Technical Brief

OpenFlex® Data24 4000 Series for Disaggregated Key-ValueCache Offloading

Standalone EBOF as a shared, network-attached KV cache tier for up to 96 H100-class GPUs

KV Cache (Key–Value Cache) is a memory optimization used in transformer-based large language models to accelerate autoregressive inference. For each token generated, K(key) and V(value) vectors are created and stored in KV Cache within the GPU's High Bandwidth Memory (HBM). Being able to access these previously computed tensors helps avoid redundant computation and dramatically reduces latency. There is a memory tradeoff to take into account. KV cache consumes significant GPU memory (dominating HBM usage) which can cause production LLM inference at scale to run straight into the KV cache memory wall.

Serving long-context, multi-turn conversations with a hundred-billion-parameter class models can leave only a thin margin of GPU VRAM for active KV state before concurrency hits a hard ceiling. Operators must choose between buying more GPU nodes, truncating context, or finding a lower-cost overflow tier that does not destroy user-perceived latency.

The Western Digital OpenFlex Data24 addresses that challenge directly.

OpenFlex Data24 4000 Series Storage Platform

The OpenFlex Data24 4000 series NVMe-oF™ storage platform extends the high performance of NVMe flash to shared storage. The 4000 series provide low -latency sharing of NVMe™ SSDs over a high-performance Ethernet fabric to deliver similar performance to locally attached NVMe SSDs. Western Digital RapidFlex™ NVMe-oF controllers allow up to six dual pathed hosts to be attached without a switch.

The OpenFlex Data24 4000 series uses three of Western Digital's RapidFlex A2000 Fabric Bridge Adapters per IOM to provide up to 12 ports of 100GbE which can connect to RDMA and/or RDMA configured host ports.

Benchmarking

Benchmark testing of a 70 Billion Parameter Model running on vLLM with LMCache KV cache management validates a clean 4:1 ratio: a single NVMe namespace on the Data24 sustains KV offload for a four-GPU tensor-parallel inference node at concurrencies up to 16 simultaneous multi-turn conversations. The Data24 houses 24 NVMe SSDs, each presented as an independent namespace over NVMe/RDMA, which means one 2U enclosure can serve as a shared, disaggregated KV cache tier for up to 96 GPUs across 24 inference nodes simultaneously. No additional compute, proprietary fabric, or licensed software is required—only standard 100 GbE networking and the NVMe-RDMA kernel module.

Challenges

- GPU VRAM increasingly exhausted by larger model weights, leaving minimal headroom for concurrent KV cache at long context lengths.
- Local NVMe capacity is coupled to GPU server chassis, preventing independent scaling of compute and storage.
- Adding KV storage (more NVMe Drives) to each GPU server creates compromise on hardware design, cost, rack space, and operational complexity per node.
- Latency-sensitive inference workloads require a storage tier whose access time does not dominate the token generation pipeline.
- Shared storage tiers must serve multiple tensor-parallel nodes concurrently without head-of-line blocking.

Highlights

- 4:1 GPU-to-drive ratio validated: one NVMe namespace supports one 4-GPU inference node at concurrency 16.
- 24 independent NVMe namespaces from a single 2U enclosure serve up to 96 H100-class GPUs across 24 nodes.
- NVMe/RDMA transport over standard 100 GbE — no RDMA NICs, no specialized switches.
- Network hop latency of 10–50 µs is attenuated by the pipelined nature of token generation, preserving user-perceived throughput.
- LMCache DRAM tier (30 GB at TP=4) absorbs hot KV entries, limiting fabric traffic to cold-path evictions and recalls.
- Composable disaggregation: GPU nodes and storage enclosures scale independently across the fleet.

The 4:1 Architecture: One Drive, Four GPUs

The benchmark establishes a straightforward sizing principle rooted in the memory arithmetic of 70B-class model inference. After loading a 70 Billion Parameter Model weights in FP16 (~140 GiB), a four-GPU node with 320 GiB total VRAM retains approximately 180 GB for KV cache and runtime overhead. At the 90% memory utilization setting used in testing, vLLM allocates roughly 148 GB for KV pages. Concurrency at the 32,768-token context ceiling exhausts this budget within a handful of active conversations, making storage-tier overflow essential for any production serving workload.

LMCache interposes a tiered cache hierarchy between GPU VRAM and persistent storage. Evicted KV entries transit first to a 30 GB host DRAM buffer, then spill to the NVMe tier when the DRAM budget is saturated. The disk tier budget per node is configured at 100 GB ; a single NVMe namespace on the Data24. With KV access dominated by 128 KB sequential reads and writes, the workload maps directly onto the bandwidth-oriented path of modern NVMe media, and the 100 GbE fabric (12.5 GB/s theoretical, Data24 150GB/s aggregate) carries that traffic without congestion at the tested concurrency levels.

Data24 NVMe Drives	GPU Inference Nodes	Total H100-class GPUs	Max Concurrent Sessions
1	1	4	16
6	6	24	96
12	12	48	192
24 (full enclosure)	24	96	384

The table contains information related to scale-out from a single Data24 at the benchmark's validated concurrency of 16 simultaneous multi-turn conversations per 4-GPU node. All 24 namespaces are addressable concurrently via multipathed NVMe/RDMA.

Solution

The Data24 presents each of its 24 NVMe SSDs as an independent NVMe namespace over NVMe/RDMA. Each GPU inference node connects to exactly one namespace using the standard Linux nvme-rdma kernel module and mounts it as a conventional ext4 filesystem with noatime($library). LMCache is configured to treat this mount as its disk-tier path, handling all eviction and recall operations transparently. From the inference engine's perspective, the offload target is indistinguishable from a locally attached SSD; the network transport is entirely beneath the filesystem layer.

The network hop introduces an average round-trip latency of 10–50 microseconds for NVMe/RDMA on a well-tuned 100 GbE fabric. For KV cache offloading, this latency sits well inside the tolerance envelope: token generation operates on a timescale of tens of milliseconds per token, and the LMCache DRAM tier absorbs the hot working set so that only cold-path evictions reach the fabric. The net effect is that NVMe-oF KV cache offloading delivers concurrency scaling that would otherwise require doubling or quadrupling the GPU node count.

Benchmarking Architecture

A single shared Ethernet segment connects up to 24 GPU inference nodes to the Data24 via 100 GbE. Each node runs a dedicated vLLM inference process with LMCache enabled, configured with its own 100 GB namespace mount point. The Data24's dual 100 GbE ports provide the uplink; at the tested concurrency levels, aggregate fabric utilization across all attached nodes remains well within link capacity. For environments requiring stricter isolation, per-node VLAN segmentation is straightforward and does not require any changes to the storage configuration.

Layer	Component	Configuration
Inference Engine	vLLM + LMCache	TP=4, GPU MEM UTIL=0.90, max-model-len 32768
GPU VRAM (Tier 0)	4x NVIDIA® H100 80 GB HBM3/HBM3e	~148 GB available for KV pages post-weights
Host DRAM (Tier 1)	LMCache CPU budget	30 GB warm cache per node (TP=4 configuration)
Storage (Tier 2)	NVMe namespace on Data24	100 GB budget, ext4 noatime, NVMe/RDMA transport
Network Fabric	100 GbE Ethernet	MTU 9000, RDMA socket buffers 134 MB, nvme-RDMA kernel module
Storage Enclosure	OpenFlex Data24	24x NVMe SSDs, dual 100 GbE, 2U rackmount

Test Optimizations for Production LLM Inference Simulation

A single shared Ethernet segment connects up to 24 GPU inference nodes to the Data24 via 100 GbE. Each node runs a dedicated vLLM inference process with LMCache enabled, configured with its own 100 GB namespace mount point. The Data24's dual 100 GbE ports provide the uplink; at the tested concurrency levels, aggregate fabric utilization across all attached nodes remains well within link capacity. For environments requiring stricter isolation, per-node VLAN segmentation is straightforward and does not require any changes to the storage configuration.

Test Optimizations for Production LLM Inference Simulation

Multi-Turn Conversational AI

- KV cache grows with each conversation turn; disaggregated storage prevents context truncation at scale.
- Prefix caching deduplicates shared system-prompt KV entries, reducing fabric traffic for common prefixes.
- DRAM tier absorbs the hot working set of recent turns, limiting cold-path I/O to truly evicted context.

Inference Infrastructure Efficiency

- One 2U enclosure replaces per-server NVMe expansion across an entire GPU pod.
- Compute nodes require no local storage for KV cache, simplifying server BOM and rack design.
- NVMe media utilization is pooled across all attached nodes rather than siloed per server.

High-Concurrency Serving

- 16 simultaneous multi-turn conversations per 4-GPU node validated at long context without KV exhaustion.
- Storage disaggregation eliminates preemption events caused by local VRAM pressure.
- Up to 384 concurrent sessions from a single Data24 across a 24-node deployment.

Fleet Operations

- NVMe namespaces can be reassigned across nodes without physical drive migration.
- Capacity scales by adding enclosures; GPU nodes scale independently.
- Standard NVMe/RDMA requires no proprietary drivers, HCAs, or fabric management software.

Performance Characteristics

Storage performance was validated using fio with O_DIRECT and 128 KB block sizes — matching the approximate I/O unit of LMCache disk operations. Four parallel jobs at iodepth=32 characterize the sustained bandwidth path relevant to KV eviction (sequential write) and recall (sequential read). All LLM inference measurements used a 70 Billion Parameter Model at the full 32,768-token context window with temperature 0.7 and 512 completion tokens per turn, across four turns per conversation.

Metric	Configuration	Value
Network latency, NVMe/RDMA	200 GbE, MTU 9000, NVIDIA Switch	10–50 µs avg round-trip
KV I/O block size	LMCache default (128 KB-class chunks)	Sequential, bandwidth-dominated
Validated concurrency	TP=4, 32K context, 4-turn conversations	16 simultaneous sessions per node
Namespaces per enclosure	Data24, 24x independent NVMe SSDs	24 (one per inference node)
Max GPU nodes per Data24	4 GPUs per node, 1 namespace per node	24 nodes / 96 GPUs
Max concurrent sessions	24 nodes × 16 sessions per node	384 simultaneous multi-turn sessions
Disk tier budget per node	LMCache LMCACHE_DISK_BUDGET	100 GB per namespace
DRAM tier per node (TP=4)	LMCache LMCACHE_CPU_BUDGET_TP4	30 GB per node

Actual inference throughput and TTFT deltas (NVMe-oF vs. local NVMe) are presented in the companion technical whitepaper: Western Digital KV Cache Benchmark Whitepaper. Representative results indicate that at concurrency 16, NVMe-oF overhead is attenuated by the pipelined token generation path, with TTFT impact concentrated in the cold-path recall latency rather than mean throughput.

All measurements used the full 32,768-token context window with 4-turn multi-turn conversations. Each turn accumulates prior context, creating growing KV cache pressure across the conversation. At concurrency 4, each run produces 16 turn-level samples; at concurrency 8, it produces 32. Results are aggregated per scenario and concurrency level.

- Tokens Per Second (TPS) measures token throughput – the number of tokens the system produces per second.
- Time to First Token (TTFT) measures the latency between sending a prompt to the model and receiving the first generated token. This can be seen to determine how responsive the model feels to users.
- P95 and P99 show percentile latency. For example, P95 means 95% of requests to complete faster than the value. P95 is a sound metric to look at system stability whilst P99 can be considered worst case behavior.

Scenario	Conc	Avg TPS	TPS Δ	Avg TTFT	TTFT Δ	P95 TTFT	P95 Δ	P99 TTFT	Fail %
Baseline	4	34.39	—	0.687s	—	4.581s	—	6.776s	0%
Local SSD	4	33.94	-1.3%	0.793s	+15.4%	4.567s	-0.3%	6.787s	0%
NVMe-oF	4	33.79	-1.7%	0.770s	+12.1%	4.564s	-0.4%	6.784s	0%
Baseline	8	30.55	—	0.552s	—	1.910s	—	6.196s	0%
Local SSD	8	30.38	-0.6%	0.591s	+7.2%	2.033s	+6.4%	6.380s	0%
NVMe-oF	8	30.00	-1.8%	0.577s	+4.6%	2.020s	+5.8%	6.364s	0%

The table contains information related to benchmark results for a 70 Billion Parameter Model (TP=4, 4x H100) at concurrency 4 and 8. Baseline = pure in-VRAM (no offload). Local SSD and NVMe-oF use the same LMCache configuration with 30 GB DRAM tier and 100 GB disk tier. Delta columns are relative to Baseline at the same concurrency level. P99 TTFT data omitted from column header for width; reported values shown in table.

