TECHNICAL ADDENDUM

HBF After the First Standard: What the August 2026 Specification Changes

A technical addendum to *High Bandwidth Flash (HBF): The Emerging NAND Tier Beneath HBM for AI Accelerators* (Shimogoryo Base LLC, July 2026)

Prepared 19 August 2026 · Information horizon: public disclosures through 19 August 2026 · Triggering event: the first HBF standard specification, released 4 August 2026 at FMS 2026

Publisher	Shimogoryo Base LLC, a Colorado limited liability company
Prepared by	Shimogoryo Base Consulting — Andy Kowles, Principal Analyst
Amends	*High Bandwidth Flash (HBF): The Emerging NAND Tier Beneath HBM for AI Accelerators* — Shimogoryo Base LLC, July 2026. Sections 2.5, 4.2, 4.4, 5, 7 and 11 are affected.
Also revises	*High Bandwidth Flash: Mapping the Chain, Pricing the Shock* — investment memorandum, July 2026. Section 2.1 (wafer-diversion model) is superseded by Section 4 below.
Also draws on	Three simulation studies published Apr–Aug 2026 (arXiv 2604.16007, 2607.10186, 2608.11668). None measured HBF silicon; see Section 5.1.
Status of parent	The July 2026 report stands as issued. This addendum records what has changed rather than rewriting it, so its Appendix B verification log remains auditable.
Edition	**Free public edition** — may be shared complete and unaltered, with attribution
Contact	shimogoryobase.com/consult.html

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

This document 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 memorandum is published at no charge by **Shimogoryo Base LLC**. You are welcome to repost, forward, and share it — complete and unaltered, with attribution. It amends the Publisher's research report, *High Bandwidth Flash (HBF): The Emerging NAND Tier Beneath HBM for AI Accelerators*, and the companion screening workbook `HBF_Investment_Screen.xlsx`, which carries the full thirty-one-name grid with per-name evidence, exposure estimates, market data, and bull/bear cases. Both are available from the Publisher. For the research report, the workbook, or bespoke technology and supply-chain research, see shimogoryobase.com/consult.html.

Important notice (summary). This addendum is provided for general informational purposes only. **It is not investment, financial, trading, legal, tax, accounting, or engineering advice**, it is not a recommendation to buy, sell, or hold any security, and it is not an offer or solicitation with respect to any security in any jurisdiction. The Publisher is not a registered investment adviser or broker-dealer and is not acting as a fiduciary to any recipient. It contains forward-looking statements and engineering estimates that are inherently uncertain and subject to change without notice; the sizing analysis in Sections 3 and 4 is the Publisher's own construction from stated assumptions, not vendor-disclosed data. Information is drawn from sources believed reliable, but accuracy and completeness are not guaranteed; this addendum is provided "AS IS," reflects information only as of the dates indicated, 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 for the full terms governing its use.**

Table of Contents

1. What was actually published
2. The interconnect inversion
3. The bandwidth grades, re-sized
4. The wafer-diversion model becomes grade-dependent
5. Power, thermal, and why SLC is structural
6. Field reliability, the replacement unit, and why retention is the wrong specification
7. Ecosystem: the gating item has moved, but not cleared
8. Corrections register
9. What to watch next
10. How this addendum was made

Important Disclosures & Disclaimer

Basis, method, and confidence labels. The primary source for this addendum is SK hynix's own announcement of 4 August 2026, corroborated against independent trade coverage. As in the parent report, every load-bearing claim carries a confidence label: **[Confirmed]** (vendor primary source plus independent corroboration), **[Reported]** (single source or trade press only), **[Analyst speculation]** (third-party projection), **[Author analysis]** (this document's own inference). The arithmetic in Sections 3 and 4 re-runs the parent report's own published method against the new specification figures; it is **[Author analysis]** throughout, and the assumptions are stated so that a reader who disagrees can re-run it. **The specification document itself has not been read.** Every figure below comes from vendor statements about the specification and from trade coverage of those statements, not from the OCP document, which was not retrievable at the time of writing.

Bottom line

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

The July report got the architecture right and the interconnect wrong. Capacity, stack heights, media, the complex base die, the read-dominated tier, OCP-not-JEDEC, and complement-not-replace all survive the first specification intact. The unique-wafers thesis in Section 3 of that report — the spine of the economic argument — is untouched and slightly strengthened.

What changed is the attach. The report's plan of record was an HBM-idiom PHY dropping into an HBM site, with UCIe ranked as *fallback number two* and described as "plausible for mid-range inference parts." The specification adopts UCIe as the interface. That ranking is inverted.

The consequence is narrower than it first appears, and it is worth stating precisely, because the obvious reading — that a standard chiplet interface means HBF no longer needs anything custom on the host — is wrong. UCIe standardises the **physical layer**. It does not standardise the protocol a flash tier needs to speak, and the report's Section 4.2 had already conceded PHY reuse while locating the real barrier in the **controller**. That barrier is intact. What UCIe genuinely removes is the interface-standardisation problem, and what it genuinely adds is that the attach point becomes *fungible* and, more consequentially, *relocatable* — an HBF stack can hang off a companion I/O die rather than consuming compute-die beachfront. That last point does more damage to the report's shoreline framing than the interconnect change itself. See Section 2.3.

The second disclosure matters more for the economics than the headlines suggest. The specification does not state one bandwidth; it states three grades spanning roughly 0.4 to 3.0 TB/s. Re-running the report's own sizing method across that range shows the grade spread is not marketing — it is an admission that the array-parallelism problem the report identified is real and has been bracketed rather than solved. Because the compound bit penalty scales with mini-plane subdivision, **the wafer-diversion model is now grade-dependent**: the memorandum's published 5.65x base case turns out to describe the middle grade, with roughly 4.3x at the specification floor and roughly 7.2x at its ceiling.

A third disclosure sits outside the specification entirely. Three academic papers published between April and August 2026 — none of which has silicon, and all of which are simulation studies on assumed parameters — are the only public source on HBF power, thermals, and media. Working through them settles the shoreline question in HBF's favour (roughly **19x HBM4's capacity per millimetre** at comparable bandwidth density), corrects a widely propagated static-power figure that is wrong in form rather than merely in magnitude, and shows that **the bandwidth grades are thermally limited as well as array-limited**: Grade 1 fits inside an HBM-class envelope at roughly 37 W, while Grade 3 needs about 192 W per stack. It also establishes that **SLC is structurally required** rather than a first-generation conservatism — read-disturb refresh alone consumes 18% of an SLC stack's rated endurance over five years, against 1,051% for MLC and 14,016% for TLC. Section 5 sets this out.

Section 6 then asks what happens when one of these stacks fails in service, and finds the question governs the media decision more forcefully than any performance argument. HBF inherits HBM's defining field property: **it is bonded to the accelerator and cannot be replaced.** An SSD that wears out costs a thousand-dollar field-replaceable unit; an HBF stack that wears out scraps a thirty-to-forty-thousand-dollar assembly, including the good stacks beside it. But the same section argues that the endurance figures everyone is quoting — this addendum's Section 5 included — are specified against the wrong requirement. NAND endurance ratings are written against a *warranty retention* of months to a year. Production KV cache has a measured **P99 lifespan of 97 seconds**. That is roughly five orders of magnitude of retention capability a KV tier pays for in endurance and never uses, and the flash literature has measured 3.2x to 12.9x lifetime gains from relaxing it. Driving HBF cells deliberately past legacy warranty limits may be the correct paradigm rather than an abuse of the medium.

The gating item still has not cleared. Google and Tenstorrent joining the consortium is the first real demand-side movement in eighteen months, but consortium membership is not a design win, and no accelerator has been announced with an HBF site on it.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

1. What was actually published

On **4 August 2026**, at the Future of Memory and Storage conference in Santa Clara, SK hynix and Sandisk released the first technical specification for High Bandwidth Flash. It arrives roughly six months after the HBF workstream was formed in February 2026, and twelve months after the two companies signed the original standardisation memorandum of understanding in August 2025 [Confirmed — SK hynix newsroom, Aug 4, 2026; EE Times Asia; StorageReview, Aug 4, 2026].

1.1 The disclosed parameters

Parameter	Specification	July report's position	Status
Capacity per stack	Up to 512 GB	512 GB, from 16 × 32 GB core dies	Confirmed
Stack heights	8-high and 16-high	"stacked 8-high or 16-high over a base die" (Sec 2.4)	Confirmed
Bandwidth	Three grades, ~0.4 to 3.0 TB/s	Single Gen1 figure of 1.6 TB/s (Sec 2.5)	Superseded — see Sec 3
Host interface	UCIe	HBM-idiom PHY; UCIe ranked fallback #2 (Sec 4.4)	**Inverted — see Sec 2**
Standards body	OCP	OCP, corrected from JEDEC in the July revision	Confirmed
Scope of document	Host interface, electrical characteristics, packaging and reliability guidance, software I/O guide	Anticipated in outline (Sec 4.2, Sec 11)	Confirmed
Positioning	Coexists with HBM, does not replace it	"Complements HBM rather than replacing it" (Sec 1)	Confirmed
Consortium	SK hynix, Sandisk, Google, Tenstorrent	SK hynix and Sandisk only	Expanded — see Sec 7
Endurance / DWPD	Not disclosed	Flagged as an open question (Sec 11)	Still open

1.2 Two things the coverage is getting wrong

The headline bandwidth figure. A good deal of secondary coverage reports "3 TB/s" as though it were the specification. It is the ceiling of a three-grade range whose floor is roughly 0.4 TB/s — a factor of 7.5 between them. Reporting only the ceiling makes HBF sound like a near-peer of HBM4 across the board, when in fact the entry grade sits below a quarter of the parent report's own Gen1 anchor. Readers relying on machine-generated summaries should be especially careful here; several of the widely circulated write-ups of this announcement are LLM-produced and explicitly not editorially reviewed, and they consistently flatten the grade structure into a single number [Author analysis].

