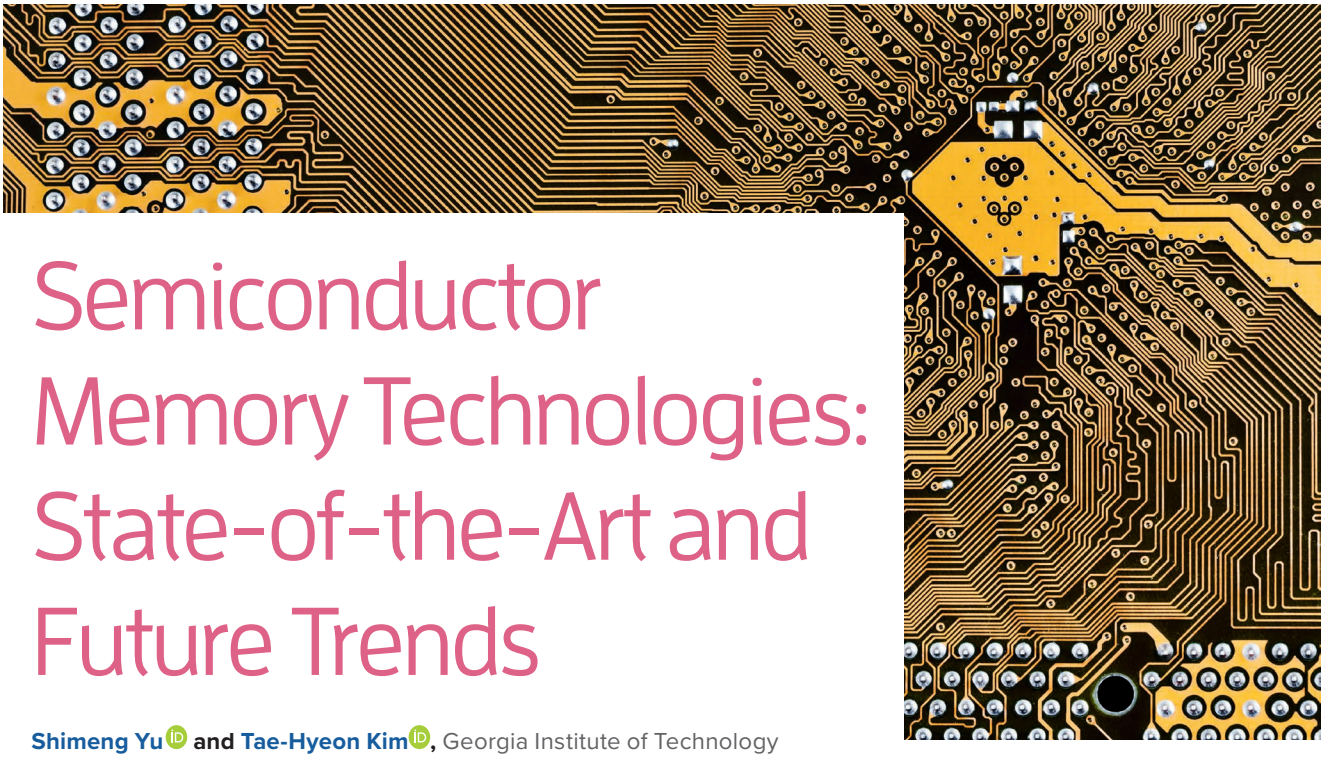

Semiconductor Memory Technologies: State-of-the-Art and Future Trends

Shimeng Yu and **Tae-Hyeon Kim**, Georgia Institute of Technology

This article surveys the recent development of semiconductor memory technologies spanning from the mainstream static random-access memory, dynamic random-access memory, and flash memory toward emerging candidates such as resistive, ferroelectric, and magnetic memories. Pathways for future technological innovations are presented.

