High Bandwidth Flash (HBF): The Emerging NAND Tier Beneath HBM for AI Accelerators

A research report on architecture, packaging, vendor roadmaps, and system implications

Prepared July 2026 · Information horizon: public disclosures through early July 2026 (live-verified July 8, 2026; academic literature incorporated July 9, 2026)

© 2026 Shimogoryo Base LLC, a Colorado limited liability company. Free to share — complete and unaltered, with attribution.

This report is the work product of Shimogoryo Base LLC (the "Publisher") and is published free of charge. ***You may read it, act on it, repost it, forward it, and share it publicly, provided you share it complete and unaltered and credit Shimogoryo Base LLC.*** *Please do not excerpt figures or tables in isolation, publish modified versions, or resell it. For commercial reuse, derivative works, or reprint rights, contact the Publisher via shimogoryobase.com/consult.html.*

About this free edition. This report is published at no charge by **Shimogoryo Base LLC**. You are welcome to repost, forward, and share it — complete and unaltered, with attribution. Two companion products build on it and remain available from the Publisher: an **investment memorandum** mapping the HBF supply chain to US-listed equities and quantifying the second-order NAND supply shock, and the **HBF_Investment_Screen.xlsx** workbook carrying the full thirty-one-name grid with per-name evidence, exposure estimates, and bull/bear cases. For those, or for bespoke technology and supply-chain research, see shimogoryobase.com/consult.html.

Important notice (summary). This report is provided for general informational purposes only. It is not investment, financial, legal, tax, accounting, or engineering advice, and it is not an offer or solicitation with respect to any security. It contains forward-looking statements and projections that are inherently uncertain and subject to change without notice. Information is drawn from sources believed reliable, but accuracy and completeness are not guaranteed; the report is provided "AS IS," reflects information only as of the publication date, and the Publisher assumes no duty to update it. All third-party trademarks are the property of their respective owners, and no affiliation or endorsement is implied. **See the "Important Disclosures & Disclaimer" section at the end of this report for the full terms governing its use.**

Methodology and sourcing note. This report was originally compiled via a custom Fable 5 agentic loop prompted by Shimogoryo Base LLC/GK, from public reporting, vendor announcements, and conference disclosures (SanDisk Investor Day, FMS 2025, OCP Global Summit 2025, vendor press releases, and trade-press coverage from Tom's Hardware, Blocks & Files, ServeTheHome, TrendForce, and Korean/Japanese business press) at a time when live web verification was not available. On **July 8, 2026** every load-bearing claim was **live-verified against primary sources**, URL citations were added inline, confidence labels were updated, and a verification log was appended (Appendix B). Each claim remains labeled by confidence: **[Confirmed]** (multiple public sources / vendor statements), **[Reported]** (single-source or trade-press reporting), **[Analyst speculation]** (third-party projection), or **[Author analysis]** this report's own inference either via Shimogoryo Base Principal Analyst or via Claude Fable 5. The final report was hand-reviewed and hand-edited by Andy Kowles of Shimogoryo Base LLC/GK. Andy has conducted research into memory and storage tiers optimized for Inference.

Revision note (July 9, 2026). This revision incorporates an **academic literature survey** (*HBF_Literature_Survey_20260709_092527.md*, prepared July 9, 2026) as a new **Section 10 ("What the Academic Literature Signals")**; the former Sections 10 and 11 are renumbered 11 and 12. Paper-level claims were verified on July 9, 2026 against the arXiv abstract pages for the cited preprints, the Microsoft Research publications page, Korea Times, and Blocks & Files. Four sources were not machine-readable at revision time — the SK hynix Research blog post on H3, the IEEE Xplore paper page, the SemiEngineering KAIST TERALAB overview, and the SanDisk interview with Prof. Joungho Kim — and claims resting on them are labeled as relying on secondary summaries. The July 8 verification history (Appendix B) is retained unchanged apart from renumbered internal section references.

Information horizon caveat (updated). The original draft's sourcing predated February 2026; live verification on July 8, 2026 extended the horizon through early July 2026 and surfaced several H1 2026 developments now incorporated in the body: (1) SanDisk and SK hynix moved HBF standardization into a **dedicated Open Compute Project (OCP) workstream** at a consortium kick-off on February 25, 2026 — not JEDEC, as earlier drafts speculated; (2) the 2H 2026 sample / early 2027 device schedule was **reaffirmed** at that event, though no public sample shipment had been announced as of July 8, 2026; (3) Samsung reportedly began early-stage HBF concept design (Oct–Nov 2025 reporting); (4) NVIDIA reportedly **removed Rubin CPX from its roadmap at GTC in March 2026**; and (5) YMTC publicly declared HBF development plans (Dec 2025).

Table of Contents

1 What Is HBF and Why Now
2 How the NAND Itself Will Be Structured
3 Will HBF Require Unique Wafers?
4 How It Will Be Attached
5 Where It Will Be Attached
6 What HBF Means for HBM Systems
7 The Beachfront (Shoreline) Question
8 Vendor-by-Vendor State of the Art
9 Competing and Adjacent Approaches
10 What the Academic Literature Signals
11 Risks and Open Questions
12 Conclusion and Outlook Table

1. What Is HBF and Why Now

High Bandwidth Flash (HBF) is a proposed memory device class that stacks 3D NAND flash dies with through-silicon vias (TSVs) over a base logic die — in the physical idiom of High Bandwidth Memory (HBM) — to deliver **HBM-class read bandwidth at NAND-class capacity and cost**, packaged on or near an AI accelerator. The concept was introduced publicly by SanDisk at its "Future FWD" Investor Day in February 2025, held as it completed its separation from Western Digital [Confirmed — Tom's Hardware, Feb 2025; TrendForce, Feb 14, 2025 — note the Investor Day slightly preceded the split's

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided “AS IS.”

formal completion], and was elevated from a single-vendor initiative to a would-be industry standard when **SanDisk and SK hynix signed a memorandum of understanding on August 6, 2025 to jointly define an HBF specification and build an ecosystem around it** [Confirmed — SanDisk press release, Aug 6, 2025; Tom's Hardware; Blocks & Files]. In **February 2026** the two companies converted the MoU into a formal **"HBF Spec. Standardization Consortium," launching a dedicated workstream under the Open Compute Project (OCP)** to make HBF a global open standard [Confirmed — SanDisk press release, Feb 25, 2026; SK hynix newsroom].

SanDisk's headline claims at introduction, repeated at the Future of Memory and Storage (FMS) conference in August 2025 (where HBF won the FMS "Best of Show — Most Innovative Technology" award in the NAND flash category [Confirmed — FMS 2025 winners list; StorageNewsletter, Aug 7, 2025]):

- **8–16x the capacity of HBM per stack at comparable cost per stack** — SanDisk has cited on the order of **512 GB per HBF stack**, versus 24–36 GB for HBM3E and roughly 48–64 GB for HBM4-generation stacks [Confirmed — SanDisk/SK hynix MoU release, Aug 2025; Tom's Hardware, Feb 2025].
- **Read bandwidth "comparable to HBM"** — achieved through massive on-die parallelism rather than fast individual cells (see Section 2). SanDisk has since put a number on Gen1: **1.6 TB/s per stack**, which it describes as "more than 50x faster than top-tier PCIe 5.0 SSDs" — below leading-edge HBM4 (~2–2.8 TB/s) but far above any SSD, with multi-generation scaling planned [Confirmed — SanDisk "Memory-Centric AI" blog, 2025; SemiEngineering]. The question of read bandwidth saturation of an HBM-like interface is relatively straightforward: NAND SSD designs have long been designed to saturate the external interface by adding sufficient channel, die, and package parallelism, increasing channel rates (e.g. ONFI3/4/5 and Toggle), and controller datapaths. The intrinsic imbalance of write B/W vs read B/W is acknowledged here even if unsaid/implied in the public announcements. The expected application of this media is for read-dominated workloads.
- An accelerator with **8 HBF stacks could present ~4 TB of directly attached memory**, versus ~192–288 GB for 8 stacks of HBM3E (up to ~512 GB for HBM4) [Confirmed — Tom's Hardware, Feb 2025].
- Roadmap: **first HBF memory samples in the second half of 2026**, and **samples of the first AI inference devices using HBF in early 2027** [Confirmed — SanDisk/SK hynix MoU release, Aug 2025; reaffirmed at the Feb 25, 2026 OCP standardization kick-off per TrendForce, Feb 26, 2026. **Status as of July 8, 2026:** the schedule was publicly reaffirmed in February 2026, but no sample shipment announcement had yet been made — consistent with the stated 2H 2026 window, which had only just opened].
- Media Requirements Latency/Bandwidth/Endurance: A high endurance, low tREAD, unique SLC NAND is expected to be the first NAND media deployed in the HBF tier. While current best-in-class SLC NAND has tREAD around ~5us, this is still much slower than DRAM. Even if all the other requirements are met for HBF, and even ignoring writes and endurance, the read latency of 5us for a 512B-4KB access will need to be improved and this will be costly.

1.1 Why now: the inference memory wall

The motivation is the shift of AI economics from training to **inference**, where the binding constraints are memory capacity and memory bandwidth rather than FLOPs:

- **Model weights keep growing.** Frontier mixture-of-experts models exceed a trillion parameters; even at 4–8-bit quantization, weights alone can exceed the HBM capacity of an 8-GPU node, forcing model sharding across many accelerators purely for *capacity*, not compute [Confirmed framing — e.g., SanDisk "Memory-Centric AI" blog; SK hynix inference-era framing, Feb 2026].
- **KV cache growth.** Long-context and agentic inference makes the transformer key-value (KV) cache a first-order memory consumer: KV cache scales with batch size x context length x layers x heads, and for large models at 100K–1M-token contexts it can reach **tens to hundreds of gigabytes per request batch**, frequently exceeding the space left in HBM after weights. The industry response — KV cache offload to DRAM and SSD tiers (NVIDIA Dynamo/NIXL, LMCache, Mooncake, Huawei UCM) — is exactly the workload HBF targets with a nearer, faster tier [Confirmed — NVIDIA Dynamo

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

documentation 2025; Huawei UCM announcement, Aug 12, 2025 (Tom's Hardware); author analysis on sizing]. The academic literature now quantifies this pressure directly: **KVNAND** notes the KV cache can exceed model weights in size as context grows and reports out-of-memory failures at 100K-token contexts unless KV moves into flash-class capacity [Confirmed — arXiv 2512.03608], while SK hynix's **H3** targets 10-million-token cache-augmented generation with precomputed KV held in HBF (Section 10).

- **HBM cost and capacity limits.** HBM is the most expensive commodity memory per bit ever shipped in volume — several times the per-bit price of DDR5, itself several times NAND. DRAM cell scaling has slowed; HBM capacity grows mainly by stacking more, thinner dies (12-high, 16-high), which raises cost and yield risk superlinearly. NAND, by contrast, is roughly **one to two orders of magnitude cheaper per bit** than HBM and scales by adding layers [Reported directionally — TrendForce/market pricing commentary 2024–2025; exact ratios fluctuate with the memory cycle and no single canonical source exists]. HBM capacity is constrained not just by DRAM scaling and stacking, but by the very high shoreline consumption of HBM sites with their thousands of pins. The later HBM interfaces doubled the number of pins and this no longer scales. The number of sites per physical xPU die are constrained by shoreline/beachfront. More on this scaling problem/opportunity later in the report.
- **Read/write asymmetry of inference.** Weights are written once and read continuously; prefix/KV caches in disaggregated serving are written once per prefill and read many times during decode and cache reuse. A **read-mostly** tier tolerates NAND's limited write endurance and slow programming, while exploiting its density [Author analysis, consistent with SanDisk's stated positioning].

[Author analysis] HBF is best understood not as "flash replacing HBM" but as the industry attempting to formalize a new rung of the memory pyramid which is nonvolatile: *HBM for hot, write-intensive working set; HBF for warm, read-mostly capacity (weights, KV/prefix caches, embeddings, RAG corpora); SSD for cold bulk.* The reason it is happening now is that inference token economics finally put a dollar value on every gigabyte within a microsecond of the GPU — and NAND is the only technology that can supply terabytes there at acceptable cost. Acceptable cost has yet to be determined but the NAND media that can meet the requirements of this new tier will, like HBM-DRAM, be the most expensive NAND by far, and still with much longer latency than DRAM. It can still make sense given the exploding need for capacity in this tier. Unique wafer starts will require manufacturing capex and/or put pressure on other media swimlanes in the same way HBM has put major pricing pressure on normal DDR DRAM by pushing the supply-demand microeconomics out of equilibrium into excess demand.

2. How the NAND Itself Will Be Structured

HBF is not "an SSD in an HBM package." Delivering HBM-class read bandwidth from NAND requires re-architecting the NAND die itself. The picture below synthesizes SanDisk's FMS 2025 disclosures, SK hynix's OCP 2025 "AIN B" description, and known BiCS/CBA technology.

