Goldman Sachs Research

Equity Research

12 October 2026 | 5:00AM EDT

GLOBAL TECHNOLOGY: SEMICONDUCTORS - MEMORY

# High-performance NAND could help crumble the AI "memory wall" while driving meaningful CapEx savings

James Schneider, Ph.D.
+1(212)357-2929 | jim.schneider@gs.com
Goldman Sachs & Co. LLC

Khalil Fenina
+1(212)357-6392 | khalil.fenina@gs.com
Goldman Sachs & Co. LLC

Giuni Lee
+65-6889-1025 | giuni.lee@gs.com
Goldman Sachs (Singapore) Pte

Shuhei Nakamura
+81(3)4587-9932 | shuhei.nakamura@gs.com
Goldman Sachs Japan Co., Ltd.

Allen Chang
+852-2978-2930 | allen.k.chang@gs.com
Goldman Sachs (Asia) L.L.C.

Daiki Takayama
+81(3)4587-9870 | daiki.takayama@gs.com
Goldman Sachs Japan Co., Ltd.

Marshall Wong
+1(646)446-0222 | marshall.wong@gs.com
Goldman Sachs & Co. LLC

Anmol Makkar
+1(212)357-1366 | anmol.makkar@gs.com
Goldman Sachs & Co. LLC

Luya You
+1(212)902-5297 | luya.you@gs.com
Goldman Sachs & Co. LLC

Liana Wadhwani
+1(212)357-3204 | liana.wadhwani@gs.com
Goldman Sachs & Co. LLC

**Key stock takeaways: We view emerging high-performance NAND-based solutions - including High-Bandwidth Flash (HBF) - as promising tools to address a worsening memory constraint in AI inference. We see a potential TAM of ~$40 - $150B at a GPU adoption rate of 15%-50% by 2030 given the rapid proliferation of read-heavy and memory-intensive AI inference workloads - despite potential challenges to adoption. We believe that HBF memory can meaningfully reduce the overall cost of inference throughput by increasing GPU utilization, resulting a ~30% lower overall cost of inference throughput - translating to as much as several hundreds of billions of hyperscaler CapEx savings.**

In a bull-case scenario that assumes rapid commercial adoption of HBF on 50% of XPUs by late 2028, we see HBF bits accounting for ~4% of global NAND bit shipments - but ~9% of NAND wafer capacity given its die size penalties - similar to the HBM/DRAM dynamic. We would point out that these NAND-based solutions remain in the early development stages at a system level, and that a wide range of outcomes exists that could drive a significantly greater or smaller impact to the overall NAND supply/demand market dynamic. Stepping back, we continue to believe that limited supply additions will prolong tight NAND supply/demand in the near term - and view HBF as potentially extending the duration of the current cycle. While HBF is the technology approach that has received the most attention in the technical literature, we would also highlight XL-Flash (from Kioxia) and zNAND-O (from Samsung) as alternative approaches to achieving similar objectives.

**We continue to favor SanDisk (SNDK, Buy) and SK Hynix (000660.KS) given their technology leadership in HBF, and strong core business with improving product mix. We also see Samsung Electronics (005930.KS, Buy) as well positioned given its strong NAND portfolio;** we see Micron (MU, Neutral) as less of a direct beneficiary given its primary focus on DRAM and lower exposure to NAND. Although we recognize challenges to widespread HBF adoption, we see promise in the technology which could drive meaningfully higher cost/performance at the system level, particularly in memory-intensive inference workloads.

Goldman Sachs does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision. For Reg AC certification and other important disclosures, see the Disclosure Appendix, or go to www.gs.com/research/hedge.html. Analysts employed by non-US affiliates are not registered/qualified as research analysts with FINRA in the U.S.

## Executive Summary

On balance, we are constructive on HBF technology, and it as a promising potential solution to the intensifying memory constraint in AI inference. Our bullish view on the technology's potential is based on the following:

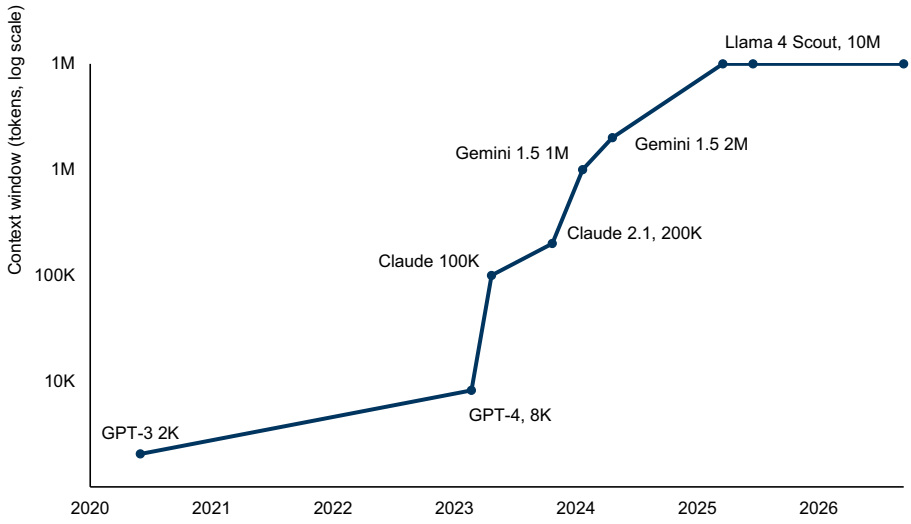
1. Proliferation of AI inference workloads will lead to structurally higher demand for high-bandwidth, near-compute storage as context windows lengthen.
2. Highly cost-competitive near-compute memory capacity at comparable sequential read bandwidth to HBM could result in attractive system-level economics and accelerate adoption of HBF among hyperscalers.
3. NAND's high density of NAND, coupled with HBF's higher resulting GPU utilization, will allow the technology to command a meaningful price premium while simultaneously lowering system-level costs for hyperscalers and AI labs.

