Breaking the GPU Memory Wall: >20x Faster LLM Inference at Multi-Million-Token Scale

Results from a joint proof-of-value: Lightbits Inferra™ orchestrating KV cache reload over a Solidigm PCIe 5.0 NVMe cluster, tested against NVIDIA B200 GPUs.

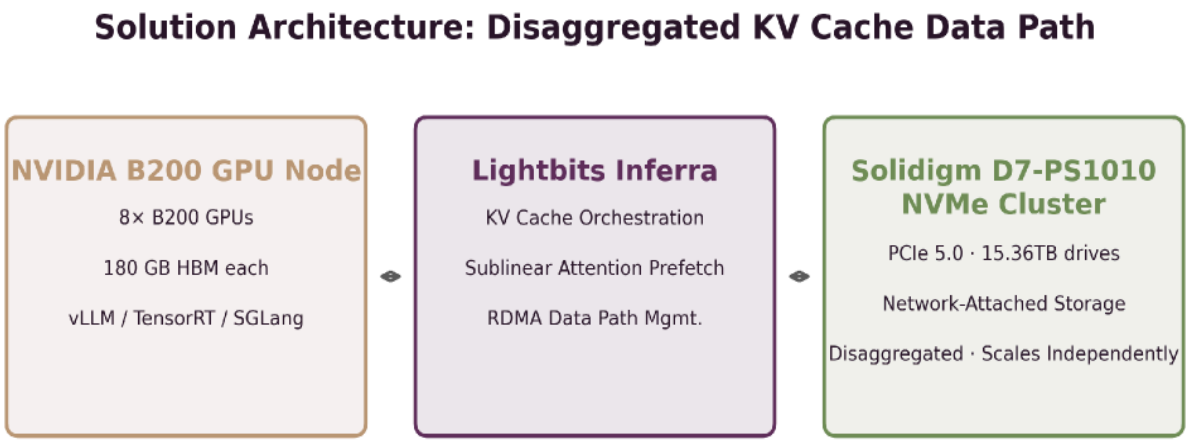
PERFORMANCE CASE STUDY

September 2026

As LLM context windows push past a million tokens and the sessions grow, the key-value (KV) cache that drives fast inference no longer fits in GPU memory. Lightbits Labs and Solidigm set out to test whether reloading that cache from a disaggregated NVMe cluster — instead of recomputing it on the GPU — could hold up as context and concurrency scale. The results so far, below, show it does.

The Architecture

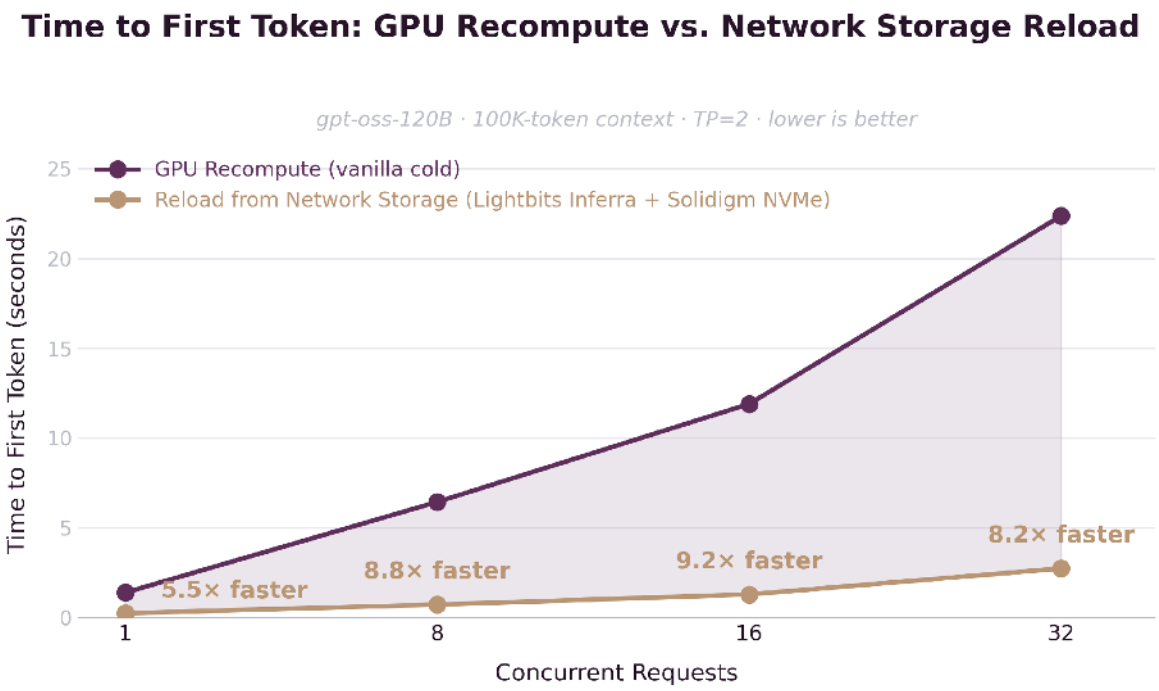
The solution pairs three purpose-built layers into a single inference data path. NVIDIA B200 GPU nodes handle compute — running vLLM, TensorRT, or SGLang — and act as the client for every KV cache reload and recompute. A disaggregated cluster of Solidigm™ D7-PS1010 NVMe SSDs (PCIe 5.0, 15.36TB) serves as network-attached storage, scaling capacity independently of GPU count. Inferra by Lightbits connects the two: KV cache orchestration software that decides what to prefetch, retain, and evict across the fabric—turning the storage tier into a practical extension of GPU memory rather than a passive backing store. The layers communicate over RDMA (RoCEv2) across multiple 200G network rails and BlueField-3 DPUs, giving the storage tier enough throughput to keep GPUs fed without becoming the bottleneck.



