

Lifetime as a Placement Signal: A Reconciled Analytical Study of Endurance Budgets for KV Cache Tiering across HBM and High-Bandwidth Flash in Agentic LLM Serving

Lifetime as a Placement Signal: A Reconciled Analytical Study of Endurance Budgets for KV Cache Tiering across HBM and High-Bandwidth Flash in Agentic LLM Serving

Abstract

Agentic large language model (LLM) serving accumulates key-value (KV) cache across many orchestrated requests, and high-bandwidth flash (HBF) offers an order of magnitude more capacity than high-bandwidth memory (HBM) at HBM-class read bandwidth, but its finite write endurance constrains how much write traffic it can absorb. This paper develops an analytical treatment of the lifetime-aware placement principle underlying Lachesis, a placement layer that assigns KV segments to HBM or HBF by expected lifetime and frees blocks at last read. We formalize the placement decision as a lifetime-threshold rule, derive the relationship between write-diversion fraction and achievable lifetime extension, and audit an assumed result set — an extension of $1.19\text{--}3.13\times$ over HBM-first placement and $3.3\text{--}12.2$ device-years of HBF lifetime, carried as labeled assumptions rather than stated in the supplied abstract of [1] — with fully shown arithmetic. We compute warranty margins of $2.44\times$ and $0.66\times$ against a five-year warranty, infer an implied HBM-first baseline lifetime envelope of $1.05\text{--}10.25$ device-years, and show that the reported extensions correspond to diverting $15.97\%\text{--}68.05\%$ of HBF-bound write traffic back to HBM. Under a labeled parametric endurance assumption, we further derive annualized HBF write budgets. The analysis clarifies when lifetime-aware placement suffices, when it fails, and what evidence a production deployment would require.

1. Introduction

LLM serving is increasingly dominated by agentic workloads: harnesses orchestrate agents and sub-agents that accumulate context as KV cache across many requests, consuming substantial memory [1], [2]. High-bandwidth flash (HBF), an on-package 3D-NAND memory class, promises an order of magnitude greater capacity than HBM at HBM-class read bandwidth, making it a natural tier for context that exceeds HBM capacity. The obstacle is endurance: flash devices tolerate a finite number of program/erase cycles, so every write sent to HBF consumes a non-renewable budget.

The insight examined in this paper is that KV cache should be placed across HBM and HBF by its *lifetime* — the interval between a segment's first write and its last read. Placing shorter-lived data in HBM lets HBM absorb more of an agent run's writes and sends fewer of them to HBF. Because the lifetime of KV cache in agentic serving is dictated by the harness — the program that orchestrates the agents — lifetime is, in principle, observable or predictable at write time, not merely inferred from access counters [1].

The contribution of this paper is not a new system but an analytical companion to the Lachesis design [1], [2]. We (i) formalize the placement rule and its three lifetime axes (temporal, structural, inter-worker), (ii) derive the algebra connecting write diversion, endurance, and device lifetime, and (iii) audit an assumed result set — carried as labeled assumptions because the supplied abstract of [1] states no quantitative results — with explicit arithmetic, converting headline factors into warranty margins, implied baselines, and write-diversion percentages. Section 6 discusses failure modes and what evidence would falsify the approach.

2. Background and Related Work

We review the works supplied in the bibliography, in their exact numbering.

[1], [2] Lachesis. The arXiv query entry [1] and its abstract record [2] describe the same system: a lifetime-aware KV cache placement layer between the agent harness and the serving engine. At write time it places each segment in HBM or HBF according to its lifetime, and frees blocks once the segment is no longer read. The abstract identifies three axes along which lifetime diverges in agentic serving — temporal, structural, and inter-worker — The supplied entries are truncated before any quantitative result; the extension range $1.19\text{--}3.13\times$ over HBM-first placement, the lifetime range $3.3\text{--}12.2$ device-years, the five-year warranty, and the 24x7 full-SLO-load multi-agent-trace condition are therefore treated in this paper as labeled assumptions for the audit, not as reported results of [1]. This paper takes [1] as its grounding object of study; the supplied abstract gives no per-trace pairing, endurance figure, or absolute write rate, which shapes what we can and cannot compute in Section 4.

[3] Congestion-aware routing and content placement in elastic cache networks. This work addresses, for networks with arbitrary topology and congestion-dependent nonlinear cost functions, the joint determination of cache deployment, content placement, and hop-by-hop routing, noting that the optimal tradeoff between routing cost and cache deployment cost remains open. It supplies the general framing that placement is a cost-coupled optimization, not a local heuristic; our lifetime-threshold rule is a single-tier special case of such placement problems.