2.1 Massive on-die parallelism: many small arrays

A conventional TLC NAND die has 4–6 planes and delivers on the order of **1–3 GB/s** sustained read — three orders of magnitude short of HBM. A single NAND page read takes tens of microseconds (TLC) down to a few microseconds (SLC/XL-Flash). The only way to bridge the gap is **parallelism**: partitioning the die into **dozens to hundreds of independently addressable mini-arrays (micro-planes)**, each with its own sense amplifiers and page buffers, all streaming concurrently into a wide TSV interface. SanDisk's HBF core-die description at FMS 2025 explicitly showed the array **subdivided into many small arrays accessed in parallel**, trading array-efficiency (more periphery per bit) for bandwidth [Confirmed — SanDisk "Scaling the Memory Wall" blog, 2025; SemiEngineering]. This will incur unit cost and require unique wafer starts. HBF NAND will be economically disruptive, and the author recommends using the HBM disruption of DRAM microeconomics as a model. Shimogoryo Base Consulting has done exactly this: We can provide a custom pricing model upon request.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

[Author analysis — sizing check, rebased to the disclosed 1.6 TB/s.] To sustain 1.6 TB/s of read bandwidth (Section 2.5) from a 16-die stack, each core die must source **~100 GB/s**. With SLC-mode reads of ~3–5 µs and 16 KB pages, one mini-array yields ~3–5 GB/s only if pipelined; realistically each die needs **on the order of 32–64+ concurrently active mini-arrays** plus deep read pipelining. (An earlier draft ran this arithmetic at ~1 TB/s → ~64 GB/s per die, before SanDisk's 1.6 TB/s figure was verified; the conclusion is unchanged, only ~1.6x harder.) This is aggressive but tractable — a layout/periphery problem rather than cell physics, which is why wafer-bonded logic (below) matters so much. The assumption of 16K pages is not verified. The questions of page size and codeword size vs. tREAD are open.

2.2 CBA / wafer-bonded logic under the array

Both leading HBF proponents build on **bonded-periphery NAND**:

- **SanDisk/Kioxia BiCS8 (218-layer) introduced CBA — "CMOS directly Bonded to Array"** — in which the NAND cell array is fabricated on one wafer and the CMOS periphery (decoders, sense amps, I/O) on a second wafer in a more advanced logic-friendly process, then hybrid-bonded face-to-face [Confirmed — Kioxia CBA technology page; Kioxia/Western Digital BiCS8 announcement, Mar 30, 2023; TechInsights BiCS8 218L CBA analysis].
- **YMTC's Xtacking** two-wafer hybrid-bonded architecture debuted at the Flash Memory Summit in **August 2018** (Xtacking 2.0 followed in 2019), making YMTC the first NAND maker to apply hybrid bonding; Xtacking 4.0 powers its newest generations [Confirmed — YMTC Xtacking unveiling at FMS 2018 (PRNewswire); TechInsights Xtacking 3.0 blog: "Xtacking 1.0 at 2018's FMS... Xtacking 2.0 in 2019"; first-to-hybrid-bond: TrendForce, Feb 24, 2025; Xtacking 4.0: TechInsights].

CBA is nearly a prerequisite for HBF: the huge number of sense-amp banks, wide internal buses, TSV landing zones, ECC, and an HBM-like interface consume far more CMOS area and speed than a conventional under-array periphery can supply. Putting periphery on its own wafer lets HBF spend transistors freely on parallelism without sacrificing bit density on the array wafer [Confirmed as SanDisk's stated approach — SemiEngineering; author elaboration on rationale].

2.3 SLC (or near-SLC) cells and read-latency targets

HBF proponents have consistently framed the device as **read-optimized**. The natural media choice is **SLC or aggressive low-latency modes**:

- Kioxia's **XL-Flash** (SLC, BiCS-based, in production since ~2020) demonstrates ~**3–5 µs read latency**, versus 40–100 µs for conventional 3D NAND [Confirmed — Kioxia XL-FLASH infographic; Tom's Hardware, 2025]. Kioxia presented a roadmap at FMS 2025 with at least 2 more generations of XL-Flash, with each technology node increasing the effective bandwidth. Even if HBF is a total failure, XL-Flash Gen2 and Gen3 will allow saturation of the latest PCIe interfaces and the production of ultra high bandwidth SSDs, with single digit microsecond read latency.
- SanDisk has indicated HBF read latency will be **microsecond-class — far slower than HBM's ~100 ns, but 100–1000x faster than SSD-path accesses** once queuing, protocol, and software stacks are included [Reported — FMS 2025 Q&A coverage; exact figures not formally specified].
- **[Analyst speculation]** Some analysts expect first-generation HBF to use SLC/MLC for latency and endurance, with TLC-based "capacity HBF" variants later if the tier bifurcates, mirroring how HBM split into bandwidth- and capacity-optimized SKUs.

The trade-off: SLC cuts density 3x versus TLC, but NAND's cost lead over DRAM is so large (~40–50x per bit for TLC) that even SLC HBF remains ~10x+ cheaper per bit than HBM [Author analysis].

There is also the question of rBER and coderate, not to mention cell wear, between these possible media configurations in an HBF slot. Given that the traditional SSD controller will need to be reimagined, we infer that the design of an HBF NAND controller will be taken as an opportunity to disrupt the legacy. See Section 4.2 ("Base/logic die with an HBM-like PHY").

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

2.4 TSVs through NAND dies and 8/16-high stacking

The stack physically mimics HBM:

- **Core dies:** NAND (array + bonded CMOS) thinned and perforated with **TSVs**, stacked **8-high or 16-high** over a base die. SanDisk's reference for the 512 GB figure is a **16-high stack of 32 GB (256 Gb-class) core dies**, using proprietary stacking with ultra-low die warpage for 16-high configurations [Confirmed — SanDisk "Memory-Centric AI" blog; SemiEngineering; the die-level split, originally author inference, was corroborated by 2025–2026 disclosures].
- **Base logic die:** hosts the external PHY (HBM-like, 1024–2048-bit wide), command/protocol handling, scheduling across mini-arrays, ECC (NAND needs far stronger ECC than DRAM — LDPC-class), wear management, and bad-block/redundancy mapping. This makes the HBF base die closer to a **flash controller** than an HBM buffer die [Confirmed in outline — SemiEngineering; author elaboration].
- TSVs through NAND are newer territory than through DRAM, but Toshiba (now Kioxia) demonstrated the **world's first 16-die TSV-stacked NAND in 2015**, explicitly targeting high-bandwidth, low-latency enterprise SSDs — so the manufacturing step is evolutionary [Confirmed — Toshiba/Kioxia press release, Aug 6, 2015. Correction of emphasis: the original draft attributed shipping TSV NAND to "Samsung and Kioxia since the late 2010s"; the verifiable anchor is Toshiba's 2015 first demonstration.]

2.5 Bandwidth targets

- **Stated framing:** "bandwidth comparable to HBM" — HBM3E delivers ~1.0–1.2 TB/s per stack; the JEDEC HBM4 standard (JESD270-4, published April 16, 2025) specifies up to **2 TB/s** per stack, with vendor implementations reaching **>2.8 TB/s** at 11+ Gb/s pin speeds [Confirmed — JEDEC JESD270-4 press release; Tom's Hardware; Micron HBM4 product page]. SanDisk's multi-generation HBF roadmap shown at FMS 2025 projects rising bandwidth and capacity and falling energy-per-bit across HBF Gen1 → Gen2 → Gen3, and SanDisk has publicly cited **1.6 TB/s** read bandwidth for HBF [Confirmed — SanDisk "Memory-Centric AI" blog].
- **[Author analysis — updated after verification.]** The original draft estimated a Gen1 landing zone of "hundreds of GB/s up to ~1 TB/s read" per stack; SanDisk's disclosed **1.6 TB/s** figure exceeds that estimate, landing between HBM3E and baseline HBM4 per-stack bandwidth. Write bandwidth is still expected to be an order of magnitude lower (NAND programming) — the read/write asymmetry analysis stands.

Attribute (per stack)	HBM3E	HBM4	HBF (Gen1 targets)	Enterprise SSD (Gen6 x4)
Capacity	24–36 GB	48–64 GB	**~512 GB (8–16x HBM)**	4–16 TB (device)
Read bandwidth	~1.0–1.2 TB/s	~2.0–2.8 TB/s	**1.6 TB/s (SanDisk-stated)** [Confirmed]	~28 GB/s
Write bandwidth	= read	= read	**Much lower (NAND program)**	~14–28 GB/s
Random access latency	~100–150 ns	~100–150 ns	**~ single-digit µs (SLC-class)**	~50–100 µs (path)
Endurance	Unlimited writes	Unlimited writes	**Limited (NAND P/E cycles)**	Limited (DWPD-rated)
Cost per bit	Highest	Highest	**~10x+ cheaper than HBM** [directional]	Cheapest
Attach	Interposer, HBM PHY	Interposer, 2048-bit	**HBM-compatible site (goal)**	PCIe/NVMe
Volatility	Volatile	Volatile	**Non-volatile** (bonus: instant weight load)	Non-volatile

(Table values: HBM figures confirmed from JEDEC/vendor specs; HBF column is SanDisk-stated goals plus author estimates as labeled.)

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

Non-volatility is an under-discussed bonus: model weights persist across resets, enabling near-instant model switching and reload — a feature SanDisk has explicitly promoted [Confirmed — SanDisk "Scaling the Memory Wall" blog, 2025].

3. Will HBF Require Unique Wafers?

Short answer: yes — HBF will use purpose-designed NAND array wafers and dedicated CMOS wafers, not repurposed commodity SSD dies. [Confirmed in direction — SemiEngineering architecture synthesis; SanDisk "Scaling the Memory Wall" blog; details below mix confirmed points and author analysis.]

3.1 Dedicated array wafers

A commodity TLC die is optimized for **cost per bit**: maximal array efficiency, few planes, minimal periphery. An HBF core die inverts these priorities:

- **Plane/mini-array count 10–50x higher**, with sense-amp and page-buffer area to match — a fundamentally different floorplan [Confirmed concept (subdivided parallel arrays) — SanDisk "Scaling the Memory Wall" blog; the 10–50x multiplier is author estimate].
- **TSV keep-out regions**: TSV fields must punch through the die stack, requiring the 3D NAND array (a monolithic block of ~200–400 wordline layers) to be partitioned around TSV corridors — etch/fill integration commodity NAND wafers lack [Author analysis; consistent with how DRAM had to be redesigned for HBM].
- Likely **SLC-optimized cell operation** and relaxed layer count in Gen1 (bandwidth, not density, is the product), possibly reusing a proven BiCS/V-NAND node rather than the bleeding-edge layer count [Analyst speculation, widely echoed in trade coverage].

3.2 Dedicated CMOS wafers on "advanced-ish" logic nodes

With CBA/Xtacking-style bonding the periphery wafer is a separate mask set anyway; HBF pushes it harder. Wide internal buses, fast TSV-facing I/O, large LDPC engines, and scheduling logic argue for a **more advanced logic node than typical NAND periphery** (historically 28–40 nm-class), with the *base* die possibly on a foundry logic node, echoing HBM4's move of base dies to TSMC [Author analysis; HBM4 base-die precedent Confirmed — SK hynix–TSMC announcement, April 2024].

3.3 Cost and fab implications

- **Cost:** HBF dies sacrifice array efficiency, may use SLC, add TSVs, and consume a bonded CMOS wafer per array wafer — so HBF cost per bit will be **several times commodity TLC**, yet still roughly an order of magnitude below HBM. SanDisk's framing — "similar cost to HBM *per stack*, 8–16x the capacity" — implies exactly this positioning [Confirmed framing — Tom's Hardware, Feb 2025; author cost decomposition].
- **Fabs:** SanDisk's NAND comes from the **SanDisk–Kioxia joint-venture fabs (Yokkaichi, Kitakami)**, which already run CBA bonding at scale for BiCS8+ — the JV is the natural HBF manufacturing base, and it puts Kioxia physically adjacent to HBF even though the announcement was SanDisk's [Confirmed JV structure — Kioxia/Western Digital joint BiCS8 announcement; author inference on HBF siting]. **SK hynix** would build HBF-class NAND in its NAND fabs (M15X and the Cheongju complex; note M15X was designated primarily for DRAM/HBM expansion — SK hynix's NAND capacity sits in M15/M11/M12 and Solidigm-era assets), with HBM-honed TSV/stacking lines providing the packaging know-how [Confirmed capacity facts; author inference on allocation].
- **[Author analysis]** The strategic consequence: HBF creates a **premium, differentiated NAND product line** insulated from commodity NAND price cycles — precisely why NAND vendors squeezed by HBM-rich DRAM competitors are motivated to make it exist. Kioxia has long produced XL-Flash on a unique wafer, they are ahead in this niche. Micron has, on the other

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

hand, relied on pSLC mode to create high endurance SSDs from the undisrupted TLC wafer flow, and have made huge gains in enterprise and data center SSD shares with mainstream G8 and now G9 TLC wafers. These two approaches are significantly different, and have major implications for the economics of HBF and, bigger picture, for the economics of performance NAND vs. capacity NAND. Kioxia–SanDisk–SK hynix is a formidable fraction of NAND fabrication and innovation and this industry collaboration introduces HBM experience and lessons learned and xPU vendor relationships into non-HBM NAND producers. Micron and Samsung are the only other DRAM/HBM/NAND producers. Samsung has their V-NAND experience for high speed cells on unique wafers. Micron does not. For industry watchers, it will be significant (even if HBF fails to launch) if and when Micron begins or announces unique high speed NAND wafer starts or capex for capacity expansion.

