BUYER'S GUIDE

Take control of your *AI Infrastructure* with Ori *Private Cloud*

Transform your enterprise AI with cloud infrastructure that is secure, highly-performant, and cost-effective—keeping you ahead of your competition.

ORI

ori.co

Contents

Rethinking AI infrastructure: Why Private Cloud is essential to unlock innovation 1

Ori Private Cloud: Enterprise AI platform to build anything imaginable 2

Inside Ori Private Cloud: Core capabilities 3

The cornerstones of your enterprise AI transformation 4

Build AI faster 4

Develop secure AI 5

Amplify your ROI 5

Comparing Ori Private Cloud to traditional public cloud solutions 6

Build secure AI,faster on Ori Private Cloud 6

Rethinking *AI infrastructure*: Why Private Cloud is essential to unlock *innovation*

As artificial intelligence becomes a central driver of innovation, enterprises face mounting challenges with traditional public cloud environments. Here are some industry insights that highlight why a dedicated, private AI cloud is not only desirable but essential:

Rapid AI adoption

78%

of enterprise leaders expect to increase their overall AI spending in 2025.[1]

Escalating security concerns

80%

of companies have encountered an increase in the frequency of cloud attacks in 2024[2], emphasizing the need for an isolated, secure environment.

Rising cloud costs

30%

Enterprise cloud costs rose an average of 30% with spending on AI applications and generative AI cited as top drivers for growing cloud spend[3], reinforcing the need for cost-efficiency.

Moreover, modern AI infrastructure faces significant challenges:

- **Infrastructure complexity**: Setting up and managing large GPU clusters with robust interconnectivity and storage is inherently complex.
- **Hyperscale cloud Costs**: Hyperscale clouds are 3-4 times more expensive than AI Native cloud providers such as Ori[4].
- **Resource underutilization**: Inefficient GPU utilization can lead to substantial financial losses (a 1,024 GPU cluster operating at 70% efficiency can incur a $6M loss over three years, worsening as efficiency drops)[5].

Resource underutilisation

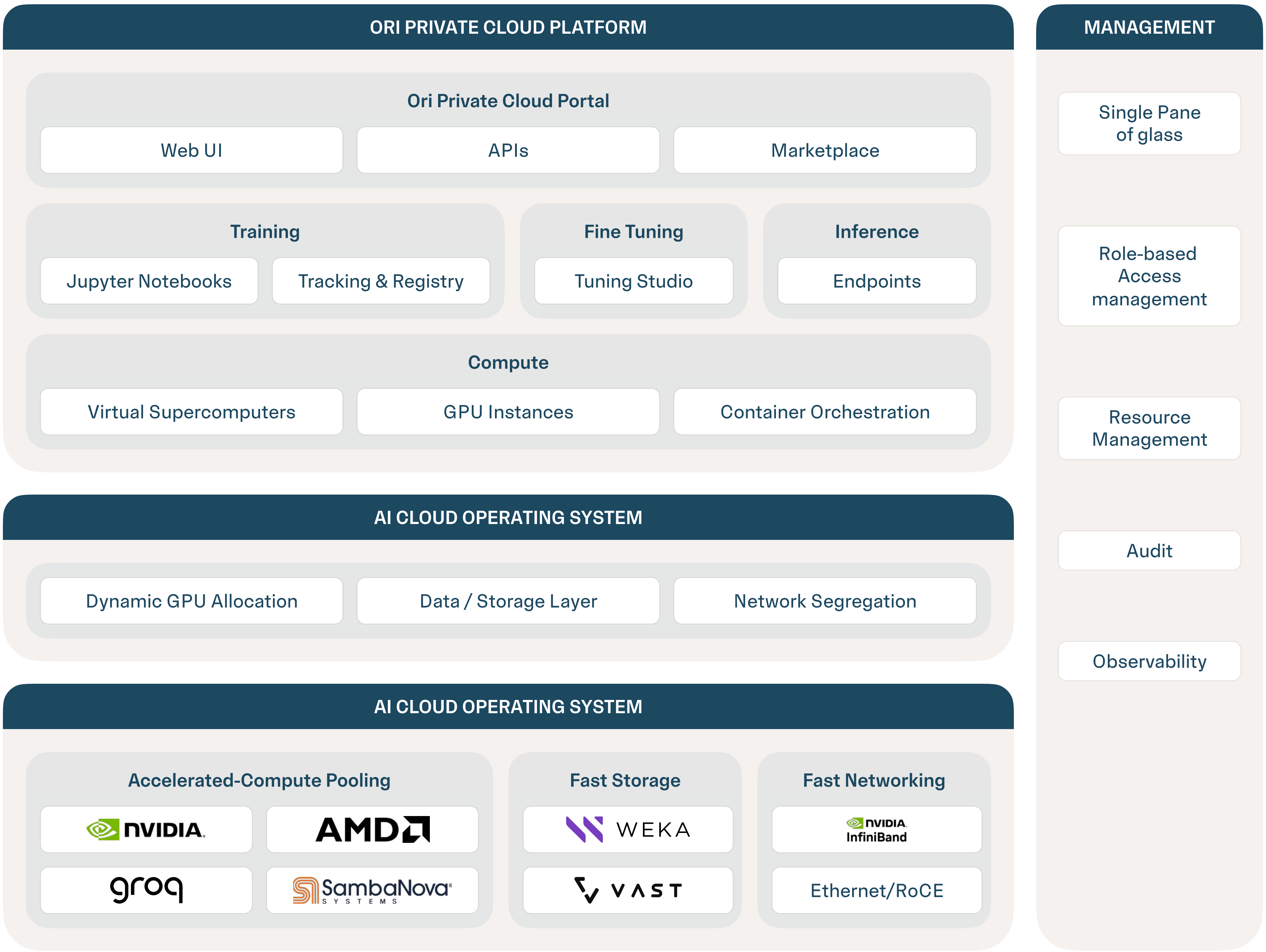
GPU Efficiency	Financial loss
70%	$6M
50%	$10M
20%	$16M