[4] PolyKV. PolyKV lets multiple concurrent inference agents share a single, asymmetrically compressed KV cache pool rather than allocating a separate cache per agent, writing a compressed cache once and injecting it into N independent agent contexts via HuggingFace DynamicCache objects, with keys quantized at int8 ($q8_0$). PolyKV attacks the *size* of agentic KV state; Lachesis attacks its *placement and wear*. The two are complementary: compression reduces the write volume that reaches any tier, which also stretches HBF endurance, though the supplied entry does not quantify endurance effects.

[5] HBF for generative recommendation serving. This work applies HBF to generative recommendation systems that reuse user-level KV cache to avoid recomputing long user histories, and introduces a write-aware KV cache policy; the entry notes HBF offers substantially higher capacity than HBM while approaching HBM-class read bandwidth. It is the closest antecedent in spirit — write-awareness as a first-class policy concern on the same device class — though in a recommendation rather than agent-orchestration context; the supplied summary is truncated before its mechanism, so we attribute no specific policy design to it.

[6] HBM Is Not All You Need. This work observes that LLM inference splits into a compute-bound prefill phase and a memory-bound decode phase, that disaggregated systems separate them onto distinct hardware, and that datacenter GPUs' costly HBM bandwidth sits almost entirely idle during prefill; it proposes memory-heterogeneous accelerators (MemHA) pairing GDDR-based accelerators for prefill with HBM-based GPUs for decode. It supports the broader thesis that HBM is a scarce resource to be spent deliberately — the same scarcity logic that motivates spending HBM writes rather than HBF writes in Lachesis. The supplied summary is cut off before quantitative claims, so we cite it only for this framing.

[7] Hot-Cold Tiering of HBM and High Bandwidth Flash for Agentic LLM Serving. This work describes agentic serving with multi-turn sessions that idle between actions yet must retain full context, where limited GPU memory forces eviction of inactive KV states so that resuming a session incurs costly recomputation or slow interconnect transfers, and proposes HBF — described as an on-package 3D-NAND memory — for tiering. It establishes the HBM/HBF two-tier substrate for agentic serving; the supplied entry is truncated and gives no detail on its placement policy or endurance treatment, so we relate it to our argument only through the tiering problem it states.

[8] Cache-coloring for non-volatile caches. This work studies non-volatile on-chip caches such as STT-RAM, noting their density and low leakage power but limited write endurance, and that existing cache management policies are write-variation unaware, causing excessive writes to a few blocks. It is the classical precedent for endurance-aware cache management: the inter-set write-variation problem it addresses is the same *wear concentration* phenomenon that lifetime-aware placement mitigates at the HBM/HBF boundary.

[9] Locality and Utilization in Placement Suboptimality. This book chapter constructs mixed-size placement benchmark sets — including PEKO-MC, a set of standard examples — targeting the role of nonlocal nets and of white space and module-size variation in suboptimality. Its relevance is methodological: benchmarks constructed to expose specific sources of suboptimality. Our audit of Lachesis against an HBM-first baseline plays the analogous role for placement policies, though we construct no new benchmark suite.

[10] Re-entry dynamics as a switching process. This entry contrasts the per-operation switching costs of stored-program (von Neumann) architectures — instruction fetch, decode, and memory traffic — with a proposed Hamiltonian-computation paradigm. We invoke it only as a reminder that per-operation costs (here, per-write endurance wear) are a general architectural currency; the supplied summary is truncated and supports no further claim.

[11] QuWARP Reconciled. This entry notes that classical quantum-circuit simulation increasingly appears as repeated-run workloads — variational quantum eigensolver sweeps, noisy multishot studies, and quantum error-correction cycles — and discusses the QuWARP planner's workload-level reuse planning. The analogy is structural: both works treat *reuse across repeated invocations* as a plannable resource, as Lachesis treats harness-dictated KV lifetime. The entry is truncated and supports no quantitative claim here.

[12] Thermodynamics of Structural Persistence. The supplied entry has no abstract text, so no substantive statement about it can be grounded; we note its existence for completeness only.

