Solution Overview

NVIDIA CMX Context Memory Storage Platform

Rearchitecting storage for the next frontier of AI.

AI-native organizations are hitting new limits as agentic, multi-turn workflows and growing context windows drive rapid growth in key value (KV) cache usage, straining existing memory and storage tiers. The NVIDIA CMX context memory storage platform introduces a dedicated context memory tier that extends effective GPU memory, boosts tokens per second, and improves power efficiency for long-context inference. CMX transforms context into a shared, high-bandwidth resource across pod-scale AI systems by treating KV cache as a new AI-native data type and using a co-design approach across compute, networking, and storage. It enables organizations to scale agentic AI workloads with higher performance, better tokens per watt, and improved cost per token.

Challenges for AI Inference at Scale

Organizations are under pressure as agentic, multi-turn workflows and growing context requirements increase KV cache footprints faster than existing memory and storage tiers can absorb them. This drives up costs, leaves GPUs underutilized, and makes it harder to reliably scale long-context inference.

- **Growing KV cache pressure from long-context workloads:** Organizations are seeing rapid KV cache growth as multi-turn, context-rich workloads expand, stressing GPU High Bandwidth Memory (HBM), system memory, and local storage capacity. This reveals limitations in existing storage hierarchies that were not designed to manage large volumes of frequently reused inference context.
- **Memory and storage tiers not designed for KV cache:** Existing tiers are optimized either for speed or for durable enterprise data, forcing KV cache into scarce HBM or general purpose storage that adds latency, cost, and power overhead. As context grows, standard storage tuned for durability and data services cannot efficiently support latency-sensitive, ephemeral KV cache.
- **Underutilized GPUs and higher cost per token:** Recomputing history and accessing KV cache from inefficient storage tiers introduce decode stalls and GPU idle time, reducing throughput and increasing cost per token. These inefficiencies translate directly into wasted GPU capacity as workloads scale.
- **Latency and inefficiency when context is not shared effectively:** Without an efficient shared KV cache layer across pods and rack-scale clusters, context reuse depends on inefficient data paths, increasing latency and limiting performance for long-context, multi-agent inference. As deployments scale out, lack of coordinated sharing further degrades consistency and responsiveness.

Key Customer Challenges

- **Rapidly growing context and KV cache:** Agentic, multi-turn workflows and context windows reaching into the millions of tokens drive rapid growth in KV cache, stretching existing GPU memory and storage hierarchies.
- **Inefficient memory and storage tiers for KV cache:** Existing storage tiers are optimized either for speed or durability, forcing KV cache into scarce HBM or general-purpose storage that adds latency, cost, and power overhead.
- **Underutilized GPUs and higher cost per token:** Recomputing history and accessing KV cache from inefficient tiers leads to decoder stalls, idle GPUs, and rising energy use, inflating cost per token at scale.
- **Latency and inefficiency in sharing context across rack-scale AI systems:** Without an efficient shared KV cache layer, context reuse across pods and rack-scale clusters depends on slow or unoptimized paths, increasing latency and limiting performance for long-context and multi-agent inference.

These challenges create significant performance and cost impacts. They limit tokens per second per GPU, increase power consumption per useful token, and make it difficult to keep accelerators fully utilized. Over time, this raises cost per token, constrains how far existing data centers can scale agentic AI workloads, and slows the rollout of new long-context applications that depend on fast, efficient KV cache reuse.

The NVIDIA CMX context memory storage platform

NVIDIA CMX is a new class of AI-native storage infrastructure that creates a dedicated context memory tier for long-context, multi-turn, and multi-agent inference. CMX is powered by the NVIDIA® BlueField®-4 processor and is part of the NVIDIA Vera Rubin platform, extending effective GPU KV cache capacity and enabling high-bandwidth sharing of context across the pod. By treating KV cache as a new class of AI data type rather than transient memory, CMX helps organizations improve responsiveness, increase throughput per GPU, and scale long-context inference with better power efficiency and cost per token.

CMX reflects NVIDIA's extreme co-design approach across compute, networking, and storage. These infrastructure components work together so inference context can be placed, moved, and reused efficiently across memory and storage tiers, plus the new pod-level context tier. This turns KV cache into a shared, high-bandwidth service for agentic workloads instead of a bottleneck that limits scale.

