DWM-TAPESTRI - An Energy Efficient All-Spin Cache using Domain wall Shift based Writes

Rangharajan Venkatesan, Mrigank Sharad, Kaushik Roy, and Anand Raghunathan
School of Electrical and Computer Engineering, Purdue University
{rvenkate,msharad,kaushik,raghunathan}@purdue.edu

***Abstract*—Spin-based memories are promising candidates for future on-chip memories due to their high density, non-volatility, and very low leakage. However, the high energy and latency of write operations in these memories is a major challenge. In this work, we explore a new approach – shift based write – that offers a fast and energy-efficient alternative to performing writes in spin-based memories. We propose DWM-TAPESTRI, a new all-spin cache design that utilizes Domain Wall Memory (DWM) with shift based writes at all levels of the cache hierarchy. The proposed write scheme enables DWM to be used, for the first time, in L1 caches and in tag arrays, where the inefficiency of writes in spin memories has traditionally precluded their use. At the circuit level, we propose bit-cell designs utilizing shift-based writes, which are tailored to the differing requirements of different levels in the cache hierarchy. We also propose pre-shifting as an architectural technique to hide the latency of shift operations that is inherent to DWM. We performed a systematic device-circuit-architecture evaluation of the proposed design. Over a wide range of SPEC 2006 benchmarks, DWM-TAPESTRI achieves 8.2X improvement in energy and 4X improvement in area, with virtually identical performance, compared to an iso-capacity SRAM cache. Compared to an iso-capacity STT-MRAM cache, the proposed design achieves around 1.6X improvement in both area and energy under iso-performance conditions.**

I. Introduction

The incessant increase in the demand for larger on-chip memories has led to the search for denser, energy-efficient memory technologies that can complement or replace SRAM/DRAM. A promising direction that has resulted from this search is the use of spin-based memories to design on-chip caches. The most common genre of spin-based memory is STT-MRAM [1]. It has a number of benefits over SRAM and DRAM including high density, non-volatility (therefore low leakage power), and energy-efficient, fast read operations. However, the latency and energy required to perform write operations are major bottlenecks. Another recently proposed spin-based memory technology is Domain Wall Memory (DWM), which stores data in the magnetic domains of a ferromagnetic wire [2]. A key characteristic of DWM is that the bits stored in the ferromagnetic wire can be shifted by applying a current pulse, utilizing a phenomenon called domain wall motion. DWM was initially envisioned as a replacement for secondary storage due to its excellent density, which is unmatched even among other emerging memory technologies. Recently, research efforts have investigated the use of DWM as on-chip memories [3]–[5]. However, the efficiency of write operations in DWM, which are similar to STT-MRAM, remains a major challenge.

In this paper, we bring a completely new and different insight to address the challenge of write energy and latency in cache design — *domain wall motion, which was originally proposed for performing shift operations in DWM, offers a fast, energy-efficient alternative for performing writes.* The concept of domain wall motion is fundamentally different from the MTJ-based write mechanism used in both STT-MRAM and traditional DWM designs. Domain wall motion based write has been experimentally demonstrated to be superior to MTJ based write in terms of energy, latency, as well as scalability, for nanoscale magnets with perpendicular magnetic anisotrophy (PMA) [6]. Leveraging this insight, we propose a new design for an all-spin on-chip cache hierarchy that uses domain wall motion based write, which we also refer to as *shift based write*.

This material is based upon work supported in part by Intel Corporation.

Our proposal, which we call DWM-TAPESTRI (Domain Wall Memory TAPE with ShifT based wRIte), consists of circuit and architecture level techniques that maximize the efficiency of a DWM-based cache hierarchy.

Our first contribution is the design of two different bit-cells, TAPESTRI-1bit and TAPESTRI-multi, which are optimized for the differing requirements of the different levels of the cache hierarchy. Both bit-cells utilize shift-based writes to improve write efficiency.

- TAPESTRI-1bit is a bit-cell that is designed to optimize performance. It retains all the benefits of STT-MRAM and can match SRAM in write efficiency. Moreover, unlike conventional DWM bit-cells, it does not require any shift operations. This allows it to be used in L1 cache, where spin memories have conventionally not been used due to their high write latency/energy.
- TAPESTRI-multi is a bit-cell that is designed to maximally utilize the density benefits of domain wall memory. It achieves much higher density than STT-MRAM (and TAPESTRI-1bit) by storing multiple bits in a single cell. However, this design, in general, requires shift operations to be performed on a cell before read/write accesses.

Our second contribution is an all-spin cache design using the proposed cells, which uses spin-based memory at all levels of the cache hierarchy, and for both the tag and data arrays within each level.

- Considering the design constraints of different levels of the cache hierarchy, we propose a cache organization wherein TAPESTRI-1bit is used for both the tag and data arrays for the L1 cache, while the L2 cache utilizes a hybrid organization with TAPESTRI-multi for the data array and TAPESTRI-1bit for the tag array.
- In order to alleviate the performance overhead imposed by the additional shift operations in TAPESTRI-multi, we propose cache pre-shifting, a technique in which bits in each cell are predictively shifted such that the bit that is expected to be accessed next is aligned with the read/write port. In this way, pre-shifting hides the extra latency imposed by shift operations.

