REFERENCE ARCHITECTURE

Ori AI Fabric: A reference architecture for *sovereign* and *private* AI clouds

ori.co

Contents

Introduction

The Ori AI Fabric platform is a single control plane that unifies AI compute, storage, networking, orchestration and machine-learning tooling. Ori uses this platform to power its public cloud offering and licenses it to third parties to build their own clouds.

This reference architecture is focused on the licensed platform. A brief overview of the Ori public cloud is included in the appendix. The underlying software architecture is the same across both instances.

1. Executive summary and guiding principles

1.1 Guiding principles

Ori's platform architecture is rooted in these foundational principles:

End to end: The Ori AI Fabric unifies AI compute, storage, networking, orchestration and machine-learning tooling on a single platform. This eliminates glue code and operational friction associated with Day 2 operations.

Multi-silicon compute: Ori supports a spectrum of AI compute options from general-purpose GPUs to AI accelerators and allows blending of different types of compute and compute in different locations (datacenter, on-prem).

Elasticity with full control: Resources can be allocated on demand but administrators retain precise governance through role-based access control (RBAC), quotas and policy enforcement.

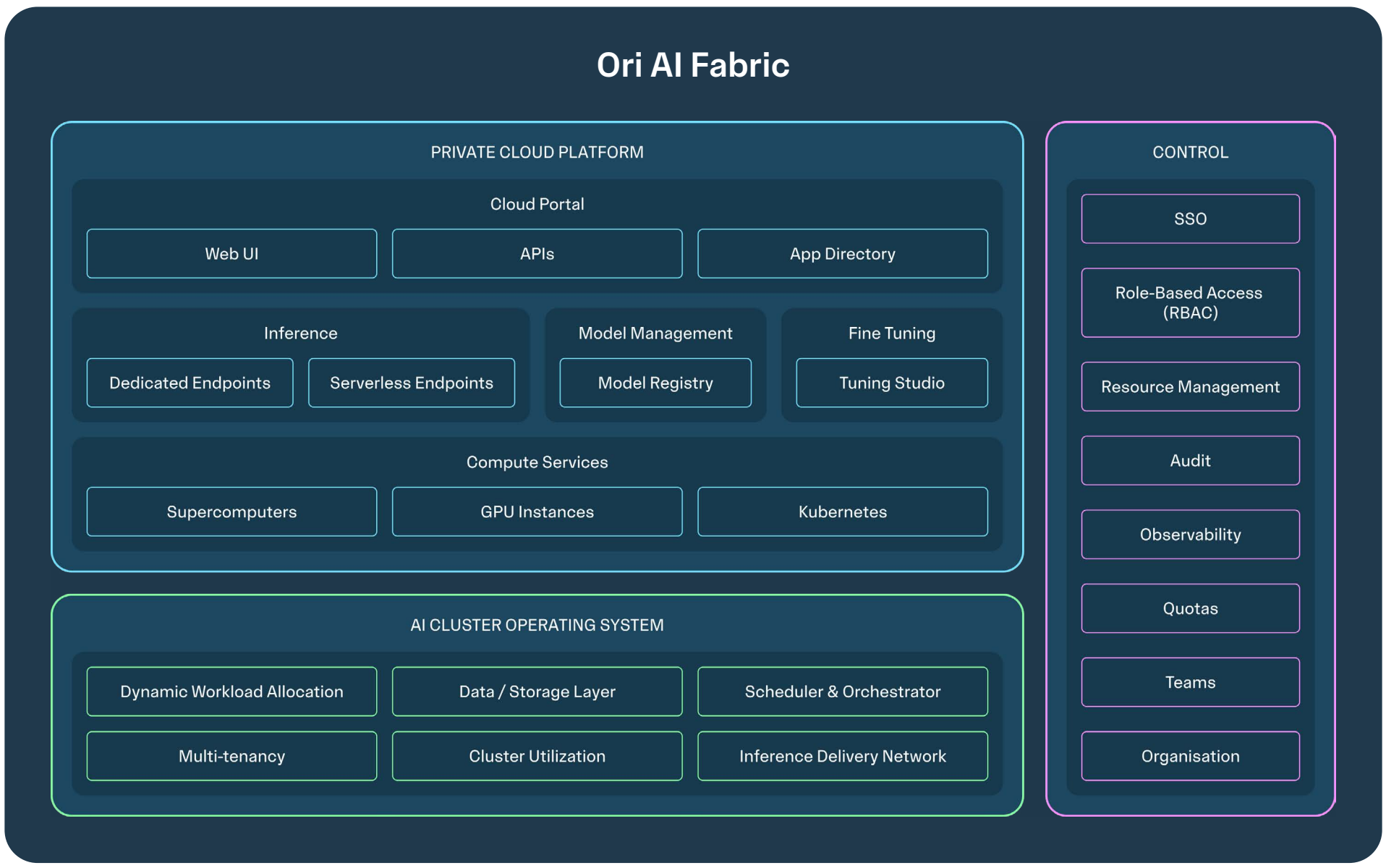
Secure multi-tenancy: The platform isolates tenants across compute, storage and networking, enabling multiple teams or customers to share infrastructure without interference. A continuum of tenancy options are supported.

