Reducing Power in Hybrid MRAM Cache through Iterative Low Voltage Writes

Reyhan Jabbarvand, Houman Homayoun
George Mason University, Technical Report, September 2013

Abstract— **As technology scales down, the static leakage power of SRAM based cache becomes a more critical source of dissipated power, particularly for large last level cache where leakage power is high. The emerging non-volatile Spin Transfer Torque (STT-RAM) is a nominee to substitute SRAM due to low leakage power. However, considerable high energy and long latency of STT-RAMs for write operations is a barrier to their commercial adoption. To address this problem, we propose a hybrid non-uniform cache architecture (NUCA) of SRAMs and STT-RAMs with different operating voltage/ pulse width settings. Operating at low voltage increases the probability of failure. To alleviate this, we propose VITERAT, a technique that reduces STT-RAM write access energy by lowering voltage while ensures correctness by retrying the failed writes. Simulation results indicates overall 30~40% power gain for various workloads in hybrid cache architecture. This comes with negligible performance cost (less than 2%) performance penalty.**

I. INTRODUCTION

In modern designs, the last level cache (LLC) is typically a shared cache, configured to be large enough to effectively cover the data of all private caches, and store enough data to protect the off-chip memory and interconnect from the vast majority of on-chip accesses of all cores. As a result, the LLC is typically one of the largest consumers of power on the processor. Moreover, leveraging the size of LLC increases the latency. Thus, power consumption and latency are the major concerns of LLC design. The latency of larges caches can be improved through Non-Uniform Cache Architecture (NUCA) [1]. In NUCA, a large cache is divided into multiple banks with different access latencies which results in reducing average access latency of the cache.

New advancements in non-volatile memory technology such as Magnetic RAM (MRAM), and Phase-change RAM (PRAM), have made them competitive not just with DRAM, but also with SRAM. These new memory technologies offer significantly different power and performance characteristics compared to standard SRAM based cache design. This is particularly true for large last-level caches where static leakage is high – replacing SRAM with non-volatile memory is attractive because it virtually eliminates the leakage energy consumed by memory cells.

Recent studies have shown that traditional SRAM NUCA can be outperformed by using different memory technologies instead of a single technology. Hybrid Cache Architecture (HCA) can improve performance, latency, and power consumption of large caches by diving it into multiple banks with different power and performance characteristic [2].

This paper examines the use of Magnetic RAM (MRAM), one of the emerging non-volatile technologies, in hybrid cache architecture. Our studied HC architecture integrates non-volatile high density STT-RAMs along with energy efficient SRAMs to build a hybrid on-chip cache architecture. Although MRAMs are attractive due to fast read access, low leakage power, high bit density, and long endurance, their high write power and slow access of write operations are a big concern. Among the

mentioned barriers, write latency is less of a concern as it can be hidden by using large buffers.

This paper introduces VITERAT a Variable Voltage and Pulse Width Iterative Writes Mechanism for low power MRAM caches. VITERAT reduces the write voltage of STT-RAM to gain significant power savings. While voltage scaling has a super-linear effect on reducing power, it exponentially increases the defect rate in STT-RAMs. By using a voltage pulse that is allowed to fail with some probability, VITERAT retry the failed writes and still save significant overall energy. This is shown with multi-core and multi-thread simulation results. With minor design modifications, we can achieve a 63% reduction in cache write power, and as much as 26% reduction in overall cache power across various workloads. This paper makes the following major contributions:

- Highlights the dependence of write error rate of MRAMs to write pulse width and write voltage. It also demonstrates how this would affect the power dissipation of large last level cache.
- Proposes the novel concept of Multiple Variable Voltage and Pulse Width (MLVW) writes that enables low power and reliable LLC with non-uniform cache access at near threshold voltage.

- Examine various uniform and non-uniform assignments of voltage and pulse width pair to different MRAM banks.
- Examine the effectiveness of the technique in a Hybrid NUCA cache with extensive architecture simulations.

The rest of the paper is organized as follows. Section II provides a background on NUCA and STT-RAM. Section III gives an overview of the entire problem and the proposed solution. In section IV, we discuss the proposed methodology by providing simulation results. Section V reviews related works on STT-RAM low-power management, and finally Section VI concludes the paper.