Semiconductor memory technologies play a pivotal role in modern computing systems, serving as the primary means of storing and retrieving digital information. These technologies encompass a diverse range of memory types, each with unique characteristics suited for specific applications. Demand for higher capacity, faster speed, and lower power consumption continues to drive innovation in memory technologies. Additionally, the proliferation of data-intensive applications like artificial intelligence (AI), machine learning (ML), and big data analytics fuels the need for more advanced memory solutions. As a result, the semiconductor memory market is expected to remain robust, with ongoing developments shaping its landscape.

The memory hierarchy traditionally refers to the organization of different types of memory in computing systems, ranging from high-speed, low-capacity registers and caches to slower but larger main memory and persistent storage. However, emerging trends like Compute Express Link (CXL) are blurring the boundaries of this hierarchy. CXL, a high-speed interconnect standard, enables processors to access various types of memory and accelerators as if they were part of the CPU's memory space. This architecture allows for more flexible and efficient data movement between

Digital Object Identifier 10.1109/MC.2024.3363269
Date of current version: 5 April 2024

0018-9162/24©2024IEEE

EDITOR **CONRAD JAMES**
Sandia National Laboratories, USA;
cdjame@sandia.gov

different memory types, including working memory, storage-class memory, and even AI accelerators like GPUs or field-programmable gate arrays. As a result, the distinction between different layers of the memory hierarchy becomes less rigid, with memory resources becoming more tightly integrated and accessible across the system. This blurring of boundaries offers opportunities for improved performance, energy efficiency, and scalability in modern computing systems, as data-intensive workloads can leverage a more unified and versatile memory architecture. Nevertheless, the fundamental building blocks of such a versatile memory architecture remain upon the underlying memory device technologies. In the following, the mainstream and emerging memory device technologies are surveyed. State of the art from the industry and future trends of these technologies are discussed.

STATIC RANDOM ACCESS MEMORY

Static random-access memory (SRAM) is widely used as the on-chip cache for microprocessors including CPU/GPU and domain-specific accelerators such as tensor processing units. SRAM is still irreplaceable owing to its subnanosecond access speed and unlimited endurance. Depending on the applications, SRAM's bit cell design features high-density or high-performance variants (mainly by sizing the number of fins in the FinFET era). Figure 1 shows the historical scaling trends in the SRAM bit cell area (for the high-density cell) from the planar transistor era to today's FinFET era. The data points are collected from the industrial reports in leading conferences. The representative microscopic views of the six-transistor bit cell are also shown. As is shown, the SRAM enjoys the scaling benefits of the logic process to the 5-nm/3-nm node, reaching the bit density around 30 Mbit/mm^2. However, the scaling rate has significantly slowed down in recent years. Taking the Taiwan Semiconductor Manufacturing Company's (TSMC's) technology as an example, from 5-nm node to 3-nm node only 5% area reduction is achieved when the high-density bit cell area reduces from 0.021 μm^2 at 5-nm node to 0.0199 μm^2 at 3-nm node.[2] Three-dimensional die stacking of SRAM by advanced packaging techniques, for example, hybrid bonding as used in Advanced Micro Devices' 3D V-Cache,[3] enables the 768-Mbit ultralarge last-level cache for high-performance computing. It is noted that the second generation of 3D V-Cache still used a less advanced 7-nm node for SRAM dies while the processor cores are on a more advanced 5-nm node. The future challenges of SRAM design require innovations in design-technology cooptimization, for example, double/triple layers of wires for wordline/bitline to reduce the parasitic interconnect resistance, backside power rail and power delivery network, stacked nanosheet transistor, folded SRAM bit cell in monolithic 3D integration with complementary field-effect transistor, and so on.

