EFFICIENT BYPASS POLICIES FOR NON-VOLATILE SHARED CACHES

ARITRA BAGCHI

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
APRIL 2025

EFFICIENT BYPASS POLICIES FOR NON-VOLATILE SHARED CACHES

by

ARITRA BAGCHI

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2025

DEDICATED TO

My parents Pradip and Rita

Certificate

This is to certify that the thesis titled "**Efficient Bypass Policies for Non-volatile Shared Caches**" being submitted by **Aritra Bagchi** for the award of **Doctor of Philosophy** in Computer Science and Engineering is a record of bonafide work carried out by him under my guidance and supervision in the Department of Computer Science and Engineering, Indian Institute of Technology Delhi. The work presented in this dissertation has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly.

Certain works included in this dissertation involved collaboration or use of measurement data obtained by other students, which have been explicitly specified/acknowledged in the corresponding chapters and the part done by those collaborators appeared in their respective reports/theses.

Preeti Ranjan Panda
Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi
New Delhi–110016

Acknowledgments

"In the middle of winter I at last discovered that there was in me an invincible summer."

—Albert Camus, *Return to Tipasa*

First and foremost, I would like to express my deepest gratitude to my advisor, **Prof. Preeti Ranjan Panda**. During challenging times, especially during my early Ph.D. years when positive outcomes were rare and setbacks frequent, his constant encouragement kept me going. In those years of despair, it was his optimism that helped me push through, motivating me to stay committed to my research. He always made himself available for detailed research discussions. I cannot recall a single instance when I wanted to discuss research and he was not willing to engage. He created opportunities for me to review research papers for top conferences and journals. This review exposure has been instrumental in my growth as a young researcher. He also has been proactive in creating opportunities for me to discuss my research with leading experts from academia and industry. Perhaps, in a book on what an ideal Ph.D. advisor should do, all these would be listed. But I am not so naïve as to take for granted the difference between what one is supposed to do and what one actually does. In addition to being a technical mentor, he provided excellent insights into many other aspects of academic and non-academic life. Thanks, **Preeti**, for being the go-to person for so many of my problems—technical, administrative, financial, and more. I would also like to express my sincere gratitude to my research committee—**Prof. Sanjiva Prasad**, **Prof. Kolin Paul**, and **Prof. Manan Suri** for providing regular feedback on my research.

I extend heartfelt gratitude to my collaborators and co-authors—**Dinesh Joshi**, **Garima Modi**, **Dharamjeet**, **Ohm Rishabh**—whose contributions have been instrumental. I would like to thank my senior colleagues, **Sakshi Tiwari**, **Lokesh Siddhu**, and **Rajesh Kedia**, for the insights and advices they provided especially during my early days when those were very relevant. I am fortunate to be part of such a vibrant research group as MARG (Memory & Embedded Architectures Research Group), where both technical and non-technical discussions are always just a call or a desk away. Thanks to **Shailja Pandey**, **Garima Modi**, **Ayushi Agarwal**, **Divya Praneetha Ravipati**, **Naman Jain**, **Dinesh Joshi**, **Sakshi Tiwari**, **Lokesh Siddhu**, **Rajesh Kedia**, **Dharamjeet**, and many others for our countless enriching conversations. Special thanks to my friends **Vipul Rathore**, **Arkajyoti De**, **Soumen Basu**, **Prashant Agarwal**, and **Vijay Kumar** for their exuberant companies throughout the journey. I deeply cherished the post-dinner walks with Vipul and Soumen around the IIT Delhi campus—a lifeline during the intense PhD journey, filled with conversations that ranged from dilemmas to resolutions.

Family plays a crucial role in the life of a Ph.D. student. My own journey had many ups and downs. The initial phase of my Ph.D. was particularly difficult—a prolonged period of struggle without concrete outputs, which felt endless and left me emotionally drained. I consider myself truly fortunate to have had my family by my side through it all. Above all, I owe everything to my parents, **Pradip** and **Rita**. No words can truly capture the depth of my gratitude. Constant support and encouragement from my parents and my younger brother **Anamitra** have been the foundation of my journey. It is their belief in me that gave me the resilience to pursue higher studies and research. Thank you, **Samapti**—my wife, and companion for the past decade—for standing by me unwaveringly. I have always admired—and admittedly envied—your invincible optimism. At times, that very optimism became the anchor I needed. Besides, you understood my silences and tolerated my anxieties. I simply could not have made it through this journey without the love and support of these remarkable people.