Sovereignty & hybrid flexibility: Workloads can run on-premises or in a particular region for data sovereignty and optionally burst to Ori's global cloud for elasticity. A unified control plane enforces region-aware scheduling and data residency policies while providing a single pane of glass to the administrator.

Compliance by design: The system enforces SSO, 2FA, audit logging and meets SOC 2 Type II, ISO 27001, HIPAA and GDPR requirements.

1.2 Architecture overview

Ori's AI Fabric was woven across hundreds of clients running on thousands of GPUs. It is a complete software stack that enables customers to exert control over their infrastructure, while maintaining flexibility over their business choices.



The Ori AI Fabric platform is built around a unified control plane, noted above as the Ori Private Cloud Portal. It can be accessed via intuitive Web UI, REST APIs, and a fully featured CLI. These access surfaces feed directly into both high-level machine learning services and infrastructure services, enabling end-user ML teams to iterate quickly on experiments, integrate AI workflows into CI/CD pipelines, take models to production and maintain full auditability of every action. It can be whitelabeled by customers who license the platform and expose it to end user customers.

Ori's machine learning suite comprises services for inference, model management and fine-tuning. For inference, customers may offer either Dedicated Endpoints and/or Serverless Endpoints that autoscale on demand. The Ori Model Registry provides storage, versioning, and easy deployment for all your models, while Ori FineTuning Studio orchestrates distributed fine-tuning jobs with built-in support for parameter-efficient techniques.

Beneath these services sits a cloud-native compute and orchestration layers. Using the AI Fabric platform, customers can configure compute is multiple ways:

Virtual Supercomputers

One-click GPU clusters connected by NVIDIA Infiniband

GPU Clusters

Bare-metal instances to be configured by the end customer/stakeholder

GPU Instances

Virtualized instances that are ready to be deployed on-demand

Managing those compute offerings, which include Serverless Kubernetes, requires sophisticated software. Ori's AI Fabric has an operating system layer that continuously optimizes resource utilization with real-time workload distribution, multi-tenant isolation and policy-driven autoscaling. This layer includes a pluggable data/storage layer that supports GPUDirect- enabled parallel storage and high-speed object storage, while a custom scheduler and inference delivery network ensure low-latency model distribution and high cluster throughput.

Underpinning the entire stack is a validated hardware ecosystem - NVIDIA, AMD, Groq, Qualcomm accelerators; InfiniBand and RoCE networking fabrics; WEKA and VAST parallel storage, combined with enterprise-grade controls that enforces SSO/federation, hierarchical RBAC, resource quotas, audit logging, and full observability. Teams and organizations map directly to isolation boundaries and billing scopes, providing governance controls while maintaining agility.

1.3 Deployment models

Ori AI Fabric supports four deployment modes. Other options are possible, but the vast majority fall into the following categories.

Ori Public Cloud: This model runs on Ori Public Cloud and is fully-managed by Ori using the Ori AI Fabric. Ori provides, manages and supports the infrastructure. This is delivered as-a-service in Ori's data centers and covers all offerings (GPUs, Kubernetes, Inference, Fine-Tuning etc).

Best for: Teams needing the fastest path to production with zero infrastructure overhead.

Customer infrastructure, Ori managed with Ori AI Fabric: In this model the customer provides the infrastructure and Ori AI Fabric software is used to manage it. The location of the infrastructure is secondary, it can be on-prem, your private datacenter or your private deployment in a public datacenter. Ori is responsible for the Ori platform and control plane, including upgrades, security patches, observability and SRE.