DYNAMIC RANDOM-ACCESS MEMORY

Dynamic random-access memory (DRAM) is used as the main memory, and it is often regarded as off-chip standalone memory with input/output (I/O) links communicating with microprocessors/accelerators. Depending on the applications, DRAM products have different I/O interface protocols such as double data rate (DDR), low power DDR (LPDDR), graphic DDR (GDDR), and high-bandwidth-memory (HBM). For HBM, multiple DRAM dies are stacked vertically with microbump and through-silicon-via and is controlled by the logic base die. High-performance computing platforms are often equipped with GDDR or HBM owing to their ultrafast bandwidth. Figure 2 shows the normalized bit density scaling of various memory technologies, and Figure 3 shows the storage capacity of various memory technologies. DRAM's scaling as of 2023 has reached 12-nm node and the bit density reaches more than 300 Mbit/mm^2 for DDR5[4] and exceed 1 Gbit/mm^2 for HBM3 that employs the 3D die stacking. Extreme ultraviolet lithography and high-k/metal-gate peripheral logic processes have been introduced for DRAM

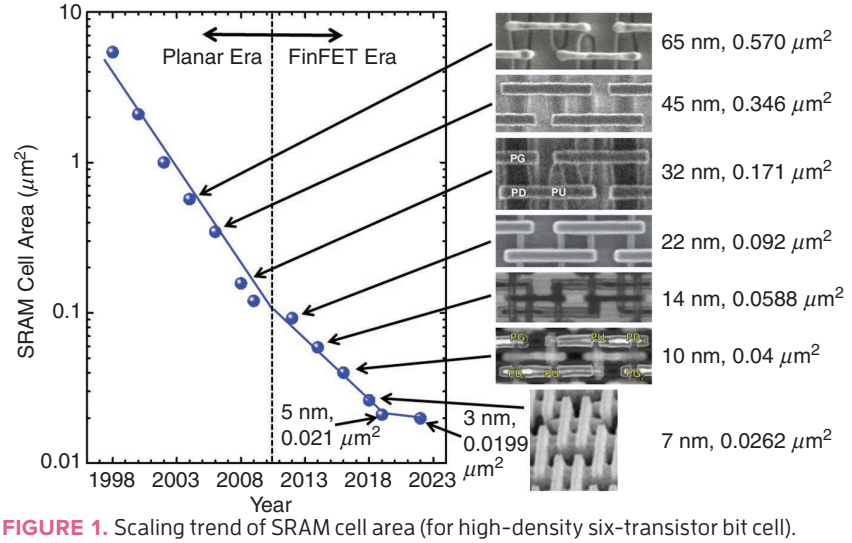

FIGURE 1. Scaling trend of SRAM cell area (for high-density six-transistor bit cell). Adapted from Yu[1] with recent years' data added.

mass production. The future scaling challenges to DRAM include maintaining the sense margin (that is, increasing storage node capacitance by high-aspect ratio stacked capacitors, or reducing the bitline parasitic capacitance), and maintaining the data retention (that is, mitigating the capacitive coupling-induced bit errors such as the row-hammer effect). The possible technological innovations to the next-generation DRAM include employing new channel materials of the cell transistor (for example, amorphous oxide semiconductors) that has intrinsically lower leakage, hiding the peripheral circuits underneath the cell array, or exploiting monolithic 3D stacked DRAM (for example, laying down the DRAM capacitors horizontally or exploiting other mechanisms such as floating-body or avalanche effects for enabling capacitorless bit cells.[5]

