Phison Pascari X200P 7.68TB SSD Test Results - LLM KV Cache Offload Tier

These results characterize a Phison Pascari X200 7.68 TB SSD acting as the external key-value (KV) cache tier for large language model inference across three storage paths. Configurations progress in complexity in the following order: a no-offload recompute baseline; three memory-constrained tiers with the KV cache on a locally attached X200; the same three tiers with the KV cache on a second, physically distinct X200 presented over NVMe-oF™ from an OpenFlex® Data24 4240; and those tiers again after that drive was moved into an OpenFlex Data24 4140. Host memory is capped per tier so the OS page cache cannot retain the KV set, which forces retrievals to be served by the storage device rather than by DRAM. Device service was verified per cell with iostat and cross-checked against whole-run /proc/diskstats deltas, and the workload is replayed byte-identically under a single stream hash so all four groups process the same token stream.

With every KV retrieval served from NAND, the worst-case local tier returns 7.47 tokens/s end to end against 6.19 tokens/s for no-offload recompute, a gain of 20.7 percent, while mean time to first token falls 19.7 percent. Moving the identical tier onto the Data24 4240 costs 6.0 to 7.4 percent against the protocol-identical local tier; moving it to the Data24 4140 costs only 1.1 to 3.6 percent and holds 18.6 to 21.3 percent above recompute. The two enclosures differ in the PCIe link they present to the drive, Gen4 x2 on the 4240 against Gen4 x4 on the 4140, and the end-to-end penalty tracks that link rather than the use of NVMe-oF. The workload is latency and hit-rate dominated rather than bandwidth dominated: a 72 percent reduction in raw path read bandwidth on the 4240 produced only a 7.4 percent end-to-end cost on the realistic tier.

Device and Platform Under Test

Test Description	Value
Device under test	Phison Pascari X200, 7.68 TB, PCIe® Gen5 NVMe™
Part number	XP208H087T68P328T1910
Controller / firmware	PS5302-X2 / X2WM6024
LBA format / filesystem	512 B factory setting; ext4 with 4 KiB blocks (ext4 retained rather than XFS for comparability with prior runs)
Server	Supermicro, 2 sockets, 64 cores, 1 TiB DDR5
GPUs	4x NVIDIA® H100
GPU driver / CUDA	610.43.02 open kernel modules / CUDA 13.3
OS / kernel	Rocky Linux® 10.1, kernel 6.12.0-124.20.1.el10_1
Inference stack	vLLM 0.22.0, LMCache 0.4.6 (LMCacheConnectorV1, kv_both), torch 2.11.0+cu130, transformers 5.12.1
Model	70B parameter instruction-tuned transformer, BF16 safetensors (131.4 GiB), TP=4, max_model_len 32768, gpu_memory_utilization 0.55
Tuning	accelerator-performance profile, GPU clocks locked (1620 MHz PCIe, 1635 MHz NVL), THP never, memlock unlimited
Fabric transport	NVMe-oF over RoCE v2 (RDMA), verified per cell from sysfs; Mellanox® mlx5 100 GbE, MTU 4500
Fabric target A	OpenFlex Data24 4240, FW 3.0.0, dual IOM; single path to IOMA-PORT1. Presents PCIe Gen4 x2 to the drive.
Fabric target B	OpenFlex Data24 4140, FW 3.0.0, dual IOM; single path to IOMA-PORT1. Presents PCIe Gen4 x4 to the drive.

Note: Device details, host platform, and fabric configuration. Both samples are the same model and firmware; the fabric sample is a second physical drive.