Our third contribution is a systematic device-circuit-architecture evaluation of the proposed design. In this work, we focus on evaluating the energy and area benefits of the proposed design at iso-performance. For this purpose, we perform iso-capacity replacement of SRAM and STT-MRAM caches with the proposed design. Our analysis shows that the DWM-TAPESTRI cache design achieves 8.2X improvement in energy and 4X improvement in area compared to SRAM cache and 1.63X improvement in energy and 1.56X improvements in area compared to STT-MRAM cache.

The rest of this paper is organized as follows. Section II presents a brief survey of related work. Section III outlines the fundamental concepts of DWM. Section IV presents the proposed TAPESTRI-1bit and TAPESTRI-multi bit-cell designs. Section V describes the proposed DWM-TAPESTRI cache architecture. Section VI presents our methodology for modeling the proposed design. In Section VII, we present experimental results and Section VIII concludes the paper.

II. Related Work

In this section, we present a brief survey of research efforts that have addressed the inefficiency of write operations in spin-based memories at different levels of abstraction.

At the device level, researchers have optimized the write operation by designing different kinds of MTJ structures such as Dual-pillar MTJ, tilted MTJ, Dual barrier MTJ *etc.* [7], [8]. Many of the proposed device structures decouple the read and write paths, thereby relaxing the read *vs.* write design conflicts that are commonly present in memory design. At the circuit level, proposals to use 2T-1R structures with dual source lines and early write termination are aimed at reducing the write energy consumption [9], [10]. At the architecture level, the impact of inefficient writes is minimized by reducing the number of write operations, using hybrid caches in which the frequently written blocks are stored in SRAM [11], [12], write-biasing to increase the residency of dirty blocks and avoid repeated writes [13], and hiding the write latency using write buffers [14]. Recent efforts have proposed volatile STT-MRAM design by relaxing the non-volatility requirement at the device level to exploit the short lifetime of data in caches and improve the write efficiency of STT-MRAM [15], [16]. While the inefficiency of write operations is a common concern of all spin-based memories, DWM introduces the additional challenge of shift operations. Architectural techniques to address the overhead of shift operations were presented in [5], including cache management policies for head selection and update.

Our contributions of shift based write and pre-shifting are different from, and complementary to previous proposals. Although ours is the first work to leverage shift-based writes in the design of caches using spin memory, the fundamental benefits of domain wall motion based write have been demonstrated previously at the device level [6], which forms the motivation for our work.

III. Background

In this section, we provide background information about the structure and operation of domain wall memory.

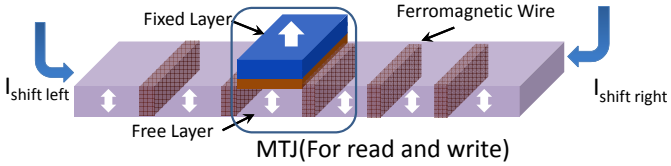


Fig. 1. Structure of a DWM Device

Figure 1 shows the structure of a DWM device consisting of a ferromagnetic wire and an MTJ. The data in a DWM is stored in the domains of the ferromagnetic wire [2]. A ferromagnetic wire can have multiple domains and therefore is capable of storing multiple bits of data. A key feature of DWM is that the data stored in the magnetic domains can be shifted by injecting a current pulse through the ferromagnetic wire. The current causes the domains to move along the strip through a phenomenon known as domain wall translation [2]. Note that there is no physical movement of the wall; instead the spins change their orientation when current is applied. For performing the read/write operations in a DWM, the MTJ is traditionally used, similar to STT-MRAM. This leads to high write energy and latency in DWM.

IV. DWM-TAPESTRI: Circuit Design

In a typical DWM, reads/writes are performed using an MTJ and domain wall shifts are only used to select the correct bit to be read/written. *However, we note that shift operations can also be used to perform writes.* This is accomplished by designing a DWM with 2 fixed domains having opposite spin orientations, as shown in Figure 2. These domains act as source of 0/1 and hence, we can perform a write operation by performing a shift in the appropriate direction. Note that the magnetic orientations of the fixed domains remain constant and cannot be modified by the shift operation if they are sized appropriately. Based on this concept, we propose two bit-cell designs – TAPESTRI-1bit and TAPESTRI-multi that are optimized for performance and area respectively. The detailed description of the two bit-cells is presented below.

A. TAPESTRI-*1bit*

Figure 2a shows the schematic view of the TAPESTRI-1bit bit-cell. It consists of a ferromagnetic wire, an MTJ and 2 access transistors. The data stored in the bit-cell is determined by the magnetic orientation of the free domain. When the magnetic orientation of the free domain is parallel to that of the MTJ fixed layer, then the MTJ offers low resistance indicating the '0' state. When the magnetic orientations are anti-parallel, the MTJ offers high resistance indicating the '1' state. The bit-cell also consists of two separate access transistors – a read access transistor (T2) and a write access transistor (T1), which are used to control the direction of currents during the read/write operations.