- **NVIDIA BlueField-4 STX Storage Processor for Context Memory Storage:** The NVIDIA BlueField-4 STX storage processor powers the CMX platform, combining NVIDIA Vera CPU cores, dedicated datapath accelerators, and high-speed NVIDIA ConnectX®-9 networking to manage context memory efficiently. In CMX target nodes, BlueField-4 acts as the storage controller for NVMe flash, offloading data movement, integrity, and encryption for KV cache with high-power efficiency. In NVIDIA Rubin compute nodes, BlueField-4 DPUs provide the initiator side for CMX, handling RDMA storage traffic over NVIDIA Spectrum-X™ Ethernet networking and enforcing secure, isolated access to the context memory tier.
- **KV-Aware Acceleration and Services With NVIDIA DOCA, NVIDIA Dynamo, and NIXL:** CMX uses NVIDIA DOCA™ Memos, NVIDIA Dynamo, and NVIDIA Inference Transfer Library (NIXL), to manage KV cache as a shared resource across compute and storage. DOCA Memos provides KV-aware services and I/O control on BlueField-4, while NVIDIA Dynamo and NIXL integrate context placement and reuse into the inference serving layer. Together, they determine which KV blocks remain in GPU memory, which move to host memory or local SSDs, and which reside in the CMX context tier. This improves tokens per second, reduces time to first token, and enables pod-wide context reuse for multi-turn, multi-agent workloads.
- **AI-Optimized Fabric With NVIDIA Spectrum-X Ethernet:** NVIDIA Spectrum-X Ethernet provides an AI-optimized RDMA fabric that connects NVIDIA Rubin compute nodes and CMX target nodes. It delivers predictable, low-latency, high-bandwidth access to KV cache through features such as advanced congestion control, adaptive routing, and lossless RoCE. By providing consistent performance at scale and integrating directly with CMX and NVIDIA Rubin pods, Spectrum-X Ethernet helps keep GPUs supplied with context, supporting high throughput and responsive long-context inference.

NVIDIA Benefits

- **AI-native storage for content memory:** BlueField-4 powers an AI-native storage infrastructure designed for large-scale inference, extending effective GPU memory and transforming KV cache into shared long-term memory across NVIDIA Vera Rubin pods.
- **Up to 5x higher tokens per second with extended context memory:** A dedicated context memory tier extends GPU memory capacity and boosts tokens per second compared to traditional storage approaches for long-context, multi-turn, and multi-agent inference.
- **Up to 5x better power efficiency:** Treating KV cache as ephemeral AI-native data avoids heavy general-purpose storage services and enables better power efficiency for long-context inference while reducing recomputation and GPU idle time.
- **Extreme co-design across compute, networking, and storage:** NVIDIA's extreme co-design tightly integrates NVIDIA Rubin GPUs, BlueField-4, DOCA, Spectrum-X Ethernet, NVIDIA Dynamo, NIXL and context-aware orchestration into a single platform that intelligently manages KV cache across rack-scale systems to boost responsiveness and reduce time to first token.

Scaling Agentic AI With Context Memory

CMX delivers up to 5x higher sustained tokens per second (throughput) and up to 5x better power efficiency for long-context and agentic inference compared with traditional storage-based approaches. This means organizations can serve more queries per GPU, reduce recomputation and idle time, and fit significantly more usable AI capacity within the same power and data center footprint. By extending effective GPU memory and turning KV cache into a shared, high-bandwidth resource, CMX improves both performance and utilization, which directly lowers cost per token and improves return on existing and future AI infrastructure investments.

These gains compound at scale. Higher tokens per second and better tokens per watt allow organizations to support larger context windows and more advanced agentic workflows without a linear increase in GPU count or rack space. At the same time, offloading KV cache management to CMX reduces pressure on general purpose storage systems and host CPUs, simplifying operations and enabling teams to plan expansions around GPU capacity and application requirements rather than storage bottlenecks.

> **Consistent AI performance at scale with an optimized fabric and ecosystem:** An AI-optimized Spectrum-X Ethernet fabric provides predictable, low-latency, high-bandwidth access to shared context, while a broad ecosystem of storage and system partners gives customers validated paths to deploy high-throughput, long-context inference at scale.

Dimension	With CMX	Business Impact
Throughput (TPS)	Up to 5x higher sustained TPS for long-context workloads	More queries served per GPU and per rack
Power Efficiency	Up to 5x better power efficiency for KV cache operations	More tokens per watt, better use of available power budget
GPU Utilization	Higher utilization through shared, pre-staged KV cache	Less idle time, higher effective GPU capacity
Cost per Token	Reduced cost per token, better efficiency and reuse	Lower unit cost for long-context and agentic AI
KV Cache Sharing	Pod-wide, shared KV cache accessible across GPUs and nodes	Faster context access, reduced duplication, and improved scalability

CMX Storage Partner Solutions

NVIDIA CMX is supported by a broad ecosystem of storage and system partners that are aligning their next-generation platforms with NVIDIA BlueField-4-powered context memory. These partners give organizations multiple validated paths to deploy shared KV cache and integrate CMX with new NVIDIA Vera Rubin deployments.

Ready to Get Started?

To learn more about the NVIDIA CMX platform, visit: www.nvidia.com/cmx, or contact an NVIDIA sales representative: www.nvidia.com/contact-sales

NVIDIA