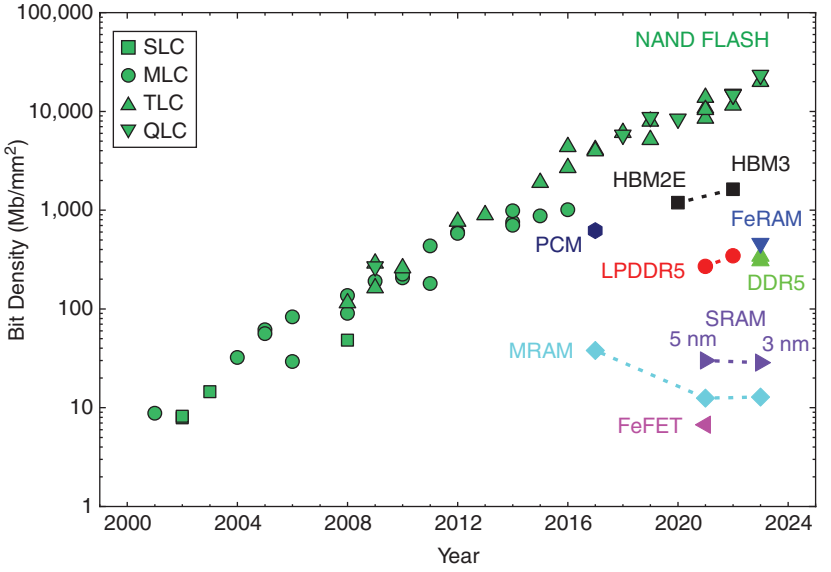

FIGURE 2. Scaling trend of memory bit density for various technologies. Adapted from Yu[1] with recent years' data added. MRAM: magnetic random-access memory.

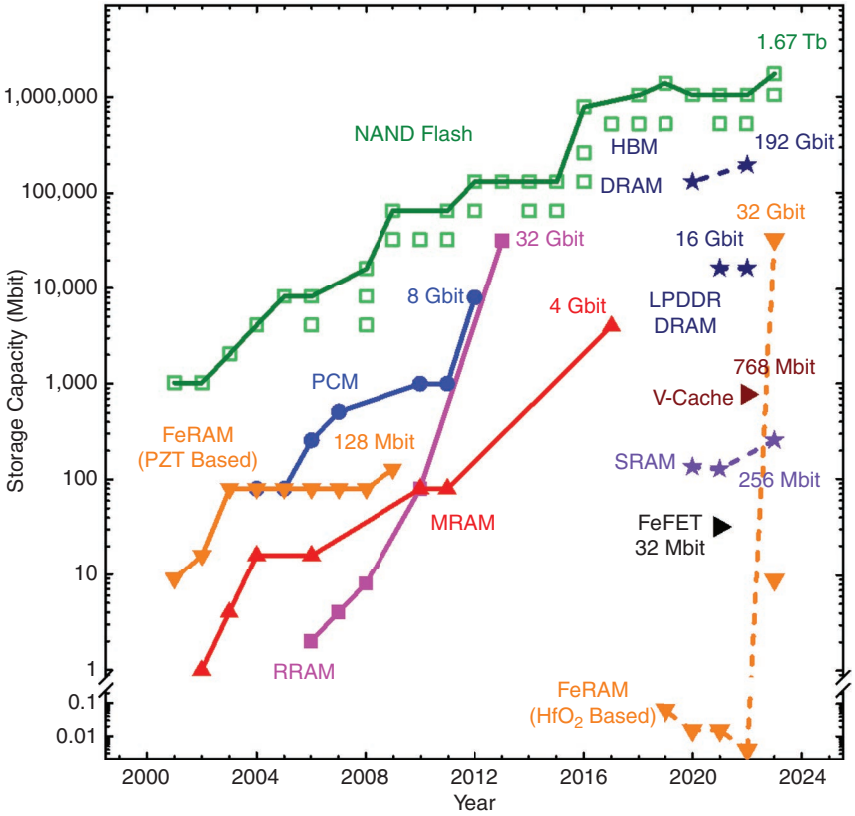

FIGURE 3. Scaling trend of memory capacity for various technologies. Adapted from Yu[1] with recent years' data added. RRAM: resistive random-access memory; FeRAM: ferroelectric random-access memory.