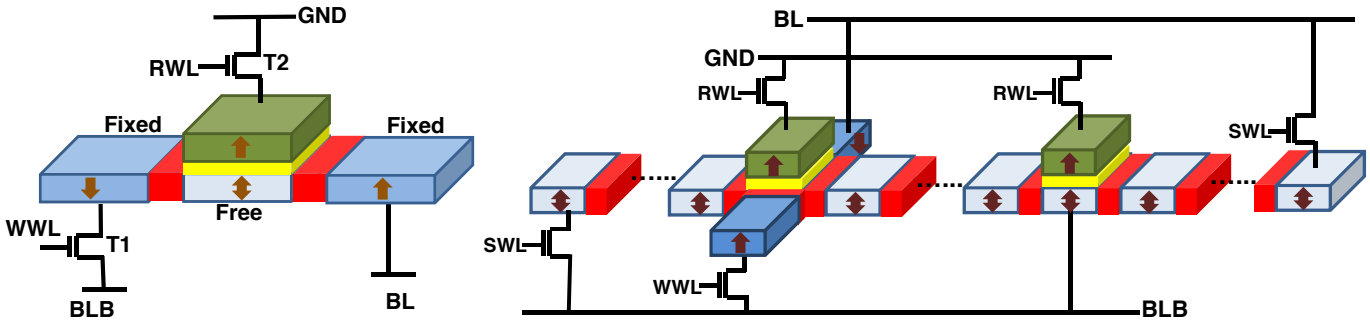
Read/Write operation: In order to read the contents of the cell, the read access transistor (T2) is turned ON and the bitline BL is driven high. The current that flows from BL to GND varies depending on the resistance offered by the MTJ, and is used to determine the value stored in the cell. The write operation in the proposed design is performed by shifting the appropriate magnetization from the fixed domains to the free domain of the ferromagnetic wire. In order to write 1, write access transistor (T1) is turned ON, bitline BL is driven high and BLB is connected to GND. This shifts the domains towards the right, thereby writing 1 into the bit-cell. For writing 0, the voltage conditions of the bitlines are reversed.

B. TAPESTRI-*multi*

The schematic of TAPESTRI-multi is shown in Figure 2b. It consists of a ferromagnetic wire capable of storing multiple bits, a read-write port, a read-only port and shift ports. The read/write port in TAPESTRI-multi is made up of 2 fixed domains, 1 MTJ and 2 access transistors, a structure that is similar to TAPESTRI-1bit. The read-only port is a 1T-1MTJ structure that can be used to perform only read-operations. In addition to read/write ports and read-only ports, the TAPESTRI-multi bit-cell has shift ports consisting of an access transistor at each end of the ferromagnetic wire. Note that such structures have been proposed and prototyped in the context of domain-wall logic [17]. In this work, we use them to achieve high density and efficient write operations simultaneously.

Read/Write/Shift Operation: Three kinds of operations can be performed in a TAPESTRI-multi bit-cell – Read, write and shift. Reading/writing of data to the domain at the read/write port is performed in a manner similar to TAPESTRI-1bit, as described above. Shifting of bits in TAPESTRI-multi is accomplished by turning ON the shift access transistors and precharging the bitlines to appropriate voltages. For shifting the bits towards the left, BL is connected to VDD and BLB to GND. For shifting in the opposite direction, the voltage conditions of the bitlines are reversed.

The key benefit of TAPESTRI-multi is that it achieves very high density by sharing the read/write ports across multiple bits that are stored in the ferromagnetic wire. However, this introduces the need to shift the bits to the appropriate read/write port before they can be accessed. Shift operations introduce additional latency for accessing the bits stored in TAPESTRI-multi. In order to address this problem of increased latency, we use multiple read/write ports and read-only ports to reduce the number of shift operations. Another implication of the shift operations is the need for extra "overflow" bits beyond the shift ports to prevent loss of data. Note that these extra bits typically



(a) TAPESTRI-1bit bit-cell schematic (b) TAPESTRI-multi bit-cell schematic

Fig. 2. Schematic of DWM-TAPESTRI cells

do not incur area penalty as the cell area is dominated by the access transistors.

C. Analysis of the proposed bit-cells

The salient features of the proposed bit-cells are as follows:

Optimized Write: The key feature of the proposed bit-cells is that domain wall motion is used for performing writes. Figure 3 compares the characteristics of domain wall motion based write with that of MTJ based write operation for nanoscale magnets with perpendicular magnetic anisotrophy (PMA). We consider a ferromagnetic wire with domains of size 32nm x 64nm and different values thickness (T) and use the device simulation model proposed in [18] for this evaluation. For MTJ based write, we consider an MTJ of size 32nm x 64nm with oxide thickness of 1.6nm. While the current density requirements for both write mechanisms are comparable, the cross-sectional area with domain wall motion based write is 32nm x T compared to 32nm x 64nm in the case of MTJ based write. This results in significant reduction in the write current requirement, thereby improving the efficiency of write operations in terms of both latency and energy consumption.