>20×
faster than GPU recompute — Llama-4-Scout, ~4.8M-token context

>9×
faster time-to-first-token along with 16x concurrency, 100K-token context

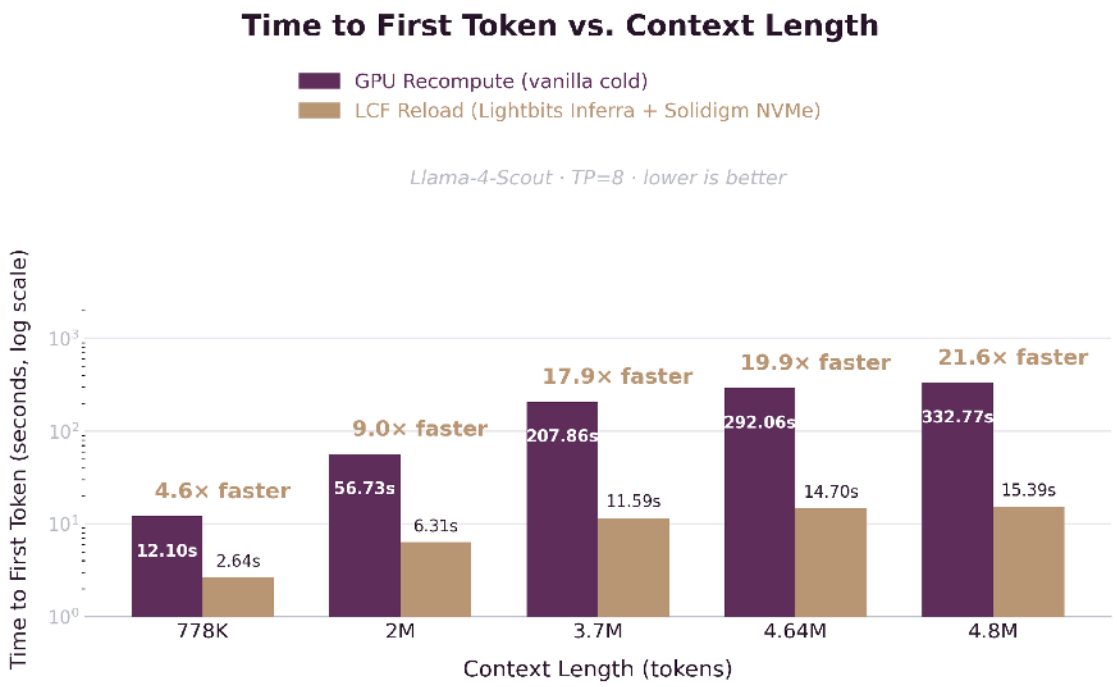
~70,000
concurrent 100K-token conversations served live from one cluster



The gap widens with load: GPU recompute cost scales with concurrency, while reload cost — bound by KV cache size, not compute — barely moves. At 32 concurrent requests, reloading from the network is 8.2x faster than GPU recomputation.

