Achieve Higher Sustained LLM Throughput with SupremeRAID™ KV Cache for Rack

How a Shareable External KV Cache Tier Improves Total Throughput by 53%

JUNE 2026

Table of Contents

Executive Summary **3**
Key Findings 4
Audience and Scope 4
Why KV Cache Belongs on a Shareable Storage Tier **5**
About SupremeRAID™ KV Cache for Rack **6**
Certified Server Platforms 7
Tested Configuration - SupremeRAID™ KV Cache for Rack 8
Validation Methodology **10**
Application Throughput Scaling **10**
Total Throughput Across Prefix Lengths 12
What The Results Mean for Deployment **13**
Conclusion **14**
Appendix A - References **15**
Appendix B - Certified Server Platforms **15**

Executive Summary

Production LLM inference is moving toward longer context windows, higher concurrency, and more interactive multi-turn workloads, which drives KV cache demand higher and makes cache placement an infrastructure design decision rather than a model-serving configuration detail. GPU servers are optimized for accelerator density, power, cooling, and high-speed networking, not for serving as the long-term cache expansion tier. When KV cache capacity is tied to local drives in each GPU server, platform teams must scale cache one compute node at a time, which limits deployment flexibility, complicates hardware standardization, and makes it harder to run different GPU server profiles in the same environment.

The SupremeRAID™ KV Cache for Rack externalizes that layer. It combines SupremeRAID™ with a certified storage server, builds a high-performance NVMe RAID array for service continuity, and exports the cache filesystem to GPU servers over NFS. GPU servers continue to run the inference stack, while the SupremeRAID™ KV Cache for Rack provides a shareable external cache tier that scales independently from compute. This paper validates that design with an application-level EvalScope benchmark on a Qwen3-235B + vLLM + LMCache serving stack across prefix lengths from 512 to 8,192 tokens, comparing no KV cache offload against a cache path hosted on the SupremeRAID™ KV Cache for Rack.

The SupremeRAID™ KV Cache for Rack improved total throughput at every prefix length tested, and the gains grew with context: from 32.7% at 512 tokens to 53.4% at 8,192 tokens, with zero failed requests at any test point.

For AI infrastructure teams running long-context inference under GPU memory pressure, the SupremeRAID™ KV Cache for Rack scales compute and cache on independent axes, without the throughput penalty that appears when the serving stack has no external cache tier to fall back on.

Key Findings

- **Sustained throughput uplift at every tested prefix length.** Total Throughput improved 32.7% at 512 tokens, 37.7% at 1,024, 47.0% at 2,048, 51.6% at 4,096, and 53.4% at 8,192; a clear, linear gain.
- **The benefit grows as context grows.** The longer the reusable prefix, the more valuable cache reuse becomes, exactly the workload shape long-context inference is moving toward.
- **Reliable under load.** All 512 requests completed successfully at every prefix length in both scenarios, including 128-way parallelism and 3-to-5-turn conversations.
- **Production-grade stack.** Qwen3-235B-A22B-Instruct-2507 on 4x NVIDIA H200 GPUs with vLLM and LMCache, NFS-mounted to a 10-drive RAID5 volume on a Supermicro SSG-221E-DN2R24R with dual 200 Gb/s Ethernet, representative of the platforms customers are deploying today.
- **Architectural value beyond the numbers.** GPU servers stay focused on inference while cache capacity scales independently on a dedicated, shareable storage tier — the deployment flexibility teams need as serving fleets and context windows both grow.
- **Cache placement is now a platform decision.** As context, concurrency, and multi-turn interactivity all grow together, KV cache capacity is an infrastructure design decision rather than a model-serving configuration detail and centralizing it on a shareable tier preserves room for heterogeneous GPU server profiles in the same fleet.

Audience and Scope

This paper is written for AI infrastructure architects, platform engineering teams, storage architects, solution engineers, and technical marketing reviewers evaluating inference infrastructure under GPU memory pressure.

The comparison focuses on a practical deployment decision: whether a shareable external NVMe cache tier helps the serving stack sustain Total Throughput as prefix length grows and the reusable KV working set outgrows GPU memory.

