# NVP: Non-uniform Voltage and Pulse width Settings for Power Efficient Hybrid STT-RAM

Reyhaneh Jabbarvand Behrouz[+], Houman Homayoun[*]

[+] Department of Computer Science, George Mason University, USA {rjabbarv@gmu.edu}

[*] Department of Electrical and Computer Engineering, George Mason University, USA {hhomayou@gmu.edu}

*Abstract*— **As technology scales down, the leakage power of SRAM based cache becomes a more critical source of power dissipation, particularly for large last level cache where leakage power is dominant. The emerging non-volatile spin transfer torque RAM (STT-RAM) is a candidate to substitute SRAM due to its low leakage power and high thermal stability. However, considerable high energy and long latency of write operations in STT-RAMs are barriers to their commercial adoption. To address this problem, we propose a hybrid non-uniform cache architecture (NUCA) by combining 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 a technique that reduces STT-RAM write access energy by lowering voltage while ensures correctness by either retrying the failed writes or increasing effective pulse width. Simulation results indicate overall 20~30% power gain for various workloads in hybrid cache architecture. This comes with less than 2% performance loss.**

***Keywords—STT-RAM; NUCA; Hybrid Cache Architecture; Low Power***

## I. INTRODUCTION

In modern processor design, the last level cache (LLC) is typically a shared cache, configured 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 cores accesses. As a result, the LLC is typically one of the largest power consumers on the processor. Moreover, leveraging the size of LLC increases the latency which makes the power dissipation the major concern of LLC design. The latency of large caches can be improved through non-uniform cache architecture (NUCA) design [1]-[2]. In NUCA, the large cache is divided into multiple banks with different access latencies which results in significant reduction in average cache access latency.

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. Exploiting these new memory technologies offers significantly different power and performance characteristics compared to standard SRAM based caches. This is particularly true for large last level cache where leakage power is relatively high – replacing SRAM with non-volatile memory is attractive as it virtually eliminates the leakage power consumed by the 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 and power consumption of large caches by diving them into multiple banks with different power and performance characteristics [3]-[10].

This paper studies the use of spin torque transfer RAM (STT-RAM), one of the emerging non-volatile technologies, in hybrid cache architecture. Our studied architecture integrates non-volatile, high density STT-RAMs along with energy efficient SRAMs to build a hybrid on-chip LLC. Although STT-RAMs 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 yet a big concern. Among these barriers, write latency is less of a concern as it can be hidden by using large cache write buffers.

In this paper we introduce NVP, a mechanism to non-uniformly assign voltage and pulse width settings to different banks of hybrid SRAM/STT-RAM cache for low power purpose. NVP reduces the write voltage of STT-RAM to achieve significant power savings. While voltage scaling has a super-linear effect on reducing power, it exponentially increases the failure rate in STT-RAMs. We studied two mechanisms to overcome this reliability issue. The first strategy is to provide longer pulse width and increase the effective time of applying voltage pulse across MTJ cell. In the second strategy, the pulse width remains untouched and by using a voltage pulse that is allowed to fail with some probability, NVP retries the failed writes and still saves significant overall power and energy. Our multi-thread simulation results show that we can achieve 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:

- Examines various uniform and non-uniform assignments of voltage/pulse width pair to different STT-RAM banks.
- Highlights the dependence of write error rate of STT-RAMs to write pulse width and write voltage. It also demonstrates how this affects the power dissipation of large last level cache.
- Proposes the concept of multiple low energy writes that enables low power and reliable LLC with non-uniform cache architecture at near threshold voltage.