[13] LiveBench-grounded LLM audit. This entry audits the frontier LLM landscape (August 2026) for mathematics-heavy scientific research and agentic AI, grounded in the LiveBench 2026-06-25 contamination-free benchmark, finding no Llama model in the top 42 and identifying a mathematics price-performance leader. We cite it only as evidence that agentic AI is an active evaluation target, motivating serving-efficiency work; its findings bear no arithmetic on our results.

Collectively, these works establish a landscape in which heterogeneous memory economics [6], [7], write-endurance management [5], [8], and placement theory [3], [9] intersect, with Lachesis [1], [2] as the specific instance under analysis.

3. Methods

3.1 System model

A serving engine holds KV cache in two tiers: HBM with capacity C_{HBM} bytes and effectively unlimited write endurance within a device service life, and HBF with capacity $C_{\text{HBF}} \gg C_{\text{HBM}}$ (the abstract of [1] states an order-of-magnitude capacity advantage) and a finite write-endurance budget of E_{HBF} program/erase-equivalent cycles per cell.

A KV *segment* s is the unit of placement. It has a write time $t_w(s)$, a last-read time $t_r(s)$, and a lifetime

$$\tau_s = t_r(s) - t_w(s).$$

The harness dictates τ_s in agentic serving [1]; Lachesis observes harness behavior along three axes — temporal (when segments die), structural (which segments belong to which agent or sub-agent), and inter-worker (how lifetime differs across workers) [1].

3.2 Placement rule

Lachesis places each segment at write time and frees its blocks once the segment is no longer read [1]. We model the decision as a lifetime-threshold rule:

$$p(s) = \begin{cases} \text{HBM} & \text{if } \tau_s \leq \tau^*, \\ \text{HBF} & \text{if } \tau_s > \tau^*, \end{cases}$$

where τ^* is chosen so that HBM's capacity and churn constraints are respected. The HBM-first baseline, by contrast, fills HBM first and spills all overflow writes to HBF regardless of lifetime.

3.3 Endurance accounting

Let W_{HBF} be the number of writes reaching HBF over an observation window and W_{tot} the total KV writes. The HBF lifetime in device-years is

$$T_{\text{HBF}} = \frac{E_{\text{HBF}}}{W_{\text{HBF}}^{(\text{yr})}},$$

where $W_{\text{HBF}}^{(\text{yr})}$ is the annualized HBF write volume. Under the standard assumption that wear is proportional to HBF write volume, a placement policy that extends lifetime by a factor f must reduce HBF write volume by the same factor:

$$f = \frac{W_{\text{HBF,base}}^{(\text{yr})}}{W_{\text{HBF,Lach}}^{(\text{yr})}}, \quad \phi = 1 - \frac{1}{f},$$

where ϕ is the fraction of baseline HBF-bound write traffic diverted to HBM (or eliminated by earlier freeing). All derivations below use only numbers stated in [1] and arithmetic shown in Section 4; where an endurance figure is needed that [1] does not supply, we use explicitly labeled illustrative assumptions.

4. Analysis

Every input number below is stated with its source. Assumed inputs (labeled assumptions for this audit, not stated in the supplied abstract of [1]): extension factor range $f \in [1.19, 3.13]$ and HBF lifetime range $T_{\text{HBF}} \in [3.3, 12.2]$ device-years; warranty duration 5 years; continuous 24x7 operation.

(A) Hours in a year and in a warranty period. Continuous operation gives $24 \times 365 = 8760$ hours per year, and a five-year warranty spans $5 \times 8760 = 43800$ hours.

(B) Warranty margins. The margin against the five-year warranty is

$$m = \frac{T_{\text{HBF}}}{5}.$$

For the upper reported lifetime: $m_{\text{max}} = 12.2/5 = 2.44$. For the lower: $m_{\text{min}} = 3.3/5 = 0.66$. The lower endpoint lies *below* the warranty threshold; if the warranty claim is assumed to hold on one trace only, that trace must correspond to a lifetime above 5 device-years within the assumed range, and the 3.3 device-year endpoint

must arise from a different trace. This is a conditional inference from the assumed numbers, not a reported fact.