"Published through OCP." SK hynix's own release states the specification is published through the Open Compute Project. At least one careful trade outlet noted at the time of writing that the document was not yet formally posted in OCP's public repository [Reported]. The distinction matters for anyone trying to design against it: an announced specification and a retrievable specification are not the same thing, and this addendum is written from vendor statements about the document rather than from the document [Author analysis].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

2. The interconnect inversion

This is the substantive correction, and it goes against the parent report.

2.1 What the report said

The July report treated an HBM-idiom PHY in an HBM site as the plan of record, and put UCIe in a list of things that would happen only if that failed. Section 4.4, "On-package vs near-package fallbacks," opened with the conditional "*If interposer sites are too scarce (Section 7), fallback attach options exist, in descending bandwidth order*," and ranked them:

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.
3. On-module, mezzanine or SOCAMM-like.
4. Across CXL/PCIe — at which point it converges with AI SSDs.

Section 7 was more explicit still, describing UCIe/die-to-die attach as something that "saves compute-die shoreline, costs latency/power per bit and adds a chiplet," and judging it "plausible for *mid-range inference parts*" — that is, a compromise for the low end rather than the mainstream design point [direct quotations from the parent report, Sections 4.4 and 7].

2.2 What the specification does

It adopts UCIe as the host interface, for the stated purpose of enabling integration with CPUs, GPUs and custom accelerators [Confirmed — SK hynix; EE Times Asia]. The report's fallback is the design point.

The important qualification is that this is not the same thing as the report's fallback number two. The report imagined a *serial, near-package* link, and priced it accordingly in latency and energy per bit. What has actually been specified appears to be advanced-package UCIe, in which the HBF stack sits on the same silicon interposer or a high-density organic substrate as the host [Reported — StorageReview; Tom's Hardware]. So HBF has not moved off the package and has not escaped the shoreline problem. It has changed the protocol spoken across the same real estate.

> **What survives and what does not**
>
> The report's *shoreline analysis* is intact. Package beachfront remains the binding constraint, the arithmetic of "how many HBM sites is a terabyte worth" is unchanged, and Section 7's framing of the adoption question as one about site economics rather than about flash is if anything reinforced.
>
> The report's *PHY prediction* is wrong, and the strategic conclusion that rested on it needs reversing.

2.3 What UCIe removes, and what it leaves

Section 7 of the parent report concluded that "HBF is therefore gated, as HBM was gated, on close GPU vendor collaboration with the media manufacturers." It is tempting to read the adoption of a standard chiplet interface as defeating that conclusion outright. It does not, and the reason is that the report was careful about *where* it located the barrier. Section 4.2 reads:

> "…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."

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

The report had already conditionalised on PHY reuse. UCIe resolves the PHY question in precisely the direction Section 4.2 allowed for, and leaves the controller question exactly where it was.

2.3.1 The protocol problem UCIe does not solve

UCIe defines a die-to-die physical and link layer, plus protocol mappings for PCIe and CXL. Neither mapping fits HBF [Author analysis]:

- **CXL.mem** is cacheline-granular load/store with a host expecting bounded, low-nanosecond, effectively deterministic completion. HBF is page-granular, asynchronous, microsecond-class, variable-latency, and radically asymmetric between read and write. Mapping HBF onto CXL.mem would either stall the host's memory pipeline on every access or require enough shim logic that the custom controller has been rebuilt under another name.
- **PCIe** carries the wrong semantics and the wrong overhead for a tier whose entire justification is being closer than an SSD.

The remaining option is UCIe's **streaming (raw) mode**, which carries an arbitrary protocol chosen by the implementer. That is almost certainly what has been specified, and it is the most likely explanation of the "xPU–HBF interface" naming discussed in 2.4: UCIe is the transport, and the HBF specification defines what rides on it. Both vendor statements are then true simultaneously.

If that reading is right, the host must still implement in silicon: page-granular asynchronous reads with out-of-order completion, read-retry and temperature-compensated reference adjustment, ECC/LDPC framing, bad-block and wear telemetry, scheduling across dozens of concurrently active mini-arrays, and demotion-write handling. No shipping accelerator contains that block. The parent report's claim that "significant logic… more sophisticated… than HBM" is required stands [Author analysis].

2.3.2 The physical envelope is still provisioned per-site

A UCIe port is not a universal socket. Hosting an HBF stack requires the port and its surrounding package to be designed for one [Author analysis]:

- **Link count.** Grade 3 needs roughly six to eight x64 links on the arithmetic in Section 3.2. That is a large, pre-committed allocation, not a spare port.
- **Power delivery** sized for a 16-high NAND stack, whose profile differs from a compute or I/O chiplet.
- **Thermal path.** This deserves more weight than the parent report gave it. NAND data retention degrades with temperature in a way DRAM refresh largely masks; placing a flash stack adjacent to a high-power accelerator die is a genuine reliability problem, not a packaging detail. It bears directly on the endurance question the report left open in Section 11.
- **Mechanical keep-out** for stack height, and package routing designed for the link topology.

2.3.3 What genuinely changes

Two things, and the second matters more than the interconnect switch itself.

Fungibility. A UCIe port within a compatible envelope can host an HBF stack, a third-party memory chiplet, or an I/O die. An HBM site hosts HBM. That asymmetry is what makes reserving the port defensible before HBF has proven itself, and it is the strongest of the adoption arguments: a UCIe site can be repurposed if HBF slips, whereas a bespoke HBF PHY is stranded silicon. The parent report came close to anticipating this, expecting vendors to "configure each of their limited number of 'sites'… to accommodate any configuration of HBM/HBF either in uniform or hybrid stacks."

Relocatability. If HBF is a UCIe chiplet, it need not attach to the compute die at all — it can hang off a companion I/O or memory-controller die. The report listed this option but filed it under fallbacks. Under a standard chiplet interface it becomes a natural topology, and it moves the binding constraint: **compute-die beachfront stops being the scarce resource and interposer area plus reticle limits become the scarce resource instead.** Both are scarce, but they are not equally scarce, and they are not owned by the same design team. This does more to undermine the parent report's central framing — "how

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

many HBM sites is a terabyte worth?" — than the interface change does, because for an I/O-die topology that question is simply the wrong one [Author analysis].

2.3.4 Net assessment

Barrier the report identified	Effect of UCIe adoption
Host-side memory PHY must be designed	**Removed.** Shipping IP, multiple vendors
Interface must be standardised before anyone can design to it	**Removed.** This is what the specification does
Host needs custom HBF controller silicon	**Unchanged.** Streaming-mode transport carries no flash semantics
HBF and xPU generations must be co-designed and co-timed	**Weakened, not removed.** The controller block still has to be validated against an HBF generation
Site allocation is a one-way bet on an unproven tier	**Removed.** Fungible port; repurposable if HBF slips
Compute-die shoreline is the binding constraint	**Displaced.** Shifts to interposer area under an I/O-die topology
Latency and energy per bit across a die-to-die link	**Worsened.** Protocol and serialisation overhead atop microsecond reads

The last row should not be waved away. For a read-streaming, bandwidth-bound tier holding weights and prefix cache, the trade is defensible. For fine-grained random access it is not, and the report's Section 11 warning that HBF's addressable share shrinks toward "weights only" if workloads prove latency-sensitive is *more* live under a die-to-die protocol, not less [Author analysis].

2.4 An unresolved question on naming

SK hynix says UCIe. Sandisk's framing refers to an "xPU–HBF host interface," and at least one outlet was unable to determine whether that is a UCIe profile, a superset, or a parallel interface described alongside UCIe [Reported — Tom's Hardware].

Section 2.3.1 offers what this document believes is the resolution: the two are different layers. UCIe is the transport, in streaming mode; "xPU–HBF" is the flash-specific protocol carried over it. On that reading both companies are describing the same architecture accurately and the apparent conflict dissolves. **This is a falsifiable prediction, not a confirmed fact** [Author analysis]. If instead "xPU–HBF" proves to be an independent parallel interface for which UCIe is merely one of several transport options, then the interface-standardisation benefit claimed in 2.3.4 is weaker than stated, because designing to HBF would no longer be a matter of instantiating known IP.