Phison Pascari X200P 7.68TB SSD Test Results - LLM KV Cache Offload Tier

Tier	Label	LMCache DRAM tier	MemoryHigh	Intent
a3 / b3	Realistic	30 GiB total (7.5 GiB x 4 TP workers)	56 G	Hot set in DRAM, spill served from NAND
a4 / b4	Happy medium	None	28 G	About a quarter cacheable, majority NAND served
a5 / b5	Worst case	None	20 G	About 6 G over the anonymous floor; every retrieval is a cold NAND read

Method

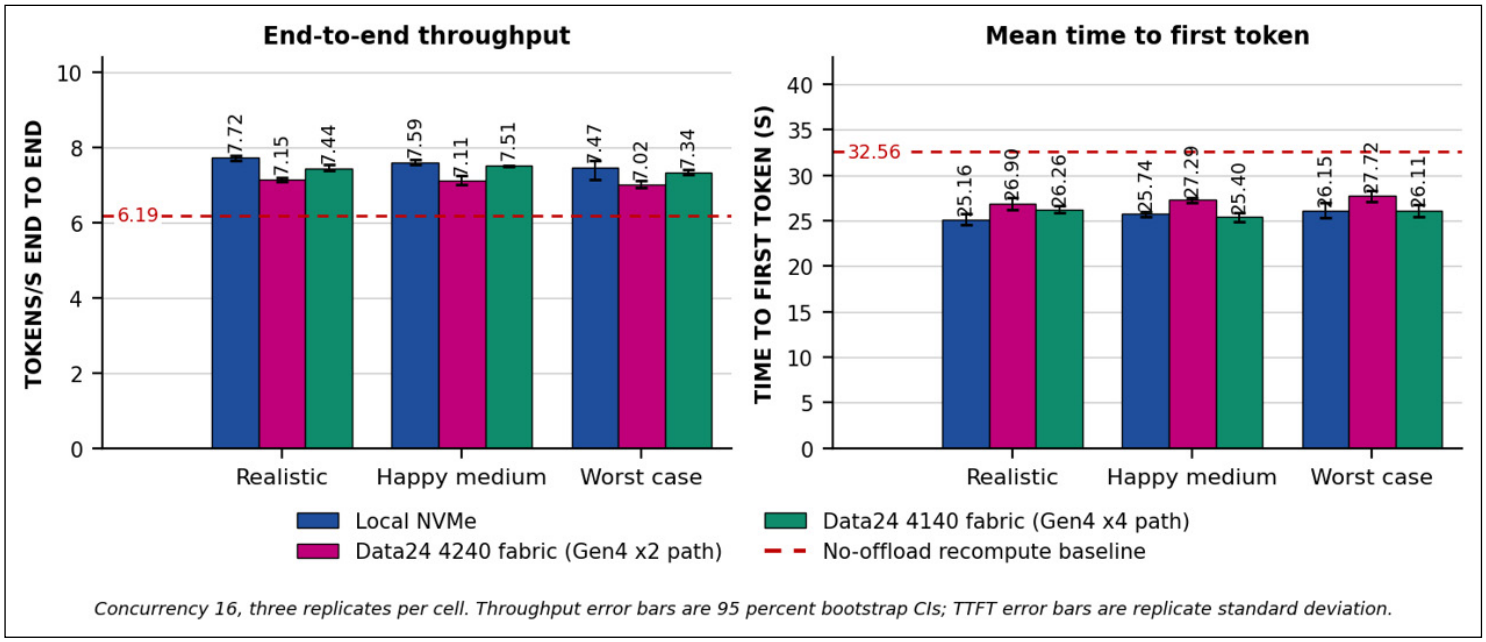
- Workload: 16 concurrent multi-turn conversations of 4 turns each, about 10,000 prompt tokens per conversation, 512 maximum completion tokens per turn, greedy decoding.
- Cell protocol: every (tier, replicate) cell wipes the KV cache directory, starts a fresh vLLM server, asserts cold state, runs a disjoint-prompt warmup, resets the GPU prefix cache, then measures 64 turns. Three replicates per tier.
- Statistics: replicate-level mean and 95 percent bootstrap confidence intervals from 10,000 resamples. Tail percentiles are pooled across replicates, n = 192 turns. Hit rate is the per-run delta of the vLLM external prefix cache counters, token weighted, about 565,000 queried tokens per cell.
- Memory constraint: the process group runs in a systemd user cgroup scope. A hard MemoryMax ceiling is applied at launch to survive the weight-broadcast transient, reclaimable startup pages are drained through cgroup v2 memory.reclaim, then MemoryHigh is tightened to the tier target. Swap is disabled. KV set size at concurrency 16 is about 51 GiB against a process anonymous floor of 10 to 14 GiB.
- BF16 KV cache and gpu_memory_utilization of 0.55 are deliberate: BF16 is the worst-case I/O volume, and the reduced GPU allocation keeps the on-GPU KV cache small so lookups reach the external tier. GPU prefix hit rate is 0.9 percent in every tier, confirming the external tier served the reuse.
- Only concurrency 16 is reported. At concurrency 4 and 8 the on-GPU KV cache absorbs the workload (GPU hit 74 to 75 percent) and external hits are 0.0 percent, so those cells do not exercise storage.