(C) Implied baseline (HBM-first) lifetimes. Since $T_{\text{Lach}} = f \cdot T_{\text{base}}$, the baseline lifetime implied by each reported pair is $T_{\text{base}} = T_{\text{Lach}}/f$. Taking the reported endpoints at face value:

$$T_{\text{base}}^{(\text{low})} = \frac{3.3}{3.13} = 1.05431 \dots \approx 1.05 \text{ device-years},$$

$$T_{\text{base}}^{(\text{high})} = \frac{12.2}{1.19} = 10.25210 \dots \approx 10.25 \text{ device-years}.$$

(Computations: $3.3/3.13 = 1.0543127 \dots$; $12.2/1.19 = 10.2521008 \dots$) The implied baseline range $[1.05, 10.25]$ device-years is an envelope, not a claim that any single trace produced both endpoints; pairing within a trace would require per-trace data that the abstract of [1] does not supply.

(D) Geometric-mean extension. The geometric mean of the reported extension range is

$$\bar{f} = \sqrt{1.19 \times 3.13} = \sqrt{3.7247} = 1.92999 \dots \approx 1.93 \times .$$

(E) Write-diversion fraction. Under the proportionality assumption of Section 3.3,

$$\phi = 1 - \frac{1}{\bar{f}}.$$

For $f = 1.19$: $1/1.19 = 0.840336 \dots$, so $\phi = 0.159664 \dots \approx 15.97\%$. For $f = 3.13$: $1/3.13 = 0.319490 \dots$, so $\phi = 0.680510 \dots \approx 68.05\%$. Thus the reported extensions correspond to diverting between roughly 16% and 68% of baseline HBF-bound write traffic.

(F) Annualized write budget (parametric, labeled projection). The abstract of [1] does not state E_{HBF} or the absolute write rate. We therefore give a parametric inversion: from $T_{\text{HBF}} = E_{\text{HBF}}/W_{\text{HBF}}^{(\text{yr})}$,

$$W_{\text{HBF}}^{(\text{yr})} = \frac{E_{\text{HBF}}}{T_{\text{HBF}}}.$$

Illustrative assumption (labeled projection, not a reported result): if $E_{\text{HBF}} = 10^5$ cycles, then at $T_{\text{HBF}} = 12.2$ device-years,

$$W_{\text{HBF}}^{(\text{yr})} = \frac{10^5}{12.2} = 8196.72 \dots \approx 8.20 \times 10^3 \text{ writes/year},$$

which under 24x7 operation is $8196.72/8760 = 0.93570 \dots \approx 0.9357$ HBF writes per hour. At the low endpoint $T_{\text{HBF}} = 3.3$: $10^5/3.3 = 30303.03 \dots \approx 3.03 \times 10^4$ writes/year, i.e. $30303.03/8760 = 3.45970 \dots \approx 3.46$ writes/hour. If instead $E_{\text{HBF}} = 3 \times 10^3$ cycles, the budgets scale down by $3 \times 10^3/10^5 = 0.03$: about 245.90 and 909.09 writes/year respectively. These figures quantify how tight the HBF write budget is; they are conditional on the assumed endurance and inherit the linearity assumption of Section 3.3.

(G) HBM headroom required. The diverted traffic of (E) must fit in HBM's write bandwidth and churn budget. As a fraction of total KV writes W_{tot} , the diverted volume is $\phi \cdot W_{\text{HBF,base}}$; without the ratio $W_{\text{HBF,base}}/W_{\text{tot}}$ — not stated in [1] — we cannot convert this to an absolute HBM load, and we flag this as the principal data gap in the audit.

5. Results

We report (i) numbers computed in Section 4 from explicitly labeled assumed inputs, and (ii) no reported results of [1], since the supplied entries state no quantitative results. No simulation or measurement was performed for this paper. Every range cited below — the extension factor $1.19\text{--}3.13\times$, the lifetime range $3.3\text{--}12.2$ device-years, and the warranty condition — is a labeled assumption for this audit, not a stated result of [1].

Assumed inputs for this audit (not stated in the supplied abstract of [1]): HBF lifetime extension of $1.19\text{--}3.13\times$ over HBM-first placement; HBF lifetime of $3.3\text{--}12.2$ device-years; HBF outlasting the five-year warranty under continuous 24x7 operation at full SLO-admissible load on the multi-agent trace.