4. How It Will Be Attached

4.1 Inside the stack: TSV + microbump first, hybrid bonding later

- **Gen1 HBF stacks are expected to use TSVs with microbump (or advanced MR-MUF/TCB-class) die attach**, directly borrowing HBM3E-era assembly — the equipment and process exist today, notably at SK hynix (MR-MUF) and in OSAT ecosystems [Reported/author analysis; SanDisk showed HBM-like stack cross-sections at FMS 2025].
- **Hybrid (bumpless Cu-Cu) bonding** is the roadmap for taller, cooler, denser stacks — the same transition HBM is making for 16-high HBM4/HBM4E. NAND vendors are unusually well positioned here: YMTC has bonded every NAND die it ships, and Kioxia/SanDisk bond every BiCS8+ die (array-to-CMOS). Extending wafer-level bonding vertically through a 16-high stack is a serious but natural progression [Confirmed vendor bonding capabilities — TechInsights on Xtacking; Kioxia CBA page; author projection on timing].

4.2 Base/logic die with an HBM-like PHY

The stack terminates in a **base die exposing an HBM-compatible electrical interface**. SanDisk's stated design goal is that HBF match HBM at the physical/electrical level — **same electrical interface, pinout, package footprint, power profile, and stack height as HBM4 — with some protocol changes** handled by the memory controller (NAND needs asynchronous, variable-latency page-granular read semantics, richer error/management signaling, and asymmetric write handling). Note the important nuance, explicit in early coverage: HBF is **not a literal drop-in replacement** for HBM — the host controller must implement the modified protocol — but it reuses HBM interposer and packaging infrastructure as-is [Confirmed — Tom's Hardware, Feb 2025: "same electrical interface... with some protocol changes," "not drop-in compatible"]. The SanDisk–SK hynix MoU — and since February 2026 the OCP workstream — exists precisely to standardize this protocol layer so xPU vendors design one controller [Confirmed — MoU, Aug 6, 2025; OCP workstream, Feb 25, 2026].

But just because the interface is electrically compatible does not mean that HBF interface is going to be trivial, and certainly we expect that xPU vendors will need to incorporate custom silicon for HBF controller functionality even if the HBF/HBM physical interface (PHY) silicon can be reused and/or shared. HBM is a DRAM interface. NAND interfaces are significantly different. Significant logic in the interposer will be necessary, likely more sophisticated requiring more gates, than HBM. This adds marginal cost and may degrade the HBF economics.

Expect xPU vendors, those designers who attach HBM today, to make design choices to be able to configure each of their limited number of "sites", with its expensive shoreline requirements, to be able to accommodate any configuration of HBM/HBF either in uniform or hybrid stacks. This is not trivial. HBF is therefore gated, as HBM was gated, on close GPU vendor collaboration with the media fabricators. Expect the media vendors who created the HBM ecosystem to leverage their existing engineer-to-engineer and other relationships. This is one reason why SanDisk needed SK hynix on board for the HBF ecosystem to have a chance at success. And the significance of Samsung and Micron participation is major. Shimogoryo Base

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

recommends close examination of the moves of Samsung and Micron with respect to HBF: Either of these companies can be kingmaker of the HBF ecosystem.

4.3 Interposer attach: CoWoS-class 2.5D

Because HBF speaks an HBM-class parallel PHY (1024–2048 bits at multi-Gb/s), it needs **interposer-class wiring density**: TSMC CoWoS-S/-L/-R, Samsung I-Cube, Intel EMIB/EMIB-T, or equivalent silicon-bridge/RDL solutions. HBF stacks would occupy sites on the same 2.5D substrate as HBM, placed and reflowed identically [Confirmed as the intended model — Tom's Hardware, Feb 2025 (HBM interposer/packaging reuse); packaging-vendor specifics are author analysis]. Nothing about HBF requires 3D (SoIC-style) attach in Gen1, though hybrid HBM/HBF 3D stacking is a discussed future (Section 6).

4.4 On-package vs near-package fallbacks

If interposer sites are too scarce (Section 7), fallback attach options exist, in descending bandwidth order [Author analysis, informed by industry discussion at OCP 2025]:

1. **On-interposer HBM site** (native plan of record).
2. **Near-package on-substrate** with a serial/UCIe die-to-die link to the xPU. This presents an opportunity to completely remove ONFI/Toggle from the equation, and demands unique die and packaging.
3. **On-module** (mezzanine/SOCAMM-like flash module) — bandwidth drops to link speeds.
4. **Across CXL/PCIe** — at which point it converges with high-IOPS AI SSDs and loses most of HBF's raison d'être.

5. Where It Will Be Attached

Four placements are in play; the first is the canonical proposal.

5.1 Same interposer as the xPU, in HBM-style shoreline sites (plan of record)

The SanDisk/SK hynix model: an accelerator with, say, 8 memory sites populates **some with HBM and some with HBF** — e.g., 6 HBM3E + 2 HBF = ~200 GB hot + ~1 TB warm; or inference-specialized parts go HBF-heavy (2 HBM + 6 HBF = ~3 TB) [Confirmed concept — Tom's Hardware, Feb 2025 (SanDisk illustrative configurations)]. **Trade-offs:** highest bandwidth and lowest latency to flash ever offered; but consumes precious shoreline, couples NAND thermally to a 1kW-class package (Section 11), and demands the drop-in electrical compatibility actually be achieved.

This has been explored in academic papers — most directly by **SK hynix Research's H3 paper** ("H3: Hybrid Architecture Using High Bandwidth Memory and High Bandwidth Flash for Cost-Efficient LLM Inference," IEEE, 2026), which evaluates exactly this placement: HBM and HBF stacks sharing a single interposer with an NVIDIA Blackwell-class GPU (modeled as a B200 carrying 8 HBM3E + 8 HBF stacks), storing read-only model weights and precomputed KV in HBF and mutable, generated KV cache in HBM behind a unified address space, and reporting up to **2.69x performance per watt** over an HBM-only baseline [Confirmed — Blocks & Files, Feb 16, 2026; paper: IEEE Xplore; vendor summary: SK hynix Research blog]. See Section 10 for the full literature synthesis.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

5.2 Hybrid stacks: HBF under/over HBM, or DRAM-cached HBF stacks

A more radical option discussed by SanDisk and analysts: **a single stack mixing DRAM and NAND dies** — e.g., a few HBM DRAM dies acting as a hardware-managed cache/write buffer atop HBF core dies, presenting one interface. This hides NAND latency for hot lines and absorbs write churn [Reported — SanDisk future-concept commentary at FMS 2025; Analyst speculation on implementations]. **[Author analysis]** This is the most elegant endpoint but the hardest: it requires DRAM and NAND vendors to co-assemble one stack (why SK hynix, positioned to build both dies in-house, is so strategic a partner), plus a substantially smarter base die.

This has been explored in academic research in South Korea and elsewhere: SK hynix Research's **H3** architecture already implements a **latency-hiding buffer** in front of its HBF stacks to compensate for flash access delays — a system-level step toward the DRAM-cached-HBF concept, published by the very vendor positioned to co-assemble mixed stacks [Confirmed — Blocks & Files, Feb 16, 2026; SK hynix Research blog] — and **KAIST TERALAB** (Prof. Joungho Kim, widely called the "father of HBM") has published roadmap work placing HBF and hybrid HBM/HBF stacking on a multi-year trajectory in its 371-page **HBM Roadmap Ver 1.7** (June 2025) and subsequent HBF workload seminars [Reported — survey summary; full PDF; SemiEngineering overview; Prof. Kim's HBF roadmap coverage Confirmed — Korea Times, Feb 3, 2026]. See Sections 10.6–10.7.

5.3 On-module: SOCAMM-like or mezzanine flash

LPDDR-based **SOCAMM** modules (Micron, Samsung, SK hynix, for NVIDIA GB300-class systems) established the pattern of high-bandwidth memory modules near the package. A flash analog ("flash SOCAMM," or Kioxia-style XL-Flash modules) provides terabytes at link-limited bandwidth without touching the interposer [Reported SOCAMM context — Micron announcements 2025 (trade coverage); flash-module analog is Analyst speculation/author analysis].

5.4 Across a link: CXL/PCIe

CXL memory expanders and GPU-direct NVMe (BaM, GPUDirect Storage, Nvidia's "Storage Next"-style initiatives) already offload KV cache today. This is the incumbent HBF must beat: it wins on ecosystem maturity and zero shoreline cost, loses ~10–100x on bandwidth and latency [Confirmed landscape; comparison is author analysis].

[Author analysis — placement verdict.] Expect Gen1 HBF on-interposer in HBM sites for inference-specialized accelerators, with hybrid HBM+HBF stacks as the Gen2/Gen3 differentiator and CXL/NVMe tiers persisting underneath as the cold layer (the deciding economics: Section 7).

---

6. What HBF Means for HBM Systems

6.1 HBM is not displaced — it is complemented

Every credible actor, including SanDisk, frames HBF as **additive**: NAND's microsecond reads, slow writes, and finite endurance make it unusable as the primary working memory for activations, gradients, or KV-cache *write* traffic. HBM remains the hot tier; HBF absorbs the **read-mostly capacity** (weights, reused prefix/KV, embeddings, RAG indices) that currently forces over-provisioning of HBM and over-sharding of models [Confirmed vendor framing — SanDisk "Memory-Centric AI" blog (HBF as HBM augmentation); author endorsement]. Notably, SK hynix — the HBM market leader — co-signing the HBF standardization effort is itself strong evidence the HBM incumbents see complement, not cannibalization [Confirmed — MoU, Aug 6, 2025].

[Author analysis] The write-asymmetry point deserves emphasis: decode-phase KV *appends* are write traffic. A naive "put the KV cache in HBF" design would burn NAND endurance quickly. Practical designs write KV to HBM/DRAM and demote

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

completed, reusable prefixes to HBF — i.e., HBF as a shared prefix/weight library, with HBM as the mutable frontier. This matches how Mooncake/LMCache/UCM-style software already tiers KV data.

6.2 Impact on the HBM roadmap

- **Custom base dies converge.** HBM4/HBM4E introduce customer-specific base dies built on foundry logic (SK hynix–TSMC collaboration, announced April 2024 — TSMC N12/N5-class processes for HBM4 base dies; Samsung and Micron equivalents). The same customization vector can host **HBF controllers or HBF-aware caching logic in an HBM base die**, or conversely DRAM caches in HBF base dies — the base die becomes the tier-fusion point [Confirmed HBM4 custom-base trend; fusion scenario is Analyst speculation/author analysis].
- **Capacity pressure eases at the margin.** If HBF succeeds, the pressure for ruinously expensive 20-high/24-high HBM capacity plays softens; HBM competes on bandwidth/watt where DRAM is unbeatable [Author analysis].
- **Packaging demand compounds.** Both tiers consume CoWoS-class capacity and TSV/bonding throughput — packaging, already the industry bottleneck, becomes more so [Author analysis].

6.3 Software: tiering becomes architectural

HBF forces the accelerator software stack to acknowledge a **non-uniform, asymmetric-write memory** inside the package: page/segment placement of weights, prefix-cache demotion policies, wear-aware allocation, and failure-domain handling (a worn HBF block must not crash a training job). Precedents exist — Huawei's **UnifiedCache Manager (UCM)**, open-sourced in 2025 for KV tiering across HBM/DRAM/SSD, and NVIDIA Dynamo's KV block manager — and HBF essentially gives such software a hardware tier worth managing [Confirmed software precedents — Huawei UCM (GitHub), NVIDIA Dynamo 2025; author framing]. The academic literature confirms the direction: Microsoft Research's **MemExplorer** treats tier placement across SRAM/HBM/LPDDR/GDDR/HBF as an automated design-space parameter for inference NPUs [Confirmed — arXiv 2604.16007], and both **H3** (weights and precomputed KV in HBF, generated KV in HBM) and **KVNAND** (page-level KV cache mapping aligned to flash organization) implement KV-tier placement policies of exactly the shape sketched in Section 6.1 [Confirmed — Blocks & Files on H3; arXiv 2512.03608]. See Section 10.