978-1-4799-6177-1/14/$31.00 

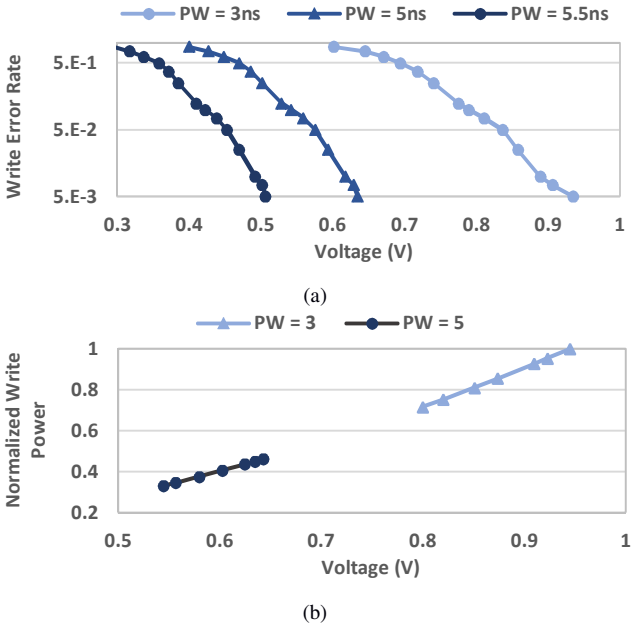


Fig. 1. a) Write error rate as a function of voltage and pulse width for write operations b) Power vs. voltage with 3ns and 5ns pulse width for single write access to cache

- Studies 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 gives an overview of the entire problem and the proposed solution. In section III, we discuss the proposed methodology by providing simulation results. Section IV reviews related works on STT-RAM low-power management, and finally Section V concludes the paper.

## II. 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 STT-RAM cell is a probabilistic function of the pulse voltage due to the thermal field torque act on the free layer magnetization. The write speed of a 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.

A number of studies show that STT-RAM cell write latency can be traded for energy and power consumption [4]-[5]. We take the 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. 1.a, as we slow down the write operation, by reducing the pulse width for instance, we can achieve 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 levels of 0.89v, 0.62 v, and 0.5v for 3ns, 5ns, and 5.5ns pulse widths, respectively. Fig. 1.b shows write power consumption of STT-RAM as a function of voltage in our studied architecture. Results from the Fig. 1.b suggests that by reducing the voltage the power decreases almost linearly for a given pulse width.

In this paper we propose NVP, a mechanism to optimize power and energy in NUCA hybrid caches by assigning different voltage/pulse width to various STT-RAM cache banks. We augment NVP with a write scheme to optimize the energy and the power in NUCA hybrid caches by lowering the write voltage.

We adapt two mechanisms to compensate the failure rate in STT-RAMs caused by voltage scaling. The first mechanism we use is to increase the effective time of applying voltage in order to address reliability issue. The second mechanism employs multiple low energy writes, with lower write voltage relative to the nominal voltage, instead of a single high energy write that guarantees near zero write error rate. For every block write on STT-RAM cells, our proposed method (a) performs a low energy write operation; (b) reads back the contents of the written cells; (c) compares the written and the original values to determine whether the write succeeded; (d) If any cell failed to store the written value, the process starts over to write remaining faulty bits, at lower write energy compared to an ordinary write.

There are several mechanisms proposed in the literature to reduce overhead of redundant writes [6]-[7]. We assume a write mechanism with bit granularity that can identify correctly written bits and writes only non-identical bits on subsequent writes. The rationale of writing non-identical bits is the high probability of redundant writes into a cache subsystem [8]. We employed early write termination strategy to identify non-identical bits in each write [9].

The early write termination strategy relies on stochastic switching feature of MTJ cell which allows STT-RAM to hold its valid value at the early stage of a write operation. In addition, write operation keeps the voltage long enough to change MTJ state. Taking into account that the write operation is much longer than a read in STT-RAM, early write termination assumes that a read operation can be performed during a write. The main objective of NVP is to minimize power dissipation without losing 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 the proposed write technique which attempts multiple writes to ensure accuracy. At architecture level, one solution is to use a store buffer that pools pending writes until they complete [10]. In this paper, we increase the size of the write buffers without imposing further overhead to the system. Our results indicate that this simple strategy ensures minimal impact on performance.

### B. *NVP*

In this section, we discuss various policies for implementing NVP, demonstrating that non-uniform settings of voltage/pulse width pair in different banks of NUCA significantly affect power and performance of the system. In fact, the power/performance trade-off depends on the mapping policy and reliability threshold that application can tolerate. The proposed assignment heuristics in this paper assume zero tolerance for reliability, which means no failure is allowed. However, in presence of fault tolerant mechanisms, more heuristics can be introduced for various voltage/pulse width assignments.

