Droop Mitigating Last Level Cache Architecture for STTRAM

Radha Krishna Aluru[†] and Swaroop Ghosh[‡]
[†] Department of Computer Science and Engineering,
University of South Florida, Tampa, FL, USA.
[‡] School of Electrical Engineering and Computer Science,
Pennsylvania State University, University Park, PA-16801, USA.
[†] aluru@mail.usf.edu, [‡] szg212@psu.edu

Abstract*—Spin-Transfer Torque Random Access Memory (STTRAM) is one of the emerging Non-Volatile Memory (NVM) technologies especially preferred for the Last Level Cache (LLC). The amount of current needed to switch the magnetization is high (~100μA per bit). For a full cache line (512-bit) write, this extremely high current results in a voltage droop in the conventional cache architecture. Due to this droop, the write operation fails especially, when the farthest bank of the cache is accessed. In this paper, we propose a new cache architecture to mitigate this problem of droop and make the write operation successful. Instead of continuously writing the entire cache line (512-bit) in a single bank, the proposed architecture writes 64-bits in multiple physically separated locations across the cache. The simulation results obtained (both circuit and micro-architectural) comparing our proposed architecture against the conventional are found to be 1.96% (IPC) and 5.21% (energy).

Keywords*—Droop, LLC, STTRAM, Bank, Latency, Energy, SPLASH benchmarks.

I. INTRODUCTION

Spin-Transfer Torque RAM (STTRAM) [1] is a promising memory technology due to high-speed, low-power, non-volatility, and low cost. It is an energy-efficient modification of MRAM [2], where switching of the magnetization is obtained by current induced spin-transfer torque. A density closer to DRAM, speed closer to SRAM, high endurance and superb retention time, makes STTRAM is widely considered to be a suitable candidate for universal memory [4-5].

Fig. 1 shows the STTRAM cell schematic, where the Magnetic Tunnel Junction (MTJ) with a free layer and a pinned magnetic layer is the storage element. The resistance of the MTJ is high (low) if free layer magnetic orientation is anti-parallel (parallel) compared to the fixed layer. Spin-transfer torque is used to flip the active elements in MRAM [3]. The MTJ configuration can be changed from parallel to anti-parallel (or vice versa) by injecting current from sourceline to bitline (or vice versa).

Spin-transfer torque lowers the write current requirements in STTRAM. However, the write current is still too high for most of the commercial applications [3]. In a conventional cache architecture, the high write current (~100μA per bit) may result in a write failure in the farthest bank due to the high voltage droop caused by the large interconnect resistance. The need to write large number of bits simultaneously i.e. 512 cache line bits in a single bank is another reason for this failure.

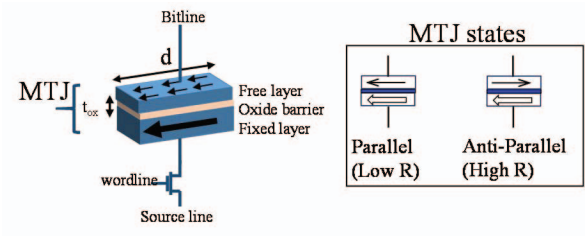


Fig. 1 (a) STTRAM bit cell showing an MTJ with different layers and the bitline, wordline; (b) two MTJ states Parallel (low resistance – logic 0) and Anti-Parallel (high resistance – logic 1)

To overcome the above challenge, we propose a new LLC architecture. Instead of writing the entire 512-bits in a single bank which draws significant current (~512x100uA) creating a large voltage droop for the last sequence of bits, we split the entire cache line into 8 parts (64*8) and write them in multiple physically separated locations across the cache. The approach reduces the current drawn per bank from $512xI_{write}$ to $64xI_{write}$. In the proposed approach, the worst case bits only experience a maximum of 10% droop at which the write operations can succeed without any errors. Note that, we still write in a single bank logically, however, the bank is spread across the cache physically to mitigate the droop.