II. BACKGROUND

A. Hybrid cache

Utilizing HCA can be on inter cache level (LHCA) or intra cache level (RHCA) [5]. In LHCA, each level of on-chip hierarchy constitutes of different memory technology. In RHCA or Region-Based Hybrid Cache Hierarchy, a single level of cache consists of multiple memory banks with diverse

TABLE I. COMPARISON OF DIFFERENT CACHE TECHNOLOGY

Features	SRAM	STT-RAM
Density	Low	High
Write Power	Low	High
Read Power	Low	Low
Write Speed	Very Fast	Slow
Read Speed	Very Fast	Fast
Leak. Power	High	Low
Non-volatile	No	Yes

technologies. In this work, we have used Region based HCA (RHCA) through LLC which have shown to reduce power consumption and improve latency of LLC. Table I lists and compares features of major memory technologies. As shown, STT-RAM has a very different read and write features in terms of latency and power consumption, with particularly high write power consumption. Compare to SRAM, STT-RAM has relatively low static power due to its non-volatile characteristic. Therefore, a hybrid cache of SRAM and STT-RAM can benefit low leakage power (STT-RAM), high density (STT-RAM), and low dynamic power (SRAM).

B. STT-RAM

STT-RAM is a new generation of Magnetic Random Access Memory (MRAM). MRAM is one of the most mature technologies among the emerging non-volatile memory technolgoies. In a conventional MRAM, a STT-RAM cell uses Magnetic Tunnel Junction (MTJ) as a basic storage element to store binary data. Each MTJ is composed of two ferromagnetic layers; one has a fixed magnetization direction and the other has a free one. The relative magnetization direction between the reference layer and free layer results in different resistance of MTJ, which is used to represent the binary data stored in the cell. Since the free layer can change its direction, the relative direction of these two is used to represent a digital '0' or '1'. The latest STT-MRAM technology uses the current flow from bit-line to source-line to change the direction of the MTJ free layer as shown in Fig 2. When writing '0' state into an STT-RAM cell, positive voltage polarity is applied between bit-lines and source-line; when writing '1', negative polarity is established.

Unlike previous MRAM designs, STT-RAM uses spin-polarized current flowing through the MTJ to change the direction of magnetic direction in the free layer. Hence, STT-RAM uses much lower switching current than conventional MRAMs, which makes it feasible for cache or memory. In addition, STT-RAM's threshold current decreases as the size of MTJ becomes smaller, resulting in better scalability than previous MRAM designs [3]. The design of STT-RAM cell array is shown in Fig. 1.

III. POWER OPTIMIZATION ON NUCA HYBRID CACHE

A. Overview

Although STT-RAM technology provides low leakage power and high density, its high power consumption which mostly relates to write operations, is still a big concern. The write operation of an STT-RAM cell is a probabilistic function of the pulse voltage due to the thermal field torque is acting on the free layer magnetization. The write speed

of an STT-RAM cell is decided by the switching behavior of its free layer, which is influenced by various torques such as spin, thermal field and easy plane anisotropy.

Studies showed that STT-RAM cell write latency can be traded for energy and power consumption [4]. We used same methodology as Nigam et al. [5] to estimate the behavior of Write Error Rate (WER) as a function of pulse width and voltage. According to Fig. 3.a, as we slow down the write operation, by reducing the pulse width for instance, we can gain lower power consumption for a given error rate. It also demonstrates that a desired WER can be achieved by different design points (voltage/pulse width). For example, for a given WER of 1e-2, we can choose the voltage either of 0.89v, 0.62 v, and 0.5v for 3ns, 5ns, and 5.5ns pulse widths, respectively.

Fig. 3.b shows write power consumption of STT-RAM as a function of voltage in our studied architecture. As the figure shows, for a given pulse width, by reducing the voltage the power reduces almost linearly.

B. VITERAT

In this paper, we propose VITERAT, which is a mechanism to optimize power and energy in NUCA hybrid caches by lowering voltage and increasing pulse width. VITERAT employs MLVW writes when WER is sacrificed for power, in order to compensate the error rate. The MLVW write scheme exploits iterative low energy writes and variable voltage/pulse width settings for STT-RAMs to meet low power. For every block write, MLVW: (a) performs a write operation; (b) reads back the contents of the written STT-RAM or SRAM cells; and (c) compares the written and the original values to determine whether the write succeeded. If any cell failed to store the written value, the process starts over to remaining faulty bits, with lower energy compared to ordinary write.

There are several mechanisms proposed in the literature to reduce overhead due to redundant writes [6]-[7]. We assume a write mechanism that can identify and bypass correctly written bits with bit granularity based on bit write mask. Initially, the stored bits are read from the array cells and compared with the head of the write buffer to form the initial write mask. We refer to set bits in the initial bitmask as the non-identical bits. If the mask is non-zero, we perform the first write operation for non-identical bits. On completion, the written bits are read back and compared again with the write buffer, updating the write mask, which now indicates which bits should be written in the second iteration of writes. The process repeats until all non-identical bits have been successfully written to the device cells. As the number of bits to be written decrease in consequent writes, we can consider lower write energy for them.

