

MiPhi AI Inference Platform: KV-Cache Offload over OpenFlex® Data24 4000

Challenges

- Throughput and latency on large models are limited by the number of concurrent users, not by the size of the model weights. Each concurrent user maintains a dedicated KV cache that grows with every round of the conversation, so the KV working set of an LLM or VLM deployment expands continuously over the life of a session and across a growing user population, while GPU memory capacity remains fixed.
- Once the working set exceeds GPU memory, previously computed KV cache must be evicted to free capacity - whether to admit a new user or to serve the next round of a user already in conversation. In a GPU-only serving stack, no additional tier exists to retain the evicted cache.
- Each eviction imposes a recomputation cost. The evicted prefix must be re-prefilled in full on the next round, increasing time to first token for concurrent users and reducing decode throughput as recomputation consumes GPU cycles otherwise available for generation. As the number of users served grows, eviction and recomputation increase proportionally, and both metrics degrade further.
- Adding GPUs purely to gain KV-cache capacity is expensive and leaves the added compute under-utilised.

Highlights

- The MiPhi AI Inference Platform manages the KV cache as a tiered resource spanning GPU memory, system RAM and shared NVMe on an OpenFlex Data24 platform, reloading evicted prefixes rather than recomputing them.
- Up to 12x lower median time to first token once the GPU begins evicting KV cache, and the advantage holds as concurrency grows.
- Up to 1.9x higher decode throughput under load: throughput stays essentially flat as concurrency rises, while the GPU-only baseline declines.
- KV-cache capacity scales on shared storage rather than on GPUs: as the user count grew 7x, the KV working set held on the OpenFlex Data24 tier more than doubled, with no additional GPUs.

Solution Details

Key Solution Components of the MiPhi AI Inference Platform

- Metadata servers, chunk servers, and client architecture
- High availability via replication and erasure coding
- Parallel I/O for high throughput
- Tiered KV cache spanning GPU memory, system RAM and OpenFlex Data24 capacity
- Prefix retention and reload on eviction in place of prefix recompute
- KV-cache-aware routing across multiple inference servers, preserving prefix reuse when requests are spread over a cluster

Key Solution Components of the OpenFlex Data24 4000 Platform

- 24-bay NVMe SSD enclosure
- 100GbE connectivity with NVMe-over-Fabrics
- Shared storage across multiple hosts
- High-Speed Fabric Network with RoCE or TCP connectivity for flexible deployment
- Composable Infrastructure Layer with Dynamic storage allocation across workloads

OpenFlex Data24 4000 Series Storage Platforms

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. WD RapidFlex™ NVMe-oF controllers allow up to six dual pathed hosts to be attached without a switch. The OpenFlex Data24 4200 series uses WD's RapidFlex C2000 Fabric Bridge Adapters to provide up to 12 ports of 100GbE which can connect to RDMA and/or TCP configured host ports.

Features

- NVMe-over-Fabrics (NVMe-oF) architecture
- Bandwidth match between PCIe Gen4 SSDs (Storage) and I/O (Network) - no oversubscription
- RESTful API support for simplified management
- N+2 fan redundancy eliminates the need for hot-swappable fan modules (Data24 4100)
- Connect directly to an Ethernet NIC in your host or through an Ethernet switch
- Device Sharing optional configuration utilizes NVMe Non-Transparent Bridging to allow all Ethernet ports to access any SSD
- High IOPS and throughput performance
- Software-defined composability
- Supports Ethernet-based disaggregated storage
- High availability with redundant components
- Ideal for AI/ML, real-time analytics, and transactional workloads

MiPhi AI Inference Platform: Tiered KV Cache

The MiPhi AI Inference Platform treats the KV cache as a tiered resource rather than a GPU-resident structure. A conversation's key-value state is distributed across GPU memory, system RAM and NVMe capacity accessed over NVMe-over-Fabrics, with prefixes migrating between these tiers as concurrency changes. When a prefix is evicted from the GPU cache, the MiPhi AI Inference Platform reloads it from RAM or NVMe rather than reprocessing it through the model, so the prefill cost of a conversation is incurred once rather than once per round. The NVMe tier consists of a Western Digital OpenFlex Data24 platform holding four PASCARI A100 2 TB PCIe Gen4 drives in RAID 0, shared over the fabric by both inference servers.