> **Best for**: Maintaining data sovereignty and hardware control, without the day-to-day operational burden.

Customer infrastructure, customer managed with Ori AI Fabric: In this model the customer provides the infrastructure and uses Ori's AI Fabric software to manage it. The location of the infrastructure is totally up to the customer. Ori provides support, guidance and expertise. This option ensures maximum sovereignty as, depending on the configuration, data never leaves the customer's premises - satisfying regulatory requirements where they exist.

> **Best for**: Teams with in-house expertise seeking total operational autonomy and deep customization.

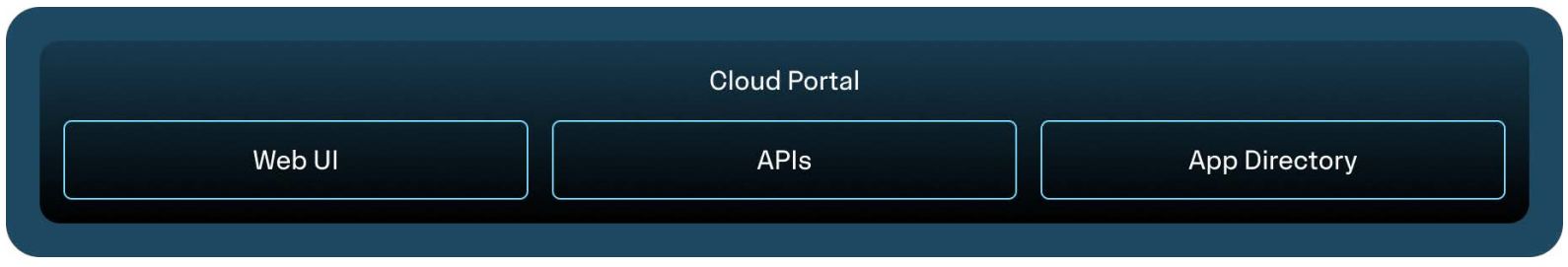
Hybrid cloud: In this model, the infrastructure is blended across public and private clouds. It is all managed through the Ori AI Fabric. The decision on who manages it is up to the customer.

> **Best for**: Unified management across distributed environments to optimize for cost, performance, and scalability.

All options are viable, however, specific design considerations may result in recommending one model over another. Ori will work with clients on a case by case basis to make those determinations.

> **Implementation tip**: If you already own accelerators or GPUs, Ori can help you leverage your compute investments with our Hybrid approach. This strategy is particularly useful if you have a mix of workloads that require different types of compute. Hybrid cloud is also a great choice if you have sensitive data that needs to stay on your premises, whereas the rest of your workloads can live on the cloud.

2. User interfaces, APIs and accessibility



The Cloud Portal serves as the unified, API-first control plane interface for all AI Fabric operations, abstracting the complexity of the underlying heterogeneous GPU infrastructure. It comprises the Web UI, comprehensive APIs and the App Directory, acting as the authoritative gateway for tenants, developers and platform operators.

