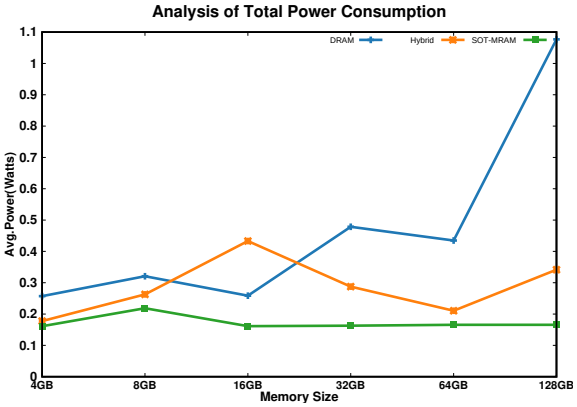
Table 5.18: Composition of Hybrid Memories.

Hybrid Memory Size	DRAM(GB)	SOT-MRAM(GB)
4GB	2	2
8GB	4	4
16GB	8	8
32GB	16	16
64GB	32	32
128GB	64	64

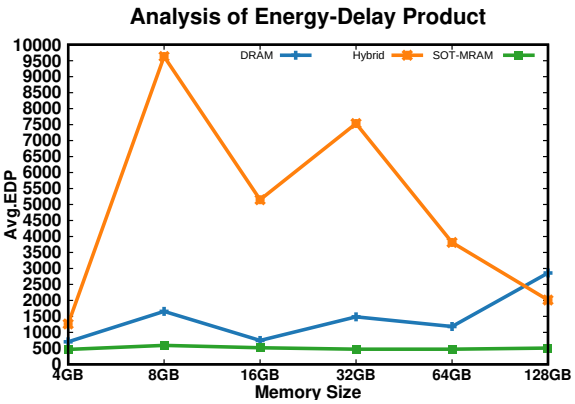
The analysis of total power consumption for the given workloads is shown in Fig.5.17a. The following points analyse the DRAM, Hybrid and SOT-MRAM memory structures.

- DRAM and Hybrid: The hybrid memory configuration shows a notable reduction in total power consumption compared to DRAM across all capacities, ranging from -30.96% to -68.28%. At 16GB capacity, the hybrid memory demonstrates a surprising 67.46% increase in power consumption, suggesting that it may not be the optimal choice for this specific capacity.

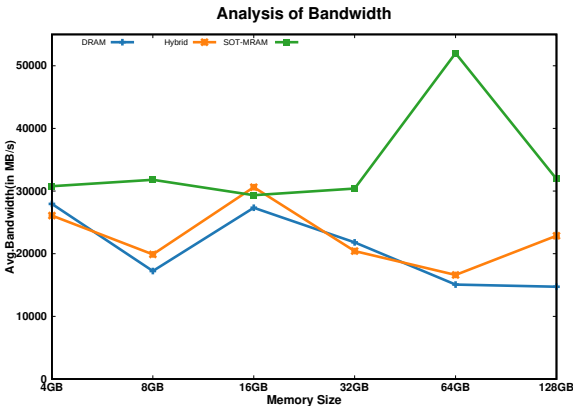
- DRAM and SOT-MRAM: The SOT-MRAM configuration consistently displays significant power savings over DRAM, with reductions ranging from -37.3% to -84.62%.At 64GB capacity, the SOT-MRAM configuration shows a 61.89% decrease in power consumption, highlighting its superiority over DRAM for this capacity.



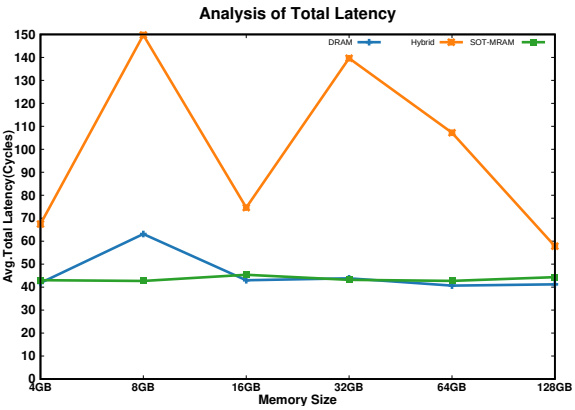
(a) Total power



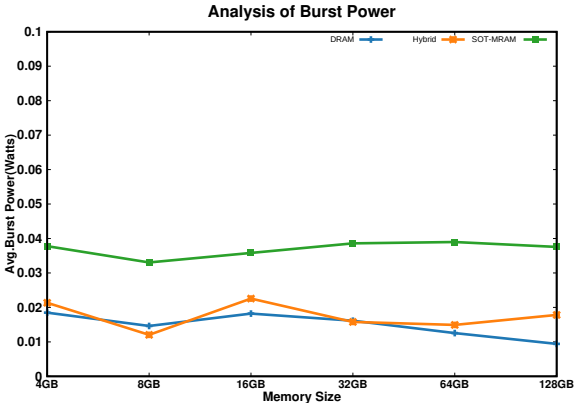
(b) Analysis of EDP



(c) Analysis of Bandwidth



(d) Analysis of Latency



(e) Analysis of Burst power

Figure 5.17: Various Capacity Memory Structures Power and Performance

- Hybrid and SOT-MRAM: The SOT-MRAM memory configuration experiences varying power savings compared to the hybrid memory configuration, ranging from -9.25% to -62.75% reduction. At 16GB capacity, the SOT-MRAM memory exhibits a 62.74% decrease in power consumption compared to hybrid memory, indicating its potential advantage over hybrid memory for this capacity.