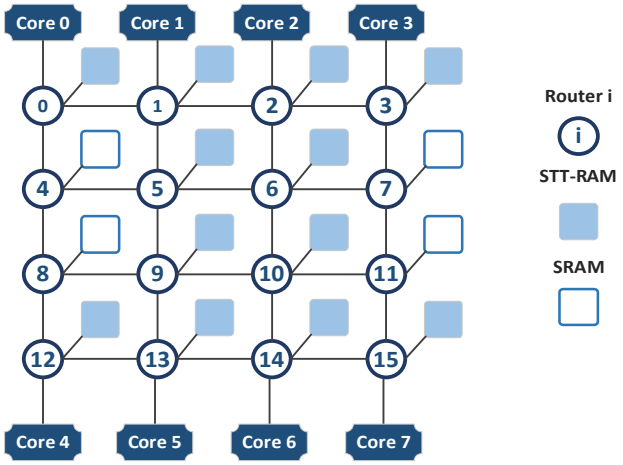


Fig. 2. Baseline hybrid cache architecture

The pair of voltage/pulse width assignment can be based on application characteristic. For example, in applications where the performance is highly sensitive to LLC latency (delay sensitive applications), assigning a high pulse width or performing iterative writes could severely affects performance. Therefore, we decrease voltage only to the extent that meets the desired reliability. On the other hand, for delay non-sensitive applications, larger pulse width and higher access latency caused by iterative writes can be tolerated which enables higher voltage scaling. This setting consequently results in lower power consumption without noticeable impact on the performance. Hence, the NVP improvement on power varies based on applications LLC characteristics.

In addition to application specific voltage/pulse width assignment, there is a more general consideration for NUCA cache designs. Banks in different locations are accessed with different latencies in NUCA. Assume a 4×4 NUCA shown in Fig. 2 which comprised of 16 cache banks and 8 cores that four cores are located on the top and the remaining are located on the bottom. The average number of hops to reach to the edge banks from each core is 4. However, the average hop to reach a bank in the middle is 3.5 which means middle banks are accessible with lower average number of hops than the edge banks. Based on this property, one heuristic is to assign settings with higher pulse width and therefore higher latency to the middle banks rather than the edge banks. Therefore, we expect more power gain by scaling down the voltage of these banks and reducing the energy. We should also note that the corresponding routers connected to the middle banks are more congested and there is a threshold on increasing the pulse width and the number of subsequent write iterations. Based on the application characteristic and NoC topology, we introduce the following heuristics for voltage/pulse width assignment:

- Application characteristic based assignment:
  1. Low voltage/High pulse width for frequently accessed STT-RAMs: Clearly, assigning lower voltage for the STT-RAMs that are accessed more frequently reduces total power consumption of the cache. Conversely, this requires higher pulse width to satisfy reliability, which in turn increases the access latency of the cache. This heuristic can be used for delay non-sensitive applications to save power while maintaining performance.
  2. Low voltage/High pulse width for infrequently accessed STT-RAM: This policy can be used for delay sensitive applications where increasing the pulse width of the infrequently accessed banks does not reduce performance significantly. Therefore, assigning low voltage to these banks can reduce power consumption while maintaining performance.
- Topology based assignment:
  3. Low voltage/High pulse width for middle STT-RAMs: Since the average hop count of middle banks to cores is less than of the edge banks, we can assign higher pulse width to the middle banks and therefore reduce their voltage further to gain more power savings
  4. Low voltage/High pulse width for edge STT-RAMs: Middle routers in the 2D mesh topology are more congested comparing to edge routers. Therefore, one can assign higher pulse width to edge banks instead of middle banks in order to control congestion of the middle buffers.

Since we are studying static NUCA (SNUCA) where the access to banks are uniform, we only consider the topology based heuristics in this paper. In addition, we can use the iterative writes mechanism instead of increasing the pulse width on all aforementioned heuristics.

