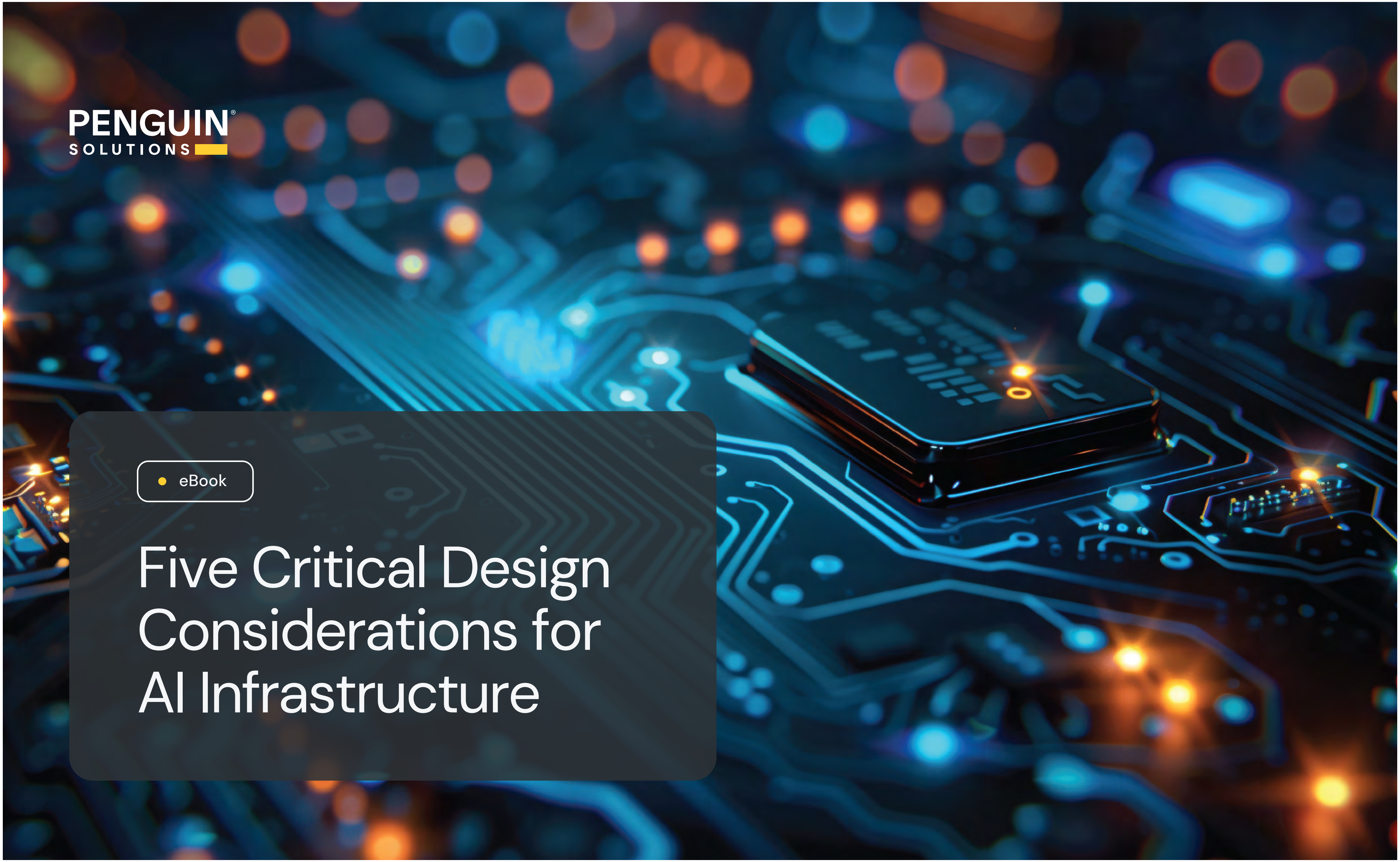


PENGUIN® SOLUTIONS

● eBook

Five Critical Design Considerations for AI Infrastructure

Introduction

The global artificial intelligence (AI) market is on a rapid rise. In 2025, it's **valued at $243 billion**, and by 2030, it's expected to more than triple, reaching $826 billion as companies leverage the practical applications of machine learning (ML), natural language processing (NLP), and generative AI (GenAI). Enterprises are increasingly turning to AI to scale operations, automate processes, and achieve transformative outcomes.

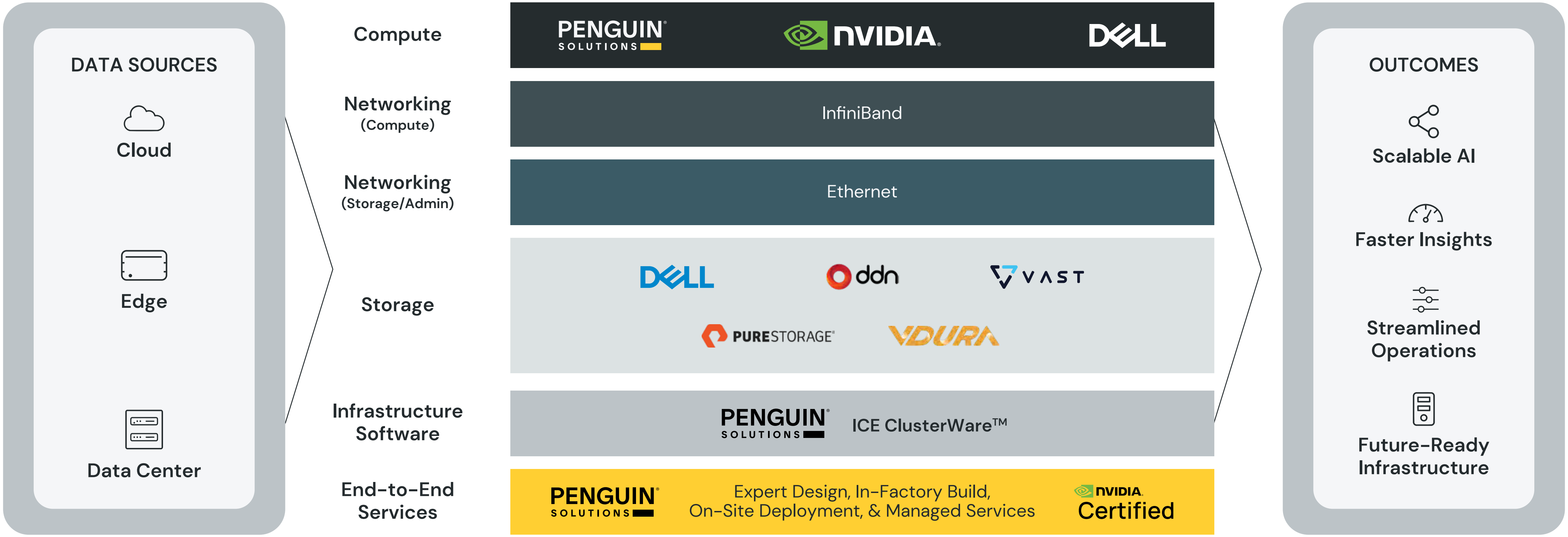
The rapid adoption of AI infrastructure sets the stage for the next evolution of enterprise operations: the AI factory. This eBook defines the fundamentals of AI factory development, explores key development phases, addresses common challenges, and outlines five critical design considerations for IT decision makers.