Overall, the analysis reveals that SOT-MRAM consistently demonstrates the most significant power savings compared to DRAM and hybrid memory across different capacities. However, it is crucial to consider specific capacity requirements when selecting the optimal memory organization to balance power consumption and performance.

Figure 5.17b illustrates the average Energy-Delay Product (EDP) analysis for optimal memory organizations suitable for 4GB to 128 GB capacity.

- DRAM and Hybrid: The hybrid memory configuration shows an increase in EDP compared to DRAM, with the percentage change ranging from 79.61% to 481.14%.At 8GB capacity, the hybrid memory exhibits the highest increase in EDP, indicating that it may not be the most energy-efficient option for this specific capacity.

- DRAM and SOT-MRAM: The SOT-MRAM configuration demonstrates a reduction in EDP compared to DRAM, with the percentage change ranging from -33.75% to -82.17%.At 4GB capacity, the SOT-MRAM configuration exhibits the highest reduction in EDP, suggesting it may offer better energy efficiency for low-capacity scenarios.

- Hybrid and SOT-MRAM: The SOT-MRAM memory configuration shows a decrease in EDP compared to the hybrid memory configurations, with the percentage change ranging from -63.11% to -93.85%.At capacities of 8GB and 32GB, the SOT-MRAM memory demonstrates the most significant EDP reduction of -93.8% and -93.6%, respectively. This suggests that, for these specific capacities, SOT-MRAM could offer enhanced energy efficiency compared to hybrid memory.

In summary, the analysis reveals that SOT-MRAM demonstrates lower EDP values compared to DRAM, making it a more energy-efficient choice. However, the hybrid

memory configuration shows higher EDP values than both DRAM and SOT-MRAM, suggesting it may not be the best option for optimizing energy efficiency. Selecting the appropriate memory organization should consider the specific capacity requirements and the desired trade-off between energy efficiency and performance.

In the following analysis, we consider Fig.5.17c the average bandwidth utilization across all workloads for three memory structures.

- DRAM and Hybrid: The hybrid memory configuration experiences a decrease in bandwidth utilization compared to DRAM, with the percentage change ranging from -6.66 to 15.36%. At 4GB capacity, the hybrid memory demonstrates the highest decrease in bandwidth utilization, indicating that it may not be the best choice for low-capacity scenarios.

- DRAM and SOT-MRAM: The SOT-MRAM configuration shows an increase in bandwidth utilization compared to DRAM, with the percentage change ranging from 10.09% to 84.45%. At 8GB capacity, the SOT-MRAM configuration exhibits the highest increase in bandwidth utilization, suggesting it may provide better performance for this specific capacity.

- Hybrid and SOT-MRAM: The hybrid memory configuration experiences varying changes in bandwidth utilization compared to SOT-MRAM, with the percentage change ranging from 17.94% to -4.18%. At 64GB capacity, the SOT-MRAM memory exhibits the highest increase in bandwidth utilization, indicating it may offer better performance for this particular capacity.

Overall, the analysis reveals that the hybrid memory shows lower bandwidth utilization compared to DRAM, whereas SOT-MRAM demonstrates improvements over DRAM depending on the capacity.

Fig.5.17d presents the analysis of the three memory technologies' average total latency at various memory capacities.

- DRAM and Hybrid: The hybrid memory configuration exhibits a considerable increase in average latency compared to DRAM, ranging from 61.24% to 137.20%.At 32GB capacity, the hybrid memory shows the highest increase in average latency, indicating that it may not be the best choice for applications.

- DRAM and SOT-MRAM: The SOT-MRAM configuration demonstrates a slight increase and decrease in average latency compared to DRAM, ranging from 2.83% to -32.37%.At 8GB capacity, the SOT-MRAM configuration displays the highest reduction in average latency by -32.37%, suggesting it may offer better performance for this specific capacity.

- Hybrid and SOT-MRAM: The hybrid memory configuration experiences varying changes in average latency compared to SOT-MRAM, ranging from -36.23% to -71.49%.At 8GB capacity, the SOT-MRAM memory exhibits the highest decrease in average latency at -71.48%, indicating it may provide better latency performance for this particular capacity. On average, the SOT-MRAM memory can reduce the latency and speed up workloads by 49.90%.

The analysis indicates that hybrid memory exhibits higher average latencies than DRAM, while SOT-MRAM showcases enhancements and comparable latencies to DRAM, contingent upon the capacity. Specific capacity needs and the intended balance between latency and performance should guide the choice of an optimal memory organization.

The analysis of average burst power across capacities from 4GB to 128GB reveals distinct power consumption patterns among the memory technologies in Fig.5.17e. DRAM exhibits the lowest burst power consumption at 0.015 watts, followed by hybrid memory at 0.017 watts. Notably, SOT-MRAM demonstrates higher burst power consumption, registering 0.036 watts. This observation underscores the varying power efficiency profiles of these memory technologies.