However, key challenges to adoption remain, namely latency concerns (particularly for random read / write operations), endurance and thermal considerations, and manufacturing resource intensity. Despite these challenges, we see a significant commercial opportunity in HBF and view the economics of its adoption as attractive - even at a significant price premium relative to traditional NAND.

## Background - Why is there a need for High-Bandwidth Flash?

**The shift to inference requires much near-compute storage capacity, and HBF's similar sequential read bandwidth vs. HBM positions it well as a potential solution to the memory constraint given NAND's high density.** As more AI workflows shift away from compute-heavy training and toward memory-intensive inference, context windows are lengthening exponentially and scaling toward 10 mn+ tokens from a ~1.5 mn baseline (see Exhibit 1). The KV (key value) cache required for these more memory-intensive workloads is growing with larger context windows, broadly shifting the memory bottleneck from purely bandwidth-related to a combined bandwidth-capacity constraint. Although we believe DRAM-based High Bandwidth Memory (HBM) - and its subsequent iterations - is likely to lead in bandwidth and performance, the technology is significantly capacity-constrained (current architectures typically include ~192 GB per GPU) given current DRAM supply constraints. AI infrastructure operators are thus deploying additional GPU units to drive incremental compute-adjacent memory capacity, resulting in meaningfully underutilized compute capacity. Through a strategic collaboration, SanDisk and SK Hynix have introduced High Bandwidth Flash (HBF) as a proposed solution to this near-compute capacity constraint.

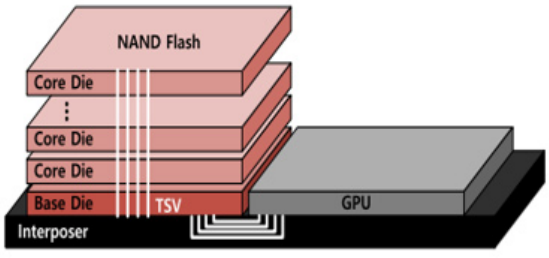
**Exhibit 1: Chart showing context window growth over time for common LLMs**




Source: Hidekazu-Konishi, data compiled by Goldman Sachs Global Investment Research

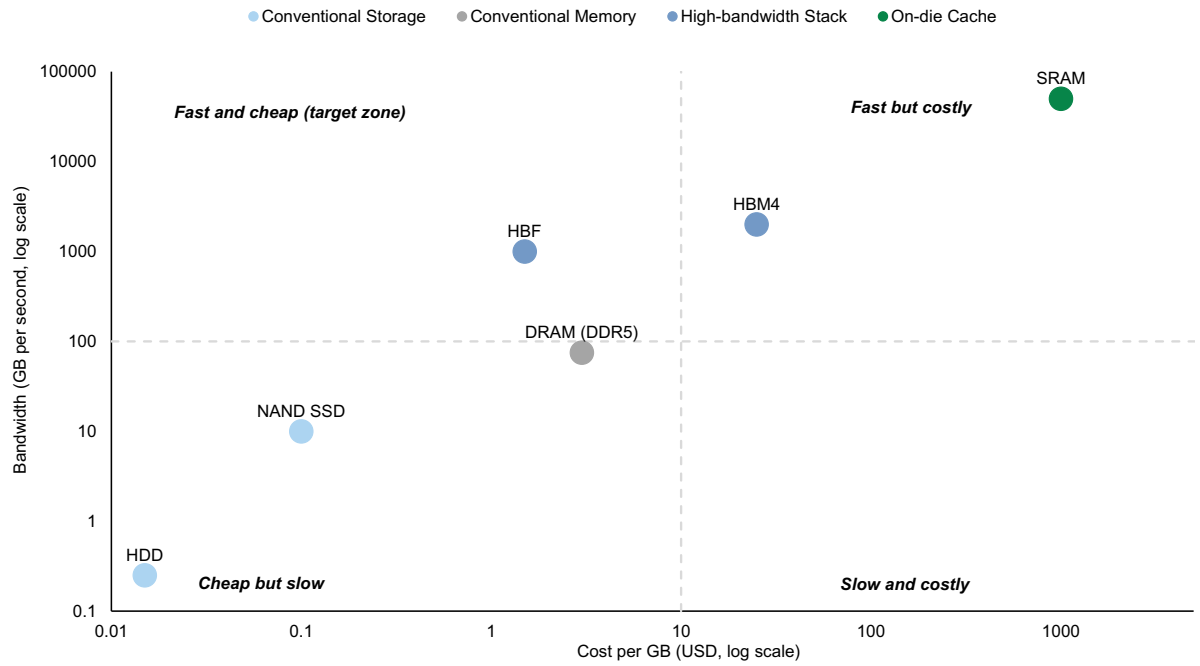
**We expect HBF pricing to be significantly below DRAM-based HBM4 contract pricing (on a per GB basis) once commercial samples begin shipping in CY27 (see Exhibit 3), despite the product's pricing currently being undisclosed.** High Bandwidth Flash (HBF) has been introduced an open-standard, near-compute memory tier that stacks high-density 3D NAND flash into the AI accelerator package via through-silicon vias (TSVs) and UCIe chiplet interconnects (see Exhibit 2). It establishes an intermediate tier in the memory hierarchy, offering a bandwidth leap over standard NAND SSDs and offers near-HBM bandwidth on a sequential read basis. Governed under the Open Compute Project (OCP) which includes SanDisk and SK Hynix, as well as ASIC customer hyperscalers (namely Alphabet and Meta), HBF delivers up to 512 GB per stack (vs. 16-32 GB per HBM stack) at sequential read bandwidths up to ~3.0 TB/s (see Exhibit 5), per the technical literature.