Fastest Where It Matters Most: Multi-Million-Token Context

The advantage compounds as context grows, because recompute cost scales with tokens while reload cost scales with cache size on a high-throughput fabric. On Llama-4-Scout at a ~4.8-million-token context — where GPU recompute takes more than 5 minutes (332 seconds) — reload from an Inferra + Solidigm NVMe cluster returned the same ~0.86 TB KV cache, byte-identical to the original output, in 15.4 seconds: 21.6x faster, at an effective ~56 GB/s.



Real Capacity, Real Concurrency

The same cluster was filled with distinct 100K-token conversations until it reached capacity: 69,244 live conversations held simultaneously — nearly 70,000 concurrent, full-context sessions — on 5 of the cluster's 6 nodes. Extrapolated to the full, correctly configured cluster, that scales to roughly 89,200 concurrent 100K-token conversations. Reload latency stayed steady even near full capacity, with mid-fill conversations reloading in under 2 seconds. Inferra does predictive prefetching of data from different memory tiers and makes it available to the compute block just when it needs it.

Cluster capacity and concurrent session scaling

Configuration	Usable Capacity	Data Drives	Concurrent 100K-Token Conversations
Tested (5 nodes)	607 TB	56	69,244 — measured
Full cluster (6 nodes)	**780 TB**	**72**	**~89,200 — projected**

Tested figures are measured; full-cluster figures are extrapolated from the same measured capacity-per-conversation factor.

Note
Test environment: single NVIDIA B200 GPU node (8× B200) as inference client; Solidigm 15.36TB PCIe 5.0 NVMe cluster (D7-PS1010) as the disaggregated storage tier, connected via RDMA over 4×200G RoCEv2 rails. The metric is client-side time-to-first-token; we measured GPU-recompute and network-reload paths on the same deployment, config-matched. Results are interim; testing continues across the full 6-node cluster.

What's Next: From Proof to Production

These numbers are early proof, not the ceiling. Lightbits Labs and Solidigm are extending this work to the full six-node cluster, widening the workload matrix, and closing the fan-in gap that already lets a single node run ahead of the cluster average. As context windows grow, idle GPU memory gets more expensive — a disaggregated, NVMe-backed KV cache turns that liability into capacity the fleet can sell.

For neoclouds and hyperscalers running GPU fleets at capacity, the opportunity is not cheaper storage — it is more revenue from the GPUs already deployed. Inferra by Lightbits turns Solidigm's high-performance PCIe 5.0 NVMe into a practical extension of GPU memory, so the same B200 cluster can serve more concurrent sessions, longer contexts, and faster time-to-first-token without adding a single GPU. **Same hardware. More revenue.**

To learn more about Inferra, go to www.lightbits.ai.

About Solidigm

Solidigm is a pioneer in enterprise data storage, headquartered in Rancho Cordova, California, and operating globally as a standalone subsidiary of SK hynix. Its D7-PS1010 — the PCIe 5.0 drive used as network-attached storage in this study — is Solidigm's flagship performance SSD, built on 176-layer TLC 3D NAND and rated for up to 3.1M random-read IOPS and 14,500 MB/s of sequential read throughput.

This testing was conducted in the Solidigm AI Central Lab, a purpose-built facility operated by AI Infrastructure provider FarmGPU that brings together storage and AI capabilities to perform cutting-edge research and improve bottom-line results. The lab features high-performance GPUs, 800 Gbps Ethernet networking, and extensive Solidigm SSD infrastructure, all designed on reference architectures that mirror what hyperscalers and enterprises are deploying in data centers worldwide, making the findings broadly applicable to customer environments.

About Lightbits Labs

Lightbits Labs (Lightbits®) invented the NVMe over TCP storage protocol, which is natively built into its industry-leading block storage, and introduced the first KV cache optimization engine for AI acceleration. Lightbits data infrastructure solutions are engineered to deliver unmatched high performance and maximum hardware efficiency for LLM inference, real-time analytics, and transactional workloads. Lightbits is backed by enterprise technology leaders [Cisco Investments, Dell Technologies Capital, Intel Capital, J.P. Morgan, Lenovo, and Micron] and is on a mission to deliver best-in-class, cost-efficient software solutions for performance-sensitive workloads at scale.

🌐 www.lightbitslabs.com ✉ info@lightbitslabs.com

US Offices
1830 The Alameda,
San Jose, CA 95126, USA

Israel Office
17 Atir Yeda Street,
Kfar Saba 4464313, Israel

LBWP27/2026/9