The paper does not attempt to rank every possible KV cache backend, model, or workload mix. It validates one practical platform question with one production-grade configuration, and the results should be read in that scope.

Where this paper sits in the SupremeRAID™ KV Cache portfolio:

SupremeRAID™ KV Cache for Server provides single-node NVMe cache acceleration on individual GPU servers and edge AI deployments.

SupremeRAID™ KV Cache for Rack, the subject of this paper, is a rack-scale, partner-validated solution that externalizes the cache tier on a dedicated storage server shared across multiple GPU hosts.

SupremeRAID™ KV Cache Platform is on the roadmap, purpose-built for NVIDIA's STX reference architecture with native BlueField-4 DPU execution.

Why KV Cache Belongs on a Shareable Storage Tier

KV cache offload changes the shape of inference infrastructure. Instead of forcing every GPU server to carry the full cache expansion footprint locally, a shared external tier lets teams size compute and cache storage separately.

This separation matters in real deployments:

- GPU servers can be selected for accelerator density and serving performance, without requiring every node to carry the same local SSD footprint.
- Cache capacity can be expanded in the storage tier as workload context length, concurrency, or retention needs grow.
- Multiple GPU servers can mount a common cache namespace when the serving stack, model identity, tokenizer, routing policy, and cache configuration are aligned.
- Service continuity, monitoring, and lifecycle operations can be centralized on a dedicated server platform.

The goal is not to move every byte of inference state out of GPU memory. GPU memory remains the fastest working tier. The role of the SupremeRAID™ KV Cache for Rack is to provide a shareable filesystem-backed tier when the cache working set exceeds what is practical to keep local to each GPU server.

About SupremeRAID™ KV Cache for Rack

SupremeRAID™ is Graid Technology's high-performance NVMe RAID technology for modern data-intensive servers. It aggregates local NVMe SSDs into RAID volumes designed to sustain high storage throughput while keeping storage management efficient for application-heavy environments. In AI infrastructure, SupremeRAID™ provides a practical storage foundation for cache, dataset, and checkpoint paths that need both bandwidth and service continuity.

The SupremeRAID™ KV Cache for Rack is a turnkey external KV cache storage architecture for AI inference deployments. It enables storage appliance vendors, system integrators, and infrastructure teams to build a dedicated KV Cache for Rack appliance that presents cache capacity as a single shared filesystem namespace outside the GPU server. Instead of tying KV cache capacity to isolated SSD pools inside each GPU server, the architecture centralizes the cache storage layer on a SupremeRAID™ NVMe appliance.

In the tested architecture, a dedicated SupremeRAID™ KV Cache for Rack storage server connects to GPU servers over high-speed Ethernet and exports an NFS-mounted cache path. LMCache consumes this path inside the inference environment, allowing GPU servers to stay focused on model serving while KV cache capacity is delivered from a dedicated storage tier.

This model gives users and system vendors deployment flexibility. They can select the server platform, NVMe SSD configuration, and network fabric that match their AI deployment requirements, while SupremeRAID™ provides the high-performance local NVMe volume used as the cache storage backend. In broader deployments, the same architecture can support a shared cache namespace across multiple GPU servers when model identity, tokenizer, LMCache configuration, filesystem behavior, and request routing policy are aligned.