**Exhibit 2: Cross-sectional architecture of High Bandwidth Flash (HBF)**




Source: SK Hynix

**Exhibit 3: Memory hierarchy Diagram**




Source: Goldman Sachs Global Investment Research

## Advantages of HBF - Cost effectiveness and other benefits

**We believe HBF can lower the overall cost of equivalent inference token output by ~30% or more, and lead to significant reductions in hyperscaler capital expenditures through combined hybrid architectures with HBM and increased GPU utilization.** Given NAND's thermal and write endurance constraints, we believe HBF stacks are likely to utilize SLC-based (single level cell) NAND to mitigate these risks - and given the higher costs and die size penalties will be priced far above MLC - similar to the relationship between HBM and conventional DRAM. HBF's interconnect specification will likely also result in significantly lower output per wafer (given the larger physical size of an HBF chip relative to an equivalent capacity standard NAND chip; we assume HBF requires ~3.5X more die/wafer area vs. traditional NAND). Even at a meaningful premium, we see significant room for HBF to significantly lower inference output costs, even at ~1mn token context length - with the benefit growing roughly linearly as sequences scale higher (see Exhibit 4).

**Exhibit 4: Illustrative per 1mn token cost comparison HBM only vs. H3 GPU memory architecture and HBF price per GB scenario analysis**

| On-Chip Memory Configuration | HBM Only | HBM-HBF Hybrid |
|---|---|---|
| GPU Model | Nvidia B200 | Nvidia B200 |
| HBM Capacity (GB) per GPU | 192 | 192 |
| HBF Capacity (GB) per GPU | 0 | 4,096 |
| | | |
| **Number of GPUs** | **8** | **4** |
| Total System HBM Capacity | 1,536 | 768 |
| Total System HBF Capacity | 0 | 16,384 |
| | | |
| HBM Cost per GB ($) | 15.00 | 15.00 |
| HBF Cost per GB ($) | 1.75 | 1.75 |
| | | |
| HBM Cost per GPU ($) | 2,880 | 2,880 |
| HBF Cost per GPU ($) | 0 | 7,168 |
| GPU Cost (incl. packaging, peripherals, but excl. memory) | 7,620 | 7,620 |
| | | |
| **Aggregate System Cost (GPUs + Memory)** | 84,000 | 70,672 |
| | | |
| **Aggregate System Price (GPUs + Memory) (70% GM)** | **280,000** | **235,573** |
| | | |
| **Illustrative Inference Throughput in TPS at a 1mn token batch size (approximate)** | 10,000 | 12,500 |
| | | |
| **Cost per 1K inference output tokens per second** | **$28.00** | **$18.85** |
| Cost Efficiency Gain | | -32.69% |

| Potential Cost Efficiency Gain for H3 at 4 GPUs | |
|---|---|
| HBF price per GB | |
| $ 0.25 | -56.10% |
| $ 0.50 | -52.20% |
| $ 0.75 | -48.30% |
| $ 1.00 | -44.40% |
| $ 1.25 | -40.50% |
| $ 1.50 | -36.59% |
| $ 1.75 | -32.69% |
| $ 2.00 | -28.79% |
| $ 2.25 | -24.89% |
| $ 2.50 | -20.99% |

Source: Company data, TrendForce, Goldman Sachs Global Investment Research

**HBF could achieve meaningful cost advantages over HBM-only systems with higher memory density, comparable sequential read bandwidth, and higher GPU utilization.** We highlight key technical points including:

- **Higher memory density:** At ~512 GB per stack and ~4 TB per GPU (in SK Hynix's proposed H3 architecture), HBF offers ~16X higher memory capacity vs. HBM. This density allows systems to process 2.6X larger batch sizes at 1M tokens and up to 18.8X larger batches at 10M tokens compared to pure-HBM setups - resulting in serving throughput advantages widen alongside sequence length. This advantage is estimated at ~1.25X at 1 mn tokens and ~6.14X at 10 mn tokens, according to research published by SK Hynix.
- **Mitigating the bandwidth gap between DRAM and NAND:** Although NAND lags DRAM in peak bandwidth, research from SK Hynix and SanDisk has shown that sequential read bandwidth for HBF significantly narrows the gap with HBM and achieves up to ~3 TB/s (see Exhibit 5). Also, in most inference-heavy workloads, the number of write operations is quite limited, reducing the significance of random read or write bandwidth of memory.
- **~2.7X energy efficiency gain at scale:** Although an individual HBF stack consumes 2X to 4X more power in absolute terms (80-160W vs. 40W), its ability to process significantly larger inference batches may yield up to ~2.7X higher throughput-per-watt, with efficiency gains widening as context lengths scale, according to research supported by SK Hynix.
- **Potential for XPU cost savings via higher utilization:** By eliminating memory-capacity bottlenecks and maximizing GPU core utilization, 4 HBF-equipped XPU accelerators can potentially match the aggregate serving throughput of an 8-XPU pure-HBM system (at 192 GB / 4 TB of HBM / HBF respectively per GPU,

according to SanDisk).

**Exhibit 5: Comparison between HBF and HBM 4**

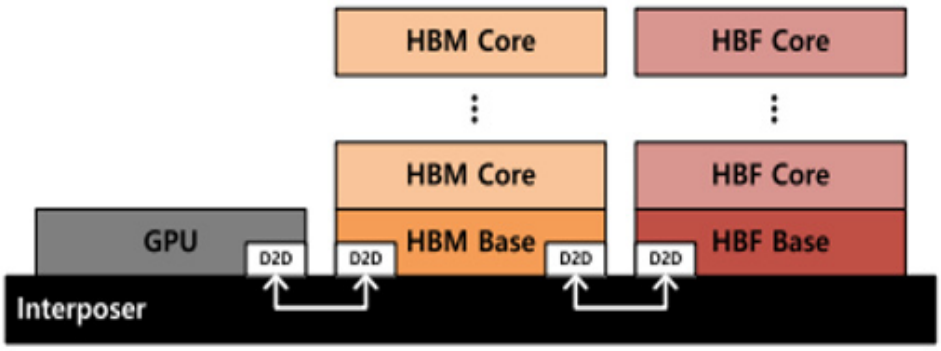
| Comparison between HBF and HBM 4 | | |
|---|---|---|
| Category | HBM 4 | HBF |
| **Capacity (GB)** | 36 | 512 |
| **Price ($)** | ~15.28 | N/A |
| **Power Consumption (Watt)** | HBF is ~2-4x HBM | |
| **Throughput per power** | HBF 1.25x higher for 1M tokens; 6.14x higher for 10M * | |
| **Sequential Read Bandwidth** | ~2.0-3.3 TB/s | ~0.4-3 TB/s |
| **Latency** | ~20ns | ~10 µs (10000ns) |
| **Read granularity (bytes per read)** | 32 | 4096 |
| **Write Endurance** | High | Low |
| **Excels in** | Write dominant workload | Read dominant workload |

Source: Silicon analyst, Hotchips conference, Goldman Sachs Global Investment Research

**HBF's benefits are enhanced in hybrid architectures that help mitigate the technology's key drawbacks.** Since NAND flash access latency may prove sub-optimal for real-time token generation in certain situations, SK Hynix has proposed its own "H³" architecture, which consists of a hybrid HBM-HBF configuration. The company claims this helps address NAND's write / random read bandwidth shortcomings by unifying both memory types into a complementary, two-tier memory hierarchy. We note the combined architecture could potentially offer:

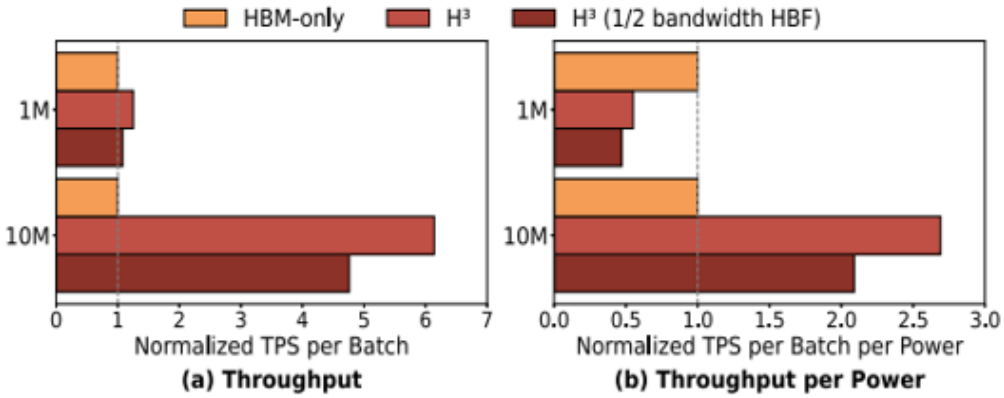
- **Zero shoreline penalty:** HBF stacks are daisy-chained directly through the HBM base die, expanding on-package capacity without occupying the GPU's limited physical edge connections (see Exhibit 6)
- **Optimized workload partitioning:** Low-cost HBF stores massive, static data structures (model weights and pre-computed KV caches), reserving valuable HBM capacity exclusively for latency-critical live decoding.
- **Commercial and operational payoff:** Offloading static memory expands concurrent batch capacity, enabling up to 2.7x higher throughput per watt (see Exhibit 7). This reduces overall GPU cluster counts and facility power draw, lowering total cost of ownership (TCO) and per-query serving costs.

**Exhibit 6: Side cross sectional view of H3**




Source: SK Hynix

**Exhibit 7: Throughput & throughput per power comparison of HBM-only vs H3**



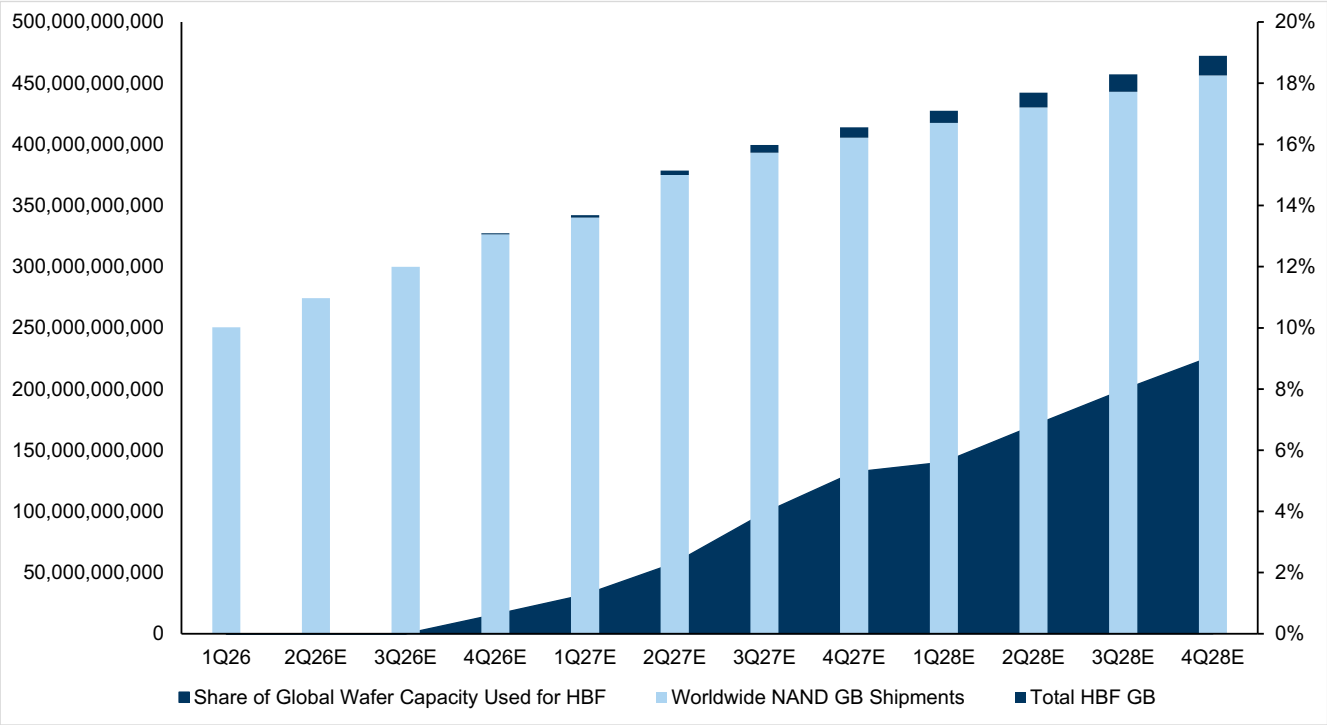

Source: SK Hynix

## HBF forecast model and potential commercial opportunity

**We see a significant potential commercial opportunity in HBF and believe it could account for up to ~4% of global NAND bit shipments by late 2028 in a bull case that assumes a rapid and elevated adoption rate of 50% by 4Q28, which could equate to an HBF TAM of ~$150 bn.** In a more measured scenario that assumes a penetration rate of ~15% of GPU units by late 2030 and a price per GB of HBF of $1.75 (which we view as conservative), we see an annualized TAM for HBF of ~$40 bn.