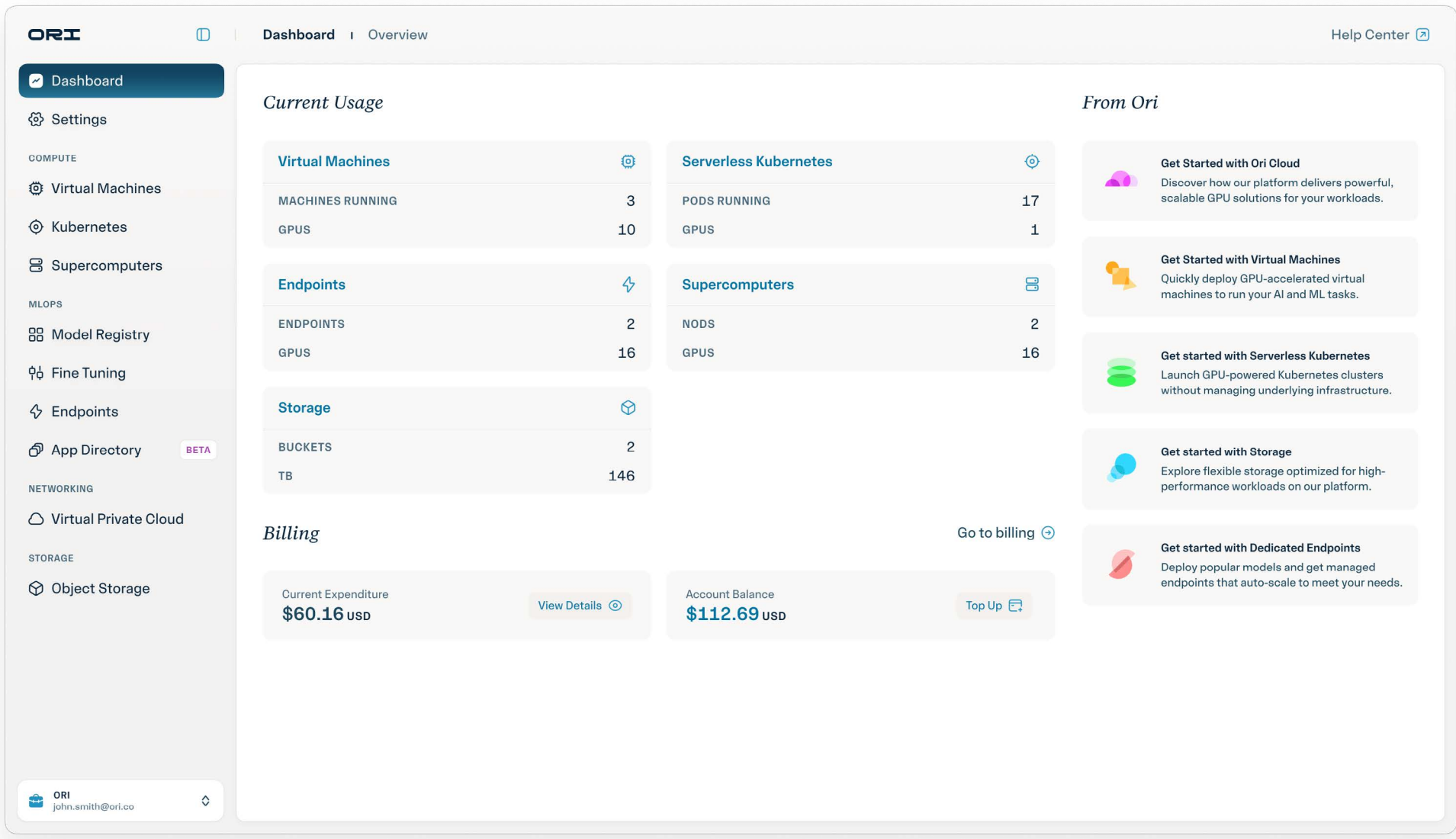
The portal operates as the cloud management platform (CMP) layer and is built on a standards-compliant architecture that prioritizes programmatic access and self-service.

APIs (Application Programming Interfaces): The platform is API-first, meaning every resource and workflow is accessible programmatically. These REST APIs are secured using API Keys passed in the Authorization header, which is linked to the Ori IAM & Federation Hub (Keycloak). This integration ensures that Role-Based Access Control (RBAC) is enforced down to the resource level.

- **Automation**: API calls initiate declarative workflows within the Automation and Orchestration Engine (AOE). This allows the platform to manage and reconfigure bare-metal, VM and containerized resources autonomously, ensuring self-service provisioning is policy-governed and auditable.

Web UI (User Interface): The Web UI is a single pane of glass built atop the very same APIs. It provides an intuitive, seamless UI that abstracts Kubernetes and infrastructure complexity away from ML scientists.

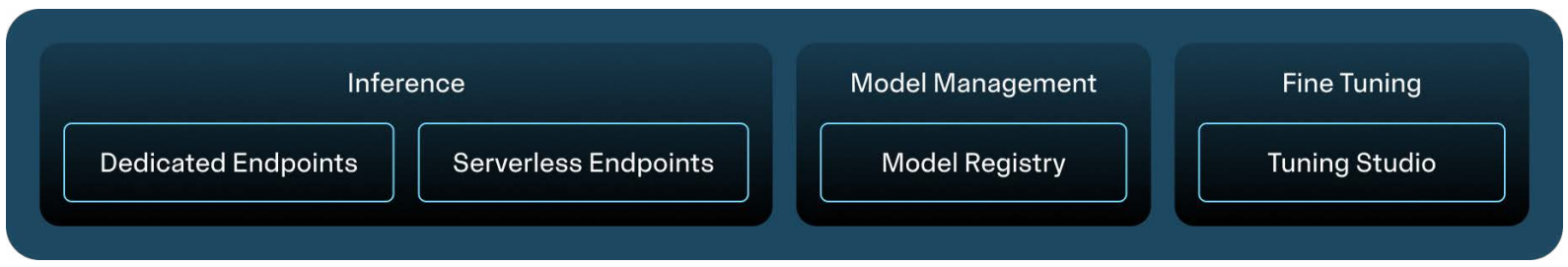
- **Functionality**: It provides real-time visibility into the Ori Observability Framework (metrics, logs, audit trails and cost metering) and is the primary interface for tenants to provision, destroy, suspend and resume GPU Instances, Supercomputers and Serverless Kubernetes clusters in under two minutes.