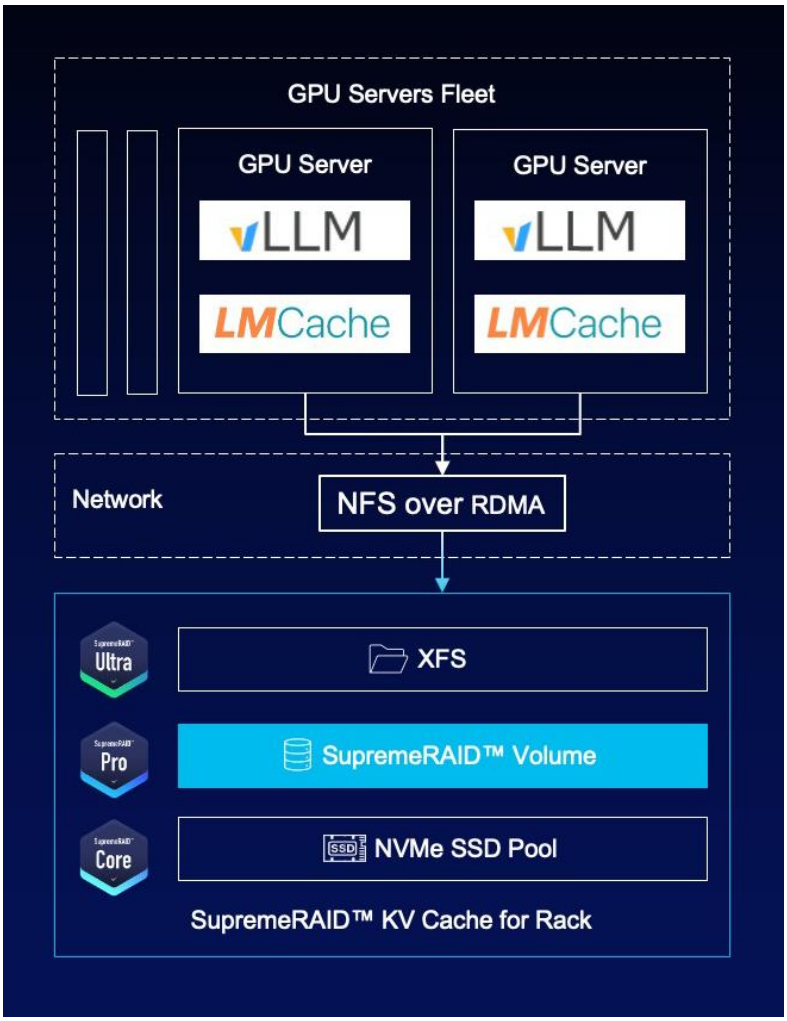

Fig. 1: SupremeRAID™ KV Cache for Rack architecture block diagram

Certified Server Platforms

SupremeRAID™ KV Cache for Rack is deployment-ready on NVMe storage server platforms from leading system vendors; including AIC, Dell, Giga Computing, Lenovo, and Supermicro. The current list of systems validated by Graid Technology can be found in Appendix B. Architects can select from this list to match form factor, processor platform, NVMe density, and expansion requirements to a specific deployment, with confidence that the SupremeRAID software stack has been qualified on the chosen hardware. The storage server used for the testing presented in this paper was the Supermicro SSG-221E-DN2R24R.

Tested Configuration - SupremeRAID™ KV Cache for Rack

The SupremeRAID™ KV Cache for Rack used for this validation was built on a Supermicro SSG-221E-DN2R24R storage server and configured as follows:

Component	Tested Configuration
Storage server model	Supermicro SSG-221E-DN2R24R
Processor	1 x Intel® Xeon® Gold 6548Y+
System memory	256 GB DDR5
Operating system	Ubuntu 24.04.3 LTS
Linux kernel	6.8.0-107-generic
SupremeRAID™ software	2.0.0-93
NVMe SSDs	10 x Kioxia CM7-V 3.2 TB
NVMe link configuration	PCIe Gen5 x2 per SSD
RAID configuration	10-drive RAID 5
RAID controller	SupremeRAID™ PAM2-032
Network adapter	Mellanox ConnectX-7 NDR, dual-port 200 Gb/s

Before running the application benchmark, the local RAID volume was validated with fio to confirm that the storage subsystem had enough bandwidth to drive the dual-port 200 Gb/s network path. The SupremeRAID™ RAID device was mounted locally for this pre-check:

```sh
mount -o noatime,nodiratime /dev/gdg0n1 /mnt/graid
```

The fio workload models the local storage behavior of KV cache offload. KV cache data is typically stored as cache-block files; each file is written sequentially during cache population and read sequentially during cache reload. The fio pre-check therefore used sequential 1 MiB writes and reads against the mounted RAID volume, with direct I/O and a 64-job, queue-depth-4 concurrency profile.

