STAG: Spintronic-Tape Architecture for GPGPU Cache Hierarchies

Rangharajan Venkatesan

Purdue University

Paper Overview

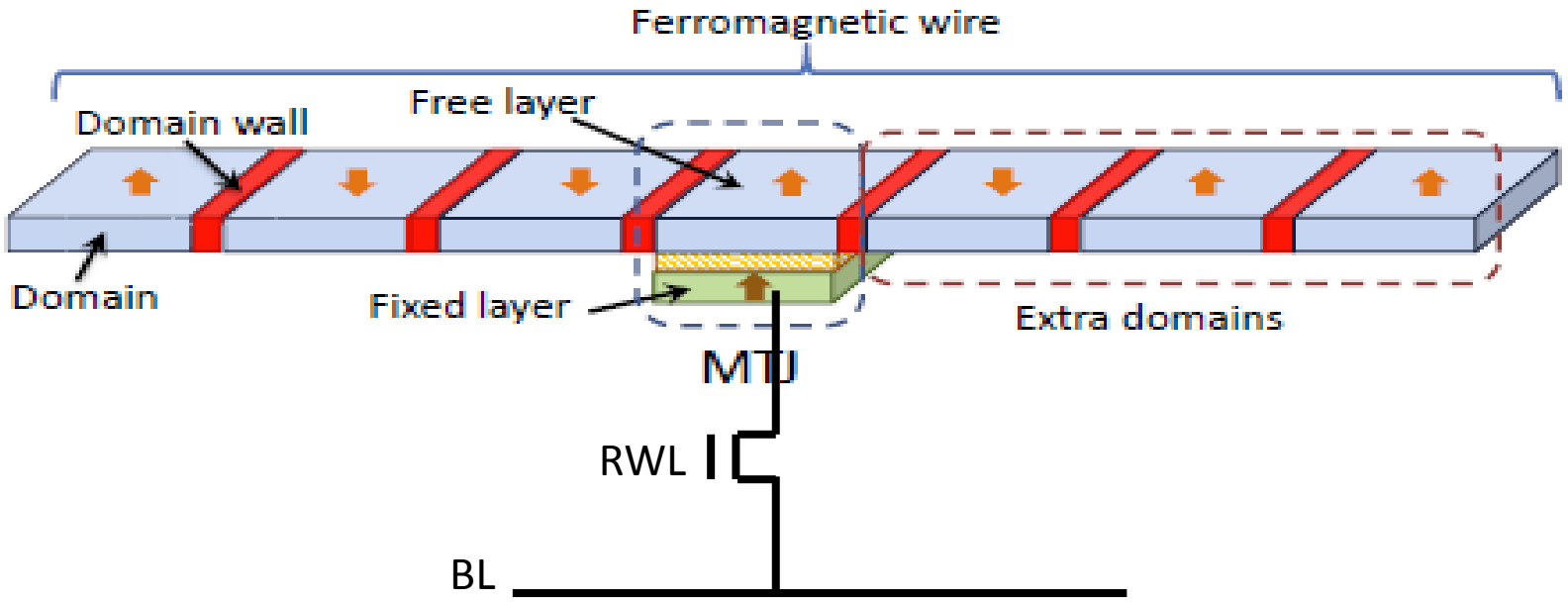
- Significance: current CMOS-based cache memories, SRAM and eDRAM, face the challenges of power consumption and scalability limit. Domain Wall Memory(DWM) has the benefit of high density, low power, and high access speed.
- Challenges: The structure of DWM logically resembles a bit-serial tape. Thus, the data in DWM has to be sequentially accessed by performing "shift" operations, resulting in high access latencies.
- Target: design a DWM-based cache hierarchies with optimized latency and density.
- Methods: L1 cache-> single-bit DWM, L2 cache's tag array-> single-bit DWM, L2 cache's data array->multiple-bits DWM.

Multiple-bits DWM-> clustered bit-interleaved organization, tape-head management policy, shift aware promotion buffer.

- Results: 12.1% and 5.8% performance benefit over SRAM and STT-MRAM; 3.3X and 2.6X energy saving over SRAM and STT-MRAM.

Background—Domain Wall Memory

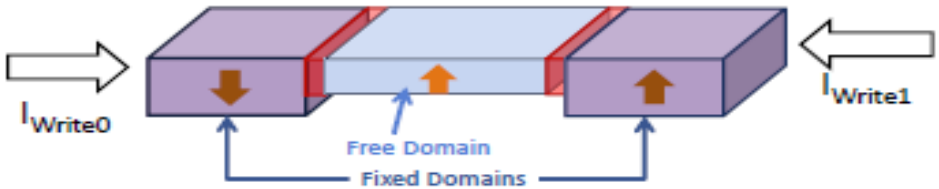
Domain Wall Memory is a spin-based memory technology, which represents information using the spin orientation of the magnetic domains in a ferromagnetic wire.



