Extended Abstracts of the 2016 International Conference on Solid State Devices and Materials, Tsukuba, 2016, pp73-74

B-2-01(Invited)

Perpendicular STT-MRAM macro embedded into 40nm CMOS logic platform

Yu Lu
Hikstor Technology Ltd.
998 Wenyi West Road, Hangzhou, Zhejiang 211133, China
Phone: +86 0571-8983-8932 E-mail: lu_yu@hikstor.com

Abstract

STT-MRAM is a versatile technology that offers a cost-effective embedded memory solution in modern CMOS technologies. Perpendicular STT-MRAM promises compelling advantages in cost, performance, and power consumption when compared with conventional embedded NVM. Moreover, STT-MRAM can also function as working memory by its fast access, superior cycle endurance and byte-addressability, and as high density OTP. Perpendicular STT-MRAM is a key component for small-scale low-power information processing platform, and may enable many new IOT applications. We will discuss the advantages of STT-MRAM and their implications to IOT device designs and applications.

1. Introduction

It is well known in the industry that the latency gaps between main memory and file storage system, and between processing unit and main memory system are widening. As a result, the processing-memory-storage architecture is entering a state of dynamic complexity in large scale systems. When discussing these challenges, the growing diversity of the computing system is often overlooked. As an example, memory related challenges in IOT computing systems are quite distinct from that of stationary, performance-driven systems (Table I). In large server systems, the sizes of storage, main memory, and on-die cache are separated by orders of magnitude. While in small-scale portable computing systems, the memory hierarchy is much simpler, and the same tiers are closer in size. Other characteristics of the computing systems, such as typical applications, physical form factor, and power and thermal constraints, also determine the technical trade-off at architecture and component levels.

Table I: Examples of computing systems of different scale

Product	Dell R930	Fitbit Flex
CPU	Intel Xeon 8core	ARM Cortex-M3
L3 Cache Size	16MB	n/a
Main Memory	8-128GB	16kB
Storage size	up to 4TB	128kB eFLASH
Storage size normalized	32	8

For IOT (including some wearable) devices, simpler memory-storage architecture offers many benefits in power consumption and cost. Simpler systems also improves design productivity and reduces time to market. Unfortunately, limitations of conventional silicon-based technologies force system architects to adopt a mixture of SRAM, embedded FLASH, EEPROM, OTP, and in many cases off-chip solutions to achieve design goals.

Embedded STT-MRAM has the potential to serve the requirements of all these memory related functions in a single cost-effective technology platform.

2. Embedded 40nm perpendicular STT-MRAM[3]

In IEDM2015, Qualcomm and TDK-Headway presented an example of high performance embedded STT-MRAM based on perpendicular MTJ technology. This 1Mb embedded STT-MRAM macro (Table II) is fabricated with standard 40nm foundry CMOS and perpendicular MTJ. All process

Table II: 40nm embedded STT-MRAM macro description[3]

CMOS Base Process	Generic 40LP with 6 metal levels
Power Supply (Core/IO)	1.2V / 1.8V, no charge pump
Density	1 Mbit
MTJ	Perpendicular MTJ
Read Access Time	20 ns
Write Cycle Time	20 – 100 ns
Clock Frequency	50 MHz
IO Width	×32 / ×64
ECC	SEC-DED
Redundancy	Rows & columns (not activated for this work)

levels above the 4th metal interconnect are fabricated at TDK-Headway. A simplified process flow is shown in figure 1. The finished chip used standard packaging method without any magnetic shielding.

Figure 1: A simplified fabrication flow and a packaged chip[3]

Figure 2 shows the distribution of TMR and parallel state resistance Rp in a typical chip. Rp distribution is about 6% in this array.

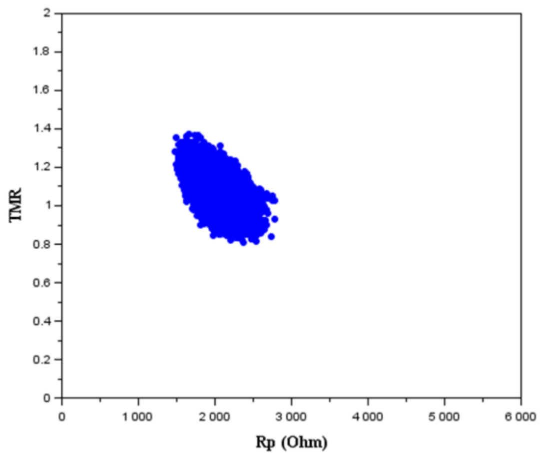

Figure 2: TMR and Rp distribution in an array [3].

This 40nm MRAM macro achieved 16ns random read access time, and can be written with a 6ns write pulse, as shown by the shmoo plots in figure 3. This performance is maintained from 0 to 70 degrees. Short write pulse is critical to reduce total write energy or average write power, since the required write current does not increase as fast as the inverse of pulse width in this regime. Fast write operation also allows the entire system to spend less time in active mode. Furthermore, fast write operation without the need for erase reduces the need for data buffering and improves overall read performance.

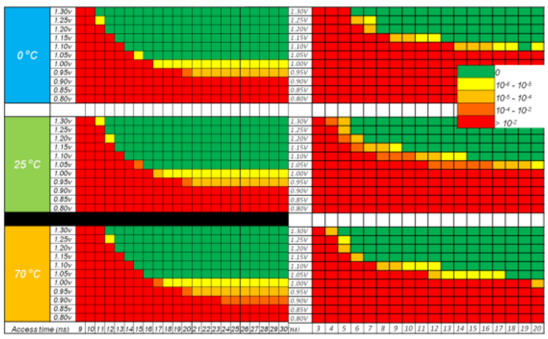

Figure 3: Read access time and write pulse width shmoo at three different temperatures.[3]

During continuous write operation (50MHz, X64, checkerboard pattern), current draw for a typical chip is around 8.6mA. This is normalized to about 3.2μW/Mbps, which is significantly lower than typical FLASH programming power requirement.

3. Conclusions

Small scale computing systems require simple memory hierarchy and versatile memory technologies. Perpendicular STT-MRAM has emerged as a cost-effective embedded memory solution that can unify many memory needs in such systems.

References

[1] K. Lee, J. J. Kan, S. H. Kang, ISLPED 2014, p.131

[2] G. Jan, L. Thomas, S. Le et al, Symp. VLSI Tech. 2015, p.164

[3] Yu Lu et al, IEDM 2015