BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK
Computer Science

COMPUTER SCIENCE RESEARCH SEMINAR

Multi-Block Data Compression Techniques for Multi-Core Systems

Madhu Mutyam, Professor
Department of Computer Science, IIT - Chennai

Friday, September 27, 2019 at noon in room T-1, Engineering Building

Abstract: With limited space available on-chip, we need to fit as much data into the caches as possible, and compression is a popular choice for doing so. Compression at the last-level cache and the DRAM play an important role in improving system performance by increasing their effective capacities. A compressed block in DRAM also reduces the transfer time over the memory bus to the caches, reducing the latency of a LLC cache miss. Usually, compression is achieved by exploiting data patterns present within a block. But applications can exhibit data locality that spread across multiple consecutive data blocks. We observe that there is significant opportunity available for compressing multiple consecutive data blocks into one single block, both at the LLC and DRAM. Motivated by these observations, we propose a mechanism, namely, MBZip, that compresses multiple data blocks into one single block (called a zipped block), both at the LLC and DRAM. At the cache, MBZip includes a simple tag structure to index into these zipped cache blocks and the indexing does not incur any redirectional delay. At the DRAM, MBZip does not need any changes to the address computation logic and works seamlessly with the conventional/existing logic. MBZip is a synergistic mechanism that coordinates these zipped blocks at the LLC and DRAM. Further, we also explore silent writes at the DRAM and show that certain writes need not access the memory when blocks are zipped. MBZip improves the system performance by 21.9%, with a maximum of 90.3% on a 4-core system. To utilize the space saved by compression, we need an efficient compaction technique that fits multiple compressed blocks together inside a cache-block. An ideal compression-compaction combination should have low internal fragmentation, low access latency, and low storage overheads. We can eliminate internal fragmentation by compacting variable sized blocks into a single cache-block but at the cost of high storage overhead. Previous works that deal with the storage overhead had to settle with fixed size compaction, which wastes valuable cache space. We propose Variable Sized Cache Block Compaction (VSCC), which compacts blocks of variable sizes together and locates them with the help of compression encodings of pre-existing blocks. We introduce a novel read/write scheme and a new base-delta-immediate compression encoding, which reduce the tracking operations by 50%. VSCC reduces internal fragmentation while maintaining lower storage overhead compared to earlier works. Experimental results reveal that VSCC with a sector of size 8 outperforms the state-of-the-art techniques such as Yet Another Compressed Cache and Decoupled Compressed Cache from the performance and energy point of view.

Bio: Madhu Mutyam is a professor in the Department of Computer Science and Engineering, Indian Institute of Technology Madras. His research interests include Computer Architechture, specifically focusing on Memory hierarchy and technologies, network-on-chip, and secure processor microarchitecture.

This event is funded by GSOCS, a subsidiary of GSO, using Student Activity Fee funds

Refreshments will be provided!