Low Power Data-Aware STT-RAM based Hybrid Cache Architecture

Mohsen Imani, Shruti Patil, Tajana Rosing

Computer Science and Engineering Department
University of California San Diego, La Jolla, CA 92093, USA
E-mail: {moimani, patil, tajana}@ucsd.edu

Abstract

Static Random Access Memories (SRAMs) occupy a large area of today's microprocessors, and are a prime source of leakage power in highly scaled technologies. Low leakage and high density Spin-Transfer Torque RAMs (STT-RAMs) are ideal candidates for a power-efficient memory. However, STT-RAM suffers from high write energy and latency, especially when writing 'one' data. In this paper we propose a novel data-aware hybrid STT-RAM/SRAM cache architecture which stores data in the two partitions based on their bit counts. To exploit the new resultant data distribution in the SRAM partition, we employ an asymmetric low-power 5T-SRAM structure which has high reliability for majority 'one' data. The proposed design significantly reduces the number of writes and hence dynamic energy in both STT-RAM and SRAM partitions. We employed a write cache policy and a small swap memory to control data migration between cache partitions. Our evaluation on UltraSPARC-III processor shows that utilizing STT-RAM/6T-SRAM and STT-RAM/5T-SRAM architectures for the L2 cache results in 42% and 53% energy efficiency, 9.3% and 9.1% performance improvement and 16.9% and 20.3% area efficiency respectively, with respect to SRAM-based cache running SPEC CPU 2006 benchmarks.

Keywords

Hybrid cache, Non-volatile memory, SRAM, NBTI

1. Introduction

Over the last few decades, scaling of conventional CMOS technology has been motivated by the need for higher integration density and performance. Static power consumption is a major concern in designing nano-scale integrated circuits, due to the exponential dependence of subthreshold current on the threshold voltage. As embedded RAM now comprises a dominant portion of chip area in contemporary microprocessors, minimizing its leakage power is a critical design objective [1].

The low leakage power and high density of Spin-Torque Transfer RAMs (STT-RAMs) make them ideal candidates to replace conventional SRAMs [2, 3]. However, STT-RAMs suffer from high write latency and energy, which restricts their direct use as low level caches. In STT-RAMs the write latency optimization incurs tradeoffs with many other cell parameter, such as write/read energy and read latency [4]. There have been several architectural proposals to improve the write energy and performance of STT-RAMs including write buffer [5], early write termination (EWT) technique [6], multi retention time cache [7] and hybrid read/write-intensive cache architecture [8]. However, in most techniques the write latency is in tradeoff with write energy or endurance of the cache.

In this paper, we design a hybrid cache with data encoding using low cost peripheral circuitry to improve energy, latency and endurance of cache simultaneously. The proposed data-aware hybrid cache splits the input data between STT-RAM and SRAM caches based on the proportion of 'ones' in input data. This data distribution stores data with majority 'zero' and 'one' bits in STT-RAM and SRAM caches respectively. A high percentage of 'zeros' on STT-RAM improves the performance and energy efficiency of the proposed cache since writing zero on STT-RAM cell consumes 3.5X-6.5X less write energy compared to writing one [9]. To exploit the high proportion of ones in SRAM, we use asymmetric 5T-SRAM cell which consumes very low power to store 'one' data. Writing majority zero (one) data on the cache containing majority zero (one) improves the number of write operations, write energy and endurance of STT-RAM (SRAM). Our evaluations on UltraSPARC-III using GEMS simulator show that the STT-RAM/6T-SRAM and STT-RAM/5T-SRAM architectures can achieve 42% and 53% energy efficiency, and 9.3% and 9.1% performance improvement on average, with respect to conventional SRAM caches running SPEC CPU 2006 benchmarks.

2. Related work

NVMs are promising technologies to reduce the static power consumption of on-chip memories and recently as an computing units [10, 11]. However, they suffer from high write latency, energy consumption and endurance. Several techniques have been introduced to address NVM dynamic energy [2, 12] write latency [4, 13] and endurance [14, 15]. We briefly outline them in this section.

Hybrid NVM/SRAM caches effectively decrease the write stress on NVMs. A hybrid cache can utilize fast SRAMs for write intensive data, and low leaky NVMs for non-write intensive workloads [2]. To monitor and control the number of read and write operations in NVM and SRAM caches, Wu. et al, [16] proposed a migration technique by adding two flag bits to tag store. Smullen, et al, [4] proposed a SRAM/STT-RAM hybrid cache which uses relaxed magnetic tunneling junction (MTJ) that decreases data retention of STT-RAM in order to speed-up STT-RAM write operations. Further, Sun, et al. [7] split the STT-RAM into low and high retention regions and utilized a refresh scheme (similar to DRAMs) to update the data in low

retention region. However, the high energy consumption of refresh schemes reduce the advantages of using NVM.

