future cache design:STT MRAMs

k.vivek veman

October 2019

Spin-transfer torque magnetic RAM (STT MRAM) has emerged as a promising candidate for on-chip memory in future computing platforms. We present a cross-layer (device-circuit-architecture) approach to energy-efficient cache design using STT MRAM

1 INTRODUCTION

• The ever-increasing gap between processor speed and main memory latency has driven the demand for larger on-chip caches in processors. • Traditionally, on-chip caches in modern processors are implemented using static random access memories (SRAM). • However, limited scalability, susceptibility to soft errors and high leakage power of SRAM pose challenges to high-density on-chip cache implementation. In order to address the limited scalability of SRAMs, several recent processors have adopted embedded dynamic RAM (EDRAM) in lower level caches. However, vulnerability to soft errors and significant standby power of EDRAM caches due to high cell leakage are still major bottlenecks in on-chip cache design • STT MRAMs compatibility with CMOS processes makes it an attractive vehicle to realize high-density low-power embedded memories in scaled technologies

STT MRAM CACHE DESIGN

1.1 STT MRAM Preliminaries

• A conventional STT MRAM cell comprises of a magnetic tunnel junction (MTJ) and an access transistor in series (Figure 1 (a-b)). The MTJ contains a pinned layer and a free layer separated by a dielectric layer (e.g. MgO). The pinned layer has a fixed magnetization, and the free layer is programmable by changing its magnetic orientation. The resistance of the MTJ depends on the relative magnetization of the free layer with respect to the pinned layer. Parallel magnetization of the free layer with respect to the pinned layer leads to a lower resistance (RP) compared to the resistance in the anti-parallel state (RAP). The two resistances of the MTJ define the binary states of the memory cell. A read operation is performed by sensing resistance difference of the two binary states. A write operation is performed by passing a current

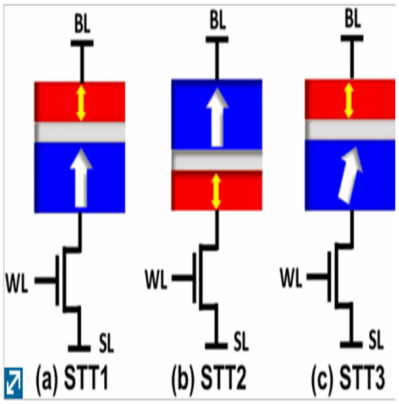

Figure 1: Schematics of an STT MRAM bitcell (a) in the standard-connected configuration (b) in the reverse-connected configuration and (c) with tilted magnetic anisotropy

(IW) through the bitcell that exceeds a critical current (IC). The direction of (IW) determines the final magnetization of the free layer (i.e., parallel or anti-parallel states of the MTJ)

1.2 STT MRAM Bitcell Design: Devices and Circuits

• Different types of MTJ stacks , and bitcell configurations provide several design choices, and can result in substantially different bitcell characteristics. Before exploring these choices, we first discuss design considerations of STT MRAM bitcells. A conventional MTJ has a large switching current density requirement, and the requirement increases dramatically with lower switching delay . The large switching current requirement for fast write operation is one of the major challenges for energy-efficient STT MRAM design. In order to address the excessive switching current requirement, an MTJ with tilted magnetic anisotropy (TMA) has been proposed in . Tilting the direction of the pinned layer, by a larger angle than what stochastic thermal noise can provide, leads to a thermal-noise-independent non-zero initial angle for precessional switching. As a result, the switching current overdrive and switching delay can be reduced significantly

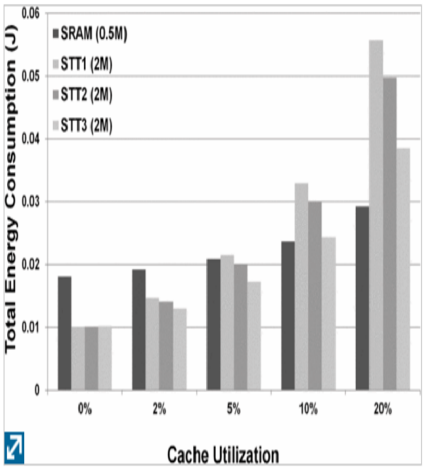

Figure 3: Total energy consumption of L2 caches at iso-area (0.5MB SRAM vs. 2MB STT MRAM)

1.3 Cache Utilization and Energy Consumption

• The contribution of active and leakage energy to total energy consumption is different for SRAM- and STT MRAM-based caches. The leakage energy in an STT MRAM cache is smaller than an SRAM cache even with 4 times larger capacity (at iso-area). On the other hand, the dynamic energy for a write operation is higher in an STT MRAM cache compared to an SRAM cache. It is important to note that the total energy dissipation in a cache depends on factors such as cache access patterns (number of read and write operations) and cache utilization (number of times a processor accesses the cache per unit cycle). The cache utilization is lower than 30 in today's processors . Moreover, for lower levels of the cache hierarchy, the cache utilization is significantly lower than 30 percentage . We have measured L2 cache utilizations for various SPEC2000 benchmarks based on the Simplescalar framework with a 32KB L1 cache configuration. Our simulation results also confirm low L2 cache utilization. For a majority of the benchmarks, L2 cache utilization is lower than 3. The highest utilization, observed for the AMMP benchmark, is about 13, and the average utilization across 16 benchmarks is only 2.2.