In the rest of this paper, we study the power and performance of the hybrid SRAM/STT-RAM cache with different voltage/pulse width assignment for STT-RAMs, according to the topology based heuristics discussed above.

## III. Evaluation

In this section, the proposed methodology is evaluated using SPEC2000 and SPEC2006 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. 2. 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 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 64 KB L1, 512 KB L2 and 6.5 MB hybrid L3 consists of 12 STT-RAM and

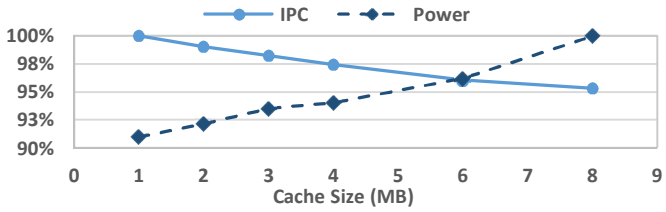


Fig. 3. Normalized LLC power and IPC sensitivity to different combination of SRAMs and STT-RAMs for 8-core hybrid architecture

4 SRAM banks with capacity of 512KB and 64 KB respectively. According to the results in [11], 64KB SRAM and 512KB STT-RAM have almost similar area. Hence, we can replace 64KB SRAMs with high bit density 512KB STT-RAMs. We chose to replace 4 out of 16 STT-RAMs with the same area SRAMs as SRAM provides lower read/write energy and latency. We found this configuration gives us best power-performance trade-off as shown in Fig. 3. We studied the power and performance of different configurations, from 8MB all STT-RAMs cache to 1MB all SRAM cache shown in Fig. 3. STT-RAMs benefit from near zero leakage power while they have high write power and latency. Moreover, we observed that most of the studied applications do not benefit significantly from a larger cache capacity, as results of replacing more SRAM banks with STTRAMs Therefore, it can be observed that the IPC drops by replacing more SRAMs with STTRAMs, as the write latency of total cache increases. In addition, the more STT-RAM banks in hybrid cache decreases the total static power while increases total dynamic power. We assumed drowsy cache technique applied to SRAM banks, which reduces their leakage power by up to 80% [15]. Hence, as we increase the number of STTRAM banks we notice an increase in the total cache power, as the dynamic power becomes dominant. There is a trade-off point where IPC and power cross each other, which we choose to construct our baseline architectural model. For an 8-core architecture, we consider 6.5MB LLC which comprised of 12 STT-RAM and 4 SRAM banks.

For cycle-accurate CMP + NUCA cache simulation we integrated SMTSIM [12], a multi-threading simulator, and BOOKSIM [13], a cycle accurate interconnection network simulator, in order to evaluate NVP mechanism. We simulate an 8-core CMP machine similar to the architecture shown in Fig. 2. 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 I shows the design parameters and configuration applied to our hybrid cache calculated by NVSIM [14]. The MTJ cell specification obtained from scaling parameters derived from the MTJ manufactured by Samsung [19].

The NoC exploits X-Y routing for delivering read and write requests from cores to cache banks. We have augmented Booksim by Orion2.0 [16] in order to investigate the power consumption of LLC. The power results are based on a 64-bit NoC implemented in 45nm technology and the frequency of 1 GHz. To model LLC cache power, we assume constant leakage power and read/write energy per access for SRAMs. For STT-RAMs on the other hand, different energy per STT- RAM write access, depending on write current and pulse width is considered.

TABLE I. ARCHITECTURAL MODEL DESIGN PARAMETERS FOR 45 NM

| | | |
|---|---|---|
| Cache Size | L1 (I/D) Private: 64 KB<br>L2 Private: 512 KB<br>8-core: 6.25 MB (12 STT-RAM, 4 SRAM) | |
| Associativity | 16 | |
| Bank Size | SRAM: 64KB | STT-RAM: 512 KB |
| Read Latency | SRAM: 3 cycles | STT-RAM: 3 cycles |
| Write Latency | SRAM: 3 cycles | STT-RAM: 15 cycles |
| Read Energy | SRAM: 0.026 nJ | STT-RAM: 0.077 nJ |
| Write Energy | SRAM: 0.026 nJ | STT-RAM: 0.264 nJ |
| Leakage Power per Bank | SRAM: 0.039 mW | STT-RAM: 0.003 mW |
| Interconnect | 8-core: 4 × 4 mesh, 2 banks for each core<br>Link Latency: 1 cycle, Router Latency: 1 cycle<br>Link Capacity: 64 bits<br>Buffer Size: 8 flits | |

We assume the leakage power to be constant for STT-RAMs. The total read/write energy is then derived by summing the read/write energy per access for total number of reads/writes. 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.

