Stacking magnetic random access memory atop microprocessors: an architecture-level evaluation

X. Dong\textsuperscript{1} X. Wu\textsuperscript{2} Y. Xie\textsuperscript{1} Y. Chen\textsuperscript{3} H. Li\textsuperscript{4}

\textsuperscript{1}Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA16802, USA
\textsuperscript{2}Advanced Technology Group, Qualcomm
\textsuperscript{3}Department of Electrical and Computer Engineering, University of Pittsburgh
\textsuperscript{4}Department of Electrical and Computer Engineering, Polytechnic Institute of New York University
E-mail: xydong@cse.psu.edu

Abstract: Magnetic random access memory (MRAM) has been considered as a promising memory technology because of its attractive properties such as non-volatility, fast access, zero standby leakage and high density. Although integrating MRAM with complementary metal-oxide-semiconductor (CMOS) logic may incur extra manufacturing cost because of the hybrid magnetic-CMOS fabrication process, it is feasible and cost-effective to fabricate MRAM and CMOS logic separately and then integrate them using 3D stacking. In this work, we first studied the MRAM properties and built an MRAM cache model in terms of performance, energy and area. Using this model, we evaluated the impact of stacking MRAM caches atop microprocessor cores and compared MRAM against its static random access memory (SRAM) and dynamic random access memory (DRAM) counterparts. Our simulation result shows that MRAM stacking can provide competitive instruction-per-cycle (IPC) performance with a large reduction in power consumption.

1 Introduction

Unlike traditional memory technologies that use electric charges as the information carrier, magnetic random access memory (MRAM) uses Magnetic Tunnel Junction (MTJ) as its binary storage. MRAM has been under development since the 1990s\cite{1}. In the last several years, MRAM has been proposed and developed by different companies\cite{2–6}. MRAM has several desired properties such as non-volatility, fast access, zero standby leakage and high density. Since the conventional toggle-mode MRAM suffers slow write speed and high write power consumption\cite{7–9}, a second-generation MRAM technique called spin-torque transfer MRAM (STT-RAM or SP-RAM) becomes the most popular design owing to its better scalability property, higher speed and lower power consumption\cite{5, 6}. In this work, MRAM refers to STT-RAM without explicit explanations.

The MRAM fabrication involves hybrid magnetic-CMOS processes (i.e. the MRAM process requires growing a magnetic layer between two metal layers), and so it may incur extra cost and additional fabrication complexity to integrate MRAM with conventional CMOS logic into a conventional 2D chip. Nevertheless, 3D integration\cite{10} provides a feasible and cost-effective solution. Multiple active device layers can be stacked in a 3D chip with through-silicon-vias (TSVs). The separation of MRAM and CMOS chip fabrications makes it a cost-effective way of integrating MRAM memory resources into a CMOS logic chip.

While most of the recent MRAM research is focused on the MRAM fabrication process, MRAM device model\cite{11, 12}, or MRAM chip design\cite{2–5}, the architectural-level benefits of using MRAM as a universal memory replacement for SRAM or DRAM are not well evaluated yet. Therefore the main objective of this work is to extend the MRAM research from circuit level to microarchitecture level. In this work, we first studied the MRAM cell properties on device and circuit levels. After extracting an abstraction of these properties, we enhanced a widely used cache simulator in the computer architecture community called CACTI\cite{13} to be MRAM aware; we then compared MRAM caches with its SRAM and DRAM counterparts with regard to performance, energy and area; finally we explored several different 3D MRAM stacking options including extending L2 cache size, adding L3 cache hierarchy and directly using MRAM as the main memory.

2 MRAM background and MRAM cell model

As illustrated in Fig. 1, MRAM cells are usually built in a one-transistor-one-MTJ (‘1T1J’) structure\cite{5, 6}, in which MTJ is the storage device and metal-oxide-semiconductor field-effect transistor (MOSFET) is the access device. Each MTJ contains two ferromagnetic layers and one tunnel barrier layer (MgO). The direction of one ferromagnetic layer called reference layer is fixed, while the direction of the other layer called free layer can be changed by passing a driving current. If these two ferromagnetic layers have the same directions, the resistance of MTJ is low, indicating a