Device and Platform Under Test

Test Description	Recompute	Local Pascari X200			Fabric, Data24 4240 (Gen4 x2 path)			Fabric, Data24 4140 (Gen4 x4 path)		
	No offload	a3	a4	a5	b3	b4	b5	b3	b4	b4
Tokens/s end to end	6.19	7.72	7.59	7.47	7.15	7.11	7.02	7.44	7.51	7.34
95 percent CI	6.05-6.29	7.66-7.79	7.53-7.67	7.14-7.65	7.09-7.20	7.02-7.24	6.94-7.11	7.38-7.55	7.50-7.51	7.27-7.41
Tokens/s decode	10.34	12.74	12.55	12.23	11.83	11.70	11.61	12.33	12.28	12.07
TTFT mean (s)	32.56	25.16	25.74	26.15	26.90	27.29	27.72	26.26	25.40	26.11
TTFT P95 (s)	60.31	63.97	62.31	62.67	64.32	61.28	63.07	n/r	n/r	n/r
TTFT P99 (s)	65.12	69.96	67.88	66.81	69.57	67.46	67.35	n/r	n/r	n/r
External KV hit rate (%)	n/a	71.78	70.80	69.33	71.78	71.08	67.81	71.78	71.11	70.36
Device KV read (GiB)	n/a	110.2	123.3	120.2	111.1	123.4	118.2	112.7	123.8	122.4
Device eviction write (GiB)	n/a	51.5	51.7	52.8	51.5	51.4	53.9	51-52	51-52	51-52
Sustained read (GiB/s)	n/a	0.92	0.97	0.94	0.75	0.87	0.82	n/r	n/r	n/r
Peak read (GiB/s)	n/a	2.94	3.14	2.73	2.02	2.51	2.16	3.05	2.89	3.39

Note: Top line performance by configuration. All cells are concurrency 16 with three replicates. Device I/O rows are per-cell means from iostat sampled at 5 s. n/r means not reported in that run.

Phison Pascari X200P 7.68TB SSD Test Results - LLM KV Cache Offload Tier



Note: End-to-end throughput and mean time to first token, by tier and storage path.

Normalized Results, Fabric Cost and Variability

Throughput multipliers above 1.00X and TTFT multipliers below 1.00X are improvements. The middle band gives the cost of each fabric tier against its protocol-identical local counterpart, which is the comparison that isolates the storage path. Coefficients of variation are reported because they compare directly across metrics with different means, which standard deviations do not. These are replicate-level values, so they describe reproducibility of the full serving stack rather than sample-interval jitter of the device.

