Expert Opinion

# With New Memories Come New Challenges

**Moinuddin Qureshi**
Georgia Institute of Technology

**The performance of** a modern computer system is increasingly getting dictated by the performance of the memory system. As we pack more cores, more threads per core, and domain-specific accelerators on the chip, the memory system must scale in both capacity and bandwidth to supply the data to all these computing units. For the past four decades, we have had mainly two technologies for building the components of the memory system: SRAM for caches and DRAM for main memories, and systems have been limited by the constraints of these two technologies. We are at a juncture where several new memory technologies are emerging that can overcome the limitations of conventional technologies and offer new features such as nonvolatility. However, with new technologies come new challenges and system designers will need innovative solutions to address these challenges in order to fully realize the benefits offered by the emerging technologies.

We are at a juncture where several new memory technologies are emerging that can overcome the limitations of conventional technologies and offer new features such as nonvolatility.

*The Quest for Large On-Chip Caches:* Spin-torque transfer magnetic random access memory (STT-RAM) is a technology that can offer 3× -4× higher density than SRAM, which can help with building much larger on-chip caches. However, the challenges with STT-RAM are that it has higher write latency than SRAM and has new failure modes (such as retention failure due to thermal strikes). While the problem of write latency can be overcome using a hybrid organization where frequently written data stay in the SRAM and less frequently written data are resident in STT-RAM, the problem of reliability is still an open challenge and this problem is likely to become severe as the technology is scaled to smaller feature sizes.[1] Innovative solutions that can mitigate the high rate of failures without requiring significant overheads in terms of performance and storage will be helpful in exploiting the density advantages offered by STT-RAM.

*The Quest for Higher Bandwidth Memory:* Three-dimensional die-stacking and silicon interposer technologies can allow for tighter integration of memory modules with the processor chip, thereby dramatically increasing the overall bandwidth between the processor and the memory substrate. Technologies such as the High Bandwidth Memory (HBM) already provide 4× -8× more bandwidth than commodity DRAM and are getting adopted by the industry in premium products. One of the limitations of the die-stacked memory is that the capacity of these modules is still limited to only a few gigabytes, and the technology is not mature yet to reach the capacity of commodity DRAM modules. Therefore, stacked-DRAM modules are usually used along with commodity DRAM and the hardware/software[2] is responsible for deciding the placement of the data that must be placed in the higher bandwidth memory versus the commodity DRAM memory.

*The Quest for Lower Latency Memory:* DRAM memory modules are primarily optimized for

*Digital Object Identifier 10.1109/MM.2019.2892195*
*Date of current version 21 February 2019.*



Published by the IEEE Computer Society

density and latency is usually considered of secondary importance. Reducing memory latency can significantly improve system performance. There are variants of DRAM technology that are optimized for latency, called reduced latency DRAM (RL-DRAM), that tries to reduce the bit-line capacitance by amortizing the sensing circuit overall a smaller data array.[3] Alternative technologies that can reduce memory latency include rethinking the organization of DRAM array using the 3-D die stacking technology.[4] The key challenge with these technologies is the added cost, as both RL-DRAM and DRAM redesigned for 3-D will likely continue to be more expensive than commodity DRAM. Therefore, these technologies will either remain limited to systems where latency is ultra critical (e.g., network switches or stock trading), or likely to be adopted along with commodity DRAM, with hardware/software responsible for doing intelligent data placement.

*The Quest for Higher Capacity Memory:* With the slowdown of DRAM scaling and the continued demand for larger memory capacity from data-intensive applications, there is a need for a memory technology that has performance close to DRAM and yet provides much higher density than DRAM. Emerging technologies, such as phase change memory (PCM) and 3-D cross point (3-D XPoint), promise to offer 4× higher density than DRAM thereby enabling large memory systems. However, these technologies suffer from the challenges of higher read latency, asymmetric read–write times, reduced write bandwidth, and limited endurance.[5] To effectively utilize the density advantages of these technologies, systems need to provision large DRAM buffers such that the application performance is dictated by DRAM instead of the latency/bandwidth of these new technologies. Furthermore, novel error-correction techniques and wear-leveling solutions are needed to mitigate the reliability challenges that are inherent in these new technologies.

*The Quest for Persistent Memory:* DRAM is a volatile memory, in that it loses states when power is turned off, therefore data that need to be preserved must be explicitly written to storage (SSD or hard-drive). The emerging technologies (PCM and 3-D XPoint) for main memory offer nonvolatility and can avoid such extra writes of persistent data from memory to storage. Storage is typically accessed at the granularity of few kilobytes and using OS calls, whereas memory can be modified at the granularity of a few bytes and directly using load/store instructions, therefore such persistent memory systems promise significantly lower latency and improved efficiency for doing updates to persistent data. The main challenges for the persistent memory system are to develop hardware solutions[6] that can provide ACID semantics efficiently and transparently, and develop software interfaces[7] so that applications can easily use persistent memory.

While emerging memory technologies promise a dramatic improvement in the capabilities for future systems, we will be able to realize these benefits only if we can efficiently address the problems and challenges of these technologies. It is an exciting time to be working on these problems both in the hardware and the software community.

## REFERENCES


1. H. Naemi, C. Augustine, A. Raychowdhury, L. Shih-Lien, and J. Tschanz, "STTRAM scaling and retention failures," *Intel Technol. J.*, vol. 39, no. 10, pp. 54–75, 2013
2. C. Chou, A. Jaleel, and M. Qureshi, "BATMAN: Techniques for maximizing system bandwidth of memory systems with stacked-DRAM," in *Proc. Int. Symp. Memory Syst.*, 2017, pp. 268–280.
3. D. Lee, Y. Kim, V. Seshadri, J. Liu, L. Subramanian, and O. Multu, "Tiered-latency DRAM: A low latency and low cost DRAM architecture," in *Proc. Int. Symp. High Perform. Comput. Archit.*, 2013, pp. 615–626.
4. G. Loh, "3D-stacked memory architectures for multi-core processors," in *Proc. Int. Symp. Comput. Archit.*, 2008, pp. 453–464.
5. M. Qureshi, S. Gurumurthi, and B. Rajendran, "Phase change memory: From devices to systems," in *Synthesis Lectures on Computer Architecture*, San Rafael, CA, USA: Morgan & Claypool, 2011.
6. A. Kolli, "Architecting persistent memory systems," Ph.D. dissertation, Univ. Michigan, Ann Arbor, MI, USA, 2017.
7. H. Volos, A. J. Tack, and M. M. Swift, "Mnemosyne: Lightweight persistent memory," in *Proc. Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, 2011, pp. 91–104.



**Moinuddin Qureshi** is a Professor of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA, USA. Contact him at moin@gatech.edu.