NAND FLASH

3D NAND dominates the Flash memory applications in solid-state drives and other mainstream storage media. Today's Flash primarily utilizes the nitride-based charge trap layer in the gate stack as the storage mechanism. 2023 marks the 10th year that the industry transitioned from the 2D NAND architecture to the 3D NAND architecture that takes advantages of the vertical channel in a cost-effective integration solution. State-of-the-art 3D NAND (as of 2023) reaches more than 300 layers and a bit density over 20 Gbit/mm^2.[2,6] The enabling technologies for realizing such high integration density include the triple-level cell (TLC) or quadruple-level cell (QLC), CMOS under array (CuA), multideck stacking (splitting the vertical channel formation into multiple steps), and so on. The future scaling challenges for 3D NAND include the diminishing sensing current along a very tall vertical channel, the degraded reliability for TLC/QLC operations, and the associated fabrication process difficulties (for example, deep trench etch) toward 1000 layers. The feature directions include possible replacement of the poly-silicon channel materials with higher mobility amorphous oxide semiconductors and possible replacement of the charge-trap layer with a ferroelectric layer in the gate stack for lower program voltage, faster program speed, and improved cycling endurance.[7]

EMERGING MEMORIES

Emerging memories have been extensively explored in the past decade with the hope of supplementing the mainstream technologies (SRAM, DRAM, and NAND Flash) as aforementioned. The tangible applications of emerging memories are mainly serving as embedded nonvolatile memories for the global buffer in microprocessors/accelerators or code storage as in microcontrollers. It is understood that emerging memories are facing difficulties directly competing against the high-density DRAM/NAND products in the standalone memory space. As of 2023, emerging memories are available from foundry platforms at mature legacy nodes. For instance, TSMC is offering resistive random-access memory (RRAM) at 40-nm/28-nm/22-nm nodes.[8] TSMC is also offering spin-transfer torque magnetic random-access memory (STT-MRAM) at 22-nm/16-nm nodes.[9] STMicroelectronics is offering phase change memory (PCM) at 28-nm node,[10] and GlobalFoundries is offering FeFET at 28-nm/22-nm nodes.[11] Sony and Micron are developing FeRAM based on HfO_2 material, and the prototype chip density increased from 64 kb[12] to 32 Gb[13] recently. The general characteristics of these emerging memories include sub-100-ns write/read speed, $>10^6$ endurance cycles, and more years of retention, while the MRAM has a unique advantage of low write voltage (<1 V), and FeFET has a unique advantage of low write energy (<10 fJ/bit). Emerging memories are attractive for certain niche markets, for example, automotive and aerospace electronics where stringent requirements exist on high/low temperature performance or immunity to radiation effects. The challenges for expanding the application space include further lowering the write voltage and making the technologies compatible with more advanced logic processes such as 7 nm or beyond, further improving the endurance and retention and supporting the reliable multilevel operation. The research community is also actively exploring using the emerging memories in the new compute paradigm such as in-memory computing or in-memory search to accelerate the data-intensive workloads such as AI/ML and combinatorial optimization.

The mainstream memory technologies such as SRAM, DRAM, and NAND Flash have benefited from the technology scaling in the past decades, and the roadmap for continued scaling (with a transition to 3D or even more 3D layers) is defined by the industry. So far, the replacement for these mainstream memory technologies remains elusive. In simple words, SRAM's foreseeable future is better SRAM, DRAM's foreseeable future is better DRAM, and NAND Flash's foreseeable future is better NAND Flash. This is because no other known memory technologies could offer the fast access speed of SRAM while not suffering from endurance degradation. Nor could those technologies provide the high density (thus ultralow cost per bit) of NAND Flash or have a balance between the cost per bit and the access speed/endurance of DRAM. Take the Intel/Micron's 3D XPoint technology[14] (that is based on PCM) as an example of a technology that had its production halted. The business model indicated a high barrier for emerging technology to serve as storage class memory due to competition with high-end NAND Flash based on single-level cell operation which offers relatively fast access speed down to approximately 1 µs. Micron's latest 32-Gbit FeRAM prototype[15] is another example in that it shows superior characteristics that almost meet DRAM specifications while providing certain nonvolatility; however, from the cost perspective, it is still quite challenging for such an emerging technology to gain advantages over mass-produced existing technologies. Therefore, the current role of emerging technologies is to augment mainstream technologies rather than to replace them. New functionalities that are offered by emerging devices and architectures such as in-memory computing or in-memory search will continue to drive further development of these technologies.