Key Capabilities

- Tiered KV-cache management across GPU memory, system RAM and OpenFlex Data24 storage
- Prefix reuse across conversation rounds, removing repeated prefill of the shared system prompt and the per-user history block
- KV working set scales with concurrency, from 70 GB at 10 users to 143 GB at 70 users, independently of GPU count
- Employs shared, disaggregated NVMe so that multiple inference hosts draw from a single capacity pool rather than captive local drives

Reference Architecture

The evaluation architecture comprises two compute servers, one storage server and one OpenFlex Data24 platform. Each compute server is equipped with a single NVIDIA RTX A6000 (48 GB) and runs a vLLM instance, with a KV-cache-aware router positioned in front of the pair to direct each request to the server that already holds the KV prefix for that conversation. A third server hosts the Data24 platform and presents the RAID 0 volume built from four PASCARI A100 2 TB PCIe Gen4 NVMe drives; both compute servers mount this volume over the fabric via NFS and offload their KV cache to it. The two configurations were measured sequentially on the same cluster - first with GPU-only vLLM prefix caching on both servers, then with the MiPhi AI Inference Platform's tiered KV cache enabled on both servers - so that hardware, router and workload remain identical across the comparison and only the location of the KV cache changes.

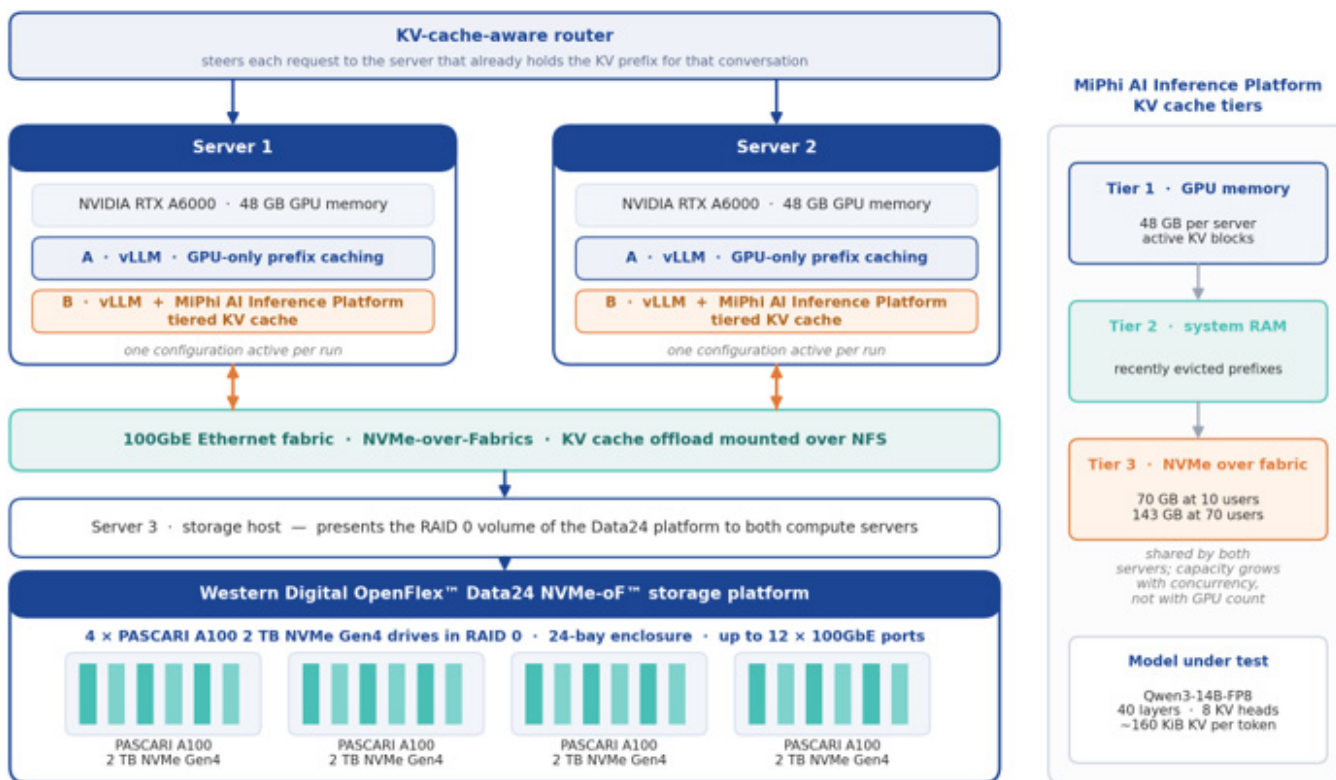
Test Configuration

Component	As Tested
Compute servers	2 servers, each with 1x NVIDIA RTX A6000 (48 GB GPU memory), each running a vLLM instance
Request routing	KV-cache-aware router in front of both compute servers, steering each request to the server that already holds its KV prefix
Storage media	4x PASCARI A100 2 TB ¹ NVMe PCIe Gen4 drives, configured in RAID 0
Storage platform	Western Digital OpenFlex Data24 platform; the four PASCARI drives reside in the Data24 enclosure and are mounted through a third server
