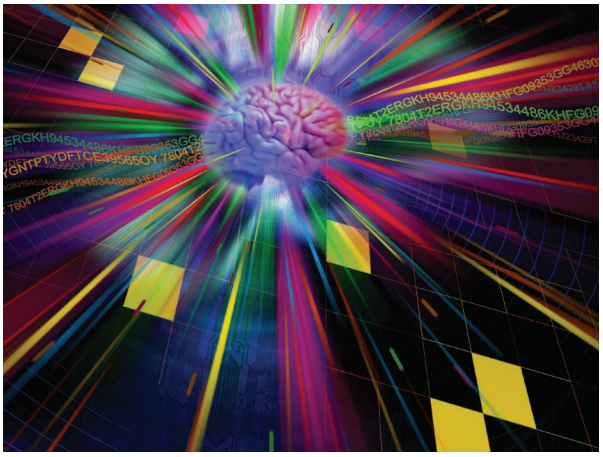

Next-Generation Memory

Greg Atwood, *Micron Technology*
Soo-Ik Chae, *Seoul National University*
Simon S.Y. Shim, *San Jose State University*

A defining component of today's mobile devices and computing systems, memory technology is both a solution and a bottleneck, and industry is racing to redefine its use for future systems.

The memory technologies in today's computing systems all emerged in the early 1970s at the dawn of the semiconductor industry. Solid-state memory—static RAM (SRAM), dynamic RAM (DRAM), and flash (initially EPROM)—is based on electron storage in transistors, while mechanical memory—tape and the hard disk drive (HDD)—relies on magnetic storage. These two storage media have had an amazingly long life, roughly doubling in density and halving in cost every two years according to Moore's law. The scalability of these technologies explains why they have been a key enabler in the emergence of increasingly complex computing devices.

Memory is a defining component in many of today's portable devices that themselves are becoming indispensable to our lives. In a smartphone or tablet, memory is typically second, and often equal, to the display as the largest cost component of the system, well above the CPU. For high-end computing systems such as servers, memory defines system performance and power more than any other component. As the demand for large amounts of data instantly accessible to millions of users continues to increase, memory technology becomes both a solution and a bottleneck, spurring the industry to redefine how these systems use memory. One of the best examples of this is the emergence of solid-state drives (SSDs) across the range of computing devices.

The rapid density growth and cost reduction of memory has been enabled by scaling the underlying technologies to ever-smaller feature sizes and thus smaller memory-bit sizes. Unfortunately, this scaling is starting to approach the limitations of the storage physics, making the task of maintaining the cost/density value curve increasingly difficult. In a state-of-the-art 20-nm NAND multilevel cell (MLC) memory unit, the memory state is stored using only a few tens of electrons. Increasing management of the memory bits is required for the system to be able to continue to use them. Fortunately, the ability of modern manufacturing tools to manipulate materials at the atomic level, combined with an improved understanding of and ability to model fundamental storage physics, is opening the door for new advances in memory. Emerging memory devices could prove to be even more scalable than current devices, ensuring continuation of the cost/density value curve.

IN THIS ISSUE

It is generally agreed that today's memory technologies, such as NAND flash, are still scalable for several generations, but it is equally clear that physical limitations will soon slow their pace of scaling and further reduce their functionality. Increasing levels of memory management will therefore be needed to prolong the life of these technologies. "NAND Flash Memory: Challenges and Opportunities," by Yan Li and Khandker N. Quader at SanDisk, examines the scaling direction for NAND and the management techniques that will be required.

0018-9162/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

New memory technologies generally do not function like today's memory technologies and have different tradeoffs for performance, power, and cost. Similar to the way SSDs are changing the hierarchy of memory usage, emerging technologies are likely to have a transformational effect on memory usage and integration in computer systems. In "What Lies Ahead for Resistance-Based Memory Technologies?," Yoon-Jong Song and his coauthors at Samsung Electronics highlight and discuss the cutting-edge physics and features that new memory technologies enable.