2 ENERGY-EFFICIENT STT MRAM CACHE DESIGN

2.1 Column Selection: SRAM vs. STT MRAM

• Since set associativity is common in modern caches, column selection in SRAM arrays is imperative. Furthermore, bit-interleaving can only be achieved by employing column selection. Bit-interleaving is a commonly adopted technique in SRAM arrays to mitigate soft errors to increase array density by bitline multiplexing . In the column selection operation of an SRAM array, all unselected bitcells in the accessed row have to be under read mode to prevent unexpected bit flips, when a wordline is asserted. This phenomenon is commonly known as pseudo read or half selection . Note that, in an STT MRAM array, the nonvolatility of bitcells can eliminate the half selection problem.

2.2 Read Energy Reduction in STT MRAM Cache

• One challenge to enable energy-efficient column selection can be to identify the selected column address during cache read operation with minimal performance penalty. We observed that the proposed technique can be easily adopted in a cache implementing sequential tag-data access. Sequential tag-data access is often employed in large, lower-level caches to improve energy-efficiency during operation . In sequential tag-data access, a cache probes the tag array first, and identifies a hit or miss. Access to the data array occurs only when there is a cache hit, and only the sub-array storing the corresponding cache line in the data array is accessed. As a result, significant energy savings can be achieved.

2.3 Write Energy Reduction in STT MRAM Cache

• Similar to the read energy reduction technique described above, improvement in write energy efficiency of STT MRAM cache can also be achieved by exploiting half-selection-free column selection. We propose partial cache line update (PLU) to reduce cache writeback energy consumption. This technique exploits data redundancy in a multi-level cache hierarchy as well as non-volatility of STT MRAM bitcells. In a writeback cache, writeback is performed when a dirty cache line in the L1 cache needs to be replaced by a new cache line. Hence, the dirty line has to be written into the L2 cache.
• Figure 6 presents the proposed PLU STT MRAM cache architecture. Each cache line is partitioned into n partial lines in order to utilize the energy-efficient column selection of STT MRAM arrays (n=4 in the given example). During writeback from the SRAM L1 cache, only the partitions in the cache line that have been updated by the processor (1 out of 4 partitions in the example) are written to the STT MRAM L2 data array. The data in the remaining partitions are identical to the data already stored in the L2 cache.

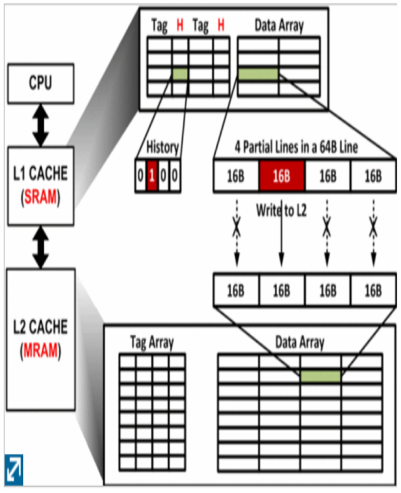

Figure 6: Partial cache line update

Therefore, writing the unchanged partitions into the L2 cache is unnecessary. The change of partitions can be tracked by using a history bit per partition. In the given example, 4 history bits to support 4 partitions are added into each tag in the L1 SRAM cache. The history data is used and reset whenever the corresponding cache line is written back into the L2 STT MRAM cache.

2.4 Total L2 Energy Consumption

• In order to analyze the energy efficiency of STT MRAM caches in comparison to SRAM caches, we measured the total energy consumption of L2 cache including leakage, read and write energy over 1 billion cycles of processor execution. The results presented in Figure 8 are obtained by averaging L2 cache energy consumption across 8 integer and 8 floating point benchmarks. SRAM-based L2 cache shows the largest energy consumption compared to STT MRAM caches with the same capacity, due to the significant leakage energy of SRAM bitcells. The energy difference is further improved for larger cache capacities. Moreover, under iso-area comparison (e.g., 0.5MB SRAM and 2MB STT MRAM caches), STT MRAM caches show significant energy benefit along with larger cache capacity (note that larger capacity improves processor performance by lowering cache misses). Our results show that a processor with

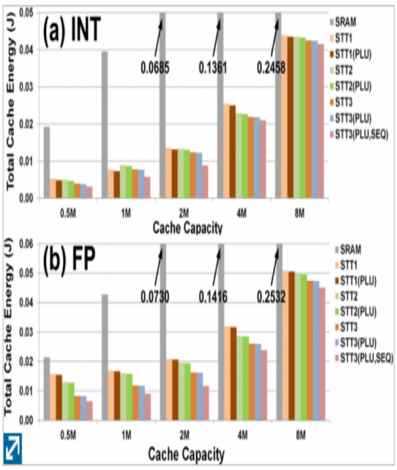

Figure 8: Total energy consumption of SRAM and STT MRAM L2 caches

2MB STT MRAM L2 outperforms one with 0.5MB SRAM L2 by 10 in IPC.

3 Conclusion

•In this work, they performed a comprehensive analysis of the performance, energy consumption and integration density of STT MRAM caches in comparison to conventional SRAM cache. they considered different genres of MTJ stacks and STT MRAM bitcell configurations in this study. Based on the detailed analysis of various bitcell characteristics including accurate area estimation from physical layout, they showed that, for large cache capacity, STT MRAM caches can have lower dynamic energy consumption and read latency compared to SRAM caches with the same capacity. Moreover, the low leakage energy consumption and high integration density of STT MRAM are highly beneficial for lower level caches (due to low utilization), and improve energy efficiency and processor performance. they also proposed read and write energy reduction techniques, namely sequential tag-data access in reads and partial cache line update in writes, which exploit the non-volatility of STT MRAM bitcells. The results show that the proposed techniques further improve the energy efficiency of STT MRAM caches.
Reference: source:ieee authors: sang phill park and sumeet gupta and niladri