© 2025 Penguin Solution

Table of Contents

Defining the AI Factory and Its Requirements

The combination of accelerated computing, storage, networking, and software into a high-performance computing asset is generally referred to as an AI factory. This specialized, infrastructure collects, curates, and analyzes internal and external data to generate actionable insights using advanced processing. These factories depend on continually updated AI models to process vast amounts of incoming data and manage the numerous possible outcomes generated by this data.



DATA SOURCES

Cloud

Edge

Data Center

Compute: PENGUIN SOLUTIONS · NVIDIA · DELL

Networking (Compute): InfiniBand

Networking (Storage/Admin): Ethernet

Storage: DELL · ddn · VAST · PURESTORAGE · VDURA

Infrastructure Software: PENGUIN SOLUTIONS ICE ClusterWare™

End-to-End Services: PENGUIN SOLUTIONS · Expert Design, In-Factory Build, On-Site Deployment, & Managed Services · NVIDIA Certified

OUTCOMES

Scalable AI

Faster Insights

Streamlined Operations

Future-Ready Infrastructure

© 2025 Penguin Solution

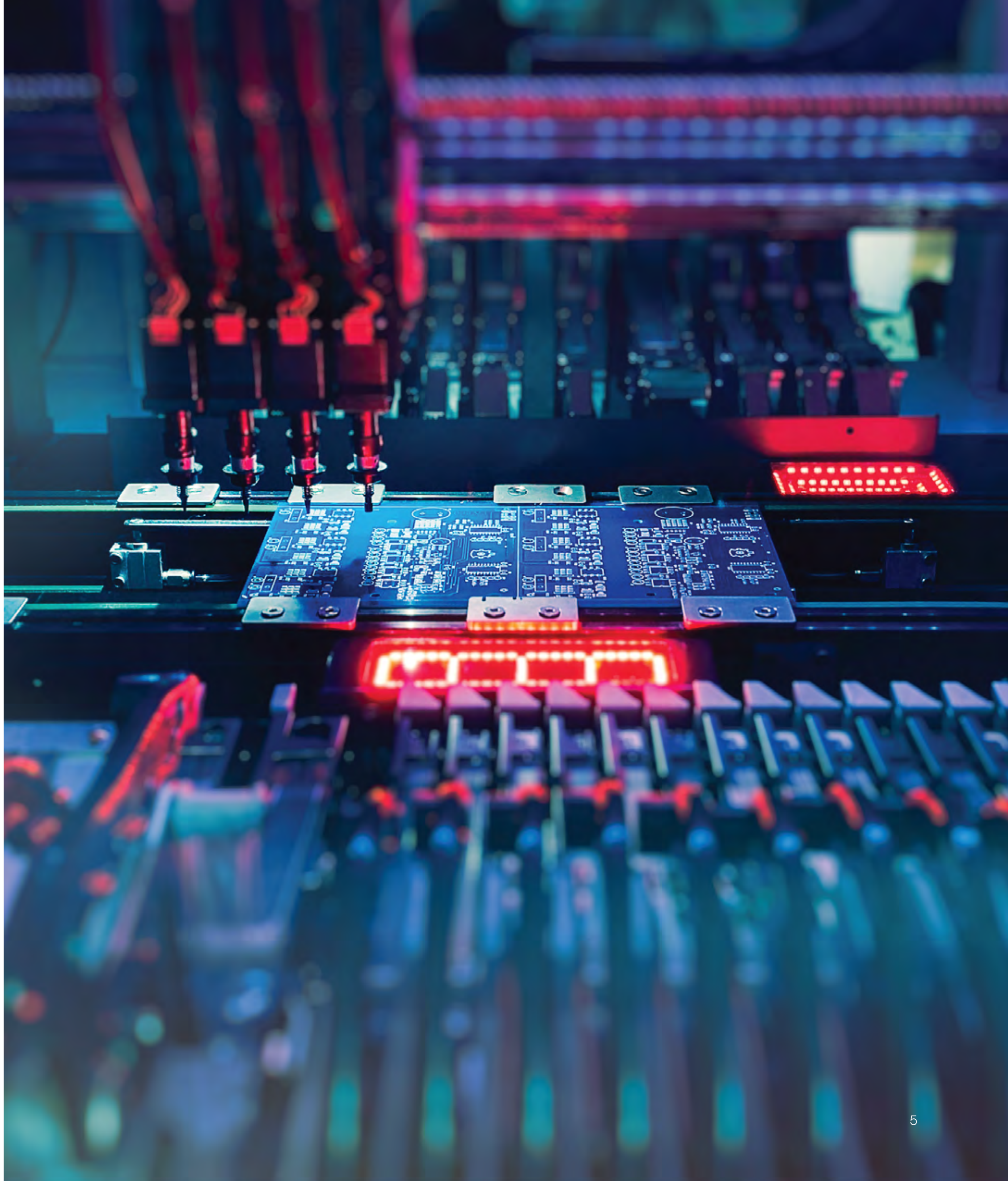
What is an AI Factory?

AI factories function much like traditional manufacturing facilities. Where a production line might automate tasks like assembling car doors, an AI factory automates the collection and verification of data sources where the products are AI-driven outcomes.

Like manufacturing factories, AI factories are component-based and built for scalability. Traditional plants ramp up production by adding machinery or lines; AI factories rely on expandable GPU-based servers and racks, edge computing frameworks, networking, and scalable storage solutions. In both factory types, regular reviews help refine processes and streamline operations.

Another key similarity is their purpose-built nature—both are designed to deliver higher performance and stability for specific workloads. For example, while networking and data processing architectures remain consistent across commercial solutions, the AI and ML tools deployed within these architectures define their specific functions.

Building a robust AI factory requires more than just advanced algorithms—it depends on essential infrastructure. Importantly, infrastructure should be scalable: starting with smaller, flexible systems that can grow as demand increases, or using cloud-based services to offload non-time-sensitive tasks.



© 2025 Penguin Solution

Phases of Deployment of an AI Factory

AI factory deployment happens in two broad phases. The first focuses on standing up the requisite accelerated computing infrastructure to run planned workloads, and the second ensures continous operation, management, and resource optimization.

Phase 1

Phase 1 is made up of three key steps: design, build, and deploy.

Design: Enterprises assess what they have, what they need, and how they want their AI factory to integrate with current processes.

Build: Businesses acquire the necessary hardware, software, and other components, then assemble the AI factory framework.

Deploy: The AI factory goes live. Deployments typically begin on a small scale, allowing IT teams to monitor factories for any issues or unexpected outcomes.

Keep it simple: In phase 1, keep two goals in mind—staying on time and within budget.

Phase 2

Phase 2 also involves three steps: manage, expand, and evolve.

Manage: Once AI factories are up and running, enterprises must shift to management mode. In practice, this means regularly reviewing power usage, total cost of ownership (TCO), and AI outputs to ensure they meet expected values.

Expand: Once factories have passed initial evaluations and demonstrated consistent performance, companies can scale up resources to tackle new projects or handle larger data volumes.

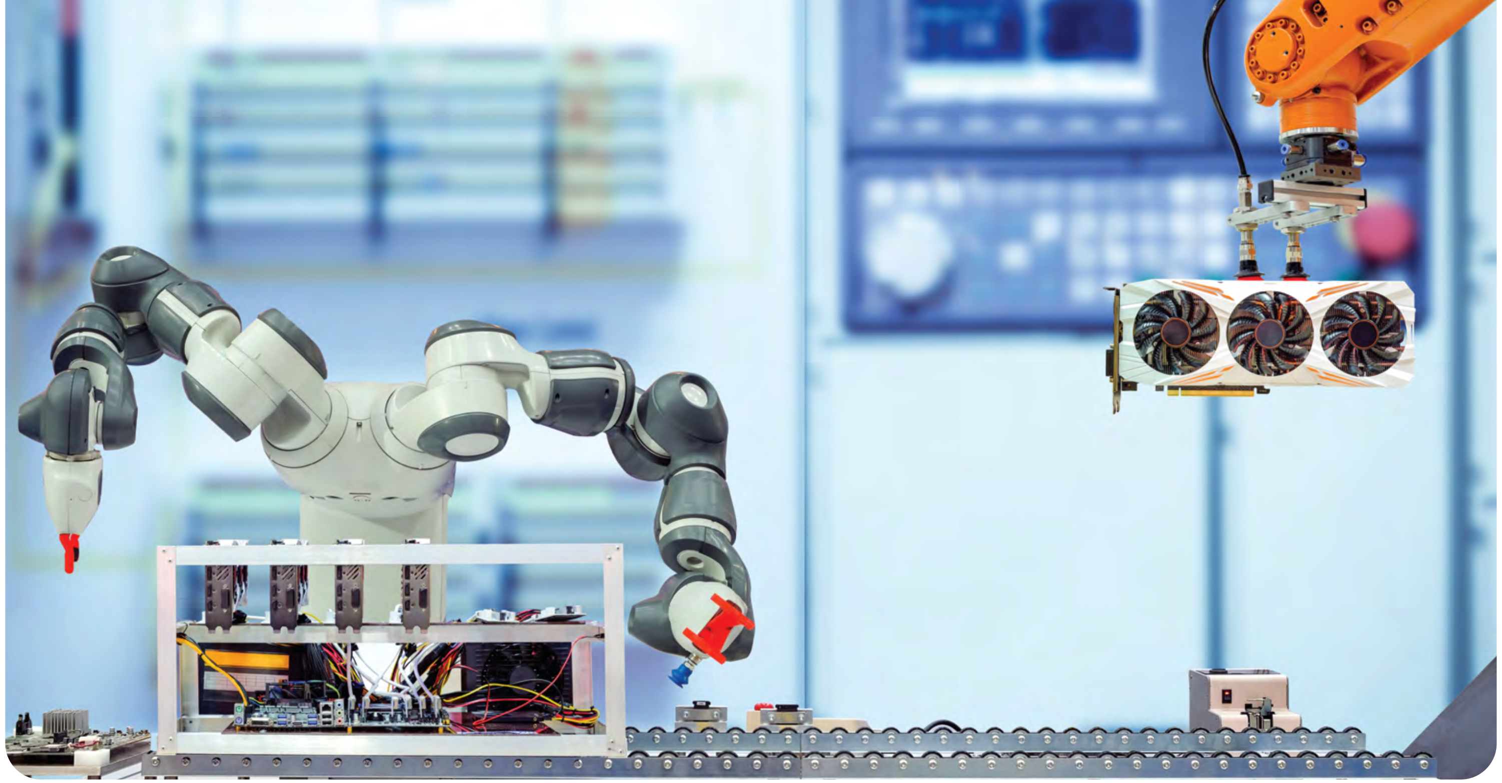
Evolve: AI is a constantly evolving discipline. Consider the rapid rise of GenAI. Cutting edge just two years ago, it's now embedded into everything from search results to chatbots to content creation.

Stay the course: In phase 2, the focus is to maximize the value AI delivers.

© 2025 Penguin Solution

Challenges in Deployment

Building an AI factory comes with challenges. In fact, one in three companies has delayed AI deployments due to skill shortages, budget constraints, and infrastructure issues.



Four common AI challenges include:

1. Complex design

Several factors contribute to AI complexity. First, new architectures are necessary to support evolving workloads, requiring the purchase of AI clusters that are highly sensitive to disruption at scale. AI factories also rely on multiple network and processor types, which can hinder interoperability unless proactively addressed.

Worth noting: **More than 80% of AI projects fail to meet their goals**, in part due to complexity.

2. Intricate build

There are no off-the-shelf options for AI factories. While some new devices feature purpose-built GPUs to handle large-scale workloads or complete tasks autonomously, building an AI factory remains an intricate, iterative process. Enterprises may encounter issues with supply chain delays that limit access to critical components or may struggle to unify deployments without software capable of handling AI processes at scale.

3. Lengthy deployment

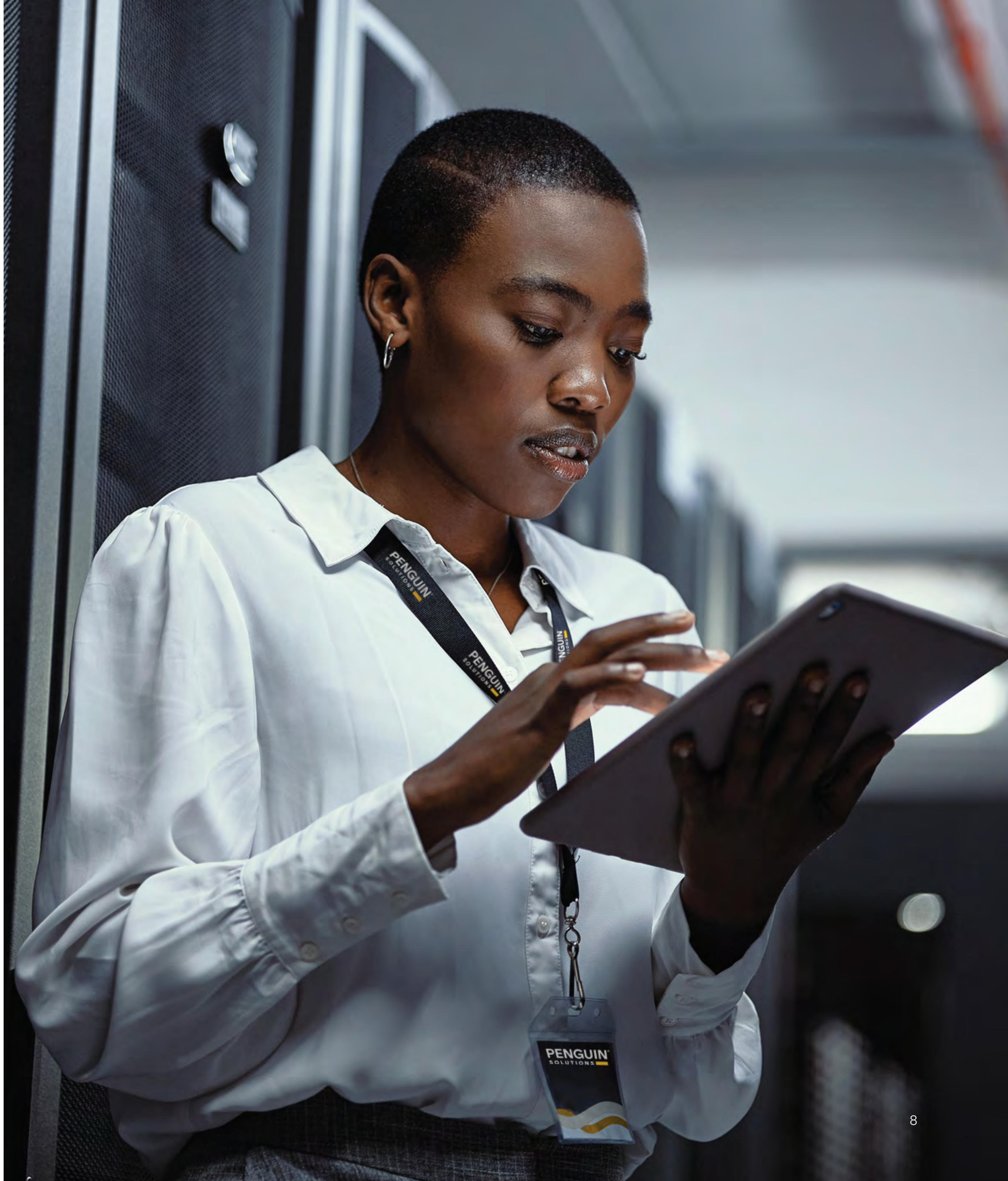
Deploying an AI factory must be a thoughtful process. First, businesses must conduct pre-production performance evaluations and throughput validation. Next, they must create a blueprint that accounts for complex on-site power, cooling, networking, and security integration. Finally, they need tools capable of continuously monitoring production clusters.

4. Precision management

AI factories are highly sensitive to disruption. If the data fed into AI toolsets is not timely or accurate, outputs will suffer. Inaccurate results may require model re-training, which can mean lost time and interrupted workloads.

In addition, AI-specific specialty components have unique failure signatures. Unlike typical systems where power failures signal obvious problems, AI tools may appear to operate within normal parameters, but their outputs may become inaccurate or unreliable.

The need for precision management can also lead to long-tail concerns. For example, if enterprises rely on AI factories to predict market trends and drive decision-making, sudden performance issues could result in significant financial losses.

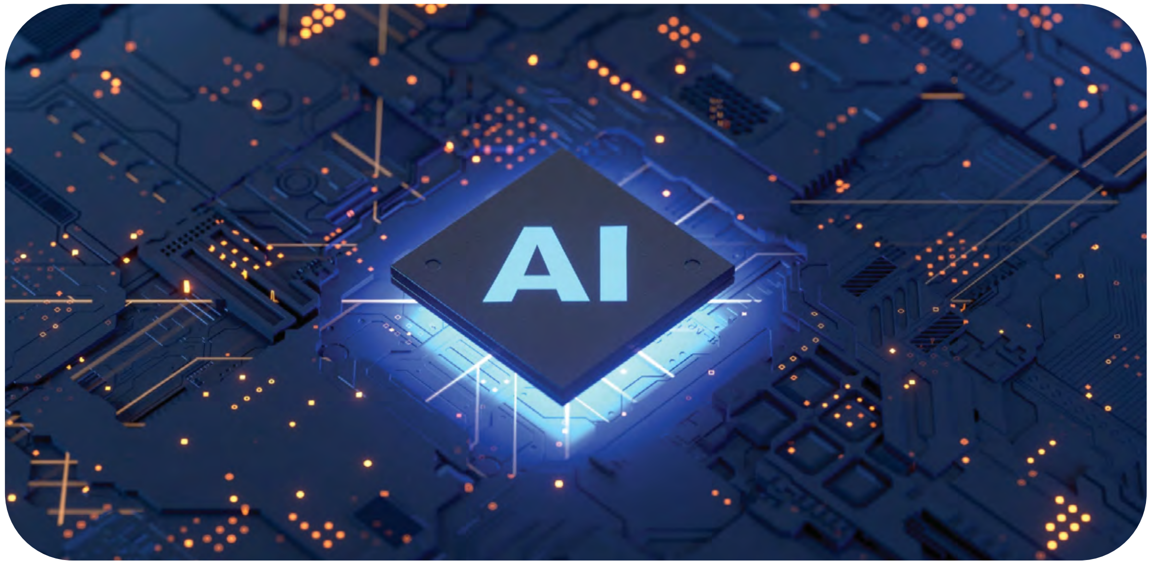


© 2025 Penguin Solution

Why Design Is Critical

Unlike traditional IT deployments, AI factories are naturally iterative. AI algorithms continuously learn by rapidly analyzing billions of data points, allowing them to develop new associations and improve outputs. This creates a cycle—more data leads to deeper analysis, which produces better outputs that are more effective at handling data.

Because of this, design is critical. AI factories don't generate value by chance—it's the result of intentional design planning that aligns infrastructure with desired productivity outcomes. Two of the most important design factors are AI node availability and AI cluster performance.



AI nodes

These processing units receive, process, and transmit information to other nodes. Thousands, millions, or billions of these nodes together create an AI network, making AI node availability critical. While 100% availability is an ideal, in reality, achieving maximum uptime requires specialized expertise, resilient design, and proactive management. Like any advanced system, maintaining high availability and performance across an AI factory is a complex challenge—and a key factor in sustained success.

AI clusters

Clusters are integrated systems composed of compute nodes, storage, networking, and orchestration software that all work together to complete tasks, automate processes, or analyze data at scale. Their performance depends on multiple factors, including the type of network infrastructure used, the bandwidth available, and the health of connected hardware. For example, servers nearing the end of their lifecycle may begin to experience random performance drops. Proactive monitoring and maintenance can help reduce this risk.

In short, high node availability and reliable cluster performance are both necessary to optimize AI infrastructure value.

Inefficient vs. Efficient Design

Companies can lose more than half the value of their infrastructure investment due to inefficient design.

Consider an AI network that has been optimized for high-speed compute and stable network performance. If the nodes on this network are only sporadically available—frequently failing to fully analyze data or becoming unexpectedly unresponsive—the iterative chain breaks down.