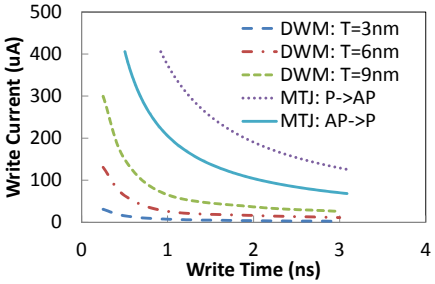
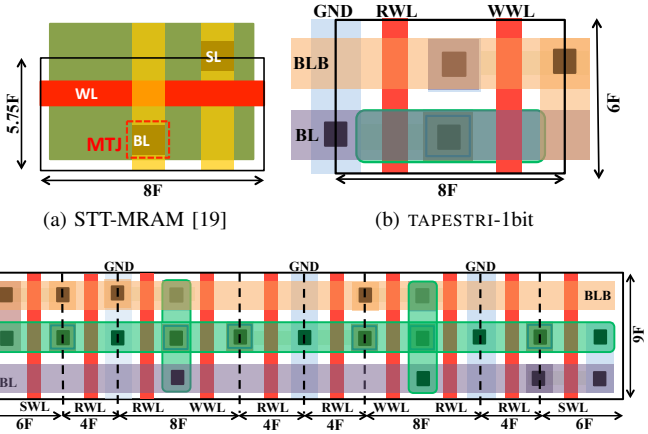


Fig. 3. Comparison of write characteristics of shift based write with MTJ based write

Density: Figure 4 shows the bit-cell layout comparison of the proposed bit-cells with STT-MRAM. The area of the TAPESTRI-1bit bit-cell is comparable to that of an STT-MRAM bit-cell. This is because the domain wall motion based write mechanism reduces the write current requirement of the proposed bit-cell considerably, enabling the use of a minimum-sized write access transistor. Both the access transistors in TAPESTRI-1bit can be minimum sized compared to a large single access transistor in the STT-MRAM bit-cell. Figure 4 also shows that a TAPESTRI-multi (which stores 16 bits) can achieve even higher densities ($24.75F^2$/bit) compared to STT-MRAM($46F^2$) and TAPESTRI-1bit ($48F^2$) bit-cell designs.

Read optimization: Another key feature of the proposed design is that it decouples the read and write current paths enabling us to optimize the bit-cell for read and write independently. The MTJ can now be optimized exclusively for read operations. This also enables us to use voltage mode sensing for faster reads.

Reliability: One of the key advantages of the proposed design is the mitigation of reliability issues related to tunnel oxide breakdown. In



(a) STT-MRAM [19] (b) TAPESTRI-1bit

(c) TAPESTRI-multi (16 bits per bit-cell)

Fig. 4. Layout of STT-MRAM, TAPESTRI-1bit, and TAPESTRI-multi bit-cells

standard STT-MRAM, the write speed is mainly limited by Time-Dependent Dielectric Breakdown (TDDB) of the tunnel oxide. With higher speed, the required write current density of the MTJ increases, which exponentially degrades the TDDB limited MTJ lifetime [20]. In the proposed design, decoupled read/write paths facilitate faster write operations without such reliability concerns.

Stability: Decoupling of read/write paths enables us to design a robust bit-cell with higher oxide thickness having higher cell Tunnelling Magneto-Resistance (TMR) and read disturb margins. Figure 5 shows that we can design the cells over a wide range of delay constraints without any read disturb margin degradation.

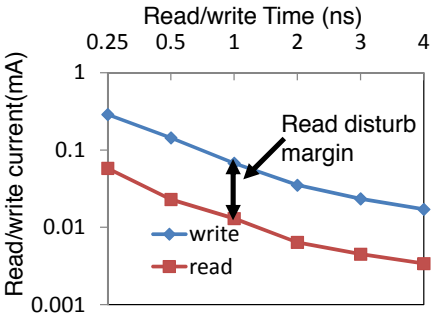


Fig. 5. Read/Write Stability of DWM-TAPESTRI cells

V. DWM-TAPESTRI: Cache Architecture

In this section, we first discuss the various design considerations for different levels of caches and then describe the proposed DWM-TAPESTRI cache architecture that uses the proposed bit-cell designs.

Cache design considerations: Different levels of caches are introduced to bridge the performance gap between the processor and main memory. L1 cache is the closest to the processor and is designed to

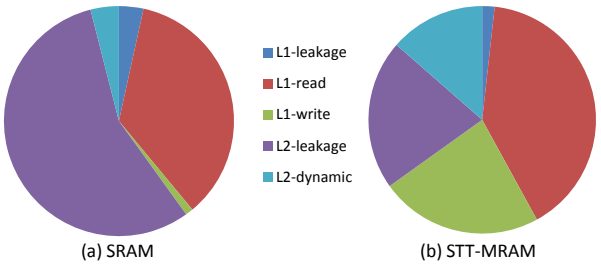


Fig. 6. Energy distribution in SRAM and STT-MRAM based caches

provide fast access to the data stored in it. The L2 cache is introduced to reduce the number of off-chip accesses to main memory. When we consider the different components of cache energy consumption of a system consisting of 32KB L1 cache and 1MB L2 cache (Figure 6), we can see that the leakage energy of L2 cache and the read energy of L1 cache contribute a significant fraction to cache energy consumption in an SRAM based cache system. Designing cache using STT-MRAM reduces the leakage energy component from L2 cache but increases the write energy from L1 and L2 caches. Therefore, we need to design a fast L1 cache with efficient read/write energies and a high density L2 cache with low leakage power.