Fabric / access	Both compute servers access the RAID 0 volume attached to the storage server over NFS, offloading their KV cache to the Data24 platform
Model	Qwen3-14B-FP8 - 40 layers, 8 KV heads, ~160 KiB of KV cache per token
Configuration A (baseline)	GPU-only vLLM with prefix caching, on both compute servers
Configuration B (evaluated)	MiPhi AI Inference Platform tiered KV cache across GPU, system RAM and NVMe over NVMe-oF, on both compute servers
Test method	The two configurations were run sequentially on the same two servers; hardware, router and workload are identical across both runs, so only the location of the KV cache differs
vLLM settings	gpu-memory-utilization 0.9, max-num-seqs 12
Benchmark	MiPhi AI Inference Platform benchmarks/multi_round_qa; 600 s run per concurrency point; concurrency swept from 10 to 70 strict concurrent users; the same offered load in both runs
Workload per user	Five-round conversation: ~1,000-token shared system prompt + ~8,000-token per-user history + ~20-token questions, answers capped at 100 tokens. ~46,400 prompt tokens and ~500 generated tokens per user; the first ~9k tokens form the reusable prefix
Offload measurement	Peak NVMe KV footprint reported by the SSD monitor for each run

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.

Architecture Diagram

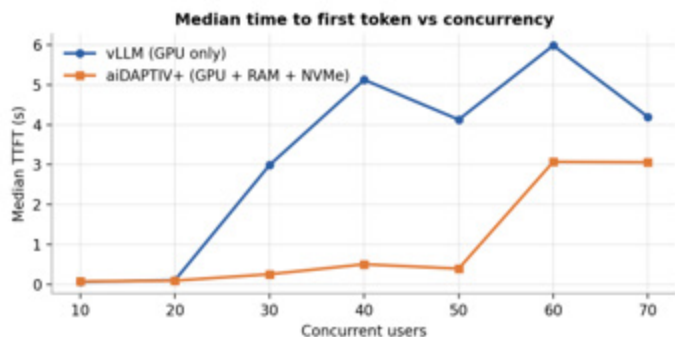
MiPhi AI Inference Platform benchmarks/multi_round_qa — 5-round QA, ~46k prompt tokens per user, 10 → 70 strict concurrent users, 600 s per point



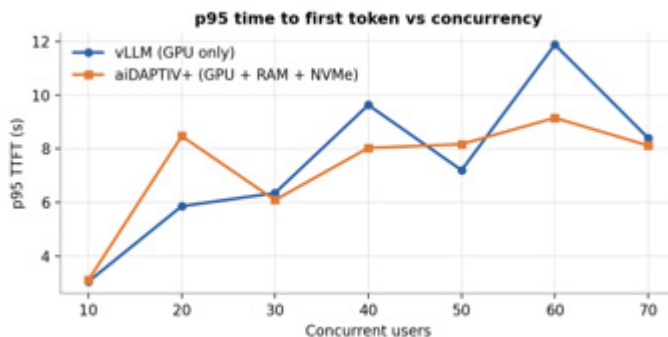
Note: Both compute servers are configured identically and operate behind a KV-cache-aware router; the two configurations were measured sequentially on the same cluster, so only the location of the KV cache differs between runs. The MiPhi AI Inference Platform extends the cache from GPU memory into system RAM and shared NVMe on the OpenFlex Data24 platform over the 100GbE fabric.

