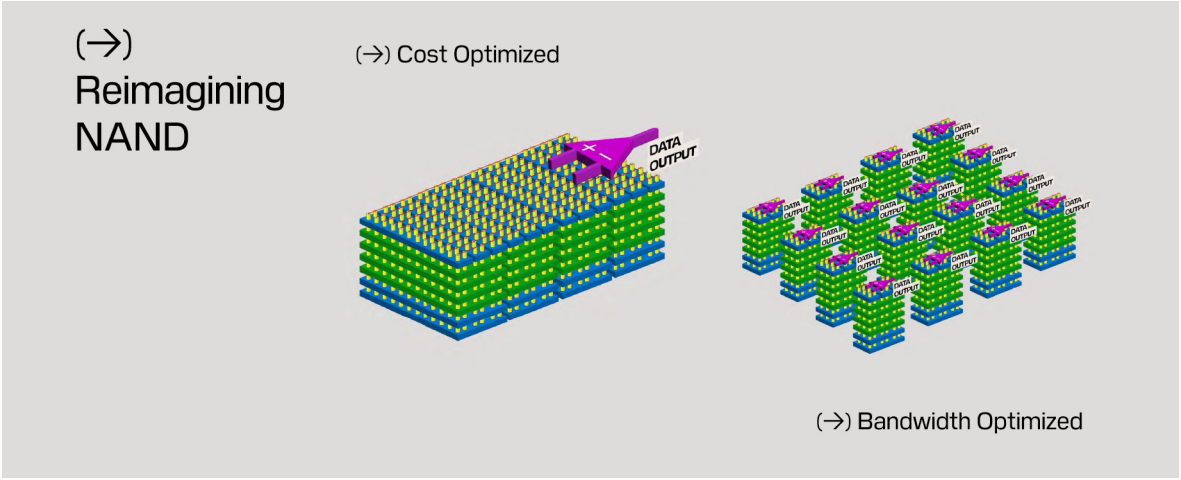
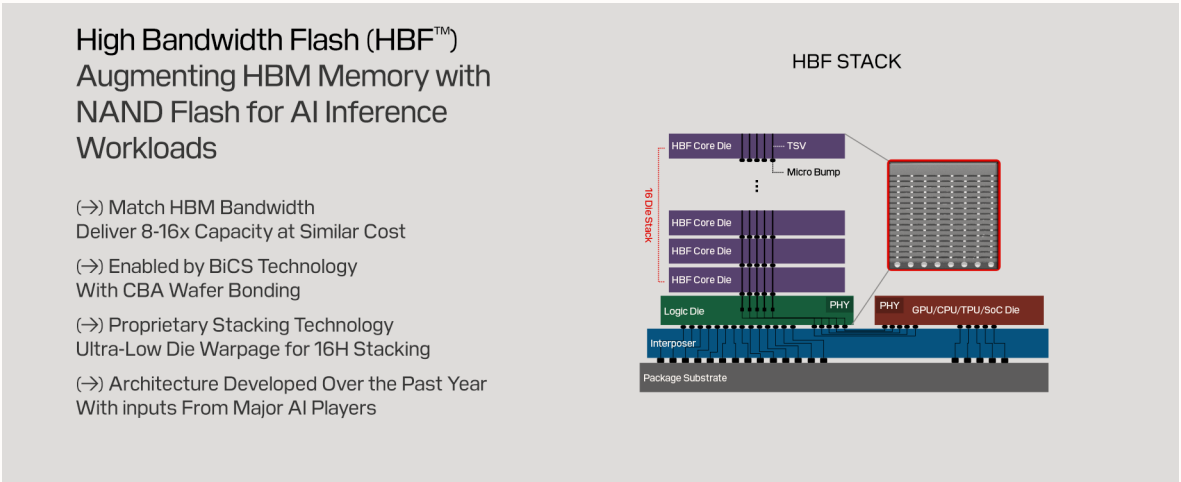
(→) SANDISK UNVEILS THE FUTURE OF MEMORY ARCHITECTURE FOR AI INTRODUCING: HIGH BANDWIDTH FLASH