7. The Beachfront (Shoreline) Question

Will AI xPUs need new specialized beachfront for HBF PHYs? The HBF camp's answer is "no — reuse HBM's," and this is arguably the single most important design decision in the whole proposal.

7.1 The shoreline scarcity problem

An HBM-class parallel interface must exit the compute die's edge ("beachfront"/"shoreline") into the interposer with escape-routing density that only the die perimeter supports. Modern accelerators are already shoreline-bound: reticle-limited compute dies devote most of two or more edges to 6–8 HBM PHYs plus die-to-die and I/O PHYs; HBM4's 2048-bit interface doubles per-stack wiring. Shoreline, not area, increasingly determines how much memory bandwidth a die can have — this is why CoWoS-L keeps growing (multi-reticle interposers, with TSMC's roadmap heading toward >5x reticle late-decade) and why multi-die xPUs partition specifically to multiply beachfront [Reported industry context — TSMC roadmap disclosures 2024–2025; author synthesis].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

7.2 HBF's answer: occupy HBM sites, don't add new ones

SanDisk's stated goal is that HBF share the **same electrical interface and physical site as HBM, with protocol differences** — so an xPU designer builds one PHY and decides *at product-definition (or even SKU) time* how many sites carry HBM vs HBF [Confirmed — Tom's Hardware, Feb 2025]. Consequences:

- **No new beachfront class is needed** — but HBF *competes with HBM for the existing sites*, making the HBM-vs-HBF split a zero-sum shoreline allocation. Inference SKUs trade bandwidth for capacity; training SKUs stay all-HBM [Author analysis].
- The memory controller behind the shared PHY must handle **variable-latency, split-transaction NAND semantics** — this is where the standardization effort (since February 2026, a dedicated OCP workstream; Section 11) matters most, because a "same-PHY, different-protocol" device is only drop-in if the protocol is stable and public [Confirmed problem statement — SanDisk/SK hynix standardization rationale, Aug 2025; OCP workstream, Feb 25, 2026].

7.3 Alternatives if sites run out

- **Bigger interposers** (CoWoS-L multi-reticle) to add sites for both HBM and HBF — the brute-force answer, already on TSMC's roadmap for other reasons [Confirmed roadmap; application to HBF is author analysis].
- **Hybrid stacks** (Section 5.2): put capacity *vertically* inside existing HBM sites instead of consuming new ones — the most shoreline-efficient endgame [Analyst speculation/author analysis].
- **UCIe/die-to-die attach**: hang HBF off a companion I/O die or use serial links; saves compute-die shoreline, costs latency/power per bit and adds a chiplet. Plausible for mid-range inference parts [Author analysis].
- **Custom HBM4E base dies with integrated HBF back-doors** — speculative, but the custom-base-die era makes stack-behind-stack topologies conceivable [Analyst speculation].

[Author analysis — verdict.] No *new* beachfront type will be required if the compatibility goal is met; the real fight is allocation of existing HBM shoreline, and it will be decided by inference token economics per millimeter of die edge.

---

8. Vendor-by-Vendor State of the Art

8.1 SanDisk — inventor and evangelist

- **Feb 2025:** Introduced HBF at its "Future FWD" Investor Day (held as the WD spin-off completed); disclosed 8–16x HBM capacity at comparable cost, ~512 GB/stack, HBM-compatible interface goal; development reportedly informed by input from major AI accelerator vendors [Confirmed — Tom's Hardware, Feb 2025; TrendForce, Feb 14, 2025].
- **Jul 24, 2025:** Formed an **HBF Technical Advisory Board** including **Prof. David Patterson** (RISC/RAID pioneer, Turing laureate) and **Raja Koduri** (ex-Intel/AMD graphics chief, founder/CEO of Oxmiq Labs) to steer architecture and ecosystem [Confirmed — SanDisk press release, Jul 24, 2025. **Correction:** the original draft dated this to February 2025; the press release is dated July 24, 2025.]
- **Aug 2025 (FMS):** Deeper architecture disclosure (mini-array parallelism, CBA-based core die, 16-high stacks, multi-generation roadmap); **Best of Show — Most Innovative Technology award**; presented HBF-attached inference concepts and reiterated **samples 2H 2026, first HBF-based inference silicon sampling early 2027** [Confirmed — FMS 2025 winners; StorageNewsletter, Aug 7, 2025].
- **Aug 6, 2025: MoU with SK hynix** to jointly standardize HBF specifications and ecosystem [Confirmed — joint release].
- **Feb 25, 2026: "HBF Spec. Standardization Consortium" kick-off** with SK hynix at SanDisk HQ (Milpitas, CA); global standardization to proceed via a **dedicated Open Compute Project (OCP) workstream** [Confirmed — SanDisk press release, Feb 25, 2026; SK hynix newsroom].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

- **Jun 2026:** A SanDisk patent surfaced describing a **processor bonded directly onto a NAND "tile" sharing an interposer with HBM stacks** — a possible beyond-Gen1 direction [Reported — TrendForce, Jun 22, 2026].
- Manufacturing base: Kioxia–SanDisk JV fabs (Yokkaichi/Kitakami) with CBA-bonded BiCS8/BiCS9 [Confirmed JV — Kioxia/Western Digital joint BiCS8 announcement; author inference on HBF siting].

8.2 SK hynix — the HBM king hedges into flash

- **Aug 2025:** Co-signed HBF standardization MoU with SanDisk — significant because SK hynix leads HBM (~50%+ share) and thus legitimizes HBF as complement [Confirmed — joint release, Aug 6, 2025].
- **OCP Global Summit, Oct 13–16, 2025:** Unveiled the **"AIN" (AI-NAND) family**: **AIN P (Performance)** — high-IOPS, small-block SSD aimed at Nvidia-class GPU-direct storage, **samples targeted by end of 2026**; **AIN B (Bandwidth)** — NAND stacked HBM-style, i.e., SK hynix's HBF implementation under the MoU; **AIN D (Density)** — petabyte-class QLC storage for AI data lakes [Confirmed — SK hynix newsroom, Oct 2025; Tom's Hardware].
- **Feb 2026:** SK hynix Research published the **H3 hybrid HBM+HBF architecture paper** (IEEE), peer-reviewed evidence of its HBF design intent — see Sections 5.1 and 10 [Confirmed — Blocks & Files, Feb 16, 2026; SK hynix Research blog].
- **Feb 25, 2026:** Co-launched the HBF standardization consortium and OCP workstream with SanDisk [Confirmed — SK hynix newsroom, Feb 2026].
- Unique position: the only company with in-house leading HBM DRAM *and* NAND *and* MR-MUF/TSV stacking at scale — the natural builder of hybrid HBM+HBF stacks [Author analysis].

8.3 Samsung — capable, quiet, now reportedly moving

- No HBF-branded product announcement through mid-2026, but reporting from October 2025 through February 2026 says Samsung has **begun early-stage concept design work on its own HBF products**, leveraging its NAND share and prior high-performance storage R&D (e.g., Z-NAND), with specifications and production timelines not yet finalized and commercial targets around 2027 [Reported — DigiTimes, Oct 7, 2025; TrendForce, Nov 11, 2025; The Investor/Korea Herald, Feb 4, 2026. **Update:** the original draft characterized Samsung as offering only "soft comments"; this reported entry into HBF concept design is a posture change.]
- Technology position is strong: TSV-stacked NAND experience, V-NAND roadmap moving to **bonded architectures around V10 (~400+ layers, ~2026)**, plus in-house HBM, 2.5D (I-Cube) and logic — Samsung could vertically integrate an HBF competitor faster than anyone if it commits [Reported roadmap direction — Samsung disclosures 2024–2025; author assessment].
- **[Analyst speculation]** Samsung's likely path: develop independently while the SanDisk/SK hynix spec is standardized through the **OCP workstream** (launched Feb 2026), then align with the open standard rather than adopt a rival-defined spec early; meanwhile push CXL and its own AI-SSD line. Korean-press reporting frames Samsung's approach as "redefining the technology's role" within a broader AI memory/storage restructuring rather than following the SanDisk/SK hynix path [Reported — The Investor, Feb 4, 2026].

8.4 Micron — HBM-first, flash-for-AI via SSDs, HBF-skeptical posture

- Micron's public AI-memory story is **HBM3E/HBM4 (36 GB 12-high HBM4 samples shipped to key customers June 10, 2025 at ~2.0 TB/s, later demonstrated >2.8 TB/s at 11+ Gb/s; high-volume production for NVIDIA Vera Rubin entered 2026), HBM4E with custom base dies, SOCAMM/LPDDR for Grace-class systems, and G8/G9 NAND in data-center SSDs** [Confirmed — Micron HBM4 press release, Jun 10, 2025; Micron HBM4 product page; Tom's Hardware on HBM4 HVM].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

- Through mid-2026 Micron made **no HBF commitment**; commentary emphasized that AI flash demand is served by high-performance SSDs and that Micron evaluates new tiers as standards mature [Reported — earnings-call tone; author paraphrase, not a direct quote; no contrary evidence found in verification on Jul 8, 2026].
- **[Author analysis]** Micron has the least incentive to legitimize a NAND tier that could shave HBM demand at the margin while it is gaining HBM share — expect it to join if/when the OCP-hosted standard and xPU design wins make HBF inevitable.

8.5 Kioxia — the adjacent arsenal (XL-Flash, CBA, AI SSD)

- **XL-Flash**: shipping SLC-class low-latency NAND (~3–5 µs reads) — the closest existing media to HBF requirements [Confirmed — Kioxia XL-FLASH literature].
- **BiCS8/BiCS9 with CBA** co-developed with SanDisk — Kioxia co-owns the exact die technology HBF builds on, and shares the JV fabs; despite not being a named HBF party, it is structurally inside the program's supply chain [Confirmed technology/JV facts — Kioxia CBA page; author inference on HBF positioning].
- **AI SSD program:** Kioxia is developing an XL-Flash-based "AI SSD" targeting **>10 million IOPS (512B) with GPU peer-to-peer access, scheduled for release in 2H 2026** [Confirmed — Tom's Hardware, 2025]; beyond that, Kioxia and Nvidia are targeting a **100 million IOPS SSD in 2027** for AI servers [Confirmed — Tom's Hardware, Sept 2025; Blocks & Files, Sept 15, 2025]. Kioxia researchers have also discussed longer-term concepts toward **multi-TB/s-class flash subsystems (~5 TB/s targets around 2027–2029)** [Reported — research-roadmap coverage; treat as directional].

8.6 Huawei — the sanction-driven parallel universe

Blocked from HBM3E-class supply, Huawei attacks the same memory wall from software and SSDs:

- **UnifiedCache Manager (UCM):** KV-cache tiering software (HBM → DRAM → SSD) announced August 12, 2025 (claiming up to 90% latency reduction and up to 22x throughput in long-sequence scenarios) and open-sourced in September 2025 into its ModelEngine ecosystem [Confirmed — Tom's Hardware, Aug 2025; TrendForce, Aug 13, 2025; GitHub: ModelEngine-Group/unified-cache-management].
- **AI SSDs:** Huawei's preparation of a UCM-paired **AI SSD for KV-cache offload** is corroborated by Western trade press [Confirmed in outline — Blocks & Files, Aug 26, 2025], but the specific **HiBL1000 / HiZQ2000 model names and the "eKV" (extended key-value) architecture** could not be corroborated in reachable primary or reputable trade sources during verification on July 8, 2026 [Reported — downgraded; Chinese tech-press provenance only; treat these product details with caution].
- **[Author analysis]** Huawei cannot build HBF (no domestic HBM-class packaging plus leading NAND at scale yet), so it is building HBF's *software layer* first. If Chinese NAND (YMTC) and packaging mature, the UCM stack gives Huawei a fast on-ramp to a domestic HBF analog around 2028+ [Speculation].

8.7 YMTC — the natural hybrid-bonding enabler, sanction-constrained

- **Xtacking 4.0** (in mass production; YMTC's shipping fifth-generation TLC NAND stacks **294 total layers, 232 active**, with hybrid bonding between logic and array): every YMTC die is already a two-wafer hybrid-bonded NAND — array wafer + full CMOS wafer — giving YMTC arguably the most production hybrid-bonding NAND experience on earth, a direct architectural head start for HBF-style parts [Confirmed — Tom's Hardware on Gen5 294-layer shipments; TechInsights Xtacking analysis].
- **Dec 2025:** YMTC publicly disclosed **plans to develop HBF technology**, positioning it as a key strategic direction and citing Xtacking's suitability for AI-accelerator integration [Reported — DigiTimes, Dec 31, 2025. **Update:** the original draft called a YMTC HBF play "Speculation"; it is now company-stated intent, upgraded to Reported.]