All previous hybrid caches tried to balance write latency, write energy and endurance of a cache. In contrast, the proposed technique combines the strengths of hybrid caches with data encoding using low cost peripheral circuitry. Proposed data-aware hybrid cache architecture exploits the asymmetric write characteristics of STT-RAMs to mitigate write stress. Our design exploits the data distribution on both STT-RAM and SRAM parts to improve both energy and performance at the same time.

3. Background and motivation

3.1. STT-RAM

An STT-RAM cell consists of an MTJ and an access transistor (see Figure 1a). A bit value is stored in the cell based on the state of the MTJ (High/low resistance). This cell uses the same path for read and write operations. STT-RAM consumes zero leakage power, and is approximately 4X denser than SRAM cell. The endurance of STT-RAM is high due to high MTJ robustness (>10^{15} cycles), which makes STT-RAM suitable for use in caches [4]. However, endurance is reduced by unbalanced write accesses to caches. STT-RAMs consume higher write energy (6X) and have longer write latency (8X) than SRAMs. These characteristics limit the direct usage of STT-RAMs in high speed cache structures with high write stress, such as the L1 cache [4, 6, 13]. Additionally, STT-RAM has asymmetric structure that makes writing '1' harder than writing '0'. For a 5ns write pulse width, writing '1' needs 3.5X more energy than writing '0' and this difference increases severely with reducing write pulse width [2]. This asymmetric write significantly affects the energy, latency and the endurance of STT-RAM cell.

3.2. SRAM

Static power is the main source of power consumption in SRAMs [17]. The conventional 6T-SRAM cell is shown in Figure 1b. The cell consists of two back-to-back inverters (M1-M4) and two access transistors (M5, M6). This cell uses the same path for both read and write operations, and most of the cell failures occur due to the read mode issues. A 5T-SRAM cell is an asymmetric type of SRAM cell (see Figure 1c). This cell leaves out one of the pull-up PMOS transistors in order to decrease the strength of back-to-back inverter, which lowers the static power. The correct functionality of this cell over a wide range of supply voltages has been verified in previous work [18-20]. We observe that this cell provides a novel opportunity to save static power when storing 'one' in Q node. During this operation, the single pull-up PMOS transistor is in OFF mode and disconnects the Vdd from storage nodes. This reduces the energy consumption of writing 'one' on 5T cells.

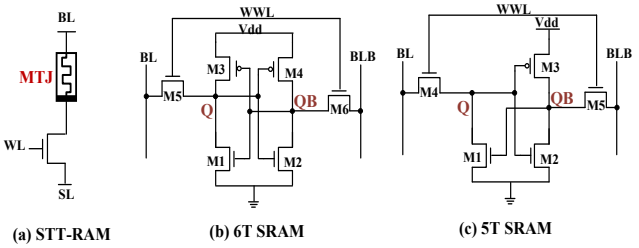

Figure 1. (a) STT-RAM standard cell, (b) conventional 6T-SRAM and (c) 5T-SRAM cells

4. Proposed hybrid cache

In this section we describe our proposed hybrid cache design with STT-RAM and SRAM partitions. We design each partition differently based on the expected data distribution within it. We also use a cache policy that determines which partition should be chosen for writing data, and how/when to migrate existing data between the partitions.

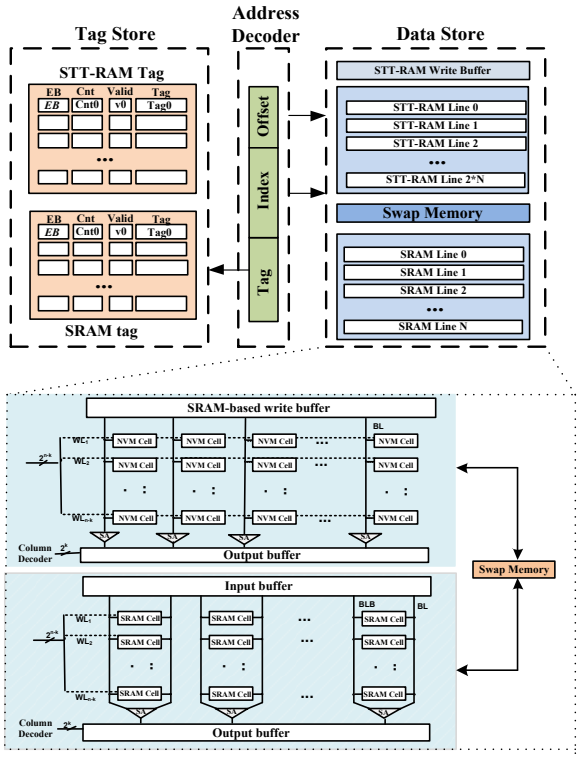

Figure 2. Proposed data-aware hybrid cache architecture

4.1. Cache architecture

We propose a hybrid cache architecture which stores data with majority 'zero' and 'one' bits in STT-RAM and SRAM cache partitions respectively. Before each write operation, a parallel counter circuitry counts the number of 'zeros' in the input data and compares it with a threshold value (*THR*). If this number is larger than *THR*, the data is written in STT-RAM cache and the *EB* flag set to '1'. Otherwise the data is written on SRAM cache, setting *EB* flag to '0'. In the read

