

DISAGGREGATED LLM KV-CACHE STORAGE REQUIRES COORDINATION-FREE CONSISTENCY ABSTRACTIONS

KALYAN PULLELA AND H B ACHARYA

Department of Computer Science, Oklahoma State University
ACM HotStorage '26 • Prague, Czech Republic • September 28–29, 2026

ARCHITECTURAL EVOLUTION

KV-Cache Reuse Across Disaggregated Storage Tiers

Disaggregated LLM serving decouples compute-bound prefill from memory-bound decoding, transferring KV state across networked memory and storage tiers. As workers concurrently place, share, and evict cached blocks, KV catalog metadata becomes a distributed storage namespace.

- Decoupled serving pools separate FLOP-intensive prefill from bandwidth-bound decode instances.
- Shared tiers HBM, host DRAM, and SSD need decentralized candidate discovery without a global-controller bottleneck.
- **Available $\neq$ Safe:** a catalog candidate is not automatically safe for decoder reuse.
- **Asymmetric penalty:** a miss incurs prefill recomputation; an invalid commit can silently perturb model output.



CALM MONOTONICITY AUDIT MAPS THE COORDINATION BOUNDARY

CALM as an operation-level diagnostic:

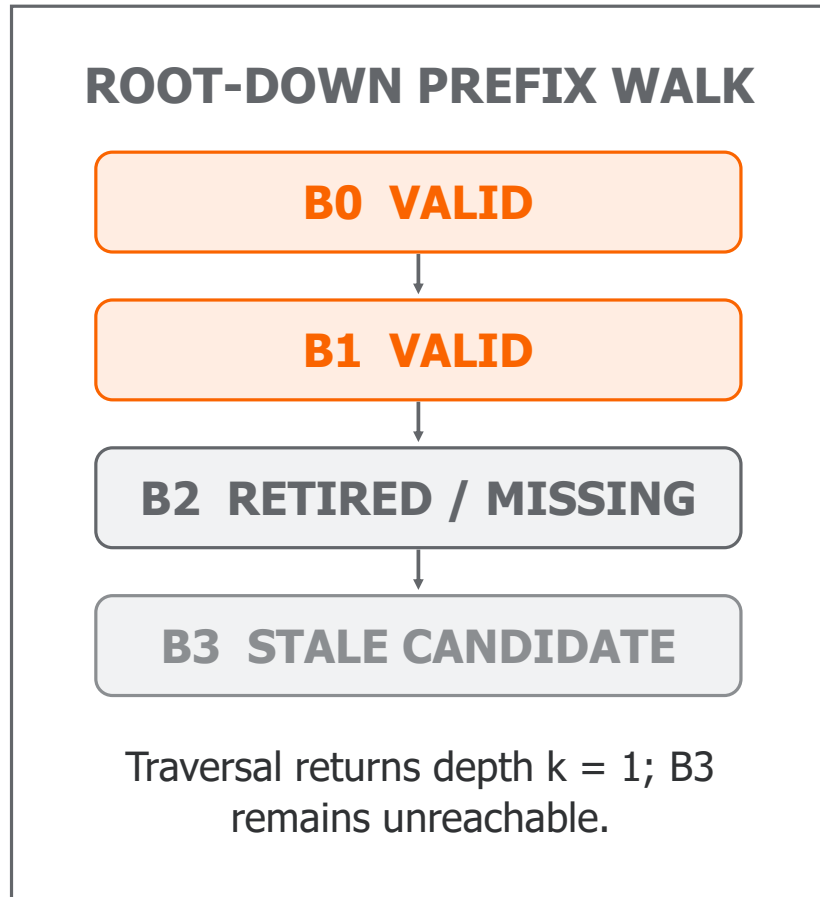
when accumulating positive input facts **cannot** invalidate an existing conclusion, replicas can make progress without global coordination.

We audit **47 normalized metadata operations** across LMCache, Mooncake, vLLM-style paging, and token-selective cache systems.