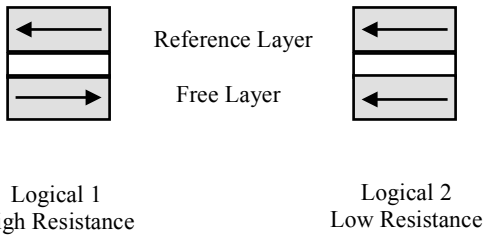


Fig. 1. Conceptual view of MTJ structure [3]

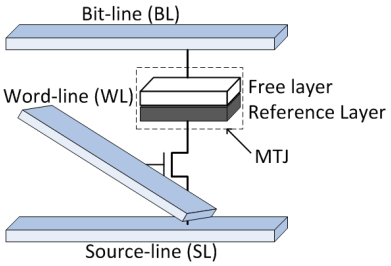
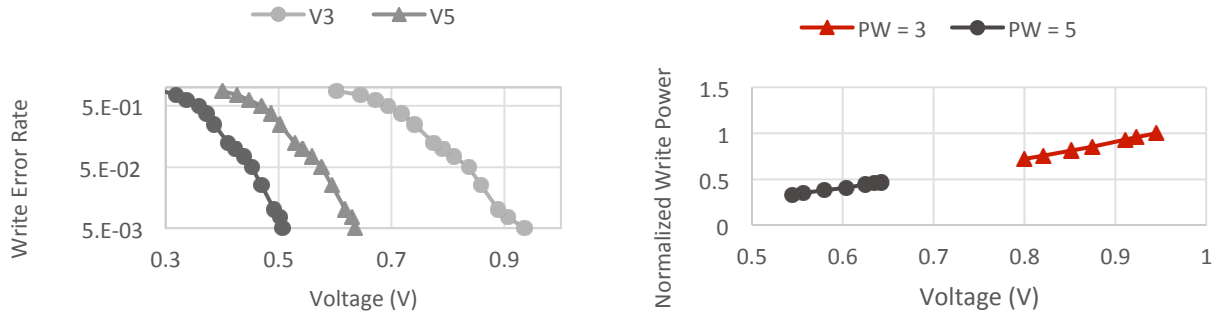


Fig. 2. The conceptual view of an STT-RAM cell.



(a) (b)

Fig. 3. a) Write error rate as a function of voltage and pulse width for write operations b) Power vs. voltage with 3ns and 5ns pulse width for single write access to cache

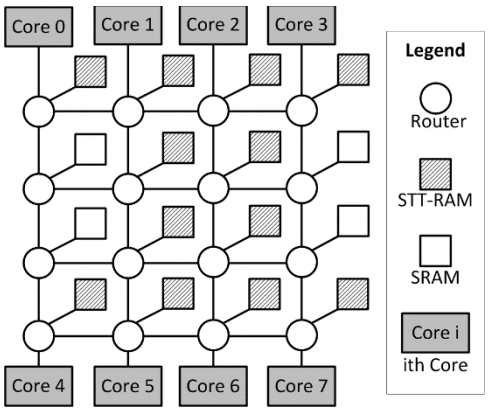


Fig. 4. Hybrid Cache Architecture

The main objective of VITERAT is to optimize power consumption without sacrificing write success and system performance. Long write latencies of STT-RAMs can have a significant impact on performance. It is essential to hide this delay, especially for MLVW technique which attempts multiple writes to ensure accuracy. In architecture level, one solution is to use a store buffer that pools pending writes until the complete [8]. In this paper, we simply add up the size of buffers for the cases use MLVW techniques without imposing further overhead to the system.

The setting of voltage and pulse width pair for different bank in NUCA cache have significant impact on power, and performance of the system. In fact, the power/performance tradeoff depends on the mapping policy and fault-tolerance threshold. Hence, several heuristics can be introduced for various voltage and pulse width pair assignment. Here are examples of such heuristics:

- Higher voltage/pulse width pair for frequently accessed STT-RAMs: one can set higher voltage and pulse width on banks which accessed frequently to avoid iterative writes, while assigns lower voltage and pulse width to less accessed banks to improve the total power consumption of the cache.
- Lower voltage/Higher pulse width pair for frequently accessed STT-RAM: For applications that are not sensitive to higher bit error rates, one can assign low voltage/high pulse width to the STT-RAM banks that are frequently accessed and define a threshold on number of iterative write which ensures correctness.
- Higher voltage/pulse width pairs for banks closer to the cores: For applications that are sensitive to latency, most of the cache accesses are made on banks that are adjacent to the cores. Therefore, the access latency of adjacent banks should not exacerbate by applying iterative writes.
- Lower voltage/Higher pulse width pair for banks closer to cores: This setting can be exploited in cache with large number of banks. The closer banks are to cores, the less access latency they will have. Hence, they can tolerate a higher write latency of iterative writes.