Read Operation: Like what STT-RAM does, the magnetic orientation of fixed layer and free layer determines the MTJ resistance. We can use the different resistance value as “1” and “0”. The latency is very similar to SRAM, and the current is also much smaller compared to SRAM.

Background—Domain Wall Memory

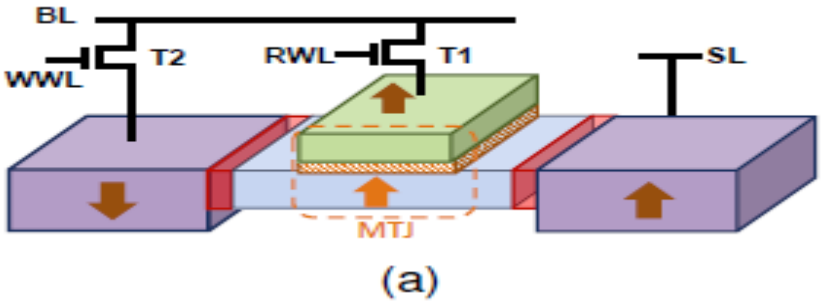
- Shift Operation: When a current pulse is applied through the ferromagnetic wire, the magnetic orientation of each domain is shifted in a direction opposite to the current. It is called as “domain wall motion”.
- Write Operation: STTRAM uses a large spin-torqued current to change the state of MTJ. But in domain wall memory, write operation based on shift costs less power and also even much faster. Two fixed domain with up-spin and down-spin are attached to both sides of free domain. A shift operation with Iwrite0 and Iwrite1 decides the new status of free domain.

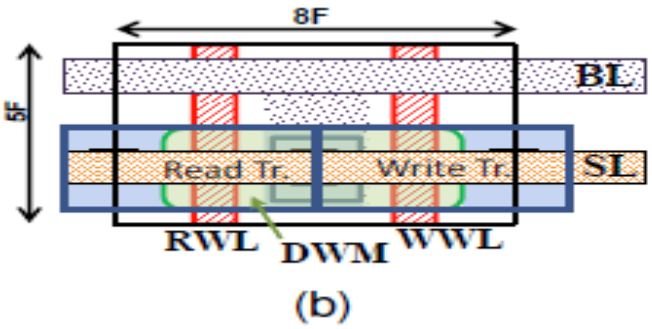


Parameter	MTJ-based write	Shift-based write
Switching current (uA)	191	30
Switching Time (ns)	2	0.25

1bitDWM and Multibit DWM

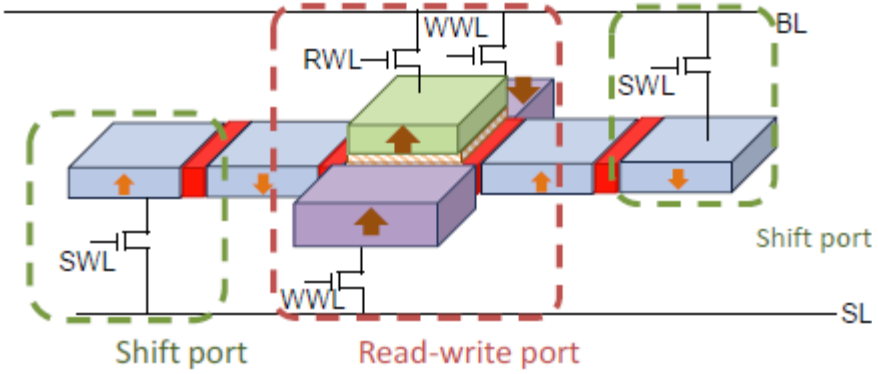
- The main component of 1bitDWM is 2 fixed domain, 1 free domain, 1 MTJ, and 2 small access transistors.

