NEWS RELEASE

Penguin Solutions Becomes an NVIDIA AI Factory Specialized Partner

2026-06-23

Recognition validates Penguin's expertise designing, building, deploying, and managing full-stack, NVIDIA-based AI factory infrastructure for enterprises, sovereign AI initiatives, and neocloud providers

FREMONT, Calif.--(BUSINESS WIRE)-- **Penguin Solutions**, Inc. (Nasdaq: **PENG**), the AI Factory Platform Company, today announced it has become an NVIDIA AI Factory Specialized Partner, joining a select group of NVIDIA Partner Network (NPN) solution providers. Penguin achieved this invitation-only NVIDIA AI Factory specialization by completing NVIDIA's comprehensive training, maintaining the relevant competencies, meeting solution requirements, and bringing proven experience in designing, building, deploying, and operating full-stack, NVIDIA-based AI factory infrastructure for enterprise and hyperscale customers.

Penguin Solutions has become an NVIDIA AI Factory Specialized Partner, joining a select group of NVIDIA Partner Network (NPN) solution providers. Penguin brings proven experience in designing, building, deploying, and operating full-stack, NVIDIA-based AI factory infrastructure for enterprise-scale AI factories that power AI training, inference and agentic AI workloads. Image courtesy of NVIDIA.

The NVIDIA AI Factory Specialized Partner designation recognizes Penguin Solutions' expertise and capabilities in delivering enterprise-scale AI inferencing and training solutions that enable organizations' agentic and AI workloads.

Penguin helps customers accelerate AI time-to-value and achieve superior token economics by delivering and optimizing NVIDIA-accelerated AI factories. Penguin's Full-Stack AI Factory Platform considers every layer of the AI environment, including the underlying power required to produce tokens; the accelerated NVIDIA processors that enable efficient computation; the specialized server, networking and storage infrastructure needed to orchestrate thousands of GPUs; and the various models and applications used to innovate and transform our world.

"For over a decade, we have worked closely with NVIDIA in delivering and operating AI factories and GPU-based solutions for leading hyperscalers, enterprises, sovereign AI, and neocloud providers," said Kash Shaikh, President and CEO of Penguin Solutions. “This designation validates Penguin's deep capabilities to design, build, deploy, and operate full-stack AI factories at scale. As demand for AI infrastructure continues to accelerate, our customers are realizing meaningful business outcomes from their AI investments in competitive, fast-paced industries where success is increasingly determined by their ability to operationalize AI at scale with superior token economics.”

As organizations look to accelerate their AI initiatives, demand for AI infrastructure that can support large-scale inference and emerging agentic AI workloads continues to grow. They require AI factory platform solutions that deliver performance, scalability, efficiency, and operational reliability. For example, **Deepgram recently worked with Penguin** to deploy a production-ready AI inference platform supporting large-scale Speech-to-Text, Text-to-Speech, and Voice Agent applications and workloads.

Penguin also collaborated with NVIDIA and SK Telecom in the creation of the **Haein** AI Factory, one of Korea’s largest, most powerful, and award-winning GPU-as-a-Service clusters, defining a new model for collaboration, execution, and sovereign AI strategy.

Penguin's Full-Stack AI Factory Platform brings together innovative products including ClusterWareAI™, MemoryAI™, ComputeAI™, OriginAI®, and End-to-end Services to help customers design, deploy, and manage AI factory environments to accelerate AI adoption while simplifying deployment and reducing operational complexity.

The NVIDIA AI Factory Specialization designation further validates Penguin’s strategy to become a leading Full Stack AI Factory Platform company for enterprises, sovereign AI initiatives, and neocloud providers.

Learn more about the **Penguin Solutions / NVIDIA partnership and AI solutions**.

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.