LightInferra Optimized Inference

March 2026

Table of Contents

Executive Summary

AI inference is entering a long-context era. As models move from chat to agentic workflows, retrieval-augmented generation, and multi-document reasoning, sequence lengths rise rapidly—and KV cache footprint grows even faster than GPU memory. The result is a familiar operational pattern for inference providers: GPU stalls, tail-latency spikes, and throughput collapse precisely when customers need consistent service levels.

LightInferra™ Optimized Inference is designed to eliminate the GPU efficiency penalties that long-context workloads impose—delivering consistent, SLA-grade performance precisely when traditional infrastructure breaks down. Through proactive data movement across memory tiers, LightInferra ensures the right KV cache is always ready at the point of need, before inference is ever impacted. The result: materially higher sustained QPS/TPS, predictable SLAs under real workloads, and dramatically better utilization from the same GPU fleet—without changing your models, your stack, or your hardware.

Built from first principles for AI inference, LightInferra treats security and QoS not as afterthoughts but as core platform capabilities—giving inference providers the controls they need to operate confidently at scale.

Problem with AI

Why general-purpose storage (even with RDMA and S3) falls short

The root issue: long-context inference requires KV cache at a scale that cannot remain entirely resident in GPU HBM (High Bandwidth Memory). Traditional “tiering” approaches rely on a small working set in HBM while the remainder is placed in DRAM, NVMe, or remote tiers. When attention needs an offloaded KV page, the system fetches it *reactively* and the GPU stalls while it waits. This creates GPU-visible KV page faults on the hot path, increasing tail latency and reducing effective throughput.

The reactive pattern plays out the same way regardless of the underlying tier: the GPU holds a sliding-window working set while overflow blocks live in DRAM, NVMe, or object storage. When attention needs an offloaded block, decoding stalls on a blocking fetch—degrading tail latency, collapsing throughput, and leaving inference providers with no path to the SLA-grade latency guarantees their customers expect.

Traditional inference stacks attempt to reduce the frequency of these stalls by reusing cached KV state across requests that share a common prefix, reducing redundant computation. Where they break down is under memory pressure. When the cache fills up, blocks are evicted—and the next request that needs them either triggers a recompute or forces a context swap, preempting an in-flight user session to reclaim HBM. The GPU stall moves later in the pipeline, but it doesn't disappear.

LightInferra was built to eliminate this failure mode entirely: KV cache is ready when the GPU needs it, protecting SLAs continuously, not retrieved after inference has already broken them.

Limitations of General-Purpose File Systems with RDMA and S3

Even when “fast” building blocks are present (RDMA networks, high-performance storage, object stores like S3), general-purpose storage stacks are not inference-aware. They tend to impose overhead and unpredictability through:

- **Reactive data movement**: fetching KV blocks only after a miss is observed - too late for SLA-grade inference.
- **Metadata bottlenecks**: page/block lookup, mapping, and eviction decisions aren’t aligned to attention’s imminent access pattern; inefficient metadata management amplifies stalls, jittering and can induce GPU-visible page faults.
- **Mismatched abstractions**: object storage and parallel filesystems were built for data lakes, checkpoints, and generic high throughput workloads - not for micro-latency, deadline-driven fetches needed to keep attention kernels fed. Lightbits highlights that a distributed key-value abstraction is the natural fit for KV cache extension, bypassing issues common in object storage and parallel file systems.
- **Security of user data**: KV cache contains real user data—prompts, context, conversation history—and traditional storage infrastructure was never designed to protect it at the session level. Parallel filesystems, S3 buckets, and self-encrypting NVMe all encrypt the wire and the disk, but expose plaintext data to operators, mount points, and every privileged component in the Trusted Computing Base by design. Employees, threat actors, closed firmware, and opaque kernel modules all sit above that boundary. KV cache can be converted back to plaintext user conversations using the model's own embeddings—no decryption required. For inference providers, that's not a theoretical risk. It's a structural exposure that exists in every standard deployment.

The Opportunity: a Better Solution, Purpose-Built for Inference Efficiency & Security

Long-context inference demands two things that general-purpose infrastructure cannot deliver: predictable, proactive KV cache readiness that sustains SLA-grade QPS/TPS under real workloads, and security controls that ensure only authorized services can ever touch user data—per-Agent cache encryption, KMS-managed keys, configurable stale data expiry that automatically purges cache after a service-provider-defined retention period, and defense-in-depth that keeps operators in control even when perimeter defenses fail. LightInferra exists to fill both gaps: an inference-first platform designed from the ground up to eliminate GPU stalls and enforce strict user data protection at every layer—including ensuring that data which no longer needs to exist doesn't, user data access is strictly controlled, and per-Agent encryption that is the industry's strongest tool for user data defense.