mode the cache transfers the value of STT-RAM or SRAM on output bus based on *EB* flag bit. The *EB* flag bit, keeps track of the exact place of each data on cache. In read mode, the *EB* signal allows the system to activate just one of the caches to effectively reduce read energy consumption. The structure of the proposed hybrid cache is shown in Figure 2.

To mitigate the write latency and energy overhead of STT-RAM cache, we optimized STT-RAM cache based on the write Energy-Delay Product (EDP). We also employed SRAM-based write buffer to reduce the write latency of the proposed design. In SRAM part, the cache is optimized based on leakage power. Our design considers the design exploration of using both 5T and 6T SRAM as a SRAM cache.

4.2. Cache policy

When a new write request comes into the cache, it may hit or miss in the cache based on the data availability. In case of a miss, the system writes the new data in SRAM or STT-RAM cache based on the percentage of ones and *THR* value. In case of a hit, the data is found in the SRAM or STT-RAM partition based on the *EB* flag bit. If a write request with majority 'ones' data *hits* in the STT-RAM cache, the data update writes majority-one data in the STT-RAM partition. Such a write degrades the energy efficiency of the system since the STT-RAM is only power efficient for data with majority 'zeros'. In order to avoid this problem, we use a cache migration policy that migrates data between the partitions in the hybrid architecture (see Figure 3). We use a two bit counter (*Cnt*) to manage migration same as reference [16]. The algorithm is as follows:

i) When the data is written in the cache for the first time, the two Cnt bits are set to 2b'11.

ii) Each time a **majority one** data is written into **STT-RAM** or **majority zero** data is written into **SRAM**, the counter value is incremented.

iii) Each time a **majority zero** data is written into **STT-RAM** or **majority one** data is written into **SRAM**, the counter value is decremented.

iv) When the *Cnt* MSB becomes zero, we swap the data between the cache partitions using an ultra-low-overhead swap memory without any performance loss. After swapping, the value of counters is initialized to 2b'11 again.

As Figure 2 shows, we add two bits Cnt (along with EB) to the tag store and during the write operation we update them based on our policy. The swap memory is dual ported memory which transfers the data between STT-RAM and SRAM data caches in non-critical write mode and with no impact on system performance. Our evaluation shows that using swap buffer with the capacity to hold 32 cache lines is sufficient to obtain zero swap latency for the studied workloads.

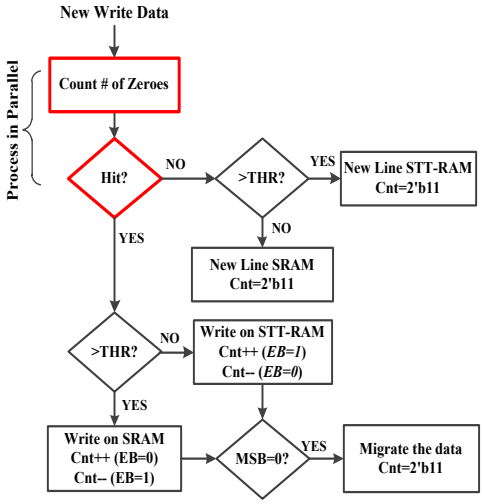

Figure 3. Proposed hybrid cache swap policy

The proposed hybrid architecture has many advantages over a conventional SRAM cache. The STT-RAM partition provides the following benefits:

- The main advantage of the proposed cache is its low leakage power due to use of NVM memory.
- Writing majority zero data on the STT-RAM cache improves the cache energy efficiency since STT-RAM cells consume 3.5X-6.5X less write energy for writing zero rather than writing one.
- Updating similar data (bit similarity) in the cache along with using EWT technique reduces the number of writes and hence cache dynamic energy
- For STT-RAMs, a prediction of up to 10^{15} write cycles is cited in previous works, but the best endurance test result for STT-RAM devices shown so far is much lower than this value [21]. Our cache decreases the number of writes on STT-RAM, improving effective lifetime of the cache.