3. The bandwidth grades, re-sized

The parent report's Section 2.5 established a method for testing whether a claimed HBF bandwidth is physically plausible: divide stack bandwidth by die count to get per-die bandwidth, then divide by the throughput of one pipelined mini-array to get the number of independently active mini-arrays each core die must sustain. At 16 KB pages and SLC reads of 3–5 µs, one mini-array yields about 4 GB/s. Applying that method unchanged to the newly disclosed grades:

Grade	Per stack	Per core die, 16-high	Mini-arrays, 16-high	Per core die, 8-high	Mini-arrays, 8-high
Grade 1 — specification floor	~0.4 TB/s	~25 GB/s	~6	~50 GB/s	~12
Grade 2 — inferred, see below	~1.6 TB/s	~100 GB/s	~24	~200 GB/s	~49
Grade 3 — specification ceiling	~3.0 TB/s	~188 GB/s	~46	~375 GB/s	~92

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

[Author analysis. Method and the 4 GB/s mini-array figure are from the parent report, Section 2.5. Only the 0.4 and 3.0 TB/s endpoints are disclosed; the middle grade is inferred, see 3.2.]

3.1 What the spread tells us

The parent report described the 1.6 TB/s case as requiring “on the order of 32–64+ concurrently active mini-arrays plus deep read pipelining” and treated that as aggressive. The new endpoints bracket it on both sides, and the implication is worth stating plainly: **the grade structure is an engineering admission, not a marketing ladder.** Grade 1 asks each die for about six concurrent mini-arrays, which is undemanding and could plausibly be built with far less array subdivision. Grade 3 asks for roughly forty-six — close to double the report’s already-aggressive central case, and that is before derating for pipelining inefficiency, bank conflicts, and ECC overhead.

Put another way: the two companies have not solved the array-parallelism problem the report identified. They have priced it, by letting the customer choose where on the curve to sit. That is a rational response to genuine uncertainty about which workloads will pay for bandwidth, and it is more informative than a single number would have been [Author analysis].

3.2 A link-count inference for the middle grade

The middle grade has not been disclosed. There is a reason to think it lands near 1.6 TB/s. A 64-lane UCIe link at 64 GT/s carries about 512 GB/s raw, or roughly 490 GB/s after protocol overhead — which is, to within rounding, Grade 1. If the grades are quantised by link count rather than chosen arbitrarily, the natural ladder is approximately one, four, and six-to-eight links, putting the middle grade in the region of 1.6–2.0 TB/s and reconciling it with the 1.6 TB/s figure Sandisk disclosed for Gen1 in 2025.

This is **[Author analysis]** resting on a lane-rate arithmetic check, not on any statement by either company, and the tables above flag the middle row accordingly. It is offered because it is falsifiable: when the document becomes readable, the middle grade either is or is not near 1.6 TB/s, and the inference can be scored.

3.3 A caution on the ceiling

It is not established whether 3.0 TB/s is specified per package or per subsystem; the trade coverage does not resolve it, and one outlet raised the question directly [Reported]. If it is per subsystem — several stacks aggregated — then the per-die arithmetic in the table above is too pessimistic at the top end by whatever the aggregation factor is, and Grade 3 becomes considerably more buildable. This is the single assumption in this addendum most likely to be wrong, and it propagates into Section 4 [Author analysis].

---

4. The wafer-diversion model becomes grade-dependent

The investment memorandum of July 2026 published a compound bit-penalty stack quantifying how many fewer sellable bits an HBF wafer yields than a commodity TLC wafer. Its base case was **5.65x**, within a range of 4.33x to 8.12x. One of its four factors — array-efficiency loss from mini-plane subdivision, bracketed at 1.25 / 1.45 / 1.70 — is a direct function of how finely the array is divided, which Section 3 has just shown to vary by a factor of roughly seven across the specified grades.

That single-figure base case can no longer stand on its own. Scaling the array-efficiency factor’s increment above unity in proportion to mini-arrays per die, and holding the other three factors at the memorandum’s published values:

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided “AS IS.”

Penalty factor	Grade 1 (~0.4 TB/s)	Grade 2 (~1.6 TB/s)	Grade 3 (~3.0 TB/s)	Basis
SLC vs TLC cell density	3.00x	3.00x	3.00x	Parent report Sec 2.3, stated exactly. Grade-invariant.
Array-efficiency loss (low / base / high)	1.06 / 1.11 / 1.18	1.25 / 1.45 / 1.70	1.47 / 1.84 / 2.31	Memorandum Sec 2.1 values, scaled by mini-arrays per die from Sec 3 above. Estimated.
TSV keep-out corridors	1.05 / 1.10 / 1.18	1.05 / 1.10 / 1.18	1.05 / 1.10 / 1.18	Memorandum Sec 2.1, unchanged. Estimated.
Stack and KGD yield loss	1.10 / 1.18 / 1.35	1.10 / 1.18 / 1.35	1.10 / 1.18 / 1.35	Memorandum Sec 2.1, unchanged. Conservative for 8-high. Estimated.
COMPOUND PENALTY (low / base / high)	3.68x / 4.33x / 5.62x	4.33x / 5.65x / 8.12x	5.09x / 7.18x / 11.05x	Multiplicative.

[Author analysis. The Grade 2 column reproduces the July memorandum's published figures exactly, which is the internal consistency check on the scaling method: the middle grade is the case the memorandum was implicitly modelling.]

4.1 What this changes, and what it does not

It does not change the direction of the thesis. Even the most benign grade destroys bits at 3.7–4.3x, which remains materially more violent than HBM's effect on commodity DRAM. The supply-destruction mechanism the memorandum identified is intact across the whole specified range.

What changes is that **the magnitude is now a function of product mix, and the mix is undisclosed.** A market that settles on Grade 1 parts for capacity-led inference destroys roughly 4.3x bits per wafer at base case. A market that settles on Grade 3 destroys roughly 7.2x. That is a spread of about 66 per cent on the central estimate, sitting entirely on a variable that no vendor has commented on. Anyone carrying the memorandum's 5.65x figure forward should now carry it as *the mid-grade case*, with explicit sensitivity to mix, rather than as a point estimate.

A second-order consequence is worth flagging. Because bandwidth and bit-destruction are coupled through array subdivision, HBF has an unusual property for a memory product: **the faster grades are disproportionately more expensive in foregone commodity supply, not merely in unit cost.** If inference workloads turn out to be capacity-hungry rather than bandwidth-hungry — which is the direction the KV-cache and long-context literature has been pointing — the market may settle toward the low grade, and the supply shock would be correspondingly milder than the memorandum's headline figure implies [Author analysis].

5. Power, thermal, and why SLC is structural

The specification says nothing public about power, thermals, endurance, or media. Three academic papers published between April and August 2026 do, and read together they change what the grade structure of Section 3 actually means. This section works through them, corrects a figure the literature has propagated, and reaches a conclusion the vendors have not stated: **SLC is not first-generation conservatism to be relaxed later. It is structurally required for any sustained high-bandwidth read tier**, and that has a direct consequence for the wafer-diversion model in Section 4.

5.1 Three sources, none of which has silicon

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

Paper	Date	What it is	Relevance
MemExplorer (arXiv 2604.16007), Cambridge / Imperial / Edinburgh / Microsoft	17 Apr 2026	Memory design-space synthesizer for heterogeneous inference NPUs	Table 1 carries the only published side-by-side of HBF against HBM on shoreline, power, and energy per bit
FlashAccel (arXiv 2607.10186), ICT / Chinese Academy of Sciences	11 Jul 2026	Co-designed HBF-in-GPU system; event-driven simulator on LLMCompass plus a plane-granularity NAND model	Models an HBF core die from a published XL-Flash die shot; reports a measured read energy from a real hybrid-bonded flash prototype
HBF Sucks! (arXiv 2608.11668), Peking University	12 Aug 2026	Serving-stack study on an extended TokenSim with four production Qwen traces; 3D-ICE thermal model and a write-endurance budget	The only published thermal model of an HBF stack

None of these measured an HBF device, because none exists. Two of the three say so explicitly; the third, despite a title and abstract that read as hardware characterisation ("a full-stack characterization… we built that system and measured it"), states plainly in its methodology section:

> "HBF remains projected hardware, so its service parameters are sourced or swept, not claimed as measured silicon."

Every figure in this section is therefore a **simulation or model output on assumed device parameters**, not a measurement, and is treated as such [Author analysis]. The single exception is FlashAccel's read energy of 8 pJ/bit, which it attributes to measurements on a real hybrid-bonded flash prototype — the only empirically grounded power number in the literature [Reported].

> **A note on citing these papers.** *HBF Sucks!* will be widely cited on its title alone. Two things should travel with it. Its headline regression (end-to-end latency up 2–5.5x) is measured on an **HBF-1** organisation in which the interposer is filled entirely with HBF and the near tier is demoted to off-package GDDR7 — a configuration the specification's own coexist-with-HBM framing rejects. And its own conclusion is far narrower than its title: *"The device is fine; the drop-in deployment is not."* The defensible claim is that swapping HBF for SSD under a transient-KV offload stack, while sacrificing HBM to do it, loses. That is close to what the parent report already said in Section 11 [Author analysis].

