SIRM: Shift Insensitive Racetrack Main Memory

Hongbin Zhang, Bo Wei, Youyou Lu, Jiwu Shu

HAL Id: hal-03770574
https://inria.hal.science/hal-03770574

Submitted on 6 Sep 2022

SIRM: Shift Insensitive Racetrack Main Memory

Hongbin Zhang[1], Bo Wei[2], Youyou Lu[1], and Jiwu Shu[1]

[1] Tsinghua University, Beijing, China
{zhanghb,luyy09,shujw}@mail.tsinghua.edu.cn
[2] Hangzhou Dianzi University, Hangzhou, China
weibo@hdu.edu.cn

Abstract. Racetrack memory(RM) is a potential DRAM alternative due to its high density and low energy cost and comparative access latency with SRAM. On this occasion, we propose a shift insensitive racetrack main memory architecture SIRM. SIRM provides uniform access latency to upper system, which make it easy to be managed. Experiments demonstrate that RM can outperform DRAM for main memory design with higher density and energy efficiency.

Keywords: Racetrack memory · Shift insensitive · Main memory.

1 Introduction

Recently, racetrack memory (RM), which is also known as domain wall memory (DWM), has attracted significant attention of researchers. Previous research has demonstrated that this approach can achieve ultra-high density by integrating multiple domains in a tape-like nanowire [1–3]. In addition, it provides SRAM-comparable access latency and high write endurance [4]. In racetrack memory, each cell has a similar architecture and access pattern to an STT-RAM cell. Each racetrack contains one or more access ports, and the data aligned with each port can be read/write by these accessing ports. In order to access other bits that are not aligned with a port, a shift operation must be performed to move these bits to the nearest access port. Owing to the comparable access latency of a SRAM, RM is a promising candidate for on-chip memory or caching [5]. Furthermore, data placement mechanisms for optimizing its access latency and energy cost have been researched intensively [6, 7]. All of these studies have focused on how to significantly reduce the shift intensity, either at the system or compiler level, in order to leverage the density, shift latency, and energy cost.

There are also approaches to compose racetrack as the main memory. A shift-sense address mapping policy (SSAM) has been proposed for reducing shift operations in racetrack-based main memory systems [8]. SSAM significantly reduces shift intensity by employing a specified address mapping policy. However, SSAM also introduces complexity to the memory management and system design because each request may have a different number of shift steps or length and the read/write latency varies.

In this work, we propose an improved solution for RM serving as main memory, shift insensitive racetrack main memory (SIRM) which successfully hide shift operations, and provide a uniform read/write interface to upper system. With its inherent advantages, SIRM can provide higher density, superior performance, and lower energy main memory compared with DRAM.

2   Motivation

In order to achieve higher bandwidth, modern commodity DRAM generally work according to DDR standard[11] in most computer architecture. The most important performance of DDR is data rate, or data burst latency. As previous work point out[1], the shift latency of RM array is related to the tape length, number of access ports and overlap layout. We propose a specific RM array with appropriate design and keep the shift latency equal or smaller than the DRAM burst cycle, then the shift latency will be covered by the DRAM burst time interval. So SIRM has a good scalability to fit for different DDR standard, which will be discussed in Section4. We design the main memory architecture with multiple RM arrays. Together with SIAM which provides a pipeline mechanism to read data from adjacent racetrack, the shift operation will be hidden under the memory level and invisible to upper system. We design and implement SIRM according to this idea and testify its effectiveness.

3   The Shift Insensitive Racetrack Main Memory Design

3.1   Basic Array of RM

Prior work[1] proposes an organization with overlapped RM cells, called Macro Unit (MU), as a basic building block of RM array, as shown in Figure 1(a). In this section, we mainly discuss the shift latency caused by MU structure and their application in SIRM.
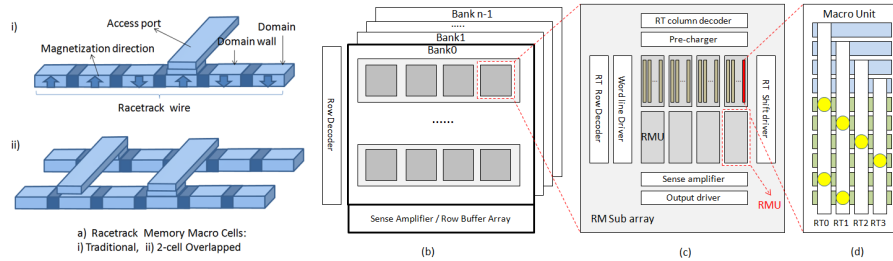