In summary, Fig. 5.17 offers a comprehensive overview of the analysis conducted on different memory technologies across various capacities, from 4GB to 128 GB. Comparing DRAM to hybrid memory, there is a noteworthy reduction in average total power consumption by 23.56%, and in comparison to SOT-MRAM, the reduction is even more significant at 53.21%. Similarly, the average total latency demonstrates a substantial increase of 115.78% when comparing DRAM to hybrid memory but a slight reduction of -2.2% when comparing to SOT-MRAM and a larger reduction of -49.90% for SOT-MRAM to hybrid memory.

Furthermore, the analysis shows improvements in bandwidth utilization. There are increases of 13.27%, 83.80%, and 62.54% for DRAM to hybrid, DRAM to SOT-

MRAM, and SOT-MRAM to hybrid comparisons, respectively. On the other hand, the Energy-Delay Product (EDP) shows an increase of 291.72% for DRAM to hybrid while experiencing reductions of -56.44% and -83.80% for DRAM to SOT-MRAM and SOT-MRAM to hybrid comparisons, respectively.

It is important to note that burst power varies across different sizes, except for SOT-MRAM, which remains consistent across all capacities. In conclusion, this analysis underscores the significance of memory technology selection based on specific capacity requirements and trade-offs between factors such as power consumption, latency, bandwidth utilization, and energy-delay products. Considering these factors, SOT-MRAM emerges as an appealing main memory candidate. Its combination of energy efficiency, competitive performance, and capacity scalability positions it as a technology that could address the evolving demands of modern computing systems while contributing to more sustainable and efficient computing practices.

5.5 Summary

This chapter explores SOT-MRAM as a primary memory alternative to DRAM in embedded systems and multi-core environments. Our analysis indicates notable advancements across various performance parameters. The circuit level analysis shows that SOT-MRAM offers a footprint that is three times smaller and access latencies that are 4 to 5 times lower than those of DRAM. Additionally, it demonstrates a 72.18% reduction in read energy and a 92.70% decrease in write energy. System-level evaluation of embedded systems shows that SOT-MRAM results in a 46.09% reduction in power consumption and a 30% performance improvement over DRAM. In multi-core setups, SOT-MRAM enhances power efficiency by 74.05%, increases bandwidth utilization by 40.10%, and reduces the EDP by 72.85% while maintaining minimal latency impacts. These findings emphasize the potential of SOT-MRAM to significantly enhance performance and energy efficiency in memory systems, positioning it as a replacement for DRAM across a broad spectrum of computing applications.

Chapter 6

Conclusion and Future work

6.1 Conclusions

The SOT-MRAM scaling roadmap uses the proposed MFS framework integrated with the PSC-VRO policy to demonstrate the end-to-end framework for evaluating power performance tradeoffs. The framework also uses the results of the scaling road map to study the density replacement of SRAM with SOT-MRAM in modern applications. The MFS framework reduces the time for the end-to-end design. It integrates and evaluates the different levels of the study to give a complete picture. The results show significant benefits of using SOT-MRAM over SRAM in various performance parameters.

Overall, SOT-MRAM offers better energy efficiency and performance improvements, making it a compelling alternative to traditional SRAM. For example, it reduces average leakage power by around 56%, while the manageable increase in write latency for smaller caches of below 1MB highlights the trade-off between power savings and latency. The density replacement study confirms these results, showing that SOT-MRAM outperforms SRAM in power consumption, latency, and bandwidth utilization at the application level. For example, for a 1:1 MB iso-capacity replacement, SOT-MRAM's total power consumption was 60% lower on average compared to SRAM. Similarly, SOT-MRAM shows an average latency improvement of around 75% and notable gains in both read and write operations. In the iso-area comparison, replacing 4MB of SRAM with 8MB of SOT-MRAM resulted in an average power reduction of 55%, with a 50% improvement in read/write latency. These findings establish

SOT-MRAM as an efficient and scalable memory solution for high-capacity cache implementations.

The Physically Split Cache with Virtual Reordering(PSC-VRO) focused on intra-set write variation and showed substantial reductions in WVAR and wear levelling using the PSC-VRO method. The PSC-VRO approach consistently demonstrated lower WVAR percentages than the baseline and the two state-of-the-art methods across various workload mixes. The average intra-set write variation was reduced by 58.8% with the PSC-VRO method, compared to a 48% reduction with the SA-2 method and a 13% reduction with the SA-1 method. This reduction in variation translates to improved reliability and longevity of the cache memory, reinforcing the benefits of adopting SOT-MRAM in modern applications. The PSC-VRO method improved LI by an average of 65.48%, indicating a significant enhancement in cache lifetime.

In summary, as CMOS technology scales down, SOT-MRAM's advantages become more prominent. SOT-MRAM exhibits reduced leakage power and area with smaller technology nodes. The framework shows SOT-MRAM and its potential to replace SRAM in AI, NLP, and general computing tasks. The MFS framework helps reduce the DSE time needed to study new designs and cache management policies. Given any new NVM design, performing an end-to-end evaluation using this framework is straightforward.

This thesis brings a fresh approach by introducing SOT-MRAM into various memory structures within multi-core systems. The work tackles the challenge of evaluating SOT-MRAM-based memory systems when specific parameters are missing. In-depth investigations at the micro-architectural level and extensive full-system simulations across diverse applications highlight SOT-MRAM's potential as a main memory technology.