5.2 The shoreline question, settled

The parent report treated package beachfront as the binding constraint but could not quantify HBF's efficiency against HBM. MemExplorer's Table 1 does:

Device	Capacity	Bandwidth	Shoreline	GB per mm	GB/s per mm
HBM3E, 8-high, 1 stack	24 GB	1 TB/s	~11 mm	2.2	91
HBM4, 12-high, 1 stack	36 GB	2 TB/s	~15 mm	2.4	133
HBF, 1 stack	**384 GB**	**1 TB/s**	**~8.25 mm**	**46.5**	**121**

HBF delivers roughly 19x HBM4's capacity per millimetre of shoreline at 0.91x its bandwidth density. In bandwidth terms the shoreline trade is close to a wash; the capacity gain is an order of magnitude and a half. This is the strongest single number available for the thesis, and it comes from a disinterested academic source rather than a vendor deck [Reported — MemExplorer Table 1; the HBF row cites vendor-derived references, so it is a modelled figure, not an independent measurement].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

Two caveats. The assumed configuration — 384 GB at 1 TB/s, implying a 12-high stack — matches neither of the specification's stack heights, because it predates the specification by four months. And it sits between Grade 1 and Grade 2, so it is not a datapoint for the grade structure, only for shoreline efficiency.

5.3 The published power figures do not reconcile

HBF is, at a glance and at the marketing level, attractive compared with HBM for read-heavy workloads. Inference-focused AI hardware systems must seek lower idle power — particularly at the edge — better thermal dissipation, and support for large context. So here we dive deeper than the original report on the subject of power, using four public sources which are in circulation. **These sources are mutually incompatible.**

Source	Figure	Implication for a 384–512 GB stack
MemExplorer	Static background ~300 mW/GB (stated as 4x HBM3E)	**115 W before any access**
HBF Sucks! 3D-ICE model	Hits 80 °C junction at 202.27 GB/s and 53.72 W	Total thermal budget ~54 W
H3 (SK hynix's own paper, via *HBF Sucks!*)	~160 W per HBF cube	Six stacks approach 960 W
FlashAccel	8 pJ/bit read, measured on a hybrid-bonded prototype	~26 W dynamic at Grade 1

The contradiction is immediate: **if static power were really 115 W, the thermally modelled stack would exceed its 80 °C junction limit while completely idle.** Both cannot be true. One of the two most-cited figures in the literature is wrong, and the following subsection argues it is the per-GB static figure — not merely in magnitude, but in form [Author analysis].

5.3.1 NAND static power is a fixed term, not a per-GB term

DRAM background power is dominated by **refresh**: every row, every 32–64 ms, indefinitely, whether or not the device is being used. Because refresh work scales with the number of rows, expressing DRAM static power in mW/GB is physically sensible.

NAND has no refresh. Stored charge persists without maintenance — that is the entire premise of non-volatile memory. An idle HBF stack therefore has no capacity-proportional background activity at all. Its static draw is set by components that do not scale with array size:

Component	Estimated idle draw	Scales with capacity?
UCIe-A PHY, link up, no traffic	~1.5 W	No
Base-die logic (scheduler, L2P, flash management)	~6 W	Weakly, via mapping-table size
ECC engine, idle	~2 W	No
DRAM buffer (~4 GB) for mapping and write staging	~0.3 W	No
Array and peripheral bias, planes idle	~1.5 W	Weakly
TOTAL, essentially capacity-independent	**~11 W**	

[Author analysis. Component estimates are engineering judgment, not vendor data. The structural claim — that the total is dominated by fixed terms — is the load-bearing part and does not depend on the individual figures.]

Against MemExplorer's implied 115 W, this is roughly **10x lower**, and the discrepancy grows with capacity precisely because the per-GB form is wrong. The error is easy to make: taking HBM3E's ~75 mW/GB, multiplying by four to reflect flash's worse energy per bit, then applying the result to an array sixteen times larger. The multiplication is defensible; the per-GB basis is not.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

This resolves the contradiction in 5.3. With ~11 W static and ~26 W dynamic, a Grade 1 stack draws roughly **37 W** — consistent with *HBF Sucks!*'s ~54 W envelope and comfortably inside HBM-class cooling. H3's ~160 W per cube must then describe peak dynamic operation at a high grade, not idle draw [Author analysis].

One caveat, and it is the right one: this holds only if the device is not performing continuous background reliability reads. Section 5.5 quantifies that scrub load and finds it costs about **10 mW** for SLC — thermally and energetically free.

5.4 The thermal ceiling, re-derived for a read-mostly tier

HBF Sucks! reports its stack reaching the 80 °C junction limit at 202.27 GB/s and 53.72 W. Backing out the energy per bit:

> 53.72 W ÷ (202.27 GB/s × 8 bits) = **33.2 pJ/bit**

That is four to five times the 6–8 pJ/bit that MemExplorer and FlashAccel give for reads. The reason is stated in the paper: the model is a **16-high TLC stack driven by write-heavy KV traffic**, and NAND programming — multi-pass incremental-step-pulse programming with verify — is far more energy-intensive than sensing. The authors say as much: *"A read-mostly workload at the same bandwidth would run cooler."*

The 202 GB/s figure is therefore not a general HBF thermal limit. It is the limit for the most punishing possible workload on the least suitable media. Holding the same 53.72 W envelope and substituting read energy:

Workload and media	Energy per bit	Sustainable bandwidth at 80 °C
TLC, write-heavy (the published case)	33.2 pJ/bit	202 GB/s
SLC, read-mostly, at FlashAccel's measured 8 pJ/bit	8 pJ/bit	**839 GB/s**
SLC, read-mostly, at MemExplorer's 6 pJ/bit	6 pJ/bit	**1,119 GB/s**

5.4.1 The grades are thermally limited, not only array-limited

Section 3 showed the bandwidth grades map onto concurrently active mini-arrays. They also map onto dissipated power, and the second constraint binds sooner:

Grade	Bandwidth	Dynamic power at 8 pJ/bit	Plus ~11 W static	Against a ~54 W envelope
Grade 1 — floor	400 GB/s	25.6 W	~37 W	**0.5x — fits**
Grade 2 — inferred	1,600 GB/s	102.4 W	~113 W	1.9x over
Grade 3 — ceiling	3,000 GB/s	192.0 W	~203 W	**3.6x over**

[Author analysis, using FlashAccel's measured read energy and the thermal envelope implied by HBF Sucks!'s 3D-ICE model.]

Grade 1 read-mostly sits at roughly half an HBM3E-class thermal budget and is genuinely buildable. **Grade 3 requires dissipating around 192 W from a single stack**; six of them is about 1.15 kW of memory before the accelerator draws anything, which is the same order as H3's 960 W for six cubes. Grade 3 is not gated on array parallelism alone. It is gated on a cooling solution that does not exist in an HBM form factor, and that is a harder problem than adding planes.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

5.4.2 Two notes on the thermal target

80 °C is a DRAM-derived limit applied to flash. *HBF Sucks!* sets its junction ceiling "matching HBM3e." But a hot DRAM merely refreshes more often — a recoverable bandwidth and power tax. A hot NAND array loses stored charge, and the remedy consumes program/erase cycles. Retention degrades roughly two-fold per 10 °C, so 40 °C to 80 °C is about a sixteen-fold reduction. HBF may need a *lower* junction limit than HBM, which would make the power problem harder than the figures above suggest, not easier [Author analysis].

The TSV area penalty buys thermal conductivity. FlashAccel's modelled die spends 12 of 149 mm² on TSV — 8.1% of array area, and a cost in the wafer-diversion model. But copper conducts heat at roughly 400 W/m·K through a stack of thinned dies whose bond interfaces are otherwise poor vertically. The corridors that cost array efficiency are also the primary vertical heat path out of the stack. The memorandum books that 8% as pure loss; it is not [Author analysis].

5.5 Read disturb under a streaming workload

A tier that sustains hundreds of gigabytes per second of reads generates read counts unlike anything in SSD practice, and read disturb is the mechanism most likely to be mis-modelled by anyone reasoning from storage intuition.

5.5.1 The mechanism, stated correctly

Reading a page applies a low reference voltage to the *selected* wordline and a high pass voltage, Vpass, to **every other wordline in the block**. Vpass weakly injects charge into those unselected cells. The cell being read is not the victim; its blockmates are. Read disturb is therefore a **block-level** phenomenon, and vendor budgets are quoted as reads-per-block between refreshes. Reasoning in reads-per-page understates exposure by the number of pages in a block.

At Grade 1, a 512 GB stack with 4 KB pages completes a full sweep of every page every **1.28 seconds**. With a 256-page block, each block receives 256 reads per sweep, giving **1.73×10^7 block-reads per day** [Author analysis].