These trends call for an AI infrastructure that is flexible enough to support innovation while delivering the security, compliance, and cost control required by today's enterprises.

1. Deloitte (2025). State of Gen AI Report: Wave 4; 2. SentinelOne (n.d.). Cloud Security Statistics; 3. Tangoe (2024, February 5); 4. Ori Pricing; 5. Ori Analysis of estimated loss (computing only) over three years of a 1,024 GPU cluster (excludes opportunity cost, impact of a slower time to market, cost to the business, etc.).

Ori Private Cloud: *Enterprise* AI platform to build *anything* imaginable

Ori Private Cloud offers a dedicated, scalable platform specifically engineered for the demands of modern AI workloads. By blending the agility of cloud computing with the control of a private environment, Ori Private Cloud addresses the critical challenges of traditional infrastructure.



ORI PRIVATE CLOUD PLATFORM

Ori Private Cloud Portal

Web UI

APIs

Marketplace

Training

Jupyter Notebooks

Tracking & Registry

Fine Tuning

Tuning Studio

Inference

Endpoints

Compute

Virtual Supercomputers

GPU Instances

Container Orchestration

AI CLOUD OPERATING SYSTEM

Dynamic GPU Allocation

Data / Storage Layer

Network Segregation

AI CLOUD OPERATING SYSTEM

Accelerated-Compute Pooling

NVIDIA

AMD

groq

SambaNova SYSTEMS

Fast Storage

WEKA

VAST

Fast Networking

NVIDIA InfiniBand

Ethernet/RoCE

MANAGEMENT

Single Pane of glass

Role-based Access management

Resource Management

Audit

Observability

Inside Ori Private Cloud: *Core* capabilities

AI-optimized compute platform

- **Flexible and Powerful Compute Options:** Choose from high-performance Compute options such as **NVIDIA**, **AMD**, **Groq**, **Samba Nova**, **Qualcomm and more**. For complex workloads with varying performance requirements, you can blend multiple silicon platforms or "**Bring Your Own Compute**".
- **Virtual Supercomputers**: High-performance compute with **multi-node**, **interconnected GPU instances**.
- **Container Orchestration**: **Kubernetes-based orchestration** to manage multiple workloads across teams efficiently.

Enterprise-grade management & security

- **Isolated and Secure**: Completely private infrastructure that only your organization can access.
- **Single-Pane-of-Glass Monitoring**: Unified dashboard for real-time AI infrastructure visibility.
- **Admin Control**: Seamlessly oversee a multi-tenant environment where each team has its own workspace, while maintaining visibility and management across your organization.
- **Role-Based Access Management (RBAC)**: Granular security controls and user access policies.
- **Audit & Compliance**: Built-in logging observability for easier **regulatory compliance**.

AI development & MLOps tools

- **Jupyter Notebooks**: Pre-configured environments for AI model development and experimentation.
- **Model Tuning & Experiment Tracking**: Tools for finetuning and AI workflow management.
- **Inference & Deployment Engine**: **Scalable API endpoints** to deploy AI models with ease.

High-performance AI infrastructure

- **Top-tier Compute**: Access powerful NVIDIA GPUs such as B200, H200 and H100 without the hassle of compute quotas.
- **Fast Storage Solutions**: Weka & VAST storage systems for ultra-fast data retrieval and processing.
- **High-Speed Networking**: **NVIDIA InfiniBand & Ethernet/RoCE** for non-blocking, high bandwidth, shared memory.