The improvement on power varies for each suggested setting. In this paper, we have used the third heuristic as it gives the best power results (results for other heuristic are not presented due to space limitation).

IV. EVALUATION

In this section, the proposed methodology is evaluated using SPEC2K and SPEC2K6 benchmarks for 8-thread workloads. We first introduce simulation environments, methodology, and design parameters. Then, we demonstrate how our methodology improves power efficiency of hybrid NUCA caches.

A. Methodology

Our baseline architecture is shown in Fig. 4. In our architecture, the LLC is divided into multiple SRAM and STT-RAM banks where banks are interconnected as a 2-D mesh via a network-on-chip (NoC). The LLC is shared across all cores. Each LLC bank includes multiple cache sets, with a set of lines (blocks with the same index) all mapped to the same bank. The baseline design assumes a NUCA LLC. With multiple banks within the LLC, we have the choice of either always putting a block into a designated bank (static mapping) or allowing a block to reside in one of multiple banks (dynamic mapping). We consider static mapping in our baseline design and model static NUCA policy for CMP architectures (CMP-SNUCA). CMP-SNUCA statically partitions the address space across cache banks connected via a 2-D mesh interconnection network.

For the three-level cache hierarchy, we use a 512KB L2 cache and 6MB hybrid L3 cache consist of 12 STT-RAM and 4 SRAM banks with capacity of 512KB and 64 KB respectively. According to the results in [9], 64KB SRAM and 512KB STT-RAM has almost similar area. Hence, we can replace SRAMs with high bit density STT-RAMs.

For cycle-accurate CMP + NUCA cache simulation we integrated SMTSIM [24], a multi-threading simulator, and BOOKSIM [22], a cycle accurate interconnection network simulator, in order to evaluate VITERAT mechanism. We simulate an 8-core CMP machine similar to the architecture shown in Fig. 4. The NoC topology is the conventional 4×4 2-D mesh and it adopts wormhole switching with 8-flit packets and 8-flit buffers for one virtual channel. Table II shows the design parameters and configuration applied to our hybrid cache which are derived from NVSIM simulator [12]. We assumed drowsy cache technique applied to SRAM banks, which reduces their leakage power by up to 80% [13]. The total capacity of the hybrid LLC that consists of 12 STT-RAM banks and 4 SRAM banks, with capacity of 512KB and 64KB respectively, is 8MB.

We have selected the X-Y routing for delivering read and write requests from cores to cache bank. Our baseline architecture employs write-forwarding technique. The motivation for write-forwarding is the high write energy of STT-RAM banks relative to SRAM banks.

We have augmented BOOKSIM by Orion2.0 [14] in order to investigate the power consumption of NoCs. The power results are based on a 64-bit NoC implemented in 45nm technology and the frequency of 500 MHz. To model cache power, we assume constant leakage power and read energy per access, but different energy per STT-RAM write access, depending on write voltage and write pulse width. The total read/write energy then derived by summing the read/write energy per access for total number of reads/writes. For simplicity, our simulation methodology assumes that every subsequent bit write operation occurs with the same latency and power overhead as the first operation. This assumption addresses the worst case scenario as subsequent write operations consume less energy due to smaller number of bits to be written. We only reduce packet size for subsequent write operations in sake of network congestion.

The results are based on simulations for several combinations of 8-thread workloads. As the significant barrier of STT-RAMs is their high power consumption of write operations, we categorized our benchmarks based on their read and write operation intensity in order to compare the impact of our proposed method. We studied the behavior of each SPEC2K and SPEC2K6 benchmark as single-thread workload simulation on SMTSIM. We then categorized them first by number of access to memory, as Memory Intensive (MI) and Memory Non-Intensive (MNI), and then by number of read/write access, as Read Intensive (RI) and Write Intensive (WI). An example of each different workload is provided on Table III. For each application in the mix we fast-forward to skip the initialization phase and then simulate until all threads execute 200 million instructions.

B. Results and Aalysis

Fig. 4.a show how WER changes by decreasing voltage for different pulse width. We choose 9 values for WER, from 5e-1 to 5e-3. There are three pairs of different voltage/pulse width for a given WER, hence requires 27 samples of simulation to compare their power.

MLVW act as a fault tolerance technique in presence of write failure. To account for write-forwarding scheme we safely assumed an access pattern of 50-50, which equally distribute access between SRAM and STT-RAM.