My old friends from school and undergraduate days have also been a source of encouragement. So, I would rather not miss this opportunity to acknowledge their presence and impact here—**Soumik Dutta**, **Samrat Sarkar**, **Soumyadeep Pal**, **Sayan Dhara**, **Snehasish Das**, **Sushobhan Chatterjee**, **Arghyadeep Bose**, **Sahil Ali**, **Rajdeep Bhaduri**, **Hironmoy Malakar**, **Sayan Chowdhury**, **Sayan Chakraborty**, **Sounak Biswas**, **Tanmay Sarkar**, **Saikat Dutta**, **Sourav Biswas**, **Ranojit Biswas**, **Prasad Ranjan Ghosh**, **Amitava Biswas**, and many others.

Our lab and office administrators—Ms. Vandana, Ms. Rekha, Mr. Hemant, Mr. Suresh, Mr. Rajesh, Ms. Manju, and many others—have worked efficiently behind the scenes to provide crucial logistical support.

I apologize to and thank all those who have helped me, but their names have inadvertently been missed.

Finally, I bow down to the mysterious One for bestowing love and blessings upon me. After all, who else can show an invincible summer within, in the middle of winter?

Camus will perhaps forgive me for interpreting his words in a spiritual way!

Aritra Bagchi

Abstract

Multi-Processor Systems-on-Chip (MPSoCs) have become increasingly popular for high-performance and energy-efficient computing, especially with a growing emphasis on AI/ML, multimedia, and real-time applications. To accommodate the demands of modern workloads, contemporary multi-core systems are integrating an increasing number of compute resources, i.e., cores and accelerators. However, the Achilles heel of modern processors and SoCs, both in terms of performance and energy, lies in the off-chip data communications. In contemporary systems, caches are typically implemented with SRAM, which encounters major challenges (e.g., high leakage power) against the technology scaling required to meet the demands of modern workloads. Hence, non-volatile memory (NVM) technologies have gained considerable research attention due to their intrinsically higher storage density and substantially lower leakage power consumption over traditional technologies like SRAM and DRAM. However, certain unique characteristics of NVMs, such as high write overheads and limited endurance, present significant challenges to their integration into modern LLCs.

Processor cores send read requests to the shared LLC, and if the requested data is not present within the LLC, the off-chip main memory is accessed and the corresponding responses are returned to the LLC. The LLC controller forwards a copy of the response to the requesting core (referred to as response forwarding) to maintain the progress of that core while retaining the original response for subsequent LLC writes. The LLC also receives the dirty cache lines evicted from higher-level caches as writebacks for corresponding cache data updates. With increasing numbers of cores and accelerators being integrated in MPSoCs, requests, responses, and writebacks from various cores (and caches) compete for the limited available bandwidth of the shared LLC. This contention creates a significant performance bottleneck even for conventional SRAM LLCs and is aggravated in the context of NVM LLCs, where cache write operations (corresponding to writebacks and responses) are considerably slower than reads. While the contention for the shared LLC capacity has been extensively addressed in prior works, the system-level implications of LLC bandwidth contention remain largely unexplored.

To first alleviate the bandwidth contention for conventional LLCs, we propose a novel LLC controller policy called **COBRRA** (COntention-aware cache Bypass with Request-Response Arbitration), which aggressively bypasses responses from main memory while efficiently arbitrating between requests from different cores and their responses to enhance the overall system throughput. Our novel bypass policy exploits an interesting trade-off between LLC data reuse and contention and improves performance by mitigating contention, even at the expense of some reuse. While the data reuse aspect is well studied by prior cache bypass strategies, the ramification of LLC bandwidth contention is largely overlooked by prior policies, and COBRRA, therefore, makes an attempt to bridge this key research gap.