High proportion of ones in SRAM cache motivates us to choose an efficient cell compatible with majority 'one' data. We experiment with both, conventional 6T-SRAM cell, and asymmetric 5T-SRAM cell in the SRAM partition. This partition provides the following benefits:

- Independent of the type of SRAM cell, updating pre-existing majority 'one' data with similar data type reduces the number of bit-level writes.
- Asymmetric 5T-SRAM cell in the SRAM cache consumes extremely low power to store one because the back-to-back invertor is in OFF mode.
- The new data distribution in 5T-SRAM cache improves the cache lifetime (reliability) due to the NBTI effect. This is because during write 'one', the supply voltage is applied to the gate of pull up PMOS transistor which reduces the PMOS stress mode.

4.3. THR and STT/SRAM aspect ratio

In the proposed hybrid cache, two design parameters affect the cache power efficiency: STT-RAM to SRAM cache aspect ratio and *THR* value. Although a large STT-RAM to SRAM aspect ratio improves the area and static power of the proposed hybrid cache, the write energy/latency of STT-RAM could slow down the cache operation or degrade the cache dynamic energy. This overhead can be controlled by setting *THR* to appropriate value. In a cache with smaller STT-RAM to SRAM aspect ratio, the SRAM partition consumes high leakage power since it occupies larger portion of the on-chip cache.

In recent workloads, it has been observed that the proportion of zero bits in the cache data is higher than the number of ones due to narrow-width feature [22], [23]. Our observation verifies this fact on several benchmarks such as SPEC CPU 2006 and Mibench and motivates us to choose the STT-RAM to SRAM aspect ratio as 2:1. In addition 2:1 STT-RAM to SRAM aspect ratio results in the same physical area for both memories in proposed cache since STT-RAM partition has ~2X higher density compared to SRAM at large sizes. We experiment with varying values of *THR* in order to balance the number of writes in STT-RAM and the number of swaps.

5. Experimental simulation

5.1. Experimental Setup

We extend the full-system cycle-accurate Simics [24] and GEMS [25] simulation platforms to model the proposed hardware support. The system configurations are listed in Table 1. We used circuit level *HSPICE* simulation in 45-nm technology to extract the cell features of STT-RAM and 5T-SRAM cell. We consider asymmetric characteristic of 5T-cell where writing 'one' is much slower than 'zero' on cell. The NVsim tool [26] is used to estimate power, performance and area of SRAM and non-volatile caches. However, general NV*sim* source code does not model the 5T-SRAM cell and asymmetric STT-RAM write characteristics. Therefore we modified the NVsim source code and verified our results using paper [27].

Table 1. Baseline processor configuration

Processor Configuration	
Frequency	UltraSPARC-III Cu processor core
L1 Instruction/ Data Cache	32KB, 2-way set-association, 64-byte block, Pseudo-LRU
L2 Cache (Hybrid Cache)	768KB (512KB STT-RAM, 256KB SRAM), 16-way set-associative (4-way SRAM, 12-way STT-RAM), 64-byte block
Main Memory	DRAM, DDR3, 4GB
Benchmarks	
SPEC CPU2006	milc, astar, Perlbench, games, leslie3d, namd, bzip2, gcc, h246ref, bvawes, sjeng, gobmk, wrf, hmmer, mcf,

5.2. Cache size

Table 2 compares the read and write latency characteristics of the STT-RAM (STT), 5T-SRAM (5T) and 6T-SRAM (6T) caches at different sizes. At small cache sizes, the read latency of STT-RAM and 6T cell is similar but 5T cell has slower read operation its weak back-to-back inverter. At larger cache sizes (>256KB), the read operation in STT-RAM is faster than both SRAM cells due to higher density and lower interconnect delay of STT-RAM. The write latency comparison between SRAM cells shows that 5T cell is faster than 6T cell due to its high controllability over the main body. We use SRAM-based write buffer in the STT-RAM partition to control the effective write latency experienced. At large cache sizes, the latency difference between SRAMs and STT-RAM reduces due to high STT-RAM density (Write latency: 5T < 6T ~ STT).

Table 2. Latency of the pure SRAM and STT-RAM caches

Cache size		**64KB**	**128KB**	**256KB**	**512KB**	**1MB**
Read Latency	STTRAM*	1.41ns	1.42ns	1.45ns	1.46ns	1.53ns
	5T-SRAM†	1.21ns	1.43ns	2.23ns	3.84ns	4.57ns
	6T-SRAM†	0.97ns	1.22ns	1.77ns	3.53ns	4.19ns
Write Latency	STTRAM	5.26ns	5.26ns	5.29ns	5.30ns	5.34ns
	5T-SRAM	0.83ns	1.04ns	1.28ns	2.28ns	3.16ns
	6T-SRAM	0.91ns	1.19ns	1.75ns	3.51ns	4.17ns

* STT-RAM optimized based on write energy-delay product
† SRAM optimized based on leakage power