The voltage droop for crossbar memories such as Resistive RAM (ReRAM) has been pointed out [7][11] however, similar issue for STTRAM has never been investigated. *In this paper, for the first time, to the best of our knowledge, we identify the voltage droop challenge for write operation in STTRAM and propose a novel micro-architectural solution.*

In summary, we make following contributions in this paper:

- Voltage droop analysis of the write operation in a conventional STTRAM LLC with a circuit.
- Propose an architecture that significantly mitigates droop.
- Impact of proposed architecture on the cache parameters like latency and energy are compared with the conventional architecture for various benchmarks.

The rest of the paper is organized as follows. In Section II, we perform detailed voltage droop analysis on conventional LLC. The proposed droop mitigating architecture and simulation

978-3-9815370-8-6/17/$31.00 ©2017 IEEE

results are discussed in Section III. Conclusions are drawn in Section IV.

II. Voltage Droop Analysis in STTRAM LLC

In this section, we describe the model and bank architecture of an existing STTRAM LLC. Next, we perform the droop analysis using the LLC model.

A. Model of STTRAM LLC

The entire cache is divided into multiple banks/slices. The entire cache line read/write is performed within this single bank [9]. Each bank is divided into a group of mats. A mat contains multiple ways (way 0-n). A group of mats together provide the output cache-line (e.g., 8 mats provide 64 bits each totaling 512 bits) [10]. Each mat consists of a group of subarrays which share a common pre-decoder to provide the requested data or perform write operation.

For this study, we have considered an 8-MB LLC with 8-way set associativity. Each subarray refers to the associativity's (ways) from 0-7 to select from. Each of these ways consists of the rows and columns (global columns are muxed with local columns) to store the individual bits. The subarray size is 16kb. Each mat is composed of 8 subarrays (SA[7:0]) amounting to a total size of 128kb each. Each bank is composed of 8 mats (mat[7:0]) of total size 1MB. There are 8 such independent banks in the cache. The cache organization is shown in detail in Fig. 2(a). Fig. 2(b) shows a proposed subarray design consisting a total of 512 WLs and 512 local columns with 64 (32*2) global columns.

B. Circuit Simulation with Droop

We have used circuit model of an 8MB cache and simulated the effective model of a cache line (512-bit) write. We write 64-bit in each of the 8-subarrays (belonging to way7) in each of the 8 mats within a single independent bank. The values of the effective resistance and capacitance of the power supply which provides the current needed to write the 512 bits (64 bits across 8 mats) is shown in Table I. Fig. 3(a) shows the circuit model of the write operation with the calculated effective values. Fig. 3(b) shows a circuit model of the cache

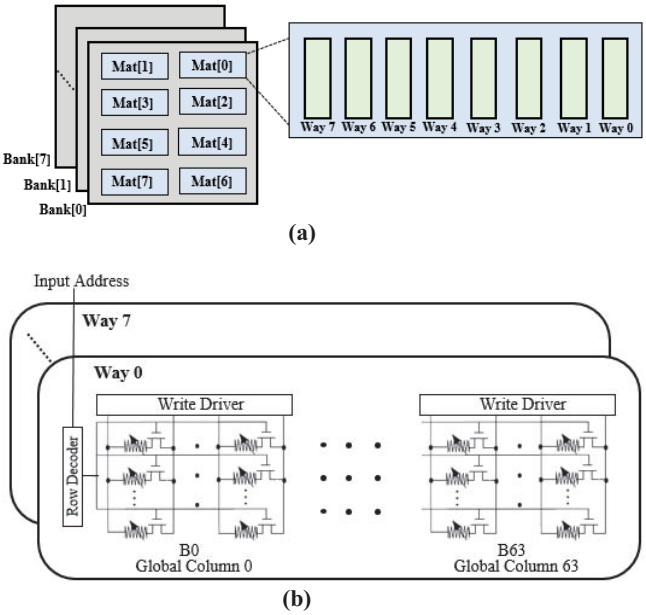


Fig. 2 (a) An 8MB LLC organization with banks, mats and ways/ subarrays; (b) subarray/way architecture of STTRAM showing the STTRAMs, the organization of bits, the global columns and the write drivers [10].