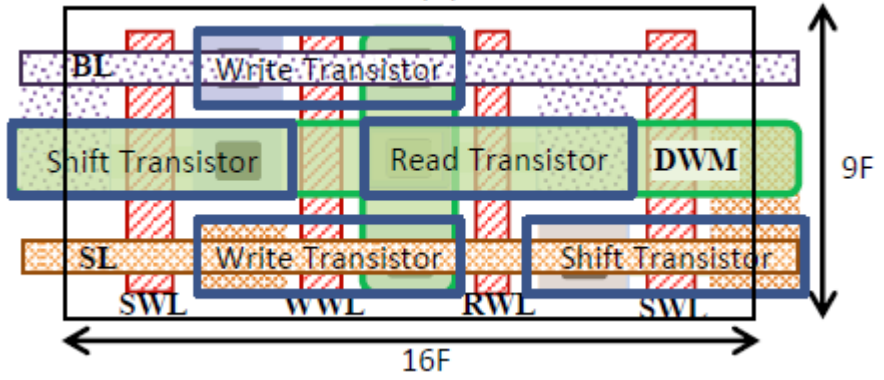




Each free domain has its own access transistors and fixed domain. It makes the access time optimized, because there is no need for shifting.

- The main components of multibitDWM includes 1 ferromagnetic wire of multiple free domains, 2 fixed domains, and 5 access transistors.





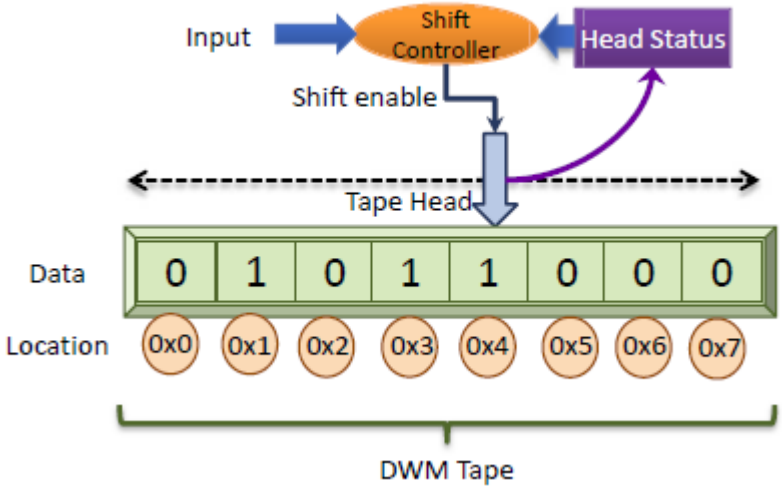
1bitDWM and Multibit DWM

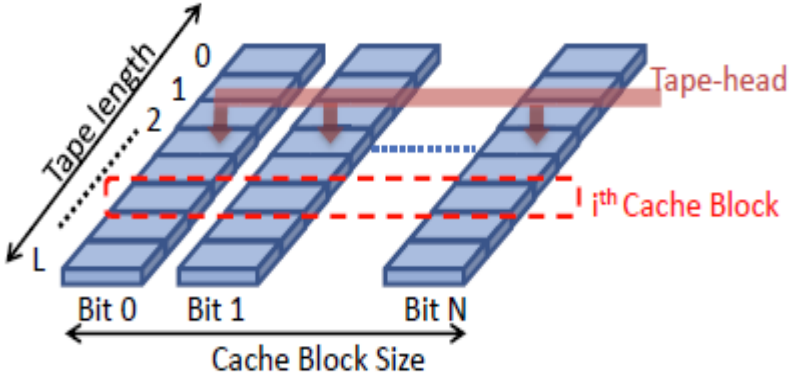
- The main difference of 1 bitDWM and Multibit DWM is its density and shift operation.
- Shift operation: MultibitDWM is impacted by shift operation. The main factor is how many free domains a ferromagnetic wire contains.
- Density: 1bitDWM has dedicated read/write port (access transistors, and fixed domains). It has similar cell size compared to STTRAM. But multibitDWM shares the same read/write port within all the free domain in a wire. So it has much higher density.

	1bitDWM	MultibitDWM	STTRAM	SRAM
Cell Size(F2)	40	4.5	40	146
Size Ratio	8.9	1	8.9	32.4

cache design overview

- L1 cache requires a short access latency and high capacity to get high performance benefit. 1bitDWM provides fast access time and also has 2X density benefit.
- L2 cache prefers a high-density design to reduce the burden of main memory. But large access latency will exhaust all active warps. 1bitDWM is utilized for tag array design for fast lookup time. MultibitDWM is utilized for data array design for high density. But the optimization of access latency is in need to reduce the extra overhead of shifting operation.
- Bit-interleaved tape cluster organization: group 1024(128) ferromagnetic wires as a cluster. A cache block restores in the same position of each wire. The benefit is, a cache block can be read at one time. The optimized case is, there is no shift operation required for data access.