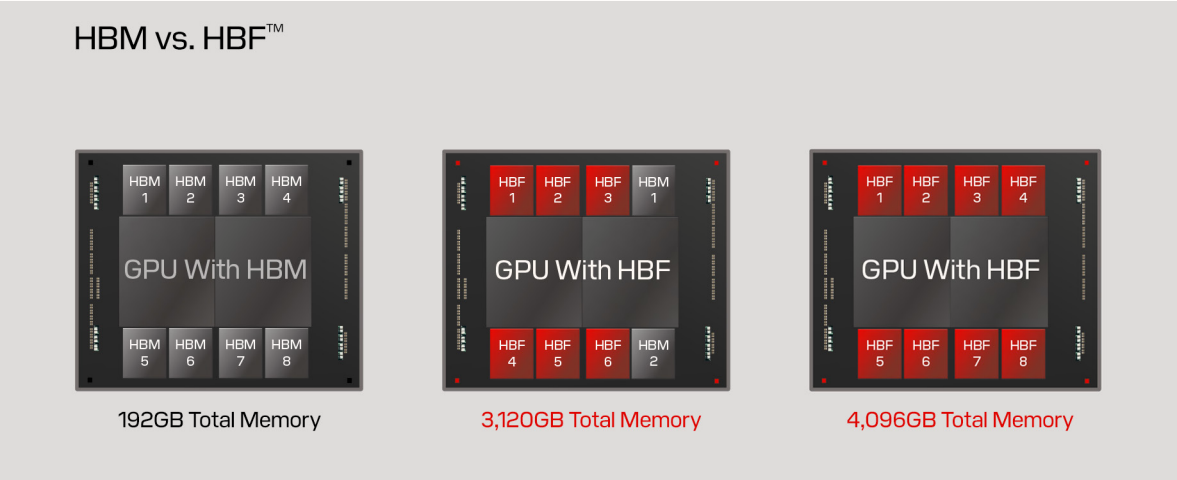
AI is pushing today's memory technologies past their limits. As AI models grow larger and more complex, inference workloads demand both massive bandwidth and significantly greater memory capacity, something traditional High-Bandwidth Memory (HBM) cannot economically or physically deliver at scale. HBM offers high speeds but is constrained by limited capacity, high power consumption, and steep costs, creating a bottleneck for developers and hyperscalers seeking to scale AI infrastructure efficiently.



To help address this critical gap, Sandisk developed a revolutionary new memory technology: High Bandwidth Flash (HBF™). HBF is built from the ground up for AI inferencing. With up to 8-16x the capacity of HBM, HBF offers industry-leading capacity at similar bandwidth and similar cost.



Based on NAND memory technology, HBF delivers performance within 2.2% of unlimited-capacity HBM[1], while offering a significant increase in memory capacity—unlocking AI services that traditional HBM simply can't support. Powered by Sandisk's CBA (CMOS directly Bonded to Array) technology, HBF sets a new standard for high-bandwidth memory, combining exceptional speed, density, and energy efficiency in a single architecture.



The first-generation product delivers blazing-fast read bandwidth of 1.6 TB/s and packs 256Gb per die, reaching 512GB[2] total capacity per 16 die stack while closely matching the physical footprint, power profile, and stack height of HBM4. HBF is purpose-built to meet the scale and intensity of next-generation AI inference workloads, offering the performance needed without compromising on capacity or efficiency.

Key highlights of HBF include:

- Our most scalable semiconductor technology in high volume manufacturing: Unlike memory solutions like DRAM, which face significant challenges on future scalability in capacity, HBF is built on Sandisk's BiCS technology, offering a clear path to greater capacity and bandwidth, as demands grow.
- High bandwidth, low cost: HBF technology offers one of the lowest cost per bit of any memory technology.
- CBA (CMOS directly Bonded to Array): Sandisk's CBA technology is a breakthrough that enables ultra-high-density, high-speed, and low-power circuits—setting a new benchmark for high-bandwidth memory performance.
- Advanced die stacking technology: HBF uses patented technology to help reduce die warpage/ stress and improve thermal conductivity, enabling 16 die stacking to deliver the targeted capacity and bandwidth.
- Non-volatile, no refresh power needed: Leveraging NAND technology, HBF retains its data even when a power source is removed and it requires no refresh power to prevent data leakage or loss.
- Thermal stability under high duress: HBF maintains thermal stability even at high temperatures expected in Data Center environments, enabling it to withstand greater operational stress without performance or reliability degradation.
- Advanced reliability and redundancy management techniques: HBF technology is designed to meet demanding requirements for concerns such as endurance and high temperature of operations in data centers and function reliably even the presence of potential failures.

HBF capabilities will only continue to improve with future generations of CBA NAND, pushing performance even further. Gen 2 and Gen 3 are expected to deliver preliminary read bandwidths exceeding 2 TB/s and 3.2 TB/s, respectively, while stack capacities are set to reach up to 1 TB and 1.5 TB[3]. Both are also projected to achieve greater energy efficiency, targeting 0.8x and 0.64x the power consumption of Gen 1.

High-Bandwidth Flash redefines what's possible in AI memory by building from Sandisk's low-cost, high-density CBA based NAND core technology, with architectural improvements in silicon technology, design technology and 3D stack packaging, enabling high bandwidth, high endurance, and energy efficiency. Purpose-built for the demands of AI inference, HBF delivers the speed, scale, and efficiency needed to fuel the next generation of intelligent systems—helping to solve one of AI's most critical infrastructure challenges.

[1]Based on internal testing and simulation.This is simulated for reading 8-bit pretrained weights on Llama 3.1 405B parameter model. One kernel is executed on the xPU performance model at a time. Actual results may vary depending on specific circumstances and contextual factors. The comparison does not reflect the capacity advantage of HBF that can fit the full model as HBM capacity is assumed to be unlimited for modeling purposes. This demonstrates that with the higher latency and larger page size of HBF compared to HBM, the system level performance is still comparable and HBF can work for AI inference workloads.

[2]1GB = 1,000,000,000 bytes and 1TB = 1,000,000,000,000 bytes. Actual user capacity less.

[3]Based on internal testing and simulation.