TABLE –I. Equivalent calculated values for the LLC cache Model

# of bits	Resistance	Capacitance	Write current(Load)
1	0.4 Ω	9 pF	100 µA
64	25.6 Ω	0.57nF	6.4 mA

with 8 banks to perform the write operation in the 8 mats of the last and the farthest bank. The idle banks are represented by shaded blocks while the unshaded bank represents a cache line written into it. The supply voltage is assumed to be 1V. Fig. 4 shows the corresponding voltages at each of these mats see from the plot that the voltage keeps on drooping down and the last 64-bit of the cache line at mat8 receives ~0.79V to perform the write operation. At a such low voltage, the STTRAM fails to switch states resulting in write failures. This is due to high write current of the STTRAM and inability of

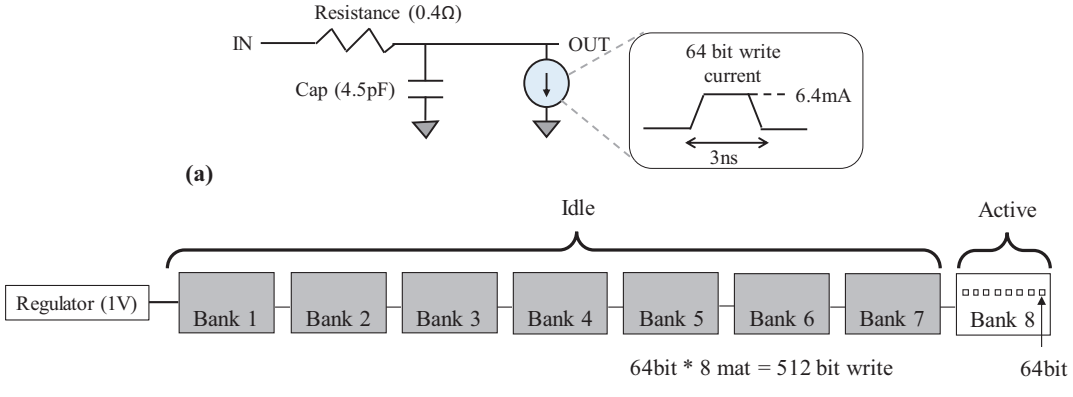


Fig. 3 (a) 1-bit write operation with effective resistance 0.4 Ohms in parallel to a 4.5pf capacitance and a pulse current load of 0.1*64mA; (b) cache circuit model with write in the last/farthest bank.

the conventional LLC bank architecture to provide reliable supply current.

We have used Hspice [12] to simulate the STTRAM flipping phenomenon for the write-1 and write-0 at various supply voltages. The STTRAM model consists of a MTJ verilogA with adjustable parameters. In a 22nm technology, we performed the write simulation (bit-0 and bit-1) at varying supply voltages. Fig. 5(a) and 5(b) show the plots of write latency with supply voltage. It is observed that the STTRAM write latency increases as the supply voltage is reduced from 1V. For write-1 the latency increases by ~150ps at 0.9V and follows an exponential trend. For write-0, less than ~0.9V will result in failure and at 0.9V latency increases by ~160ps.

