Extended Abstracts of the 2010 International Conference on Solid State Devices and Materials, Tokyo, 2010, pp1144-1145

Hierarchical Nonvolatile Memory with Perpendicular Magnetic Tunnel Junctions for Normally-Off Computing

Keiko Abe, Kumiko Nomura, Sumio Ikegawa, Tatsuya Kishi, Hiroaki Yoda and Shinobu Fujita

Toshiba Corporation, Corporate R&D center,
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan
Phone: +81-44-549-2192 E-mail: keiko2.abe@toshiba.co.jp

1. Introduction

Static power due to a MOSFET leakage current has been a dominant component in power consumption of microprocessor since 90nm CMOS technology. The power gating can effectively decrease the static power. However, the power supply can not be cut off to volatile memories in hierarchical memory such as DRAM-based main memory and SRAM-based cache memories. The main power consumed in microprocessor is, therefore, static power for memory data retention. If these memories can be replaced by nonvolatile (NV-) RAM with high-speed operation, the static power will be reduced dramatically to close to 0W, which means that "Normally-Off computers" expected as an idealistic computing [1] is realized. This paper presents hierarchical nonvolatile memory with perpendicular magnetic tunnel junctions (p-MTJ) for the normally-off computing, as shown in Fig.1.

Since magnetoresistive random access memory with p-MTJ based on spin-transfer torque (P-STT-MRAM) can realize ultra high-density NV-RAM with high speed, low power and unlimited endurance [2], it can replace DRAM. P-STT-MRAM embedded in a CPU die can replace high capacity cache memory like level-3(L3) or lower level cache. However, P-STT-MRAM cannot replace higher speed L2 cache due to its operation speed limit. This paper describes the 6-transistor (6T-) nonvolatile (NV-) SRAM cells that can be used as a SRAM in a normal operation mode and the data can be stored into p-MTJs without power supply in a power-down state for NV-L2 cache, as described in the next section. When a processor enters into the power-down state, the power supply is cut off to the logic unit, register files and L1 cache at the same time, leakage current of these memories does not contribute the static power in the CPU core. Therefore, the presented nonvolatile hierarchical memory systems are composed of MRAM for main memory, embedded-(e-) P-STT-MRAM for L3 (or lower level) cache and 6T-NV-SRAM cell for L2 cache with p-MTJs, as shown in Fig.1.

2. 6T-Nonvolatile SRAM Cell for L2 Cache

We had previously proposed the *nonvolatile flip flop* with cross coupled inverters serially connected with two p-MTJs [3, 4]. These cross coupled inverters (latch) with two p-MTJs can be used as the 6T-NV-SRAM cell as shown in Fig. 2 (a).It is advantageous that the number of the transistors is the same as a conventional 6T-SRAM cell. Based on our precise physical layout, it has 28% area overhead. However, its area overhead is expected to be canceled, since area of e-P-STT-MRAM for L3 cache is smaller than that of conventional SRAM.

In a normal mode, the 6T-NV-SRAM cell writes data (WRITE) and reads data (READ) like a conventional SRAM cell (Fig.2 (b)). Data in the 6T-NV-SRAM cell are stored (STORE) in two p-MTJs before power shutdown. When the power is turned on again, the cell data is automatically recalled (RECALL) from two p-MTJs. Each operating condition is shown in Fig. 3. The two p-MTJs are set to the AP-state to initialize the 6T-NV-SRAM cell before WRITE (RESET). For this RESET operation, two MTJs in the 6T-NV-SRAM cell are programmed to high-resistance (anti-parallel (AP)) state by applying low level signal (L) to BL and BLB. The conditions in WRITE and READ are the same as those for conventional SRAM. The simulated delay overhead of WRITE and READ due to insertion of p-MTJs is 0ps and 27ps, respectively, in a 45nm CMOS technology. However, read delay overhead can be neglected in total delay of SRAM.