- **11 monotone in isolation:** placements, hash chains, chunked prefill, copy migrations.
- **36 non-monotone:** evictions, negative assertions, unique-source choices, and mutable counter thresholds.
- **Scope:** the 11/36 split is a boundary diagnostic, not a production-bug census.
- **Takeaway:** decentralize placement admissions; re-encode observed retirements as tagged facts (make them monotone).



ROOT-DOWN TRAVERSAL & THREE-TIER IDENTITY



Root-Down Prefix Safety: a reader returns depth k only after verifying blocks $0 \dots k$; an orphan descendant cannot be committed.

THREE IDENTITY LAYERS MUST AGREE

- **Logical** — (σ, h, p, π) : tenant namespace, prefix hash, parent pointer, and logical position map.
- **Catalog** — $(t, \text{inst}, \text{loc})$: unique admission tag, instance ID, and physical location.
- **Physical** — (g, δ) : sealed allocation generation and payload digest.



TAGGED OBSERVED-REMOVE PLACEMENT CATALOG

ADMIT RULE

admit(a)

Publish only after physical generation g is fully written and sealed.

RETIRE RULE

retire(T)

Append the concrete observed tag set T to R ; never infer deletion from stale local absence.

STATE $S = (A, R)$

A is the grow-only set of admitted records; R is the grow-only set of retired tags.

MERGE = UNION

$(A_1 \cup A_2, R_1 \cup R_2)$ is associative, commutative, and idempotent.

ADD-WINS

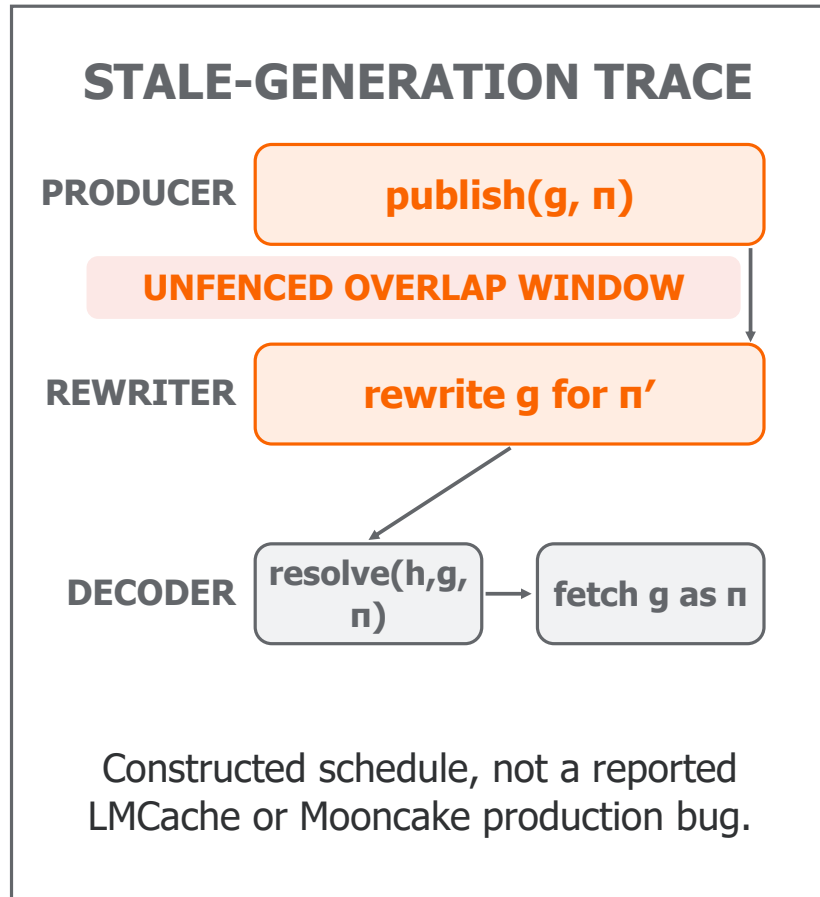
A concurrent or later admission receives a fresh tag $t \notin R$.

SAFE GC

Reclaim retired tags only after acknowledgments, not a fixed TTL.



