Graid Technology
Agentic AI Storage Portfolio

Purpose-built KV Cache Solutions for Inference at Scale

April 2026

Table of Contents

The KV Cache Crisis:
Agentic Inference's Missing Memory Tier

Your GPUs aren't slow. They just have a short memory.

There's a persistent myth in AI infrastructure circles: AI is stateless. Each prompt is independent, models don't need persistent memory, and storage is an afterthought. It's a comfortable assumption; and it's completely wrong.

Every time a large language model processes a conversation, it builds a KV cache: a record of the keys and values from previous tokens that allows the model to maintain context without recomputing everything from scratch. Think of it as the model's working memory. Without it, your AI system wakes up to every conversation with no recollection of what came before.

The scale of the challenge is hard to overstate. Microsoft processed over 50 trillion tokens in a single month. Google went from 9.7 trillion to over 480 trillion; a 50x increase in a single year. Agentic AI, where models run continuously, perform multi-step tasks, and maintain session state across hours of operation, consumes up to 1,000,000x more tokens than a standard query. All those tokens need KV cache. And KV cache needs somewhere fast enough to live.

Where It Breaks

Take a real example: Llama 3-70B on an H100 80GB GPU with 128K context windows. A single user requires approximately 40GB of KV cache; manageable. A third user pushes the requirement to 120GB, and GPU HBM overflows entirely. When KV cache spills out of HBM3, it cascades down a performance cliff: Time to First Token latency spikes up to 18x. Throughput drops up to 10x. GPU utilization falls to 50%; not because the hardware is idle, but because it is burning compute cycles recomputing KV vectors it has already calculated. When evicted tokens are needed again, the GPU cannot retrieve them; it must regenerate the key and value vectors from scratch, wasting capacity on work it has already done instead of generating new output.

The result is wasted resources and delays that make production SLAs nearly impossible to guarantee. The GPU/compute side of AI infrastructure has scaled dramatically. The memory architecture supporting it has not, and when memory runs out, it's compute that pays the price.

The Cost Compounds the Performance Penalty

The financial impact is just as severe as the technical one. When HBM overflows and KV cache is evicted, the GPU doesn't sit idle, it recomputes key and value vectors for every evicted input token from scratch. That recomputation extends Time to First Token, directly reduces output and consequently raises the $ cost per token. The typical response is to add more GPUs to absorb the extra compute burden, but this treats the symptom rather than the cause: more GPUs don't eliminate the recomputation, they just give it more cycles to run on.

The real cost is larger than most teams realize. A single recomputation event can require up to 1 trillion compute operations; capacity that should be generating output tokens for users, now consumed entirely by recovery work. In the emerging economics of AI factories, where cost per token has become the defining measure of inference efficiency, every eviction quietly raises that number. Teams that treat memory architecture as a line item rather than a performance variable routinely overspend on GPU capacity while the real constraint sits at a tier below HBM, and the cost per token they report to the business reflects it.

Beyond Latency: What the Model Actually Loses

While infrastructure utilization metrics like latency and throughput are easy to measure, the more damaging consequences of KV cache eviction occur below the surface, at the model level. When context is evicted from GPU memory, the model loses more than compute cycles; it loses the pre-computed attention state it had already built. Even when that state is reconstructed, the result may be inconsistent with the original. Users will experience this as stalled tokens and uneven response times.

Managing long context windows with limited HBM often forces permanent eviction — and when tokens are permanently evicted, the model loses access to that information entirely. In practice, this produces three failure modes that are difficult to distinguish from each other:

a. **Hallucination**, where the model fabricates facts or instructions it can no longer retrieve.

b. **Inconsistency**, where long-range dependencies break and the model contradicts earlier outputs or fails to connect current generation to prior context; and

c. **Reasoning Degradation**, where intermediate 'thinking' tokens that were evicted force the model to reconstruct its reasoning verbosely from visible outputs alone, consuming more tokens and more time to arrive at a less reliable answer.

For agentic workloads, where a single session may span hundreds of steps, millions of tokens, and hours of continuous operation, any one of these failure modes can silently corrupt the entire task.