**We see HBF bits occupying up to ~9% of global NAND wafer capacity in a bull-case adoption scenario given its SLC structure and larger die size.** Relative to traditional NAND, we assume a 3.5X higher wafer space usage per GB equivalent for HBF (roughly similar to the DRAM/HBM ratio) and thus see it accounting for a larger share of global NAND wafer capacity relative to its share of total NAND bits. Assuming a bull case scenario of a rapid ramp in HBF's deployment to 50% of GPU units by late 2028, and based on current industry forecasts for global NAND bit output and wafer capacity, we see HBF accounting for ~4% of global NAND bits while simultaneously consuming over 9% of global NAND wafer capacity (see Exhibit 8).

**Exhibit 8: HBF share of global NAND GB and global NAND wafer capacity**




Source: TrendForce, Goldman Sachs Global Investment Research

**Exhibit 9: Global HBF forecast model**

| | 1Q26 | 2Q26E | 3Q26E | 4Q26E | 1Q27E | 2Q27E | 3Q27E | 4Q27E | 1Q28E | 2Q28E | 3Q28E | 4Q28E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NVDA | 1,592 | 1,866 | 2,067 | 2,265 | 2,525 | 2,393 | 3,093 | 3,345 | 3,402 | 3,447 | 3,575 | 3,588 |
| AMD | 127 | 135 | 146 | 244 | 266 | 298 | 343 | 377 | 358 | 362 | 365 | 376 |
| AVGO | 1,254 | 1,379 | 1,614 | 1,904 | 1,828 | 2,193 | 2,522 | 2,951 | 2,951 | 3,305 | 3,438 | 3,541 |
| MRVL | 110 | 118 | 148 | 195 | 152 | 178 | 208 | 239 | 206 | 237 | 278 | 333 |
| Total (k Units) | 3,082 | 3,499 | 3,975 | 4,608 | 4,771 | 5,062 | 6,166 | 6,912 | 6,917 | 7,351 | 7,656 | 7,838 |
| % w/ HBF | 0.00% | 0.00% | 0.00% | 5.00% | 10.00% | 17.50% | 25.00% | 30.00% | 35.00% | 40.00% | 45.00% | 50.00% |
| k Units w/ HBF | 0 | 0 | 0 | 230 | 477 | 886 | 1,541 | 2,074 | 2,421 | 2,940 | 3,445 | 3,919 |
| HBF GB per Unit | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 | 4,096 |
| Total HBF GB | 0 | 0 | 0 | 943,769,934 | 1,954,354,278 | 3,628,252,884 | 6,313,882,638 | 8,493,715,308 | 9,916,832,398 | 12,044,213,194 | 14,111,129,075 | 16,051,941,696 |
| Global NAND GB Shipped | 250,502,593,784 | 274,274,827,172 | 299,837,605,810 | 326,345,845,374 | 340,103,507,470 | 374,978,692,941 | 393,082,944,774 | 405,412,980,472 | 417,575,369,886 | 430,102,630,983 | 443,005,709,912 | 456,295,881,210 |
| QoQ Growth (%) | | 9.32% | 8.84% | 4.22% | 10.25% | 4.83% | 3.14% | 3.00% | 3.00% | 3.00% | 3.00% | |
| % of WW NAND GB as HBF | 0.00% | 0.00% | 0.00% | 0.29% | 0.57% | 0.97% | 1.61% | 2.10% | 2.37% | 2.80% | 3.19% | 3.52% |
| Estimated SanDisk GB Shipped | 80,147,536,031 | 80,147,536,031 | 80,147,536,031 | 80,147,536,031 | 1.01133E+11 | 1.01133E+11 | 1.01133E+11 | 1.01133E+11 | 1.11246E+11 | 1.11246E+11 | 1.11246E+11 | 1.11246E+11 |
| % of SanDisk GB Shipped (illustrative, 100% market share) | 0.00% | 0.00% | 0.00% | 1.18% | 1.93% | 3.59% | 6.24% | 8.40% | 8.91% | 10.83% | 12.68% | 14.43% |
| Global Wafer Starts in NAND | 4,011,000 | 4,026,000 | 4,092,000 | 3,987,000 | 4,071,000 | 4,191,000 | 4,263,000 | 4,293,000 | 4,722,300 | 4,816,746 | 4,913,081 | 5,011,343 |
| Max Trad. NAND GB per 12in. Wafer | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 | 131,072 |
| HBF to Trad. NAND Wafer Space Ratio | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x | 3.5x |
| Max HBF GB per 12in. Wafer | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 | 37,449 |
| Wafers Needed for HBF | - | - | - | 25,201 | 52,187 | 96,885 | 168,599 | 226,807 | 264,808 | 321,615 | 376,808 | 428,633 |
| Share of Global Wafer Capacity Used for HBF | 0.00% | 0.00% | 0.00% | 0.63% | 1.28% | 2.31% | 3.95% | 5.28% | 5.61% | 6.68% | 7.67% | 8.55% |

Source: TrendForce, Goldman Sachs Global Investment Research

## Challenges to HBF adoption

Despite the promise of High-Bandwidth Flash, we note several key challenges to adoption.

