<u>Title:</u> Efficient bypass policies for non-volatile shared caches

<u>Author</u>: Aritra Bagchi (2018CSZ8295)
Dept. of Computer Science and Engineering

<u>Abstract:</u>

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.

Viva details:
Date: Monday, 7 April 2025
Time: 10:00-11:00 AM
Location: Khosla School of IT, Mathur Seminar Hall, SIT 001 (Ground Floor)

Photograph: