Technical Brief

KV Cache Offload Across Storage Tiers

A Controlled Benchmark on an 8x NVIDIA® B200 Inference Host

Executive Summary

This study's conclusion is a tiering rule, and the measurements size it. WD® Ultrastar Data Center hard drives are a viable KV cache tier for LLM inference when a buffering layer, host DRAM or QLC flash, sits in front of it to absorb the GPU eviction stream and serve the hot recall set. WD and Solidigm™ measured that architecture and its alternatives on a single host with 8x NVIDIA B200 GPUs running a leading 70B parameter open-weights model at BF16 under vLLM 0.22.0 with LMCache 0.4.6, driven by a byte-identical multi-turn replay with an intrinsic prefix reuse ceiling of 71.7 percent. Three remote tiers were tested over NFS: Solidigm QLC flash, a 24-drive Ultrastar HDD array in a WD Ultrastar Data60 Hybrid Storage Enclosure with the same array with asynchronous write-back exports, and synchronous (durable) exports.

Behind a 30 GiB host DRAM tier, all three configurations were statistically indistinguishable from each other and from a local PCIe® Gen 5 NVMe™ SSD at every concurrency tested. At concurrency 32 every configuration reached the full 71.7 percent ceiling at 92.0 to 92.4 tokens/s decode, and the largest between-arm difference anywhere in the buffered scenario was 0.5 tokens/s and 0.11 s TTFT. With the buffer in place, the choice of capacity media behind it is a cost-per-terabyte and capacity decision rather than a performance one.

The buffer is required rather than optional, and the reason is write absorption, not read speed. The eviction stream runs at roughly 0.7 GB/s sustained per 8-GPU host at 320 KiB per token, which exceeds what 24 spindles can reliably commit synchronously, even though the same array reads at 6.34 GB/s in isolation. Removing the buffer therefore produced the expected result for rotating media, and measured it: durable Ultrastar HDD exposed directly to the eviction stream fell to a 19.0 percent hit rate, roughly one quarter of the ceiling, with decode at 34.7 tokens/s, because chunks had not persisted when they were re-queried. Solidigm QLC flash under the same load held the full 71.7 percent ceiling at 78.4 tokens/s, which is why QLC flash also serves as the buffering layer where DRAM capacity is the constraint. Two rules follow: size the buffer to the eviction rate and the hot recall set, and never place a synchronous write path from a GPU eviction stream onto spindles.

Why KV Cache Offload

During inference, a transformer stores per-token attention state, the KV cache, in GPU memory. For the model tested here at BF16, that state costs about 320 KiB per token across the tensor-parallel group. Long prompts and high concurrency inflate the working set past what HBM can hold, at which point the serving stack must either recompute prefixes, burning GPU time, or spill cache entries to a slower tier and recall them when a conversation returns. Multi-turn workloads make recall valuable: each turn re-presents the accumulated prefix, and a tier that can serve it avoids a full prefill.

The external prefix cache hit rate is the fraction of queried tokens served from the cache hierarchy rather than recomputed. The workload used here has an intrinsic reuse ceiling of 71.7 percent. No configuration can exceed it, and reaching it means the tier served every servable lookup. Hit rates below the ceiling represent recompute work pushed back onto the GPUs.

Test Design

Platform: one Supermicro® inference host with 8x NVIDIA B200 GPUs (183 GiB HBM3e each), tensor parallelism 8, vLLM 0.22.0 with LMCache 0.4.6, PyTorch 2.11 (CUDA 13). The model is a leading 70B parameter open-weights model at BF16, the worst-case KV I/O volume. Quantization is orthogonal and stacks with storage tiering rather than competing with it.

Storage configuration: all three remote configurations are identical except for media and export durability. Each is NFSv4.2 only, over TCP, on 400GbE with MTU 9000, with identical client mount options (rsize/wsize 1 MiB, 16 connections, hard mounts), 128 server NFS threads, md RAID0, and XFS with 1 MiB stripe alignment.

1. **Solidigm QLC flash**: 23x 122.88TB[1] Solidigm QLC NVMe SSDs (about 2.8 PB raw), NFS server with 503 GiB RAM, synchronous (durable export.

2. **Ultrastar HDD async**: 24x 18TB 7200 rpm WD Ultrastar DC HC550 SAS HDDs populating a WD Ultrastar Data60 (UD60) 60- bay JBOD (432TB raw as configured), NFS server with 123 GiB RAM, with an asynchronous export, so server RAM acts as a write-back buffer that destages to the spindles in the background. This simulates the write-back cache layer production software-defined storage deployments put in front of HDD pools.

3. **Ultrastar HDD sync**: the same array in the same Data60 enclosure, on the same server, with synchronous export.

The Data60 carries 60 bays and was populated with 24 drives for this study, so the results below describe a partially loaded shelf rather than the platform's ceiling. The raw NFS exports simulate the data path of a software-defined storage layer. Results are an upper bound on media plus protocol, not a product benchmark.

Scenarios: two per arm. B-1 places a 30 GiB DRAM (CPU memory) tier in front of the remote disk. B-2 exposes the remote disk directly with no DRAM tier. B-1 is the recommended tiered deployment, and the operating point the design guidance is built on. B-2 is a deliberate stress case that removes the buffer to isolate what each medium absorbs on its own, and it is the scenario that sizes the buffer. A local anchor, one 15.36TB PCIe Gen 5 NVMe SSD in the inference host, ran the same two scenarios (A-1 and A-2)[CF3.1] inside every arm's pass as a cross-run drift control.

Workload and pressure: multi-turn conversations with roughly 10,000-token prompts, 4 turns, 512 max completion tokens, greedy decoding (temperature 0), recorded once and replayed byte-identically across every arm and replicate. GPU memory utilization was deliberately constrained to 0.15, giving a GPU KV pool of 162,224 tokens, so concurrency 32 and 64 oversubscribe the pool roughly 2x and 4x while concurrency 16 fits the pool and serves as a below-pressure control.

Protocol: independent cold cells (cache wiped, inference server restarted, cold state asserted per cell), 3 replicates per cell, 95 percent bootstrap confidence intervals, per-cell provenance manifests, and telemetry-verified tier reads. The campaign comprised 144 runs across the three arms with zero failures: 36 cells with 3 analyzed replicates each, plus one telemetry attribution run per cell.

Results

At concurrency 16, below pressure, all arms are identical: hit rate 0.0, about 108 tokens/s decode, 0.55 s mean TTFT, storage idle by design. Removing the DRAM tier (B-2) exposes each medium directly to the eviction stream. This is not a recommended deployment. It is the stress case that shows what each medium absorbs on its own and, for the Ultrastar tier, quantifies why the buffer is required. Values are means across 3 replicates with 95 percent confidence intervals in brackets.

B-1, the buffered scenario with the 30 GiB DRAM tier in front, is the headline result and the operating point the design guidance is built on. All three arms were statistically indistinguishable at every concurrency, and indistinguishable from the local NVMe anchor (A-1). The largest between-arm difference anywhere in B-1 was 0.5 tokens/s and 0.11 s TTFT.

B-1 (buffered), all arms	Decode (tokens/s)	TTFT mean (s)
Concurrency 32	92.0 to 92.4	4.13 to 4.17
Concurrency 64	90.6 to 91.1	13.23 to 13.35

At concurrency 32 every arm hit the full 71.7 percent ceiling. The A-1 local anchor at concurrency 32 returned 91.9 to 92.5 tokens/s and 4.17 to 4.20 s TTFT.

B-2 (unbuffered), concurrency 32:

Arm	Hit rate (percent)	Decode (tokens/s)	TTFT mean (s)	TTFT P95 (s)
Solidigm QLC flash	71.7	78.4 [78.1, 78.9][CF1.1]	5.03 [4.97, 5.07]	10.4
Ultrastar HDD async	64.3 [64.0, 64.4]	72.5 [72.4, 72.6]	5.79 [5.73, 5.83]	10.8
Ultrastar HDD sync	19.0 [17.9, 19.7]	34.7 [32.2, 36.2]	14.65 [13.83, 15.43]	26.7

B-2 (unbuffered), concurrency 64:

Arm	Hit rate (percent)	Decode (tokens/s)	TTFT mean (s)	TTFT P95 (s)
Solidigm QLC flash	71.7	75.8 [75.3, 76.3]	16.02 [15.87, 16.14]	22.4
Ultrastar HDD async	61.9 [61.3, 62.5]	50.6 [50.2, 50.8]	33.58 [33.22, 34.04]	59.2
Ultrastar HDD sync	17.2 [16.4, 17.6]	35.2 [34.4, 36.1]	41.36 [39.47, 43.04]	82.7

For context, the local NVMe disk-only anchor (A-2) hit 71.7 percent with decode of 85.2 tokens/s at concurrency 32 and 83.5 to 84.4 tokens/s at concurrency 64.

The between-arm Welch 95 percent confidence intervals are unambiguous. QLC minus Ultrastar sync at concurrency 32 is +43.7 [+41.1, +46.3] tokens/s decode, -9.62 [-10.59, -8.65] s TTFT, and +52.7 [+51.5, +53.9] hit-rate points. At concurrency 64 the decode gap holds at +40.6 [+39.4, +41.8] tokens/s while the TTFT gap widens to -25.34 [-27.53, -23.15] s.