At the circuit level, SOT-MRAM offers advantages like a 3x smaller footprint, 4x to 5x lower access latencies compared to DRAM at various capacities, 72.18% less read energy usage, and an impressive 92.70% reduction in write energy. The system-level results conclusively demonstrate that SOT-MRAM, when used as main memory in embedded systems, offers substantial power savings of up to 46.09% and a 30% performance improvement over traditional DRAM. Furthermore, at the system level, during multi-core evaluations with real workloads, it demonstrates substantial

power efficiency with a remarkable 74.05% reduction, a 40.10% increase in bandwidth utilization, and a significant 72.85% reduction in EDP. Crucially, it maintains minimal latency impact, which is vital for real-time applications. The comprehensive evaluation in this work, encompassing circuit-level and system-level analyses, underscores SOT-MRAM's superiority over traditional DRAM. It suggests significant enhancements in performance, energy efficiency, and minimal latency penalties. These findings position SOT-MRAM as a technology with the potential to revolutionize memory systems across a wide spectrum of computing applications.

The application of these results extends to various computing domains, where SOT-MRAM can offer substantial benefits. Its reduced power consumption and enhanced bandwidth utilization can enhance the performance of energy-sensitive applications in fields like IoT and mobile devices. In high-performance computing, the minimal latency impact and favourable power reduction make it a promising candidate for optimizing memory-intensive tasks. Additionally, real-time systems, which demand low latency and energy efficiency, can greatly benefit from the superior attributes of SOT-MRAM. The potential for SOT-MRAM to replace or complement DRAM in diverse computing applications is a significant step toward improved system efficiency and performance.

6.2 Future Directions

One future work direction is to optimize sense amplifiers in the memory array and evaluate performance against modern workloads in future work. Furthermore, device parameters may differ, and manufacturing and operating conditions can influence results. As memory technologies continue to evolve, our findings require updates to remain relevant. Despite these constraints, our research provides valuable insights into integrating SOT-MRAM into memory systems with reliable simulations.

A potential enhancement for PSC would be to make the reordering interval and partition ratio dynamic, further optimizing performance. The fault maps can be created once the cells wear out, and a byte/frame level handling mechanism helps extend the reliability and lifetime of LLC further.

Bibliography

AMD processors. https://www.amd.com/en/products/processors/server/epyc.html. [Accessed 18-05-2024].

Intel Core. https://www.intel.com/content/www/us/en/products/details/processors.html. [Accessed 18-05-2024].

List of Nvidia graphics processing units. https://www.nvidia.com/en-in/geforce/graphics-cards/. [Accessed 18-05-2024].

Agarwal, S. and H. K. Kapoor (2017). Towards a better lifetime for non-volatile caches in chip multiprocessors. In *2017 30th International Conference on VLSI Design and 2017 16th International Conference on Embedded Systems (VLSID)*, pp. 29–34. IEEE.

Agarwal, S. and H. K. Kapoor (2019). Improving the lifetime of non-volatile cache by write restriction. *IEEE Transactions on Computers 68*(9), 1297–1312.

Agarwal, S. and H. K. Kapoor (2020). Linovo: Longevity enhancement of non-volatile last level caches in chip multiprocessors. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 194–199. IEEE.

Alhalabi, R., E. Nowak, I. L. Prejbeanu, and G. di Pendina (2018). High density sot-mram memory array based on a single transistor. *2018 Non-Volatile Memory Technology Symposium (NVMTS)*, 1–3.

Alhalabi, R., E. Nowak, I.-l. Prejbeanu, and G. D. Pendina (2018). High density sot-mram memory array based on a single transistor. In *2018 Non-Volatile Memory Technology Symposium (NVMTS)*, pp. 1–3.

Asifuzzaman, K. (2019). *Evaluation of STT-MRAM main memory for HPC and real-time systems*. Ph. D. thesis, Department of Computer Architecture,Universitat Politècnica de Catalunya.

Asifuzzaman, K., M. Fernandez, P. Radojković, J. Abella, and F. J. Cazorla (2019). Stt-mram for real-time embedded systems: Performance and wcet implications. In *Proceedings of the International Symposium on Memory Systems*, MEMSYS '19, New York, NY, USA, pp. 195–205. Association for Computing Machinery.

Asifuzzaman, K., R. S. Verdejo, and P. Radojković (2022, jan). Performance and power estimation of stt-mram main memory with reliable system-level simulation. *ACM Trans. Embed. Comput. Syst. 21*(1).

Bienia, C., S. Kumar, J. P. Singh, and K. Li (2008a). The parsec benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, PACT '08, New York, NY, USA, pp. 72–81. Association for Computing Machinery.

Bienia, C., S. Kumar, J. P. Singh, and K. Li (2008b). The parsec benchmark suite: characterization and architectural implications. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, PACT '08, New York, NY, USA, pp. 72–81. Association for Computing Machinery.

Binkert, N., B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood (2011, aug). The gem5 simulator. *SIGARCH Comput. Archit. News 39*(2), 1–7.

Cargnini, L. V., L. Torres, R. M. Brum, S. Senni, and G. Sassatelli (2014). Embedded memory hierarchy exploration based on magnetic random access memory. *Journal of Low Power Electronics and Applications 4*(3), 214–230.

Chen, X., E. H.-M. Sha, W. Jiang, C. Yang, T. Wu, and Q. Zhuge (2017). Refinery swap: An efficient swap mechanism for hybrid dram–nvm systems. *Future Generation Computer Systems 77*, 52–64.