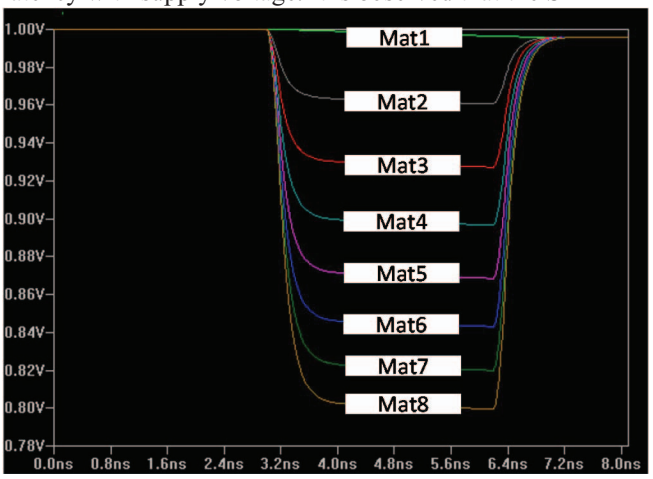
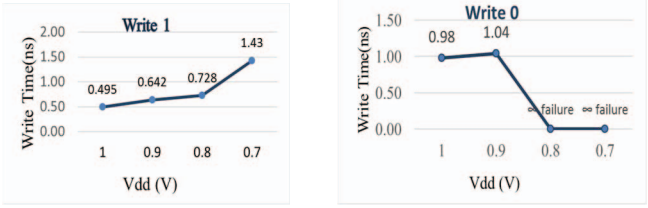


Fig. 4 Voltage plot showing the available voltage for each of the 64-bit write operations in each mat.



(a) **(b)**

Fig. 5(a) & (b) plots showing the results obtained from Hspice simulations of the STTRAM write time for wrote-1 and write-0 against different supply voltages.

III. Droop Mitigating Bank Architecture

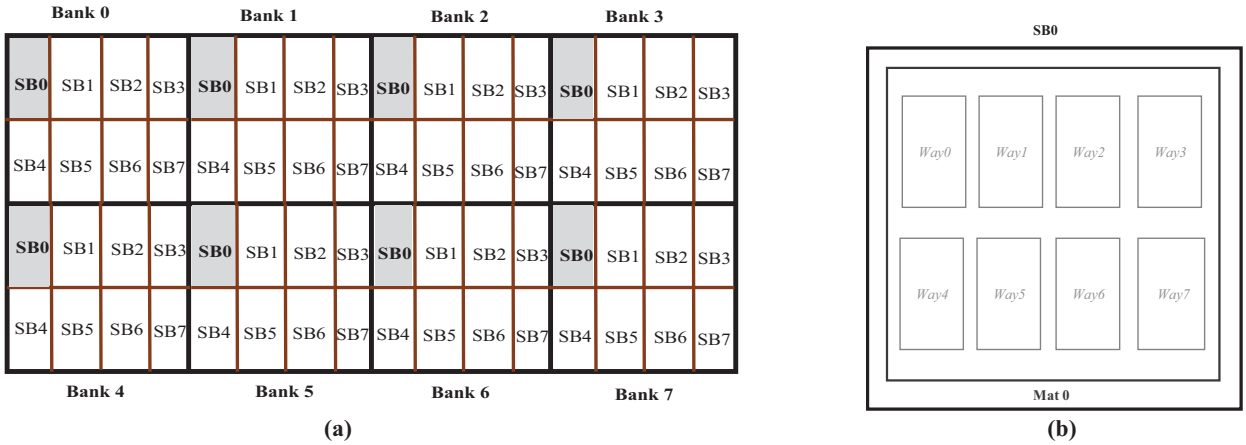
In this section, we present the proposed LLC architecture and circuit analysis to overcome the write voltage droop problem.

A. Architecture Model

We propose a new bank architecture which distributes the current drawn during write operations at different physical locations. Therefore, the effective droop at a location is reduced. Fig. 6(a) shows the physical bank architecture of the droop mitigating LLC. Each of physical banks divided into 8 sub-banks (SB0–SB7) are represented in shaded. Shaded sub banks located at various physical locations throughout cache represent a complete logical bank i.e. though separated physically, each of these sub-banks are still logically continuous in terms of physical addresses. For example, consider the sub-bank SB0 is distributed in 8 locations, forms a continuous logical bank, bank0. Fig. 6(b) shows the inside look of a sub-bank which contains a single mat unlike 8 mats before. Internal structure of this mat with the subarrays/ways remains the same. Likewise, each of these 8 sub-banks now contributing 64 bit each form the 512bit cache line. Each of the 64bit write occurs at different physically separated locations mitigating the droop and allowing write operation to succeed.