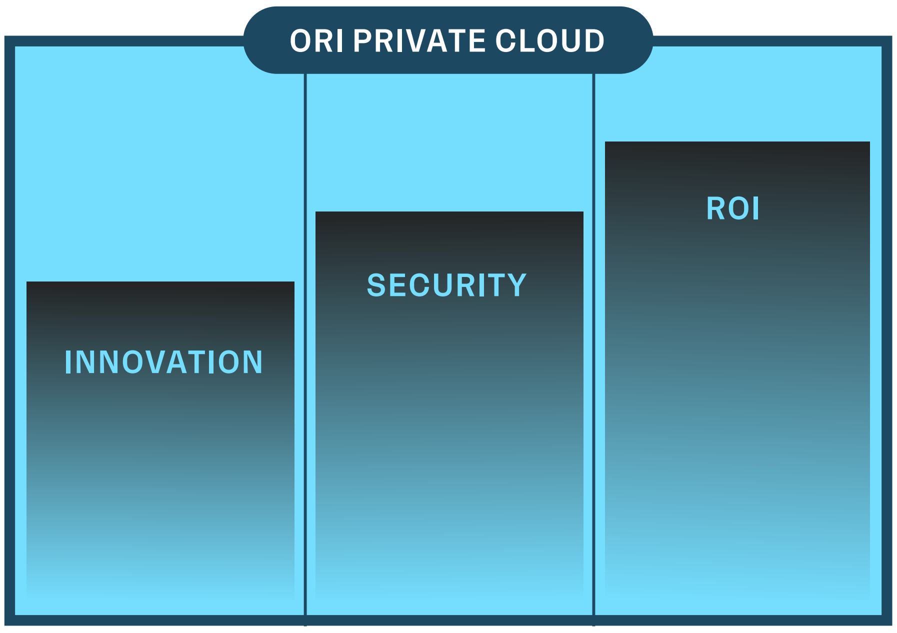
AI cloud operating system

- **Dynamic GPU Allocation**: Intelligent scheduling between services and workloads ensures **efficient GPU utilization**, reducing idle compute time.
- **Integrated Data & Storage Layer**: High-throughput, low-latency storage optimized for AI workloads.
- **Network Segregation**: Secure, isolated networking for data privacy and regulatory compliance.

Chart your own AI *reality*

Talk to a Private Cloud Expert →

The cornerstones of your *enterprise* AI transformation



Ori Private Cloud is designed around three core pillars that drive enterprise AI success:

- ✓ Accelerating innovation
- ✓ Ensuring robust security
- ✓ Optimizing return on investment.

Build AI *faster*

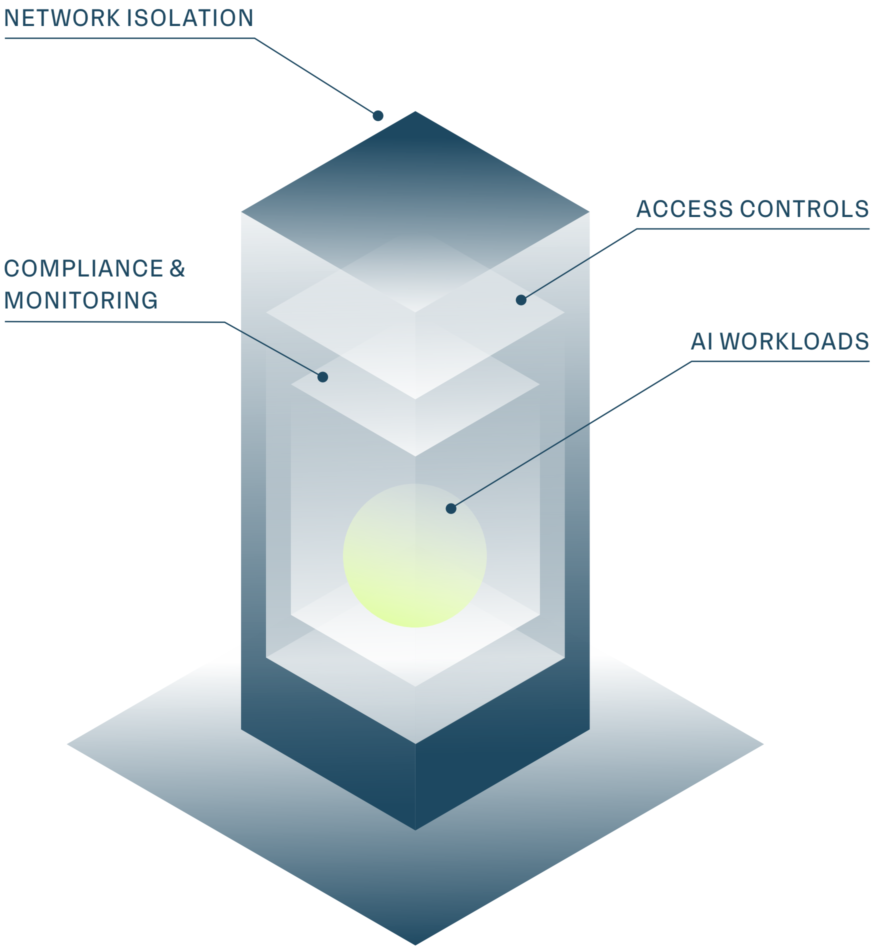
- **Get Your AI to Market Faster with Powerful GPU Clusters**: Deploy dedicated GPU clusters that dramatically reduce model training times, helping your AI projects move from concept to production with speed.
- **Purpose-built Storage and Networking for peak performance**: Purpose-built NVMe storage and low-latency Infiniband networking ensure that data flows unimpeded across your infrastructure, eliminating performance bottlenecks.
- **End-to-end platform to train, scale and serve your AI models and apps**: From experimenting with models on GPU instances to training them on GPU clusters and deployment via the Ori Inference Engine, enjoy a seamless, integrated platform that supports every stage of your AI lifecycle.

