NEWS RELEASE

# Penguin Solutions Expands ClusterWareAI Operating System Software for AI Factories with AI-Powered Operations and Automated Remediation

2026-06-25

New AI Factory Operations Agent, Automated GPU Remediation for Kubernetes Workloads, and Expanded Health Visibility Improve AI Factory Performance and Resilience

FREMONT, Calif.--(BUSINESS WIRE)-- **Penguin Solutions**, Inc. (Nasdaq: **PENG**), the AI Factory Platform Company, today announced its latest ClusterWareAI™ AI Factory Platform Operating System Software. The updated ClusterWareAI software enables AI operators to optimize AI factory performance, improve workload resilience, and simplify operations across an entire AI factory.

Penguin Solutions latest ClusterWareAI AI Factory Platform Operating System software enables AI operators to optimize AI factory performance, improve workload resilience, and simplify operations across the entire AI factory. Pictured is the new AI Factory Operations Agent, which provides administrators with a conversational interface for gaining insights into GPU cluster performance using natural language queries.

As enterprises, sovereign AI initiatives, and neocloud providers scale infrastructure to support AI inference at scale, maintaining performance, resilience, and service-level objectives becomes increasingly complex. The latest ClusterWareAI release from Penguin Solutions addresses these challenges with new capabilities that enhance hardware-level visibility and helps maintain peak performance across GPU clusters running inference workloads.

“At scale, AI infrastructure demands a new level of operational intelligence,” said Ian Colle, SVP and Chief Product Officer at Penguin Solutions. “The ClusterWareAI AI Factory Platform Operating System Software from Penguin Solutions provides the unified operational control plane that helps transform compute, memory, networking, storage, and software into an AI factory. This release advances our vision of intelligent, self-managing AI

infrastructure through AI-driven operations, automated remediation, and deep infrastructure awareness."

ClusterWareAI software is the operating system for AI factories, providing a hardware-agnostic control plane that unifies deployment, observability, automation, governance, and performance optimization across training, inference, and Agentic AI environments. This gives operators a single view to manage the entire AI infrastructure lifecycle. By combining deep infrastructure telemetry, operational intelligence, and AI-assisted workflows, ClusterWareAI software enables organizations to accelerate AI deployment, improve resilience, scale operations efficiently, and reduce total cost of ownership.

The new AI Factory Operations Agent provides administrators a conversational interface for gaining insight into GPU cluster performance using natural language queries. The agent accelerates root cause analysis, streamlines troubleshooting, and reduces reliance on specialized expertise, enabling faster issue resolution and improved operational efficiency. It is the first in a planned family of AI-powered agents designed to simplify cluster operations and increase administrator productivity.

This release also extends automated remediation capabilities to Kubernetes®-based inference environments and expands its built-in hardware-level monitoring, ensuring that only GPUs operating at optimum performance are available in inference worker pools. Leveraging precise bare-metal telemetry, ClusterWareAI software delivers real-time visibility into cluster health and proactively detects and resolves hardware degradations before they impact application performance. The timely identification of "fail-slow" conditions, where components degrade without fully failing, helps maintain consistent performance and maximize GPU utilization. Together, these advancements reflect a critical shift in how operators manage AI at scale.

As organizations scale generative AI models from model development and training to production inference and Agentic AI, maximizing infrastructure availability and GPU utilization becomes critical to achieving business value and return on investment. By automating diagnostics, remediation, and performance optimization, ClusterWareAI software helps organizations maximize GPU utilization, reduce operational overhead, and operate AI factories more efficiently at scale.

Learn more about the general availability of this latest release of Penguin Solutions **ClusterWareAI AI Factory Platform Operating System software**.



## About Penguin Solutions

Penguin Solutions is a leading provider of memory and AI infrastructure, powering the AI factories of the future for enterprises, sovereign AI initiatives, and neocloud providers.

Built on decades of engineering expertise at the intersection of memory and AI/HPC infrastructure, we bring together differentiated infrastructure software, advanced memory, compute systems, end-to-end services, and industry-leading partner solutions in a full-stack AI factory platform designed to help customers deploy and scale AI workloads with speed and precision.

Headquartered in Silicon Valley, California, we operate globally through our network of R&D, manufacturing, and sales locations. Learn more at **PenguinSolutions.com**.

## PR Contact

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

Source: Penguin Solutions, Inc.