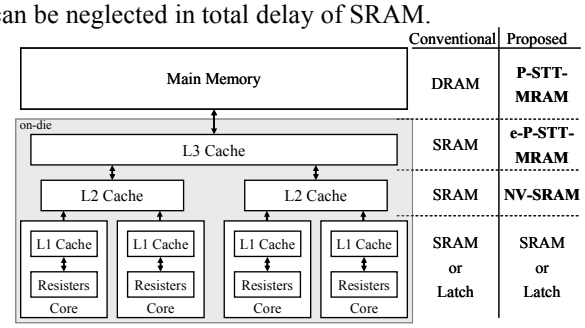

Fig. 1 Proposed NV-memory hierarchy in a multi-core processor.

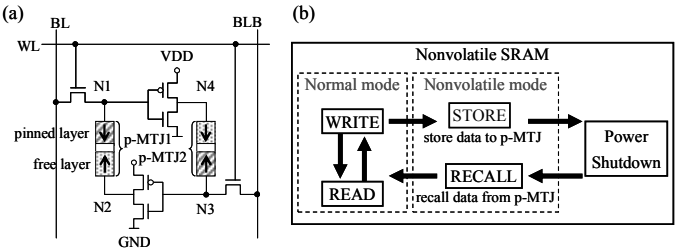

Fig. 2 (a)Circuit diagram and (b)operation of 6T-NV-SRAM cell.

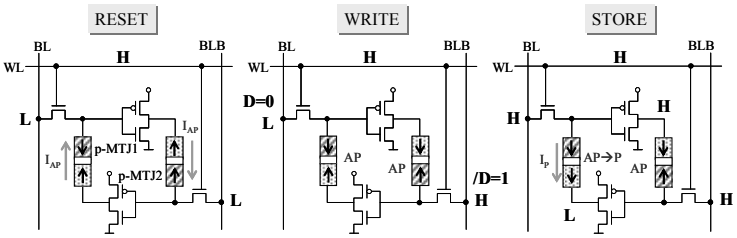

Fig. 3 RESET, WRITE and STORE operation conditions.

Fig. 4(a) shows the simulated waveforms of RECALL operation for 6T-NV-SRAM cell where the states of p-MTJ1 and p-MTJ2 are 'P' and 'AP', respectively. In RECALL operation, the BL and BLB are set to 'H' level, and then the power to cell (V_{DD}) is turned on after the WL is set to 'H' level. Consequently, the difference between resistances of the two p-MTJs is converted to the difference of voltages in N1 and N3. Thus, the cell data is recalled after V_{DD} is turned on. It should be noted that these circuit operations are designed to be simple so that operation speed becomes fast.

There is a possibility that local Vth variation of MOSFET may result in incorrect RECALL operation. RECALL operations in some conditions of magnetoresistance (MR) ratios and MRs are simulated by Hspice with Vth variations for the six transistors. Fig. 4 shows the Vth shift accepted for stable RECALL operation. It indicates that the 6T-NV-SRAM cell with higher resistance p-MTJs and larger MR ratio is highly-tolerant to the shift of Vth. P-MTJ can, in principle, meet such requirements.

3. Power saving by hierarchical nonvolatile memory

The power of microprocessor with L2 cache based on 6T-NV-SRAM cells was estimated by simulations in 45nm CMOS technology. In the case of conventional L2 cache, average powers in an active mode and a power-down mode of a micro processor are 2.4W and 0.5W, respectively. In the case of NV-SRAM-based L2 cache, the additional powers required per unit access are 5.1×10^{-13}J for RESET, 2.6×10^{-13}J for STORE, and 0 J for the power-down mode. Fig.5 shows the average power versus average ratio of power-down time for a microprocessor. The longer power-down time can reduce the power of microprocessor with NV-SRAM-based L2 cache. When the power-down time ratio is 90%, which is a general usage case for mobile applications, 65% power saving has been achieved.