Benchmark Results

LightInferra Platform

The following benchmarks compare a baseline cache-regen mode versus the LightInferra path across increasing sequence lengths. The most important customer-facing takeaway is the scale behavior: turn-two TTFT remains dramatically lower under long context, protecting throughput and SLAs.

Benchmark A: (DeepSeek-R1-Distill-Llama-70B, FP8 Dynamic)

At 130,939 tokens (turn 2):

- Baseline (cache-regen): TTFT 70,838 ms, latency 77,245 ms
- LightInferra: TTFT 465 ms, latency 6,855 ms
- Improvement: **152x faster TTFT**, **11.3x lower latency** (turn 2)

This is the practical difference between an inference service that times out or violates SLA versus one that remains usable and monetizable under extreme context.

Benchmark B: (Qwen2.5-7B-Instruct-1M)

At 409,600 tokens (turn 2):

- Baseline: TTFT 72,561 ms, latency 74,481 ms
- LightInferra: TTFT 793 ms, latency 2,714 ms
- Improvement: **91.5x faster TTFT**, **27.4x lower latency** (turn 2)

At 1,009,867 tokens (turn 2):

- Baseline: TTFT 372,259 ms, latency 375,661 ms
- Improved path: TTFT 1,301 ms, latency 4,704 ms
- Improvement: **286x faster TTFT**, **79.9x lower latency** (turn 2)

In customer terms: these results translate to dramatically higher usable capacity from the same GPU fleet during long-context requests - protecting QPS, reducing tail-latency penalties, and increasing revenue per deployed GPU.

LightInferra: Summary

LightInferra is a bespoke platform built specifically to optimize AI inference efficiency end-to-end—not a legacy storage system retrofit for inference. Built from first principles, it gives inference providers the controls that actually matter at scale: per-agent in-flight and at rest encryption, robust auditing and data compliance controls, enforceable SLA guarantees, and strong QoS policy controls that maintain SLA commitments across all inference workloads.

Benefits of LightInferra

Designed from the Ground Up for Inference Requirements

LightInferra is designed to convert long-context inference from a stall-prone workflow into a predictable, SLA-driven service. Key customer benefits include:

1) Higher GPU efficiency and lower cost per token

By preventing GPU stalls before they happen, LightInferra increases the fraction of time GPUs spend generating billable tokens—not waiting on KV movement. Lightbits highlights "cost-efficiency" gains and the ability to serve more requests on the same GPUs (e.g., up to ~3× request capacity in product materials).

2) SLA-ready performance under long context

LightInferra includes strong SLA/QoS policy controls so that latency-sensitive workloads can be protected even in mixed environments with trials and background jobs—helping managed inference providers deliver consistent TTFT and inter-token latency.

3) Predictable behavior across tiers

LightInferra supports tiered cache use across HBM, DRAM, and NVMe, including hyper-converged and disaggregated configurations, managing full-spectrum memory and network tail latencies.

4) Security by design for multi-tenant inference

LightInferra enables encryption enforced between sessions and tenants, with integration to KMS systems and configurable data management policies for access and deletion.

5) Future-proof foundation for inference-first infrastructure

Lightbits describes the platform as inference-first and built on a distributed key-value abstraction that aligns naturally with KV cache needs, avoiding performance issues seen in object storage and parallel file systems.

About Lightbits Labs™

Lightbits Labs (Lightbits) is leading the digital data center transformation by making high-performance elastic block storage available to any cloud. Creators of the NVMe® over TCP (NVMe/TCP) protocol, Lightbits software-defined storage is easy to deploy at scale and delivers performance equivalent to local flash to accelerate cloud-native applications in bare metal, virtual, or containerized environments. Backed by leading enterprise investors including Cisco Investments, Dell Technologies Capital, Intel Capital, JP Morgan Chase, Lenovo, and Micron, Lightbits is on a mission to make high-performance elastic block storage simple, scalable and cost-efficient for any cloud.

🌐 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

The information in this document and any document referenced herein is provided for informational purposes only, is provided as is and with all faults and cannot be understood as substituting for customized service and information that might be developed by Lightbits Labs ltd for a particular user based upon that user's particular environment. Reliance upon this document and any document referenced herein is at the user's own risk.
The software is provided "As is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. In no event shall the contributors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the software or the use or other dealings with the software.
Unauthorized copying or distributing of included software files, via any medium is strictly prohibited.

LBWP23/2026/3