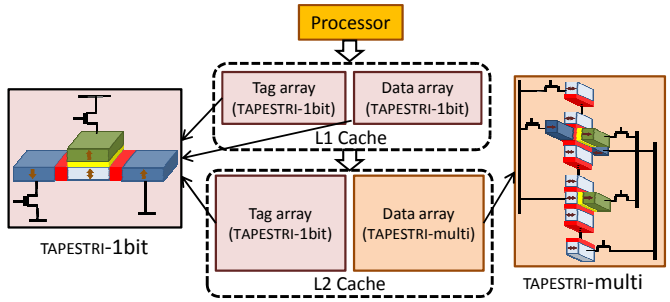


Fig. 7. DWM-TAPESTRI cache organization

A. *DWM-TAPESTRI Cache Organization*

Figure 7 shows the organization of the proposed DWM-TAPESTRI cache. In order to optimally exploit the benefits of the proposed bit-cells, we design an L1 cache consisting of TAPESTRI-1bit based data and tag arrays and a hybrid L2 cache consisting of a TAPESTRI-1bit based tag array and a TAPESTRI-multi based data array. As described in Section IV, TAPESTRI-1bit offers superior read/write performance with low read/write energy consumption, making it an ideal candidate for L1 cache. TAPESTRI-multi has high density and low leakage power making it an attractive candidate for designing lower level caches. However, designing both the data and tag arrays of an L2 cache with TAPESTRI-multi will first require shift operations to determine hits/misses from tag array followed by shift operations for accessing the cache block from data array[1]. This would considerably degrade the performance of the cache. Fortunately, the tag array contributes only a small fraction of the total area and energy consumption of a cache. Therefore, designing a hybrid L2 cache with a TAPESTRI-1bit based tag array and a TAPESTRI-multi based data array enables us to achieve most of the density benefits of DWM with minimal performance degradation due to shifts. In the proposed design, the data array organization and addressing policy are assumed to be similar to that of TapeCache [5].

B. *Cache management policy*

In order to reduce the performance impact of shifts from TAPESTRI-multi, we propose the concept of "preshifting" in which we predict the bit that is likely to be accessed next and align it with the read/write port to hide the impact of shift latency from the next cache access. The concept of "preshifting" is analogous to prefetching but is unique to DWM, which requires shift operations to access data.

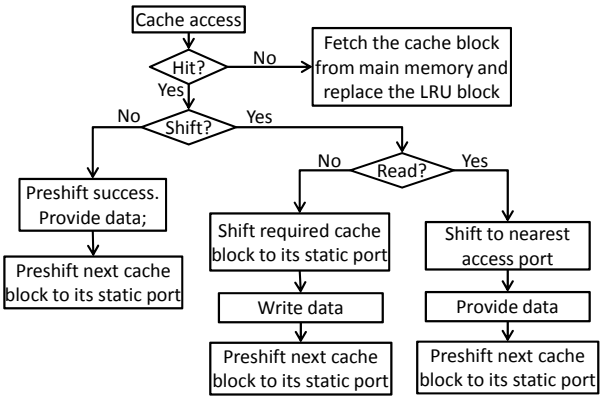


Fig. 8. Preshifting based cache management policy

Figure 8 shows the different steps involved in the proposed cache management policy. Initially, each cache block is assigned a read/write port statically based on its location. After a cache access, we predict the next cache block likely to be accessed and preshift the block to its statically assigned read/write port. If the prediction is successful, then no shift operation will be required, resulting in faster cache access. However, if the preshift fails and the cache access is a read operation, then we determine the nearest read port to the required cache block and use it to perform the required read access. This enables us to reduce the performance penalty due to misprediction for the performance critical read operation. On the other hand, if the cache access is a write operation, we use the statically allocated write port. This is because, it ensures that the number of extra bits required to avoid loss of data during shifting is small. Note that, a TAPESTRI-multi bit-cell storing N_b bits and N_{rw} read/write ports would require $2N_b/N_{rw}$ extra bits when we use statically allocated ports for writes compared to $2N_b$ extra bits if we use the nearest port for performing writes[2].

VI. MODELING

DWM-TAPESTRI differs significantly from traditional memories in terms of the device structure, circuit design as well as cache architecture. In order to accurately evaluate the characteristics of the proposed cache design, we developed a device-circuit-architecture modeling framework that is shown in Figure 9. The framework for

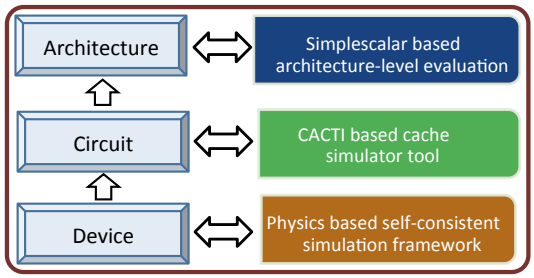


Fig. 9. Device-Circuit-Architecture Modeling Framework

modeling domain wall motion at the device level is based on the self-consistent simulation framework proposed in [18]. The device level characteristics were then used by a modified CACTI tool to evaluate stand-alone cache characteristics. This tool takes parameters such as the write current, write latency, number of read/write ports,

[1]In lower level caches, tag and data array accesses are serialized so as to avoid the energy overheads associated with reading all ways of a cache simultaneously.

