INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

LiNoVo: Longevity Enhancement of Non-Volatile Caches by Placement, Write-Restriction & Victim Caching in Chip Multi-Processors

by

Sukarn Agarwal

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Department of Computer Science and Engineering

Under the supervision of

Prof. Hemangee K. Kapoor

March 2020

Abstract

The ever increasing demand for higher processing speed with hiked data parallelism force the computer architects to increase the number of processing cores on a single chip called Chip-Multi-processors (CMPs). Towards meeting the performance goals, these CMPs are equipped with larger on-chip Last Level Caches (LLCs) to enhance the probability of the presence of data on-chip during process execution. The existing literature portrays that conventional LLCs built-in charge-based memory technologies are although faster but fall short in fulfilling these demands, especially in terms of increased power consumption. Furthermore, stagnation in process technology drove memory architects and researchers towards investigating Non-Volatile Memories (NVMs) for designing on-chip LLCs due to their promising scalability, reduced leakage power consumption, and compatibility with the conventional CMOS. However, many of these NVMs suffer from costly write operations with lower write endurance.

To achieve the best of both conventional as well as emerging NVMs, Hybrid Cache Architecture (HCA) has been evolved where different memory technologies are fabricated to build up a single level of cache. In particular, in this thesis, we adopt HCA based LLC, in which a large portion of LLC is built-in NVM for stimulating energy efficiency and the remaining smaller part is engineered with the conventional faster SRAM. In such an HCA, the block placement to the appropriate region is the key challenge from the energy-efficiency perspective. Towards this, we proposed a private block-based block placement technique that allocates dataless entries in the NVM portion of HCA. In this approach, additional savings in the number of writes to the NVM portion are governed by employing a Reuse Distance Aware Write Intensity Predictor. Besides the block placement approach, the fields of the predictor are used to improve the victim replacement decision for different portions of HCA. From this contribution, we get 34.5% reduction in writes and $16-19.6\%$ savings in energy over prior works. Towards a performance perspective, to overcome the effect of costly write latency operations, in the next contribution, the victim cache is explored with pure NVM and HCA based cache. With NVM cache, the victim cache is used to retain both victims as well as the write-intensive live blocks to save on the time to exchange and subsequent slow writes. By experimental analysis, we achieved 5.88% speedup over the baseline. With HCA, two policies are proposed to manage the victim cache effectively, where former one decides the placement of the block upon a victim cache hit in the different regions of HCA, whereas latter one gives a substantial amount of space

for the victims evicted from each region using dynamic region-based victim cache partitioning. These couple of approaches improve the overall performance of HCA by 4.43% and reduce the miss rate by 7.81%.

According to the available literature, due to lower write endurance, the lifetime of NVM caches is limited. Additionally, the run-time behavior of the applications, working set sizes, and cache replacement policies altogether lead to write variations across and inside the sets in the cache. In that, some sets and ways (inside the same set) get written heavily compared to others which are termed as inter-set, and intra-set write variations, respectively. These variations are one of the biggest design concern for HCA as they further limit the longevity and lifetime of its NVM-portion. To mitigate these unwanted write variation, two wear-leveling techniques: inter-set and intra-set are further proposed. The intra-set wear leveling works on the basic concept of the write restriction by partitioning the cache horizontally and vertically and is able to reduce the intra-set write variation in the range of $80-86.5\%$ with 7.27 times improvement in lifetime over the prior works. The inter-set wear-leveling technique exploits the concept of fellow sets and the dynamic associativity management to overcome the write variation across the cache set. With these approaches, write variation is reduced by $27.6-34\%$, and the lifetime is further improved by $14.7-20.7\%$ over the baseline.

The thesis has thus demonstrated the effective management techniques for longevity enhancement of the NVM cache for an optimal lifetime and controlling the effect of costly write operations.