INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

DOCTORAL THESIS

Content and Coherence Based Strategies for Optimizing Refreshes in Volatile Last Level Caches

Sheel Sindhu Manohar
Roll No: 156101027

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of the Indian Institute of Technology Guwahati.

Supervisor: Prof.Hemangee K. Kapoor

Department of Computer Science Engineering
Indian Institute of Technology Guwahati
Guwahati (Assam), India, Pin- 781039

December 2021

Abstract-TH-2943_156101027

Abstract

With each process generation, Moore's law offers us an exponential growth in the transistor budget on the chip. Technically, these extra transistors were used to improve processor architecture speed by adding more complicated and simple pipelines and better arithmetic and floating-point units. The futher advances included multi-core systems which demanded larger on-chip caches to support the data demands. Larger last-level caches are deployed across the chip to meet the increasing need for higher cache capacity due to included CMPs in processing cores. LLCs play an important function in the cache hierarchy by giving necessary data to hungry CMPs. SRAMs are not scalable and require advancements in power, performance, and scalability. In order to deploy massive LLCs, researchers are focusing on the construction of caches using alternative technologies that have advantages over traditional SRAM. High scalability, lower leakage power, and higher capacity in the same area footprint as SRAM are among the benefits of these technologies. However, we must investigate the best of these alternatives because they are not without flaws.

In order to control the leakage generated by SRAM based caches, researchers have explored avenues in emerging eDRAM and STTRAM based designs. These new memories come with their own challenges, in that eDRAM requires refreshing the data at regular intervals while low retention STTRAM needs restores and refreshes. This thesis aims to reduce the number of refreshes required for such type of last level caches. Our goal is to address the reliability of data over the volatile caches by ensuring the retention of data blocks. Here, we deal with volatile memory technologies' challenges and make them suitable candidates for cache hierarchy. Similarly, in STTRAM, the asymmetrical read and write with the costly write operation. Thus, it can be addressed by the relaxed versions of STTRAM.

Towards this aim, the thesis makes the following contributions: (1) In the first contribution we note that eDRAM based caches need to be refreshed at regular intervals due to their volatile nature. Here we use content based information to decide which

blocks can be avoided from refreshing. In particular, blocks has zero values need not be refreshed. (2) Second contribution uses Coherence based messages to identify the cache blocks which get loaded for getting written/updation by upper level caches. These exclusively hold block get stale eventually Copies of such private blocks in the LLC need not be refreshed as they will get updated in due course. (3) In our third contribution we deal with volatile STTRAM caches. At deep sub-micron technology these caches have read and write current values close to each other leading to read disturbance error (RDE). This makes each read destructive forcing us to restore the data after every read. We handle this by identifying read intensive blocks and moving them to a SRAM buffer. Future reads will be served from the buffer thus saving read related restores. (4) Our fourth contriution is deals with issues arising out from multi-retention STTRAM caches. STTRAM caches can be optimised to reduce their write energy by reducing their retention interval. This has disadvantage of blocks expiring at the end of retention interval. To avoid this, we use coherence based messages to identify the best partition to place a block. The idea behind this is that if the cache block is accessed before it expires we can avoid refreshing that block. All proposals have saved considerable refreshes, saved total energy as well as improved system performance.

Abstract-TH-2943_156101027