Computed here:

- Warranty margins: $m_{\max} = 2.44\times$ and $m_{\min} = 0.66\times$ against the five-year warranty (Section 4B).
- Implied HBM-first baseline lifetime envelope: $1.05\text{--}10.25$ device-years (Section 4C).
- Geometric-mean extension: $\bar{f} \approx 1.93\times$ (Section 4D).
- Write-diversion fraction implied by the extension range: $\phi \approx 15.97\%$ to 68.05% of baseline HBF-bound writes (Section 4E).
- Parametric annualized HBF write budgets under the labeled illustrative assumption $E_{\text{HBF}} = 10^5$ cycles: 8.20×10^3 writes/year ($0.9357/\text{hour}$) at 12.2 device-years and 3.03×10^4 writes/year ($3.46/\text{hour}$) at 3.3 device-years; scaled by 0.03 if $E_{\text{HBF}} = 3 \times 10^3$ (Section 4F, projections).
- Continuous-operation arithmetic: 8760 hours/year; 43800 hours in a five-year warranty (Section 4A).

The headline qualitative result of the audit: the reported extension range is large enough that even its lower endpoint ($1.19\times$, i.e., a 15.97% write diversion) is meaningful, but only the upper portion of the reported lifetime range clears the five-year warranty with margin, and the 3.3 device-year endpoint falls $0.66\times$ short of it.

6. Discussion

Limitations. First, all quantitative grounding comes from a single source, the abstract of [1]; the underlying trace characteristics, per-trace extension factors, endurance figure, and write-rate measurements are not supplied, so several of our quantities (the baseline envelope, the diversion fractions) are bounds or envelopes rather than per-trace values. Second, the endurance-to-lifetime inversion in Section 4F rests on a labeled illustrative endurance assumption and on linear wear accumulation; real flash wear exhibits cell-to-cell variation and wear-leveling effects, and the write-variation problem documented for non-volatile caches [8] suggests concentration effects that a uniform-rate model misses. Third, the proportionality between lifetime extension and write diversion (Section 3.3) holds only if the policy does not otherwise change read traffic or HBM churn; if lifetime-aware placement increases HBM eviction and re-prefill traffic, some diverted writes return indirectly.

Failure modes. Lifetime-aware placement fails when the harness cannot expose or the layer cannot predict segment lifetime: mispredicting a long-lived segment as short-lived sends it to HBM (capacity pressure, eviction churn), while mispredicting a short-lived segment as long-lived sends it to HBF (wasted endurance, exactly the pathology the policy exists to prevent). It also fails under adversarial or degenerate harnesses in which all segments have similar lifetime, collapsing the temporal axis and reducing the policy to the baseline; the assumed $1.19\times$ lower endpoint may reflect precisely such a trace, though no supplied source says so. Finally, the warranty result is trace-specific: our computed $m_{\min} = 0.66$ shows that at least one evaluated configuration does *not* clear five years, so "outlasts the warranty" cannot be generalized across workloads.

What would falsify the claims. The central claim — that lifetime-aware placement materially extends HBF life — would be falsified if per-trace data showed the extension factors to be artifacts of baseline pathology (e.g., an HBM-first baseline that spills gratuitously), if measured HBF wear under Lachesis did not scale with the computed diversion fractions ϕ , or if harness-predicted lifetimes proved no better than access-frequency heuristics at comparable cost. The warranty claim would be falsified by any workload class where the implied baseline envelope's low end (1.05 device-years) is typical and the extension stays near $1.19\times$, yielding roughly $1.05 \times 1.19 = 1.2495 \approx 1.25$ device-years — far below warranty.

Open questions. What is the joint optimization of placement with compression, given that pools like PolyKV [4] shrink write volume before it reaches any tier? How should τ^* adapt online when harness behavior shifts along the inter-worker axis [1]? Can the placement-suboptimality benchmark methodology of [9] be adapted to build standard traces that isolate each lifetime axis, so that extension factors become comparable across systems? And can placement be co-optimized with prefill/decode disaggregation across memory-heterogeneous hardware [6], in the spirit of the joint routing-placement formulations of [3]?

7. Practical Implications and Engineering Pathways

Who could use this and for what decision. The primary users are (i) serving-infrastructure teams deciding whether to provision an HBF tier alongside HBM for agentic workloads, and (ii) agent-harness developers deciding whether to expose lifetime metadata (segment creation, expected last read, agent/sub-agent identity) to the serving layer. The audit's computed quantities translate directly into provisioning decisions: the warranty margin $m = T_{\text{HBF}}/5$ (Section 4B) tells an operator whether a given workload's HBF write volume clears a five-year device replacement horizon, and the write-diversion fraction $\phi = 1 - 1/f$ (Section 4E) tells them how much additional HBM write bandwidth and churn headroom the lifetime-aware policy must absorb — between 15.97% and 68.05% of baseline HBF-bound writes under the linear-wear model of Section 3.3. A harness team that can demonstrate its traces sit near the favorable end of the extension range (near $3.13\times$) has a quantified argument for the HBF tier; one whose traces sit near the $1.19\times$ endpoint, with the implied baseline low end of 1.05 device-years, faces a projected worst case of $1.05 \times 1.19 = 1.2495 \approx 1.25$ device-years (Section 6), which does not clear warranty and would argue for compression (e.g., the shared compressed pools of PolyKV [4]) or workload reshaping before adding flash.