- **Bandwidth constraints and latency concerns:** According to SanDisk, SK Hynix, and other studies, HBF offers comparable sequential read bandwidth to HBM4, although

we note that (1) HBM4E is currently in the sampling phase and delivers at least ~50% higher bandwidth; (2) although inference workloads are read-dominant overall, any write operations significantly reduces the effective bandwidth of HBF given its NAND-based design; (3) HBF's structurally lower write performance and endurance relative to DRAM-based HBM results in its practical benefits being highly dependent on its read performance keeping pace with the latest generations of HBM.

- **Variability of read/write ratio in inference:** As context windows expand and workloads shift to read-heavy inference, occasional write operations in KV cache remain essential: if no prefix cache hit occurs (i.e., the prompt / prefix is not found in the existing KV cache), the read/write ratio is effectively zero. Although this ratio will likely become increasingly rare over time and inference remains highly read-dominant, the variability of write operation frequency (and the resulting meaningful bandwidth penalty for NAND-based HBF) could render the effective bandwidth of HBF unpredictable in the early phases of any rollout.
- **Endurance, energy intensity, and thermal concerns:** Because it is based on NAND flash, HBF tolerates a finite number of program/erase cycles - whereas DRAM-based HBM offers far greater write endurance. Also, the higher density of a HBF cube result drives higher energy consumption per stack (~160 W per cube in SK Hynix's $H^3$ architecture, ~4X more power than HBM3E vs. ~2.5X more power than HBM4). Given the peak power of most flagship accelerators of ~1,000-1,200 W, and the utilization of 8 HBF stacks in $H^3$, the power footprint of HBF alone can easily approach the accelerator's own constraints and result in new thermal issues. This excess heat may further accelerate wear on the flash cells and complicate lifespan considerations.
- **Manufacturing resource intensity:** As with HBM, HBF uses design changes to maximize bandwidth, including increased internal structural parallelism and 3D packaging methods. It also uses ultra-wide on-package interconnect to alleviate the external data bottleneck. These advanced packaging techniques, the use of hybrid bonding, and the ultra-wide interconnect all contribute to a more resource-intensive and higher-cost manufacturing process, including a significantly higher wafer consumption (most estimates suggest a 3X-4X higher cost relative to traditional NAND, similar to the DRAM/HBM "trade ratio").

In aggregate, we view HBF's potential bandwidth limitations and resource intensity as the most significant challenges facing wide adoption of the technology. As context windows lengthen and the share of inference workflows grows, we see improved visibility on the read/write ratio in KV cache and on pricing as key variables.

## Alternative approaches to scaling the memory wall - XL Flash and zNAND-O

**We are constructive on the broad set of NAND-based technologies set to address memory constraints in AI inference - and see alternative approaches such as XL Flash and zNAND-O ultimately driving similar implications.** While HBF remains one of the leading approaches to addressing current AI system memory constraints given public backing from SanDisk and SK Hynix (and ongoing development from Samsung) and significant peer-reviewed publications, other approaches have emerged, including

"XL Flash." Introduced by Kioxia and built on the company's BiCS Flash 3D NAND architecture, the approach includes key architectural and design differences relative to HBF which could potentially offer easier packaging integration and/or lower cost, potentially at the expense of some performance. We would also highlight Samsung's development of zNAND-O - similar in concept such as using TSV to connect 4 or 8 layer vertical stacks of V-NAND dies, while targeting a different market segment in on-device AI (smartphones, PCs, etc.). Samsung is also developing zHBM for data centers where it stacks the memory on top of the AI accelerator rather than next to it, to improve the performance and power efficiency. Although it is too early to tell which approach might come to market first, we believe that all three approaches are ultimately aimed at solving the "memory wall" problem by applying NAND-based system approaches to increase context length while offering significant potential cost savings at the system level.

**XL flash may hold write, cost, and certain other advantages over HBF, although based on publicly-available literature HBF appears well-suited for read-heavy AI inference applications.** XL Flash utilizes standard PCIe, NVMe, and CXL system buses, but relies on less dense SLC-based NAND, smaller transfer blocks, and larger die architectures relative to traditional 3D NAND flash to maximize bandwidth, significantly lower random read latency, and meaningfully increase write endurance; characteristics that may render XL Flash better suited for write operations relative to HBF. The technology's reliance on existing approaches to 3D NAND manufacturing, its lack of TSVs (through-silicon vias, required for the stacking associated with HBF), and physical separation from the GPU (vs. HBF using the same silicon interposer) may also lower the upfront transition cost and perceived investment risks.