CONDITIONAL HAZARD: POSITIONAL REWRITE UNDER SHARED GENERATION

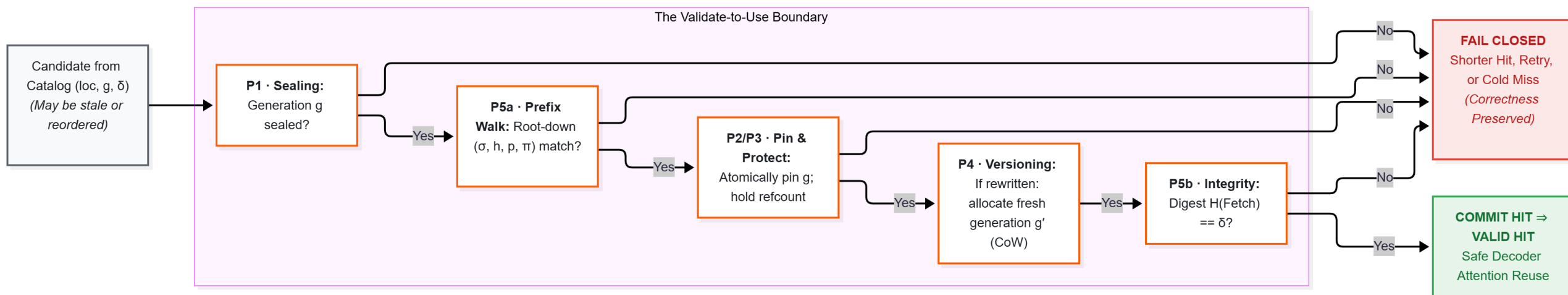


In RoPE-based models, cached key tensors depend on the position at which they were encoded.

- **The race:** rewriting an allocation in place from n to n' under generation g lets overlapping readers resolve the old position map but fetch rewritten bytes.
- **Joint hypotheses H1–H6:** position dependence, in-place reinterpretation, reachable generation, reader overlap, no lease or fence, and no payload-integrity check.
- **Scope:** a constructed hazard showing that catalog convergence alone cannot provide object-level safety.



VALIDATE-TO-USE BOUNDARY



FAILED CLOSED SYMMETRY: Any gate failure safely aborts to a prefix miss; only sealed, pinned, and digest-verified blocks reach decoder attention.



FAULT INJECTION & MODEL-DEPENDENT DIVERGENCE

QWEN2.5-1.5B · HIGH SENSITIVITY

- Eight fixed prompts with injected in-place positional rewrites.
- **8 / 8** unsafe outputs diverged from the fresh-prefill reference.
- **Mean first-token KL: 8.24**
- Every injected task changed decoder output under the unsafe rewrite.

SMOLLM3-3B · OUTPUT RESILIENCE

- The same eight prompts and injected race condition.
- **1 / 8** top-1 tokens changed; all eight final generation texts matched.
- **Mean first-token KL: 0.73**
- The downstream effect of stale state is model- and task-dependent.

**SAFE COPY-ON-WRITE AND PREFIX-EVICTION CONTROLS
MATCHED FRESH PREFILL IN BOTH MODELS**



PROTOTYPE CONVERGENCE & GARBAGE COLLECTION

NO-GC BASELINE

0

RECLAMATION

Tombstones accumulate monotonically, avoiding resurrection but growing without bound.

UNSAFE FIXED-TTL GC

**DELAYED ADD
RESURRECTED**

Premature reclamation reproduced the fault in the 90-case matrix.

4-REPLICA PROTOTYPE

2,000 faulted steps with delays, duplicates, partitions, and rejoins.

0 invalid or cross-tenant committed hits

33.8 MB updates · 19.2 MB full sync

ACK / CAUSAL-STABILITY GC

Reclaim only after all active replicas acknowledge, or causal stability is established.

0 resurrecting faults in the tested execution



AVAILABLE $\neq$ SAFE

DECENTRALIZE DISCOVERY, COORDINATE COMMIT

TLA+ • 1,733 generated / 634 distinct states • depth 16 • 0 invariant violations under the contract
Removing reader protection exposed the validate-to-use race.

