NEWS RELEASE

Penguin Solutions Introduces Industry's First Production-Ready CXL-Based KV Cache Server

2026-03-16

Penguin Solutions MemoryAI KV cache server, an 11TB memory appliance, enables efficient deployment of enterprise-scale AI inference

FREMONT, Calif.--(BUSINESS WIRE)-- **Penguin Solutions**, Inc. (Nasdaq: **PENG**), the AI factory platform company, today announced the industry's first production-ready KV cache server that utilizes CXL memory technology to address the critical "memory wall" challenge in AI inferencing—Penguin Solutions MemoryAI™ KV cache server. This innovative solution delivers up to 11 TB of CXL-based memory engineered to optimize performance of enterprise scale inference, including agentic AI. The result is lower latency, higher throughput, increased efficiency of GPU clusters, consistent achievement of stringent service-level agreements (SLAs), and faster time-to-first-token (TTFT).

Penguin Solutions MemoryAI KV cache server is the industry's first production-ready KV cache server that utilizes CXL memory technology to address the critical "memory wall" challenge in AI inferencing. The innovative solution delivers up to 11 TB of CXL-based memory engineered to optimize performance of enterprise scale inference, including agentic AI.

While model training and tuning is primarily compute-bound and occurs episodically, the continuous memory-bound and latency-sensitive inference workloads required for inference and agentic AI are complex and fundamentally different. Inference demands are typically 30% compute driven (GPU) and 70% memory driven (RAM), elevating the need for greater memory capacity and causing performance bottlenecks and GPU idle time. Accelerating memory-dependent AI processes, Penguin's MemoryAI KV cache server increases memory capacity by integrating 3 TB of DDR5 main memory and up to eight 1 TB CXL Add-in Cards (AICs).

"CXL-enabled KV cache technology delivers faster time-to-first-token, reduced time per output token, and increased overall end-to-end token throughput," said Phil Pokorny, chief technology officer at Penguin Solutions. "These

critical performance improvements enable enterprise-scale inferencing across many users who expect low latency and timely access to AI-generated insights. The introduction of Penguin's MemoryAI KV cache server is designed to help enterprises sustain these performance improvements and consistent service standards as model size, context windows, precision requirements, and concurrency demands continue to grow."

By significantly expanding the memory available to GPUs, the server enables organizations to mitigate GPU memory bandwidth limits, reduce redundant re-compute operations, and optimize clusters for inference performance. This increased system efficiency also enables organizations to train larger models and process expansive datasets faster.

Benefits of Penguin Solutions MemoryAI KV cache server in Cluster Design

With expanded, disaggregated memory, the server offers several operational benefits:

- Support for larger context size and concurrency: Penguin's MemoryAI KV cache server is particularly crucial for enterprise-scale tasks requiring large context windows and minimal latency, including real-time financial news parsing, retrieval-augmented generation (RAG) over massive 10-K datasets, and regulatory compliance analysis.
- Flexibility to tier cluster memory: CXL-based KV cache delivered by the server creates a new tier of cluster memory to supplement existing high bandwidth memory (HBM) and system DRAM, delivering speeds 10x faster than NVMe-based approaches. This provides new flexibility in offloading KV data for faster access.
- Compatibility with NVIDIA Dynamo: The solution is compatible with NVIDIA Dynamo, NVIDIA's software architecture for KV cache memory offloading.
- Cost and power efficiency: The server enables organizations to maximize the efficient use of GPUs by adding large memory pools and optimizes clusters by right-sizing GPUs and memory. Additionally, the solution provides efficient operation, drawing less power than equivalent GPU servers.

The Penguin Solutions MemoryAI KV cache server builds upon Penguin Solutions' legacy of innovation in high-performance computing expertise, with customers already deploying the solution to optimize cluster performance and meet demanding latency SLAs for production AI workloads.

Explore Penguin Solutions' MemoryAI KV cache server **page** or visit booth #1031 at the NVIDIA GTC AI Conference and Expo March 16-19, 2026, in San Jose, Calif.

About Penguin Solutions

The most transformative technological advancements are often the hardest to deploy and optimize. Penguin Solutions, the AI factory platform company, has the innovative technologies, skills, experience, and partnerships needed to turn your AI ambitions into reality.

In addition to our AI capabilities, Penguin Solutions offers memory and LED solutions serving a wide range of high-performance and specialized applications.

For more information, visit **https://www.penguinsolutions.com**.

PR Contact

Maureen O'Leary
Corporate Communications, Penguin Solutions
1-602-330-6846
pr@penguinsolutions.com

Source: Penguin Solutions, Inc.