In our subsequent work, we introduce a controller for NVM LLC named **POEM** (Performance Optimization and Endurance Management for Non-volatile Caches), where we employ aggressive bypass strategies for different sources of NVM writes: writebacks from higher-level caches and responses from main memory. While existing bypass strategies prioritize cache data reuse, advocating more frequent cache writes, POEM aggressively bypasses NVM writes to effectively mitigate the impact of LLC contention on critical reads while also leveraging LLC data reuse through highly selective cache writes. We demonstrate that such an aggressive bypass is beneficial for overall system performance, even if it comes at the expense of some cache reuse. The aggressive bypassing of NVM writes also favors NVM endurance, with an overall reduction in the LLC write stress. In this work, we also propose an efficient controller policy to migrate data between hot and cold cache lines to achieve a more balanced write distribution, thereby improving NVM LLC lifespan and overall system reliability. Because such migrations interfere with normal LLC accesses, POEM exercises careful control over data migration. Additionally, POEM, as a comprehensive policy, offers a balance between the two crucial system-level objectives: enhancing overall system performance and improving NVM cache endurance, allowing assignment of different priorities to these objectives through parameterized control of a threshold.

While aggressive bypassing of NVM writes can enhance overall system performance, it negatively impacts off-chip memory energy consumption by sacrificing too much cache reuse to mitigate NVM LLC contention. System-level energy efficiency is crucial in modern computing paradigms, and off-chip data movements contribute significantly to the total system energy. The static and refresh components of this energy are closely tied to overall system performance, while the dynamic component is influenced by the volume of off-chip memory traffic, which, in turn, is controlled by LLC data reuse. Aggressive bypassing of NVM writes can increase memory dynamic energy to the point where it outweighs reductions in other energy components, demonstrating that performance-optimal solutions may not always be

energy-optimal. In our next work, we introduce the **NOVELLA** (Non-Volatile Last-Level Cache Bypass for Optimizing Off-chip Memory Energy) to exploit various trade-offs between different components of the off-chip memory energy. By intelligently bypassing different sources of NVM writes, NOVELLA aims to reduce the overall off-chip memory energy consumption, facilitating next-generation processors and SoCs scale against the memory power wall.

सारांश

मल्टी-प्रोसेसर सिस्टम-ऑन-चिप (**MPSoCs**) उच्च-प्रदर्शन और ऊर्जा-कुशल कंप्यूटिंग के लिए तेजी से लोकप्रिय हो गए हैं, खासकर **AI/ML**, मल्टीमीडिया और वास्तविक समय के अनुप्रयोगों पर बढ़ते जोर के साथ। आधुनिक कार्यभार की मांगों को पूरा करने के लिए, समकालीन मल्टी-कोर सिस्टम बढ़ती संख्या में कंप्यूट संसाधनों, यानी कोर और एक्सेलेरेटर को एकीकृत कर रहे हैं। हालाँकि, आधुनिक प्रोसेसर और **SoCs** की कमजोरी, प्रदर्शन और ऊर्जा दोनों के मामले में, ऑफ-चिप डेटा संचार में निहित है। समकालीन प्रणालियों में, कैश को आमतौर पर **SRAM** के साथ लागू किया जाता है, जो आधुनिक कार्यभार की मांगों को पूरा करने के लिए आवश्यक प्रौद्योगिकी स्केलिंग के खिलाफ बड़ी चुनौतियों (जैसे, उच्च रिसाव शक्ति) का सामना करता है। इसलिए, गैर-वाष्पशील मेमोरी (**NVM**) तकनीकों ने **SRAM** और **DRAM** जैसी पारंपरिक तकनीकों की तुलना में उनके आंतरिक रूप से उच्च भंडारण घनत्व और काफी कम रिसाव बिजली की खपत के कारण काफी शोध ध्यान आकर्षित किया है। हालाँकि, **NVM** की कुछ विशिष्ट विशेषताएँ, जैसे उच्च लेखन ओवरहेड्स और सीमित धीरज, आधुनिक **LLC** में उनके एकीकरण के लिए महत्वपूर्ण चुनौतियाँ प्रस्तुत करती हैं।