B. Circuit Analysis of the new Model

We have created a circuit model of the proposed 8MB LLC (Fig. 7). The shaded portions in each of the banks represent an equivalent mat/sub-bank with the 64-bit being written in them. The remaining unshaded portions in each of the banks are the idle sub-banks. The supply voltage is kept at 1V. Fig.



(a) **(b)**

Fig. 6 (a) Droop mitigating physical bank architecture of 8MB LLC; (b) inside look of the bank0: sub-bank(SB0)

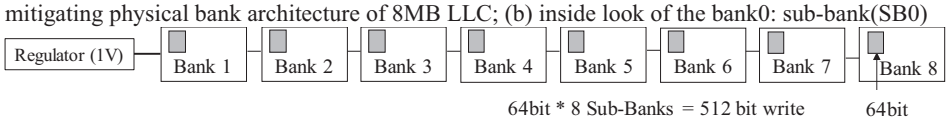


Fig. 7 The cache circuit model with write operation in the droop mitigating architecture.

8 shows the corresponding voltages at each of sub-banks for the write operation. It can be observed that the voltage droop is greatly reduced and the last 64-bit of the cache line in bank8 now receives close to 0.9V to perform the write operation. Even though the write latency increases, the failure could be avoided.

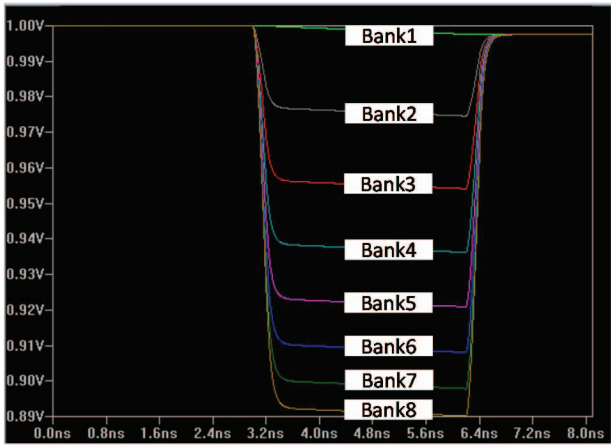


Fig. 8 Voltage plot showing the available voltage for each of the 64-bit write operations in each of the banks (i.e. sub-banks)

C. *Micro architecture Evaluation and Results*

We have used Cacti [13] [14] at a footprint of size $4F^2$ to simulate a model of an 8MB STTRAM LLC for the architectures. We have calculated the values of read/write latency and the dynamic read/write energy per access. Fig. 9(a) show the normalized latency comparison of these proposed architecture results of with the conventional architecture.

We used the Gem5 [15] [16] to plug-in the result values of latency and energy obtained from Cacti. Table II shows the processor configuration table. Using these values in modified Gem5, we ran Full system simulations against various benchmarks from SPLASH suite for both the conventional LLC and the proposed LLC. Mcpat [17] tool was used to plug-in these benchmark stats obtained and generate the values of dynamic and leakage energies of the LLC.

Fig. 9(b)- (d) show the comparison between droop mitigating architecture with conventional architecture with respect to IPC and energy. The proposed architecture is clearly results in minor (an average of 1.96%) overhead in terms of IPC and 5.21% energy. The reason for this is the nature of the non-volatile memories, where the bit write/read time takes the dominant part of the latency while impact of the hop latencies in a cache is very small. The benefit of proposed architecture is observed from ~50% improvement in worst case droop (~100mV droop compared to ~200mV droop in conventional architecture).

