Extended Abstracts of the 2016 International Conference on Solid State Devices and Materials, Tsukuba, 2016, pp103-104

B-5-02(Invited)

Near-future memory hierarchy with emerging nonvolatile memories and a case study of e-STT-MRAM applications

Shinobu Fujita, Hiroki Noguchi, Kazutaka Ikegami, Susumu Takeda, Kumiko Nomura and Keiko Abe
Toshiba Corporation, Corporate R&D Center
1 Komukai-Toshiba, Kawasaki, Kanagawa, 2128582 Japan
Phone: +81-44-549-2315, e-mail : shinobu.fujita@toshiba.co.jp

Abstract

This paper describes prospects of near-future memory hierarchy with emerging nonvolatile memories. One of NVM candidates is spin torque transfer (STT)-MRAM having potentials to cover working memory applications due to its high access speed and novel endurance. This paper also describes a case study of eSTT-MRAM applications.

1. Future memory hierarchy depending applications

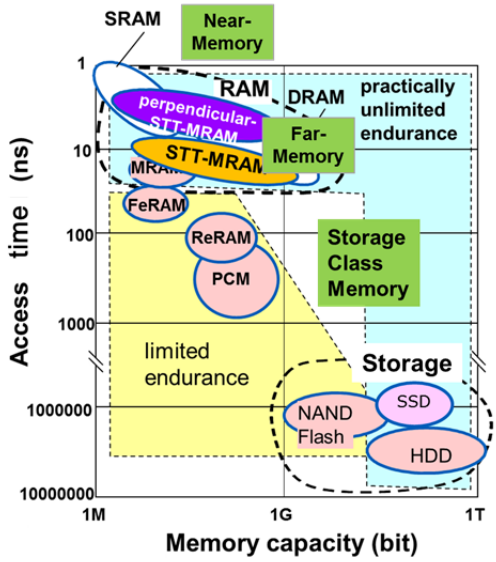

Fig.1. Memory capacity and access speed of various nonvolatile memories.

As shown in Fig.1, there are various kinds of semiconductor memories to cover wide applications. Conventional volatile memories like SRAM and DRAM are used as working memories and flash memories are used as storages. Other nonvolatile memories such as FeRAM, MRAM and PRAM have been used in niche applications. Very recently, 3D-stacked memories have been released based on NAND-flash, PRAM and DRAM to increase memory density per units. In addition, next generation nonvolatile memories such as STT-MRAM and ReRAM have been continuously developed. These new semiconductor memory trends open the opportunities to add new memory layers that can fill various memory bandwidth gaps, such as near memory, far memory, storage class memory (SCM, M-type) and S-type SCM. As a result, there are more than 7 memory layers for near-future computing, as shown in Fig.1.

However, even for high-end server applications, too many additional memory layers are not affordable because of increasing cost. It is noted that suitable memory hierarchy should be different from application to applications, which should be selected for each application. Figures 2 show some examples of expected memory hierarchy with respective application from cloud server to IoT/wearables. For example in cloud servers, CPU-GPU/near memory/storage will be used, since the bottle neck of computing performance is main memory access. Here, embedded STT-MRAM can be used in this near memory as described later. In another example, in high-speed database or big data analytic applications, CPU/far memory/SCM/storage will be used, since the storage access is a bottle neck of computing performance. When we look at edge devices such as IoT/wearables, memory hierarchy will be much simpler. The most preferable hierarchy is only CPU and “nonvolatile working memory (NWM)” combining with storage for embedded software. NVM candidate is eSTT-MRAM [6] or FeRAM. For fog-computing between the cloud and the edge devices, CPU/near memory/SCM or CPU/nonvolatile-far memory will be the most effective, if applications systems can simply customized.

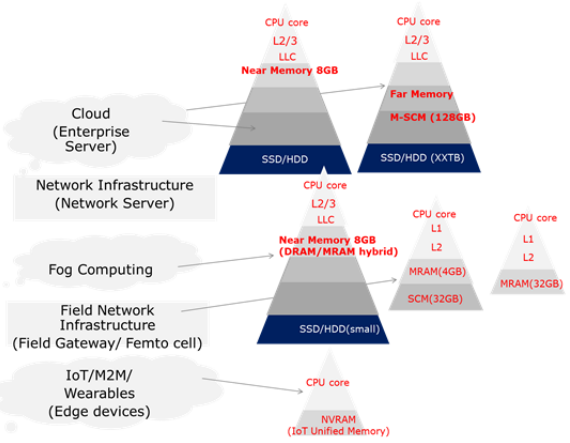

Fig 2. Near-future memory hierarchy expected for various applications.

2. New memory hierarchy with eSTT-MRAM

As shown in Fig.1, spin torque transfer (STT)-MRAM has potentials to cover working memory applications due to its high access speed and novel endurance. Furthermore, since perpendicular (p-) STT-MRAM has potentials to have higher access speed, p-STT-MRAM is expected to replace SRAM used as embedded working memory.

In advanced processors, thanks to the power gating (PG) technique to the CPU cores (including local cache L1, L2) , low-power and high-performance operation has been achieved. Volatile memories, SRAM, in the CPU cores (L1, L2 cache) aggressively use PG. Hence, average processor power depends strongly on last level cache memory capacity. Also, cache memory capacity has been dramatically increased to reach over 1Gb using volatile eDRAM, since

cache memory capacity has been simply increased for improving the performance. p-STT-MRAM based nonvolatile LLC and SRAM hybrid cache for L1/L2 is effective for power reduction in high-end processors. We reported CPU performance (execution time) and LLC energy was simulated using reported data and our measurement data, as shown in Figs.3. There was no performance degradation, and consumed energy of LLC was decreased with p-STT-MRAM by 59.6% compared with SRAM based LLC[2]. Further, LLC energy was decreased by more than 90% with fast power gating for peripheral circuits[3].