Local Storage Pre-Check	Result
Mount point	/mnt/graid
Mount options	noatime,nodiratime
fio I/O engine	libaio
I/O mode	direct=1
Block size	1 MiB
Concurrency profile	64 jobs, iodepth=4
Sequential write bandwidth	49.7 GiB/s (53.3 GB/s)
Sequential read bandwidth	69.4 GiB/s (74.5 GB/s)
Network headroom target	Dual 200 Gb/s Ethernet links (400 Gb/s)

The fio profile used for the local storage pre-check is shown below. The write and read phases use the same job shape and are listed with descriptive names for clarity.

```ini
[global]
directory=/mnt/graid
randrepeat=0
ioengine=libaio
direct=1
random_generator=tausworthe64
cpus_allowed_policy=split
group_reporting=1
cpus_allowed=0-63
size=10G
wait_for_previous=1

[sequential_write]
rw=write
bs=1m
numjobs=64
iodepth=4

[sequential_read]
rw=read
bs=1m
numjobs=64
iodepth=4
```

The local RAID 5 fio result exceeds the aggregate line-rate bandwidth of dual 200 Gb/s links, which is 400 Gb/s, or about 46.6 GiB/s (50.0 GB/s) before protocol overhead. This pre-check confirms that the tested storage server configuration had sufficient local storage bandwidth for the NFS-based KV cache validation.

The architectural value is separation. GPU servers stay focused on inference, while cache capacity is delivered from a dedicated SupremeRAID™ KV Cache for Rack that can be expanded and operated independently from the compute fleet while supporting service continuity for the shared cache tier.

Validation Methodology

The validation focuses on application throughput, because that is the outcome platform teams ultimately need to maintain as context windows grow. EvalScope drives a high-concurrency multi-turn workload through the vLLM OpenAI-compatible API while LMCache uses either no external cache tier or an NFS-mounted filesystem path backed by the SupremeRAID™ KV Cache for Rack.

Validation Focus	Purpose	Primary Metric
Qwen + vLLM + LMCache with EvalScope	Measures application throughput as prefix length increases	EvalScope Total Throughput (tok/s)

Application Throughput Scaling

The application benchmark uses a Qwen + vLLM + LMCache serving stack with KV cache offload to the SupremeRAID™ KV Cache for Rack. EvalScope drives a multi-turn OpenAI-compatible workload and reports Total Throughput as the primary application metric. The validation compares no KV cache offload against the SupremeRAID™ KV Cache for Rack while increasing prefix length.

Category	Configuration
CPU	2 x Intel Xeon Platinum 8562Y+
System memory	1 TB DDR5
GPUs used for inference	4 x NVIDIA H200

Inference engine	vLLM
Container image	vllm/vllm-openai:v0.18.1
Model	Qwen3-235B-A22B-Instruct-2507
Tensor parallel size	4
GPU memory utilization	0.9086
KV cache software	LMCache
KV cache backend	NFS-mounted filesystem path backed by SupremeRAID™ KV Cache for Rack
Host KV cache path	/mnt/graid/lmcache
Container KV cache path	/lmcache

The workload uses EvalScope random_multi_turn to exercise repeated multi-turn prompt processing under high concurrency.

Workload Parameter	Value
Benchmark tool	EvalScope perf
API	OpenAI-compatible chat completions
Dataset	random_multi_turn
Parallelism	128
Request count	512
Turns per request	3 to 5
Maximum output tokens	128
Prefix length sweep	512, 1024, 2048, 4096, 8192
Minimum prompt length	256
Primary metric	EvalScope Total Throughput (tok/s)

Total Throughput Across Prefix Lengths

Prefix Length	No KV Cache Offload Total Throughput (tok/s)	SupremeRAID™ KV Cache for Rack Total Throughput (tok/s)	Throughput Uplift
512	6,645.0404	8,817.5742	32.70%
1024	6,788.3417	9,346.9200	37.70%
2048	6,869.9656	10,098.3141	47.00%
4096	6,829.0544	10,355.9133	51.60%
8192	6,600.9698	10,126.4631	53.40%