Figure 4 shows the dynamic energy and static power consumption of hybrid caches at different sizes. STT-RAMs are slow and power hungry for the write operation. At larger sizes, the dynamic energy growth tends to slow down due to higher STT-RAM density with respect to SRAMs. Storing majority zero and one data on STT-RAM and SRAM caches improves power efficiency of the proposed design because: (i) Asymmetric write in STT-RAM and 5T-SRAM cells makes them efficient cells for majority zero and one data respectively. (ii) The number of writes in STT-RAM and SRAM caches decreases due to writing similar data bits in caches and using write optimizations. (iii) Based on the *EB* flag bit value, the system can activate only the SRAM or STT-RAM partition for the read operation. These reasons improve the dynamic energy efficiency of the proposed cache compared to a conventional SRAM cache. This energy improvement is greater with STT/5T cache because data activity has more impact on 5T-SRAM energy efficiency than 6T-SRAM (Dynamic energy: STT/5T < STT/6T < SRAM).

In contrast to high leaky 6T cell, the leakage power of 5T cell is data dependent. Therefore storing 'one' in 5T-SRAM results in 1.7X lower power consumption than storing 'zero' because the pull-up PMOS gates the supply voltage from the storage node. In contrast, STT-RAM being a non-volatile memory cell consumes zero leakage power to save the data. Figure 4 compares the static power of the hybrid and SRAM caches at different sizes. The graph shows that the proposed STT/6T and STT/5T caches consume 67% and 78% lower static power compared to conventional SRAM cache (Static power: STT/5T < STT/6T << SRAM).

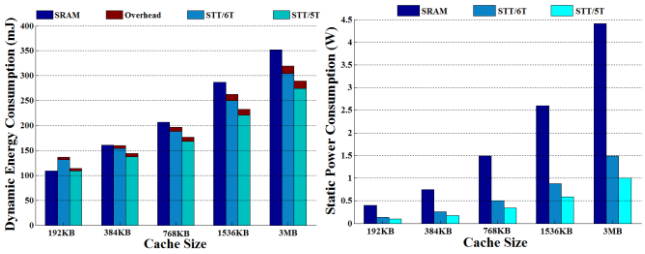

Figure 4. Energy dynamic and static energy of hybrid cache in different size

Figure 5 shows the normalized performance of the proposed hybrid caches at different sizes. The total latency comprises of write latency, read latency and the latency overhead of parallel counter, comparator and swap mechanism. Large cache sizes improve the performance of both STT/6T and STT/5T caches due to high density of STT-RAM with respect to SRAM. This improvement is less with STT/5T cache because of the slow read operation of the 5T cell. As Figure 5 indicates, the proposed cache achieves higher performance than pure SRAM cache for caches larger than 384KB. This improvement is 17.9% and 12.6% (11.3% and 4.1%) with 3MB (768KB) STT/6T and STT/5T caches respectively, with *THR*=50%.

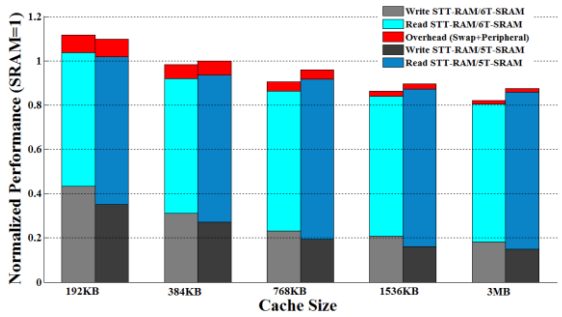

Figure 5. Normalized Performance of hybrid cache in different sizes

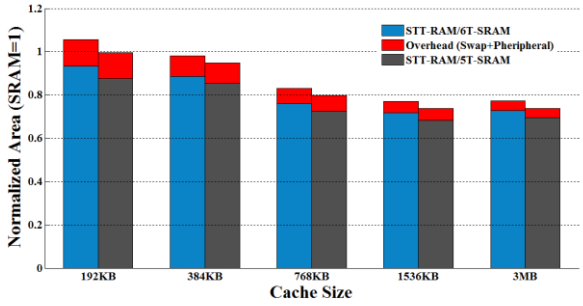

Figure 6. Normalized area of hybrid cache in different sizes

5.3. Area efficiency

Figure 6 shows the normalized area of hybrid cache at different sizes. This area consists of STT-RAM, SRAM and area overhead of peripherals containing swap memory, data splitting (counter & comparator), and *EB* flag bit and two bit *Cnt* in the tag store. We used *Synopsis Design Compiler* to determine the area overhead of the parallel counter and comparator in 45-nm technology. The area of added flag and *Cnt* bits is estimated by modifying the NV*sim* source code. For 5T cell we draw the layout of the cell in 45-nm rule and added the features to NV*sim* tool. As results show, the area efficiency of the proposed caches improves for large memory sizes. We observe that 768KB STT/6T and STT/5T have 16.9% and 20.3% area efficiency with respect to the SRAM cache.

