Extended Abstracts of the 2018 International Conference on Solid State Devices and Materials, Tokyo, 2018, pp119-120

A New Architecture of Store Energy and Latency Reduction for Nonvolatile SRAM Based on Spintronics/CMOS-Hybrid Technology

D. Kitagata, S. Yamamoto, and S. Sugahara

FIRST, Tokyo Inst. of Tech.
4259-J3-14, Nagatsuta-cho, Midori-ku, Yokohama, 226-8502, Japan
Phone: +81-45-924-5456 Email: kitagata.d@isl.titech.ac.jp

Abstract

In this paper, we develop a new architecture of store energy/latency reduction for NV-SRAM based on spintronics/CMOS-hybrid technology, called hierarchical store-free (HSF) architecture. This architecture employs addresses accessed during writ operations in the normal SRAM operation mode for judgment of store-free domains (subarray and finer block levels), and this architecture can be easily implemented by adding a simple hardware to the peripherals. Power-gating with the HSF architecture for NV-SRAM can dramatically reduce the store energy and latency depending on store-free portion of the array. The HSF architecture would be applicable to caches using NV-SRAM (or nonvolatile retention), which can lead to highly energy-efficient core-level power-gating of microprocessors and SoCs.

1. Introduction

Cache applications of nonvolatile retention have attracted considerable attention for reducing their static power dissipation in microprocessors and SoCs. Spin-transfer-torque magnetoresistive RAM (STT-MRAM) is a candidate for these applications. STT-MRAM caches are expected to achieve excellent static leakage reduction, high bit density, and moderately fast readout. However, the high store energy and long store latency for their nonvolatile memory elements (magnetic tunnel junctions; MTJs) restrict them to lower- or last-level cache (LLC) applications [1]. Nonvolatile SRAM (NV-SRAM) using MTJs is an alternative architecture using nonvolatile retention for reducing cache leakage. The normally-off and nonvolatile power-gating (NVPG) architectures have been proposed for NV-SRAM caches [2]. There are various types of NV-SRAM cell using MTJs for the normally-off architecture, and these cells can effectively shut off static leakage [3]. Nevertheless, the normally-off NV-SRAM architecture also suffers from high store energy and long store latency. Furthermore, it costs extra energy and latency for on/off-control of power-supply for every read/write operation. Recently, we have proposed a NV-SRAM cell (see Fig. 1) using MTJs for the NVPG architecture that enables to reduce static power via power gating using nonvolatile retention [4]. The cell can electrically separate the normal SRAM operations (without nonvolatile retention) and the shutdown mode with nonvolatile retention. Therefore, the cell can show fast readout/write operations comparable to ordinary SRAM caches in the normal SRAM operation mode. The write latency for the MTJs has no influence on the normal SRAM operation mode, since the nonvolatile retention is executed only for the shutdown. From these features, the NV-SRAM NVPG architecture is applicable to higher level caches including 1st level cache.

In ordinary core-level power-gating, data-transfer for important data in caches to always-on back-up memory is required. This costs huge extra energy and long latency and thus restricts the granularity (energy efficiency) of core-level power-gating. Typically, the break-even time (BET) that is a performance index used for temporal granularity evaluation of power-gating has a range between a few ms and several hundreds of ms [5]. The NV-SRAM NVPG architecture requires no data-transfer to back-up memory, and it can show moderately short BET. To further reduce BET, the store energy and latency for the MTJs need to be diminished. Various methods for reducing the store energy/latency have been proposed for STT-MRAM caches [1]. In this paper, we develop an alternative architecture of store energy/latency reduction for NV-SRAM NVPG systems, called hierarchical store-free (HSF) architecture. The HSF architecture is based on judgment of store-free domains using addresses accessed during write operations in the normal SRAM operation mode, and it can be implemented by only a simple additional hardware. The energy performance of NV-SRAM using the HSF architecture is computationally analyzed and experimentally verified from a fabricated NV-SRAM TEG.

2. NV-SRAM architecture for HSF

Fig. 2 shows the subarray organization of the proposed NV-SRAM system. The cell architecture for the NV-SRAM was shown in elsewhere [6]. The cell array is divided into 8 blocks (or more) and each block has power switches for power-gating management. The store-free indicator (SFI) memorizes information of accessed subarrays/blocks during write operations in the normal SRAM operation mode. The resulting store-unnecessary subarrays and blocks are shut down based on the information in SFI at the beginning of

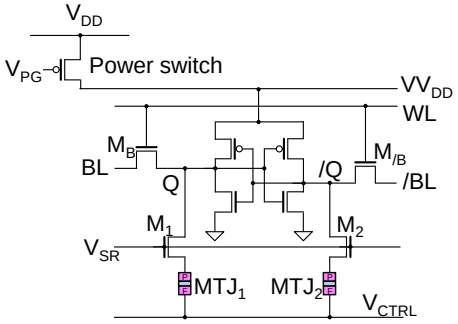

Fig.1 NV-SRAM cell configuration

the shutdown mode. The HSF architecture proceeds as follows: Firstly, the store-free subarrays that are indicated in the SFI circuits are shut down (Fig.3(a)). Then, the store-free blocks in the remaining subarrays that are also specified in the SFI circuit in the corresponding subarray are shut down (Fig.3(b)). Finally, the residual blocks are sequentially stored (Fig.3(c)) and then shut down. The store operation continues until all the blocks and subarrays are completely shut down.