In general, e-P-STT-MRAM-based L3 cache has additional delay overhead to the conventional one. We evaluated total delay overhead of the presented hierarchical NV-memory by a hierarchical memory simulator [5]. The access time of conventional SRAM in 45nm CMOS technology was calculated by CACTI 5.3 [6]. For e-P-STT-MRAM-based L3 cache, the read and write (R/W) access times are longer by 0.5nsec and 1.0nsec, respectively, than those of conventional SRAM. Fig.6 shows simulation results on R/W access times of hierarchical NV-memory as a function of interval time between access requests of CPU cores. These results indicate that delay overhead for the R/W access of the hierarchical NV-memory is small enough to be neglected. This is because the frequency to access L3 cache is very small in general. Hence, the hierarchical NV-memory is a highly effective memory system to save power without noticeable delay overhead.

4. Conclusions

Hierarchical NV-memory composed of P-STT-MRAM for main memory, e-P-STT-MRAM for L3 (or lower level) cache and NV-SRAM for L2 cache with p-MTJ are proposed for the normally-off computing. The NV-SRAM with 6 transistors and two p-MTJs can reduce the static power of L2 cache and compensate delay overhead of MRAM, since it acts as a conventional SRAM in a normal mode, and as a NV-memory in a power-down mode. The power saving of a microprocessor with the hierarchical NV-memory achieved is about 65% in the case of power-down time ratio of 90%. We have also shown that the overheads of area and R/W access delay for the hierarchical NV-memory are small enough to be neglected.

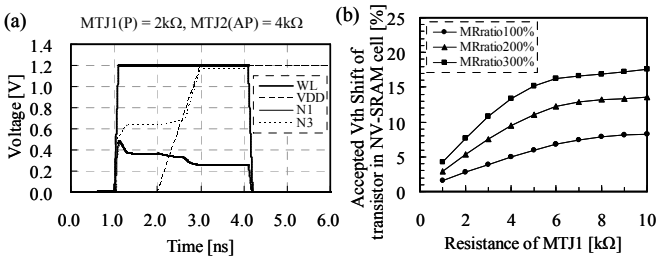

Fig. 4 (a) Simulated waveforms and (b) maximum accepted Vth shift of 6T-NV- SRAM cell in RECALL operation.

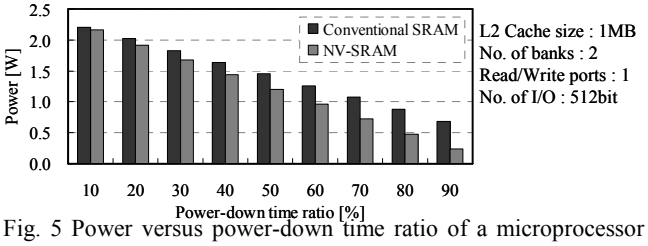

Fig. 5 Power versus power-down time ratio of a microprocessor with NV- SRAM-based L2 cache.

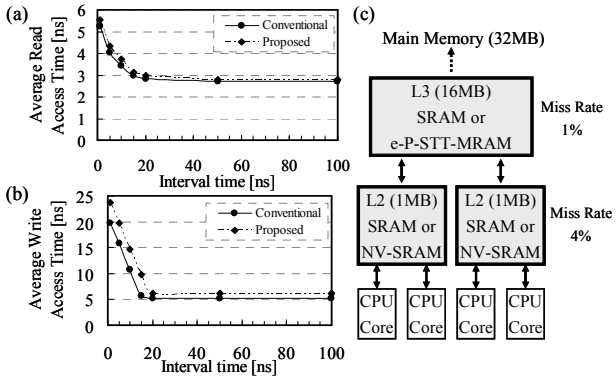

Fig. 6 Average (a) read and (b) write access times and (c) illustrated memory hierarchy.

Acknowledgements

The authors would like to thank Dr. Koji Ando for AIST for valuable discussions on the normally-off computing.

References

[1] K. Ando, AIST Today No.27 2008-winter, 16.
[2] T. Kishi et al., IEEE IEDM Tech. Dig., 12.6, 2008.
[3] K. Abe et al., European Micro and Nano Systems 2004.
[4] K. Abe et al., NSTI Nanotechnology, 3, pp. 203 – 206, 2005.
[5] K. Nomura et al., DATE'09, W5, 2009
[6] http://quid.hpl.hp.com:9081/cacti/