## Individual stock implications:

**Exhibit 10: Select stocks within our coverage positioned to be affected by HBF**

| Company | Ticker | Rating | 12 month price target | Current Price | Upside/ downside |
|---|---|---|---|---|---|
| Advanced Micro Devices | AMD | Buy | $700 | $611 | 14.5% |
| Broadcom | AVGO | Buy | $540 | $363 | 48.9% |
| Kioxia | 285A.T | Buy | ¥38,667 | ¥17,505 | 120.9% |
| Micron | MU | Neutral | $1,250 | $1,027 | 21.7% |
| Nvidia | NVDA | Buy | $300 | $229 | 30.7% |
| Samsung | 005930.KS | Buy | ₩500,000 | ₩262,000 | 90.8% |
| Sandisk | SNDK | Buy | $2,200 | $1,607 | 36.9% |
| SK Hynix | 000660.KS | Buy | ₩3,500,000 | ₩1,681,000 | 108.2% |

Note: Current Price is as of October 9, 2026.

Source: Goldman Sachs Global Investment Research

- **SanDisk (SNDK, Buy):** We remain constructive on SNDK and view the company as a well-positioned beneficiary of the potential widespread adoption of HBF given its technology leadership in collaboration with SK Hynix and its low capital intensity through its Kioxia manufacturing partnership - which we see as part differentiating given the resource intensity of HBF. We also believe simultaneously retain its solid core business in traditional NAND, wh continue to outperform on improving product mix and further prog roadmap.

人在草木间

- **Kioxia Holdings (285A.T, Buy):** We reiterate our Buy rating on Kioxia given our expectations on ongoing supply shortages in the NAND market and constructive pricing outlook, which we believe should lead to positive street estimate revisions. We also believe Kioxia is poised to announce a series of shareholder return, including the resumption of dividend payment as well as continued share buyback. Regarding XL-Flash, which is intended to expand the memory capacity of GPUs in inference GPU servers, Kioxia plans to ship samples of a product achieving over 10 million IOPS (storage processing speed) by the end of CY26. It also plans to ship samples of a second-generation product with processing speed further increased to 100 million IOPS in CY27. Compared to HB Flash (HBF), Kioxia's SSD-based solution is differentiated in that it does not require additional manufacturing processes such as TSV.
- **SK Hynix (000660.KS, Buy):** We maintain our positive view on Hynix and see the company as a potential leader in the HBF market. While the company's focus remains on higher-margin DRAM and HBM, Hynix is uniquely positioned to be a prime beneficiary of HBF market expansion given its proactive initiatives in co-development and industry standardization of HBF with SanDisk. Through leveraging its advanced NAND technology, vast experience in producing various types of NAND, and the advanced packaging expertise from HBM leadership, we believe Hynix possesses the core competencies required for HBF commercialization.
- **Samsung Electronics (005930.KS, Buy, on CL):** We maintain our Buy rating (on the APAC Conviction List) on SEC. Although SEC has not yet released specific details on its work on HBF (although it did recently file a trademark application for "zHBF"), recent announcements on its technology milestones—such as the TSV-integrated zNAND-O architecture—demonstrate its active R&D in high-bandwidth NAND. Supported by its largest NAND capacity globally and technology leadership, we believe SEC will likely be ready to compete in the high-bandwidth NAND market, also leveraging its superior cost management and packaging competitiveness. We highlight SEC being the only company in the world that can produce DRAM and NAND while having the foundry/logic/advanced packaging capabilities, which we believe along with the industry's largest memory capacity (for both DRAM and NAND), will help the company to continue to lead across an increasingly multi-tiered AI memory hierarchy.
- **Micron (MU, Neutral):** While we expect the company to participate in any significant future HBF market, we view its focus on HBM and DRAM, its lack of participation in the consortium governing HBF, and its prioritization of DRAM greenfield capacity overall as likely to temper its potential future market share in the novel NAND technology. We are Neutral-rated on MU, primarily on the risks of HBM pricing momentum slowing in 2028 given the prospects of meaningful supply additions by the industry, as well as possible de-specing of HBM as HBF adoption ramps. We nevertheless see continued tailwinds for Micron's business overall, but view SK Hynix and SanDisk as more direct and nearer-term beneficiaries of any HBF-related inflection in NAND demand.
- **Semi Cap Equipment:** We view HBF as a modest incremental positive for semiconductor equipment demand, though the implications are more nuanced than

a simple increase in NAND content. On one hand, HBF introduces additional manufacturing complexity through TSVs, hybrid bonding, advanced packaging, inspection, and test, which could benefit equipment suppliers with exposure to packaging and integration workflows. On the other hand, HBF's value proposition is rooted in reducing memory capacity bottlenecks and improving GPU utilization, which could ultimately lower the number of accelerators required for a given inference workload. We nevertheless see the added packaging, integration, and memory intensity as likely to offset any reduction in system-level hardware demand.

- **Digital Semis:** We see an incrementally negative impact for a range of XPU accelerator makers, including Nvidia, Broadcom, AMD, and others - as the success of any high-performance NAND approach is likely to reduce the number of accelerators ultimately deployed at a similar performance level given the potential cost savings involved. We would point out that Nvidia is currently the only provider of accelerators which controls the entire system stack from software to hardware components - and hence it may be able to optimize the system-level design to mitigate this impact, similar to what it has done by optimizing the HBM-NAND balance in the current Rubin generation.