L2 Cache Optimization

- Challenge: The multibitDWM experience a variable and long shift latency because of various distance of data location and tape head. If we can predict the next data location and shift tape head close to it, it has short shift latency.
- Tape Head Management Policies

1. Restored head policy: always shift the tape heads of a DWM tape cluster (DTC) to the middle of DWM tapes after each access to the DTC.
2. Leaved-in-place head policy: dynamically decide the place to put tape head based on the most recently access policy.
3. Preshifted head policy:

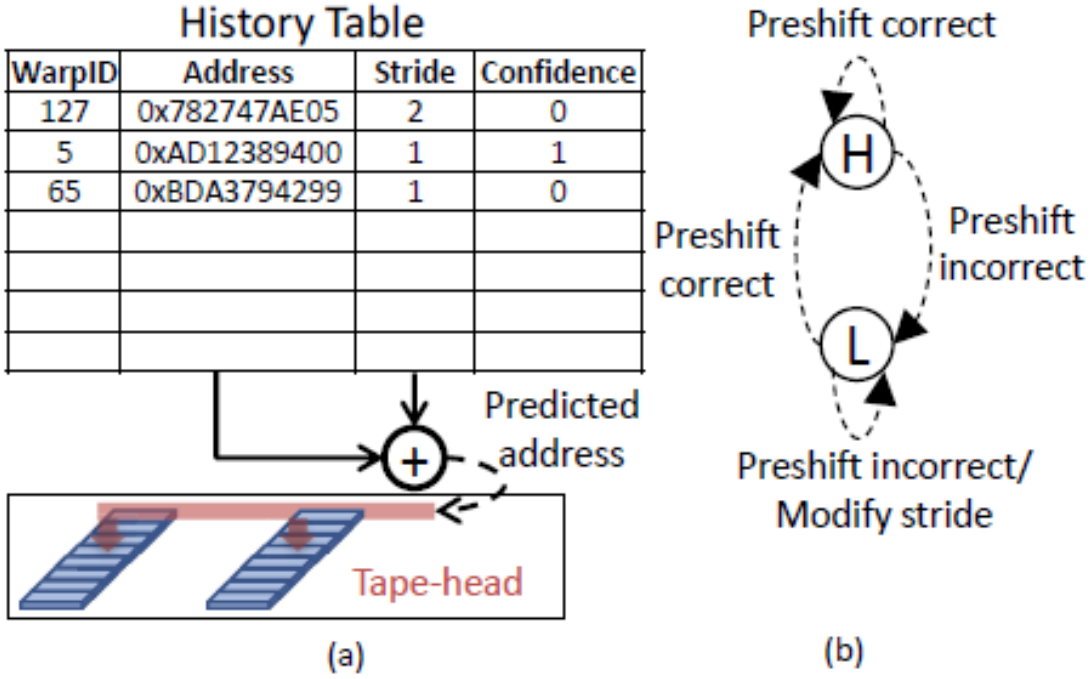
Observation: data access from single warp has high access locality, but the total multiple warps may pollute this data access locality.

Methods: Introduce a history table to predict next data location.

L2 Cache Optimization

3. Preshifted head policy:

Methods: Introduce a history table to predict next data location.



An entry records the status of one warp.

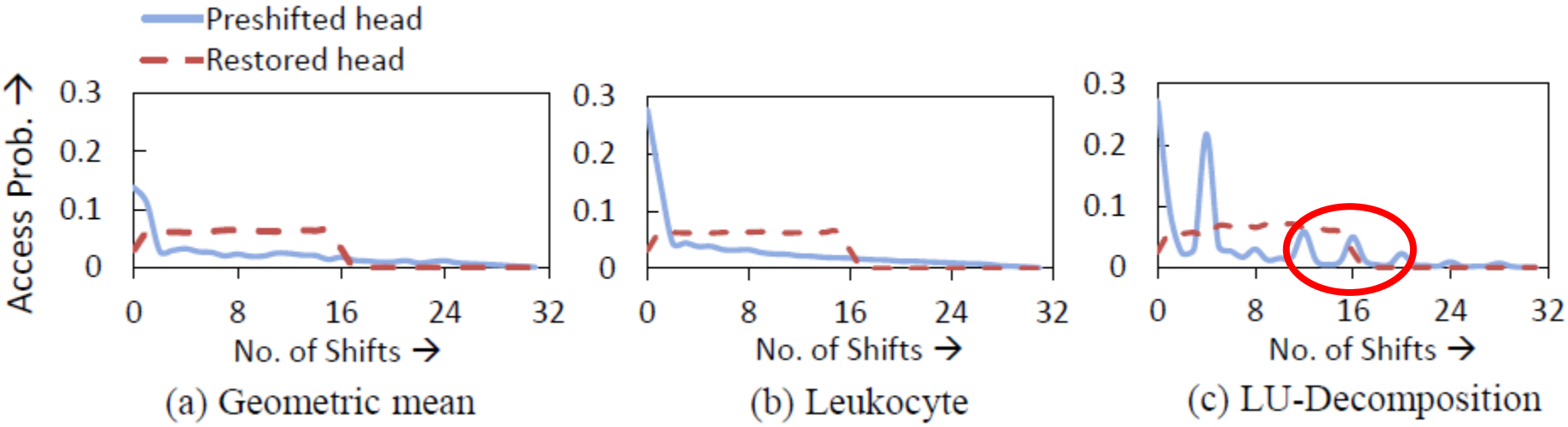
"Address" stores the last access address from the exact warp.

