NEWS RELEASE

Penguin Solutions Expands ComputeAI Portfolio with NVIDIA Vera Rubin NVL72 Rack-Scale Systems

2026-11-17

Pre-validated, liquid-cooled rack-scale system simplifies deployment, reduces integration complexity and accelerates time to value for next-generation AI workloads.

FREMONT, Calif.--(BUSINESS WIRE)-- **Penguin Solutions, Inc.** (Nasdaq: **PENG**), the AI Factory Platform Company, today announced the availability of ComputeAI™ NML72-VR200, a pre-validated, liquid-cooled rack-scale platform built on the NVIDIA Vera Rubin NVL72 architecture. The system provides a deployment-ready foundation for advanced AI workloads, including agentic AI, inference-time scaling, and gigascale model development, supporting AI cloud providers, enterprises, and sovereign AI initiatives.

Penguin Solutions Expands ComputeAI Portfolio

As AI workloads grow in scale and complexity, particularly agentic AI systems that execute multi-step tasks and reason across large context windows, organizations require greater compute density, memory capacity, networking performance, and power and cooling capabilities. Building infrastructure to meet these requirements from smaller systems increases integration complexity, extends deployment timelines, and delays time to production.

Penguin Solutions' ComputeAI NML72-VR200 addresses these challenges with a fully integrated, pre-validated rack-scale system featuring 72 NVIDIA Rubin GPUs, 36 NVIDIA Vera CPUs, sixth-generation NVIDIA NVLink, NVIDIA ConnectX-9 networking, NVIDIA BlueField-4 DPUs, high-capacity memory and storage, rack-level power, and direct liquid cooling. This integrated architecture simplifies deployment, reduces integration risk, and provides a repeatable foundation for scaling AI services and proprietary model development.

“As AI workloads grow more complex, organizations need infrastructure that can move from deployment to

production quickly without compromising performance or scalability," said Ranjeet Sudan, Vice President of Product Management at Penguin Solutions. "With the ComputeAI NML72-VR200, we combine advanced NVIDIA technology with Penguin Solutions' deep infrastructure expertise to deliver a pre-validated rack-scale system that simplifies deployment and enables customers to scale next-generation AI workloads with greater speed and confidence."

Designed to support demanding AI workloads, including high-concurrency agentic applications, long-context reasoning, and large-scale model training, ComputeAI NML72-VR200 expands Penguin Solutions' ComputeAI portfolio of advanced compute systems optimized for AI data centers. As part of the company's Full-Stack AI Factory Platform, the solution combines advanced compute technologies with Penguin Solutions' infrastructure expertise to help organizations simplify deployment, scale operations and maximize the value of their AI investments.

ComputeAI NML72-VR200 is available now. To learn more, **click here**.

Penguin Solutions and ComputeAI 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

Penguin Solutions is the AI Factory Platform Company. We design, build, and manage next-generation data centers for enterprises, sovereign AI initiatives, and AI cloud providers.

With deep design expertise at the intersection of data center AI infrastructure and memory solutions, our Full-Stack AI Factory Platform combines differentiated infrastructure software, advanced memory, compute systems, end-to-end services, and industry-leading partner technologies to help customers accelerate deployment, optimize token economics, and maximize the return on their AI investments.

Learn more at **PenguinSolutions.com**.

Media Relations

Lerin O'Neill
Sr. Director, Corporate Communications & Public Relations
Penguin Solutions
pr@penguinsolutions.com

Source: Penguin Solutions, Inc.