Performance consideration

Exponential increases in data and demand for improved performance to process that data has spawned a variety of new approaches to processor design and packaging, but it also is driving big changes on the memory side.

While the underlying technology still looks very familiar, the real shift is in the way those memories are connected to processing elements and various components within a system. That can have a big impact on system performance, power consumption, and even the overall resource utilization.

Many different types of memories have emerged over the years, most with a well-defined purpose despite some crossovers and unique use cases. Among them are DRAM and SRAM, flash, and other specialty memories. DRAM and SRAM are volatile memories, meaning they require power to maintain data. Non-volatile memories do not require power to retain data, but the number of read/write operations is limited, and they do wear out over time.

All of these fit into the so-called memory hierarchy, starting with SRAM — a very fast memory that typically is used for various levels of cache. SRAM is extremely fast, but its applications are limited due to the high cost per bit. Also at the lowest level, and often embedded into an SoC or attached to a PCB, NOR flash typically used for booting up devices. It's optimized for random access so it does not have to follow any particular sequence for storage locations.

Moving up a step in the memory hierarchy, DRAM is by far the most popular option, in part because of its capacity and resilience, and in part because of its low cost per bit. That is partially due to the fact that the leading DRAM vendors have fully depreciated their fabs and equipment, but as new types of DRAM come online, the price has been rising, opening the door to new competitors.

There has been talk of replacing DRAM for decades, but DRAM has proved to be much more resilient from a market standpoint than anyone would have anticipated. In 3D configurations of high-bandwidth memory (HBM), it has proven to be an extremely fast, low-power option, as well.

JEDEC defines four main types of DRAM:

- Double data rate (DDRx) for standard memory;
- Low-power DDR (LPDDRx), primary used in mobile or battery-operated devices;
- Graphics DDR (GDDRx), which initially was designed for high-speed graphic applications, but which also is used for other applications, as well, and
- High-bandwidth memories (HBMx), which primary for high-performance applications such as AI or inside of data centers.

NAND flash, meanwhile, is typically used as removable storage (SSD/USB stick). Due to longer erase/write cycle and lower lifespan, flash is not suitable for CPU/GPU and system applications.