TABLE II. BASELINE DEIGN PARAMETERS

Size	6MB
Banks	64KB SRAM (4 banks), 512 KB STT-RAM (12 banks)
Associativity	8
Read Latency	SRAM: 3 cycles STT-RAM: 9 cycles
Write Latency	SRAM: 3cycles STT-RAM: 31 cycles
Read Energy	SRAM: 0.037 nJ STT-RAM: 0.078 nJ
Write Energy	SRAM: 0.037 nJ ST-RAM: 0.265 nJ
Leakage Power per bank	SRAM: 17 mW STT-RAM: 20 mW

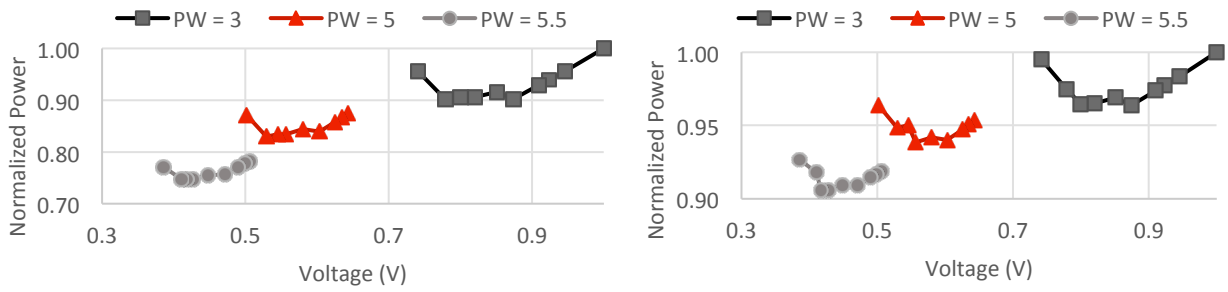
TABLE III. EIGHT THREAD WORKLOAD EXAMPLE

MI	mcf_06, art_470, applu, lbm_06, leslie3d_06, libquantum_06, mgrid, gcc_06_typeck
MNI	fma3d, h264ref_sss_encoder_main_06, namd_06, sphinx3_06, equake, povray_06, crafty, gzip_log
WI	namd_06, mesa, applu, vpr_route, lucas, apsi, gap, sphinx3_06
RI	vortex_3, gobmk_06_nngs, crafty, galgel, omnetpp_06, art_470, sixtrack, povray_06

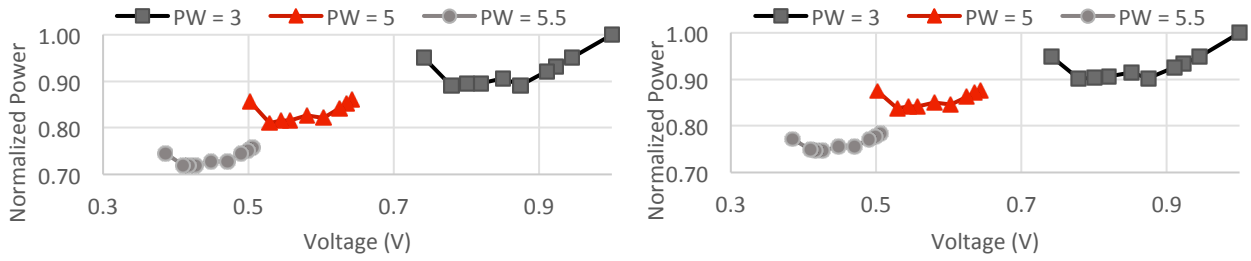
Most of the power consumption of the LLC-NoC stems from the STT-RAM and SRAM banks. Fig. 5 demonstrates the relationship between power consumption of hybrid SNUCA cache and the write voltage for different pulse width. Note that all markers on the figure indicate same WER as we applied MLVW technique to tolerate the failures. Note that, as reported in Fig. 3.a, for the studied architecture with low write pulse width, the slope of write error rate vs. voltage is not steep. Therefore, we only need to repeat writes twice to reach to above 97% write success. There are the following observations on Fig. 5:

- We gain more power by lowering the voltage for a given pulse width without sacrificing accuracy. This gain diminishes by increasing pulse width while decreasing voltage.
- The increase in power for a given pulse is due to multiple write operation. In error rates less than 75% for a given pulse width, the increase is negligible as the numbers of erroneous bits to be rewritten is small and the iterative writes needs slight energy to perform. For higher error rates in other hand, there will be a spike on power by applying MLVW. However, we can trade power with error rate for higher WER and gain same power, or even lower, utilizing MLVW.

