Penguin Solutions OriginAI®
Solution Brief

PENGUIN™
SOLUTIONS

Deploying & Managing AI Infrastructure at Scale

More than 80% of enterprises will be using Generative AI in production environments by 2026, up from less than 5% in 2023. GenAI will be ubiquitous in 3 years.[1] As AI continues to transform industries, businesses are deploying more resources towards building and integrating AI solutions.

Yet, most companies lack the expertise to build AI infrastructure at scale. 75% of companies are struggling to hire the AI talent they need.[2] There's a big knowledge gap, and upskilling is slow going. And it's not getting any better. By 2026, more than 90% of organizations will suffer from an AI and tech talent crisis, amounting to $5.5 trillion in losses from product delays, loss of business, and an inability to compete effectively.[3]

Even for enterprises experienced with large-scale systems, managing AI infrastructure can be exceptionally resource-intensive, diverting key personnel from other critical projects. **AI infrastructure is incredibly expensive, complex, and challenging to build, deploy, and manage, requiring a significantly different approach compared to traditional enterprise IT infrastructure. Most AI projects fail.** The Harvard Business Review estimates that the failure rate is as high as 80% — about twice the rate of corporate IT project failures.[4]

AI solutions are evolving constantly and the need for reliable, proven AI infrastructure to exploit market opportunities is greater than ever.

Penguin Solutions OriginAI® is uniquely designed for customers who are looking to rapidly deploy AI infrastructure at scale — utilizing hundreds or tens of thousands of GPUs. Whether you're beginning your AI journey or you've hit a roadblock, OriginAI is the perfect solution.

Solution at a Glance

Features

- Preconfigured, validated, and tested AI infrastructure architectures
- Scalable designs from hundreds to more than 16,000 GPU clusters
- Scyld ClusterWare® for automated cluster management and predictive analysis for cluster failure with automated ticketing and resolution
- Full suite of professional and managed services

Benefits

- Reduce the complexity and risk of AI implementation by working with a partner who has deployed and managed AI clusters totaling over 75,000 GPUs
- Accelerate time to market and time to value with a comprehensive, scalable AI infrastructure solution using proven designs and architectures
- Realize predictable performance and optimal ROI at deployment with factory integration, validation and testing services
- Maximize GPU availability and utilization with advanced software tools and managed services
- Focus on AI production and deployment instead of worrying about AI infrastructure

Penguin Solutions OriginAI

Penguin Solutions OriginAI provides a scalable AI infrastructure solution for foundational model training, model tuning and generative inference workloads. OriginAI enables organizations to rapidly deploy AI infrastructure at scale and avoid the pitfalls of stablishing, expanding, or managing an AI factory.

Built upon proven, pre-defined AI infrastructure architectures, OriginAI is an AI factory infrastructure solution that simplifies deployment and implementation and maximizes GPU availability and utilization for predictable performance and optimal ROI. OriginAI integrates validated technologies backed by Penguin's intelligent, intuitive cluster management software and expert services for designing, building, deploying, and managing AI infrastructure.

Key Features and Benefits

Reducing the Complexity and Risk of AI Implementation

Penguin Solutions is an NVIDIA-certified Elite OEM and DGX AI Compute Systems Solution Provider, and a DGX-Ready Managed Services partner. We have successfully built, deployed, and managed AI infrastructures since 2017 with more than 75,000 GPUs deployed and managed.

Our unmatched AI experience powers OriginAI, based on the proven architectures and methodologies that we've developed for numerous customers across multiple industries including Financial Services, Energy, Life Sciences, Government and Higher Education. When you put OriginAI to work for you, you get an AI architecture that's ready-to-go based on proven models to accelerate your time to market.

Delivering Predictable Performance and Optimal ROI at Deployment

OriginAI solutions leverage our factory testing and simulation environment to help confirm production readiness and validate AI cluster performance. Starting with a validated architecture and then testing it in the factory helps to ensure performance and ROI at deployment.

Plus, you get the benefit of some of the most experienced HPC and AI experts in the industry to help you manage your deployment, along with Penguin Solutions Scyld ClusterWare for rapid provisioning, automated cluster management, and high availability.