How the Industry Is Responding, and Why It Falls Short

The instinctive response to an AI performance problem is to add more GPUs. It's an understandable reflex; GPUs are the engine of inference, and more engines should mean more capacity. But when the bottleneck is memory, not compute, adding GPUs makes the problem worse, not better. Each additional GPU increases KV cache demand, amplifying the overflow and eviction patterns already driving performance loss. Overprovisioning GPUs to compensate for a memory constraint is the infrastructure equivalent of widening a highway to fix a traffic jam at the toll booth; the congestion just moves upstream, and the cost is prohibitive.

The best practice response to KV cache overflow is offloading, moving keys and values out of GPU HBM into a lower tier rather than dropping them. Two approaches dominate the current landscape:

1. The first is DRAM offloading: evicted cache is moved to CPU host memory, preserving context but introducing the CPU memory bus as a bottleneck. Retrieval times climb, latency increases under load, and at current DRAM pricing, building enough host memory to absorb production-scale KV cache overflow is prohibitively expensive; often more costly than the GPU capacity it is meant to protect.
2. The second approach is NVMe offloading: evicted cache moves to solid-state drives, which are far cheaper than DRAM and offer substantially more capacity. The tradeoff is performance. Commercial NVMe drives at up to 14 GB/s are not fast enough to serve KV cache at inference speed without introducing their own latency penalty.

Neither approach is wrong in principle. Both fail in practice because they were not designed for this workload. DRAM offloading trades cost for latency. Legacy NVMe offloading trades latency for cost, but it also introduces a failure mode the industry has largely ignored: a single drive loss means KV cache is gone, the GPU recomputes from scratch, and your SLA breaks for every session that drive was serving. Solving this requires more than legacy NVMe alone. It requires a new approach: NVMe architected for KV cache: aggregated for HBM-class bandwidth, GPU-direct to eliminate the CPU bottleneck, and RAID-protected so a single drive failure cannot break the SLA.

That is precisely what the Graid Technology KV Cache portfolio delivers.

Rethinking the Memory Tier for Agentic AI:
The KV Cache Portfolio by Graid Technology

Solving the KV cache bottleneck requires more than faster drives. It requires rethinking how storage connects to the GPU: eliminating the CPU bounce buffer, aggregating NVMe bandwidth to match HBM overflow rates, and doing it without consuming the GPU cycles needed for inference. Graid Technology's SupremeRAID™ product line delivers exactly that: 32 NVMe drives aggregated to 280 GB/s, GPU Direct Storage bypassing the CPU entirely, and RAID protection that keeps SLAs intact when drives fail. All with negligible compute overhead.

By the Numbers

The performance gap between accelerated and unaccelerated KV cache storage is not incremental, it is structural. SupremeRAID delivers KV cache reads at 1.3ms versus 100ms or more with standard NVMe, a 77x reduction in latency. Aggregate bandwidth of 280 GB/s across 32 drives matches the overflow rates that production inference clusters generate. GPU Direct Storage eliminates the CPU entirely from the data path, so the GPU receives its KV cache directly, no bounce buffer, no system memory contention, no CPU bottleneck. The result is more concurrent users on the same hardware, predictable SLAs under load, and GPU utilization that stays where it belongs: above 90%.

The **Graid Technology KV Cache portfolio** brings this capability to market across three deployment tiers, from individual inference servers to enterprise rack deployments to NVIDIA's next-generation STX reference architecture:

The Graid Technology KV Cache Portfolio

Solution	Description	Availability
KV Cache Server *Single-Node NVMe Acceleration* › **Entry Point**	Purpose-built for individual inference servers and edge AI deployments. SupremeRAID™ transforms up to 32 NVMe drives into a single 280 GB/s pool; seamlessly absorbing KV cache overflow from GPU HBM with GPU Direct Storage and zero CPU bottleneck. Ideal for on-premises AI, edge inference, and developer clusters.	***AVAILABLE NOW***
KV Cache Rack *Rack-Scale, Partner-Validated* › **Scale-Out**	A rack-scale KV Cache solution co-engineered with Supermicro, AIC, and Gigabyte. SupremeRAID runs inside validated platforms, delivering shared, high-bandwidth NVMe storage across the entire AI cluster in a single rack. Designed for enterprises deploying multi-GPU inference at scale without building custom infrastructure.	***AVAILABLE NOW***
KV Cache Platform *NVIDIA STX-Native Architecture* › **Platform-Native**	Aligned to NVIDIA's STX reference architecture and CMX context memory platform. SupremeRAID serves as the G3.5 storage performance engine, the NVMe acceleration layer beneath BlueField-4 DPUs and DOCA Memos, making instant agentic context handoff between GPUs viable at inference speed.	***COMING H2 2026*** *Native BlueField-4 execution*