There is a reason why existing memory technologies have been so successful and no significant new ones have been introduced in 40 years. Introducing a new memory type is complex: the storage physics are not well understood, designing large-density storage devices and state-of-the-art lithography techniques is challenging, and manufacturability at high volumes and yields is unproven. Consequently, it can take up to a decade for a new memory technology to evolve from a basic concept demonstration to a finished commercial product.

New memory types likely will first supplement existing memory technology to help overcome the latter's scaling deficiencies. They might also find an entry point in applications that leverage their unique set of features. Phase-change memory (PCM), magnetic memory, and ferroelectric memory have all emerged at various levels of density and lithography. Of these, only PCM has been demonstrated at Gbyte-level densities and "near state of the art" lithography. PCM combines some of the properties of DRAM and NAND, providing a new set of features. Although PCM is the first, several other technologies under active development promise similar features.

The industry now faces two challenges: developing novel memory types to enable further density scaling, and integrating these into computer systems. These challenges are compounded by the high likelihood that the new memory technologies will have different functionality than the existing ones. In "The Nonvolatile Memory Transformation of Client Storage," Intel's Amber Huffman and Dale Juenemann examine the transformation that SSDs enable in client systems and the potential for emerging memory types beyond the SSD.

As new memory technologies are introduced into computer systems, so changes the memory hierarchy, which has been defined by the evolution of memory-type capabilities over the past four decades. How computer hardware and software deal with memory has been defined by its access latency, access granularity, volatility, power, and cost. The memory hierarchy has evolved into a multitier system with various levels of caching in SRAM, main memory in DRAM, "fast" storage in the HDD and now SSD, and "slow" storage in tape.

The memory hierarchy, and the resulting hardware and software wrapped around it, are as much defined by the perceived deficiencies of the memory types as by their advantages. For example, DRAM, the most common memory for main code and data storage, will "forget" what is stored when the power is removed. Worse, it will forget even when the power is on, resulting in the requirement to periodically refresh the data. The refresh consumes time, power, and computer resources to manage it. Certainly these are not desirable properties, but because DRAM could be built for a reasonable cost and was scalable, system designers learned how to deal with its inadequacies. Similar observations can be made about other memory technologies, none of which are ideal, and system infrastructure has been built up around extracting value out of each of them.

Hence, new memory types will significantly impact software design, which takes advantage of the memory hierarchy. In "How Persistent Memory Will Change Software Systems," Anirudh Badam of Microsoft Research addresses the potential exploitation and effect of emerging memory technologies on the OS. And in "Refactor, Reduce, Recycle: Restructuring the I/O Stack for the Future of Storage," Steven Swanson and Adrian M. Caulfield of the University of California, San Diego, focus on the software-to-hardware interface, evaluating the potential repartitioning of functions across this boundary enabled by such technologies.

Many new memory types are in various stages of commercial development. PCM, commercialized by Micron Technology and Samsung, is on hold until market conditions support mass production. The industry faces significant challenges over the next 10 years as it determines how these emerging memory technologies will evolve from both a manufacturing and memory hierarchy point of view. The impact of these developments will be extremely long lasting.

Greg Atwood *is a senior fellow in the Research and Development Group at Micron Technology. His primary focus is on emerging memory technologies and systems. Atwood received an MS in physics from Purdue University. Contact him at gatwood@micron.com.*

Soo-Ik Chae *is a professor in the Department of Electrical Engineering and Computer Science at Seoul National University. His research interests include VLSI implementation and video systems. Chae received a PhD in electrical engineering from Stanford University. He is a member of the IEEE Computer Society. Contact him at chae@snu.ac.kr.*

Simon S.Y. Shim *is a professor in the Computer Engineering Department at San Jose State University. His research interests include Internet computing, high-performance computing, and databases. Shim received a PhD in computer science and engineering from the University of Minnesota. He is a member of the IEEE Computer Society. Contact him at simon.shim@sjsu.edu.*