5.5.2 What sequential access buys, and what it does not

The metric that governs is **disturb per unique byte delivered**, and here the workload matters enormously:

- **Weights streaming is the best case that exists.** Every read returns a page of unique data, so every Vpass application is paid for. Reuse factor R = 1.
- **Transient KV with hot prefixes is the pathological case.** A page read R times inflicts R disturbs on its blockmates while delivering one page of unique data — R times the disturb per useful byte. This is the regime *HBF Sucks!* measured, and it is a property of the access pattern rather than of the device.

Sequential access buys two further advantages. Refresh becomes **deterministic**: the sweep order is known, so the moment each block crosses threshold is known, without per-block read counters or conservative margins. And refresh becomes **nearly free**, because the block has just been read and the data is already in flight — the scrub costs only the program, not the read.

There is a third advantage specific to model weights, and it is not present in any of the three papers. **Weights are a cached copy of authoritative data held elsewhere.** A degraded block does not require in-place refresh; it can simply be re-fetched from host memory or storage. That decouples read disturb from program/erase consumption entirely. KV cache is unique data and has no such escape. This is a structural asymmetry between the two workloads and it favours exactly the use case the parent report identified as HBF's natural one [Author analysis].

What layout does not buy: total block-reads per sweep is conserved. If every page is read once, each block receives exactly its page count in reads regardless of ordering or striping. Access order cannot reduce the total. The design lever is **block geometry**:

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

Pages per block	Refresh interval (SLC, Grade 1)	Scrub volume
64	16.7 hrs	0.74 TB/day
128	8.3 hrs	1.47 TB/day
256 (XL-Flash-like)	4.2 hrs	2.9 TB/day
2,048 (conventional 3D NAND)	0.5 hrs	23.6 TB/day

A 32-fold swing from block geometry alone. XL-Flash's small blocks and 4 KB pages are conventionally explained as a latency optimisation; they are also a read-disturb optimisation, which is a further reason it is the right starting die alongside its plane count and sense time [Author analysis].

The cost of that scrub, at the SLC rate, is approximately **10 mW** — 2 mW of read and 8 mW of program. It does not perturb the power analysis of 5.4 [Author analysis].

5.6 Why SLC is structural, not first-generation conservatism

Running the same Grade 1 streaming workload against three media, and charging each only for read-disturb refresh — no user writes at all:

Media	Read-disturb budget	Refresh interval	Scrub volume	P/E consumed over 5 years	Verdict
SLC	~3×10^6 reads/block	4.2 hrs	2.9 TB/day	10,512 cycles — **18% of rated**	**Viable**
MLC	~3×10^5	25 min	29.5 TB/day	105,120 — **1,051% of rated**	10x over
TLC	~7.5×10^4	6 min	118 TB/day	420,480 — **14,016% of rated**	140x over

[Author analysis. Read-disturb budgets are order-of-magnitude engineering figures from general 3D NAND practice, not vendor datasheet values. The ratios between media are robust; the conclusion survives an order of magnitude either way. See the caveat at the end of this section.]

MLC and TLC exhaust their entire rated endurance on scrubbing before a single user write lands. That single row disqualifies both for a sustained high-bandwidth read tier. It also means *HBF Sucks!*'s endurance finding is **more** pessimistic than reported: the authors charged their TLC stack for KV writes against a "deliberately favorable" budget, without also charging it for read-disturb refresh.

Three further arguments reinforce the same conclusion.

ECC throughput. Grade 1 pushes **3.2 Tb/s** through the error-correction engine; Grade 3 pushes **24 Tb/s**. SLC requires light BCH at high throughput per instance. TLC requires LDPC with soft decision, at roughly 10–50 Gb/s per decoder instance — implying **64 to 320 instances at Grade 1**, and up to 2,400 at Grade 3. That is not a base die anyone builds inside a stack power budget, and it compounds the parent report's Section 4.2 observation that HBF's logic die is more gate-heavy than HBM's.

A feedback loop with no SLC equivalent. TLC's higher raw bit error rate triggers soft-decision retries. Each retry is an *additional read operation*, applying *more* Vpass, raising disturb, raising RBER further. At a threefold retry rate under heat and wear, block-reads per day rise from 1.73×10^7 to 5.19×10^7. SLC reads hard-decision only; the loop does not exist.

Cross-temperature. Threshold voltages drift roughly 2–4 mV per °C. Enterprise SSDs qualify across a wide cross-temperature window because they sit in thermally stable chassis. A stack bonded beside an accelerator does not: it swings some 40 °C between idle and load on second timescales, so data programmed cool is routinely read hot. SLC has one wide gap between two states and absorbs this. TLC has seven narrow gaps and cross-temperature becomes a first-order error term.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

The one axis where SLC's advantage is smaller than expected is raw bandwidth per plane. SLC at a 4 KB page and 4 µs sense yields about 1.02 GB/s per plane; TLC at a 16 KB page and 60 µs yields about 0.27 GB/s — a factor of 3.75, not the order of magnitude the latency ratio alone would suggest, because the larger TLC page partly offsets the slower sense. Plane bandwidth is not the decisive argument. Read disturb and ECC are.

5.7 Consequence for the wafer-diversion model

Section 4 revised the compound bit penalty to be grade-dependent while holding the SLC-versus-TLC density factor at a flat 3.00x. The analysis above says something stronger about that factor: **it is permanent for the performance grades, not a first-generation artifact.**

The common expectation — voiced in the parent report's own Section 2.5 as analyst speculation, and hinted at by SanDisk in discussion of a later "capacity HBF" — is that HBF begins in SLC for latency and migrates to TLC for density once the tier matures. The read-disturb arithmetic says that migration cannot happen at high bandwidth. A TLC variant is possible only at bandwidths low enough that sweep frequency falls by two orders of magnitude, which places it well below Grade 1 and arguably outside the specification's range altogether.

For the model in Section 4 this means the 3.00x factor should be treated as **fixed across all three grades** rather than as a parameter that relaxes over time. It also means any bull case resting on TLC-based capacity HBF arriving to soften the supply shock should be discounted: the media that would soften it cannot survive the workload that motivates it [Author analysis].

> **What would falsify this section.** The read-disturb budgets are the weakest input, and they are not vendor figures. Three numbers would settle it: an SLC read-disturb specification in reads-per-block from Kioxia or SK hynix; a static-power figure for an HBF stack, which decides whether Grade 1 is a 37 W device or a 140 W one; and confirmation of whether H3's ~160 W per cube is peak dynamic or total. All three are reasonable requests to put to the specification's contributors, and each is a single number that would move this analysis materially.

6. Field reliability, the replacement unit, and why retention is the wrong specification

Neither the specification nor the parent report addresses what happens when an HBF stack fails in service. That question turns out to govern the media decision more forcefully than any of the performance arguments in Section 5, because HBF inherits a property of HBM that has nothing to do with flash: **it is bonded to the accelerator and cannot be replaced.**

This section establishes what HBM's field behaviour actually is, what a failure costs, how that extends to a medium that wears out, and then argues that the endurance figures everyone is quoting — including this addendum's own Section 5 — are specified against the wrong requirement for the workload HBF is aimed at.

6.1 What HBM's field rate actually is

Component-level fleet data is scarce. Two public anchors exist, and they measure different things.

Meta's Llama 3 405B run remains the only published component breakdown at scale: **419 unexpected interruptions over 54 days on 16,384 H100s** — one every ~3 hours cluster-wide — of which **72 (17.2%) were attributed to HBM3**, second only to GPU faults at 148 (30.1%) [Confirmed — widely reported from the Llama 3 paper].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

Derived quantity	Value
HBM3 events per GPU-day	8.14×10^{-5}
Annualised per GPU	**~2.97%**
Annualised per HBM stack (H100 SXM5 carries 5 active)	**~0.59%**
P(≥1 HBM interruption per GPU across a 5-year rack life)	~14%

Databricks, writing in July 2026, gives an operational planning figure: GPU hardware failure event rates "can be an order of magnitude higher than CPUs," and they use **1% annualised per GPU** as a conservative all-cause assumption — from which a 256-GPU job over 30 days has a ~19% chance of an event, and a 1,024-GPU job 57% [Reported].

> **Three caveats that must travel with these numbers.** An *interruption* is not an *RMA*: most are correctable or contained errors that halt a job and consume a spare row, not hardware condemnations. A 54-day window on a newly commissioned cluster sits in the infant-mortality region of the bathtub curve and likely overstates steady state. And a five-year life reaches the wear-out region, which linear extrapolation understates. Treat 2.97% per GPU-year as an interruption rate with wide error bars, not an AFR. **No inference-fleet component breakdown has been published at all**, which is a real gap given that inference duty cycles differ materially from training [Author analysis].

6.2 The repair ladder, and what the replacement unit is

HBM has more in-field repair depth than is generally assumed. NVIDIA documents the escalation explicitly:

Stage	Mechanism	Exhaustion
1	On-die ECC and error containment	transparent
2	Dynamic page offlining	retire the affected page
3	Row remapping into spare rows	**up to 512 total remaps**
4	HBM channel repair *(Blackwell)* — swap in a spare channel on a third remap attempt against a bank; if repair succeeds the GPU is recovered	spare channels exhausted
5	**RMA the GPU**	—

RMA criteria are met when any of the following sets the row-remap failure flag: eight uncorrectable rows remapped on one bank, the *same* row remapped twice, or 512 total remaps [Confirmed — NVIDIA GPU Memory Error Management, RMA Policy].