Fig. 1. The RM based main memory Architecture. a)RM Macro Cells; b)Overview of a bank; c)A RM subarray; d)A Macro Unit;

The basic array composed by MUs with different parameters has different shift latency. Theoretically, the longer the racetrack is, the longer shift latency

is. The more access ports the racetrack has, she shorter shift latency is. And the more racetracks MU has, the longer the shift latency is because it needs more energy to sense the data out. According to[1], the area optimized solution for RM data array is MU-64-32-4. According to[8], the MU-64-32-4 and MU-64-16-4 have similar performance. This paper use MU-64-32-4 as basic MU to simulate main memory. MU-64-32-4 has 4 racetracks, 32 access ports and each racetrack has 64 storage domains and 8 access ports. In order to compose an appropriate RM data array which has comparable shift latency with DRAM burst cycle, we simulate several RM data array with different capacity and test their performance with NVsim[10]. According to the result, we choose 8MB as the basic size of array because data shows that the 8MB array cost 1.25ns to read data and shift a step, which is equal to the burst cycle of DDR4-1600MHz modules, which has a I/O bus of 1600MHz. Theoretically, the RM array can provide the same read latency with DDR4-1600 and shift operations will be covered.

In this paper, according to Microns data sheet[9], we extend this model to simulate a 128Mb RM memory chip with 16 banks, and 256B row buffer in 45nm technology, which is the most advanced one we can get. The RM device level parameters in this paper are similar to the previous work[1].

3.2   Main Memory Architecture

A single rank of main memory contains multiple memory chips, which typically has 4,8 or 16 data output pins[12]. As shown in Figure1, a RM based chip can be organized as three levels: bank, mat, and subarray. Bank is the top level unit, mat is the building block of bank, and subarray is the elementary structure. Subarray is composed by 8 RMU that is described in Figure2. One RMU is composed by 16 MU and a MU has a structure of 64-32-4 as described above. Then, one RMU has 4K bits and an subarray has 32K bytes. One bank has 256 subarray and 8M bits. One chip has 16 banks and 128M bits. Multiple RMUs in one subarray share the same corresponding periphery circuitry in order to shrink the energy cost.

3.3   Shift Insensitive Address Mapping

We propose SIAM policy to cover the shift operation through pipeline operation. We take the part of subarray to simplify the discussion. As is shown in Figure2, an array has eight RMUs in lateral and each RMU has 64*8bytes in vertical. Data are numbered in cacheline which has 64bytes. For example, cacheline1 is numbered one and cacheline8 is numbered eight. As conventional, data is addressed sequentially along the RMU, as Figure 2(1) shown. In SSAM, data are addressed across the RMUs and cacheline is spread across RMUs as Figure 2(2) shown. In SIAM, data also is addressed in cacheline, but each of them is distributed along the diagonal across RMUs as Figure 2(3) shown. Each cacheline is divided into eight parts and can be read through eight phase in pipeline. As described above, we design the racetrack array which has the exact shift latency equal or smaller than the burst time span in DDR4-1600. Thus, one cacheline

can be read out through eight phase, costing the same latency just as DRAM read the data through eight bursts. Then, the shift operations are covered to the upper system.
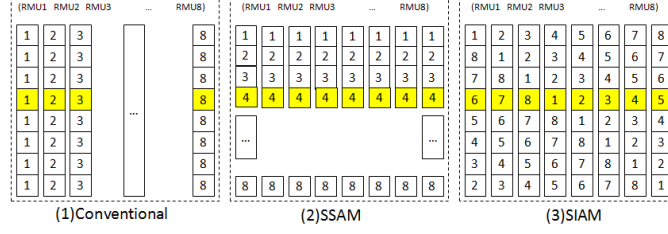

Fig. 2. Shift Insensitive Address Mapping.

4 Experimental results

4.1 Experimental setup

We evaluate the SIRM with a full system cycle accurate simulator gem5[13]. For workload, we select 13 workloads from Parsec3 benchmarks.