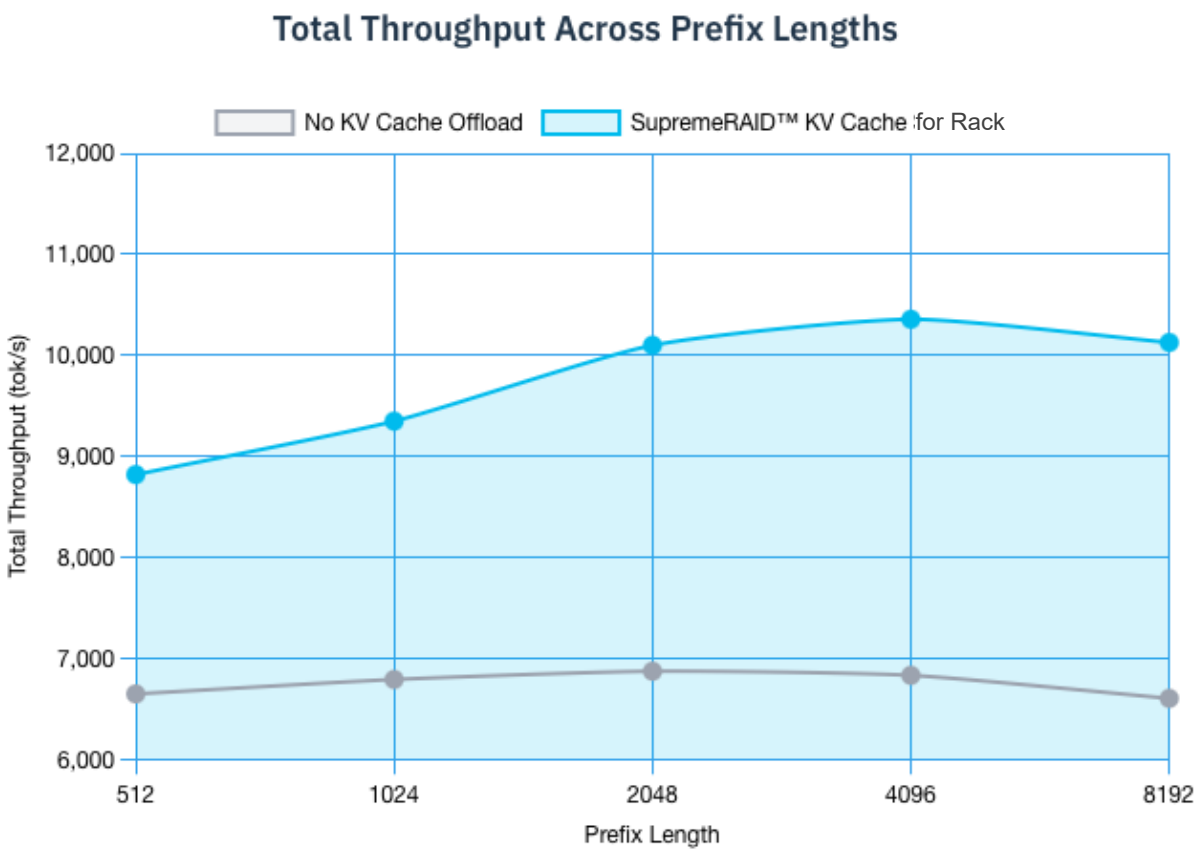

Fig. 2: SupremeRAID™ KV Cache for Rack vs. No KV Cache Offload

The SupremeRAID™ KV Cache for Rack improved Total Throughput at every tested prefix length. The benefit increased as prefixes became longer: at 2048 tokens, Total Throughput improved from 6869.9656 tok/s to 10098.3141 tok/s; at 8192 tokens, it improved from 6600.9698 tok/s to 10126.4631 tok/s. All test points completed with zero failed requests in both scenarios.

The result captures full serving-stack behavior, including model execution, request scheduling, cache reuse, and storage access through the external cache tier. The stronger improvement at longer prefixes is consistent with the role of KV cache offload: as reusable context grows, the ability to reload cached KV data becomes more valuable.

What The Results Mean for Deployment

The EvalScope result shows how the full inference stack behaves as prefix length increases. This is the customer-facing question for long-context inference: can the deployment continue to deliver strong token throughput when the prompt gets longer and the reusable KV working set grows?

The tested SupremeRAID™ KV Cache for Rack configuration consistently outperformed the no-cache baseline. The uplift was already visible at short prefixes and became more meaningful as prefix length increased:

- At 512 prefix length, Total Throughput improved by 32.7%.
- At 2048 prefix length, Total Throughput improved by 47.0%.
- At 4096 prefix length, Total Throughput improved by 51.6%.
- At 8192 prefix length, Total Throughput improved by 53.4%.