"Stride" estimates the distance of predicted address from last access address from the same warp.

"Confidence" is 1-bit flag to filter out rare random accesses.

L2 Cache Optimization

- Challenge: preshifted head policy can reduce the intra-warp data distance, but cannot reduce inter-warp data distance.



- Methods: introduce a “Shift Aware Promotion Buffer”.

It is a small, full associative buffer built with 1bitDWM, which is put in front of L2cache. The main role of SAPB is to contain cache block which may incur large number of shifts.

L2 Cache Optimization

- Methods: introduce a “Shift Aware Promotion Buffer”.

It is a small, full associative buffer built with 1bitDWM, which is put in front of L2cache. The main role of SAPB is to contain cache block which may incur large number of shifts.

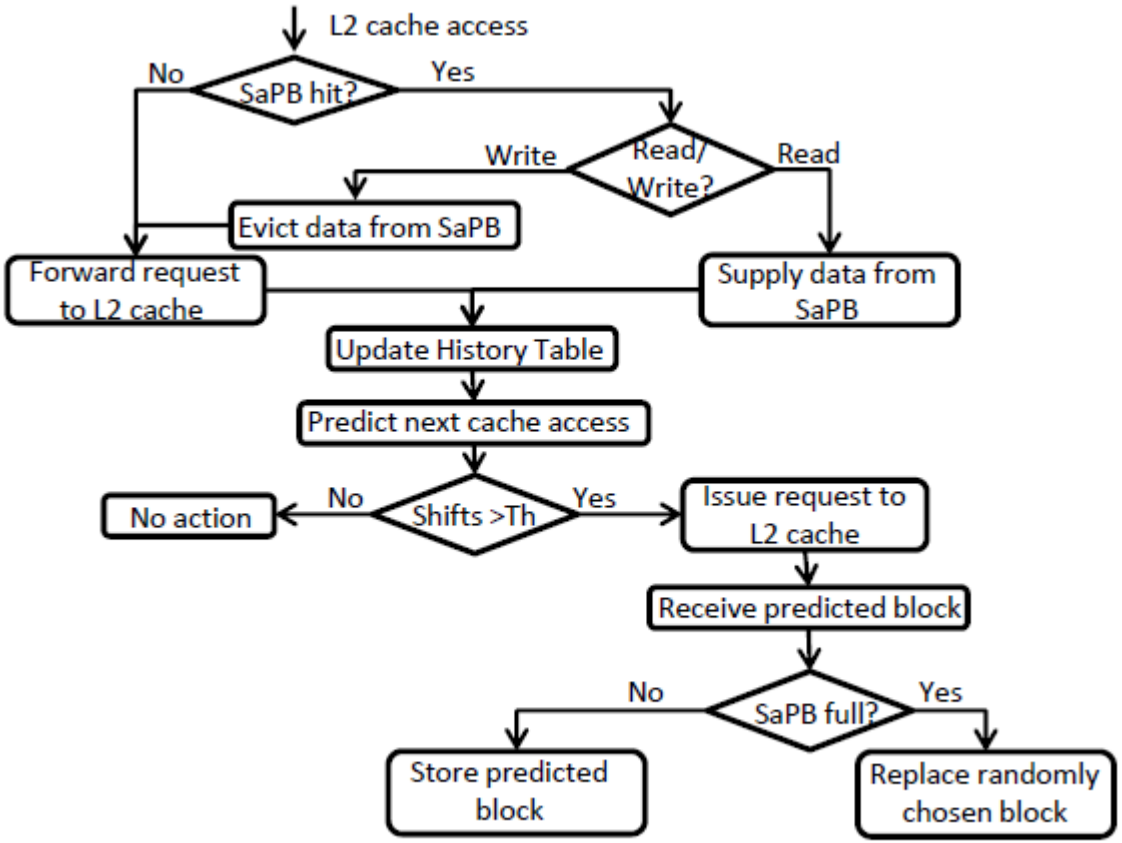


Figure 13: SaPB operation