Performance Details

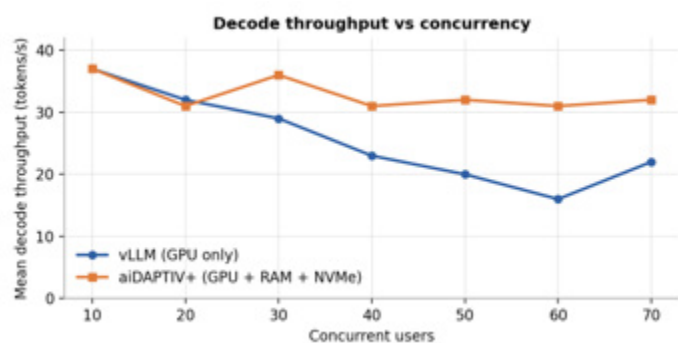
Concurrency was swept from 10 to 70 strict concurrent users at 600 s per measurement point, with each configuration evaluated in a separate run on the same two servers. At 10-20 users the two configurations are indistinguishable, as the KV working set still fits within GPU memory. From 30 users onward the GPU cache overflows and the two configurations diverge: GPU-only vLLM evicts and recomputes conversation prefixes, while the MiPhi AI Inference Platform reloads them from RAM and NVMe.



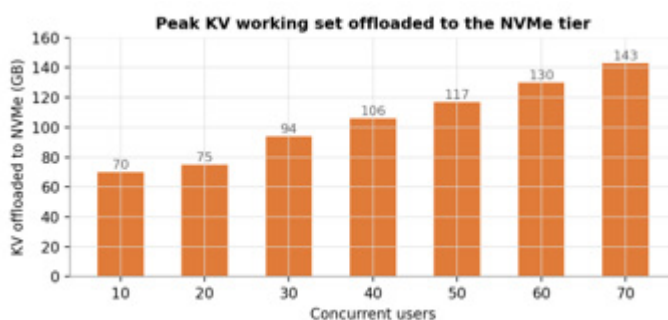
Note: Median TTFT vs concurrency. The two configurations diverge at 30 users, where the GPU cache first overflows.



Note: p95 TTFT vs concurrency. Tail latency converges at peak load while the medians stay separated.



Note: Mean decode throughput vs concurrency. The MiPhi AI Inference Platform holds 31-36 tok/s across the sweep; GPU-only vLLM declines to 16-22 tok/s as recomputation consumes GPU cycles.



Note: Peak KV working set offloaded to the NVMe tier, growing from 70 GB at 10 users to 143 GB at 70 users.

At QPS 0.3, the MiPhi AI Inference Platform maintains a consistent median-TTFT advantage across the full sweep. Up to 50 users, its median TTFT remains low (0.25-0.50 s) while GPU-only vLLM has already risen to 3-5 s. The sweep was deliberately extended beyond 50 users to increase memory pressure; in this stress region the MiPhi AI Inference Platform's median rises to approximately 3 s (3.07 s at 60 users, 3.06 s at 70 users) and both configurations exhibit greater fluctuation in TTFT and throughput, though the MiPhi AI Inference Platform remains at or below vLLM (5.99 s at 60 users, 4.20 s at 70 users). The decode-throughput gap persists: the MiPhi AI Inference Platform sustains 31-32 tok/s while vLLM declines to 16-22 tok/s.