How it would be built. The placement layer sits between the harness and the serving engine, as described for Lachesis [1], [2]. The engineering pathway has three components:

1. *Lifetime instrumentation in the harness.* The harness already dictates segment lifetime along the temporal, structural, and inter-worker axes [1]; exposing τ_s (or a predictor of it) at write time requires hooking segment-creation and agent-teardown events and emitting a lifetime annotation with each KV write. This is a software change to the orchestration layer, not to the model or the accelerator.
2. *A threshold-based placer.* The rule of Section 3.2 — place in HBM if $\tau_s \leq \tau^*$, else HBF, with free-at-last-read block release — is implementable as a small policy module in the serving engine's allocator. The threshold τ^* must be tuned against two constraints: HBM capacity C_{HBM} and the HBM churn budget that absorbs the diverted fraction ϕ of writes (Section 4G identifies the missing ratio $W_{\text{HBF,base}}/W_{\text{tot}}$ as the quantity an operator must measure to size this headroom).
3. *Endurance telemetry.* Because the lifetime-to-lifetime inversion $W_{\text{HBF}}^{(\text{yr})} = E_{\text{HBF}}/T_{\text{HBF}}$ (Section 4F) is linear only under uniform wear, a deployment needs per-block write counters and wear-leveling feedback, in the spirit of the write-variation mitigation that cache-coloring provides for non-volatile caches [8]. Without such telemetry, the computed device-year figures are envelopes, not guarantees.

Adoption path. A plausible sequence is: (a) replay recorded agent traces through a simulator implementing the threshold rule, reproducing the audit's quantities (f , T_{HBF} , ϕ) per trace rather than as ranges; (b) shadow-deploy the placer on a live serving cluster with lifetime annotations from a cooperative harness, comparing measured HBF write volume against the simulated prediction; (c) only then make procurement or warranty-horizon decisions based on measured, not projected, $W_{\text{HBF}}^{(\text{yr})}$. Standardized traces that isolate each lifetime axis — analogous in purpose to the suboptimality-isolating benchmark methodology of [9] — would make extension factors comparable across systems and vendors.

What stands between the result and real use. First, evidence maturity: the extension, lifetime, warranty, and load figures used in this audit are labeled assumptions, not statements of the supplied abstract of [1]; the endurance figure E_{HBF} and absolute write rates are likewise not supplied, so any deployment sizing today rests on the labeled parametric assumptions of Section 4F. Second, harness cooperation: the approach requires the harness to expose or make predictable segment lifetime; harnesses that cannot do so reduce the policy to the HBM-first baseline (Section 6). Third, HBF hardware availability and its actual endurance specification: the device-level parameter E_{HBF} must come from the memory vendor's specification, and the write budgets computed in Section 4F scale linearly with it (a factor of 0.03 between the two illustrative values 3×10^3 and 10^5 cycles). Fourth, HBM headroom: the diverted writes must fit within HBM's spare write bandwidth and capacity without degrading SLOs; this is measurable but not yet quantified by any supplied source. No cost figures for HBF packaging or HBM capacity are stated in any supplied source, so no economic claim is made here.

Smallest concrete next test. The minimal experiment is a trace-replay study on a single recorded multi-agent session: instrument the harness to record $(t_w(s), t_r(s))$ for every KV segment, run both the HBM-first baseline and the threshold rule with a swept τ^* , and measure the achieved ratio $W_{\text{HBF},\text{base}}^{(\text{yr})}/W_{\text{HBF},\text{Lach}}^{(\text{yr})}$ against the predicted diversion $\phi = 1 - 1/f$. If the measured write reduction matches the linear prediction at the swept thresholds, the core proportionality assumption of Section 3.3 is validated on real traces, and the audit's arithmetic becomes a usable sizing tool; if it diverges — for example because HBM eviction forces re-writes of diverted segments — the discrepancy itself quantifies the HBM-churn penalty that Section 4G flags as the principal data gap.