4.2 Comparison of Experimental Results

1) Performance Evaluation: We compare the normalized performance between DRAM, SSAM and SIRM in Figure 3. The results are normalized to baseline of DRAM main memory. SIRM and SSAM has obvious shorter access time than DRAM because of the RM's access characteristic. The SIRM has similar or better access latency than SSAM in most benchmarks except *facesim* and *x264*. Mainly because these two program is data centric and the CPU read several words from memory each time, just as we discussed in the fourth section.
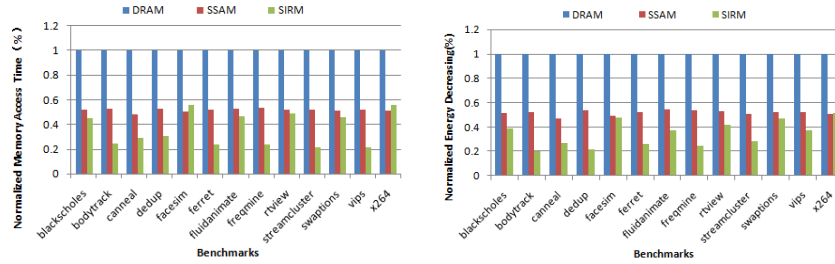

Fig. 3. Normalized memory access time and energy decrease.

2) Energy Evaluation: We compare normalized energy overhead between DRAM, SSAM and SIRM in Figure 3. All results are normalized to baseline of DRAM main memory. It is obvious that SIRM reduced much energy than

DRAM and SSAM in most benchmarks. For *facesim* and *x264*, the energy cost is similar with SSAM, mainly because SIRM uses more shift operation than other benchmarks.

5   Conclusion

Racetrack memory is attractive because of its high density and comparable read/write latency with SRAM, and non-volatility. It has the potential to be the replacement of DRAM. In this work, we presents a SIRM architecture based on specific RM array design, SIAM policy and special timing restriction in pipeline. In SIRM, the shift operations are covered and shift latency is insensitive to system level, which make RM memory easy to be managed by operating system. Experimental results show that in most applications SIRM can outperform DRAM or SSAM based racetrack main memory in performance and energy efficiency.

References

1. Chao Zhang, et al. "Quantitative Modeling of Racetrack Memory, A Tradeoff among Area, Performance, and Power," in Proceedings of the 20th Asia and South Pacific Design Automation Conference, January 2015, Chiba, Japan, pp. 100-105.
2. Yue Zhang, et al. "Perspectives of Racetrack Memory for Large-Capacity On-Chip Memory: From Device to System," IEEE Transactions on Circuits and Systems, Vol. 63, No. 5, pp. 629-638, May 2016.
3. Guangyu Sun, et al. "From Device To System: Cross-Layer Design Exploration of Racetrack Memory," in Proceedings of the 18th Design, Automation and Test in Europe (DATE 2015), March 9-13, 2015, Grenoble, France, pp. 1018-1023.
4. Parkin S S, Hayashi M, Thomas L. Magnetic domain-wall racetrack memory.[J]. Science, 2008, 320(5873):190-4.
5. Venkatesan R, et al. TapeCache: a high density, energy efficient cache based on domain wall memory[C]// Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design. ACM, 2012:185-190.
6. Haiyu Mao, et al. "Exploring Data Placement in Racetrack Memory based Scratchpad Memory," in Proceedings of the 4th IEEE Non-Volatile Memory System and Applications Symposium, August 2015, Hong Kong, China, pp. 1-5.
7. Chen X, et al. Optimizing data placement for reducing shift operations on Domain Wall Memories[C]. Design Automation Conference. ACM, 2015:1-6.
8. Qingda Hu, et al. "Exploring Main Memory Design Based on Racetrack Memory Technology," in Proceedings of the 26th ACM Great Lakes Symposium on VLSI (GLSVLSI 2016), May 18-20, 2016, Boston, MA, USA, pp. 397-402.
9. Micron. 8Gb: x4, x8, x16 DDR4 SDRAM Description. www.micron.com, 2016.
10. X. Dong et al. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 31(7), 2012.
11. $https://en.wikipedia.org/wiki/DDR4\_SDRA\sharp Mcite\_note - JESD79 - 3F - 3$
12. B. Jacob et al. Memory systems: cache, DRAM, disk. Morgan Kaufmann, 2010.
13. Binkert, N and Beckmann, B and Black, G and others. The gem5 simulator. SIGARCH Comput. Archit. 2011:1-7.