App Directory (Self-Service Catalog): This centralized service catalog exposes parameterized blueprints for common AI workloads, transforming complex infrastructure into consumable, deployable services.

- **Orchestration**: The catalog offers one-click deployment for core MLOps tools (e.g., Jupyter Notebooks, MLFlow), foundational models and custom pipelines (training, fine-tuning, RAG). This process triggers the underlying AOE to deploy all necessary components—including segregated network, storage and compute—while binding the deployment to the tenant's specific RBAC and quota policies.

3. MLOps and tooling



Ori's end-to-end machine-learning tooling is integrated directly into its infrastructure platform and is available in all deployment modes. This eliminates the need to piece together disparate services.

3.1 Model management: Model Registry

The Ori Model Registry acts as a central hub to store, version and deploy your AI models. It serves as the single source of truth for stakeholders. It is tightly integrated with both the Fine-Tuning Studio and Inference Endpoints. Models are automatically cached globally and distributed via the control plane to the point of compute, ensuring they are available where needed for fast, compliant and location-aware inference. The Ori Model Registry serves as the central hub for all AI assets, automatically registering and versioning models—regardless of whether they are uploaded by the customer or generated internally via the Fine-Tuning Studio—with full metadata for auditability.

3.2 Fine tuning: Tuning Studio

The Ori Fine Tuning Studio enables customers to fine-tune foundation models on your own datasets, abstracting the complexity of scripts and commands. The Studio fully automates scheduling resource provisioning retries and multi-stage validation monitoring. When training completes, your model is automatically stored and versioned in the Model Registry. This integrated workflow removes the need for manual orchestration and glue code, providing a simple, integrated path from custom data to a production-ready model.

3.3 Serverless Kubernetes: The pipeline orchestrator

Ori's Serverless Kubernetes delivers a fully managed, AI-centric Kubernetes environment by abstracting away node pools, load balancers and cluster configuration, enabling rapid AI/ML deployments with native Helm support. It offers precise autoscaling from zero to thousands of GPUs, enforced isolation via a dedicated control plane and pay-as-you-go billing to eliminate idle GPU costs.

3.4 Inference delivery: Dedicated and Serverless Endpoints

The Ori Inference Endpoints provide production-grade serving for models. While customers who license the Ori AI Fabric will ultimately determine their own deployment topology, they can optionally leverage the Ori global Inference Delivery Network (IDN) architecture to cache and route models for fast compliant and location-aware serving across their federated sites.

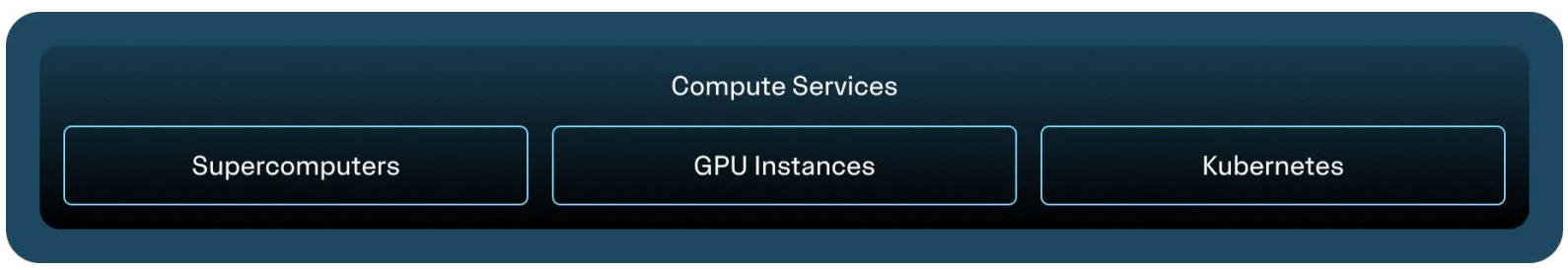
Two distinct modes address all latency, cost and compliance requirements:

- **Dedicated Endpoints**: Models are deployed on dedicated GPU instances with predictable performance. Customers retain full control over the endpoint hardware profile and can deploy open source models from our catalog or bring their own. This mode is suitable for latency-sensitive workloads or when compliance requires dedicated hardware.
- **Serverless Endpoints**: A fully managed offering where requests are routed to an autoscaling pool of GPUs or AI accelerators. Users are billed per token or per request. The IDN globally caches models and routes requests to the nearest endpoint, ensuring low latency and high availability for thousands of concurrent clients across regions.

3.5 Integration with external tooling

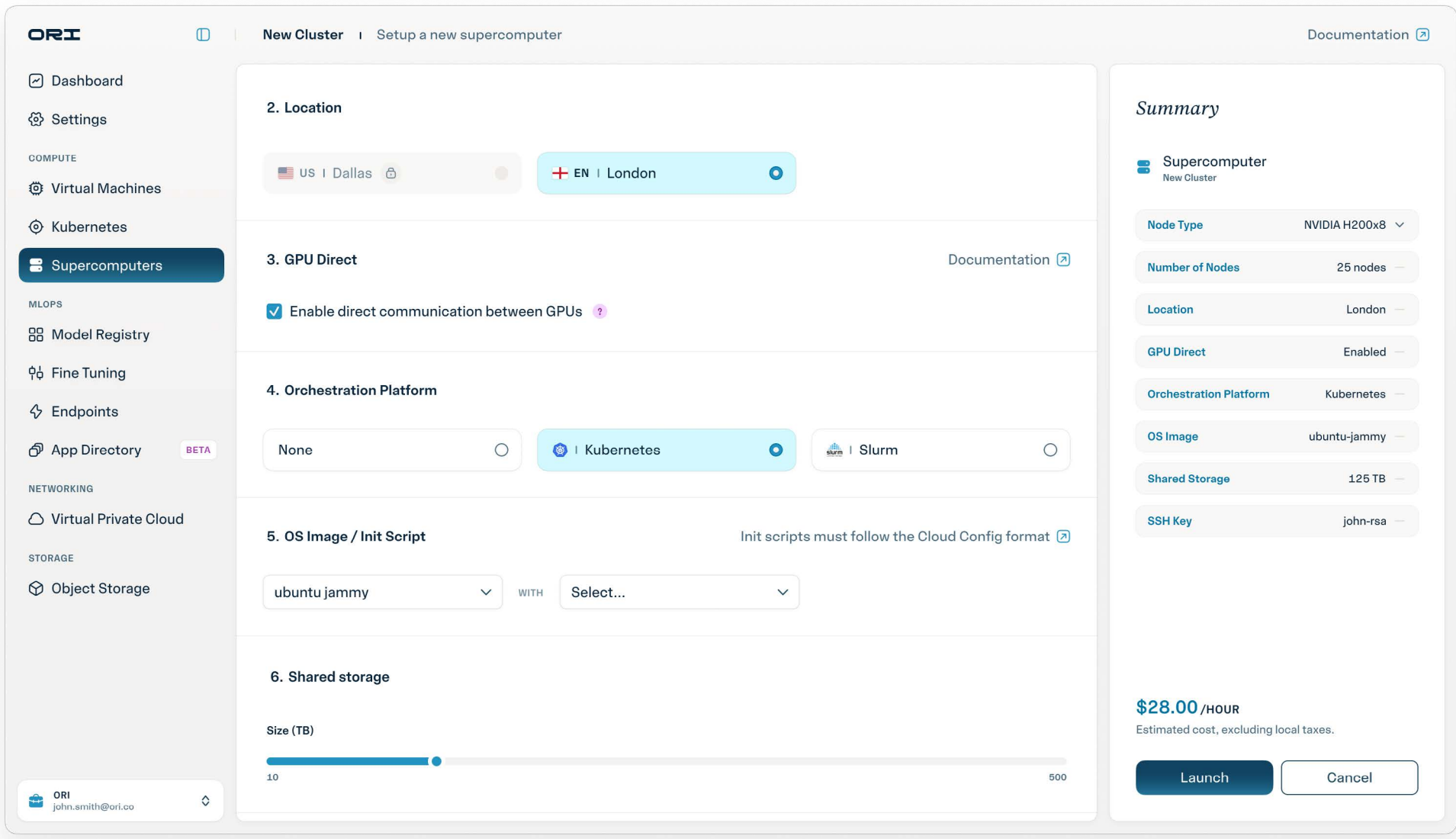
Ori exposes a comprehensive REST API to integrate all platform services with existing DevOps and MLOps workflows. Teams can manage infrastructure via GitOps pipelines and integrate with standard CI/CD tools (Jenkins, GitLab). Furthermore, all platform-level metrics are easily exported to popular observability tools such as Grafana, Prometheus and Datadog.