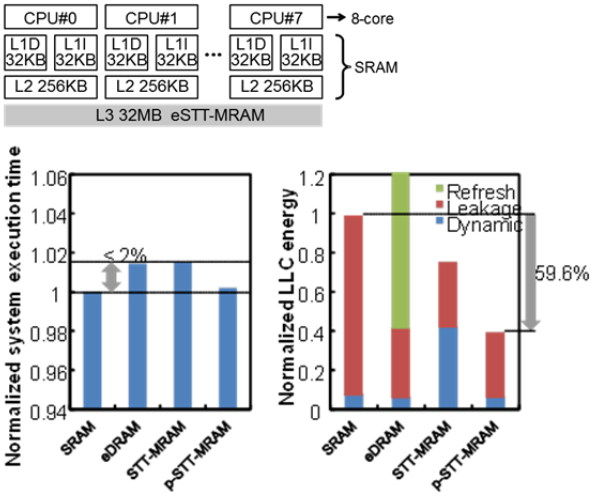

Fig. 4. Block diagram of an 8-core processor for high-end server with L1.L2-SRAM/LLC-MRAM hybrid caches and simulation results of nor execution time of CPU and consumed energy in LLC.[2]

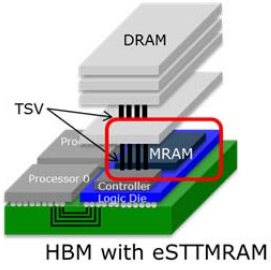

Fig. 5. Near-memory, HBM, with eSTT-MRAM.

For near-memory applications, recently 3D stacked high-memory bandwidth memories (HBM) have been commercially released. e-STT-MRAM can be implemented into the control CMOS-logic die at the bottom of HBM. This eSTT-MRAM is used as both nonvolatile cache and nonvolatile near-memory that can save DRAM energy consumption and increase data reliability. For this application, we have designed 1Gb e-STT-MRAM in a 28nm CMOS technology using 2T-2MTJ[1] cells. Cell area in the layout is about 1/4 of that of SRAM.

Suitable memory hierarchy is, thus, different from application to applications. Their respective memory hierarchies are shown in Fig.6. It is noted that a “primary” nonvolatile memory layer should be “one” (The **Sole Nonvolatile Memory** design concept.). Required retention time of additional nonvolatile memory between CPU core and the primary nonvolatile memory is not long, since it is needed to be just longer than the time while the stored data can be written back to the lower memory layer. Nonvolatile LLC with p-STT-MRAM is a typical example, and it is a “semi-nonvolatile”. A way to build up suitable memory layers for other various applications based on the “solo NVM” design concept will also be presented.

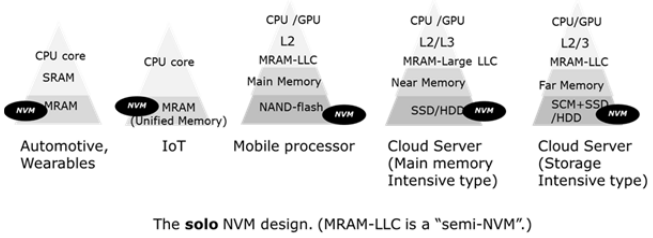

Fig.6. Near-future memory hierarchies for various applications with STT-MRAM.

Securing data in the NVM and storage is crucial due to information security. For IoT edge devices, direct attack to hardware is extremely risky like the side channel attack. To avoid such risk, streaming cryptograph for NVM outputs is much effective at bus interface using random numbers. It has been reported that STT-MRAM can be used for truly random number generators due to spin quantum physics[7]. However, the outputs of STT-MRAM based are analog signals and 0/1 balance is not 50%:50%. It was reported that a modified ECC circuit (Fig.7) of STT-MRAM can convert 0/1 balance of RNG to perfectly 50% : 50%. Therefore, it is easy to accommodate RNG using part of MRAM circuit with ECC.

Thus p-STT-MRAM is expected to play important role for various kinds of ICT key devices.

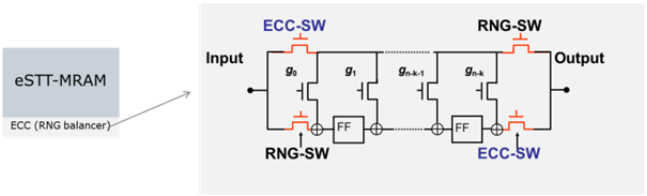

Fig. 7. ECC and RNG circuit for MRAM[7].

References

[1] H. Noguchi et. al, VLSI Technology Symposium, p.108, 2013.
[2] H. Noguchi et. al, VLSI Technology Symposium, p.97, 2014.
[3] H. Noguchi et. al, ISSCC Technical Digest, 7.2, 2016.
[4] S. Fujita et al., ISSCC 2015, Forum 2: Memory Trends: From Big Data to Wearable Devices.
[5] S. Fujita et al., IEEE International Memory Workshop 1-8, 2015.
[6] Y. Lu et al, International Electron Device Meetings, 26.1 2015
[7] S. Yuasa et. al, International Electron Device Meetings Technical Digest 3.1.1, 2013.
[8] T. Tanamoto, S. Fujita et al. Japanese Journal of Applied Physics 50, 04DM01 , 2011 .