5.4. THR effect

A. THR effect on energy

The proposed hybrid cache changes the data distribution in SRAM and STT-RAM partitions depending on *THR* value. Figure 7 and Figure 8 show the effect of *THR* value on the energy consumption and performance of proposed hybrid cache normalized to pure 6T-SRAM cache in 768KB size. A large *THR* value increases the percentage of zeros in STT-RAM cache. This significantly affects the power efficiency of hybrid cache due to: (i) asymmetric write operation in STT-RAM where writing zero needs much lower energy than writing one; (ii) writing data with high bit similarity on STT-RAM and SRAM caches. In addition, the EWT technique further decreases the write energy of STT-RAM partition by checking the input data with pre-stored cache value bit by bit before each write operation. Figure 7 shows that *THR* value greater than 75% does not have positive impact on STT/6T cache energy efficiency since it pushes more random bit distribution in SRAM partition. This unbalanced data distribution increases the probability of data migration due to limited SRAM cache capacity.

The power overhead of data migration is the main reason of high dynamic energy consumption of hybrid cache in large *THR* values. Table 3 shows the percentage of time the data migration scheme is active at different *THR* values. Large *THR* stores data with high percentage of zeros in the STT-RAM part, so that a large portion of data is saved in the SRAM cache. This unbalanced data distribution between caches increases the probability of data migration since there is higher probability that new incoming data will be updated with low percentage of zero values (< *THR*) in the STT-RAM partition. One possible way to handle the migration is to use larger STT-RAM to SRAM cache ratio. However, as we explained before the SRAM cache increases the static power of SRAM cell and degrades the advantages of using STT-RAM.

Table 3. Swap memory activation time in different THR

THR	40%	50%	60%	70%	80%	90%
Average swap activation (%)	1.73	2.15	3.27	4.18	7.63	9.27

This energy tradeoff becomes more complicated for STT/5T hybrid cache where large *THR* reduces the percentage of ones and degrades 5T-SRAM energy efficiency. This is because writing ‘one’ in this cell needs more time but less energy than writing ‘zero’. This extra energy term pushes the minimum energy point of STT/5T cache to a smaller *THR* value than STT/6T (see Figure 7). The error bar in the Figure 7 shows that the variation in energy consumption of STT/5T is larger than STT/6T cell

which is due to data dependency and asymmetric 5T-cell characteristic. Our future work is to set *THR* value dynamically based on the available STT-RAM and SRAM free capacity and application type so that the proposed cache can operate at a minimal energy point while executing each application.

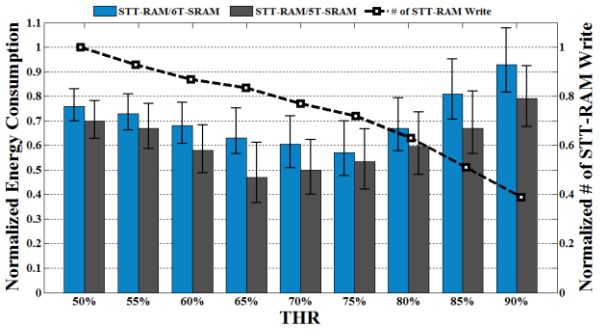

Figure 7. Normalized energy consumption of the proposed hybrid cache in different THR

B. THR effect on latency

Both STT-RAM and 5T-SRAM cells have asymmetry and data dependency in their write characteristics. Therefore, in STT/6T cache, a high *THR* results in faster write operation because majority of the data is pre-stored in the SRAM partition, which is fast for both read and write operations. This also improves the performance by reducing the average time that the STT-RAM write buffer is full. The hybrid cache uses the low write latency of SRAM and low read latency of STT-RAM to improve the total performance. This performance improvement is not linear with *THR* value because after a while the slope of performance improvement gets saturated due to limited SRAM capacity and high number of swaps. These two factors degrade the performance of hybrid cache for *THR* value larger than 65%. Although the number of swaps increases at large *THR* values (> 65%), due to low write energy of STT-RAM cache, the total cache energy improvement continues up to *THR*=75% (see Figure 7 & Figure 8).