Escuin, C., P. Ibañez, T. Monreal, J. M. Llaberia, and V. Viñals (2022). Forecasting lifetime and performance of a novel nvm last-level cache with compression. *arXiv preprint arXiv:2204.03512*.

Escuin, C., A. A. Khan, P. Ibañez, T. Monreal, V. Viñals, and J. Castrillon (2022). Hycsim: A rapid design space exploration tool for emerging hybrid last-level caches. In *System Engineering for Constrained Embedded Systems*, DroneSE and RAPIDO, New York, NY, USA, pp. 53–58. Association for Computing Machinery.

Escuin, C., A. A. Khan, P. Ibáñez, T. Monreal, J. Castrillon, and V. Viñals (2023). Compression-aware and performance-efficient insertion policies for long-lasting hybrid llcs. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 179–192.

Evenblij, T., M. Perumkunnil, F. Catthoor, S. Sakhare, P. Debacker, G. Kar, A. Furnemont, N. Bueno, J. I. Gómez-Pérez, and C. Tenllado (2019). A comparative analysis on the impact of bank contention in stt-mram and sram based llcs. In *2019 IEEE 37th International Conference on Computer Design (ICCD)*, pp. 255–263.

Fu, Y., Y. Lu, Z. Chen, Y. Wu, and N. Xiao (2022, jul). Design and simulation of content-aware hybrid dram-pcm memory system. *IEEE Trans. Parallel Distrib. Syst. 33*(7), 1666–1677.

Garello, K., F. Yasin, S. Couet, L. Souriau, J. Swerts, S. Rao, S. Van Beek, W. Kim, E. Liu, S. Kundu, et al. (2018). Sot-mram 300mm integration for low power and ultrafast embedded memories. In *2018 IEEE Symposium on VLSI Circuits*, pp. 81–82. IEEE.

Gholami, A., Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer (2024). Ai and memory wall. *IEEE Micro*, 1–5.

Gupta, M., M. Perumkunnil, K. Garello, S. Rao, F. Yasin, G. Kar, and A. Furnémont (2020). High-density sot-mram technology and design specifications for the embedded domain at 5nm node. In *2020 IEEE International Electron Devices Meeting (IEDM)*, pp. 24.5.1–24.5.4.

Guthaus, M., J. Ringenberg, D. Ernst, T. Austin, T. Mudge, and R. Brown (2001). Mibench: A free, commercially representative embedded benchmark suite. In *Proceedings of the Fourth Annual IEEE International Workshop on Workload Characterization. WWC-4 (Cat. No.01EX538)*, pp. 3–14.

Han, S. and Y. Jiang (2023). Risc-v-based evaluation and strategy exploration of mram triple-level hybrid cache systems. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems 31*(7), 980–992.

Han, S. and Y. Jiang (2024, 01). Advanced hybrid MRAM based novel GPU cache system for graphic processing with high efficiency. *AIP Advances 14*(1), 015058.

Hosomi, M., H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano (2005). A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram. In *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.*, pp. 459–462.

Inc., E. T. Emd3d256m-256mb spin-transfer torque mram. `https://www.everspin.com/family/emd3d256m`, publisher=Everspin Technologies Inc,, year=2018.

Inc., M. (2006). 1gb:x4,x8,x16 ddr3 sdram. `https://media-www.micron.com/-/media/client/global/documents/products/data-sheet/dram/ddr3/1gb_ddr3_sdram.pdf`.

Inci, A., M. M. Isgenc, and D. Marculescu (2022). Deepnvm++: Cross-layer modeling and optimization framework of nonvolatile memories for deep learning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 41*(10), 3426–3437.

Jang, Y. and J. Park (2022). Area and energy efficient joint 2t sot-mram-based on diffusion region sharing with adjacent cells. *IEEE Transactions on Circuits and Systems II: Express Briefs 69*(3), 1622–1626.

Jia, G., G. Han, J. Jiang, and L. Liu (2017). Dynamic adaptive replacement policy in shared last-level cache of dram/pcm hybrid memory for big data storage. *IEEE Transactions on Industrial Informatics 13*(4), 1951–1960.

Jing, X. and H. Li (2022). Construction and optimization of heterogeneous memory system based on numa architecture. *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 767–772.

Kallinatha, H. D., S. Rai, and B. Talawar (2024). A detailed study of sot-mram as an alternative to dram primary memory in multi-core environment. *IEEE Access 12*, 7224–7243.

Kalllinatha, H. D. and B. Talawar (2023, 06). Comparative analysis of non-volatile memory on-chip caches. *AIP Conference Proceedings 2705*(1), 040008.

Kim, T., S. Jamil, J. Park, and Y. Kim (2020). Optimizing heap memory object placement in the hybrid memory system with energy constraints. *IEEE Access 8*, 130323–130339.

Komalan, M., O. H. Rock, M. Hartmann, S. Sakhare, C. Tenllado, J. I. Gómez, G. S. Kar, A. Furnemont, F. Catthoor, S. Senni, D. Novo, A. Gamatie, and L. Torres (2018). Main memory organization trade-offs with dram and stt-mram options based on gem5-nvmain simulation frameworks. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 103–108.

Komalan, M. P., M. Gupta, S. Rao, W. Kim, F. Yasin, S. Couet, A. Furnemont, and G. S. Kar (2022). Feasibility analysis of embedded mram solutions at advanced process nodes. In *2022 6th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, pp. 73–75.