Why the Buffer is Required

The unbuffered sync result is the expected behavior for rotating media under a synchronous, create-heavy write stream at GPU eviction rates, and the telemetry shows exactly why. The limiter is write absorption, not read speed. The Ultrastar array reads well in isolation: 6.34 GB/s raw sequential and roughly 2.2 to 2.9 GB/s under concurrent multi-stream reads, separately measured. But the KV eviction stream runs at roughly 0.7 GB/s sustained at concurrency 32 and above, a single durable 1 MiB write to the array costs about 13 ms, and the durable create-heavy write ceiling is about 692 MB/s at ideal commit granularity, with durable write throughput spanning about 9x depending on commit granularity.

In matched measurement windows the sync Ultrastar arm absorbed only 10 to 18 GiB of KV writes while the async Ultrastar arm absorbed 89 GiB and the Solidigm QLC arm absorbed 84.6 GiB. The read-back telemetry confirms the consequence: the QLC arm served 79.9 GiB of tier reads in-window with full-prefix retrievals, while the sync Ultrastar arm served 4.2 GiB with few, partial retrievals, because chunks had not yet persisted when re-queried. The 19.0 percent hit rate is not a read-latency artifact. The cache simply never contained the data.

The async write-back variant closes the absorption gap entirely, but read-back then competes with destaging on the same spindles: at concurrency 64 its mean TTFT doubles versus Solidigm QLC (33.6 s vs 16.0 s) even though its hit rate stays within 10 points of the ceiling. Absorbing the writes is necessary but not sufficient. The spindles must also serve reads while destaging.

Implications for System Design

Behind a modest DRAM tier, a 24-drive Ultrastar array in a Data60 shelf delivers KV cache offload performance statistically indistinguishable from Solidigm QLC flash and from local NVMe. Recalls at these operating points are served from DRAM, and the media difference disappears entirely. Behind the buffer, the choice of capacity media is a cost-per-terabyte and capacity decision rather than a performance one. Exposed directly to the eviction stream of 8 B200-class GPUs, Solidigm QLC flash holds the workload's reuse ceiling while unbuffered durable HDD, as expected, falls to about one quarter of it.

Tiering is not optional at this scale. One buffering layer, DRAM or QLC flash, in front of capacity Ultrastar HDD recovers nearly all of the performance at a fraction of the all-flash cost. The numbers quantify the sizing: the buffer must absorb a sustained eviction stream of roughly 0.7 GB/s per 8-GPU host at 320 KiB per token, and a 30 GiB DRAM tier was sufficient for full parity at these operating points. Where DRAM capacity is the constraint, QLC flash is the alternative front layer, since the same eviction stream that overwhelms durable spindles was absorbed by the Solidigm QLC arm at 84.6 GiB in matched windows while it held the reuse ceiling.

Two planning rules follow directly. Never put a synchronous write path from a GPU eviction stream onto spindles. And size the buffer to the eviction rate and the hot recall set rather than to the capacity tier behind it.

Limitations

These results come from a single host, a single model configuration, and one workload shape (long-prompt multi-turn), with n=3 replicates per cell. All remote traffic was NFS over TCP; no RDMA path was measured. Server page cache participates in both arms as it would in production, and the two NFS servers have different RAM sizes (503 GiB versus 123 GiB), so in-window read-back on the QLC arm may be partially DRAM-served.

GPU memory utilization was deliberately constrained to create KV pressure on a 1.4 TiB aggregate-HBM platform; production pressure would instead come from larger working sets and longer contexts. The Data60 was populated with 24 of its 60 bays, and a higher drive count in the same enclosure would raise both aggregate read bandwidth and durable absorption.

The asynchronous Ultrastar variant trades durability of the most recent writes for absorption, exactly as write-back caches in software-defined storage do. KV cache contents are reconstructable, which makes this trade acceptable in this application, but it must be stated. The raw NFS exports are an upper bound on media plus protocol, not a claim about any storage product.

Method Notes

Every cell was an independent cold start (cache wiped, inference server restarted, cold state asserted), with the recorded workload replayed byte-identically and greedy decoding removing sampling variance. Headline figures are means across 3 replicates with 95 percent bootstrap confidence intervals, between-arm differences use Welch 95 percent confidence intervals, per-cell provenance manifests tie every number to its run, and tier reads were telemetry-verified rather than inferred. The local NVMe anchor, repeated inside every arm's pass, bounds cross-run drift at 1.3 tokens/s decode and 0.10 s TTFT across the 144-run, zero-failure campaign.

[1] One terabyte (TB) is equal to one trillion bytes and one petabyte (PB) is equal to 1,000 TB. Actual user capacity may be less due to operating environment.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com