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

- Constraints: US entity-list status limits EUV/advanced deposition-etch tool access and advanced packaging partnerships; capacity and layer-count scaling continue domestically, but an export-market HBF play is implausible; a domestic-market HBF analog for Ascend-class accelerators is the scenario to watch [Confirmed sanctions context; scenario timing is Speculation].

8.8 CXMT (context)

CXMT, China's DRAM/HBM developer (DDR5 in volume; HBM2E-class in development), is relevant here as the potential DRAM half of an all-domestic Chinese HBM+HBF stack later in the decade — a linkage made more concrete by a reported **YMTC–CXMT partnership pairing YMTC's hybrid-bonding IP with CXMT's DRAM for domestic HBM production** [Reported — Tom's Hardware; the HBF linkage remains Speculation].

9. Competing and Adjacent Approaches

Approach	What it is	BW to xPU	Latency	Capacity	Threat/complement to HBF
HBM4/4E scaling	More/taller DRAM stacks, custom base dies	2–3+ TB/s/stack	~100 ns	48–96 GB/stack (96 GB = HBM4E 16-high)	Complement (hot tier); caps HBF's value if HBM $/GB falls fast
Nvidia "Storage Next" / GPU-direct AI SSDs	High-IOPS small-block SSDs, GPU P2P access (Nvidia and Kioxia targeting a ~100M-IOPS drive for 2027) [Confirmed — Tom's Hardware, Sept 2025]	~tens of GB/s per drive	~20–100 µs	TBs	Nearest competitor; wins on ecosystem, loses ~10–100x on BW — HBF's job is to make this the *cold* tier
Kioxia AI SSD (XL-Flash)	>10M IOPS SLC SSD, release 2H 2026 [Confirmed]	~tens of GB/s	~5–20 µs	TBs	Adjacent/feeder technology for HBF media
CXL memory expansion	DRAM (or flash-backed) pools over CXL 2.0/3.x	~64–128 GB/s per link	~200–600 ns (DRAM)	TBs (pooled)	Competes for the "warm" tier at node level; loses on BW density, wins on shareability
LPDDR SOCAMM	**Modular LPDDR5X/6 near-package (Micron et al., GB300-class)**	**~hundreds of GB/s aggregate**	**~100+ ns**	**up to ~TB/system**	**DRAM-based warm tier; volatile and pricier per bit than HBF, faster and endurance-free**
KV-offload software (Dynamo/NIXL, LMCache, Mooncake, Huawei UCM)	Tiering/orchestration across whatever hardware exists	n/a	n/a	n/a	Pure complement — this software is HBF's demand generator
Rubin CPX-style split accelerators	Context-phase chips with cheaper GDDR7 (128 GB) for prefill [Confirmed — Nvidia announcement, Sept 9, 2025]. **Update: Nvidia reportedly removed Rubin CPX from its roadmap at GTC in March 2026** [Reported — The Elec]	~2 TB/s (GDDR7, 32 Gbps x 512-bit) [Reported — SemiAnalysis]	~100 ns	128 GB	Shows xPU vendors will re-architect around memory cost — a mindset that favors HBF adoption, though CPX's own fate shows how fast such SKUs can churn

[Author analysis — why HBF sits between HBM and SSD.] The memory pyramid's gap is stark: from HBM (~100 ns, ~$10s/GB, ~100s of GB/node) the next rung today is effectively NVMe (~50+ µs path, <$0.10/GB, ~100s of TB/node) — a 3-decade latency gap and 2-decade cost gap with nothing in between except thin CXL/DRAM layers. Inference monetizes exactly that gap: weights and reusable KV want *terabytes at microseconds*. HBF is the first proposal that fills the gap with the industry's cheapest bits (NAND) at the industry's densest wiring (2.5D shoreline) — which is why, despite all execution risk, it keeps attracting partners rather than losing them.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

10. What the Academic Literature Signals

New in the July 9, 2026 revision. Between December 2025 and mid-2026, HBF crossed a threshold: it stopped being a vendor-slideware concept and became a **subject of peer-reviewed and preprint research**. This section synthesizes an academic literature survey (Shimogoryo Base, July 9, 2026 — *HBF_Literature_Survey_20260709_092527.md*) into predictions about how HBF will be **used, designed, applied, packaged, accessed, and attached** — and, just as informative, examines **who is sponsoring and authoring the work**. Paper claims were verified against the arXiv abstract pages and reachable secondary coverage on July 9, 2026; where a primary page was not machine-readable, the claim is labeled as relying on the survey's summary.

10.1 The papers at a glance

Paper	Authors / sponsor	Venue, date	Core HBF-relevant claim
H3: Hybrid Architecture Using HBM and HBF for Cost-Efficient LLM Inference	SK hynix Research	IEEE journal, Feb 2026	HBM+HBF on one interposer with an NVIDIA Blackwell B200 (8 HBM3E + 8 HBF stacks); weights and precomputed KV in HBF, generated KV in HBM; **2.69x perf/W** vs HBM-only; 6.14x throughput at 10M tokens; 18.8x more simultaneous queries with 2 GPUs instead of 32 [Confirmed — Blocks & Files; IEEE Xplore; SK hynix Research blog]
HAVEN: High-Bandwidth Flash Augmented Vector Engine	Po-Kai Hsu, Weihong Xu, Qunyou Liu, Tajana Rosing, Shimeng Yu (academic)	arXiv 2603.01175, Mar 2026	HBF + near-storage search unit as **on-package complement to HBM**; full-precision billion-scale vector database resides on-device; up to **20x reranking throughput**, up to 40x latency reduction (IVF-PQ ANNS) [Confirmed — arXiv abstract]
MemExplorer: Navigating the Heterogeneous Memory Design Space for Agentic Inference NPUs	Microsoft Research (Junyi Liu, Kai Shi) + academic co-authors incl. Nicholas D. Lane, Timothy Jones, Aaron Zhao, Robert Mullins	arXiv 2604.16007, Apr 2026	HBF modeled alongside SRAM/HBM/LPDDR/GDDR in automated NPU memory design-space exploration for agentic prefill/decode disaggregation; up to **2.3x energy efficiency** vs baseline NPU and **3.23x vs H100** (prefill), 1.93x/2.72x (decode) [Confirmed — arXiv abstract; Microsoft Research page]
NVLLM: A 3D NAND-Centric Architecture Enabling Edge on-Device LLM Inference	Mingbo Hao et al. (Southeast University, Nanjing, China per survey)	arXiv 2604.25699, Apr 2026	FFN weights and compute offloaded **into** flash via wafer-to-wafer stacking; GEMM/GEMV decomposed to dot-product primitives on raw NAND reads; **16.7–37.9x speedup** vs A800 out-of-core inference (OPT/LLaMA to 30B); 2.7% CMOS area overhead [Confirmed — arXiv abstract; affiliation Reported — survey summary]
KVNAND: DRAM-Free In-Flash Computing for On-Device LLM Inference	Lishuo Deng et al. (same research cluster — overlapping authors with NVLLM)	arXiv 2512.03608, Dec 2025	Weights **and KV cache** entirely in compute-enabled 3D NAND; head-group parallelism; **page-level KV mapping aligned to flash organization**; ~2x geomean speedup at 128/1K/10K contexts; resolves OOM at 100K contexts [Confirmed — arXiv abstract]
Challenges and Research Directions for LLM Inference Hardware	Xiaoyu Ma, **David Patterson**	arXiv 2601.05047, Jan 2026	Survey identifying HBF as one of four key architecture research opportunities: **"10X memory capacity with HBM-like bandwidth"** for inference [Confirmed — arXiv abstract]

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

Paper	Authors / sponsor	Venue, date	Core HBF-relevant claim
KAIST TERALAB HBM Roadmap Ver 1.7	KAIST TERALAB, Prof. Joungho Kim	Jun 2025, 371 pages	HBM/HBF architecture roadmap; Prof. Kim predicts **HBF commercialization late 2027–early 2028** and **HBF demand surpassing HBM by 2038** [Commercialization prediction Confirmed — Korea Times, Feb 3, 2026; page count and 2038 crossover Reported — survey summary; full PDF; SemiEngineering overview (unreachable at revision time, HTTP 403)]

Adjacent preprints identified by the survey but not analyzed in depth here: **NASiC** (3D NAND-based CAM-selected multibit CIM for on-device mixture-of-experts inference, arXiv 2605.23294); **Voxel** (3D-stacked AI chip architecture for LLM inference, arXiv 2604.26821); **GNStor** (GPU-native remote all-flash array, arXiv 2606.04908); **ZettaLith** (extreme-scale AI inference acceleration, arXiv 2507.02871) [Reported — survey listing; contents not independently verified].

10.2 How HBF will be used (workloads)

[Author analysis, built on Confirmed paper claims.] The literature converges, from four independent directions, on the read-mostly workload map this report sketched in Sections 1.1 and 6.1:

- **Weights-in-HBF, KV-in-HBM split (H3).** SK hynix Research's H3 puts **read-only data (model weights and precomputed KV caches) in HBF** and **mutable, generated KV in HBM**, with both presented as GPU main memory through a unified address space. That is precisely the "HBF as shared prefix/weight library, HBM as mutable frontier" pattern of Section 6.1 — now with peer-reviewed numbers: 2.69x perf/W, and at 10-million-token contexts, 6.14x higher throughput and 18.8x more simultaneous queries served by 2 GPUs instead of 32 [Confirmed — Blocks & Files, Feb 16, 2026]. The paper frames the target as read-heavy inference such as cache-augmented generation.
- **RAG / vector databases on-device (HAVEN).** HAVEN uses HBF's terabyte-scale capacity to hold a **full-precision, billion-scale vector database on the GPU package**, eliminating PCIe and DDR bottlenecks during reranking, with a near-storage search unit doing ANNS work next to the flash — up to 20x reranking throughput over GPU-DRAM and GPU-SSD baselines [Confirmed — arXiv 2603.01175]. This validates the "embeddings, RAG corpora" entry in this report's workload list, and extends HBF's addressable market beyond LLM serving into retrieval infrastructure.
- **Agentic inference and prefill/decode disaggregation (MemExplorer).** MemExplorer's premise is that **prefilling and decoding devices want different memory systems**, and that agentic workloads (long tool-use loops, heavy context reuse) stress the hierarchy differently again; HBF enters as a capacity/bandwidth/power point in that design space [Confirmed — arXiv 2604.16007]. This is the academic mirror of the Rubin-CPX-style disaggregation trend discussed in Section 9.
- **Long-context KV capacity (KVNAND).** KVNAND observes that KV cache **can exceed model weights in size** as context grows, and demonstrates that flash-resident KV resolves out-of-memory failures at 100K-token contexts on device [Confirmed — arXiv 2512.03608] — quantitative support for the KV-growth argument of Section 1.1.

The prediction that falls out: HBF's first-order use is **weights and reusable prefix/KV**, its second-order use is **retrieval corpora and embeddings**, and its design point will be pulled by **disaggregated, agentic serving** — none of the papers propose HBF as a write-hot working memory, consistent with this report's endurance analysis.

10.3 How HBF will be designed (architecture)

[Author analysis, built on Confirmed paper claims.] Three architectural strata are visible in the literature:

1 **The standardized tier (H3, MemExplorer).** H3 is a straight instantiation of the SanDisk/SK hynix plan of record — HBF stacks in HBM-style sites on one interposer, with a **latency-hiding buffer** to mask flash access delays and a unified address space spanning HBM and HBF [Confirmed — Blocks & Files]. The buffer is notable: it is the system-level precursor

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

of the DRAM-cached hybrid stack discussed in Section 5.2, and it confirms Section 4.2's expectation that meaningful controller/caching logic sits between the xPU and raw NAND semantics. MemExplorer, meanwhile, signals that memory-hierarchy composition — how many sites of what — is becoming an **automated co-optimization variable** alongside NPU microarchitecture rather than a fixed platform decision [Confirmed — arXiv 2604.16007]; this is exactly the per-SKU HBM/HBF site-allocation flexibility predicted in Sections 4.2 and 7.2.

2 **Near-storage compute beside HBF (HAVEN).** HAVEN pairs HBF with a near-storage search unit so vector-distance work happens without hauling data to the GPU cores [Confirmed — arXiv 2603.01175]. This rhymes with the June 2026 SanDisk patent bonding a processor onto a NAND tile (Section 8.1) — the literature and vendor IP are moving toward the same "compute at the flash" endpoint.

3 **In-flash compute (NVLLM, KVNAND, NASiC).** The most radical stratum abandons the memory-device abstraction entirely: NVLLM integrates compute pipelines, ECC, and buffers with multi-plane 3D NAND via wafer-to-wafer stacking and executes GEMM/GEMV as dot-product primitives on raw NAND reads at 2.7% CMOS area overhead; KVNAND goes DRAM-free with both weights and KV in compute-enabled NAND [Confirmed — arXiv 2604.25699; arXiv 2512.03608]. These are edge/on-device designs, not HBF-standard devices — but they map the long-run design space the HBF base die could evolve into, and they demonstrate that the bonded-CMOS wafer (Section 2.2) is where the industry will be tempted to put compute, not just periphery.

