PENGUIN SOLUTIONS | OriginAI®

## Overview

Pre-configured, validated AI infrastructure architectures backed by intelligent cluster management software and expert services


## Key Features

- Preconfigured, validated, and tested AI infrastructure architectures
- Scalable designs from hundreds to more than 16,000 GPU clusters
- Tailored approach to help customers anywhere in the design, build, deploy or manage phase of their AI journey
- Flexible architectures with multiple GPU, networking and storage architecture options to best fit your workload and deployment requirements
- Scyld ClusterWare® for automated cluster management and predictive cluster failure analysis with automated ticketing and resolution
- Full suite of support, and professional and managed services

## Benefits

- Reduce the complexity and risk involved in AI implementation by leveraging Penguin's experience from deploying over 85,000 GPUs and spending over 2 billion hours managing large scale AI clusters
- Accelerate time to value with a comprehensive, scalable AI infrastructure solution by leveraging proven designs and architectures, and extensive expertise from Penguin Solutions
- Optimize performance and ROI of initial workloads at deployment following factory integration, validation, and testing services
- Realize predictable infrastructure performance and availability while meeting data center environmental requirements
- Achieve up to 95%* infrastructure availability with Penguin's Scyld ClusterWare and managed services

* OriginAI solutions offer up to 95% infrastructure availability when combined with Penguin Solutions managed services and ClusterWare software.

# Deploying & Managing AI Infrastructure at Scale

By 2026, over 80% of enterprises will use generative AI in production environments—a sharp increase from less than 5% in 2023.[1] As AI reshapes industries like energy, financial services, and cloud service providers (CSPs), organizations are investing heavily in building and integrating AI solutions as a core enterprise capability and strategy.

However, despite the increasing reliance on AI to stay competitive—driving innovation, boosting efficiency, and delivering faster, data-driven insights at speed—more than 90% of organizations will struggle to find the talent they need to implement their AI infrastructure at scale.[2]

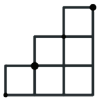
Penguin Solutions OriginAI is a portfolio of pre-configured, validated, and tested AI infrastructure architectures backed by Penguin's intelligent, intuitive cluster management software and expert services, offering organizations a powerful solution to rapidly deploy AI infrastructure at scale, whether utilizing hundreds or tens of thousands of GPUs.

# Simplify AI Factory Deployment and Management with Penguin Solutions OriginAI

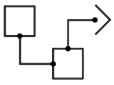
Penguin Solutions OriginAI enables organizations to efficiently scale AI infrastructure for model development, training, and tuning, and for generative inference, helping meet the demands of highly complex AI workloads.

Matched with Penguin Solutions' proven design, build, deploy, and manage approach, OriginAI provides a path for organizations to build AI factory infrastructure, accelerate time to value, and optimize performance.

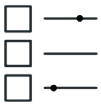
- **Design:** Penguin Solutions collaborates closely with customers to design AI infrastructures tailored to specific workloads, leveraging industry best practices and advanced technologies to meet the precise needs of each environment.



- **Build:** Penguin Solutions' build process ensures AI infrastructure is fully assembled, tested, and validated—including in-factory burn-in testing and performance validation—to minimize setup time and optimize reliability upon delivery, speeding deployment and user productivity.



- **Deploy:** Penguin Solutions' deployment services ensure that AI infrastructure is operational and optimized quickly, providing full support to maximize performance and time to value.



- **Manage:** Penguin Solutions' management services provide ongoing support to ensure that AI infrastructure remains fully operational, scalable, and optimized so that organizations can focus on AI innovation and achieving business goals.

# Proven, Pre-Defined AI Infrastructure Architectures

OriginAI is an AI factory infrastructure solution built on proven, pre-defined AI architectures that scale from hundreds to more than 16,000 GPU clusters. OriginAI integrates validated best-fit technologies with Penguin's intelligent, intuitive cluster management software and expert services for designing, building, deploying, and managing AI infrastructure at scale.

As a hardware-agnostic solution, OriginAI supports multiple GPU technologies—including AMD and NVIDIA—and a wide range of storage and networking options to meet customers' specific workload requirements and business objectives.

# Key Features and Benefits

## Proven Expertise with Large-Scale AI Factories

Penguin Solutions is a Dell Technologies Authorized Partner, an NVIDIA-certified Elite Solution Provider for Networking, DGX AI Compute Systems, and Compute, and an NVIDIA DGX Managed Service Provider. Since 2017, Penguin Solutions has successfully designed, built, deployed, and managed AI infrastructures totaling more than 85,000 GPUs and over 2 billion hours of GPU runtime across a wide range of industries, including financial services, energy, cloud service providers, life sciences, government, and higher education.

## Reduced Complexity and Lower Risk for AI Implementation

Penguin Solutions helps enterprises streamline and safeguard AI implementations by utilizing proven AI cluster architectures for optimal performance and ROI. OriginAI system designs and architectures are planned for long-term operation, security, and scalability. In addition, the proven assembly and integration methods used in Penguin Solutions' factories are specifically designed to optimize the operational efficiency of AI and HPC infrastructures.

## Predictable Performance and Optimal ROI at Deployment

OriginAI leverages Penguin Solutions' factory testing and simulation environment to confirm production readiness, validate AI cluster performance, and deliver fully staged, operational systems with pre-loaded, burnt-in images. Starting with a validated architecture that is tested first in the factory helps ensure performance and ROI from day one. In addition, Penguin's extensive service offerings give customers access to the HPC and AI expertise they need to help manage deployments.

## Maximum Infrastructure Availability

Penguin Solutions Scyld ClusterWare® intelligent management solution provides rapid provisioning, automated cluster management and deep health monitoring capabilities. Coupled with Penguin's expert managed services and spares management services, customers achieve up to 95% availability* of computing resources and maximize the return on their AI infrastructure investment.

## Penguin Solutions: Proven AI Infrastructure Expertise

Penguin Solutions helps organizations expand their AI capabilities with a comprehensive solution to design, build, deploy, and manage AI infrastructure at scale, empowering them to harness AI's full potential with confidence. With more than 25 years of HPC experience and more than 85,000 GPUs deployed since 2017, Penguin Solutions and its OriginAI architectures reduce the complexity of designing and deploying AI factory infrastructure allowing teams to stay focused on core business objectives. With OriginAI, Penguin Solutions provides a streamlined path to success at every stage of the AI journey.

## Penguin AI Infrastructure Expertise

- Building and managing AI factories since 2017
- Over two billion hours of GPU runtime
- Up to 95% infrastructure availability*
- More than 85K GPUs deployed and under management

* OriginAI solutions offer up to 95% infrastructure availability when combined with Penguin Solutions Managed Services and Scyld ClusterWare software.

## Contact Us

Contact Penguin Solutions today to explore how OriginAI® can accelerate your AI factory design and deployment.

## Sources

1. Gartner, "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026" (https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026).

2. IDC, "IT Skills Shortage Expected to Impact Nine out of Ten Organizations by 2026" (https://www.idc.com/getdoc.jsp?containerId=prUS52128824).

PENGUIN™ SOLUTIONS