Test Description	Recompute	Local Pascari X200			Fabric, Data24 4240 (Gen4 x2 path)			Fabric, Data24 4140 (Gen4 x4 path)		
	No offload	a3	a4	a5	b3	b4	b5	b3	b4	b4
Normalized to the no-offload baseline (1.00X)										
Tokens/s end to end	1.00X	1.25X	1.23X	1.21X	1.16X	1.15X	1.13X	1.20X	1.21X	1.19X
Tokens/s decode	1.00X	1.23X	1.21X	1.18X	1.14X	1.13X	1.12X	1.19X	1.19X	1.17X
TTFT mean	1.00X	0.77X	0.79X	0.80X	0.83X	0.84X	0.85X	0.81X	0.78X	0.80X
Against the protocol-identical local tier										
Tokens/s end to end	n/a	ref	ref	ref	-7.4%	-6.3%	-6.0%	-3.6%	-1.1%	-1.7%
Coefficient of variation, replicate level (CoV = SD / mean, n = 3)										
Tokens/s end to end	2.100%	0.777%	0.922%	3.882%	0.839%	1.547%	1.282%	1.344%	0.000%	0.954%
Tokens/s decode	0.580%	0.549%	0.558%	3.189%	0.592%	1.111%	0.775%	0.487%	1.221%	0.497%
TTFT mean	1.259%	2.544%	0.971%	3.021%	2.491%	0.989%	2.237%	1.447%	1.890%	2.451%
External KV hit rate	n/a	0.000%	0.791%	3.563%	0.000%	0.999%	2.006%	0.000%	0.844%	0.370%

Note: Normalized performance, fabric cost against local, and coefficient of variation. Hit rate is not normalized because the baseline configures no KV tier.

Phison Pascari X200P 7.68TB SSD Test Results - LLM KV Cache Offload Tier

Device Baseline

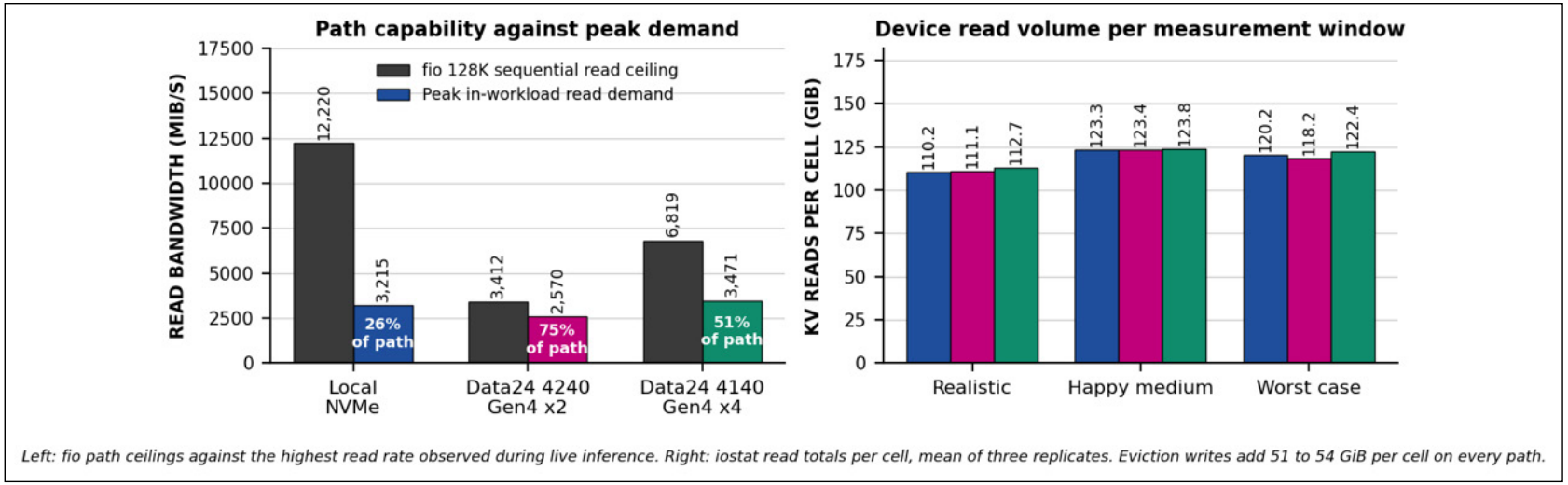
Pattern	Local	Data24 4240	Data24 4140
128K sequential read	12,218	3,412	6,819
1M sequential read	12,235	3,412	6,820
128K sequential write	8,864	3,165	6,244
1M random read	6,911	3,412	6,820
128K mixed read + write	5,127 + 5,128	3,017 + 3,018	5,260 + 5,260

Note: Bandwidth in MiB/s across all three paths, from a separate later sweep taken after the enclosure swap, in which all three were measured together. Its local column differs from Table 5 at 1M; see notes.