REFERENCES

1. S. Yu, *Semiconductor Memory Devices and Circuits*. Boca Raton, FL, USA: CRC Press, 2022.
2. C.-H. Chang et al., "Critical process features enabling aggressive contacted gate pitch scaling for 3nm CMOS technology and beyond," in *Proc. Int. Electron Devices Meeting (IEDM)*, 2022, pp. 27.1.1–27.1.4, doi: 10.1109/IEDM45625.2022.10019565.
3. J. Wuu et al., "3D V-Cache™: The implementation of a hybrid-bonded 64MB stacked cache for a 7nm x86-64 CPU," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2022, pp. 428–429, doi: 10.1109/ISSCC42614.2022.9731565.
4. W. Kim et al., "A 1.1V 16Gb DDR5 DRAM with probabilistic-aggressor tracking, refresh-management functionality, per-row hammer tracking, a multi-step precharge, and core-bias modulation for security and reliability enhancement," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 1–3, doi: 10.1109/ISSCC42615.2023.10067805.
5. W.-C. Chen et al., "A 3D stackable DRAM: Capacitor-less three-word-line gate-controlled thyristor (GCT) RAM with >40µA current sensing window, $>10^{10}$ endurance, and 3-second retention at room temperature," in *Proc. Int. Electron Devices Meeting (IEDM)*, 2022, pp. 26.3.1–26.3.4, doi: 10.1109/IEDM45625.2022.10019464.
6. B. Kim et al., "A high-performance 1Tb 3b/cell 3D-NAND flash with a 194MB/s write throughput on over 300 layers i," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 27–29, doi: 10.1109/ISSCC42615.2023.10067666.
7. S. Yoon et al., "QLC programmable 3D ferroelectric NAND Flash memory by memory window expansion using cell stack

engineering," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2023, pp. 1–2, doi: 10.23919/VLSITechnologyandCir57934.2023.10185294.
8. C.-X. Xue et al., "A 22nm 4Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7 TOPS/W for tiny AI edge devices," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2021, pp. 245–247, doi: 10.1109/ISSCC42613.2021.9365769.
9. P.-H. Lee et al., "A 16nm 32Mb embedded STT-MRAM with a 6ns read-access time, a 1M-cycle write endurance, 20-year retention at 150°C and MTJ-OTP solutions for magnetic immunity," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 494–496, doi: 10.1109/ISSCC42615.2023.10067837.
10. F. Arnaud et al., "High density embedded PCM cell in 28nm FDSOI technology for automotive micro-controller applications," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2020, pp. 24.2.1–24.2.4, doi: 10.1109/IEDM13553.2020.9371934.
11. S. Muller et al., "Development status of gate-first FeFET technology," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2021, pp. 1–2.
12. J. Okuno et al., "SoC compatible 1T1C FeRAM memory array based on ferroelectric $Hf_{0.5}Zr_{0.5}O_2$," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265063.
13. N. Ramaswamy et al., "NVDRAM: A 32Gbit dual layer 3D stacked non-volatile ferroelectric memory with near-DRAM performance for demanding AI workloads," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2023, pp. 1–4.
14. R. Smith. "Intel to wind down Optane memory business—3D XPoint storage tech reaches its end." AnandTech. Accessed: Jul. 28, 2022. [Online]. Available: https://www.anandtech.com/show/17515/intel-to-wind-down-optane-memory-business
15. C. Mellor. "Micron NVDRAM may never become a product." Blocks and Files. Accessed: Jan. 9, 2024. [Online]. Available: https://blocksandfiles.com/2024/01/09/micron-nvdram-might-never-become-a-product/

SHIMENG YU is a professor at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. Contact him at shimeng.yu@ece.gatech.edu.

TAE-HYEON KIM is a postdoctoral fellow at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. Contact him at thkim@gatech.edu.

Digital Object Identifier 10.1109/MC.2024.3371851