Fig. 6 compares total power and latency of the baseline configuration and MLVW configuration for different workloads. It can be observed that for NMI workloads, power gain is negligible comparing to other workloads and the latency overhead exceeds power gain. The reason is that for such workloads, the leakage power dominates dynamic power and

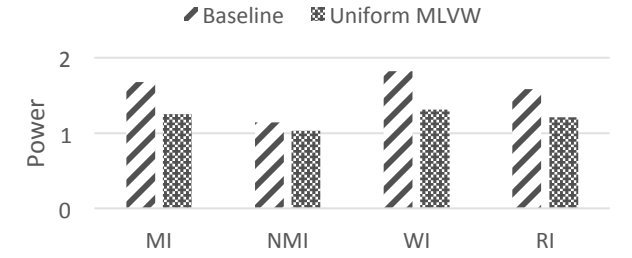


(a) (b)

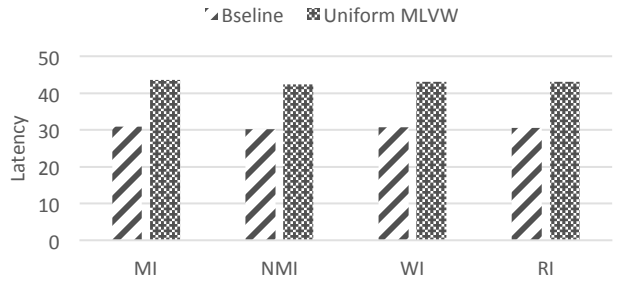


(c) (d)

Fig. 5. Comparison of power consumption vs. voltage for different pulse width a) MI workload b) MNI workload c) WI workload d) RI workload



(a)



(b)

Fig. 6. Comparison of a) power and b) latency for baseline, 1v voltage/3ns pulse width and MLVW, 0.4v voltage and 5.5ns pulse width

therefore it eliminates the benefit of MLVW. According to Fig. 6.a, we can gain 25%, 9%, 28%, and 24% power reduction in MI, MNI, WI, and RI workloads, respectively. Fig. 6.b shows the latency of the network increased by 29% after applying MLVW.

We then studied the impact of MLVW technique on performance. Simulation results indicate for most applications, the performance is not sensitive to latency. For MI, NMI, WI, and RI workloads, we observed MLVW technique improves power with affecting performance, measured by IPC, by 3%, 0.4%, 1.9%, and 2.7% respectively.

So far, we assumed all of STT-RAM banks on the hybrid cache to have the same setting for the write voltage and the pulse width (uniform setting). To reduce the overhead of MLVW on latency, we assumed non-uniform setting where STT-RAM banks can have different voltage and pulse width pair. One heuristic to reduce the overhead of MLVW is to decrease the voltage of STT-RAM banks, which are close to cores (banks 0-3 and 12-15 in Fig. 4) as low as possible. The reason is that the traffic on corresponding routers of these banks are significantly lower than STT-RAMs in the middle (banks 5, 6, 9, and 10) and MLVW will slightly impact the latency (Note that while this is true for our SNUCA architecture for DNUCA, the congestion pattern might be different due to swapping and migration). For the STT-RAM banks in the middle then, we lower the voltage so that there would be a need to apply multiple write to compensate loss in accuracy. New result illustrates that exploiting non-uniform settings improves total power by 26%, 10%, 29%, and 24% for MI, NMI, WI, and RI workloads. It also alleviate the loss on cache

access latency by reducing the growth of latency to 15% instead of 29%. It also affects performance in terms of IPC for less than 1% across all workloads.

Taking into account that changes in overall cache power depends on how effectively write operations are forwarded to SRAM banks, we also studied the behavior of MLVW technique on hybrid caches with different access pattern, with respect to the write-forwarding mechanism. We assumed the portion of writes forwarded to SRAMs or STT-RAMs defines different access patterns. For example, the 20-80 access pattern indicates 20% of writes are forwarded to SRAMs and the remaining 80% are forwarded to STT-RAMs. Note that these could be achieved by deploying architecture technique.