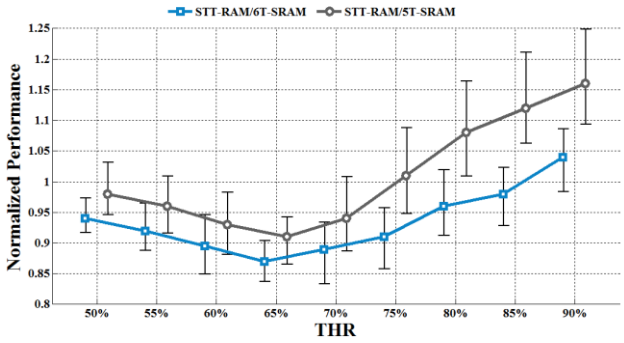

Figure 8. Normalized performance of the proposed hybrid cache in different THR

In STT/5T cache, higher *THR* value improves cache write performance since writing zero in 5T-cell is faster than writing one. However, the read operation of 5T-SRAMs is much slower than 6T-SRAM cells. This, combined with number of data migrations, SRAM capacity, and number of writes in STT-RAM cache, results in performance degradation for *THR* greater than 65%. Figure 9 compares the energy consumption of the proposed caches with respect to a pure SRAM cache running SPEC CPU 2006 benchmarks. In 768KB size, hybrid STT/6T and STT/5T caches achieve to minimum energy points of 42% and 53% respectively which are corresponding with THR=75% and THR=65%. These energy savings are achieved when the hybrid caches operate 9% faster on average compared to SRAM cache, and the performance of all benchmarks meet the SRAM latency.

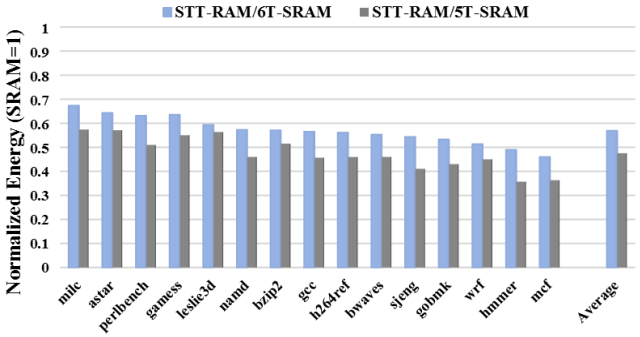

Figure 9. Normalized energy consumption of proposed hybrid cache for SPEC benchmarks

5.5. Cache lifetime

NBTI degrades the reliability and lifetime of digital devices. In case of SRAM, NBTI decreases both SNM stability and the cell performance depending on the average time that the PMOS gates are biased with zero input. The cache soft-error is a function of the average SNM [28]. In 6T-SRAM cell in Figure 1a, the Q and QB nodes save complementary values. When the Q node is zero, the gate of M3 and M4 transistors are biased with “0” and “1” respectively. In other words, when one of these transistors is in stress mode, the other one is in recovery mode. We used the NBTI model in paper [29] to evaluate the change in the threshold voltage of PMOS transistors. In 5T-SRAM cell, NBTI has huge impact on the stability because the cell just has one pull-up PMOS transistor. The average SNM stability improvement of 5T-SRAM after three years NBTI effect with 1000 Monte Carlo simulation and 10% variations are listed on Table 4. We used statistical evaluation to extract the data activity of SRAM part of the proposed cache in different THR running SPEC CPU 2006 benchmarks. Data in proposed cache increases the percentage of the ones in SRAM part. Although this increment does not have any effect on 6T-SRAM lifetime due to its symmetry, this improves the SNM degradation of the 5T-cell by 35.1% after three years NBTI effect with *THR*=65%.

To further reduce the write energy on STT-RAM cache, we use early write termination techniques. Before write operation, this technique compares the bitline values with input data bit by bit and neglects writing the same bits on the cache. In addition, we are writing mostly zero data in the cache line with majority zero. So, EWT significantly reduces

the number of writes on STT-RAM cache depending on the value of *THR*. As Table 4 shows, the proposed cache decreases the number of writes in STT-RAM by 28.9% using *THR*=75%.

Table 4. Endurance improvement of 5T-SRAM and STT-RAM on different THR values

THR	50%	60%	70%	80%	90%
5T-SRAM SNM improvement	40.2%	37.3%	32.1%	26.3%	21.7%
STT-RAM endurance improvement	23.4%	26.4%	28.7%	32.1%	37.6%

Acknowledgment

This work was sponsored by NSF grant #1527034.

6. Conclusion

We propose a new data-aware hybrid cache architecture which saves the data on STT-RAM or SRAM cache based on the percentage of ones. Data splitting reduces dynamic energy consumption of the proposed cache since STT-RAMs are so efficient for majority zero data. In addition, we examine the impact of utilizing 5T-SRAM cell in SRAM part which has ability to work efficiently for majority 'one' data (~1.7X energy saving). Our evaluation shows the proposed cache improves the cache energy efficiency while running variety of benchmarks by (i) changing the data distribution in STT-RAM and SRAM caches (ii) decreasing the number of writes by writing majority zeros (ones) data on cache with majority zeros (ones) and (iii) using low leakage power and high density STT-RAMs. In addition, the proposed cache improves the cache endurance in both STT-RAM and 5T-SRAM parts by storing suitable data in them appropriately.