That Blackwell added an entire new repair tier is the tell. Vendors are spending package area on redundancy precisely because the alternative is so expensive.

The replacement unit is the GPU. HBM is bonded to the interposer through CoWoS; there is no HBM-level field service. A single condemned stack therefore scraps the GPU die, the interposer and substrate, the assembly and test value, **and the good HBM stacks alongside it**.

The good HBM is not recovered. Thermocompression-bonded and hybrid-bonded stacks cannot be de-bonded from an interposer non-destructively at production scale; rework risk is a named concern in advanced-packaging practice, and it applies before mould and lid attach, not to fielded parts. A condemned package is materials reclamation, not component reuse [Author analysis].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

The industry response is more redundancy rather than recovery. Samsung showed base-die SRAM remapping at Hot Chips 2026 in which spares on *one* die can repair defects on *another* — pooling redundancy across the stack instead of stranding it per-die [Reported].

6.3 How this extends to HBF

Most of it carries across unchanged: the same interposer, the same non-replaceability, the same scrap consequence, and the same joint-fatigue and warpage exposure — worse at 16-high, which is why SanDisk has singled out ultra-low die warpage as an enabling capability.

Two things differ, and they pull in opposite directions.

NAND's repair ladder ends somewhere better. DRAM exhausts spare rows and the part is condemned. NAND retires bad blocks from a spare pool it carries natively, and when that pool depletes the device can **shrink rather than die**. For a component soldered to a $30,000-plus assembly, graceful capacity degradation is a far better failure mode than HBM's cliff. But it only pays off if the runtime tolerates a tier that loses capacity mid-life — a software-stack requirement nobody has specified, and which would have to live in the specification's "software user guide for read and write operations" [Author analysis].

NAND has a wear-out clock DRAM does not. DRAM has no normal-operation wear mechanism; its failures are stochastic. NAND's are partly deterministic — a function of write volume. HBF's end of life is therefore plannable, and also guaranteed.

6.4 The consequence asymmetry that sets the endurance target

Set the wear-out clock against a five-year rack life — the depreciation horizon most hyperscalers now use for AI server equipment, though Amazon shortened a subset back to five years in 2025 citing the pace of AI development [Reported]:

Workload	Implied life	Basis
Weights-only, read-mostly SLC	**~28 years**	read-disturb scrub alone consumes 18% of rated endurance in 5 years (Sec 5.6)
KV tier, as modelled by *HBF Sucks!*	**~2.8 years**	0.56x an SSD pool rated for 5 years

A KV-heavy HBF tier on those assumptions is condemned *inside* the rack life, and the remedy is scrapping the accelerator.

> ### The asymmetry
>
> An SSD wears out and you replace a field-replaceable unit costing on the order of a thousand dollars. An HBF stack wears out and you scrap an accelerator assembly costing on the order of thirty to forty thousand — roughly a thirty-fold difference in the consequence of the same physical event.
>
> **HBF's endurance target is therefore not "as good as an SSD." It must be effectively unlimited across the assembly's life.** This is a stronger argument for SLC than anything in Section 5.6: the objection to TLC is not that it is inconvenient, but that it converts a wear-out event into a capital write-off.

6.5 Retention is the wrong specification for KV cache

Everything above — and the endurance arithmetic in Section 5.6 — rests on an assumption inherited from storage practice that does not survive contact with the workload. It is worth stating plainly because it may be the single largest analytical error in the current HBF literature.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

6.5.1 Endurance is specified against retention

A NAND endurance rating is not a statement about when cells stop working. It is a statement about **when cells can no longer hold data for a warranted period after the rated cycle count.** JEDEC's SSD endurance specifications define the requirement explicitly: client parts must retain data for **one year at 30 °C** after rated endurance, enterprise parts for **three months at 40 °C**.

The dominant wear mechanism — tunnel-oxide degradation from repeated programming, producing charge leakage — manifests *as retention loss*. So the endurance number and the retention requirement are two ends of the same specification. Relax one and you buy the other.

6.5.2 KV cache needs 97 seconds

The most detailed public characterisation of production KV-cache behaviour is *KVCache Cache in the Wild* (USENIX ATC 2025, arXiv 2506.02634), built on traces from a leading commercial LLM service. Its third headline finding is unambiguous:

> "The lifespan of KV$ is **ephemeral**: the **P99 lifespan of KV$ in to-B workloads is 97 seconds**, so a small cache is typically sufficient."

The same study finds 54% of requests are single-turn, median turn count is one, and single-turn requests account for 97% of KV reuses in business workloads — so the ephemerality is not an artifact of short conversations but a property of how the tier is actually used, agentic multi-turn contexts included [Confirmed — KVCache Cache in the Wild, §1 and §3].

Against the specification the endurance rating is written to:

Requirement	Duration	Ratio to KV$ P99
Measured KV$ P99 lifespan (to-B)	**97 seconds**	1x
JEDEC enterprise retention (3 months @ 40 °C)	7.8×10^6 s	**~80,000x**
JEDEC client retention (1 year @ 30 °C)	3.2×10^7 s	**~325,000x**

Roughly five orders of magnitude of retention capability that a KV tier pays for in endurance and never uses.

6.5.3 What relaxing it actually buys

This is not a novel idea; it is a well-established line of flash research, and it has measured results:

Work	Mechanism	Measured lifetime gain
WARM (Luo, Cai, Ghose, Choi, Mutlu — MSST 2015)	Group write-hot data physically, apply retention relaxation selectively	**3.24x** alone; **12.9x** with adaptive refresh
HeatWatch (Luo et al. — HPCA 2018, real 3D NAND)	Exploit device-level self-recovery and temperature awareness	**3.85x** lifetime; RBER reduced **93.5%**

Both were demonstrated relaxing retention to the scale of *days*. A KV tier needs *seconds*. The direction is unambiguous and the magnitude is large, but **no published figure exists for seconds-scale retention on 3D SLC**, and this addendum does not extrapolate one [Author analysis].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

6.5.4 The thermal inversion

Section 5 treated an 80 °C junction as a reliability liability: retention degrades roughly two-fold per 10 °C, and cross-temperature swings widen threshold-voltage distributions. **For a tier that needs 97 seconds of retention, that cost is approximately zero.**

And the same physics runs the other way. HeatWatch's result rests on the **self-recovery effect**: trapped electrons gradually de-trap during idle time between program/erase cycles, partially restoring the tunnel oxide — and the effect is *accelerated at elevated temperature*. A hot stack heals faster between writes.

> The thermal environment that makes HBF look dangerous for a storage-grade workload makes it look *better* for a KV-grade one. The same 80 °C that destroys warranty retention accelerates the recovery that buys back endurance — and the retention being destroyed is retention the workload never needed [Author analysis].

6.5.5 What this does to the published endurance finding

HBF Sucks! derives a 21.7 TB/day write budget for a five-year target using what it calls a "deliberately favorable" TLC assumption, then measures 48–140 TB/day across four production traces — 2.2x to 6.5x over, giving 0.56x an SSD pool's life. But that budget has a warranty-grade retention requirement baked into it. Applying the measured relaxation factors above:

Relaxation applied	Implied budget	Covers the measured traces?
None (as published)	21.7 TB/day	No trace
WARM, relaxation only (3.24x)	70.3 TB/day	Some traces
HeatWatch, self-recovery (3.85x)	83.5 TB/day	Some traces
WARM + adaptive refresh (12.9x)	**279.9 TB/day**	**All four, with ~2x headroom**

[Author analysis. The multipliers are measured results from the cited work, applied to a different device generation and a different retention target than they were demonstrated at. They indicate direction and rough magnitude, not a prediction.]

The published endurance objection may therefore be an artifact of specifying a cache against a storage warranty. That does not rescue TLC — Section 5.6 disqualifies it on read disturb alone, which relaxation does not address — but it substantially weakens the claim that a KV-resident HBF tier is doomed to wear out inside a rack life.