Per-Load Observations

Load - KV on NVMe ²	Observation
10 users 70 GB on NVMe	Both configurations perform identically. Median TTFT is 0.06 s (vLLM) vs 0.07 s (MiPhi AI Inference Platform), p95 is approximately 3.1 s for both, and decode throughput is tied at 37 tok/s. The working set of approximately 74k tokens fits within the GPU cache, so no eviction occurs. The 70 GB recorded on the drive represents the baseline KV footprint.
20 users 75 GB on NVMe	The configurations remain effectively tied. Medians fall below 0.1 s after approximately 100 s. The MiPhi AI Inference Platform shows a heavier tail (p95 8.47 s vs 5.86 s), attributable to the one-time cost of warming the RAM/NVMe tiers. Decode throughput remains level at 32 vs 31 tok/s; only an additional 5 GB is offloaded to the drive.
30 users 94 GB on NVMe	This is the first load at which the GPU cache overflows. The vLLM median settles near 3 s with periodic spikes as prefixes are evicted and recomputed; the MiPhi AI Inference Platform's median drops to 0.25 s once its cache warms (approximately 140 s into the run). Decode throughput moves ahead at 36 vs 29 tok/s, as GPU cycles are no longer consumed by recomputation.
40 users 106 GB on NVMe	The gap is now clear and sustained. vLLM holds at 4-5 s for the entire run while the MiPhi AI Inference Platform drops to approximately 0.5 s. By round, vLLM rises from 5.1 to 6.1 s as recomputation compounds, while the MiPhi AI Inference Platform falls to 0.2-0.3 s through prefix reuse. The decode gap widens to 31 vs 23 tok/s.
50 users 117 GB on NVMe	The MiPhi AI Inference Platform remains sub-second (median 0.39 s) while vLLM stays elevated at 4.13 s and drifts upward late in the run. The decode gap reaches its widest point: 32 vs 20 tok/s, an approximately 1.6x throughput advantage.
60 users 130 GB on NVMe	Both configurations come under memory pressure. The MiPhi AI Inference Platform's median rises to 3.07 s, while vLLM reaches 5.99 s with p95 of 11.9 s vs 9.2 s. The decode difference is pronounced: 31 vs 16 tok/s. The vLLM generation rate has nearly halved under recomputation load, while the MiPhi AI Inference Platform holds steady.
70 users 143 GB on NVMe	At peak load, median TTFT is 3.06 s (MiPhi AI Inference Platform) vs 4.20 s (vLLM), with p95 essentially tied (8.1-8.4 s). By round, vLLM rises from 4.2 to 4.6 s while the MiPhi AI Inference Platform falls from 3.7 s to 0.2-0.3 s. Decode throughput holds at 32 vs 22 tok/s. Sustaining this load required the largest offload of the sweep: 143 GB of KV cache on the drive.

Note: Summary of each per-load comparison panel (TTFT over time, median TTFT by round, decode throughput and TTFT cumulative distribution).

² Actual results depend on CPU memory bandwidth, RNIC count/link speed, PCIe layout, and fabric configuration.

Advantages

- Up to ~12x lower median time to first token once the GPU cache overflows, keeping interactive sessions responsive as the user count grows.
- Up to 1.9x higher decode throughput under load, as GPU cycles are directed to generation rather than repeated prefill.
- Concurrency headroom on existing hardware: two 48 GB RTX A6000 servers served 70 concurrent five-round conversations by offloading 143 GB of KV cache to shared NVMe.
- KV-cache capacity scales independently of GPU count, so expanding the working set no longer requires provisioning compute that remains under-utilised.
- Shared, disaggregated NVMe on the OpenFlex Data24 platform serves multiple inference hosts from a single pool, with 100GbE NVMe-oF maintaining access latency close to that of locally attached NVMe.

Conclusions

On the same two servers, behind the same router and under the same workload, the only difference between the two configurations is the location of the KV cache. GPU-only vLLM is bounded by the 48 GB of GPU memory in each server: beyond 30 concurrent users, an increasing share of every round is spent recomputing prefixes that have already been computed, and both latency and throughput degrade. The MiPhi AI Inference Platform retains those prefixes in system RAM and on NVMe accessed over the fabric, so prefill work is performed once and reloaded thereafter. The result is 0.25-0.50 s median TTFT through 50 concurrent users and 31-36 tok/s decode throughput across the full sweep, sustained by a KV working set that grew to 143 GB on four PASCARI A100 drives in an OpenFlex Data24 platform. For interactive multi-user inference, adding shared NVMe capacity behind the MiPhi AI Inference Platform shifts the limit from GPU memory to storage capacity, and shared NVMe-oF capacity is both cheaper per gigabyte and reusable across inference hosts.