[2]Assuming that the number of shifts required for writes will be higher than that for reads due to the presence of read-only ports.

and number of read-only ports, along with other input parameters of the CACTI tool, to compute the cache characteristics. Finally, in order to evaluate the benefits of the proposed design at the system-level, we modeled the proposed DWM-TAPESTRI cache architecture in SimpleScalar [21] and evaluated the performance in terms of instructions per cycle (IPC). We also obtained the trace of the cache accesses and used the cache characteristics computed above to compute the energy consumed by the DWM-TAPESTRI cache.

VII. Experimental Results

In this section, we first present the experimental setup used to evaluate the proposed DWM-TAPESTRI cache design and then present the results comparing the area, energy and performance of DWM-TAPESTRI with SRAM and STT-MRAM based designs. We then present the results comparing the L1 and L2 cache characteristics of DWM-TAPESTRI with other memory technologies. Finally, we present architecture level results comparing the energy and performance of the proposed design across a wide range of benchmarks.

A. Experimental Setup

In our experiments, we perform iso-capacity replacement of L1 and L2 cache and compare the area, energy and performance of the proposed design with that of CMOS SRAM and STT-MRAM. All memory technologies considered are based on a 32nm technology node. The processor configuration used in our analysis is provided in Table I. We evaluate SRAM memories using CACTI [22], and spin-based memories (STT-MRAM and DWM) using a modified CACTI tool. In our analysis, we consider a TAPESTRI-multi bit-cell capable of storing 16 bits with 2 read/write ports, 4 read-only ports for designing the data array of the L2 cache. We perform architectural simulations over a wide range of benchmarks from the SPEC 2006 suite using SimpleScalar [21] for 1 billion instructions. In all our simulation runs, we warm up the cache by fast forwarding for 1 billion instructions.

TABLE I
System Configuration

Processor Core	Alpha 21264 pipeline, Issue Width - 4
Processor Frequency	2 GHz
Functional Units	Integer - 8 ALUs, 4 Multipliers Floating Point - 2 ALUs, 2 Multipliers
L1 D/I-Cache	32KB, direct mapped, 32 byte line size, 2 cycle hit latency
L2 Unified Cache	1MB, 4-way associative, 64 byte line size, hit latency depends on technology

B. Results Summary

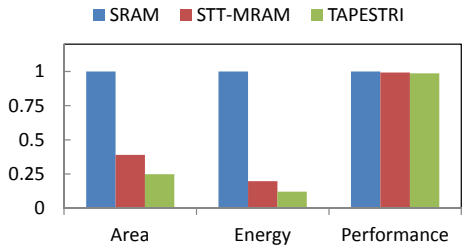
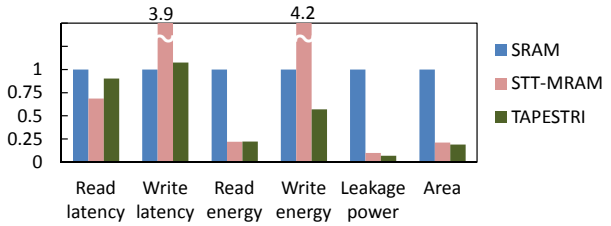


Fig. 10. Comparison of area, energy and performance across different memory technologies

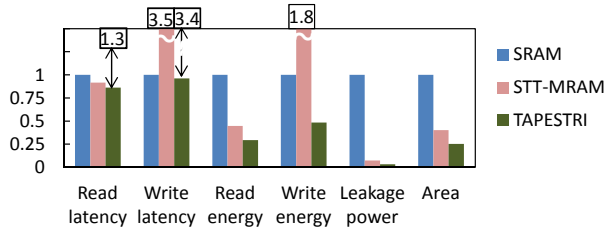
Figure 10 summarizes the benefits of DWM-TAPESTRI compared to SRAM and STT-MRAM based caches. Compared to SRAM based cache, DWM-TAPESTRI achieves 8.2X improvement in energy and 4X improvement in area at iso-performance. When we compare the results with STT-MRAM based cache, DWM-TAPESTRI achieves 1.63X improvement in energy and 1.56X improvement in energy with virtually identical performance. Next, we will examine the benefits of DWM-TAPESTRI in greater detail.

C. Cache Characteristics

In this section, we present the results comparing the characteristics of the proposed L1 and L2 cache designs with that of SRAM and STT-MRAM based caches.