6.5.6 What the paradigm costs

Three prices, and they should be stated as plainly as the benefit [Author analysis]:

- **Non-volatility, which is the thing HBF markets.** A cell driven to seconds-scale retention does not survive a power cycle. SanDisk's positioning emphasises flash persistence; for the KV use case that persistence is precisely what you would deliberately trade away. A relaxed-retention HBF region is a volatile cache that happens to be built from flash.
- **Partitioning, and therefore FTL complexity.** Weights must persist; KV need not. A stack holding both cannot run in one retention mode, so the flash-management layer must maintain physically separated regions at different retention targets — which is exactly WARM's write-hotness grouping, now mandatory rather than optional. More logic on a base die the parent report already flagged as gate-heavy.
- **A specification that does not exist.** No public HBF endurance or retention figure has been disclosed at all. Whether the contributors intend a relaxed-retention mode — and whether the interface exposes retention as a settable property per region — is unknown and is the single most consequential open question in this section.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

6.6 Anticipated HBF AFR, and why the question needs decomposing

No HBF AFR has been published, and this addendum declines to invent one. What can be said is that three mechanisms are routinely conflated and only one of them is a failure rate:

Mechanism	Physics	Is it AFR?
Retention loss	Arrhenius; roughly 2x per 10 °C	**No** — recoverable by scrub, and per 6.5 largely irrelevant for KV
Raw bit error rate	Expected, budgeted, absorbed by ECC	**No** — HBF's high RBER relative to DRAM says nothing about failure rate if ECC is sized correctly
Hard failure	Spare-pool exhaustion; TSV, microbump and solder-joint fatigue following Coffin–Manson on thermal *cycling*	**Yes**

The operative distinction is that the packaging component is driven by **ΔT excursion and cycle count, not by absolute temperature**. A stack held steadily at 80 °C is in better shape than one swinging 40 °C with accelerator load — which is exactly the duty cycle an on-package part experiences, and exactly the mechanism behind the cross-temperature RBER problem in Section 5.6. One system variable, two failure modes. It is notable that Databricks already monitors the *thermal gradient between GPU core and HBM* as a standing fleet health check [Reported].

> **The defensible formulation.** HBF's packaging-driven AFR should be comparable to or somewhat worse than HBM's, for structural reasons: sixteen-high versus eight-to-twelve, more joints, more warpage. Its media-driven contribution is **not an AFR at all** — it is a scheduled death whose date is set by write volume and retention target, which are **workload and firmware properties, not device properties**. Any vendor quoting a single HBF AFR without stating the workload should be asked which one they mean.

7. Ecosystem: the gating item has moved, but not cleared

The July report named one thing as the gate on the entire thesis: a disclosed xPU design win. Its position on 8 July 2026 was that none existed. That remains true. What has changed is that the consortium now has demand-side members.

Google and Tenstorrent joined during the standardisation process [Confirmed — SK hynix; StorageReview]. Google DeepMind is additionally on the FMS panel discussing HBF's role in the memory wall. This is the first participation by organisations that actually design accelerators, and it is the most meaningful ecosystem development since the August 2025 MoU. It is not a design win. A consortium seat commits nobody to silicon.

The absences remain conspicuous. **NVIDIA, AMD, Broadcom, Marvell, Intel, Qualcomm, Micron and Samsung are all outside the consortium** as of this writing [Author analysis, from the disclosed membership list]. NVIDIA's absence is the one that matters most, and it is consistent with the parent report's Section 8 note that Rubin CPX — the part whose profile best fit an HBF-like tier — was reportedly removed from the roadmap at GTC in March 2026.

Samsung has responded with an alternative rather than joining. At the same conference it previewed zHBM and zNAND-O concepts aimed at the same inference memory wall [Reported — StorageReview]. The parent report anticipated a competitive response on roughly this timeline. The strategic logic of publishing early and openly through OCP is now visible: HBF is trying to become the default before a rival specification can define the category [Author analysis].

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

Two adjacent disclosures from the same week bear on the report's manufacturing sections. SK hynix showed a tenth-generation **375-layer 4D NAND** claiming 2.5x performance per watt, with enterprise SSD mass production targeted for early 2027 [Confirmed — SK hynix]. Sandisk and Kioxia announced **BiCS10 QLC** at 332 active layers and more than 37 Gb/mm², built on the CBA bonded-wafer process [Reported — StorageReview]. Both sit inside the 200–400 layer band the report assumed for TSV-perforated arrays, and the continued centrality of bonded-wafer processing supports the report's Section 3 argument that each HBF array wafer consumes a companion CMOS wafer.

8. Corrections register

For readers auditing the parent report against this addendum:

Report section	Claim	Disposition
2.3, 3.x	Unique wafers; SLC; bonded CMOS wafer per array wafer	**Confirmed**, and reinforced by BiCS10 CBA and V10 disclosures
2.4	512 GB from 16-high stacks of 32 GB core dies; 8- or 16-high	**Confirmed**
2.5	Gen1 read bandwidth of 1.6 TB/s per stack	**Superseded.** Three grades, ~0.4 to 3.0 TB/s. 1.6 TB/s is probably the middle grade [Author analysis]
4.2	Base logic die materially more complex than HBM's	**Confirmed**
4.4	UCIe/serial die-to-die ranked as fallback #2 behind an on-interposer HBM site	**Inverted.** UCIe is the specified interface. The report's ranking was wrong
5	Placement on the same interposer as the xPU	**Confirmed in substance.** Advanced-package UCIe keeps HBF on the interposer or a high-density organic substrate
7	Shoreline is the binding constraint; site economics govern adoption	**Confirmed**
4.2	Host needs custom controller silicon "even if the PHY can be reused"; logic more gate-heavy than HBM	**Confirmed, and load-bearing.** UCIe supplies the PHY, not the flash protocol. This is the claim that survives UCIe adoption
7	"HBF is gated on close GPU vendor collaboration" because the host needs bespoke silicon	**Partially weakened.** PHY and interface-standardisation barriers removed; controller barrier intact; co-timing requirement reduced but not eliminated
7	Compute-die shoreline is the binding constraint on adoption	**Displaced.** A UCIe chiplet can hang off a companion I/O die, moving the constraint to interposer area and reticle limits
2.3, 3.x	SLC as the first media, expected to relax toward TLC for a later "capacity HBF"	**Revised.** SLC is structurally required at these bandwidths; a TLC variant is viable only far below Grade 1 (Sec 5.6–5.7)
7	Shoreline efficiency of HBF vs HBM left unquantified	**Resolved.** ~19x HBM4 capacity per mm at 0.91x bandwidth density (Sec 5.2)
—	Power and thermals not treated in the parent report	**New.** Grades are thermally limited; static power is a fixed term, not per-GB (Sec 5.3.1–5.4)
11	No endurance or DWPD guarantee published	**Still open.** Not addressed in any disclosed summary of the specification
—	Field reliability and the replacement unit not treated	**New.** HBF is not field-replaceable; wear-out scraps the accelerator, which sets the endurance target (Sec 6.1–6.4)

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

Report section	Claim	Disposition
11	Endurance framed against storage-grade retention	**Challenged.** KV$ P99 lifespan is 97 s against a months-to-years warranty spec; relaxed-retention operation is the more appropriate paradigm (Sec 6.5)
11	Latency sensitivity could shrink HBF toward “weights only”	**Still open, and more live** under a die-to-die protocol
12	2H 2026 samples, early 2027 device samples, 2028 revenue silicon	**Unchanged.** No sample shipment has been announced

9. What to watch next

Window	Item	Why it matters
Near term	The OCP document becoming publicly retrievable	Resolves the “xPU–HBF vs UCIe” naming question, the per-package vs per-subsystem ambiguity at 3.0 TB/s, and whether the middle grade is near 1.6 TB/s
Near term	Whether the HBF protocol rides UCIe streaming mode	Tests the Section 2.3.1 prediction, and determines whether designing to HBF means instantiating known IP or absorbing a new interface
2027	Whether first implementations attach HBF to the compute die or to a companion I/O die	Decides whether compute-die beachfront or interposer area is the real ceiling on capacity per accelerator — the Section 2.3.3 relocatability question
Near term	Whether UCIe-S on organic substrate is genuinely supported	If it is, HBF partially escapes CoWoS-class interposer scarcity — which the parent report treated as a binding constraint. This would be a larger cost story than the interface change itself
2H 2026	**First HBF sample shipment announcement**	**The schedule has been reaffirmed twice without a sample. The window is now open**
Any time	An accelerator announced with an HBF site	The gating item. Still unmet. Google and Tenstorrent are the likeliest sources
Any time	Whether the spec exposes retention as a settable per-region property	The single most consequential open question in Sec 6.5. Decides whether relaxed-retention KV operation is available at all
Any time	Any published inference-fleet component reliability breakdown	All current field data is from a 2024 training cluster; inference duty cycles differ materially (Sec 6.1)
Any time	An SLC read-disturb figure in reads-per-block	The weakest input to Sec 5.6. Sets the refresh interval and therefore whether SLC’s 18% endurance headroom is right
Any time	A static-power figure for an HBF stack	Decides whether Grade 1 is a 37 W device or a 140 W one — the difference between dropping into an HBM site and needing new cooling
Any time	Whether H3’s ~160 W per cube is peak dynamic or total	Reconciles the three incompatible power models in Sec 5.3
Any time	Any endurance or DWPD figure	Determines whether HBF can hold KV cache or only weights, which sets the addressable share
Any time	Disclosed grade mix or per-grade pricing	Directly sets which column of the Section 4 table governs the supply shock
2027	Samsung zHBM / zNAND-O progressing to specification	A competing standard would fragment the category and undercut the OCP land-grab

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided “AS IS.”

10. How this addendum was made

The method is the same as the parent report's, and the sequence is the same: domain expertise first, machine assistance second, verification third, human judgment last.