3. Simulation and experimental results

Fig. 4 shows minimum BET and store-exit latency as a function of store-free proportion for the NV-SRAM system with the HSF architecture. The HSF architecture can dramatically reduce BET depending on the store-free portion of the array, as shown in the figure. This is due to in-advance shutdown of the store-free subarrays and blocks. The store-exit latency that is required to complete the store operation before entering the shutdown mode can also be effectively reduced using the HSF architecture.

Results for the simple store skipping (SSS) architecture (that is a sequential address-order store-free shutdown technique) are also shown in the figure. The effect of BET reduction is fatally weakened with increasing array size, owing to leakage currents of the store-waiting subarrays/blocks. Although, in this architecture, the store process can be skipped for the store-free subarrays/blocks, all the addresses need to be asserted. Therefore, the SSS architecture cannot reduce the exit latency.

The above-discussed results were verified using a fabricated NV-SRAM TEG. The TEG for a cell array with peripherals was implemented using 65nm CMOS process [6]. Circuit parameters for energy performance analysis were successfully extracted from the TEG, which were consistent with HSPICE simulation results. Measured BET and average static power were almost identical with simulated results, as shown in Fig. 5. NVPG with the HSF architecture would be effective at reducing BET and average static power of caches.

4. Conclusions

The HSF architecture of store energy and latency reduction for NV-SRAM using MTJs is investigated. The HSF architecture for NV-SRAM-based NVPG can effectively reduce the store energy and latency depending on store-free portion of the array. This architecture would lead to highly energy-efficient core-level power-gating of microprocessors and SoCs.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant and developed based on the previous study supported by JST. The VLSI chip in this study has been fabricated in the chip fabrication program of VDEC, the University of Tokyo in collaboration with Renesas Electronics Corp. The authors would like to thank Dr. Y. Shuto, ISEL, Tokyo Institute of Technology.

References

[1]J.Boukhobza *et al*., ACM Trans. Design Automat. Electron. Syst. **23**, p.14, 2017

[2]S.Yamamoto *et al*., Jpn. J. Appl. Phys. **48**, 4, pp.043001/1-7, 2009.

[3]K.Abe *et al*., SSDM2010, paper F-9-3.

[4]Y.Shuto *et al*., JJAP **51**, pp.040212/1-3, 2012.

[5]V.George *et al*., ASSCC2007, pp.14-17.

[6]Y.Shuto *et al*., IEEE ESSCIRC 2016, pp.95-98.

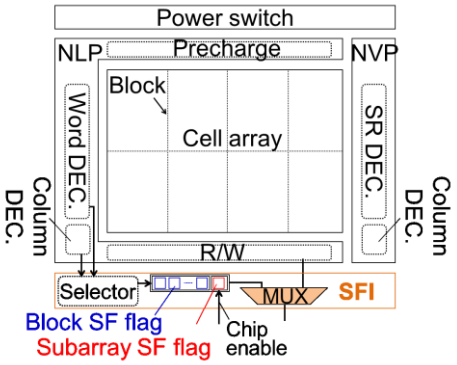

Fig.2 Subarray organization

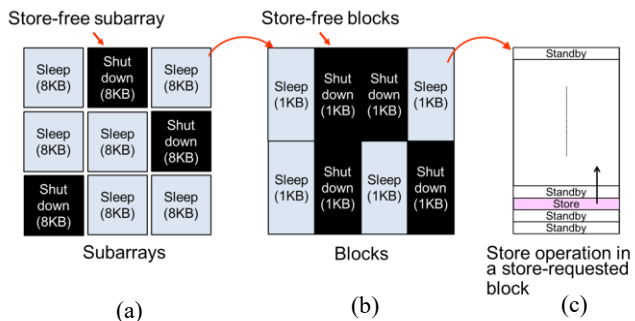

Fig.3 Schematics of the HSF architecture. (a) Subarray-level store-free shutdown. (b) Block-level store-free shutdown in a subarray. (c) Store operation in a block.

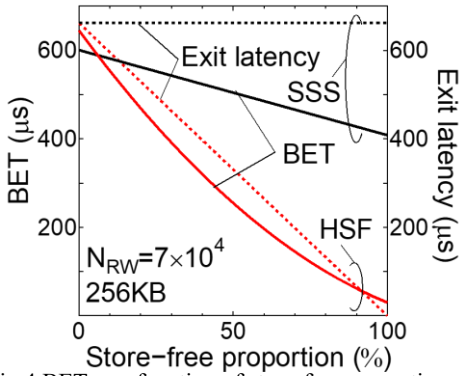

Fig.4 BET as a function of store-free proportion

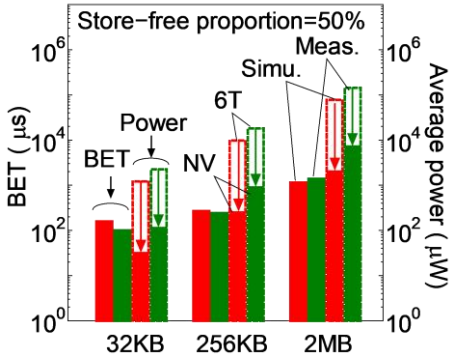

Fig.5 BET and average static power for NV-SRAM with HSF architecture