The results presented in this paper are based on simulations for several combinations of 8-thread workloads. As the significant barrier in deploying STT-RAMs is the 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 SPEC2000 and SPEC2006 benchmark as single-thread workload simulation on SMTSIM. We then categorized them first by the number of access to LLC, as LLC Access Intensive (I) and LLC Access Non-Intensive (NI) benchmarks, and then by the number of read/write access, as Read Intensive (RI) and Write Intensive (WI). For each application in an 8-thread workload, we fast-forward 500M instructions to skip the initialization phase and then simulate until each threads execute 500 million instructions.

### *B. Results and Analysis*

Fig. 1.a shows 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. The proposed iterative write scheme acts as a fault tolerance technique in presence of write failure.

Most of the power consumption of LLC-NoC comes from the STT-RAM and SRAM banks. Fig. 4 demonstrates the relationship between power consumption of the studied hybrid SNUCA cache and the write voltage for different pulse width. All markers on the figure indicate same WER since we applied iterative write technique to tolerate the failures. Note that, as reported in Fig. 1.a, the slope of write error rate vs. voltage is not steep for the studied memory architecture with small write pulse width (pulse width is lower than 10ns). Therefore, we only need to repeat writes twice to reach to above 97% write success rate. We report the following observations from Fig. 4:

- We gain more power by lowering the voltage for a given pulse width without trading accuracy. This gain diminishes by increasing pulse width while decreasing voltage.

- The rise in power for a given pulse width is due to multiple write operation. For error rates less than 75% and for a given pulse width, the increase is negligible as the number of erroneous bits to be rewritten is small and the iterative writes adds a small power overhead. For higher error rates on the other hand, there will be a spike on power by applying iterative write technique. However, we can trade power with error rate for higher WER and gain same power or even lower, utilizing this write scheme.

Fig. 5 compares total power and latency of the baseline configuration and a configuration with iterative write scheme for different workloads. It can be observed that for NI workloads, power gain is negligible comparing to other workloads and the latency overhead exceeds power gain. The reason is that for this class of workloads, due to infrequent memory access the leakage power dominates dynamic power and therefore it eliminates the benefit of iterative low energy writes. According to Fig. 5.a, we can gain 25%, 9%, 28%, and 24% power reduction in I, NI, WI, and RI workloads, respectively. Fig. 5.b shows that the latency of the network increased by 29%, on average when applying iterative low energy writes. We also study the impact of iterative writes technique on performance. Our simulation results indicate that for most of the studied applications, the performance is not significantly sensitive to last level cache latency. For I, NI, WI, and RI workloads, we observed our proposed write scheme improves power with affecting performance, measured by IPC, by 3%, 0.4%, 1.9%, and 2.7% respectively.

So far, we assumed all STT-RAM banks in hybrid cache have the same write voltage and pulse width (uniform setting). To reduce the overhead of iterative writes on latency, we also study non-uniform setting where different STT-RAM banks can have different voltage level and pulse width value. To study the impact of NVP on hybrid SNUCA cache power and performance, we developed the following scenarios:

**Case 1- Baseline:** We assume uniform assignments of pulse width and voltage values for all STT-RAM banks. The baseline pulse width is 3ns and voltage is 1v. The baseline architecture is similar to Fig. 2.

**Case 2 (Heuristic 3):** In this scenario, we study the non-uniform assignment of pulse width and voltage to different STT-RAM banks in SNUCA. We assign 3ns pulse width to edge STT-RAM banks and 5ns pulse width to middle STT-RAMs. We also use different scaling factor for edge and middle STT-RAMs voltage. The voltage for banks with pulse width of 5ns is derived from Fig. 4 so that the total cache power can be minimized.