4. GPU offerings



For customers opting to leverage the Ori AI Fabric through the Ori Cloud, there are multiple pathways to support diverse AI workflows:

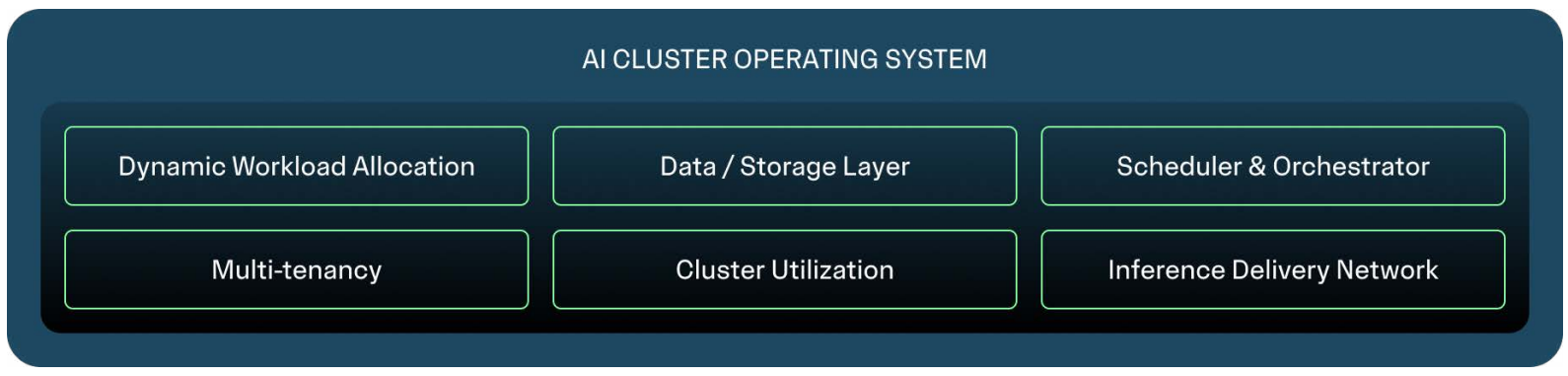
- Bare-Metal GPU Clusters: Direct access to GPU servers with full control over OS, drivers and scheduling. Suitable for training foundation models and large-scale inference.
- GPU Instances: Virtual machines with dedicated GPUs for agile experimentation and small-scale training. Users can select from a range of GPU generations (H100, H200, B200) or accelerators. Instances are provisioned in seconds through the Ori console or API.
- Virtual Supercomputers: Supercomputers provide on-demand, bare-metal clusters interconnected with high-speed NVIDIA Quantum InfiniBand or RoCE fabrics. This ultra-low-latency fabric allows hundreds of GPUs to operate as a single, massive accelerator, enabling efficient data transfer with GPUDirect and maximizing performance. One-click GPU clusters can be set up on-demand and are ideal for multi-node training runs and high speed fine-tuning.



- Serverless Kubernetes: Fully managed, autoscaling Kubernetes service optimized for AI/ML workloads that abstracts away node pools, load balancers and cluster configuration. It delivers instantaneous cold starts, scale-to-zero flexibility and a dedicated control plane so you can fully utilize the power of Kubernetes, without any of the complexity.

For more information, please see the Appendix section on the Ori Cloud.