10.4 Where HBF will be applied (market segments)

[Author analysis, built on Confirmed paper claims.] The papers partition cleanly by segment, and the partition itself is a market forecast:

- **Datacenter LLM inference (H3):** the headline economics — 2.69x performance per watt and an 18.8x consolidation of concurrent queries onto 16x fewer GPUs — are datacenter TCO arguments, published by the vendor that sells both HBM and (prospectively) HBF [Confirmed — Blocks & Files].
- **Hyperscaler NPUs (MemExplorer):** Microsoft co-authorship places HBF inside the design methodology of a first-party accelerator program; the paper benchmarks against H100 and references NVIDIA's Vera Rubin platform as a comparison point [Confirmed — Microsoft Research page].
- **Retrieval/RAG infrastructure (HAVEN):** billion-scale ANNS on-device targets the vector-database layer of the AI stack — a segment this report had treated only implicitly [Confirmed — arXiv 2603.01175].
- **Edge and on-device LLM (NVLLM, KVNAND, NASiC):** the in-flash-compute cluster targets resource-constrained devices running up to ~30B-parameter models without datacenter memory budgets [Confirmed — arXiv 2604.25699; arXiv 2512.03608].

Prediction: HBF-standard devices land first in **datacenter inference and hyperscaler NPUs** (where H3-class economics and MemExplorer-class design tooling already exist), while the **edge** is served by non-standard NAND-centric derivatives before any "mobile HBF" appears.

10.5 How HBF will be packaged, attached, and accessed

[Author analysis, built on Confirmed paper claims.]

- **Packaged.** H3 co-packages 8 HBM3E + 8 HBF stacks with a Blackwell B200 on a single interposer [Confirmed — Blocks & Files] — the CoWoS-class 2.5D model of Section 4.3, at the high end of the site counts discussed in Section 5.1. NVLLM's **wafer-to-wafer stacking** of NAND with compute pipelines [Confirmed — arXiv 2604.25699] shows the 3D endgame beyond 2.5D. KAIST's roadmap work frames HBF as vertically stacked NAND explicitly reusing manufacturing techniques developed for HBM [Confirmed — Korea Times].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

- **Attached.** H3 is the first published validation of the **same-interposer, HBM-site attach model with an NVIDIA Blackwell-class GPU** — the plan of record of Section 5.1, modeled by the co-standardizer itself. HAVEN independently treats HBF as an **on-package complement to HBM** rather than a replacement [Confirmed — arXiv 2603.01175] — the complement framing of Section 6.1, arrived at by authors outside the vendor camp.
- **Accessed.** Every paper assumes **read-mostly semantics**: H3 routes writes (generated KV) to HBM and reads (weights, precomputed KV) to HBF; HAVEN's vector database is a read-dominated corpus with search **offloaded near the storage**; KVNAND introduces **page-level KV cache mapping that aligns token access patterns with flash organization** and head-group parallelism to extract read bandwidth [all Confirmed — sources above]. The access-semantics prediction: the HBF protocol layer being standardized at OCP (Section 7.2) will be judged on how well it exposes page-granular, variable-latency, massively parallel reads — the exact properties these papers engineer around — and KVNAND's page-alignment result suggests the **software mapping layer** (page/codeword-aware tensor layout) will matter as much as the PHY.

10.6 The KAIST TERALAB roadmap: an academic timeline for HBF

Prof. Joungho Kim of KAIST — widely called the "father of HBM" for his pioneering role in the technology, and originator of "Kim's Law" on 3D-stacking cadence — leads TERALAB, whose roughly 80 alumni populate Samsung, SK hynix, Apple, and Google [Confirmed — Korea Times, Feb 3, 2026; TERALAB]. His lab's HBF-relevant output:

- **HBM Roadmap Ver 1.7** (June 2025): a comprehensive next-generation memory architecture roadmap reported at **371 pages**, covering HBM generations and HBF [Reported — survey summary; full PDF; the SemiEngineering overview was unreachable (HTTP 403) at revision time, and the TERALAB milestone page reachable on July 9, 2026 reflected an earlier HBM3–HBM5 era without HBF content].
- **2026 HBF workload analysis and roadmap seminar** (February 2026): a multi-part recorded seminar on HBF workload analysis and technology roadmap [Reported — survey summary; Part 1, Part 2, Part 4].
- **Predictions:** HBF commercialized **late 2027 to early 2028**, accelerated by HBM-derived expertise, with Samsung and SK hynix collaborating with Nvidia, Google, and AMD toward that timeline; HBM as "a bookshelf near the GPU" for speed and HBF as "a library near the GPU" for capacity; and "memory will become even more important than GPUs" [Confirmed — Korea Times]. He further predicts **HBF demand surpassing HBM by 2038** [Reported — survey summary; treat as an [Analyst speculation]-grade long-range projection].

[Author analysis] Kim's late-2027/early-2028 commercialization window sits almost exactly on this report's independently derived timeline (samples 2H 2026 → device samples early 2027 → first meaningful revenue 2028; Section 11, item 8) — convergence from an authority with deep visibility into the Korean memory complex. The 2038 demand-crossover claim should be handled with far more caution: it is a 12-year extrapolation of inference-capacity demand, but even discounted it signals that the person who called HBM's trajectory correctly believes HBF is a *decade-scale* platform, not a niche SKU. His bookshelf/library metaphor is also, usefully, the vendor-neutral restatement of this report's pyramid-rung thesis (Section 1.1).

10.7 Who sponsors the research — and what that reveals

[Author analysis, on confirmed authorship facts.] In an emerging device class, *who writes the papers* is as much a market signal as *what the papers say*:

- **H3 is authored by SK hynix Research** — the HBF co-standardizer publishing peer-reviewed validation of the hybrid HBM+HBF model on an NVIDIA GPU [Confirmed authorship — SK hynix Research blog; Blocks & Files]. This is **vendor-sponsored evidence for the plan of record**: the numbers should be read as a best-case advocacy result, but the paper's existence is itself disclosure — it reveals SK hynix's **design intent for AIN B** (Section 8.2), including the latency-hiding buffer, the unified address space, and the weights/precomputed-KV placement policy. Vendors do not usually publish architecture papers about products they intend to abandon.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

- **MemExplorer is a Microsoft Research collaboration** (Microsoft Research authors Junyi Liu and Kai Shi, with a large academic co-author team including Nicholas D. Lane, Timothy Jones, Aaron Zhao, and Robert Mullins) [Confirmed — Microsoft Research page]. This is the **demand-side signal** this report has been watching for: a hyperscaler that designs its own NPUs (Maia-class) is already modeling HBF as a first-class option in its memory-hierarchy design tooling. It bears directly on Risk 5 (Section 11) — "xPU design-in is the gating item" — because it shows at least one major xPU customer doing the design-space homework, even though it is not a product commitment.
- **The SanDisk-published interview with Prof. Joungho Kim** [Reported — survey summary; SanDisk blog — page content not machine-readable at revision time] shows the HBF inventor **cultivating academic legitimacy from the "father of HBM"** — a deliberate echo of how HBM itself was legitimized. TERALAB is not a disinterested party in the conventional sense: it is deeply woven into the Korean memory ecosystem (≈80 alumni at Samsung, SK hynix, Apple, Google [Confirmed — Korea Times]), which makes its roadmap simultaneously less independent and *more* predictive — it reflects what the Korean memory complex intends to build.
- **NVLLM and KVNAND come from a Chinese academic cluster** — NVLLM from Southeast University, Nanjing [Reported affiliation — survey summary], with overlapping authors across the two papers (Changwei Yan and Weiwei Shan appear on both [Confirmed — arXiv author lists, 2604.25699, 2512.03608]), and NASiC extending the same in-flash-compute theme. A sustained **China-academic in-flash-compute research program** is exactly the intellectual feedstock the YMTC/Huawei domestic-HBF-analog scenario (Sections 8.6–8.7) would draw on: sanction-constrained access to HBM pushes Chinese research toward making NAND itself do more, and wafer-to-wafer stacking is YMTC's home turf.
- **The independent survey is less independent than it looks — instructively so.** arXiv 2601.05047, which names HBF one of four key architecture research opportunities ("10X memory capacity with HBM-like bandwidth"), is co-authored by **David Patterson** [Confirmed — arXiv 2601.05047] — the same Turing laureate who sits on **SanDisk's HBF Technical Advisory Board** (Section 8.1). The endorsement is genuinely authoritative and the survey venue is neutral, but readers should know the field's most-cited independent validation is one degree of separation from the HBF camp.
- **Who is absent is also informative.** No NVIDIA-, AMD-, or Broadcom-authored HBF paper had surfaced as of this revision; the demand side speaks only indirectly — through Microsoft's design-space modeling and through SK hynix's choice of a Blackwell B200 as H3's host platform. Samsung and Micron are likewise absent from the HBF literature, consistent with their standoff postures (Sections 8.3–8.4). The first NVIDIA/AMD-co-authored HBF paper, or a Samsung HBF architecture paper, would be a stronger adoption signal than most press releases — add it to the watch list (Section 11, item 9).

[Author analysis — synthesis.] The sponsorship map recapitulates the HBM playbook: *vendor publishes the validating architecture paper (H3), the standard-bearer academic publishes the roadmap (KAIST), a hyperscaler quietly tools up (MemExplorer), and adjacent academia radicalizes the concept (in-flash compute).* That is what a memory class looks like roughly two years before volume — and it is also, cautionarily, what Optane's literature looked like. The difference to watch: HBF's papers are about *inference economics* (a priced, growing market), not about a device in search of a workload.

11. Risks and Open Questions

1. **Endurance vs KV churn.** SLC endurance (~50–100K P/E) is generous, but a mismanaged KV tier could still chew through it; weights-only usage is trivially safe. The spec must expose wear telemetry, and software must confine writes to demotion events. Open question: what write-bandwidth and DWPD-equivalent will the standard guarantee? [Author analysis; no public endurance spec as of July 2026.]
2. **Read latency reality.** Microsecond reads are fine for streaming weights (bandwidth-bound) and bulk KV restore, but poor for fine-grained random access; accelerators need prefetch/scratchpad discipline (or a DRAM cache layer in-stack) to hide it. If real workloads turn out more latency-sensitive than modeled, HBF's addressable share shrinks toward "weights only"

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

[Author analysis]. SK hynix's **H3** offers the first published evidence that this is manageable: its architecture interposes a latency-hiding buffer in front of HBF and still delivers 2.69x perf/W on read-heavy inference [Confirmed — Blocks & Files, Feb 16, 2026] (Section 10.3).

3 **Thermals.** NAND dislikes heat in a specific way: hot programming/cold read (and vice versa) cross-temperature effects degrade margins, retention worsens exponentially with temperature, and an HBF stack will sit millimeters from a 1,000W+ compute die under the same lid. HBM already throttles for refresh at ~95 °C; HBF will need aggressive thermal isolation, temperature-aware refresh/scrub of flash, and possibly rated hot-operation modes. This is, in the author's view, the most underdiscussed engineering risk [Author analysis; vendors have acknowledged thermal work without publishing limits].

4 **Standardization path — resolved in February 2026, but adoption still open.** Earlier drafts of this report (and much 2025 commentary) assumed the SanDisk–SK hynix MoU would feed **JEDEC** (where HBM lives). That is not what happened: on **February 25, 2026 the two companies launched an "HBF Spec. Standardization Consortium" and a dedicated workstream under the Open Compute Project (OCP)** to standardize HBF globally [Confirmed — SanDisk press release, Feb 25, 2026; SK hynix newsroom; VideoCardz]. OCP is hyperscaler-native and faster-moving than JEDEC, which suits HBF's AI-datacenter audience; but without Samsung/Micron formally aboard the workstream and at least one major xPU vendor publicly committed, HBF still risks being a two-vendor interface — historically a fragile position (cf. Z-NAND, XPoint) [historical analogy is author analysis]. Samsung's reported independent HBF development (Section 8.3) cuts both ways: it validates the tier while risking spec fragmentation.

5 **xPU design-in is the gating item.** HBF only ships if Nvidia/AMD/Broadcom-class designers burn controller area and shoreline on it. SanDisk has said HBF was developed "with input from leading AI industry players" [Reported — vendor assertion in the Aug 6, 2025 MoU release], and the early-2027 "first inference devices with HBF" sampling claim implies at least one committed design — identity undisclosed as of July 8, 2026 [Confirmed roadmap claim — same release; partner unknown]. Two literature signals now bear on this risk: SK hynix's H3 was modeled around an **NVIDIA Blackwell-class B200** [Confirmed — Blocks & Files], and **Microsoft Research co-authored MemExplorer**, which already prices HBF into hyperscaler NPU memory-hierarchy design space [Confirmed — Microsoft Research page] — neither is a commitment, but both show demand-side engineering work underway (Section 10.7).

