8

“Overcoming the Challenges of Crossbar Resistive MemoryArchitectures”, HPCA, 2015.

Technical Summary：

In this paper, the author explored the optimization possibilities of ReRAM-based memory systems from circuit-level and architecture-level compared with DRAM and PCM. Sneak current and voltage drop leading to write performance degrade and latency were mainly discussed.

To solve these problems, double-sided ground biasing, multi-phase write operations, compression based encoding scheme and latency aware write scheduling were proposed. It finally evaluated the performance based on conservative and aggressive baseline and got a performance improvement.

Description of Contribution:

First, the author studied the crossbar architecture and described trade-offs involving voltage drop, write latency, and data pattern for both a conservative baseline design and an aggressive design. The author (he/she) then analyzed microarchitectural enhancements such as double-sided ground biasing and split the long latency RESET operations to sub-phases(hRESET) to further reduce the impact of voltage loss due to the sneak current to improve write performance.

At the architecture level, the author proposed a simple compression based encoding scheme with negligible storage overhead to speed up most of the write operations by limiting the worst-case voltage drop across the selected cells. As the compressibility of a block varies based on its content, write latency was not uniform across blocks. To mitigate the impact of performance led by writes through blocking a bank and delaying subsequent reads to that bank, the author proposed and evaluate a novel scheduling policy that considers the varying latency of ReRAM writes along with pending activity of a bank when flushing writes to the memory. The architecture proposed in this paper improved the performance of a system using ReRAM as main memory. It also compared with ideal DRAM-only system to show the advantage of ReRAM technology.

Major Critiques:

ReRAM is a promising cost-efficient replacement for DRAM as main memory because of its high density, low write energy, and high endurance. But it also has some challenges. This paper analyzed design constraints for an ReRAM based memory architecture because of crossbar architecture, and proposed circuit-level and architecture-level optimizations to enable the adoption of this emerging technology for future memory systems design. The paper described the crossbar-specific innovations that different from that of PCM designs as main memory. It pointed out the significance of sneak current in a crossbar architecture and its impact on latency. In addition, the compression-based encoding, its effect on write latency, and the proposed scheduling techniques are novel additions to the non-volatile main memory literature. Note that the

author is not using compression to improve capacity or bandwidth. Compression is only being leveraged to create space to store flip bits.

2.L. Zhang et.al “Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs”, ISCA, 2016.

Technical Summary:

In this paper, because the disadvantage of using one single write latency for different workloads and the abundance of memory idle time, L. Zhang et.al aimed to increase memory lifetime while minimizing performance impact by using slow writes strategically.

In order to use slow writes strategically, the author mainly proposed two Mellow Write schemes here: Bank-Aware Mellow Writes and Eager Mellow Writes. The second scheme made up for the shortcoming of the first scheme. There is one another scheme is Wear Quota, which is a safe scheme to ensure extending the lifetime.

Description of Contribution:

The author presented three microarchitectural mechanisms (Bank-Aware Mellow Writes, Eager Mellow Writes, and Wear Quota) that selectively perform slow writes to increase memory lifetime while minimizing performance impact. Bank-Aware Mellow Writes inspects the current set of write requests, sending slow writes to banks that have only one current write request. With taking advantage of bank idle time in the presence of no write requests, this paper introduced Eager Mellow Writes, which goes one step further, identifying useless dirty lines in the last level cache (LLC) to write back to banks with no requests. In order to protect against memory-intensive workloads that do not have enough idle times, the author proposed Wear Quota to guarantee the minimal lifetime (e.g., 8 years in our experiment) of the ReRAM memory system at the cost of performance. From the experiments, this paper shows the advantage of their design that compared with the lifetime benefit, it requires minimal hardware overhead and a moderate increase in main memory energy consumption.

Major Critiques:

The author pointed out that limited write endurance is a major drawback for such resistive memory technologies. Wear leveling (balancing the distribution of writes) and wear limiting (reducing the number of writes) have been proposed to mitigate this disadvantage, but both techniques only manage a fixed budget of writes to a memory system rather than increase the number available. For many previous work, they always found ways to reduce the number of writes, however, this paper proposed to reduce the impact of some writes on the endurance by performing a slower write. One advantage of the Bank-Aware Mellow Writes technique is that it requires minimal changes to the memory controller. The only modifications are a mechanism to detect bank conflict in the read and write queues and implementation of the slow write technique. However, this paper found that the scheme they proposed has its drawback, which is when there are

no writes for a bank in the write queue, it is unable to take advantage of the bank idle time.

3.W. Wen et.al “Speeding Up Crossbar Resistive Memory by Exploiting In-memory Data Patterns”, ICCAD, 2017.

Technical Summary:

In this paper, the author proposed to use a new low overhead runtime profiling technique to dynamically acquire the data patterns in different bitlines and reduce the number of LRS cells by employing data compression and row address dependent data layout. Instead of conservatively using the worst-case access latency of all cells in ReRAM arrays, this paper proposed to dynamically speed up ReRAM RESET operations for the rows that have small numbers of LRS cells. These are based on the observation of relation between RESET latency and # of cells in LRS.

It also did the circuit-level simulation and finally proved to be energy saved and improve the speed and performance degradation from IR drop, in particularly the RESET operation.

Description of Contribution:

It exploited the bitline data patterns like the percentage of LRS cells, to speed up RESET operations in crossbar architecture for ReRAM. Based on the founding of the correlation between the RESET latency of a ReRAM row and the number of cells in low resistance state (LRS) on selected bitlines, this paper proposed to dynamically speed up the ReRAM RESET operations for the rows that have small numbers of LRS cells. In order to achieve further performance improvement, this paper also employed data compression and row address biased data layout.