(a) L1 Cache



(b) L2 Cache

Fig. 11. Comparison of L1 and L2 cache characteristics

Figures 11a and 11b compare the L1 and L2 cache characteristics, respectively, across different memory technologies. As shown in the figure, the density of the TAPESTRI-1bit based L1 cache is similar to STT-MRAM and that of the hybrid DWM-TAPESTRI L2 cache is higher than both SRAM and STT-MRAM due to the higher density of TAPESTRI-multi. When we compare the leakage power, we can see that spin-based memory technologies can achieve significant reduction in the leakage power consumption compared to SRAM due to their non-volatility. When we compare DWM-TAPESTRI with STT-MRAM, the reduction in leakage power consumption is due to smaller peripheral circuitry. Comparing the TAPESTRI-1bit based L1 cache with STT-MRAM based L1 cache, the wordline and bitline drivers in the STT-MRAM cache need to be sized larger due to the increased capacitive load from the large access transistors. This marginally increases the leakage power consumption of the STT-MRAM cache.

When we compare the access latencies of different L1 caches, we can see that both the read and write latencies of TAPESTRI-1bit are comparable to SRAM cache. Due to the inefficiency of MTJ based writes, the STT-MRAM based L1 has very high write latency. On the other hand, shift based write is highly efficient and enables us to improve the write latency significantly. When we consider the access latencies of different L2 caches, the access latency of the DWM-TAPESTRI cache varies due to the variable access latency of TAPESTRI-multi, with the best case being comparable to SRAM. The effectiveness of preshifting implies that average access latencies are close to the best case.

Next, when we consider the read energies, all spin-based memories achieve significant benefits due to reduced bitline and wordline capacitances arising from improved density. Moreover, the shift based write is highly energy efficient and this enables DWM-TAPESTRI to achieve significant reduction in write energy compared to SRAM and STT-MRAM based caches.

D. Architectural Evaluation

In this section, we present the architecture level results comparing the energy and performance of DWM-TAPESTRI with SRAM and STT-MRAM caches across a wide range of benchmarks.

Energy consumption of DWM-TAPESTRI: Figure 12 compares the energy consumed by DWM-TAPESTRI with SRAM and STT-MRAM caches, normalized to the STT-MRAM cache. As we can see from the figure, DWM-TAPESTRI achieves significant reduction in the total cache energy consumption compared to both SRAM and STT-MRAM. STT-MRAM based cache reduces the leakage and read energy while increasing the write energy. DWM-TAPESTRI achieves reduction in all the three energy components- leakage, read and write. In addition, the proposed design achieves even higher reduction in leakage and read energy compared to STT-MRAM caches as shown earlier. As a result, DWM-TAPESTRI enables us to achieve maximum benefits in the total energy consumption of cache.

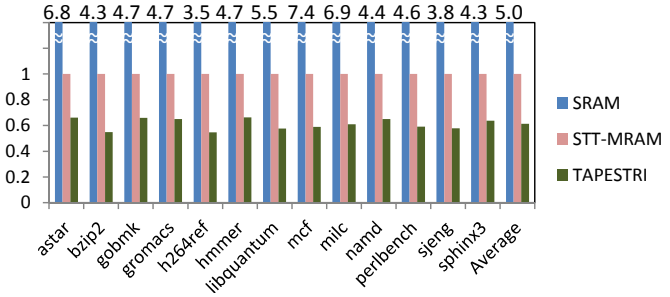


Fig. 12. Comparison of energy consumption of cache across different memory technologies

Performance evaluation: Figure 13 presents a comparison of

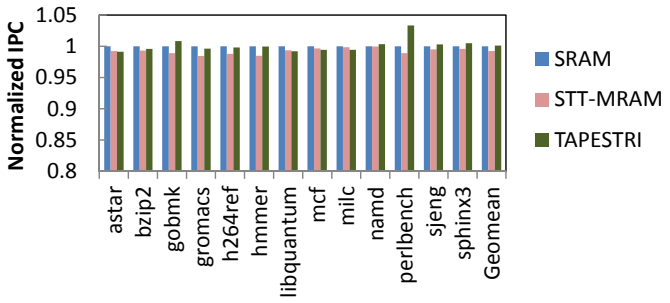


Fig. 13. Performance comparison across different memory technologies

the performance of different cache designs in terms of normalized instructions per cycle (IPC) normalized to SRAM performance. In our design, DWM-TAPESTRI was designed to achieve iso-performance and as we can see from the figure, DWM-TAPESTRI achieves virtually identical performance across all the benchmarks. This also validates the fact that the proposed cache management policy along with the introduction of multiple read/write ports and read-only ports can effectively compensate for the performance penalty arising due to shift operations in TAPESTRI-multi.

Comparison to TapeCache: We also evaluated the benefits of the proposed design compared to TapeCache [5]. For this purpose, we considered a TapeCache design with 16 bits, 4 read-only ports and a read/write port that uses the Static-Lazy (SL) management presented in policy [5]. Our evaluation showed that DWM-TAPESTRI achieves 1.2X improvement in L2 cache energy compared to TapeCache due to the use of efficient domain wall shift based writes. In terms of total energy consumption, DWM-TAPESTRI achieves 3.6X energy improvement due to significant reduction in L1 cache energy[3]. Also, the shift based write enables the use of minimum-sized transistors in DWM-TAPESTRI, which results in 1.4X improvement in cache area compared to TapeCache.

[3]Note that TapeCache uses an SRAM based L1 since the latency of DWM without shift based write is prohibitive.