6 **Ecosystem software.** Compilers, runtimes, and serving stacks must place tensors by tier; CUDA-class toolchains have no public HBF abstraction yet. The Dynamo/UCM precedents help, but on-package flash with wear constraints is a new OS-level object [Author analysis].

7 **Cost trajectory competition.** If HBM4E/5 per-bit cost falls faster than expected (or capacity-optimized 16-high HBM gets cheap), HBF's window narrows; conversely a NAND downcycle widens it. HBF's economics are cycle-dependent in a way HBM's demand is not [Author analysis].

8 **Realistic adoption timeline.** [Author analysis] Samples 2H 2026 (SanDisk-stated, reaffirmed Feb 2026) → first inference-device samples early 2027 (stated) → **first meaningful revenue silicon 2028** → mainstream inference attach and a possible OCP-standard second generation **2029–2030**. Slips of 6–12 months from stated dates would be historically normal for a new memory class; note that analyst commentary has cautioned HBF volume demand may not accelerate until ~2030 [Reported — Blocks & Files, Nov 27, 2025; TrendForce, Feb 26, 2026]. Independently, Prof. Joungho Kim (KAIST) predicts HBF commercialization in **late 2027–early 2028** [Confirmed — Korea Times, Feb 3, 2026] — squarely consistent with this timeline; his lab's longer-range projection that **HBF demand surpasses HBM by 2038** should be treated as an [Analyst speculation]-grade extrapolation [Reported — survey summary] (Section 10.6).

9 **What to watch in H2 2026.** [Author analysis — updated at verification, July 8, 2026] (a) SanDisk actually shipping HBF samples in the now-open 2H 2026 window (no shipment announced as of July 8, 2026); (b) output and membership growth of the **OCP HBF workstream** launched February 2026 — in particular any formal Samsung or Micron engagement (Samsung is so far reported to be developing HBF independently, Section 8.3); (c) SK hynix AIN P sampling by end of 2026 (stated); (d) disclosure of the xPU design win implied by the early-2027 device-sample claim — still undisclosed as of July 8, 2026; (e) *[added July 9, 2026]* the academic pipeline — a first NVIDIA/AMD-co-authored HBF paper, a Samsung HBF architecture paper, or follow-on SK hynix Research publications extending H3 toward AIN B productization (Section 10.7).

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

12. Conclusion and Outlook Table

12.1 Vendor status matrix (as of July 2026)

Vendor	HBF stance	Key asset	Public milestones	Timeline signals
SanDisk	Inventor, evangelist	BiCS + CBA, JV fabs, HBF IP, advisory board (Patterson, Koduri)	HBF unveil Feb 2025; TAB Jul 2025; FMS 2025 Best of Show; MoU w/ SK hynix Aug 2025; OCP standardization kick-off Feb 2026	Samples 2H 2026; inference devices w/ HBF early 2027 [stated, reaffirmed Feb 2026; no sample shipment announced as of Jul 8, 2026]
SK hynix	Co-standardizer	HBM leadership, MR-MUF/TSV, in-house DRAM+NAND	MoU Aug 2025; AIN family (P/B/D) at OCP Oct 2025; H3 hybrid HBM+HBF paper (IEEE) Feb 2026; OCP HBF workstream Feb 2026	AIN B (HBF) spec work 2026; AIN P samples by end 2026 [stated]
Kioxia	Adjacent (not a named party)	XL-Flash, CBA co-owner, JV fabs	AI SSD >10M IOPS (2H 2026); 100M-IOPS SSD w/ Nvidia (2027)	De facto inside HBF supply chain via JV [author inference]
Samsung	Reportedly developing own HBF	Bonded V-NAND (~V10), HBM, I-Cube, TSV NAND	Early-stage HBF concept design reported Oct–Nov 2025	Commercial HBF target ~2027 [reported]; OCP-standard alignment open
Micron	HBM-first, no HBF commitment	HBM4/4E (HVM for Vera Rubin), G9 NAND, SOCAMM	AI flash = fast SSDs, for now	Fast follower at best [speculation]
Huawei	Software analog + AI SSDs	UCM tiering stack; AI SSD for KV offload (model names unverified)	UCM announced Aug 2025, open-sourced Sept 2025	Domestic HBF analog possible 2028+ [speculation]
YMTC	**Declared HBF development plans (Dec 2025)**	**Xtacking 4.0 hybrid bonding at scale**	**HBF named a key strategic direction [DigiTimes, Dec 2025]**	**Domestic-market HBF later in decade [reported intent; timing speculative]**
Nvidia/AMD (demand side)	Undeclared	Shoreline, controllers, software	"Storage Next"-class SSD push; Rubin CPX (Sept 2025; reportedly dropped at GTC Mar 2026) shows memory-cost re-architecting — and churn	The real kingmakers [author analysis]

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

12.2 Final assessment

[Author analysis] HBF is the most credible new memory-class proposal since HBM itself, for three reasons. First, it is **demand-pulled**: inference economics have created a priced, growing market for terabytes at microseconds that nothing currently serves. Second, it is **technologically evolutionary**: every ingredient — CBA/hybrid-bonded NAND, TSV stacks, HBM PHYs, 2.5D interposers, KV-tiering software — is already in production somewhere; HBF is an integration, not an invention. Third, it made the **right strategic compromise**: reusing HBM's interface and shoreline rather than demanding new beachfront, which converts the adoption question from "will xPUs redesign for flash?" to "how many HBM sites is a terabyte worth?" To these three, the first half of 2026 added a **fourth credibility signal**: an academic ecosystem is now forming around HBF — vendor-sponsored peer-reviewed validation (SK hynix's H3), hyperscaler design-space modeling (Microsoft's MemExplorer), the KAIST TERALAB roadmap from HBM's own progenitor, and independent survey work naming HBF a first-order research opportunity — publishing quantitative evidence for the tier before first silicon even ships (Section 10).

The risks are equally clear: unproven thermals of NAND-on-a-kilowatt-package, endurance policy, a spec that — although now hosted in an open OCP workstream (February 2026) — Samsung and Micron have not yet formally joined (Samsung is reportedly building its own HBF instead), and the absence (publicly) of a committed xPU design win as of July 2026. The sober forecast: **samples in 2026, first devices 2027, judgment day 2028** — when we learn whether High Bandwidth Flash becomes the pyramid's missing rung or joins Optane in the museum of beautiful tiers. The presence of the NAND price umbrella and the HBM incumbent on the same MoU — now joined by an open standardization venue and reported entries from Samsung and YMTC — suggests, this time, the rung gets built.

Appendix A: Principal sources referenced

Note: An earlier draft of this report omitted URLs because live web verification was unavailable during preparation. Live verification was performed on July 8, 2026; the anchor documents below now carry their URLs, and inline citations were added throughout the body. See Appendix B for the claim-by-claim verification log.

- SanDisk HBF introduction, February 2025 Investor Day — coverage: Tom's Hardware; TrendForce, Feb 14, 2025; SanDisk blogs: Memory-Centric AI, Scaling the Memory Wall.
- SanDisk HBF Technical Advisory Board press release (Patterson, Koduri), July 24, 2025.
- SanDisk–SK hynix MoU to standardize High Bandwidth Flash, August 6, 2025 — coverage: Tom's Hardware; Blocks & Files.
- SanDisk–SK hynix HBF Spec. Standardization Consortium kick-off and **OCP workstream**, February 25, 2026 — SanDisk press release; SK hynix newsroom; TrendForce.
- FMS 2025 Best of Show awards — FMS official winners list; StorageNewsletter; technical synthesis: SemiEngineering.
- SK hynix AIN P / AIN B / AIN D at OCP Global Summit, October 2025 — SK hynix newsroom; Tom's Hardware.
- Kioxia: CBA technology page; BiCS8 announcement with Western Digital, March 2023; XL-FLASH infographic; AI SSD: Tom's Hardware (10M IOPS, 2H 2026), Tom's Hardware (100M IOPS w/ Nvidia, 2027).
- Micron: HBM4 36GB 12-high sample shipments, June 10, 2025; HBM4 product page; HBM4 high-volume production for Vera Rubin (Tom's Hardware).
- SK hynix–TSMC HBM4 base-die collaboration, April 2024; JEDEC JESD270-4 HBM4 standard, April 16, 2025.
- Samsung HBF posture: DigiTimes, Oct 7, 2025; TrendForce, Nov 11, 2025; The Investor, Feb 4, 2026.
- Huawei UCM: Tom's Hardware, Aug 2025; TrendForce; open-source repository; AI SSD: Blocks & Files, Aug 26, 2025.
- NVIDIA: Rubin CPX announcement, September 9, 2025; The Elec on Rubin CPX roadmap removal, 2026.
- YMTC: Xtacking unveiling at FMS 2018 (PRNewswire); TechInsights Xtacking 3.0 blog; TechInsights Xtacking 4.0 analysis; DigiTimes on YMTC HBF plans, Dec 31, 2025; Tom's Hardware on YMTC–CXMT HBM partnership.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

- Analyst caution on HBF timing: Blocks & Files, Nov 27, 2025; SanDisk beyond-HBF patent: TrendForce, Jun 22, 2026.
- **Academic literature (see Section 10)** — from the HBF literature survey of July 9, 2026 (*HBF_Literature_Survey_20260709_092527.md*):
  - H3: Hybrid Architecture Using HBM and HBF for Cost-Efficient LLM Inference (SK hynix Research, IEEE 2026) — IEEE Xplore; SK hynix Research blog summary; ResearchGate; coverage: Blocks & Files, Feb 16, 2026.
  - HAVEN: High-Bandwidth Flash Augmented Vector Engine for Large-Scale ANNS Acceleration — arXiv 2603.01175.
  - MemExplorer: Navigating the Heterogeneous Memory Design Space for Agentic Inference NPUs (Microsoft Research collaboration) — arXiv 2604.16007; Microsoft Research page.
  - NVLLM: A 3D NAND-Centric Architecture Enabling Edge on-Device LLM Inference — arXiv 2604.25699.
  - KVNAND: Efficient On-Device LLM Inference Using DRAM-Free In-Flash Computing — arXiv 2512.03608.
  - Challenges and Research Directions for LLM Inference Hardware (Ma & Patterson) — arXiv 2601.05047.
  - KAIST TERALAB HBM Roadmap Ver 1.7 (June 2025) and HBF roadmap coverage — full PDF; SemiEngineering overview; TERALAB research page; Korea Times, Feb 3, 2026; 2026 HBF workload seminar: Part 1, Part 2, Part 4; BigGo Finance coverage; SanDisk interview with Prof. Joungho Kim.
  - Adjacent: NASiC — arXiv 2605.23294; Voxel — arXiv 2604.26821; GNStor — arXiv 2606.04908; ZettaLith — arXiv 2507.02871.

Appendix B: Verification log (July 8, 2026)

All load-bearing claims were live-checked on **2026-07-08** using vendor primary sources plus one reputable trade source per claim where available. Verdicts: **Verified** (primary/multiple sources agree), **Partially verified** (core fact holds, detail adjusted or nuanced), **Could not verify** (no reachable corroboration), **Corrected** (body text changed).