Here's a snapshot of one of Ori's recent Private Cloud deployments:

1024 GPUs	NVIDIA H100
32 PFLOPS	PERFORMANCE
1 PB	PARALLEL STORAGE
3.2 Tbps	GPU INTERCONNECT

Develop *secure* AI

- **Flexible Deployment Environments**: Choose the environment that best suits your requirements—whether public, hybrid, or on-premises—to ensure your data remains where you need it.
- **Unified management across teams, exclusive to your organization**: Run and manage multiple AI workloads across teams with role-based access controls (RBAC), continuous monitoring and single-pane-of-glass observability, but in a fully isolated environment that is exclusive to your organization.
- **Achieve compliance and data sovereignty requirements**: With built-in adherence to standards like ISO 27001 and SOC2, Ori Private Cloud simplifies compliance with data sovereignty and regulatory requirements, ensuring audit readiness.



Ori Private Cloud

Compared

78% less expensive

Than traditional hyperscalers (3-year commitment)

$5M + potential savings

From optimal GPU utilization over 3-year period.[1]

Amplify your ROI

- **Private Cloud as a Service**: Transition from heavy capital investments (CapEx) to a flexible operating expense (OpEx) model, ensuring predictable costs and minimizing upfront expenditure.
- **Maximum Resource Utilization**: Intelligent scheduling and dynamic scaling ensure that every GPU is used efficiently, reducing the financial impact of underutilization.
- **Significantly lower operational costs than self-managed infrastructure**: Enjoy predictable pricing without hidden fees, enabling you to accurately forecast expenses. Our reference architecture is designed to vastly reduce your total cost of ownership (TCO) making Ori Private Cloud much more cost-effective than building and maintaining your own infrastructure.

[1] Assuming a 1,024 H100 GPU cluster at 70% utilization.

Comparing Ori Private Cloud to traditional public cloud *solutions*

Aspect	Ori Private Cloud	Traditional Public Cloud
INFRASTRUCTURE	Dedicated, single-tenant hardware ensuring total control and isolation.	Multi-tenant environments with shared resources.
DEPLOYMENT FLEXIBILITY	Options for public, hybrid, or on-premises deployments tailored to your needs.	Primarily restricted to the provider's data centers.
PERFORMANCE	Customizable for AI workloads with top-tier GPUs, fast storage, and low-latency networking.	Non-customizable tech stack with potential performance trade-offs.
SECURITY & COMPLIANCE	Robust, enterprise-grade security with complete data isolation and compliance.	Shared environment with inflexible security options.
COST EFFICIENCY	Transparent, usage-based pricing that minimizes CapEx and maximizes ROI.	Complex billing with variable costs and hidden fees.
SUPPORT & MANAGEMENT	Fully managed service with 24/7 expert support and Tier-4 data centers.	Self-managed or expensive Premium Support which many times is not AI-first.

Build secure AI, *faster* on Ori Private Cloud

Talk to a Private Cloud Expert →

Investing in purpose-built AI infrastructure is critical for enterprises that aspire to lead in innovation while maintaining stringent security and cost control. Ori Private Cloud delivers a fully flexible, enterprise-grade solution that empowers you to:

Accelerate AI development and market *readiness*

Develop AI in a secure, compliant *environment*

Optimize operational costs and *maximize* your ROI

About Ori

Ori is the first AI Infrastructure provider with the native expertise, comprehensive capabilities and end-to-endless flexibility to support any model, team, or scale. We're building the backbone of the AI era so that the technology of tomorrow can advance our world.

Ori believes that the promise of AI will be determined by how effectively AI teams can acquire and deploy the resources they need to train, serve, and scale their models. By delivering comprehensive, AI-native infrastructure that fundamentally improves how software interacts with hardware, Ori is driving the future of AI.

Learn more at www.ori.co →