Fig. 7 compares power gain of MLVW technique on non-uniform hybrid caches with different access pattern. Fig. 7 indicates a growth on total power as we increase the portion of writes to STT-RAM. Compared to the baseline, we can

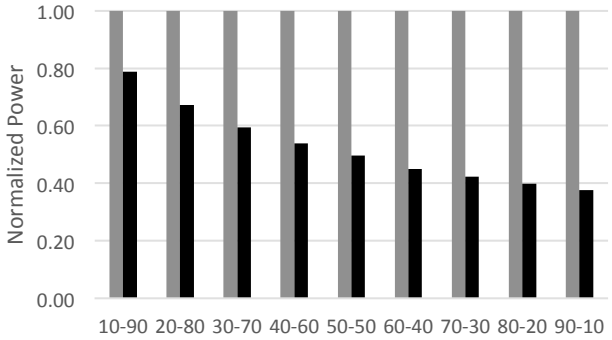


Fig. 7. Comparison of total power gain using MLVW technique for different access patterns

observe that for various access patterns VITERAT can achieve significant savings. The power savings gain increases as more writes are forwarded to STT-RAM. This imply that VITERAT can work cooperatively with other write-forwarding architectural techniques such as [9] to improve power efficiency even further.

V. Related Works

Many approaches have been proposed to optimize power and energy in hybrid caches and STT-RAM. Li et al. [9] proposed to integrate SRAM with STT-RAM to construct a novel hybrid cache architecture with a micro-architectural mechanisms to make the hybrid cache robust to workloads with different write patterns for CMPs. The proposed hybrid cache in [9] is a traditional NUCA cache utilizes the varied access latency of cache banks, due to their physical locations. However in this work, we improve the state of the art by proposed a NUCA cache architecture, which integrates SRAM and STT-RAM with different pairs of voltage/pulse width to improve power consumption further while operating at low voltage.

There are several works on improving cache reliability and reducing operating voltage, mainly for SRAM based cache [10, 11, 15, 21, 23, 25, 26, 27]. A number of these research proposed circuit level solutions [28], while there are others address fault challenge at architectural level [16]-[17]. There are also studies that proposed architectural technique to improve reliability of on-chip cache while lowering the voltage to minimum achievable operating voltage [18]-[19]. No previous work has studied the conflicting constraints of performance, power, and *reliability* in design of Hybrid on-chip caches. Sun et al. [20] proposed an iterative write-read-verify mechanism to tolerate high bit error rates exhibited by MTJ cells in extremely small scales. However, their research focuses on performance, while our work focuses mostly on power and we deliberately reduce the reliability of cell write operation to benefit from reduced write power. Moreover, we used a hybrid cache memory consist of SRAM and STT-RAM banks instead of MRAM-based cache memory.

VI. Conclusion

Hybrid Cache Architecture (HCA) with STT-RAM technology has shown to improve performance, latency, and power consumption of large caches by diving it into multiple banks with different power and performance characteristic. However, high power dissipation of write operation in STT-RAMs still remains as a major challenge in deploying them as a power-efficient solution in HC architecture. This paper introduces VITERAT a Variable Voltage and Pulse Width Iterative Writes Mechanism for low power STT-RAM in hybrid caches. VITERAT reduces the write

voltage of STT-RAM to gain significant power savings. While voltage scaling has a super-linear effect on reducing power, it exponentially increases the defect rate in STT-RAMs. By using a voltage pulse that is allowed to fail with some probability, VITERAT retry the failed writes and still save significant overall energy. This is shown with multi-core and multi-thread simulation results. With minor design modifications, we can achieve a 63% reduction in cache write power, and as much as 26% reduction in overall cache power across various workloads.

REFERENCES

[1] J. Huh, et al, "A NUCA Substrate for Flexible CMP Cache Sharing," in Proceedings of ICS, pp. 31–40, November 2005.

[2] X. Wu, et al, "Hybrid Cache Architecture with Disparate Memory Technologies," in Proceedings of ISCA, pp. 34-45, June 2009.

[3] X. Dong, and et al., "Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement," in DATE, pp. 554-559, March 2008.

[4] M. Hosomi, et al, "A Novel Nonvolatile Memory with Spin Torque Transfer Magnetization Switching: iSpin-RAM," in IEDM Technical Digest, 2005.

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

[6] W. Wu, X. Zhu, S. Kang, K. Yuen, and R. Gilmore, "Probabilistically Programmed STT-RAM," in IEEE J. on Emerging and Selected Topics in Circuits and Systems, vol.2 pp. 42-51, March 2012.

[7] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "A Durable and Energy Efficient Main Memory Using Phase Change Memory Technology," in ISCA, pp. 14-23, June 2009.

[8] G. Sun, X. Dong, Y. Xie, J. Li, Y. Chen, "A Novel Architecture of the 3D Stacked MRAM L2 Cache for CMPs," in HPCA, pp. 239-249, 2009.

[9] J. Li, C.J. Xue, ad Y. Xu, "STT-RAM Based Energy-Efficiency Hybrid Cache for SMPs," in VLSI-SoC 2011.

