PENGUIN™
SOLUTIONS

NEWS RELEASE

Penguin Solutions’ OriginAI Factory Platform Delivers Optimized Performance for AI Inference

2026-03-16

Breakthrough KV cache technology provides low latency, high throughput inference for AI, accelerated by NVIDIA RTX PRO 6000 Blackwell Server Edition and NVIDIA B300 GPUs

FREMONT, Calif.--(BUSINESS WIRE)-- **Penguin Solutions**, Inc. (Nasdaq: **PENG**), the AI factory platform company, today announced the expansion of its **OriginAI® portfolio** to include solutions that address the need for more GPU memory to solve context size and concurrency, and meet low latency demands of enterprise-scale AI inference. Penguin Solutions’ OriginAI inference solutions seamlessly add large memory appliances to NVIDIA RTX PRO 6000 and NVIDIA B300 GPU designs, helping to shorten time to value and mitigate performance limitations of AI inference. Designed to improve key operational metrics such as GPU utilization, deployment velocity, and infrastructure reliability, OriginAI enables organizations to run AI workloads with predictable performance at scale.

Penguin Solutions’ OriginAI Factory Platform delivers optimized performance for AI inference with the expansion of its OriginAI portfolio with solutions that address the need for more GPU memory to solve context size and concurrency, and meet low latency demands of enterprise-scale AI inference.

OriginAI inference solutions are designed leveraging Penguin Solutions 3.3+ billion hours of GPU runtime experience and more than 30 years of expertise delivering advanced memory solutions. OriginAI delivers production-level inference, where memory capacity and availability, not only GPU compute power, affect latency, system throughput, and overall user experience.

“Penguin Solutions operationalizes and optimizes AI inferencing by delivering the performance, scalability, and reliability required to realize fully actionable insight and discovery,” said Phil Pokorny, chief technology officer at Penguin Solutions. “Organizations must understand the factors that impact inference performance—which differ significantly from training—to productize AI and deliver accurate and fast outcomes. Whether it’s for deep research or agentic applications, we optimize infrastructure for real-world workloads and enable organizations to turn AI innovation into measurable business outcomes."

Penguin's MemoryAI™ KV Cache Server Matched with NVIDIA GPUs Optimizes OriginAI Solutions for Scalable AI Inference

Penguin Solutions OriginAI solutions also offer the flexibility to incorporate Penguin's **CXL-based** MemoryAI KV cache server, designed to support customers' KV strategies by expanding KV cache capacity, enabling low-latency, high-concurrency inference and extended context lengths for the most demanding applications. Use of Penguin's MemoryAI KV cache server, which is compatible with the NVIDIA Dynamo framework, provides cost-efficiency and optimal design for the next wave of AI deployment.

OriginAI AI factory solutions also include Penguin Solutions **ICE ClusterWare™ software**, an intelligent management layer that transforms validated hardware into a fully-tuned AI cluster. ICE ClusterWare software delivers health monitoring and auto-remediation, to ensure sustained peak performance at scale. It also enhances data security in multi-tenant environments by isolating workloads and protecting sensitive information.

The OriginAI portfolio offers a range of configurations to address diverse customer needs. NVIDIA RTX PRO 6000-based architecture targets enterprise-class copilots, retrieval-augmented generation (RAG) systems, code assistance, and document summarization, delivering a lower acquisition cost, flexible deployment, and power-efficient performance for mid-sized models. NVIDIA B300-based architecture is designed for enterprise-wide AI platforms, long-context assistants, frontier model hosting, and agentic workloads, providing massive memory bandwidth and future-proof scalability for large, shared services.

Enterprise Inference for Financial Services, Healthcare, and Retail

OriginAI inference architectures help provide the flexibility to scale out and avoid overprovisioning by combining expert infrastructure design with meticulous in-factory builds and on-site deployment. This approach enables enterprises as well as cloud service providers (CSPs) and neoclouds to cost-efficiently deploy infrastructure tailored for use case and inference applications at scale. For example:

- Financial Services: AI-driven applications in financial services, such as fraud detection and algorithmic or high-frequency trading, require ultra-low latency to process transactions in real time, optimize trading opportunities, and ensure security.
- Healthcare: Precision in AI-powered diagnostics, patient monitoring, voice-enabled applications, and real-time medical translations depends on minimal latency to deliver timely and accurate insights, often in life-critical situations.

- Retail: AI-driven personalization, inventory management, and agentic decision-making systems enable real-time customer engagement and operational efficiency, helping businesses stay competitive.

AI is reshaping how organizations achieve efficiency, accuracy, and innovation. Penguin Solutions has delivered solutions that address customers' inference objectives and KV strategies, helping them meet evolving demands and achieve measurable results.

To learn more, explore Penguin Solutions' **OriginAI inference solutions** and/or visit booth #1031 at the NVIDIA GTC AI Conference and Expo March 16-19, 2026, in San Jose, Calif.

MemoryAI, OriginAI, and ICE ClusterWare are trademarks or registered trademarks of Penguin Solutions, Inc. or its affiliates. All other trademarks are the property of their respective owners.

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 **www.penguinsolutions.com**.

PR Contact

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

Source: Penguin Solutions, Inc.