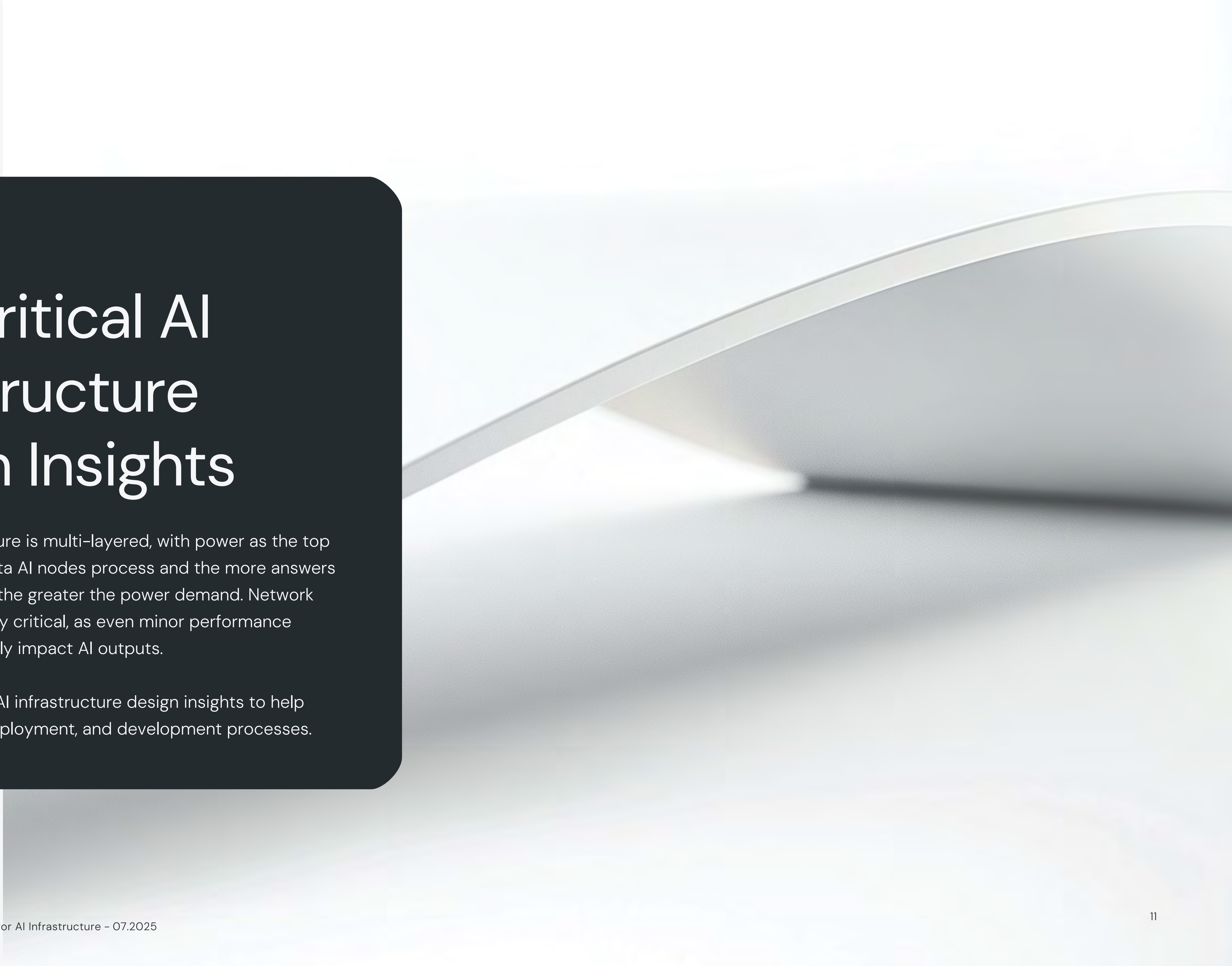
Instead of continually ingesting and analyzing data, the process is interrupted when one or more nodes go dark. This can create bottlenecks in compute performance or network synchronization, which, in turn, reduces ROI.

© 2025 Penguin Solution

Five Critical AI Infrastructure Design Insights

AI factory infrastructure is multi-layered, with power as the top priority. The more data AI nodes process and the more answers AI clusters generate, the greater the power demand. Network connectivity is equally critical, as even minor performance issues can significantly impact AI outputs.

Here are five critical AI infrastructure design insights to help streamline design, deployment, and development processes.



1. Start Your Design with the Data Center

The data center is the foundation of AI infrastructure. While ad hoc or reactive approaches to design may work in the short term, they often lead to complexity and inefficiency over time.

Many IT professionals are familiar with server closets or legacy data centers that have turned into tangled wires and connections, built to meet immediate needs rather than the long-term strategic requirements of business operations.

The rapid adoption of AI presents an opportunity to design data centers that not only support current requirements but can also handle future expansions.

Power dictates everything

Designing an AI factory starts with a simple rule: Power dictates everything. Because compute processes are handled directly by AI nodes and connected across AI clusters, power requirements for intelligent operations are much higher than those of traditional data centers.

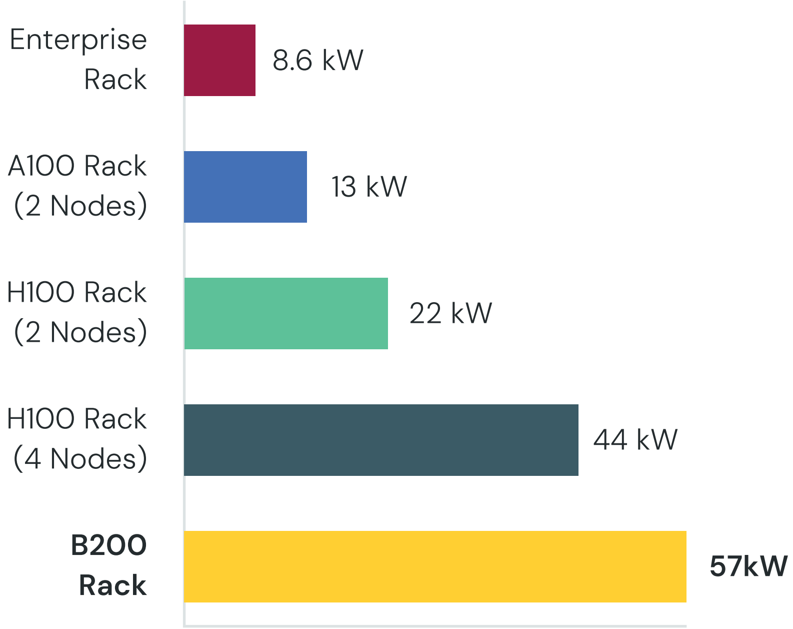
Failing to account for power requirements puts businesses at an immediate disadvantage—while they may be equipped to leverage entry-level AI solutions, the expansion of nodes or clusters can create power problems.

Recent data from Uptime Institute found that 35% of data centers struggle with power limitations due to AI workloads. Ideally, data centers should aim for a density of 2-4 GPU-based nodes per rack, but teams must also recognize that variations in power availability drive configuration diversity. Essentially, the physical power profile—where connections are located and the type and consistency of power provided—will dictate the physical layout of your data center floor.

© 2025 Penguin Solution

Examples of Different Power Environments

Power environments differ based on use case, with AI-enabled accelerator racks far outpacing traditional server solutions.