[10] A. Makhzan, H. Homayoun, A. Eltawil, F. J. Kurdahi. "Process Variation Aware Cache for Aggressive Voltage-Frequency Scaling". Design, Automation & Test in Europe, DATE 2009, Nice, France.

[11] Nikolaos Strikos, Vasileios Kontorinis, Xiangyu Dong, Houman Homayoun and Dean Tullsen. Low-Current Probabilistic Writes for Power-Efficient MRAM Caches, 31st IEEE International Conference on Computer Design, 2013.

[12] X. Dong, C. Xu, Y. Xie, N.P. Kouppi, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," in IEEE Tran. on Computer-Aided Design of Integerated Circuits and Systems, vol. 31, pp. 994-1007, July 2012.

[13] K. Flautner, N.S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy Caches: Simple Techniques for Reducing Leakage Power," in ISCA, pp.148-157, 2009.

[14] A. B. Kahng, B. Li, L. Peh, K. Samadi, "ORION 2.0: A Fast and Accurate NoC Power and Area Model for Early-Stage Design Space Exploration," in DATE, pp.423-428, March 2009.

[15] Avesta Makhzan, Houman Homayoun, Ahmed Eltawil, Fadi J. Kurdahi, "Inquisitive Defect Cache: A Means of Combating Manufacturing Induced Process Variation". IEEE Transactions on Very Large Scale Integration (VLSI) Systems, (TVLSI), 2010 (TVLSI), VOL. 19, NO. 9, SEPTEMBER 2011.

[16] C. Wilkerson, and et al., "Reducing Cache Power with Low-Cost, Multi-Bit Error Correcting Codes," in ISCA, pp. 83-93, June 2010.

[17] J. Kim, N. Hardavellas, K. Mai, B. Falsafi, and J. Hoe, "Multi-Bit Error Tolerant Caches Using Two-Dimensional Error Coding," in MICRO, pp. 197-209, December 2007.

[18] S. Ozdemir, D. Sinha, G. Memik, J. Adams, and H. Zhou, "Yield-Aware Cache Architectures," in Proceedings of MICRO, pp. 15-25, December 2006.

[19] C. Wilkerson, and et al., "Trading Off Cache Capacity for Reliability to Enable Low Voltage Operation," in ISCA, pp. 203-214, June 2008.

[20] H. Sun, C. Liu, N. Zheng, T. Min, and T. Zhang, "Design Techniques to Improve the Device Write Margin for MRAM-Based Cache Memory," in GLSVLSI, pp. 97-102, May 2011.

[21] Abbas Banayian, Houman Homayoun, Vasikeios Kontorinis, Dean Tullsen, Nikil Dutt. "REMEDIATE: A Scalable Fault-tolerant Architecture for Low-Power NUCA Cache in Tiled CMPs." International Green Computing Conference. IGCC 2013.

[22] BookSim: A Cycle-Accurate Interconnection Network Simulator, Version 2.0, https://nocs.stanford.edu/cgi-bin/trac.cgi/wiki/resources/booksim

[23] A. Makhzan, H. Homayoun, A. Eltawil, F. J. Kurdahi, "A Fault Tolerant Cache Architecture for Sub 500mv Operation Resizable Data Composer Cache (RDC-Cache)". In Proceedings of the 2009 International Conference on Compilers, Architecture, and Synthesis for Embedded Systems, CASES 2009. Grenoble, France.

[24] D.M. Tullsen, "Simulation and Modeling of a Simultaneous Multithreading Processor," in Annual Computer Measurment Group Conference, pp. 392-403, June 1995.

[25] A Chakraborty, H Homayoun, A Khajeh, N Dutt, A Eltawil, F Kurdahi, "E< MC2: less energy through multi-copy cache", Proceedings of the 2010 international conference on Compilers, architectures (CASES).

[26] Abbas Banaiyan, Houman Homayoun and Nikil Dutt. "FFT-Cache: A Flexible Fault-Tolerant Cache Architecture for Ultra Low Voltage Operation" In Proceedings of the 2011 International Conference on Compilers, Architecture, and Synthesis for Embedded Systems, CASES 2011. Taipei, Taiwan.

[27] Avesta Makhzan, Kiarash Amiri, Houman Homayoun, Ahmed Eltawil, Fadi J. Kurdahi, "Variation Trained Drowsy Cache (VTD-Cache): A History Trained Variation Aware Drowsy Cache for Fine Grain Voltage Scaling" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2010 (TVLSI). VOL. 20, Issue 4, pp: 630-642. April 2012).

[28] B. H. Calhoun and A. Chandrakasan, "A 256kb Sub-Threshold SRAM in 65nm CMOS," in ISSCC, pp. 2592-2601, 2006.