A Study on Performance and Power Efficiency of Dense Non-Volatile Caches in Multi-Core Systems

Amin Jadidi, Mohammad Arjomand, Mahmut Kandemir, and Chita Das
Pennsylvania State University

ABSTRACT

This paper presents a novel cache design based on *Multi-Level Cell Spin-Transfer Torque RAM* (MLC STT-RAM). Our design exploits the asymmetric nature of the MLC STT-RAM to build cache lines featuring heterogeneous performances, that is, half of the cache lines are read-friendly, while the other half are write-friendly – this asymmetry in read/write latencies are then used by a migration policy in order to overcome the high latency of the baseline MLC cache. Furthermore, in order to enhance the device lifetime, we propose to dynamically deactivate ways of a set in underutilized sets to convert MLC to *Single-Level Cell* (SLC) mode. Our experiments show that our design gives an average improvement of 12% in system performance and 26% in last-level cache (L3) access energy for various workloads.

ACM Reference format:

Amin Jadidi, Mohammad Arjomand, Mahmut Kandemir, and Chita Das. 2017. A Study on Performance and Power Efficiency of Dense Non-Volatile Caches in Multi-Core Systems. In *Proceedings of SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA*, 2 pages. <https://doi.org/http://dx.doi.org/10.1145/3078505.3078547>

1 INTRODUCTION

STT-RAM has zero leakage power, accommodates almost 4× more density than SRAM, and has small read access latency and high endurance (compared to other non-volatile memories). Two types of STT-RAM cell prototypes can be realized: *Single-Level Cell* (SLC) STT-RAM and *Multi-Level Cell* (MLC) STT-RAM. The SLC STT-RAM cell consists of one storage component (MTJ) which is used to store one bit information. The MLC STT-RAM device is typically composed of multiple MTJs, which are connected either serially or in parallel, and are used to store more than one bit information in a single cell. Such increased density comes at the cost of linear increase in access latency and energy consumption. For instance, the read (or write) latency and energy consumption of a 2-bit STT-RAM cell is two times higher than that of a SLC STT-RAM device under same fabrication technology. Also, MLC STT-RAM usually has lower endurance (in terms of write cycles) than SLC.

In this paper, we propose a novel 2-bit MLC STT-RAM cache design to tackle the challenges brought by an MLC STT-RAM cache. We will discuss and experimentally show that our design achieves the best features of SLC and MLC STT-RAM configurations in a unified architecture.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS '17, June 5–9, 2017, Urbana-Champaign, IL, USA

© 2017 Copyright held by the owner/author(s).

ACM ISBN ACM ISBN 978-1-4503-5032-7/17/06.

<https://doi.org/http://dx.doi.org/10.1145/3078505.3078547>

2 THE PROPOSED MLC STT-RAM CACHE

Our design is mainly based on two policies: (i) the dynamic associativity policy, and (ii) the cache line swapping policy. The former tries to reduce unnecessary write traffic to the cache and hence improve its lifetime (as well as performance), while the latter tries to improve read and write access latencies. To achieve these goals, we use the stripped data-to-cell mapping, described below. A detailed description of our proposed cache design is available at [1].

2.1 Stripped Data-to-Cell Mapping

In a 2-bit MLC device, both bits are written or read together. We call this *stacked* data-to-cell mapping, in which the latency and energy consumption of the cache accesses are roughly twice as that of a SLC. An alternative way of storing data in an MLC cell is the *stripped* mapping, which basically exploits the read and write asymmetry of the two MTJs in MLC to build a fast read and fast write storage device. These two schemes are shown in Figure 1. In a read operation, the MSB can be read from the *hard-domain* in a single read cycle. On the other hand, the cost of writing into the *soft-domain* is lower. The *stripped* mapping groups the *hard-domains* together to form *Fast Read High-Energy write* (FRHE) lines, and groups the *soft-domains* to form *Slow Read Low-Energy write* (SRLE) lines. Figure 2 compares the memory access latency (seen by a missed data in L2 in our settings) for a 4MB stripped MLC cache compared to the same sized stacked cache and a 2MB SLC cache – that shows great amount of reduction in access latencies for all kinds of workloads.

Figure 1: Stacked and stripped data-to-cell mapping..

Figure 2: Comparison of different data-to-cell mappings.

2.2 The Need for Dynamic Associativity

Memory references in general purpose applications are often non-uniformly distributed across the sets of a set-associative cache,

SIGMETRICS '17, , June 5-9, 2017, Urbana-Champaign, IL, USA Amin Jadidi, Mohammad Arjomand, Mahmut Kandemir, and Chita Das

Figure 3: Percentage of IPC improvement for the proposed cache architecture with respect to the baselines.

Figure 4: Total energy consumption of the cache architectures normalized to the SLC baselines.

Figure 5: Distribution of the missed accesses in xalancbmk.

which can lead to a high number of local conflict misses in specific sets, while other sets are underutilized. This behavior is shown in Figure 5. Moreover, these opposite-behaving sets vary from one program phase to another. As a result, our proposed architecture involves an *on-demand associativity policy* which dynamically modulates the associativity of each set. To mitigate the effects of slow reads and high-energy writes, when a cache line needs to be turned off, an FRHE and SRLE pair is merged into an SLC line.

2.3 The Need for a Cache Line Swapping Policy

To further enhance our design, we propose a swapping policy to dynamically promote write-dominated data blocks to SRLE lines (i.e., low power write operations) and read-dominated ones to FRHE lines (i.e., fast read accesses).

3 EVALUATION RESULTS

Methodology: We perform a microarchitectural simulation using the Gem5 simulator. Each core also has private L1 and L2 caches. The STT-RAM L3 cache is shared among 8 cores. We also use PARSEC-2 and SPECCPU2006 (either in rate mode or multi-programmed mode).

Performance Evaluation: Figure 3 shows an improvement of up to 29% in IPC of the high associativity caches (i.e., MLC configurations and 8 MB SLC) with respect to the 5 MB SLC baseline. Our scheme also outperforms the 8 MB stacked 2-bit cache by 10% on average thanks to being able to construct FRHE and SRLE lines. Comparing the results with the 8 MB SLC shows that, the performance of our proposed cache structure is within 5% of the performance observed for the 8 MB SLC cache.

Energy Consumption: The percentage of reduction in total memory energy is shown in Figure 4. This evaluation includes the energy consumptions of both the LLC and off-chip main memory. The energy consumption of our proposed scheme is better than the SLC baseline in applications with high and medium miss-rates due to the higher hit ratio of LLC. Besides, the energy consumption of our scheme is better than the MLC baseline with the stacked data-to-cell mapping, as it constructs lines with low write energy and trying to allocate them write-dominated blocks.

Lifetime Evaluation: In this work, we assume that reliable writes into the SLC storage is limited to 10^{12} write operations, and it is linearly scaled down for 2-bit STT-RAMs (i.e., exactly one-tenth). Overall, our proposed scheme achieves a lifetime larger than 70% of lifetime of a 5 MB SLC cache, with identical ECC strength.

4 ACKNOWLEDGMENT

This work is supported in part by NSF grants 1302557, 1213052, 1439021, 1302225, 1629129, 1526750, and 1629915, a grant from Intel.

REFERENCES

- [1] Amin Jadidi, Mohammad Arjomand, Mahmut T. Kandemir, and Chita R. Das. A Study on Performance and Power Efficiency of Dense Non-Volatile Caches in Multi-Core Systems, 2017. <http://arxiv.org/abs/1704.05044>