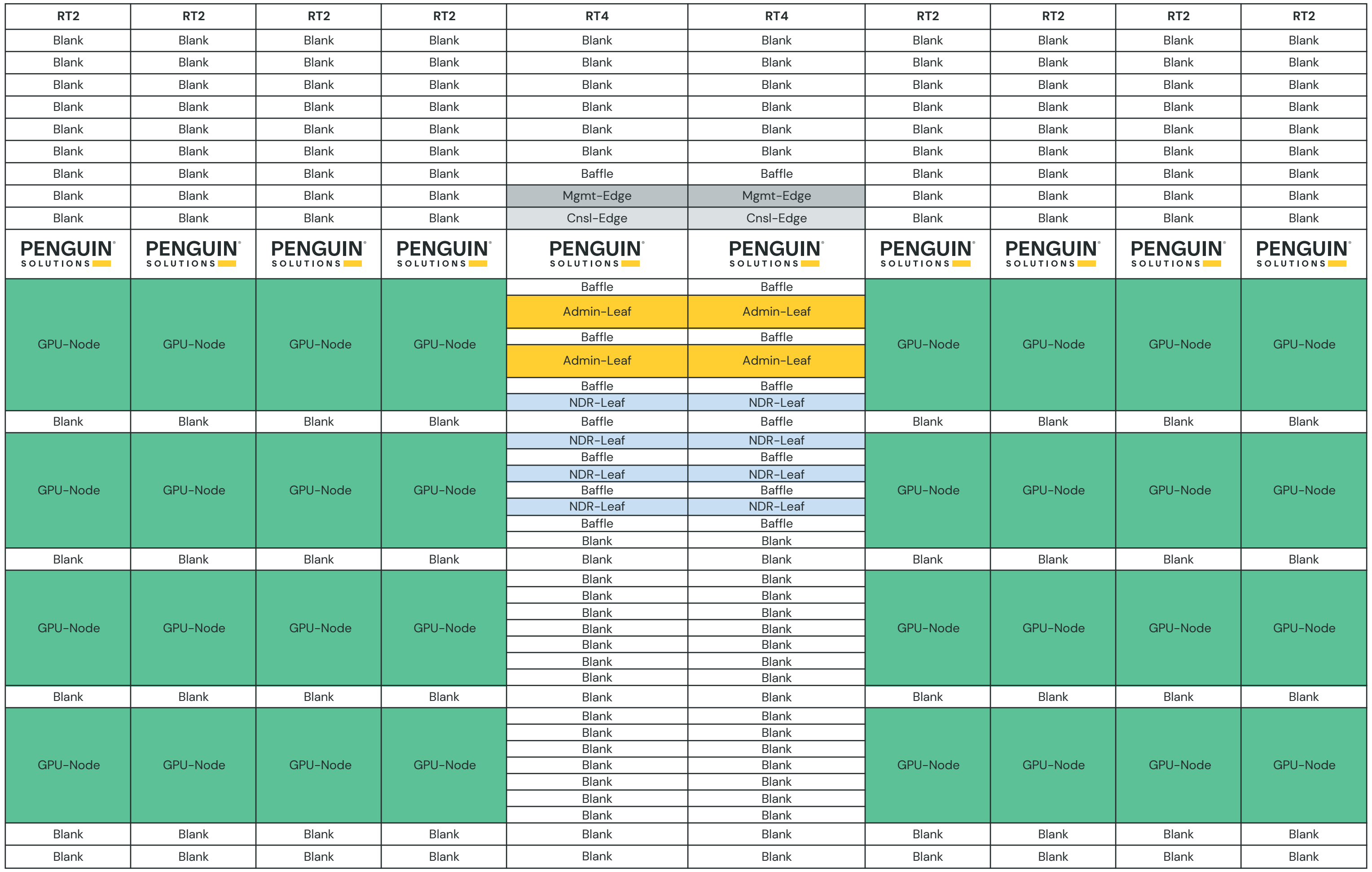


Note that the B200 Rack uses more power than the next comparable solution. This means that standard server configurations won't work for new AI systems.

© 2025 Penguin Solution

Here are Potential Power Environments for an AI Factory:

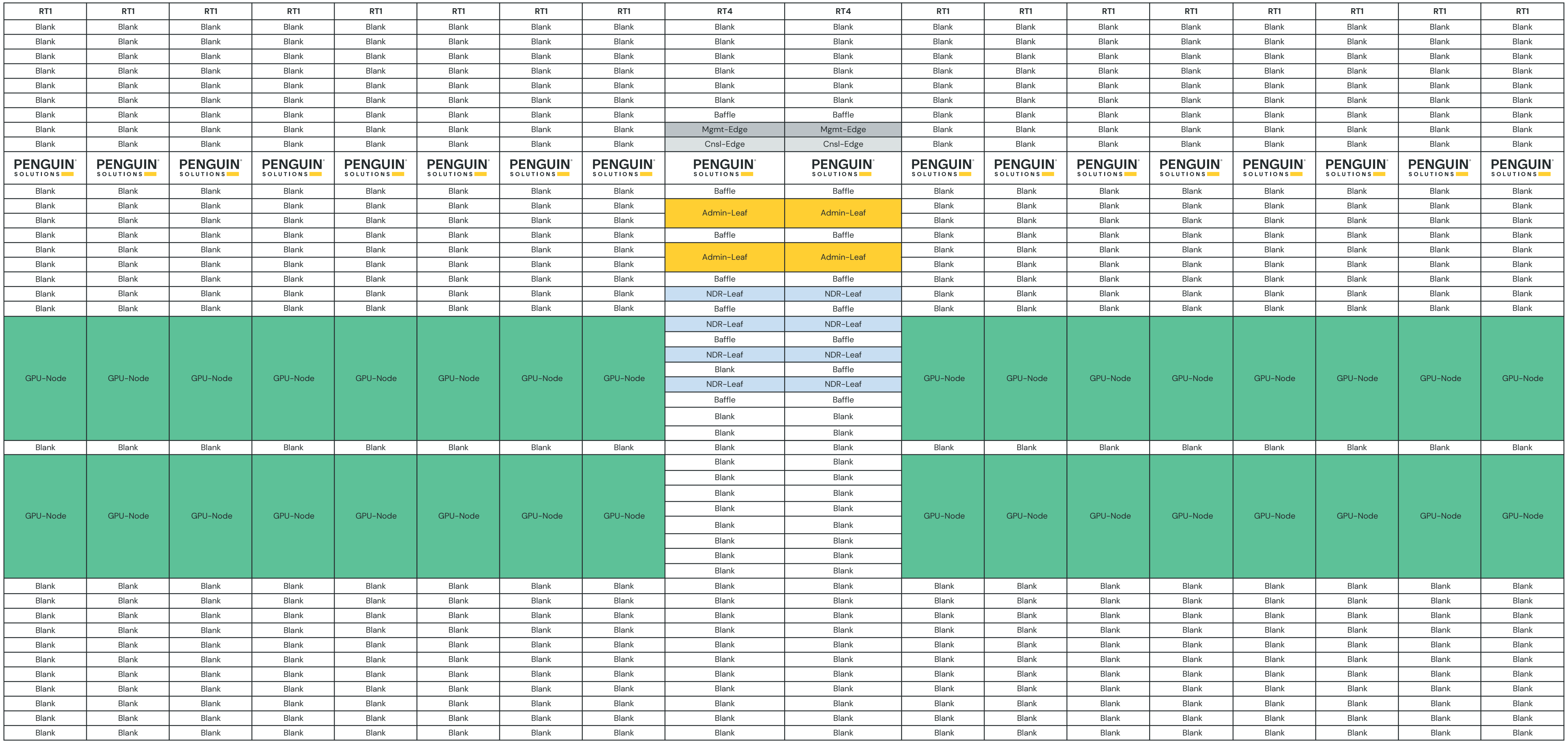
This diagram illustrates a scalable AI factory POD architecture using 4-node racks to support dense GPU-powered workloads. By aligning system layout with power and cooling requirements, this compact 10-rack configuration enables a 32-node compute pod—providing an efficient path from infrastructure design to AI workload delivery.
This example is great for data centers that can handle high power and cooling loads. Enterprises use less floor space and get more performance per rack.