**Case 3 (Heuristic 3):** We follow similar approach to case 2, however with different values for pulse width and voltage levels. In this scenario, we choose higher pulse width values of 5.5ns and 6ns for edge and middle STT-RAMs respectively. The voltages are scaled accordingly, same as case 2.

**Case 4 (Heuristic 4):** In this scenario, STT-RAMs located in the middle and corners have lower pulse width of 5.5ns while the remaining have 6ns pulse width. We scale the voltage accordingly, same as case 2.

Fig. 6 compares the (a) power and Energy Delay Product, EDP, and (b) latency of above scenarios for access intensive (I)

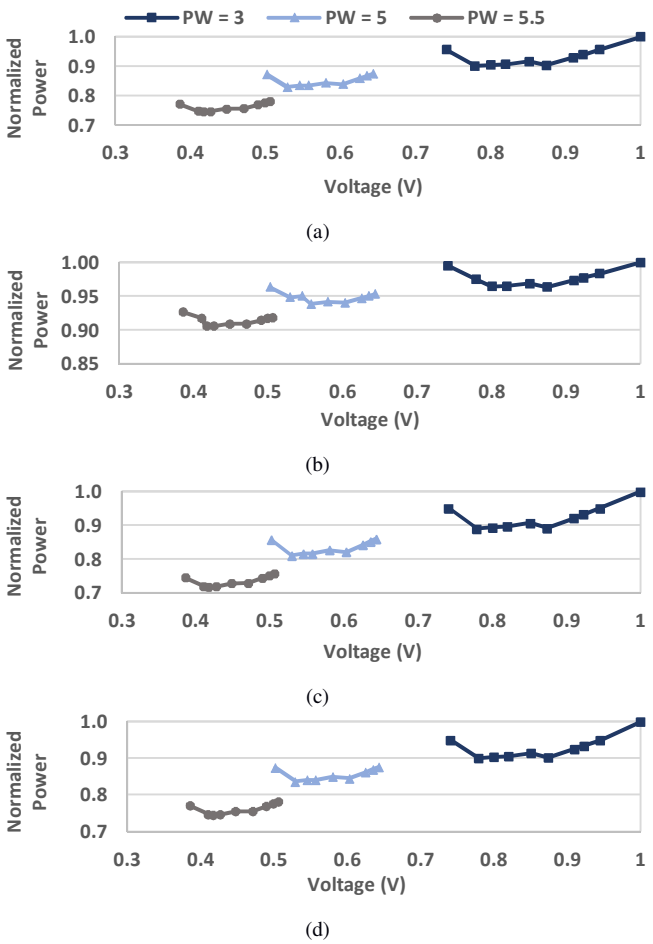


Fig. 4. Comparison of power consumption vs. voltage for different pulse width a) I workload b) NI workload c) WI workload d) RI workload

applications. The result illustrates that exploiting non-uniform settings improves total power by 26%, 10%, 29%, and 24% for I, NI, WI, and RI workloads. It also alleviates LLC access latency impact. The results show 15% increase in LLC access latency as oppose to 29% increase for uniform setting. This translates to less than 1% reduction in performance, in terms of IPC across all workloads.

## IV. Related Work

Many approaches have been proposed to optimize power and energy in hybrid caches and STT-RAM in particular. Li et al. [11] 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 [11] 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 hybrid cache architecture, which integrates SRAM and STT-RAM with non-uniform voltage/pulse width settings to improve power consumption further while operating at low voltage.