प्रोसेसर कोर साझा एलएलसी को रीड रिक्वेस्ट भेजते हैं, और यदि अनुरोधित डेटा एलएलसी के भीतर मौजूद नहीं है, तो ऑफ-चिप मुख्य मेमोरी तक पहुंच बनाई जाती है और संबंधित प्रतिक्रियाएं एलएलसी को वापस कर दी जाती हैं। एलएलसी नियंत्रक बाद के एलएलसी लेखन के लिए मूल प्रतिक्रिया को बनाए रखते हुए उस कोर की प्रगति को बनाए रखने के लिए अनुरोध करने वाले कोर (जिसे प्रतिक्रिया अग्रेषण के रूप में जाना जाता है) को प्रतिक्रिया की एक प्रति अग्रेषित करता है। एलएलसी को संबंधित कैश डेटा अपडेट के लिए राइटबैक के रूप में उच्च-स्तरीय कैश से निकाली गई गंदी कैश लाइनें भी प्राप्त होती हैं। **MPSoCs** में एकीकृत किए जा रहे कोर और एक्सेलेरेटर की बढ़ती संख्या के साथ, विभिन्न कोर (और कैश) से अनुरोध, प्रतिक्रियाएं और राइटबैक साझा **LLC** की सीमित उपलब्ध बैंडविड्थ के लिए प्रतिस्पर्धा करते हैं। यह विवाद पारंपरिक एसआरएएम एलएलसी के लिए भी एक महत्वपूर्ण प्रदर्शन बाधा पैदा करता है और एनवीएम एलएलसी के संदर्भ में यह बढ़ गया है, जहां कैश लिखने का संचालन (राइटबैक और प्रतिक्रियाओं के अनुरूप) पढ़ने की तुलना में काफी धीमा है। जबकि साझा एलएलसी क्षमता के विवाद को पिछले कार्यों में बड़े पैमाने पर संबोधित किया गया है, एलएलसी बैंडविड्थ विवाद के सिस्टम-स्तरीय निहितार्थ काफी हद तक अस्पष्ट हैं।

सबसे पहले पारंपरिक एलएलसी के लिए बैंडविड्थ विवाद को कम करने के लिए, हम एक नई एलएलसी नियंत्रक नीति का प्रस्ताव करते हैं जिसे **COBRRA** (अनुरोध-प्रतिक्रिया मध्यस्थता के साथ सामग्री-जागरूक कैश बाईपास) कहा जाता है, जो अलग-अलग अनुरोधों के बीच कुशलतापूर्वक मध्यस्थता करते हुए मुख्य मेमोरी से प्रतिक्रियाओं को आक्रामक रूप से बायपास करता है। समग्र सिस्टम थ्रूपुट को बढ़ाने के लिए कोर और उनकी प्रतिक्रियाएँ। हमारी नई बाईपास नीति एलएलसी डेटा के पुन: उपयोग और विवाद के बीच एक दिलचस्प व्यापार-बंद का फायदा उठाती है और कुछ पुन: उपयोग की कीमत पर भी विवाद को कम करके प्रदर्शन में सुधार करती है। जबकि डेटा पुन: उपयोग पहलू का पूर्व कैश बायपास रणनीतियों द्वारा अच्छी तरह से अध्ययन किया गया है, एलएलसी बैंडविड्थ विवाद के प्रभाव को पूर्व नीतियों द्वारा काफी हद तक नजरअंदाज कर दिया गया है, और इसलिए, **COBRRA** इस प्रमुख अनुसंधान अंतर को पाटने का प्रयास करता है।