RT2	RT2	RT2	RT2	RT4	RT4	RT2	RT2	RT2	RT2
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Baffle	Baffle	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Mgmt-Edge	Mgmt-Edge	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Cnsl-Edge	Cnsl-Edge	Blank	Blank	Blank	Blank
PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS
GPU-Node	GPU-Node	GPU-Node	GPU-Node	Baffle	Baffle	GPU-Node	GPU-Node	GPU-Node	GPU-Node
				Admin-Leaf	Admin-Leaf				
				Baffle	Baffle				
				Admin-Leaf	Admin-Leaf				
				Baffle	Baffle				
				NDR-Leaf	NDR-Leaf				
Blank	Blank	Blank	Blank	Baffle	Baffle	Blank	Blank	Blank	Blank
GPU-Node	GPU-Node	GPU-Node	GPU-Node	NDR-Leaf	NDR-Leaf	GPU-Node	GPU-Node	GPU-Node	GPU-Node
				Baffle	Baffle				
				NDR-Leaf	NDR-Leaf				
				Baffle	Baffle				
				NDR-Leaf	NDR-Leaf				
				Baffle	Baffle				
				Blank	Blank				
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
GPU-Node	GPU-Node	GPU-Node	GPU-Node	Blank	Blank	GPU-Node	GPU-Node	GPU-Node	GPU-Node
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
GPU-Node	GPU-Node	GPU-Node	GPU-Node	Blank	Blank	GPU-Node	GPU-Node	GPU-Node	GPU-Node
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
				Blank	Blank				
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank

© 2025 Penguin Solution

This next layout represents a 32-node AI factory POD configured with 2-node racks—an approach often driven by lower power availability within a data center. While requiring more physical space and blanking for thermal management, this design is a practical and common choice when power constraints shape system deployment.



RT1	RT1	RT1	RT1	RT1	RT1	RT1	RT1	RT4	RT4	RT1	RT1	RT1	RT1	RT1	RT1	RT1	RT1
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Baffle	Baffle	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Mgmt-Edge	Mgmt-Edge	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Cnsl-Edge	Cnsl-Edge	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS	PENGUIN SOLUTIONS
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Baffle	Baffle	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Admin-Leaf	Admin-Leaf	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank			Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Baffle	Baffle	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Admin-Leaf	Admin-Leaf	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank			Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Baffle	Baffle	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	NDR-Leaf	NDR-Leaf	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Baffle	Baffle	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	NDR-Leaf	NDR-Leaf	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node
								Baffle	Baffle								
								NDR-Leaf	NDR-Leaf								
								Blank	Baffle								
								NDR-Leaf	NDR-Leaf								
								Baffle	Baffle								
								Blank	Blank								
								Blank	Blank								
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	Blank	Blank	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node	GPU-Node
								Blank	Blank								
								Blank	Blank								
								Blank	Blank								
								Blank	Blank								
								Blank	Blank								
								Blank	Blank								
								Blank	Blank								
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank
Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank	Blank

2. Focus on the Networks

For AI factories to generate actionable outputs, continuous model training is necessary. The more data AI nodes process and the more they interact with other nodes in the cluster, the better the results.

However, AI training networks execute at the speed of the slowest node. Much like a cross-country running team, true success depends on collective performance. The finish line isn't truly crossed until every member is across it—even one delay can mean the difference between winning and being left behind.

The network is the platform

Up to **30% of wall clock time in AI and ML training** is spent waiting for the network. This is because even a single slow connection can drag down the performance of an entire AI training workload. Given that companies are now making big investments in AI architecture, even small improvements in network performance can yield valuable results.

Requirements for network tuning

Network tuning can help improve AI factory performance. One of the most effective tuning approaches is addressing large cluster underperformance. A **recent Meta analysis** found that small cluster performance can reach 90%+ out of the box—including overall communication bandwidth and utilization—while large clusters often have poor utilization, ranging anywhere from 10% to 90%. After optimizing network and software solutions, large clusters returned to the 90% range.

3. Plan to Isolate Root Causes

Identifying poor performance is easy. Addressing the symptoms is straightforward, but finding and fixing the root cause is far more complex.

Companies searching for performance issues are looking for a needle in a digital haystack. Given that each node in an AI cluster can have 4-8 GPUs, 10+ network connections, and multiple components—including memory, CPUs, and firmware—it becomes difficult to identify the exact cause.

Because of this, it's critical to focus on "why" issues occur rather than simply identifying "what" is wrong.

Adopting a systematic approach to debugging

Uncovering the "why" starts with a systematic and methodical approach. This means implementing error handling based on proven, best-practice workflows, such as:

- Leveraging a state machine framework
- Integrating with alerting subsystems
- Enabling automated triage and remediation
- Managing the full lifecycle—from detection to phase-out, RMA, requalification, and return to production

With Penguin Accumulator Technology, enterprises can identify the source of cluster performance issues before errors show up in logs. This is accomplished using bi-directional performance evaluations of GPUs and InfiniBand network adapters.

© 2025 Penguin Solution

4. Fully Define "Production Ready" at the Start

Few things undermine enthusiasm for AI faster than slow performance and failed user jobs. To prevent this, it’s essential to define what "production-ready" means for your organization before deploying AI.

False starts are costly and time-consuming. To avoid early setbacks, begin with the end in mind. This means clearly identifying:

- What you want to achieve with AI
- What resources are in place to accomplish strategic goals
- What potential roadblocks could arise

Create a production-ready checklist

A production-ready checklist helps you streamline initial deployments and minimize complexity for subsequent expansions. Key checklist items include:

- ✓ Cluster stability
- ✓ Overall performance
- ✓ Average throughput
- ✓ Impact of system outages
- ✓ Power availability
- ✓ Typical bandwidth

© 2025 Penguin Solution

5. Plan for Production Management and Expansion

Finally, plan for long-term management and AI factory expansion to ensure continued success.

Once your system is operational, achieving peak performance becomes the new priority. AI factories aren't static solutions—they evolve alongside your business. This means reaching production is just the first step.

Real-life changes require ongoing management of AI operations, such as:

- Firmware and software updates
- Failure minimization
- Consistent scalability

Examples of ongoing operations

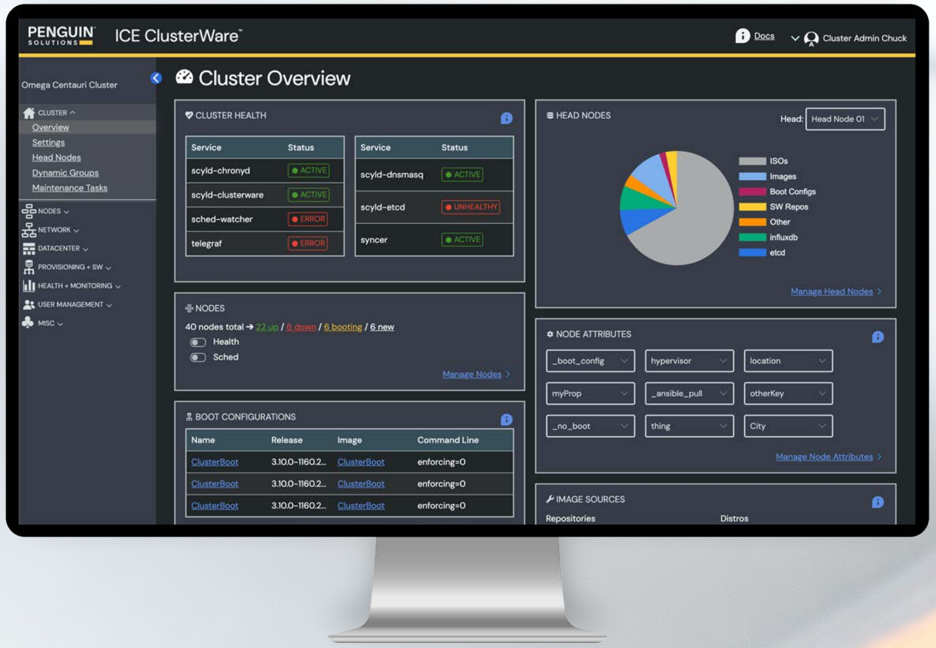
Consider firmware updates. These updates are necessary for maintaining peak AI infrastructure performance but can introduce system downtime risks. Job execution issues carry the same risk—jobs abruptly failing can lead to user frustration and reduced ROI.