A Natural Scale Story

The three tiers follow a clear progression: Server → Rack → Platform. Each tier targets a distinct deployment model, individual node, shared cluster, and fabric-native, while sharing a common technology core in SupremeRAID. Customers can start with a single KV Cache Server and scale to a KV Cache Rack as inference workloads grow, with a clear path to NVIDIA STX alignment via the KV Cache Platform.

Looking Ahead

Graid Technology's roadmap extends the KV Cache Platform's significantly on multiple fronts:

- **Native execution on the BlueField-4 DPU (2H 2026)** will expand SupremeRAID's deployment model to run directly within the STX storage chassis; extending the platform from GPU-adjacent to DPU-native, and giving infrastructure teams a fully integrated STX storage node either with or without a discrete accelerator.
- **Expanded drive count support** will allow a single SupremeRAID instance to span multiple CMX chassis, serving an entire rack of STX storage nodes at significantly greater aggregate bandwidth from one virtualized pool, simplifying DOCA Memos namespace management and delivering rack-scale throughput from a single logical storage resource.

Why Now

Agentic AI has broken the memory model that single-shot inference was built on. Models that reason across dozens of steps, coordinate between agents, maintain session state for hours, and grow context windows into the millions of tokens require a fundamentally different infrastructure approach, not more GPU, but a new memory tier: fast enough to feed the GPU, large enough to hold persistent session state, and economical enough to deploy at scale.

These workloads are not hypothetical, they are already in production. Autonomous coding agents maintain active context across multi-hour sessions, reading entire codebases, running tests, and iterating on results without ever resetting state. Enterprise document processing agents reason across thousands of pages in a single unbroken thread. Customer service platforms sustain parallel long-context conversations for thousands of simultaneous users. Scientific research agents coordinate across specialized sub-agents, each holding its own context while drawing on a shared memory pool. What all of

these have in common is a hard dependency on persistent, low-latency KV cache storage, and an inability to tolerate the performance cliff that standard NVMe delivers the moment HBM overflows.

At GTC 2025, Jensen Huang made a prediction:
"For the very first time, your storage system will be GPU-accelerated."

There, he described a vision that Graid Technology is proud to have pioneered. A year later, we see that vision now becoming real-life products with the announcement of the STX architecture and CMX Context Memory Platform as the infrastructure blueprint for agentic AI at scale.

Conclusion

The Graid Technology KV Cache Portfolio represents the next step of our progression, built for precisely this moment.

It is the only NVMe acceleration solution purpose-aligned to the G3 and G3.5 storage tier at the heart of the STX and CMX architecture. This shift to agentic AI is not a future event; it is already reshaping production infrastructure requirements today. The teams that solve the storage layer first will deploy more agents, serve more users, and do it on the hardware they already own. They will also do it at a fraction of the cost of alternatives: NVMe-based KV cache acceleration delivers HBM-class performance at storage-tier economics, eliminating the need to overprovision GPUs, avoid expensive DRAM expansion, and rebuild lost context after every drive failure.

Better performance and lower TCO are not a tradeoff; with the right storage architecture, they are the same outcome. Graid Technology's KV Cache portfolio exists to make that possible.

To learn more about Graid Technology's AI offerings, visit graidtech.com/ai. If you have further inquiries or would like to discuss your deployment, reach out to us at info@graidtech.com.

Read the blog: Your GPUs Aren't Slow. They Just Have a Short Memory.

About Graid Technology

Graid Technology is building the storage backbone for the future of AI, enterprise, and high-performance computing. As the creator of SupremeRAID™, the world's first and only GPU-based RAID, and the global steward of Intel® Virtual RAID on CPU (Intel® VROC), Graid Technology delivers flexible RAID solutions that maximize NVMe performance while ensuring resilient, scalable data protection for modern data infrastructure. Headquartered in Silicon Valley with global operations and R&D in Taiwan, Graid Technology is advancing RAID innovation for the next generation of data-intensive workloads. To learn more, visit graidtech.com.