Due to dynamically track the number of LRS cells in each bitline, the author proposed a novel low overhead runtime profiling technique. But it may cause large overhead if reading all memory lines from the mat for detection, so this paper combined a method which has been widely exploited that used the current aggregation feature of crossbar to avoid this problem. And in this paper, it implemented and compared its design with the conventional and state-of-art ReRAM designs.

Major Critiques:

Previously Many of papers explored the possibility of using ReRAM as an alternative of DRAM as main memory. They analyzed the challenges from different orientations. We all know that, the crossbar cell structure suffers from large sneak leakage and IR drop on long wires. To ensure operation reliability, ReRAM writes, in particular, RESET operations, conservatively use the worst-case access latency of all cells in ReRAM arrays, which leads to significant performance degradation and dynamic energy waste. In this paper, they studied the correlation between the RESET latency and the number of cells in LRS along bitlines. They proposed a novel profiler to periodically track the number of LRS cells, and then speed up ReRAM RESET operations for the rows that

have small numbers of LRS cells to replace using worst case access latency of all cells. It brings the performance improvement compared to the baseline and the state-of-art crossbar design.

4.M.V.Beigi and G.Memik “THOR: THermal-aware Optimizations for extending ReRAM lifetime”, IPDPS, 2018.

Technical Summary :

In this paper, M. Valad Beigi aimed to solve the problem of ReRAM endurance limitations related to increase of temperature and decrease of lifetime, which is a deficiency for achieving a scalable and high-performance memory system architecture. It mostly implemented 2.5D design.

The author used THOR-LA and THOR-SA to reduce the frequency and the number of accesses to the hot ReRAM banks, then the endurance would be increased and then finally lower the total temperature and extend the lifetime. The author concluded that THOR can achieve lifetime enhancement and peak temperature reduction over a baseline design with little performance degradation.

Description of Contribution:

In this paper, the author explored the impact of temperature on the lifetime of ReRAM in 2.5D and 3D designs. They investigated how the endurance of ReRAM (in terms of number of writes) reduces as temperature rises. And they presented a novel solution named THOR (Thermal-aware Optimizations for extending ReRAM lifetime) to reduce the temperature and enhance the lifetime of ReRAM. Their system reduces the impact of temperature on the endurance by reducing the rate of accesses to hot banks which have lower endurance compared to cooler ones. Specifically, their technique reduced the frequency and number of accesses to hot banks by 1) delaying the requests to hot banks and 2) keeping the blocks from hot banks in the LLC longer. Hence, thier technique can reduce the temperature of hot banks and achieve cooler banks with higher endurance. This, in turn, can increase the ReRAM lifetime and reduce power consumption.

Their results show that THOR can achieve significant temperature reduction and lifetime enhancement over the baseline design.

Major Critiques:

3D integration of ReRAM crossbar layers (i.e., 3D crossbar) is a potential method for further improving ReRAM density. However, 3D architectures typically suffer from high operating temperatures, which adversely impact ReRAM reliability and device performance. The objective of this study is to address ReRAM endurance limitations, which is a major drawback for such resistive memory technologies. Although the relationship between temperature and ReRAM characteristics and the impact of

temperature on ReRAM cells have been discussed before, none of the previous studies have explored the impact of temperature on the endurance of ReRAM.

The disadvantage in this paper is that it doesn't have clear descriptions of contributions. Which made people hard to recognize what is novel in their paper. They also don't present the detailed descriptions of how they get the temperature for the cells or crossbar.

5.A. Shafiee et.al "MemZip: Exploring Unconventional Benefits from Memory Compression", ISCA, 2014.

Technical Summary:

In this paper, the author targeted the unconventional metrics like complexity, energy, bandwidth and reliability instead of capacity discussed more before. He/She proposed a new and simple memory compression architecture designed for bandwidth efficiency but not for page fault rate reduction.

In addition, the space saved by compression were planned to improve reliability and energy by using ECC and DBI codes with no extra cost. There was a limitation that this architecture was based on a memory system that uses rank subsetting.

Description of Contribution:

Instead of using memory compression to improve capacity and page fault. This paper proposed the secondary benefits from compression: lower energy access 你说 lower bandwidth demand. The author designed a new compression architecture that is explicitly for energy and bandwidth efficient operation. It is useful even when applications don't stress the memory capacity or don't exhibit spatial locality.

The MemZip achieve low complexity, reliability and energy efficiency by combing compression with rank sunsetting, and with now data layouts and integrates the lay out with embedded-ECC codes as well as DBI codes. They proposed a new data layout named Generalized Rank Susetting (GRS) which allows the fine-grained memory access while supporting relatively low data transfer times. It is especially used for MemZip. And with the organization (compression + embedded-ECC), we don't need extra COL-RDs to retrieve their ECC codes. And MemZip can alleviate the overhead of embedded-ECC

Major Critiques:

It provided a promising opening topic for the future. It is potential to improve performance with intelligent schedulers that can prioritize the shortest job or deprioritize ECC code fetches. Many of system components like cache, memory and disks are capacity-constraints. Many of papers have shown the effectiveness of data compression for cache, memory and disks. In this paper, it focuses on compression applied to main memory and the organization of compressed data within the memory system.

Compared with previous work, in this work, the author designed a new simple compression architecture that is designed explicitly for energy and bandwidth-efficient operation. They first show how energy and bandwidth-efficient can be saved with

compression in a DDR3 memory system. The rank subsetting is an energy saved method proposed before, however, the author change the rank subsetting to GRS by using a new data layout.