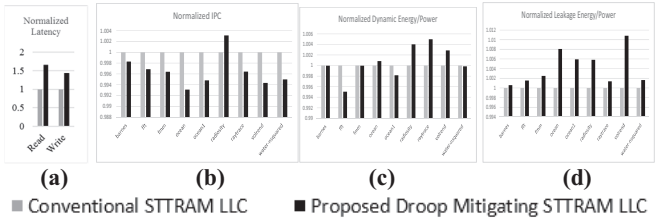


Fig. 9 (a) Normalized read/write latency comparison; (b) normalized comparison of IPC; (c) & (d) normalized comparison of dynamic and leakage energy for various benchmarks

TABLE –II. Processor Configuration

Processor	Alpha, O3, 4 cores, 2GHZ, Detailed CPU
SRAM L1 Cache	Private, Icache=32KB, Dcache=64KB, 64B Cacheline, 2 cycle Read/Write latency, Write back.
SRAM L2 Cache	Private, Size=2MB, 64B Cacheline, 8 cycle Read/Write latency, Write back.
STTRAM LLC/L3 Cache	Shared, Size=8MB, 8 banks, 8ways, 64B Cacheline, Write back, Read/Write latency based on the Architecture Model
Main Memory	4GB, DDR3, 200-cycle latency

IV. CONCLUSIONS

The high current causes voltage droop for full cache line write operation resulting in write failures especially for farthest bank/cache. In this paper, we have proposed a new droop mitigating bank architecture of LLC to mitigate the droop and enable successful write operation.

Acknowledgement: We acknowledge the support of NSF grants CNS- 1441757 and SRC grant 2442.001.

REFERENCES

[1] Driskill-Smith, Alexander. "Latest Advances and Future Prospects of STT-RAM." Non-Volatile Memories Workshop, 2010.

[2] Zhu, Jian-Gang. "Magnetoresistive Random Access Memory: The Path to Competitiveness and Scalability." *Proceedings of the IEEE,* 2008.

[3] "Spin-transferTorque." https://en.wikipedia.org/wiki/Spin-transfer_torque.

[4] Kultursay, Emre et al. "Evaluating STT-RAM as an Energy-efficient Main Memory Alternative." *ISPASS*, 2013.

[5] Chen, E. et al. "Advances and Future Prospects of Spin-Transfer Torque Random Access Memory." *IEEE Transactions on Magnetics,* 2010.

[6] Yarom, Y., Qian Ge, Fangfei Liu, Ruby B. Lee and Gernot Heiser. "Mapping the Intel Last-Level Cache." *NICTA, ICT Research*.

[7] Niu, Dimin et al. "Design Trade-offs for High Density Cross-point Resistive Memory." *ISLPED*, 2012.

[8] Tabrizi, Farhad. "Non-volatile STT-RAM: A True Universal Memory. " Flash Memory Summit, Santa Carla, CA, USA August 2009.

[9] Muralimanohar, Naveen and Rajeev Balasubramonian. "CACTI 6.0: A Tool to Understand Large Caches. " Tech. Rep, 2009.

[10] Motaman, Seyedhamidreza, and Swaroop Ghosh. "Adaptive Write and Shift Current Modulation for Process Variation Tolerance in Domain Wall Caches." *IEEE Trans. VLSI Syst.* 24, no. 3 (2016): 944-53.

[11] Xu, Cong, Dimin Niu, Naveen Muralimanohar, Rajeev Balasubramonian, Tao Zhang, Shimeng Yu, and Yuan Xie. "Overcoming the Challenges of Crossbar Resistive Memory Architectures." *HPCA*, 2015.

[12] https://www.synopsys.com/tools/Verification/AMSVerification/CircuitSimulation/HSPICE/Pages/default.aspx

[13] "Cacti. " http://www.cacti.net/download_cacti.php

[14] Muralimanohar, N. et al. "Optimizing NUCA Organizations and Wiring Alternatives for Large Caches with CACTI 6.0." *MICRO*, 2007.

[15] "Sims. " http://www.m5sim.org/Documentation

[16] "Gem5. " http://www.gem5.org/Main_Page

[17] "Mcpat. " http://www.hpl.hp.com/research/mcpat/