There are several work on improving cache reliability and reducing operating voltage, mainly concentrated on SRAM based caches. There are also studies that proposed architectural technique to improve reliability of on-chip cache while lowering

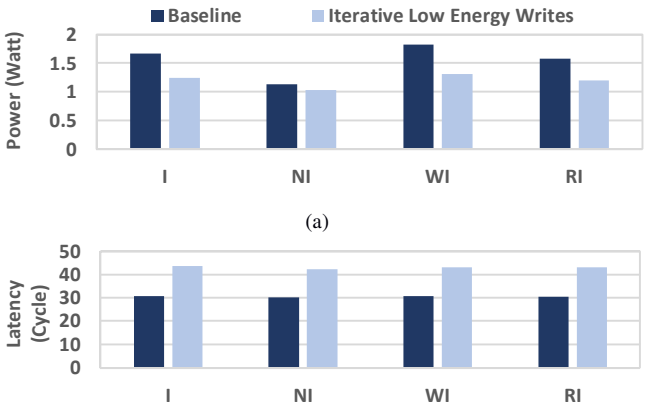


Fig. 5. Comparison of a) power and b) latency between baseline with uniform assignment of (1v,3ns) voltage/pulse-width and a configuration with iterative write scheme and uniform assignment of (0.4v,5.5ns) voltage/pulse-width

the voltage to minimum achievable operating voltage [17]. No previous work has studied the conflicting constraints of performance, power, and reliability in design of Hybrid on-chip caches. Sun et al. [18] 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. In addition, our work is different as we trade STT-RAM cell reliability with 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 a homogeneous STTRAM-based cache memory.

## V. CONCLUSION

Hybrid Cache Architecture (HCA) with STT-RAM technology has shown to improve performance and power consumption of large cache by diving it into multiple banks with different technologies and characteristics. The STT-RAM is a candidate due to its fast read access, low leakage power, high bit density, long endurance, and high thermal stability. However, high power dissipation of write operations in STT-RAMs still remains as a major challenge in deploying this new technology as a power-efficient solution in HCAs. This paper introduces NVP, a mechanism to assign Non-uniform Voltage and pulse width setting to different banks of hybrid SRAM/STT-RAM caches for low power purpose. NVP 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 failure rate in STT-RAMs. We studied two mechanisms to overcome this reliability issue. The first strategy is to provide longer pulse width and increase the effective time of applying voltage pulse across MTJ cell. In the second strategy, the pulse width remains untouched and by using a voltage pulse that is allowed to fail with some probability, NVP retries the failed writes and still saves significant overall power and energy. It is shown with multi-core and multi-thread simulation results that with minor design modifications, we can achieve 63% reduction in cache write power, and as much as 26% reduction in overall cache power across various workloads.

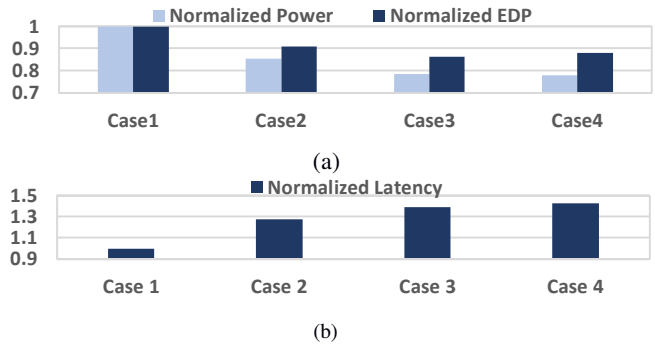


Fig. 6. Comparison of a) power, EDP and b) latency of uniform and non-uniform assignment of pulse width-current pairs

## REFERENCES


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

[2] A. BanaiyanMofrad, and et al., "REMEDIATE: A Scalable Fault-tolerant Architecture for Low-power NUCA Cache," in IGCC, pp. 1-10, 2013.

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

[4] N. Strikos, V. Kontorinis, X. Dong, H. Homayoun, and D. Tullsen, "Low-current probabilistic writes for power-efficient STT-RAM caches," *in ICCD*, pp. 511-514, October 2013.

[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] P. Zhou, B. Zhao, J. Yang, Y. Zhang, "Energy reduction for STT-RAM using early write termination," *in ICCAD*, pp. 264-268, 2009.

[9] T. Zheng, J Park, M. Orshansky, M. Erez, "Variable-Energy Write STT-RAM Architecture With Bit Wise Write Completion Monitoring," *in ISLPED*, pp. 229-234, September 2013.

[10] 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.

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

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

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] W. Kim and et al., "Extended Scalability of Perpendicular STTMRAM Towards Sub-20nm MTJ Node," *in IEDM*, 2011.