VIII. Conclusion

Spin-based memories have tremendous potential for use in future computing platforms. However, the inefficiency of write operations is a major bottleneck. In this work, we proposed a new design – DWM-TAPESTRI – for an all-spin cache that uses domain wall motion for performing writes. Our analysis shows that the proposed DWM-TAPESTRI cache achieves considerable improvements in area and energy compared to SRAM and STT-MRAM based caches.

References

[1] K. Lee and S. H. Kang. Development of Embedded STT-MRAM for Mobile System-on-Chips. *IEEE Trans. Magnetics*, 47(1):131 –136, January 2011.

[2] S. Parkin, M. Hayashi, and L. Thomas. Magnetic domain-wall racetrack memory. *Science*, 320(5873):190–194, April 2008.

[3] R. Venkatesan, V. K. Chippa, C. Augustine, K. Roy, and A. Raghunathan. Energy Efficient Many-core Processor for Recognition and Mining using Spin-based Memory. In *Proc. NANOARCH*, pages 122 –128, June 2011.

[4] W. Zhao, D. Ravelosona, J. Klein, and C. Chappert. Domain Wall Shift Register-Based Reconfigurable Logic. *IEEE Trans. Magnetics*, 47(10):2966 –2969, October 2011.

[5] R. Venkatesan, V. Kozhikkottu, C. Augustine, A. Raychowdhury, K. Roy, and A. Raghunathan. TapeCache: A High Density, Energy Efficient Cache Based on Domain Wall Memory. In *Proc. ISLPED*, pages 185–190, July 2012.

[6] S. Fukami et al. Low-Current Perpendicular Domain Wall Motion Cell for Scalable High-Speed MRAM. In *IEEE Symp. on VLSI Technology*, pages 230 –231, June 2009.

[7] N. N. Mojumder and K. Roy. Switching current reduction and thermally induced delay spread compression in tilted magnetic anisotropy spin-transfer torque (STT) MRAM. *IEEE Trans. Magnetics*, 2011.

[8] C. Augustine, A. Raychowdhury, D. Somasekhar, J. Tschanz, K. Roy, and V. K. De. Numerical analysis of typical STT-MTJ stacks for 1T-1R memory arrays. In *Proc. IEDM*, pages 22.7.1 –22.7.4, December 2010.

[9] Y. Kim, S. K. Gupta, S. P. Park, G. Panagopoulos, and K. Roy. Write-optimized reliable design of STT MRAM. In *Proc. ISLPED*, pages 3–8, 2012.

[10] P. Zhou, B. Zhao, J. Yang, and Y. Zhang. Energy reduction for STT-RAM using early write termination. In *Proc. ICCAD*, pages 264 –268, November 2009.

[11] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie. Hybrid cache architecture with disparate memory technologies. In *Proc. ISCA*, pages 34–45, June 2009.

[12] A. Jadidi, M. Arjomand, and H. S. Azad. High-endurance and performance-efficient design of hybrid cache architectures through adaptive line replacement. In *Proc. ISLPED*, pages 79–84, August 2011.

[13] M. Rasquinha, D. Choudhary, S. Chatterjee, S. Mukhopadhyay, and S. Yalamanchili. An energy efficient cache design using Spin Torque Transfer (STT) RAM. In *Proc. ISLPED*, pages 389 –394, August 2010.

[14] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen. A novel architecture of the 3D stacked MRAM L2 cache for CMPs. In *Proc. HPCA*, pages 239 –249, February 2009.

[15] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan. Relaxing non-volatility for fast and energy-efficient STT-RAM caches. In *Proc. HPCA*, pages 50 –61, February 2011.

[16] A. Jog, A. K. Mishra, C. Xu, Y. Xie, V. Narayanan, R. Iyer, and C. R. Das. Cache revive: Architecting volatile STT-RAM caches for enhanced performance in CMPs. In *Proc. DAC*, pages 243 –252, June 2012.

[17] D. A. Allwood, G. Xiong, C. C. Faulkner, D. Atkinson, D. Petit, and R. P. Cowburn. Magnetic Domain-Wall Logic. *Science*, 309(5741):1688–1692, 2005.

[18] C. Augustine, A. Raychowdhury, B. Behin-Aein, S. Srinivasan, J. Tschanz, V. K. De, and K. Roy. Numerical analysis of domain wall propagation for dense memory arrays. In *Proc. IEDM*, pages 17.6.1 –17.6.4, December 2011.

[19] S. K. Gupta, S. P. Park, N. N. Mojumder, and K. Roy. Layout-aware optimization of STT MRAMs. In *Proc. DATE*, pages 1455 –1458, march 2012.

[20] A. A. Khan, J. Schmalhorst, A. Thomas, O. Schebaum, and G. Reiss. Dielectric breakdown in CoFeB/MgO/CoFeB magnetic tunnel junction. *Journal of App. Physics*, 103:123705–123705–5, June 2008.

[21] T. Austin, E. Larson, and D. Ernst. SimpleScalar: An Infrastructure for Computer System Modeling. *Computer*, 35:59–67, February 2002.

[22] CACTI. http://www.hpl.hp.com/research/cacti/.