Preface

The analysis of big data at exascale (10^{18} bytes or flops) has introduced the emerging need to reexamine the existing hardware platform that can support intensive data-oriented computing. A big-data-driven application requires huge bandwidth with maintained low-power density. For example, web-searching application involves crawling, comparing, ranking, and paging of billions of web pages with extensive memory access. At the same time, the analysis of such a huge data at exascale is a national interest due to cybersecurity need. One needs to provide scalable big-data storage and processing solution that can detect malicious attack from the sea of data, which is beyond the capability of a pure software-based data analytic solution. The key bottleneck is from the current data storage and processing hardware, which has not only the well-known memory wall and power wall with limited accessing bandwidth but also large leakage power at advanced CMOS technology nodes. One needs to design an energy-efficient hardware platform for future big-data storage that can also support data-intensive processing for data recognition in both image and security applications.

Memory is any physical device that is able to temporarily or permanently hold the state of information. Memories can be generally classified into two categories: volatile and nonvolatile. Static and dynamic random-access memories (SRAM and DRAM) are examples of volatile memories that can be accessed in nanosecond of speed, but the stored data will be lost when powered off. The flash and hard disk drive (HDD) are examples of nonvolatile memories. Imagine the life where one can start a computer in the blink of an eye, without having to wait for the operation system to load, or transfer full-length high-definition movie by memory stick in seconds rather than hours. Such a life would happen if a universal nonvolatile memory could be developed that not only can retain information without external power but also can be accessed in high speed. In general, the following criteria examine the new memory technologies: (1) Scalability for high-density integration (2) Low energy consumption for mobile access (3) High endurance capable of 10^{12} writing/erasing cycles

The existing memory technologies have critical challenges of scaling at nanoscale due to process variation, leakage current, and I/O access limitations.

Recently, there are two research trends that attempted to alleviate the memory-wall and power-wall issues for future big-data storage and processing system. Firstly, the emerging nonvolatile memory technologies such as the resistive RAM (ReRAM), spin-transfer torque RAM (STT-RAM), and domain-wall nanowire racetrack memory have shown significantly reduced standby power and increased integration density, as well as close to DRAM/SRAM access speed. Therefore, they are considered as promising candidates of universal memory for big-data applications. Secondly, due to high data-level parallelism in big-data applications, a large number of application-specific accelerators can be deployed for data processing. However, such a memory–logic integration approach will still incur I/O overhead. Instead, an in-memory-based domain-specific computation will be highly desired with less dependence on I/Os.

In order to achieve low power and high throughput (or energy efficiency) in big-data computing, one can build an in-memory nonvolatile memory (NVM) hardware platform, where both the memory and computing resources are based on NVM devices with instant switch-on as well as ultralow leakage current. This can result in significant power reduction due to the nonvolatility. Moreover, one can develop NVM logic accelerator that can perform domain-specific computation such as machine learning in a logic-in-memory fashion. In contrast, for the conventional memory–logic integration architecture, the storage data must be loaded into the volatile main memory, processed by logic, and written back afterwards with significant I/O communication overhead.

In this book, we plan the following research studies in this regard. Firstly, we introduce a nonvolatile big-data storage design platform that can evaluate both the current NVM technology and future NVM technology. We develop a SPICE-like simulator NVM SPICE, which implements physical models for nonvolatile devices in a similar way as the BSIM model for MOSFET. We further develop an advanced NVM design platform that can provide evaluation of various memory cell structures as well as corresponding readout circuits. As such, one can perform an accurate and efficient estimation of memory performance at microarchitecture level. Secondly, we study in-memory NVM computing architecture for domain-specific big-data storage and processing. We illustrate the NVM-based basic memory and logic components and find significant power reduction. We further illustrate in-memory machine learning such as extreme learning machine (ELM) for big-data image recognition as well as security classification, which can be evaluated based on both developed NVM design platforms.

This book provides a state-of-the-art summary for the latest literature on emerging nonvolatile memory technologies and covers the entire design flow from device, circuit, to system perspectives, which is organized into five chapters. Chapter 1 covers the basics of memory and review of existing memory technologies and emerging nonvolatile memory technologies. Chapter 2 introduces the physics of the emerging nonvolatile memory as well as the agreeing computing architecture. Chapter 2 details the device characterization for the emerging nonvolatile memory by nonelectrical states. Chapter 4 explores the circuit level design techniques for the emerging nonvolatile memory. Chapter 5 presents the system-level architectures

with applications for the emerging nonvolatile memory. This book assumes that readers have basic knowledge of semiconductor device physics. This book will be a good reference for senior undergraduate and graduate students who are performing researches on nonvolatile memory technologies.

Finally, the authors would like to thank their colleagues at CMOS Emerging Technology Group at Nanyang Technological University: Wei Fei, Yang Shang, Xiwei Huang, Chun Zhang, Sai Manoj Pudukotai Dinakarrao, and Shuai Chen. The authors also owe their grateful discussion to Prof. Roy Kaushik, Prof. Dennis Sylvester, Prof. Kevin Cao, Prof. Weisheng Zhao, Prof. Yuan Xie, Prof. Yiran Chen, Prof. Hai Li, Dr. Tanay Karnik, Dr. Jing Li, Prof. Wei Zhang, Prof. Tony Kim, Prof. Wen-Siang Lew, Prof. Chip-hong Chang, and Prof. Kiat-Seng Yeo. Their support is invaluable to us during the writing of this book. The relevant research is funded by MOE Tier-2 (MOE2010-T2-2-037), A*STAR PSF (1120120 2015), and NRF CRP (NRF-CRP9-2011-01) from Singapore as well as industry research collaboration fund from Huawei Shannon Lab.

Singapore — Hao Yu
Singapore — Yuhao Wang
January 1, 2014

http://www.springer.com/978-1-4939-0550-8

Design Exploration of Emerging Nano-scale Non-volatile Memory

Yu, H.; Wang, Y.

2014, X, 192 p. 377 illus., Hardcover
ISBN: 978-1-4939-0550-8