For infrastructure teams, the value is deployment flexibility with performance headroom. GPU servers can remain focused on inference compute, while KV cache capacity scales on a shareable SupremeRAID™ KV Cache for Rack designed for service-continuity-oriented cache storage. As context length grows, the external cache tier helps reduce the throughput penalty that would otherwise appear when the serving stack has to operate without an external KV cache tier.

These results validate the tested configuration and workload. Production outcomes depend on model architecture, prompt shape, cache hit rate, GPU memory allocation, client concurrency, network topology, RAID layout, drive selection, and serving-stack behavior.

Conclusion

A SupremeRAID™ KV Cache for Rack provides a shareable external storage tier for KV cache data. It addresses a practical deployment problem: GPU servers may not have enough local SSD capacity for growing KV cache workloads, and expanding local SSDs across every GPU server can limit fleet flexibility.

The EvalScope benchmark shows that the SupremeRAID™ KV Cache for Rack helps sustain higher Total Throughput as prefix length grows. In the tested Qwen + vLLM + LMCache workload, the external cache tier improved throughput at every tested prefix length and delivered a 53.4% uplift at 8192 prefix length.

For AI infrastructure teams, this design provides a cleaner way to scale long-context inference: compute capacity stays with the GPU servers, while KV cache capacity scales independently on a dedicated SupremeRAID™ KV Cache for Rack.

Learn more at graidtech.com.

Appendix A - References

- Qwen3-235B-A22B-Instruct-2507-FP8 configuration
- LMCache local storage documentation

Appendix B - Certified Server Platforms

Graid Technology has certified SupremeRAID™ KV Cache for Rack operation on these leading storage server platforms[1]

MODEL	FORM FACTOR	PROCESSOR PLATFORM	DRIVE BAYS	EXPANSION
AIC				
SB102-CA (U.S. only)	1U	1 × AMD EPYC™ 9005 / 9004	12 × Gen5 U.2 NVMe	3 × HHHL PCIe Gen5 x16 1 × OCP 3.0 PCIe Gen5 x16
SB201-SU	2U	2 × Intel® Xeon® 6 6700E / 6500P / 6700P series	24 × Gen5 U.2 NVMe	2 × FHHL / HHHL PCIe Gen5 x16 1 × OCP 3.0
HA2026-HC	2U 2N	1 × AMD EPYC™ 9000 series per node	26 × Dual-port Gen5 U.2 NVMe	3 × FHHL PCIe Gen5 x16 via Cable CEM per node
DELL				
PowerEdge R750	2U	2 × 3rd Gen Intel® Xeon® Scalable	24 × Gen4 U.2 NVMe	Up to 8 × PCIe Gen4 slots (up to 6 × FHFL x16)
PowerEdge R760	2U	2 × 4th / 5th Gen Intel® Xeon® Scalable	16 × Gen5 E3.S NVMe (front) 4 × Gen5 E3.S NVMe (rear); 24 × Gen4 U.2 NVMe	Up to 8 × PCIe Gen5 slots 1 × OCP NIC 3.0 PCIe Gen5 x16
PowerEdge R770	2U	2 × Intel® Xeon® 6	36 × Gen5 E3.S NVMe (front) + 4 × Gen5 E3.S NVMe (rear) = up to 40 total	Multiple PCIe Gen5 slots (FHFL / FHHL x16) OCP NIC 3.0
PowerEdge R7625	2U	2 × AMD EPYC™ 9004 series	32 × Gen5 E3.S NVMe; 24 × Gen5 U.2 NVMe	Up to 8 × PCIe Gen5 slots 1 × OCP NIC 3.0 PCIe Gen5 x16