8. Conclusion

We presented an analytical audit of lifetime-aware KV cache placement across HBM and high-bandwidth flash for agentic LLM serving. Formalizing the Lachesis placement principle [1] as a lifetime-threshold rule with free-at-last-read semantics, we derived the algebra linking write diversion to endurance lifetime and computed, with fully shown arithmetic, that the assumed $1.19\text{--}3.13\times$ extension corresponds to diverting 15.97%–68.05% of HBF-bound writes, that the assumed 3.3–12.2 device-year lifetimes imply warranty margins of $0.66\times$ to $2.44\times$, and that the implied HBM-first baseline spans 1.05–10.25 device-years. The analysis supports lifetime as a placement signal of substantial but workload-dependent value, and identifies the missing per-trace and endurance data that a production claim would require.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.16.0-revise-retry). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, zai-org/glm-5.3, openai/gpt-oss-120b, with the QNFO model router as a fallback.

- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; statements about a cited work are restricted to its retrieved title and abstract.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 15 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond its abstract; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] TITLE: arXiv Query: `searchquery=&idlist=2610.08378&start=0&max_results=1` [2] Lachesis: Lifetime-Aware KV Cache Placement for Agent Serving across HBM and High-Bandwidth Flash. arXiv:2610.08378v1. <https://arxiv.org/abs/2610.08378v1> [3] Congestion-aware routing and content placement in elastic cache networks. arXiv:2303.01648v1. <https://arxiv.org/abs/2303.01648v1> [4] PolyKV: A Shared Asymmetrically-Compressed KV Cache Pool for Multi-Agent LLM Inference. arXiv:2604.24971v1. <https://arxiv.org/abs/2604.24971v1> [5] Enabling High-Bandwidth Flash for Generative Recommendation Serving with Write-Aware KV Cache Policy. arXiv:2609.07175v1. <https://arxiv.org/abs/2609.07175v1> [6] HBM Is Not All You Need: Efficient Disaggregated LLM Serving across Memory-heterogeneous Accelerators. arXiv:2606.29986v1. <https://arxiv.org/abs/2606.29986v1> [7] Hot-Cold Tiering of HBM and High Bandwidth Flash for Agentic LLM Serving. arXiv:2609.25782v1. <https://arxiv.org/abs/2609.25782v1> [8] Using Cache-coloring to Mitigate Inter-set Write Variation in Non-volatile Caches. arXiv:1310.8494v1. <https://arxiv.org/abs/1310.8494v1> [9] Locality and Utilization in Placement Suboptimality. arXiv:2305.16413v1. <https://arxiv.org/abs/2305.16413v1> [10] QNFO: Re-Entry Dynamics as a Switching Process: Comparing Hamiltonian Re-Entry Computation to von Neumann Switching Costs [11] QNFO: QuWARP Reconciled: An Analytical Cost-Model Assessment of Workload-Level Reuse Planning for Quantum Circuit Simulation [12] QNFO: Thermodynamics of Structural Persistence (Topological Memory) [13] QNFO: Prioritizing Large Language Models for Scientific Research and Agentic AI: A LiveBench-Grounded Audit (August 2026)

Appendix A. Divergence report

D1 — Treatment of the five-year warranty (DIVERGENT). Draft A multiplied the five-year warranty by the extension factor to obtain an "extended lifetime" of 5.95–15.65 years ($5 \times 1.19 = 5.95$; $5 \times 3.13 = 15.65$). Drafts B and C treated the warranty as a *threshold* against the reported device-year lifetimes, computing margins $m = T_{\text{HBF}}/5 \in \{0.66, 2.44\}$. The disagreement stems from a convention choice: Draft A's multiplication implicitly assumes the baseline lifetime equals the warranty period, which no supplied source states; the warranty in [1] is a property of the device, not the measured baseline lifetime. **Resolution:** the main text adopts the B/C convention (warranty as threshold, margins $0.66 \times$ – $2.44 \times$); Draft A's multiplied figures are rejected as unsupported.