Kryder, M. H. and C. S. Kim (2009). After hard drives—what comes next? *IEEE Transactions on Magnetics 45*(10), 3406–3413.

Kültürsay, E., M. Kandemir, A. Sivasubramaniam, and O. Mutlu (2013). Evaluating stt-ram as an energy-efficient main memory alternative. In *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 256–267.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR abs/1909.11942*.

Leskovec, J. and A. Krevl (2014, June). SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

Li, Y., W. Kang, K. Zhou, K. Qiu, and W. Zhao (2023, jan). Experimental demonstration of stt-mram-based nonvolatile instantly on/off system for iot applications: Case studies. *ACM Trans. Embed. Comput. Syst. 22*(2).

Liao, Y.-C., C. Pan, and A. Naeemi (2020). Benchmarking and optimization of spintronic memory arrays. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits 6*(1), 9–17.

Liu, C., L. Chen, X. Hao, and M. Ni (2022). Optimized fast data migration for hybrid dram/stt-mram main memory. *IEICE Electronics Express 19*(1), 20210493–20210493.

Liu, E., K. Li, A. Shen, and S. He (2023). Area and energy efficient sot-mram bit cell based on 3 transistors with shared diffusion regions. *IEEE Transactions on Circuits and Systems II: Express Briefs*.

Lu, A., J. Lee, T.-H. Kim, M. A. U. Karim, R. S. Park, H. Simka, and S. Yu (2024). High-speed emerging memories for ai hardware accelerators. *Nature Reviews Electrical Engineering 1*(1), 24–34.

Ma, H., Y. Wang, R. Ali, Z. Hou, D. Zhang, E. Deng, G. Wang, and W. Zhao (2021). Spinsim: A computer architecture-level variation aware stt-mram performance evaluation framework. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. IEEE.

Mahdavi, N., F. Razaghian, and H. Farbeh (2022, jul). Data block manipulation for error rate reduction in stt-mram based main memory. *J. Supercomput. 78*(11), 13342–13372.

Marinelli, T., J. I. G. Pérez, C. Tenllado, M. Komalan, M. Gupta, and F. Catthoor (2022, jan). Microarchitectural exploration of stt-mram last-level cache parameters for energy-efficient devices. *ACM Trans. Embed. Comput. Syst. 21*(1).

McKee, S. A. (2004). Reflections on the memory wall. In *Proceedings of the 1st Conference on Computing Frontiers*, CF '04, New York, NY, USA, pp. 162. Association for Computing Machinery.

Micron, T. I. (2018). 4gb:x4,x8,x16 ddr3 sdram. Technical report.

Mishra, R., T. Kim, J. Park, and H. Yang (2021, Feb). Shared-write-channel-based device for high-density spin-orbit-torque magnetic random-access memory. *Phys. Rev. Appl. 15*, 024063.

Mittal, S. and J. S. Vetter (2014a). Ayush: A technique for extending lifetime of sram-nvm hybrid caches. *IEEE Computer Architecture Letters 14*(2), 115–118.

Mittal, S. and J. S. Vetter (2014b). Equalchance: Addressing intra-set write variation to increase lifetime of non-volatile caches. In *2nd workshop on interactions of NVM/flash with operating systems and workloads (INFLOW 14)*.

Mittal, S. and J. S. Vetter (2015). Equalwrites: Reducing intra-set write variations for enhancing lifetime of non-volatile caches. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems 24*(1), 103–114.

Mittal, S., J. S. Vetter, and D. Li (2014). Writesmoothing: Improving lifetime of non-volatile caches using intra-set wear-leveling. In *Proceedings of the 24th edition of the great lakes symposium on VLSI*, pp. 139–144.

Mittal, S., J. S. Vetter, and D. Li (2015). A survey of architectural approaches for managing embedded dram and non-volatile on-chip caches. *IEEE Transactions on Parallel and Distributed Systems 26*(6), 1524–1537.

Mittal, S., R. Wang, and J. Vetter (2017). Destiny: A comprehensive tool with 3d and multi-level cell memory modeling capability. *Journal of Low Power Electronics and Applications 7*(3).

Mondal, D., A. Singh, S. Bhatt, and R. Mishra (2023). Hybrid spin-orbit torque/spin-transfer torque-based multibit cell for area-efficient magnetic random access memory. *IEEE Transactions on Electron Devices 70*(12), 6318–6323.

Oboril, F., R. Bishnoi, M. Ebrahimi, and M. B. Tahoori (2015). Evaluation of hybrid memory technologies using sot-mram for on-chip cache hierarchy. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 34*(3), 367–380.

Oh, B., N. Abeyratne, N. S. Kim, J. Ahn, R. G. Dreslinski, and T. Mudge (2023). Rethinking dram's page mode with stt-mram. *IEEE Transactions on Computers 72*(5), 1503–1517.

Peng, I. B., K. Wu, J. Ren, D. Li, and M. B. Gokhale (2020). Demystifying the performance of hpc scientific applications on nvm-based memory systems. *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 916–925.

Pentecost, L., A. Hankin, M. Donato, M. Hempstead, G.-Y. Wei, and D. Brooks (2022). Nvmexplorer: A framework for cross-stack comparisons of embedded non-volatile memories. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 938–956.

Poremba, M. and Y. Xie (2012). Nvmain: An architectural-level main memory simulator for emerging non-volatile memories. In *2012 IEEE Computer Society Annual Symposium on VLSI*, pp. 392–397.