7. References

[1] H. Qin, et al., "SRAM leakage suppression by minimizing standby supply voltage," *IEEE ISQED*, pp. 55-60, 2004.

[2] C. J. Xue, et al., "Emerging non-volatile memories: opportunities and challenges," *IEEE/ACM CODES+ISSS*, pp. 325-334, 2011

[3] Y. Kim, et al., "CAUSE: critical application usage-aware memory system using non-volatile memory for mobile devices," in Proceedings of the IEEE/ACM ICCAD, pp. 690-696, 2015.

[4] C. W. Smullen, et al., "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," *IEEE HPCA,* pp. 50-61, 2011.

[5] S. Kang, et al., "Performance trade-offs in using NVRAM write buffer for flash memory-based storage devices," *IEEE Computers*, vol. 58, pp. 744-758, 2009.

[6] P. Zhou, et al., "Energy reduction for STT-RAM using early write termination," *IEEE/ACM ICCAD*, pp. 264-268, 2009.

[7] Z. Sun, et al., "Multi retention level STT-RAM cache designs with a dynamic refresh scheme," *IEEE/ACM Microarchitecture*, pp. 329-338, 2011.

[8] J. Li, et al., "STT-RAM based energy-efficiency hybrid cache for CMPs," *IEEE VLSI-SoC*, pp. 31-36, 2011.

[9] A. Nigam, et al., "Delivering on the promise of universal memory for spin-transfer torque RAM (STT-RAM)," IEEE/ACM ISLPED, pp. 121-126, 2011.

[10] M. Imani, et al., "Resistive Configurable Associative Memory for Approximate Computing," *IEEE DATE*, 2016.

[11] M. Imani, et al., "MASC: Ultra-Low Energy Multiple-Access Single-Charge TCAM for Approximate Computing," *IEEE DATE*, 2016.

[12] J. Wang, et al., "A coherent hybrid SRAM and STT-RAM L1 cache architecture for shared memory multicores," *IEEE ASPDAC*, pp. 610-615, 2014.

[13] A. Jog, et al., "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs," *IEEE DAC*, pp. 243-252, 2012.

[14] S. Mittal, et al., "AYUSH: A Technique for Extending Lifetime of SRAM-NVM Hybrid Caches," *IEEE Computer Architecture Letters*, 2014.

[15] A. Jadidi, et al., "High-endurance and performance-efficient design of hybrid cache architectures through adaptive line replacement," *IEEE/ACM ISLPED*, pp. 79-84, 2011.

[16] X. Wu, et al., "Power and performance of read-write aware hybrid caches with non-volatile memories," IEEE DATE, pp. 737-742, 2009.

[17] M. Imani, et al., "Hierarchical design of robust and low data dependent FinFET based SRAM array," IEEE/ACM NANOARCH, pp. 63-68, 2015.

[18] A. Teman, et al., "A 40-nm sub-threshold 5T SRAM bit cell with improved read and write stability," *IEEE Circuits and Systems II: Express Briefs*, vol. 59, pp. 873-877, 2012.

[19] C. T. Chuang, et al., "Back-gate controlled asymmetrical memory cell and memory using the cell," *Google Patents*, 2007.

[20] M. Jafari, et al., "Analysis of power gating in different hierarchical levels of 2MB cache, considering variation," International Journal of Electronics, pp. 1-15, 2015.

[21] Y. Chen, et al., "Processor caches built using multi-level spin-transfer torque ram cells," IEEE ISLPED, pp. 73-78, 2011.

[22] D. Brooks, et al., "Dynamically exploiting narrow width operands to improve processor power and performance," *IEEE HPCA*, pp. 13-22, 1999.

[23] G. Duan, et al., "Exploiting narrow-width values for improving non-volatile cache lifetime," *IEEE DATE*, p. 52, 2014.

[24] P. S. Magnusson, et al., "Simics: A full system simulation platform," *IEEE Computer*, vol. 35, pp. 50-58, 2002.

[25] M. M. Martin, et al., "Multifacet's general execution-driven multiprocessor simulator (GEMS) toolset," *ACM SIGARCH Computer Architecture News*, vol. 33, pp. 92-99, 2005.

[26] X. Dong, et al., "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE ICCAD*, vol. 31, pp. 994-1007, 2012.

[27] R. Bishnoi, et al., "Architectural aspects in design and analysis of SOT-based memories," *IEEE ASPDAC*, pp. 700-707, 2014.

[28] E. H. Cannon, et al., "The impact of aging effects and manufacturing variation on SRAM soft-error rate," *IEEE Transactions on Device and Materials Reliability*, vol. 1, pp. 145-152, 2008.

[29] T. Grasser, et al., "The paradigm shift in understanding the bias temperature instability: from reaction–diffusion to switching oxide traps," *IEEE Electron Devices*, vol. 58, pp. 3652-3666, 2011.