Semiconductor Technology from A to Z

Digital technology

Why New Memory Technologies?

The Limits of Flash

Flash memory (see the flash memory cell chapter) has dominated the non-volatile memory market for decades, but is increasingly running into technical limits. As the floating-gate cell continues to scale down, the number of stored electrons per cell keeps shrinking – at current feature sizes, sometimes only a few dozen electrons decide between “0” and “1”. This makes the cell more susceptible to interference from neighboring cells (cell-to-cell interference via parasitic capacitance), to read disturb, and to charge loss over time, since losing even a single electron can already distort the read result. In multi-level-cell and triple-level-cell schemes (see the flash memory cell chapter), where several bits are encoded via fine gradations of the threshold voltage within one cell, this problem is further aggravated, since the voltage windows between the individual charge states keep narrowing.

Added to this is limited endurance: every write/erase cycle drives electrons through the cell's thin tunnel oxide (Fowler-Nordheim tunneling or hot-electron injection), which gradually fills the oxide with traps and stresses it mechanically. Depending on the cell type, NAND flash therefore survives only a few hundred cycles (QLC, four bits per cell) up to around 100,000 cycles (SLC, one bit per cell) before the cell's error rate can no longer be compensated by error-correction coding (ECC). Write speed is also fundamentally limited: unlike byte-wise writing, erasing in NAND flash happens block-wise (typically several hundred kilobytes at once) and, at several milliseconds, takes considerably longer than the actual write operation of a page.

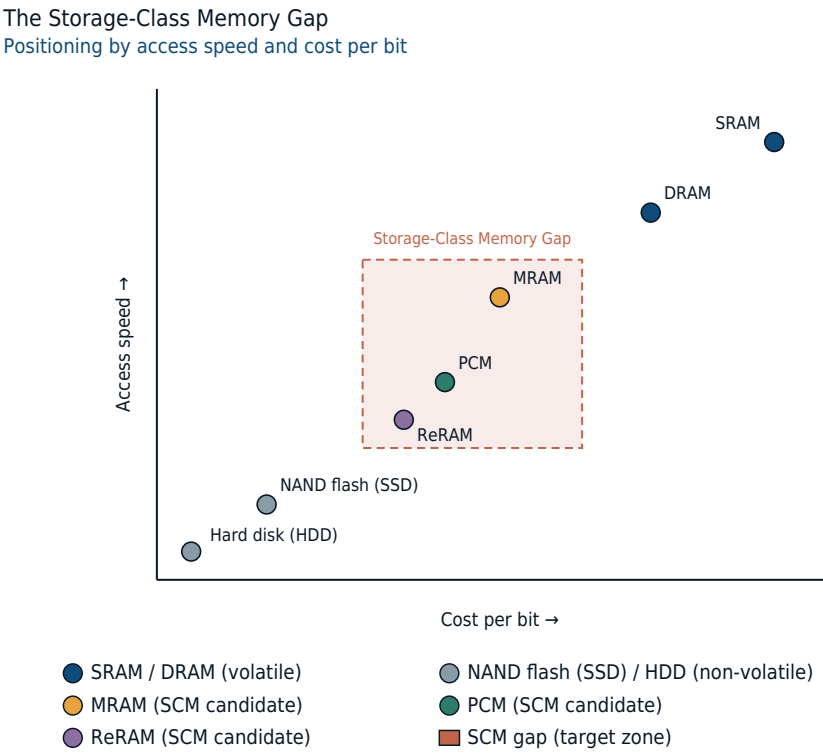
The Gap Between Working Memory and Mass Storage

Classical memory architectures have a sharp break in the memory hierarchy. DRAM (see the DRAM cell chapter) is extremely fast – access times are in the range of a few nanoseconds – but volatile (the content is lost without a refresh cycle and ultimately without power) and comes with comparatively high cost per stored bit, since each cell needs its own capacitor and access transistor. Flash-based SSDs, by contrast, are non-volatile and considerably cheaper per bit, but several orders of magnitude slower (microseconds to milliseconds instead of nanoseconds) and limited in the number of possible write cycles.

In between lies a gap referred to in the literature as storage-class memory (SCM): a memory type that combines the speed of DRAM with the non-volatility and bit cost of flash. Such a memory could be operated directly on the memory

bus within the system architecture, instead of being connected – as SSDs are today – via a comparatively slow I/O controller (such as NVMe over PCIe), and would thereby render entire software layers of operating-system and application logic unnecessary, layers that today exist solely to bridge the speed gap between working memory and mass storage (caching, paging, write-back strategies).

Positioning by speed and cost per bit



The technologies covered in this chapter – MRAM, ReRAM, and PCM – each pursue fundamentally different physical storage principles to close exactly this gap: magnetic states in MRAM, resistive filaments in ReRAM, crystalline phase states in PCM. None of them has so far displaced flash as the dominant mass-storage technology, and given the enormous scale and cost advantages NAND flash has achieved through decades of process optimization and mass production, a complete replacement is unlikely even in the medium term. All three technologies, however, have already achieved commercial relevance in specialized niches: MRAM as a cache replacement and in automotive and industrial applications, PCM in the form of Intel Optane as a fast tier ahead of NAND SSDs, and ReRAM mainly in embedded applications with low power requirements.