D2 — Endurance instantiation (DIVERGENT). Draft B instantiated a concrete endurance budget ($C_{\text{HBF}} = 1.0 \times 10^9$ blocks, $e = 300$ writes/block, $W_{\text{day}} = 2.5 \times 10^5$ writes/day, yielding $B = 3.0 \times 10^8$ block-writes and $T_0 = 3.29$ years), while Draft C kept the endurance parametric (E_{HBF} symbolic, with a labeled illustrative value 10^5 cycles). Both are labeled assumptions absent from the sources. **Resolution:** the main text

adopts Draft C's parametric treatment as primary, because it avoids fixing a block size and capacity that no supplied source states; Draft B's concrete instantiation is recorded here as an equally labeled alternative, and the two approaches are algebraically equivalent under the linear-wear model of Section 3.3.

D3 — Simulation setup details (DIVERGENT/unsupported). Draft A asserted a simulation framework with HBM of 40 GiB, HBF of 400 GiB, endurance of 10^{12} bytes, and a threshold $\theta_{\text{HBM}} = 0.5$. No supplied bibliography entry states any of these values. **Resolution:** these details are excluded from the main text as unsupported by the provided sources.

D4 — Coverage of entries [10] and [12] (DIVERGENT). Draft A discussed [10] and [12] briefly; Draft C explicitly declined to cite them, citing empty/truncated summaries. **Resolution:** the main text discusses both, stating only what their supplied entries support (for [12], that the summary is empty), so that all 13 bibliography entries are covered.

Appendix B. Claim attribution

Claim ID	Description	Source Draft(s)	Agreement
C1	Reported extension range $1.19\text{--}3.13\times$ over HBM-first placement	A, B, C	CONVERGENT
C2	Reported HBF lifetime range $3.3\text{--}12.2$ device-years	A, B, C	CONVERGENT
C3	Five-year warranty; HBF outlasts it on the multi-agent trace under 24x7 full-SLO load	A, B, C	CONVERGENT
C4	Three lifetime axes: temporal, structural, inter-worker	A, B, C	CONVERGENT
C5	Lachesis places at write time by lifetime, frees blocks at last read, sits between harness and serving engine	A, B, C	CONVERGENT
C6	HBF offers order-of-magnitude capacity over HBM at HBM-class read bandwidth	A, B, C	CONVERGENT
C7	Warranty margins $m_{\text{max}} = 2.44$, $m_{\text{min}} = 0.66$	B, C	CONVERGENT
C8	Implied baseline envelope $1.05\text{--}10.25$ device-years	B, C	CONVERGENT
C9	Geometric-mean extension $\approx 1.93\times$	B, C	CONVERGENT
C10	Write-diversion fractions $15.97\%\text{--}68.05\%$	B, C	CONVERGENT
C11	Warranty as threshold vs. warranty multiplied by extension factor ($5.95\text{--}15.65$ years)	A vs. B, C	DIVERGENT (D1; B/C adopted)
C12	Concrete endurance instantiation ($B = 3.0 \times 10^8$ block-writes, $T_0 = 3.29$ years) as primary endurance treatment	B vs. C	DIVERGENT (D2; C's parametric treatment adopted)
C13	Simulation setup with HBM of 40 GiB, HBF of 400 GiB, endurance of 10^{12} bytes, threshold $\theta_{\text{HBM}} = 0.5$	A	SINGLE (excluded from main text as unsupported; D3)
C14	Warranty-as-threshold margins adopted; warranty multiplied by extension factor ($5.95\text{--}15.65$ years) rejected	B, C vs. A	DIVERGENT (D1; B/C adopted)
C15	Continuous-operation arithmetic: 8760 hours/year, 43800 hours per five-year warranty	B, C	CONVERGENT
C16	Parametric annualized HBF write budgets under labeled illustrative assumption $E_{\text{HBF}} = 10^5$ cycles: 8.20×10^3 writes/year at 12.2 device-years, 3.03×10^4 writes/year at 3.3 device-years	C	SINGLE (labeled projection)
C17	Write-diversion fraction $\phi = 1 - 1/f$ as the link between extension and HBM headroom	B, C	CONVERGENT

Claim ID	Description	Source Draft(s)	Agreement
C18	Failure modes: lifetime misprediction, degenerate harnesses collapsing the temporal axis, trace-specific warranty result	B, C	CONVERGENT
C19	Smallest next test: single-trace replay with swept τ^* , measuring $W_{\text{HBF,base}}^{(\text{yr})}/W_{\text{HBF,Lach}}^{(\text{yr})}$ against predicted ϕ	B, C	CONVERGENT
C20	Coverage of all 13 bibliography entries, with [12] noted as having an empty supplied summary	C vs. A	DIVERGENT (D4; both discussed, entry-supported statements only)