- These are fresh-drive, single-pass figures measured in the SLC-cache regime, not steady-state preconditioned results.
- Fabric bandwidth is flat across every block size on both enclosures, marking a path ceiling rather than a device characteristic. The midplane wires PCIe Gen4 x2 per IOM on the 4240 and x4 on the 4140, and the ceilings of roughly 3,412 and 6,819 MiB/s follow that difference almost exactly. Neither limits the drive (12,219.8 MiB/s local) or the 100 GbE NIC. Fabric latency is therefore path latency, not drive latency.

Device Activity During Inference

Each cell moves roughly 160 to 176 GiB of combined KV I/O through the drive in an approximately five-minute window while serving 16 concurrent conversations, and that volume is effectively identical on all three paths because the replayed token stream is identical. The read stream and the eviction write stream are concurrent, not sequential phases.



Left: fio path ceilings against the highest read rate observed during live inference. Right: iostat read totals per cell, mean of three replicates. Eviction writes add 51 to 54 GiB per cell on every path.

Note: Path capability against observed peak demand, and measured device read volume per cell.

The share of each path consumed by peak demand separates the three configurations, and the end-to-end penalties order the same way: local peak read reaches 26 percent of its ceiling, the 4140 reaches 51 percent, and the 4240 reaches 75 percent. Sizing follows from the same measurement: the 16-conversation working set is about 51 GiB, so a 7.68 TB device holds roughly 140 such working sets.

Notes and Limitations

- Enclosure provenance: the 4240 and 4140 figures come from separate campaigns, the 4140 measured after the drive was moved between enclosures with data intact. Local reference cells are the re-measure under the current two-phase cap protocol, so every local-against-fabric comparison is protocol-identical tier for tier; superseded one-phase cells are archived and appear nowhere here.
- The unconstrained a1 and a2 cells (8.32 and 7.94 tokens/s) are omitted from the tables. Device telemetry shows zero reads there: the 1 TiB host page cache served them, since LMCache local_disk uses buffered I/O. They establish the DRAM-rich ceiling, and worst-case NAND serving at 7.47 tokens/s sits about 10 percent below it while remaining 20.7 percent above recompute.
- Attribution is by device telemetry rather than per-chunk cache accounting; the LMCache 0.4.6 audit-replicate parse is a known gap, so no DRAM-versus-disk per-chunk numbers are claimed.
- Cells marked n/r were not reported in the 4140 run: tail percentiles, sustained read rate, and per-tier eviction write totals. Its device figures are per-tier means only, without the per-replicate detail available for the other paths.
- On the 4140 the happy-medium tier (7.51) matches or exceeds the realistic tier (7.44), and its replicate standard deviation is 0.00 across three cells. Both are stated as measured: with a fast enough path the 30 GiB DRAM tier stops separating the tiers at this concurrency. Do not generalize beyond concurrency 16.
- External hit rate matches across all three paths at the realistic setting (71.78 percent, sd 0.00), because deterministic replay and fixed hashing make the tier lookup stream reproducible. The worst-case tier recovered on the faster path: 70.36 percent, sd 0.26 on the 4140 against 67.81 percent, sd 1.36 on the 4240. That is consistent with path latency interacting with the 3000 ms LMCache lookup timeout, but the mechanism is untested and is not asserted here.
- Dual-path operation was not measured and is not possible with this drive. Both IOMs answer discovery over RDMA and the enclosure is cabled and provisioned for sharing, but the installed X200 reports single-port CMIC, verified against the identical directly attached twin, so only one IOM can own it at a time. Enabling sharing produced ownership migration rather than concurrent access; dual-path measurement awaits a dual-port drive.
- Scope: one drive model, one host, one inference model, one concurrency point. No multi-drive scaling claims are supported.

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