Prenat, G., K. Jabeur, P. Vanhauwaert, G. D. Pendina, F. Oboril, R. Bishnoi, M. Ebrahimi, N. Lamard, O. Boulle, K. Garello, J. Langer, B. Ocker, M.-C. Cyrille, P. Gambardella, M. Tahoori, and G. Gaudin (2016). Ultra-fast and high-reliability sot-mram: From cache replacement to normally-off computing. *IEEE Transactions on Multi-Scale Computing Systems 2*(1), 49–60.

Rai, S., K. H. D, and B. Talawar (2022). Sot-mram based main memory: An alternative to dram. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–6.

Saha, R., Y. P. Pundir, and P. K. Pal (2022). Comparative analysis of stt and sot based mrams for last level caches. *Journal of Magnetism and Magnetic Materials 551*, 169161.

Santhalia, D. G. and M. R. Dahiya (2021). Sot-MRAM Research Overview Challenges and Current Market Trends. [Online; accessed 2023-04-05].

Sarkar, A., N. Singh, V. Venkitaraman, and V. Singh (2021). Dam: Deadblock aware migration techniques for stt-ram-based hybrid caches. *IEEE Computer Architecture Letters 20*(1), 62–4.

Schwierz, F. and J. J. Liou (2020). Status and future prospects of cmos scaling and moore's law - a personal perspective. In *2020 IEEE Latin America Electron Devices Conference (LAEDC)*, pp. 1–4.

Seo, Y. and K.-W. Kwon (2020a). Area-optimized design of sot-mram. *IEICE Electronics Express 17*(21), 20200314–20200314.

Seo, Y. and K.-W. Kwon (2020b). Area-optimized design of sot-mram. *IEICE Electronics Express 17*(21), 20200314–20200314.

Shao, Q., P. Li, L. Liu, H. Yang, S. Fukami, A. Razavi, H. Wu, K. Wang, F. Freimuth, Y. Mokrousov, M. D. Stiles, S. Emori, A. Hoffmann, J. Åkerman, K. Roy, J.-P. Wang, S.-H. Yang, K. Garello, and W. Zhang (2021). Roadmap of spin–orbit torques. *IEEE Transactions on Magnetics 57*(7), 1–39.

Singh, I., B. Raj, M. Khosla, and B. K. Kaushik (2020). Comparative analysis of spintronic memories for low power on-chip caches. *SPIN 10*(04), 2050027.

Sivakumar, S. and J. Jose (2023). Self adaptive logical split cache techniques for delayed aging of nvm llc. *ACM Transactions on Design Automation of Electronic Systems 28*(6), 1–24.

Sivakumar, S., M. Mannampalli, and J. Jose (2023). Enhancing lifetime of non-volatile memory caches by write-aware techniques. In *Emerging Electronic Devices, Circuits and Systems: Select Proceedings of EEDCS Workshop Held in Conjunction with ISDCS 2022*, pp. 109–123. Springer.

Smith, A. J. (1982, sep). Cache memories. *ACM Comput. Surv. 14*(3), 473–530.

SPEC. Spec2006 and spec2017. `https://www.spec.org/cpu2017/,https://www.spec.org/cpu2006/`. [Accessed 20-06-2024].

Sura, A. and V. Nehra (2021). Performance comparison of single level stt and sot mram cells for cache applications. In *2021 25th International Symposium on VLSI Design and Test (VDAT)*, pp. 1–4.

Van Beek, S., K. Cai, F. Yasin, H. Hody, G. Talmelli, V. Nguyen, N. F. Vergel, A. Palomino, A. Trovato, K. Wostyn, S. Rao, G. Kar, and S. Couet (2023). Scaling the sot track – a path towards maximizing efficiency in sot-mram. In *2023 International Electron Devices Meeting (IEDM)*, pp. 1–4.

Wang, C., Z. Wang, B. Wu, and W. Zhao (2019). Design and optimization of an area-efficient sot-mram. *2019 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, 1–3.

Wang, G., Y. Zhang, J. Wang, Z. Zhang, K. Zhang, Z. Zheng, J.-O. Klein, D. Ravelosona, Y. Zhang, and W. Zhao (2019). Compact modeling of perpendicular-magnetic-anisotropy double-barrier magnetic tunnel junction with enhanced thermal stability recording structure. *IEEE Transactions on Electron Devices 66*(5), 2431–2436.

Wang, J., X. Dong, and Y. Xie (2014). Enabling high-performance lpddrx-compatible mram. *2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 339–344.

Wang, J., X. Dong, Y. Xie, and N. P. Jouppi (2013). *i*^{2}*wap*: Improving non-volatile cache lifetime by reducing inter-and intra-set write variations. In *2013 IEEE 19th international symposium on high performance computer architecture (HPCA)*, pp. 234–245. IEEE.

Wu, B., C. Wang, Z. Wang, Y. Wang, D. Zhang, D. Liu, Y. Zhang, and X. S. Hu (2020a). Field-free 3t2sot mram for non-volatile cache memories. *IEEE Transactions on Circuits and Systems I: Regular Papers 67*(12), 4660–4669.

Wu, B., C. Wang, Z. Wang, Y. Wang, D. Zhang, D. Liu, Y. Zhang, and X. S. Hu (2020b). Field-free 3t2sot mram for non-volatile cache memories. *IEEE Transactions on Circuits and Systems I: Regular Papers 67*(12), 4660–4669.