#	Claim (as originally drafted)	Verdict	Source URL(s)	Date checked	Notes
1	SanDisk introduced HBF at Feb 2025 Investor Day, post-WD-spin-off	Partially verified	Tom's Hardware; TrendForce	2026-07-08	Introduction at "Future FWD" Investor Day, Feb 2025 confirmed; wording adjusted — event slightly preceded formal completion of the WD split rather than following it. Body [Confirmed] labels apply to the restated (verified) wording, not the original claim
2	~512 GB per HBF stack; 8–16x HBM capacity at comparable cost	Verified	SanDisk MoU PR; SanDisk blog	2026-07-08	16-high stack of 32 GB (256 Gb-class) dies; 1.6 TB/s read bandwidth also disclosed — added to body
3	SanDisk–SK hynix MoU signed August 2025	Verified	SanDisk PR, Aug 6, 2025; Blocks & Files	2026-07-08	Exact date Aug 6, 2025
4	HBF won Best of Show / Most Innovative Technology at FMS 2025	Verified	FMS winners list; StorageNewsletter	2026-07-08	Category: "Most Innovative Technology – NAND Flash Solution"
5	Roadmap: HBF samples 2H 2026; first AI inference devices early 2027	Verified	SanDisk MoU PR; TrendForce, Feb 26, 2026	2026-07-08	Schedule reaffirmed at Feb 25, 2026 OCP kick-off; no sample shipment announced as of check date (2H 2026 window just opened); analysts flag volume demand ~2030
6	HBF Technical Advisory Board (Patterson, Koduri) formed Feb 2025	Corrected	SanDisk PR, Jul 24, 2025	2026-07-08	Board formed **July 24, 2025**, not Feb 2025; membership (Patterson, Koduri) confirmed; body text corrected
7	SK hynix AIN P/B/D unveiled at OCP Oct 2025; AIN P sampling ~late 2026	Verified	SK hynix newsroom; Tom's Hardware	2026-07-08	OCP Global Summit Oct 13–16, 2025; AIN P samples "by end of 2026"; AIN B = HBF; AIN D petabyte-class QLC

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

#	Claim (as originally drafted)	Verdict	Source URL(s)	Date checked	Notes
8	HBF drop-in electrical compatibility with HBM sites	Partially verified	Tom's Hardware, Feb 2025	2026-07-08	"Same electrical interface... some protocol changes" confirmed; coverage explicitly notes HBF is *not* a literal drop-in for HBM (controller protocol changes required); pinout/footprint/power/stack-height match HBM4. Body wording tightened to the verified form, so its [Confirmed] label refers to the restated claim
9	Kioxia/SanDisk BiCS8 CBA architecture	Verified	Kioxia CBA page; TechInsights	2026-07-08	218-layer BiCS8 with CMOS directly Bonded to Array confirmed
10	Kioxia XL-Flash ~3–5 µs reads; AI SSD >10M IOPS 2026–27	Verified	Kioxia XL-FLASH; Tom's Hardware	2026-07-08	AI SSD release scheduled 2H 2026; new: Nvidia–Kioxia 100M-IOPS target for 2027 added to body
11	YMTC Xtacking hybrid bonding since 2018; Xtacking 4.0 current	Verified	YMTC PR at FMS 2018; TechInsights Xtacking 3.0 blog; TechInsights Xtacking 4.0	2026-07-08	Xtacking unveiled at FMS **August 2018**, Xtacking 2.0 in 2019 (TechInsights); YMTC first NAND maker to hybrid-bond (TrendForce — note TrendForce dates broad adoption to ~2021); Xtacking 4.0 in mass production — Gen5 TLC at 294 total / 232 active layers (Tom's Hardware). New: YMTC declared HBF development plans (DigiTimes, Dec 2025)
12	HBM3E ~1.0–1.2 TB/s, 24–36 GB; HBM4 ~2.0–2.8 TB/s, 48–64 GB; JEDEC HBM4 April 2025	Verified	JEDEC JESD270-4 PR; Tom's Hardware	2026-07-08	JESD270-4 published April 16, 2025: 2 TB/s at 8 Gb/s, up to 64 GB (16-high, 32 Gb); vendor parts exceed 2.8 TB/s at 11+ Gb/s
13	SK hynix–TSMC HBM4 base-die collaboration, April 2024	Verified	SK hynix newsroom	2026-07-08	MoU announced April 2024; TSMC N12/N5-class base-die processes
14	Micron HBM4 36 GB 12-high sampling 2025 (~2.8 TB/s); no HBF commitment	Verified	Micron PR, Jun 10, 2025; Micron HBM4 page	2026-07-08	Initial samples ~2.0 TB/s; >2.8 TB/s demonstrated later in 2025; HVM for NVIDIA Vera Rubin in 2026; no HBF commitment found as of check date
15	Samsung: no HBF-branded announcement; bonded V-NAND ~V10	Partially verified / updated	DigiTimes, Oct 7, 2025; TrendForce, Nov 11, 2025	2026-07-08	Still no HBF-branded product announcement, but Samsung reportedly began early-stage HBF concept design (Oct–Nov 2025), targeting ~2027 — posture change incorporated in Section 8.3
16	Huawei UCM announced/open-sourced Aug 2025; HiBL1000/HiZQ2000 eKV SSDs	Partially verified	Tom's Hardware; GitHub; Blocks & Files	2026-07-08	UCM verified (announced Aug 12, 2025; open-sourced Sept 2025). AI SSD program corroborated in outline, but **HiBL1000/HiZQ2000 model names and "eKV" architecture could not be verified** in reachable sources — downgraded with caution note
17	NVIDIA Rubin CPX announced Sept 2025 (128 GB GDDR7, prefill-specialized)	Verified / updated	NVIDIA newsroom; The Elec	2026-07-08	Announcement (Sept 9, 2025) verified. **Post-horizon update:** Rubin CPX reportedly removed from NVIDIA's roadmap at GTC March 2026 — noted in Section 9. **Corrected:** draft's "~1.7 TB/s" per-GPU GDDR7 figure was a conflation with NVIDIA's rack-level 1.7 PB/s number; per-GPU bandwidth is ~2 TB/s (32 Gbps x 512-bit) per SemiAnalysis
18	Standardization: "JEDEC-style / no JEDEC output public as of early 2026"	Corrected	SanDisk PR, Feb 25, 2026; SK hynix newsroom	2026-07-08	Standardization is proceeding via a **dedicated OCP workstream** (consortium kick-off Feb 25, 2026, Milpitas), not JEDEC; the Feb 2026 releases do not mention JEDEC. All JEDEC-speculation passages (Sections 7.2, 8.3, 11, 12 — numbered 10 and 11 prior to the July 9, 2026 revision) rewritten

Important Disclosures & Disclaimer

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided "AS IS."

The following disclosures and terms of use apply to this report in its entirety, including all sections, tables, figures, and appendices (collectively, the "Report"). By accessing, reading, or otherwise using the Report, you acknowledge that you have read, understood, and agreed to be bound by these terms.

Publisher; copyright and permitted use

The Report is published by **Shimogoryo Base LLC, a Colorado limited liability company** (the "Publisher"). © 2026 Shimogoryo Base LLC, a Colorado limited liability company.

Permission to share. This is a free public edition. The Publisher grants any recipient a worldwide, royalty-free, non-exclusive licence to **reproduce, repost, forward, transmit, and publicly share the Report**, on any platform and at no charge, on three conditions: (i) it is shared **complete and unaltered**, including this disclosures section; (ii) **Shimogoryo Base LLC is credited** as the publisher and author; and (iii) it is **not sold, licensed for a fee, or placed behind a paywall**.

What is not granted. This permission does not extend to publishing modified, abridged, or derivative versions; to reproducing individual tables, figures, or passages in isolation in a way that changes their meaning or strips their confidence labels and stated caveats; to presenting the analysis as your own or as that of any other person or firm; or to incorporating the Report into any database, information-retrieval system, or machine-learning training corpus. Except as expressly permitted here, all rights are reserved. For commercial reuse, derivative works, syndication, or reprint rights, contact the Publisher via shimogoryobase.com/consult.html.

Companion products are not covered. The HBF investment memorandum and the `HBF_Investment_Screen.xlsx` workbook are separate products and are **not** included in this permission.

Informational purposes only; no professional advice

The Report is provided for general informational purposes only. Nothing in the Report constitutes, or should be construed as, investment advice, financial advice, trading advice, legal advice, tax or accounting advice, or engineering, design, procurement, or product-qualification advice, nor is it a substitute for professional advice tailored to your particular circumstances. You should not rely on the Report as the sole basis for any investment, procurement, design, or other business decision, and you should consult your own qualified advisers before acting on any information contained herein.

No offer or solicitation

Nothing in the Report constitutes an offer to sell, or the solicitation of an offer to buy, any security, financial instrument, or other asset, in any jurisdiction. The Report is not, and should not be treated as, a recommendation regarding the securities of any company mentioned herein — including, without limitation, SanDisk, SK hynix, Samsung, Micron, Kioxia, NVIDIA, AMD, Broadcom, Microsoft, Huawei, YMTC, or CXMT. The Publisher is not a registered investment adviser, broker-dealer, or fiduciary to any recipient of the Report.

Positions and conflicts of interest

Position disclosure. As of the publication date the Publisher holds long positions in SKHY, KLIC, RMBS, and MU, and no position in any other security mentioned.

The Publisher, its members, managers, officers, contractors, and analysts may acquire or dispose of positions in the securities of companies discussed in the Report at any time after the publication date and without notice, including positions inconsistent with the analysis presented, and undertake no obligation to disclose subsequent transactions or to update this disclosure.

Commercial interest. No compensation has been received from any company mentioned in the Report in exchange for its preparation, its conclusions, or its distribution. The Publisher does, however, have a commercial interest in the distribution of this free edition: Shimogoryo Base LLC sells research and consulting services, and this Report is published in part to demonstrate the firm's work and attract clients. Readers should weigh it accordingly.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided “AS IS.”

Sources; accuracy and completeness

The Report is based on publicly available information — vendor press releases and announcements, conference disclosures, trade-press reporting, academic preprints and publications — together with the Publisher's own analysis, obtained from sources believed to be reliable as of the dates indicated. The Publisher has not independently verified all third-party information and makes no representation or warranty, express or implied, as to the accuracy, completeness, timeliness, merchantability, or fitness for a particular purpose of any information contained in the Report. The confidence labels used throughout the Report ([Confirmed], [Reported], [Analyst speculation], [Author analysis]) reflect the Publisher's sourcing judgment at the time of verification and are not guarantees of accuracy.

Forward-looking statements

The Report contains forward-looking statements, projections, scenarios, timelines, and estimates — including, without limitation, statements regarding vendor product roadmaps, sampling and commercialization schedules, standardization outcomes, market adoption, pricing, cost trajectories, and competitive strategy. Forward-looking statements are inherently uncertain; they are based on assumptions that may prove incorrect, involve known and unknown risks outside the Publisher's control, and actual outcomes may differ materially from those expressed or implied. Such statements speak only as of the publication date, and the Publisher undertakes no obligation to publicly update or revise them, whether as a result of new information, future events, or otherwise.

Third-party trademarks

All product names, company names, logos, brands, and trademarks referenced in the Report — including, without limitation, SanDisk, Western Digital, SK hynix, Samsung, Micron, Kioxia, XL-FLASH, BiCS, NVIDIA, Blackwell, Rubin, CUDA, AMD, Broadcom, Microsoft, Huawei, YMTC, Xtacking, TSMC, CoWoS, Intel, JEDEC, and Open Compute Project (OCP) — are the property of their respective owners. Such names and marks are used solely for identification, reference, and editorial purposes. Their use does not imply any affiliation with, sponsorship by, or endorsement by the respective trademark owners, and no such affiliation, sponsorship, or endorsement should be inferred. No company mentioned in the Report has reviewed, approved, sponsored, or endorsed the Report.

No warranty; limitation of liability

THE REPORT IS PROVIDED "AS IS" AND "AS AVAILABLE," WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING, WITHOUT LIMITATION, ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE, ACCURACY, OR NON-INFRINGEMENT. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, IN NO EVENT SHALL SHIMOGORYO BASE LLC OR ITS MEMBERS, MANAGERS, OFFICERS, EMPLOYEES, CONTRACTORS, ANALYSTS, LICENSORS, OR AGENTS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, CONSEQUENTIAL, SPECIAL, EXEMPLARY, OR PUNITIVE DAMAGES — INCLUDING, WITHOUT LIMITATION, LOST PROFITS, LOST REVENUE, LOSS OF DATA, OR BUSINESS INTERRUPTION — ARISING OUT OF OR IN CONNECTION WITH THE REPORT OR ANY USE OF, OR RELIANCE ON, ITS CONTENTS, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. IN ANY EVENT, THE PUBLISHER'S AGGREGATE LIABILITY ARISING FROM OR RELATING TO THE REPORT SHALL NOT EXCEED THE GREATER OF (A) THE AMOUNT ACTUALLY PAID BY YOU TO THE PUBLISHER FOR THE REPORT, WHICH FOR THIS FREE EDITION IS NIL, AND (B) USD 100.

Publication date; no duty to update

The Report reflects information available to the Publisher as of the publication date (July 9, 2026, incorporating live verification performed July 8–9, 2026). Developments occurring after that date may materially affect the analysis, conclusions, and projections contained herein. The Publisher assumes no obligation to update, supplement, or correct the Report.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
For informational purposes only; not investment, legal, or engineering advice. The Publisher holds positions in securities of companies discussed — see Positions and conflicts of interest. Provided “AS IS.”

Distribution

The Report is intended for free public distribution and is not confidential. It is nonetheless directed only at recipients who are capable of evaluating its contents independently, and it does not take account of the objectives, financial situation, or particular needs of any recipient. If you received it in a jurisdiction where its distribution would be unlawful, or where the Publisher would thereby become subject to any registration or licensing requirement, no offer or solicitation is made to you and you should not rely on it.

Contact

Shimogoryo Base LLC · shimogoryobase.com/consult.html

If this was useful, share it. If you want the HBF investment memorandum, the 31-name screening workbook, or research of this kind on your own question, get in touch.

© 2026 Shimogoryo Base LLC, a Colorado limited liability company. Free public edition — share complete and unaltered, with attribution.