हमारे बाद के काम में, हम एनवीएम एलएलसी के लिए **POEM** (गैर-वाष्पशील कैश के लिए प्रदर्शन अनुकूलन और सहनशक्ति प्रबंधन) नाम से एक नियंत्रक पेश करते हैं, जहां हम एनवीएम लेखन के विभिन्न स्रोतों के लिए आक्रामक बाईपास रणनीतियों को नियोजित करते हैं: उच्च-स्तरीय कैश से राइटबैक और मुख्य मेमोरी से प्रतिक्रियाएँ। जबकि मौजूदा बाईपास रणनीतियाँ कैश डेटा के पुन: उपयोग को प्राथमिकता देती हैं, अधिक बार कैश लिखने की वकालत करती हैं, पीओईएम आक्रामक रूप से महत्वपूर्ण रीड्स पर एलएलसी विवाद के प्रभाव को कम करने के लिए एनवीएम राइट्स को बायपास करता है, जबकि अत्यधिक चयनात्मक कैश राइट्स के माध्यम से एलएलसी डेटा पुन: उपयोग का लाभ उठाता है। हम प्रदर्शित करते हैं कि इस तरह का आक्रामक बाईपास समग्र सिस्टम प्रदर्शन के लिए फायदेमंद है, भले ही यह कुछ कैश पुन: उपयोग की कीमत पर आता हो। एनवीएम राइट्स को आक्रामक तरीके से दरकिनार करने से एलएलसी राइट्स तनाव में समग्र कमी के साथ एनवीएम सहनशक्ति को भी बढ़ावा मिलता है। इस कार्य में, हम अधिक संतुलित लेखन वितरण प्राप्त करने के लिए गर्म और ठंडे कैश लाइनों के बीच डेटा को स्थानांतरित करने के लिए एक कुशल नियंत्रक नीति का भी प्रस्ताव करते हैं, जिससे एनवीएम एलएलसी जीवनकाल और समग्र सिस्टम विश्वसनीयता में सुधार होता है। क्योंकि इस तरह के माइग्रेशन सामान्य एलएलसी एक्सेस में बाधा डालते हैं, पीओईएम डेटा माइग्रेशन पर सावधानीपूर्वक नियंत्रण रखता है। इसके अतिरिक्त, पीओईएम, एक व्यापक नीति के रूप में, दो महत्वपूर्ण सिस्टम-स्तरीय उद्देश्यों के बीच संतुलन प्रदान करता है: समग्र सिस्टम प्रदर्शन को बढ़ाना और एनवीएम कैश सहनशक्ति में सुधार करना, एक सीमा के पैरामीटरयुक्त नियंत्रण के माध्यम से इन उद्देश्यों के लिए विभिन्न प्राथमिकताओं को आवंटित करने की अनुमति देना।

जबकि एनवीएम राइट्स की आक्रामक बाईपासिंग समग्र सिस्टम प्रदर्शन को बढ़ा सकती है, यह एनवीएम एलएलसी विवाद को कम करने के लिए बहुत अधिक कैश पुन: उपयोग का त्याग करके ऑफ-चिप मेमोरी ऊर्जा खपत को नकारात्मक रूप से प्रभावित करती है। आधुनिक कंप्यूटिंग प्रतिमानों में सिस्टम-स्तरीय ऊर्जा दक्षता महत्वपूर्ण है, और ऑफ-चिप डेटा मूवमेंट कुल सिस्टम ऊर्जा में महत्वपूर्ण योगदान देता है। इस ऊर्जा के स्थिर और ताज़ा घटक समग्र सिस्टम प्रदर्शन से निकटता से जुड़े हुए हैं, जबकि गतिशील घटक ऑफ-चिप मेमोरी ट्रैफ़िक की मात्रा से प्रभावित होता है, जो बदले में एलएलसी डेटा पुन: उपयोग द्वारा नियंत्रित होता है। एनवीएम राइट्स को आक्रामक रूप से दरकिनार करने से मेमोरी डायनेमिक ऊर्जा उस बिंदु तक बढ़ सकती है जहां यह अन्य ऊर्जा घटकों में कटौती से अधिक है, यह दर्शाता है कि प्रदर्शन-इष्टतम समाधान हमेशा ऊर्जा-इष्टतम नहीं हो सकते हैं। अपने अगले काम में, हम ऑफ-चिप मेमोरी ऊर्जा के विभिन्न घटकों के बीच विभिन्न ट्रेड-ऑफ का फायदा उठाने के लिए **NOVELLA** (ऑफ-चिप मेमोरी एनर्जी को अनुकूलित करने के लिए गैर-वाष्पशील अंतिम-स्तर कैश बाईपास) पेश करते हैं। एनवीएम के विभिन्न स्रोतों को समझदारी से दरकिनार करके, नोवेल्ला का लक्ष्य समग्र ऑफ-चिप मेमोरी ऊर्जा खपत को कम करना है, जिससे मेमोरी पावर वॉल के खिलाफ अगली पीढ़ी के प्रोसेसर और **SoCs** स्केल की सुविधा मिलती है।

Contents

List of Figures

List of Tables