© 2025 Penguin Solution

Improve Infrastructure Management with Purpose-Built Tools

Purpose-built tools, such as **Penguin Solutions ICE ClusterWare™** software, can help improve factory operations. ICE ClusterWare simplifies cluster deployment, administration, monitoring, and scaling while delivering:

- ✓ Streamlined access management
- ✓ Advanced automation
- ✓ Real-time insights
- ✓ Foundational simplicity
- ✓ Scalability from day one



Service	Status
scyld-chronyd	● ACTIVE
scyld-clusterware	● ACTIVE
sched-watcher	● ERROR
telegraf	● ERROR

Service	Status
scyld-dnsmasq	● ACTIVE
scyld-etcd	● UNHEALTHY
syncer	● ACTIVE

40 nodes total → 22 up / 6 down / 6 booting / 6 new

○ Health

○ Sched

Name	Release	Image	Command Line
ClusterBoot	3.10.0-1160.2...	ClusterBoot	enforcing=0
ClusterBoot	3.10.0-1160.2...	ClusterBoot	enforcing=0
ClusterBoot	3.10.0-1160.2...	ClusterBoot	enforcing=0

_boot_config	hypervisor	location
myProp	_ansible_pull	otherKey
_no_boot	thing	City

Learn More

Unlocking AI Factory Success

Unlocking AI factory success doesn't happen by accident—it's the result of deliberate planning, strategic alignment, and continuous execution.

1. Balance design, technology, and expertise

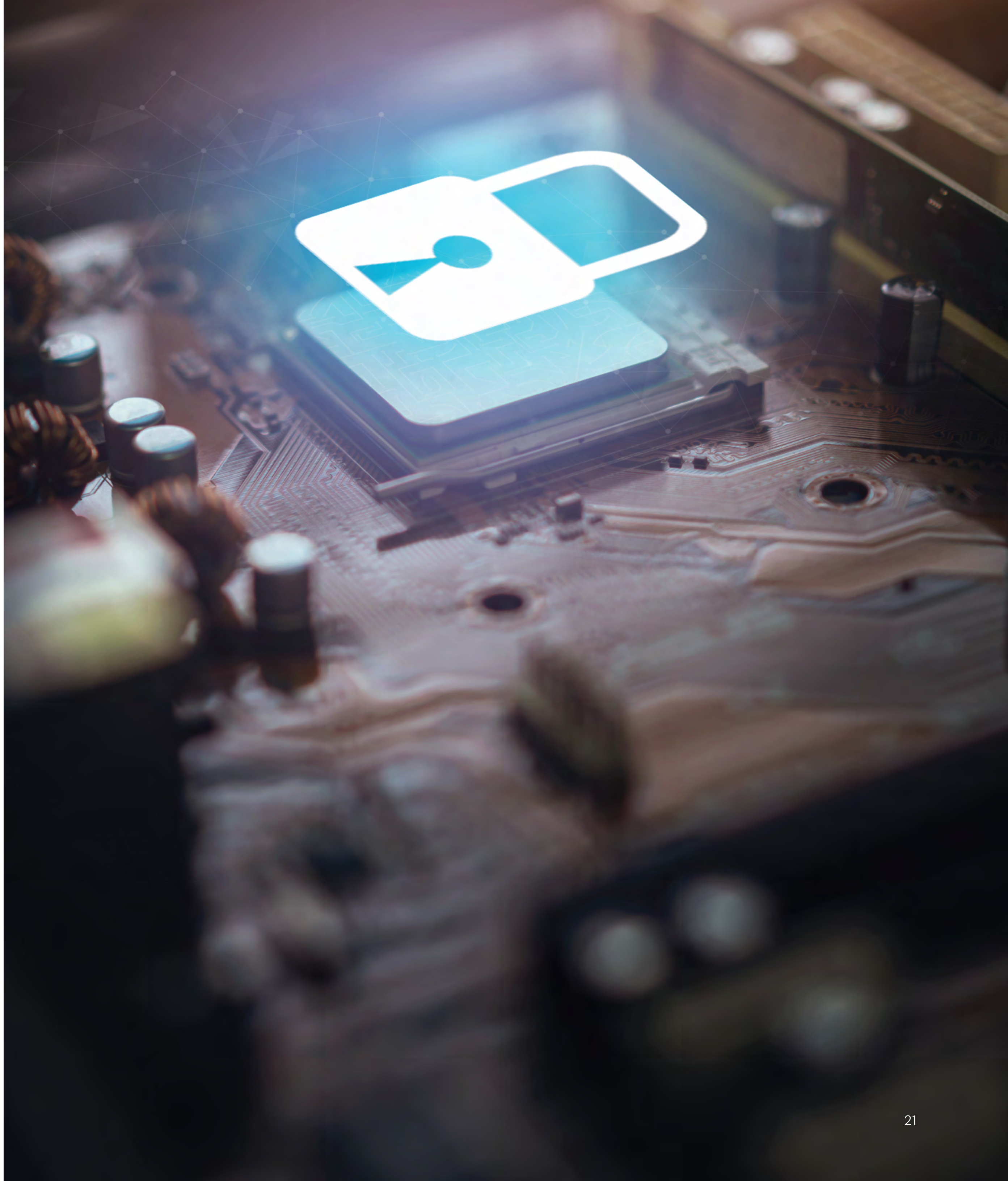
High-performing AI factories balance thoughtful data center design, advanced technology, and deep operational expertise.

2. Implement robust and scalable infrastructure

To enable long-term success, businesses must deploy and maintain robust infrastructure designed with scalability in mind.

3. Work with a trusted partner

The complexity of AI factory design and deployment makes a trusted AI partner essential for navigating workload and infrastructure design nuances.



© 2025 Penguin Solution

Partner with Penguin Solutions for AI Success

Achieving sustained AI success takes more than infrastructure alone—it calls for a partner who brings technical depth, operational excellence, and a long-term vision. Penguin Solutions delivers all three.

Penguin Solutions applies more than 25 years of HPC experience to designing, building, deploying, and managing AI factories to operationalize the use of AI. We have applied best practices and leveraged our strong and long-term relationship with our technology partners to build highly efficient AI infrastructure.

Our proven track record includes:

Managing and deploying 85,000 GPUs supporting demanding workloads across diverse industries.

Delivering more than 2.2 billion GPU runtime hours, powering real-world AI innovation in industries including Financial Services, Energy, Life Sciences and Healthcare, Government, and Higher Education.

Penguin Solutions helps organizations confidently navigate the complexity of AI infrastructure—so you can accelerate time to value, reduce risk, and future-proof your AI strategy.

Ready to take the next step? Visit us at **Penguin Solutions** or **contact our team today**.

PENGUIN®
SOLUTIONS

© 2025 Penguin Solutions, Inc. All rights reserved. Penguin Solutions, Penguin Computing, OriginAI, and ICE ClusterWare are trademarks or registered trademarks of Penguin Solutions. A ll other product names, trademarks, and registered trademarks are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, trademarks, and brands does not imply endorsement.