The analyst. Shimogoryo Base LLC is owned and operated by **Andy Kowles**, its Principal Analyst and the author of this addendum and of the parent report. The analytical frame applied here — testing a vendor bandwidth claim by decomposing it into per-die and per-mini-array requirements, and coupling array subdivision to bit yield — is his own, developed through independent research into memory and storage tiers optimised for AI inference, and published in the July report before this specification existed. That prior frame is what made it possible to say something about the August announcement beyond restating the press release.

The instruments. Large language models were used, under direction, for retrieval across the trade coverage and for drafting. They were not asked to supply the analysis. The two substantive contributions here — the grade-to-mini-array sizing and the grade-dependent penalty table — are applications of the parent report's own published method to new inputs, and the arithmetic was computed and checked independently rather than generated as prose.

Verification. Every factual claim about the specification was checked against SK hynix's own announcement and at least one independent outlet, and is labelled by confidence. Where sources conflict — the OCP publication status, the interface naming, the per-package question — the conflict is reported rather than resolved by preference. **The specification document itself was not read**, and every conclusion here is provisional on that.

> **On being wrong in public, and on over-correcting.** The parent report ranked UCIe as a fallback and called it plausible only for mid-range parts. That was wrong, and this addendum says so in the same document that carries the corrected analysis rather than quietly restating the position.
>
> An earlier draft of this addendum then made the opposite error, asserting that UCIe adoption "substantially defeats" the report's gating argument. That was too strong. On review, the report had separated the PHY barrier from the controller barrier and conditionalised explicitly on PHY reuse — so the claim UCIe actually defeats is one the report never leaned on, while the claim it identified as load-bearing survives untouched. Section 2.3 now states the narrower and more defensible version, and Section 2.3.4 itemises which barriers fall and which do not.
>
> A third correction was made during review of Section 5. An earlier working draft computed read disturb on a per-page basis. That is the wrong level: reading a page applies Vpass to every *other* wordline in the block, so disturb is a block-level effect and vendor budgets are quoted per block. The corrected framing raises the modelled exposure by the block's page count — and, because SLC absorbs it while MLC and TLC do not, strengthens rather than weakens the section's conclusion. A fourth judgment was reversed twice: the critique of MemExplorer's static-power figure was made, retracted on an arithmetic coincidence, then reinstated when the thermal model showed that a 115 W idle draw would exceed the modelled junction limit before the device did any work.
>
> These corrections are recorded because a reader has no way to calibrate an analyst who publishes only confirmations, and no way to trust a correction that overshoots. The confidence labelling exists so that claims like these — **[Author analysis]**, not **[Confirmed]** — can be scored after the fact.

Important Disclosures & Disclaimer

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

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

Publisher; copyright and permitted use

The Memorandum is published by **Shimogoryo Base LLC, a Colorado limited liability company** (the "Publisher"), and was prepared by Shimogoryo Base Consulting. © 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 Memorandum**, 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 stated caveats; to presenting the analysis as your own or as that of any other person or firm; or to incorporating the Memorandum 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.

The companion workbook is not covered. `HBF_Investment_Screen.xlsx` and the underlying research report are separate, licensed products and are **not** included in this permission.

Not investment advice; no recommendation

The Addendum is provided for general informational and educational purposes only. **Nothing in the Addendum constitutes, or should be construed as, investment advice, financial advice, trading advice, a personal recommendation, or an opinion regarding the suitability of any security or investment strategy for any particular person.** Nothing herein constitutes legal, tax, accounting, or engineering, design, procurement, or product-qualification advice, nor is it a substitute for professional advice tailored to your particular circumstances.

The Publisher is **not** a registered investment adviser, broker-dealer, commodity trading advisor, research analyst, or fiduciary to any recipient, and is not acting in any such capacity in preparing or distributing the Addendum. The Addendum has not been prepared in accordance with legal requirements designed to promote the independence of investment research and is not subject to any prohibition on dealing ahead of its dissemination.

The company names, tickers, roles, and technical assessments appearing in the Addendum — including any language such as "opportunity," "thesis," "exposure," "tier," or "verdict" — are analytical labels describing supply-chain position and sensitivity to the mechanism analysed. They are **not** buy, sell, hold, overweight, underweight, or price-target recommendations, and they should not be read as any assessment of the investment merit, valuation, or expected return of any security. You should not rely on the Addendum as the sole or primary 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. **Investing involves risk, including the possible loss of principal; past performance is not indicative of future results.**

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

No offer or solicitation

Nothing in the Addendum constitutes an offer to sell, or the solicitation of an offer to buy, any security, financial instrument, or other asset, in any jurisdiction, nor an offer to provide investment advisory services in any jurisdiction in which such an offer would be unlawful. The Addendum 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, Western Digital, Seagate, Rambus, Kulicke & Soffa, Advantest, Teradyne, Entegris, NVIDIA, AMD, Broadcom, Microsoft, Huawei, YMTC, or CXMT. It is directed only at recipients who are capable of evaluating its contents independently, and it does not take account of the investment objectives, financial situation, tax position, or particular needs of any recipient.

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 Memorandum 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 Memorandum 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 Memorandum is published in part to demonstrate the firm’s work and attract clients. Readers should weigh it accordingly.

Sources, estimates, accuracy and completeness

The Addendum is based on the Publisher’s own prior research report, together with publicly available information — vendor press releases and announcements, conference disclosures, regulatory filings, trade-press reporting, academic preprints and publications, and market data — 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 Addendum.

Modelled figures are estimates, not disclosures. The wafer-diversion model, the compound bit-penalty factors, the merchant-supply scenarios, and any derived percentages are the Publisher’s constructions built on stated assumptions, several of which are expressly labelled as the author’s own estimates rather than vendor-disclosed values. They are presented as ranges to convey that uncertainty. Small changes in those assumptions can produce materially different results. No vendor, exchange, index provider, or data provider has reviewed, endorsed, or confirmed the model or its outputs.

Market data. Prices, market capitalisations, trading volumes, valuation multiples, and performance figures are as of the dates stated, are drawn from third-party sources, are unaudited, and will change. Figures relating to ADRs and OTC-quoted instruments may reflect limited liquidity, wide spreads, delayed quotes, currency effects, and ratio changes, and may diverge materially from the issuer’s home-market listing.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided “AS IS.”

Forward-looking statements

The Addendum contains forward-looking statements, projections, scenarios, timelines, and estimates — including, without limitation, statements regarding vendor product roadmaps, sampling and commercialization schedules, standardization outcomes, wafer allocation, bit supply, market adoption, pricing, cost trajectories, competitive strategy, and the prospective sensitivity of particular companies to those developments. 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 dates indicated, 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, ticker symbols, and trademarks referenced in the Addendum — including, without limitation, SanDisk, Western Digital, Seagate, SK hynix, Samsung, Micron, Kioxia, XL-FLASH, BiCS, Rambus, Kulicke & Soffa, Advantest, Teradyne, Entegris, NVIDIA, AMD, Broadcom, Microsoft, Huawei, YMTC, CXMT, TSMC, Intel, JEDEC, Open Compute Project (OCP), NYSE, and Nasdaq — 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 Addendum has reviewed, approved, sponsored, or endorsed the Addendum.

No warranty; limitation of liability

THE MEMORANDUM 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, TRADING LOSSES, LOST PROFITS, LOST REVENUE, LOSS OF DATA, OR BUSINESS INTERRUPTION — ARISING OUT OF OR IN CONNECTION WITH THE MEMORANDUM 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 MEMORANDUM SHALL NOT EXCEED THE GREATER OF (A) THE AMOUNT ACTUALLY PAID BY YOU TO THE PUBLISHER FOR THE MEMORANDUM, WHICH FOR THIS FREE EDITION IS NIL, AND (B) USD 100.

Publication date; no duty to update

The Addendum reflects information available to the Publisher as of 28 July 2026, resting on research verified 8–9 July 2026, and issued in this edition on 14 August 2026. Developments occurring after those dates — including price movements, corporate actions, vendor announcements, and standardization decisions — may materially affect the analysis, conclusions, and projections contained herein. The Publisher assumes no obligation to update, supplement, or correct the Addendum.

Recipients and jurisdiction

The Addendum is intended for distribution only where such distribution is lawful. It is not directed at, and is not intended for use by, any person or entity who is a citizen or resident of, or located in, any jurisdiction where such distribution, publication, or use would be contrary to law or regulation, or would subject the Publisher to any registration or licensing requirement. Recipients are responsible for informing themselves of, and observing, any applicable restrictions.

© 2026 Shimogoryo Base LLC · Free edition — share complete and unaltered, with attribution.
Not investment advice; not a recommendation or an offer to buy or sell any security. The Publisher holds positions in securities mentioned — see Positions and conflicts of interest. Contains estimates and forward-looking statements. Provided "AS IS."

Distribution

The Memorandum 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 investment objectives, financial situation, tax position, 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 · Shimogoryo Base Consulting · shimogoryobase.com/consult.html

If this was useful, share it. If you want the full research report, 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.