Maximizing GPU Availability and Utilization

GPU and cluster management is complex and the more GPUs you run, the more likely something will fail. If it happens in the middle of a job, it's costing you time and money.

Scyld ClusterWare, combined with Penguin's managed and support services, ensures you get the maximum availability of nodes while driving higher GPU cluster performance. Using predictive intelligence, OriginAI maximizes cluster utilization and health alongside AI node availability. Automated ticketing and resolutions stop node failures before they occur with no human involvement — keeping your AI jobs running smoothly.

OriginAI maintains a long-run availability of 95% or greater with a turnkey AI infrastructure combining hardware, software, and management.

Scaling AI in the Real-World

Designing, deploying, and operating AI factories is an incredibly complex endeavor, requiring specialized skills, tools, and knowledge. Mistakes at any stage are common and can cost you time and money. Companies spend millions of dollars to develop their own AI infrastructure only to find it doesn't perform up to specifications. That's a big risk — and an expensive one.

OriginAI's pre-defined AI infrastructure architectures, backed by intelligent, intuitive software and Penguin's expert services, help you quickly deploy and maximize your AI investments.

Use cases include:

- **Enterprise businesses:** Rapidly deploy a large-scale GPU cluster for AI-powered research and growth initiatives.
- **Managed service providers (MSPs):** Offer OriginAI as a managed service, allowing customers to quickly provision and scale high-performance, reliable AI resources on demand.
- **Financial services:** Rapidly deploy and manage an AI infrastructure with high availability and maximized GPU utilization for applications like fraud detection and algorithmic trading.
- **Energy businesses:** Powerful, scalable, and customizable computing that addresses high-stakes exploration, simulation, and management needs.
- **Life sciences:** Accelerate AI applications like medical image analysis, drug discovery, and personalized medicine while ensuring regulatory compliance.
- **Government & academic research:** Leverage OriginAI's pre-defined architectures and expert services to manage a high-performance AI infrastructure for research projects.

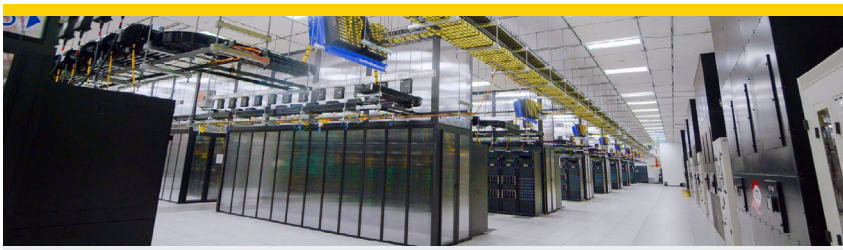
Penguin Solutions: Proven AI Infrastructure Experience

With proven experience in real-world deployments, Penguin Solutions has deep expertise in designing, building, deploying, and managing some of the world's largest AI infrastructures. With more than 25 years of HPC experience and more than 75,000 GPUs deployed since 2017, Penguin Solutions is a trusted partner for some of the largest organizations in the world.

Meta's AI Research SuperCluster

Meta (Facebook) partnered with Penguin Solutions and NVIDIA for the co-development of one of the world's largest AI factories. When built, the Meta Research SuperCluster was the most advanced NVIDIA solution on the market, incorporating more than 40,000 links to connected GPUs, working together on a single workload, including:

- 16,000 NVIDIA A100 GPUs
- 500 petabytes of storage
- 200 Gb/s HDR InfiniBand per GPU
- 5 exaFLOPS of mixed precision compute
- Penguin Solutions Managed Services



"We added a 10K GPU cluster while running multiple research projects. We now have a template for building large GPU clusters that is repeatable and reliable."

Meta

In addition to Meta, Penguin Solutions is the trusted strategic partner for AI and HPC solutions for organizations like, Sandia Labs, the U.S. Navy, and Georgia Tech.

Contact Penguin Solutions to Accelerate Your AI Factory Deployment With OriginAI

SOURCES

1. *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. *https://www.a.team/mission/ai-skills-gap*
3. *https://www.idc.com/getdoc.jsp?containerId=prUS52128824*
4. *https://hbr.org/2023/11/keep-your-ai-projects-on-track*

PENGUIN SOLUTIONS