MODEL	FORM FACTOR	PROCESSOR PLATFORM	DRIVE BAYS	EXPANSION
PowerEdge R960	4U	4 × 4th / 5th Gen Intel® Xeon® Scalable	16 × Gen5 E3.S NVMe; 24 × Gen4 U.2 NVMe	Up to 12 × PCIe Gen5 slots (mix of FHFL / FHHL x16)
GIGA COMPUTING				
R163-Z35-AAH1	1U	1 × AMD EPYC™ 9005 / 9004	12 × Gen5 U.2 NVMe	2 × FHHL PCIe Gen5 x16 2 × OCP NIC 3.0 PCIe Gen5 x16
R163-ZG6-AAL2	1U	1 × AMD EPYC™ 9005 / 9004	16 × Gen5 U.2 NVMe	1 × FHFL PCIe Gen5 x16 (GPU) 1 × FHFL PCIe Gen5 x16 1 × FHHL PCIe Gen5 x16 1 × OCP NIC 3.0 PCIe Gen5 x16
R283-Z96-AAJ1	2U	2 × AMD EPYC™ 9005 / 9004	24 × Gen5 U.2 NVMe	1 × FHHL PCIe Gen5 x16 1 × OCP NIC 3.0 PCIe Gen5 x16
R284-S91-AAJ1	2U	2 × Intel® Xeon® 6700 / 6500 series	32 × Gen5 E3.S NVMe	1 × FHHL PCIe Gen5 x16 1 × OCP NIC 3.0 PCIe Gen5 x16
LENOVO				
ThinkSystem SR630 V4	1U	2 × Intel® Xeon® 6 processors	12 × 2.5" or 16 × EDSFF E3.S	3 × PCIe Gen5 slots 2 × OCP 3.0 for I/O flexibility
ThinkSystem SR645 V3	1U	2 × 4th or 5th Gen AMD EPYC™ processors	4 × 3.5", 12 × 2.5", or 16 × EDSFF drives for high-performance storage	3 × PCIe 4.0 2 × PCIe 5.0 slots 1 × OCP 3.0 adapter slot
ThinkSystem SR650 V4	2U	2 × Intel® Xeon® 6 processors	Front: 24 × 2.5", 32 × E3.S, or 12 × 3.5" Mid: up to 8 × 2.5" Rear: up to 8 × 2.5" or 4 × 3.5"	Up to 10 × PCIe Gen5 slots 2 × OCP 3.0 slots
ThinkSystem SR655 V3	2U	1 × 4th Gen AMD EPYC™ or 5th Gen AMD EPYC™ processor	20 × 3.5" or 40 × 2.5" drives	Supports PCIe 4.0 and PCIe 5.0 slots
ThinkSystem SR665 V3	2U	2 × 4th or 5th Gen AMD EPYC™ processors	20 × 3.5" 40 × 2.5" drives	Up to 12 × PCIe slots (9 × PCIe 5.0) 1 × OCP 3.0 adapter slot
ThinkSystem SR780a V3 GPU Server	5U	2 × 5th Gen Intel® Xeon® Scalable processors	GPU server platform drive configuration per chassis option	Up to 10 × PCIe Gen5 x16 adapters (8 front, 2 rear)
SUPERMICRO				
AS-1116CS-TN	1U	1 × AMD EPYC™ 9005 / 9004	12 × Gen5 U.2 NVMe	2 × FHHL PCIe Gen5 x16
ASG-1115S-NE316R	1U	1 × AMD EPYC™ 9005 / 9004	16 × Gen5 E3.S NVMe	2 × PCIe Gen5 x16 2 × AIOM

MODEL	FORM FACTOR	PROCESSOR PLATFORM	DRIVE BAYS	EXPANSION
AS-2126HS-TN	2U	2 × AMD EPYC™ 9005 / 9004	23 × Gen5 U.2 NVMe	1 × FHHL PCIe Gen5 x16 1 × OCP NIC 3.0 PCIe Gen5 x16
ASG-2115S-NE332R	2U	1 × AMD EPYC™ 9005 / 9004	32 × Gen5 E3.S NVMe	2 × PCIe Gen5 x16 2 × AIOM
SSG-221E-DN2R24R	2U 2N	1 × Intel® Xeon® 5th / 4th Gen per node	24 × Dual-port Gen5 U.2 NVMe	2 × HHHL PCIe Gen5 x16 per node 2 × HHHL PCIe Gen5 x8 per node

[1] Platform qualification is performed by Graid Technology. Inclusion on this list does not imply endorsement, validation, or certification of SupremeRAID™ KV Cache for Rack by the listed system vendors.