Observations

Throughput: NVMe-oF is Within Statistical Noise of Local NVMe

At concurrency 4, NVMe-oF throughput of 33.79 tokens/sec represents a 1.7% reduction versus the 34.39 baseline, compared to 1.3% for local SSD. The 0.4-point difference between local and remote storage is not operationally significant. At concurrency 8, the gap narrows further: NVMe-oF at 30.00 tokens/sec carries 1.8% overhead versus local SSD's 0.6%, a spread of 1.2 percentage points in absolute terms. The cause is not fabric latency on the critical path — it is the probabilistic interaction between growing KV pressure, DRAM tier saturation timing, and the slightly higher cold-path recall latency of a fabric hop. Both offloaded scenarios deliver throughput that is functionally equivalent to the baseline for any workload where the difference between 30.55 and 30.00 tokens/sec is below the quality-of-service threshold.

TTFT: NVMe-oF Beats Local SSD at Both Concurrency Levels

The most notable result in the dataset is that NVMe-oF average TTFT is consistently lower than local SSD TTFT. At concurrency 4: 0.770s (NVMe-oF) versus 0.793s (local SSD), a 2.9% advantage for the remote tier. At concurrency 8: 0.577s versus 0.591s, a 2.4% advantage. Both are still elevated versus the baseline, but the direction of the delta is counterintuitive for anyone expecting fabric latency to dominate TTFT. The most likely explanation is I/O scheduling behavior: LMCache's asynchronous eviction path flushes dirty KV chunks to the disk tier in the background, and local NVMe's higher queue depth sensitivity can introduce brief contention during the prefill phase of a new request when background writes are in flight. The RDMA path to the Data24 provides queue depth isolation that insulates the prefill path from eviction traffic.

Tail Latency: P95 and P99 Are Effectively Flat Across Storage Tiers

P95 TTFT for NVMe-oF is marginally lower than local SSD at concurrency 4 (4.564s vs. 4.567s), and essentially tied at concurrency 8 (2.020s vs. 2.033s). P99 TTFT follows the same pattern: NVMe-oF at 6.784s and local SSD at 6.787s at concurrency 4; 6.364s versus 6.380s at concurrency 8. The implication is that the tail of the TTFT distribution is not driven by storage tier at all — it is driven by GPU scheduling, vLLM's PagedAttention eviction logic, and multi-turn context accumulation. The RDMA fabric hop to the Data24 does not introduce measurable tail latency amplification at the tested concurrency levels.

Reliability: Zero Failures Across All Configurations

Every request completed successfully across all six measurement configurations. Zero failures at concurrency 8 with NVMe-oF — the configuration most likely to surface instability from fabric reconnects, namespace contention, or LMCache eviction race conditions — validates that the disaggregated storage path is stable under load. The automated test framework enforced fresh vLLM processes and cleared cache directories for each scenario, ruling out state contamination as a reliability factor.

Conclusion

The benchmark data resolves the central question for KV cache disaggregation in production LLM serving: does the network hop to shared NVMe storage materially degrade inference performance? For the OpenFlex Data24 over RDMA at the tested concurrency levels, the answer is no. NVMe-oF throughput is within 1.8% of the in-VRAM baseline, NVMe-oF TTFT is measurably lower than local SSD TTFT, P99 tail latency is statistically indistinguishable between local and remote storage tiers, and the failure rate is zero. The RDMA fabric does not appear on the critical path of token generation at these workload parameters.

The operational consequence is straightforward: a single OpenFlex Data24 in 2U can serve as the dedicated KV cache offload tier for an entire 24-node, 96-GPU H100 pod, delivering up to 384 simultaneous multi-turn sessions at the validated concurrency, with no per-server NVMe expansion, no specialized fabric, and no proprietary software. The 4:1 GPU-to-drive ratio is an implementation artifact of the LMCache tiered architecture and the bandwidth characteristics of Gen4 SSD Interface, and 100 GbE RDMA, not a theoretical ceiling. Higher concurrency testing is ongoing and results will be published in the companion whitepaper.

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