Yang, X., Y. Zhang, Y. Zhang, and P. Wang (2022). A universal compact model for spin-transfer torque-driven magnetization switching in magnetic tunnel junction. *IEEE Transactions on Electron Devices 69*(11), 6453–6458.

Zhang, Y., W. Zhao, Y. Lakys, J.-O. Klein, J.-V. Kim, D. Ravelosona, and C. Chappert (2012). Compact modeling of perpendicular-anisotropy cofeb/mgo magnetic tunnel junctions. *IEEE Transactions on Electron Devices 59*(3), 819–826.

Zheng, Z., Y. Zhang, V. Lopez-Dominguez, L. Sánchez-Tejerina, J. Shi, X. Feng, L. Chen, Z. Wang, Z. Zhang, K. Zhang, et al. (2021). Field-free spin-orbit torque-induced switching of perpendicular magnetization in a ferrimagnetic layer with a vertical composition gradient. *Nature communications 12*(1), 4555.

Zitong Zhang, Wenjie Wang, P. Y. and Y. Jiang (2022). Cache performance of nv-stt-mram with scale effect and comparison with sram. *International Journal of Electronics 109*(3), 391–409.

List of Publications

Journal Publications

- Kallinatha H D, Sadhana Rai, Basavaraj Talawar. "A Detailed Study of SOT-MRAM as an Alternative to DRAM Primary Memory in Multi-Core Environment" in IEEE Access, vol. 12, pp. 7224-7243, 2024, doi: 10.1109/ACCESS.2024.3352151.***[SCI/Q1 Indexed]***

- Kallinatha HD, Basavaraj Talawar, "A Multi-Factor Scaling Framework for Efficient SOT-MRAM Cache Design and Lifetime Extension ", Elsevier Performance Evaluation Journal. *[Under review]* ***[SCI Indexed]***

- Kallinatha HD, Basavaraj Talawar, "An End-to-end Model for SOT-MRAM Scaling with Write Variation and Lifetime extension of Physically Split Cache", IETE Journal of Research *[Under review]****[SCI Indexed]***

Conference Publications

- H. D. Kalllinatha and Basavaraj Talawar, "Spintronics Main Memory Alternative to DRAM with Reliable Simulations", **HiPC 2023** Student Research Symposium (HiPC SRS 2023)," 2023 IEEE 30th International Conference on High Performance Computing, Data and Analytics Workshop (HiPCW), Goa, India, 2023, pp. 79, doi: 10.1109/HiPCW61695.2023.00021.***[Scopus indexed]***

- Kallinatha HD and Basavaraj Talawar,"Comparative Analysis of Non-Volatile Memory On-chip Caches", Presented in International Conference on Applied Computational Intelligence and Analytics, on February 26th -27th organized by NIT Raipur, AIP Conf. Proc. 16 June 2023; 2705 (1): 040008.doi:10.1063/5.0133350 ***[Scopus indexed]***

- S. Rai, Kallinatha H. D and B. Talawar, "SOT-MRAM Based Main Memory: An Alternative to DRAM," 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2022,pp.1-6,doi: 10.1109/CONECCT55679.2022.9865703.***[Scopus Indexed]***

- Kallinatha HD and Basavaraj Talawar," A Framework for SOT-MRAM Scaling Road-map with Density and Application Evaluation",4th IEEE International Conference on Computer Systems will be held in Hangzhou, China.*[accepted and presented]*

- Kallinatha HD and Basavaraj Talawar, "SOT-MRAM-Based Devices in Memory Hierarchy for Next-Generation Computing",29th Springer International Conference on Advanced Computing and Communications(ADCOM 2024) @ PhD Forum will be held in IIIT Bangalore, India.*[Under review]*

CURRICULUM VITAE

Kallinatha H D

kallinatha@gmail.com @KalliKalli1984

kallinathahd https://sites.google.com/site/kallinatha/home

Personal Details

Name:	Kallinatha Harogere Doddaiah
Date of Birth:	15/12/1984
Address for Communication:	"Manogna",1st Main Road,3rd Cross ,1st Parallel Road, Behind Dattatreya Temple,Jayanagar West, Tumakuru, Karnataka, 572102, India
Phone:	+91-9986295070
Email:	kallinatha@gmail.com,hdk@sit.ac.in

Employment History

2010 – Till Date **Assistant Professor**
Dept. of CSE at Siddaganaga Institute of Technology (SIT), Tumkur, Karnataka, India.

2009 – 2010 **Lecturer and Senior Lecturer**
CSE Department, SIET Tumkur, Karnataka, India.

Education

2007 – 2009 **M.Tech. Computer Science and Engineering, VTU Belagavi**
Thesis title: *High Dimensional Databases for Web Recommendation Systems*.

2002 – 2006 **B.E. Computer Science and Engineering, VTU Belagavi**
Thesis title: *Voice Enabled ATM for Physically and Visually Challenged*.

Awards and Achievements

2013 & 2018 **SITAA Golden Jubilee Award for Teachers & QIP Scholarship by AICTE,New Delhi**.

Professional

- **Services:** TPC member and Reviewer for several international conferences and journals.
- **Memberships in:**The Internet Society(ISOC), Cryptology Research Society of India, International Association of Engineers (IAENG), Hong Kong, The Institution of Engineers (India), IEEE, ACM and ACCS, IISC.