5. Cluster operating system



The AI Cluster Operating System is the core abstraction and execution layer for the entire Ori AI Fabric. Its primary role is to virtualize and manage underlying physical segments, ensuring dynamic secure and efficient delivery of GPU resources to multiple concurrent tenants. It harmonizes compute, storage and network resources beneath a single control plane.

5.1 Dynamic workload allocation

This component is the executive function of the orchestration layer, responsible for real-time resource adjustment and workload placement across the federated cluster. It works in conjunction with the Scheduler & Orchestrator to ensure high utilization and service availability.

- **Functionality**: The system dynamically manages nodes, allocating them across different services (Bare Metal, Virtual Machines, Containers) based on demand. It features advanced logic to shift node capabilities (e.g., from VM-focused to container-focused workload configurations) to minimize customer impact from failures or surges in demand.
- **Intelligent placement**: It utilizes logic for bin-packing and Topology-Aware Scheduling that models the network fabric itself (NVLink, InfiniBand or RoCE) to place distributed training jobs on nodes with the fastest inter-node communication.
- **Fault tolerance**: The Node Manager element continuously monitors compute storage and networking usage. In case of a node failure, it facilitates fast and automated recovery by detecting the fault and reallocating the associated compute resources to healthy nodes, ensuring job resilience and maximum cluster uptime.

5.2 Data / storage layer

AI training and inference consume massive datasets. Without high-throughput storage, GPUs sit idle - extending training times. Ori's AI Fabric storage layer is designed to support the following requirements:

- **High data throughput**: A storage system must deliver **several gigabytes per second (GB/s) of sustained read/write bandwidth** to keep GPUs busy. Scale-out NVMe flash arrays connected by high-bandwidth fabrics ensure data moves as fast as compute can consume it.
- **Low latency**: Distributed training and real-time inference require micro-second-scale I/O latency. Bypassing CPU memory via **GPU Direct Storage** reduces latency by transferring data directly from storage to GPU memory.
- **Scalability**: Datasets grow from terabytes to petabytes. A distributed, scale-out architecture spreads data across many servers and eliminates single bottlenecks. When more capacity or throughput is required, additional storage nodes can be added seamlessly.
- **Multi-tenant isolation**: Multiple teams need to share the same storage infrastructure. Each tenant must have dedicated namespaces or volumes with robust authentication and authorisation.
- **Storage agnosticism**: The platform is designed to integrate with a complete range of industry-leading solutions including parallel file systems (like Weka, Vast Data, DDN) and object stores (MinIO, Ceph).
- **Performance acceleration**: The entire data path is optimized using GPU Direct Storage (GDS) and network technologies to bypass the CPU. This enables GPUs to receive data at extremely high throughput and low latency, ensuring models are continually fed and accelerating multi-node training. It is scalable multi-petabyte datasets running parallel workloads on dozens of GPUs. This enables direct and efficient data transfer from the storage and network interfaces to GPU memory.

- **Hierarchical performance**: The storage module manages a tiered storage model optimized for performance across all major AI pipeline phases:
  - **Hot Tier (Parallel FS)**: Optimized for metadata-intensive and quick checkpointing tasks, this tier delivers ultra-low latency and millions of IOPS over high-speed interconnects (InfiniBand or RoCE). These file systems are built on NVMe flash and comply with NVIDIA DGX SuperPOD requirements for resilience and speed.
  - **Massive Capacity Tier (Object Store)**: This is the primary solution for AI throughput and scalability. It uses S3-compatible object storage (e.g., MinIO) built on NVMe flash and is optimized for massive sequential bandwidth (TB/s). This ensures a steady stream of large training datasets and Retrieval Augmented Generation (RAG) objects over the high-speed Ethernet fabrics.
  - **Local Storage (NVMe)**: Used for host-level caching, data shuffling and temporary checkpointing. This capability is sourced from local NVMe scratch drives on each compute tray to provide the lowest latency, off-network scratch space for performance-critical stages.

- **Orchestration & segmentation**: The data storage layer leverages Kubernetes CSI (Container Storage Interface) drivers for secure volume provisioning. Storage is isolated at the volume and namespace level with encryption at rest and policy-based access controls to enforce strict tenancy.
- **Alignment with compute**: To get the maximum storage performance, it is critical to make sure the storage solution is aligned with appropriate units of compute. For example, if you choose NVIDIA DGX GPUs, Ori will work with the customer to ensure that the storage matches NVIDIA SU (scalable units) which helps scale storage performance as compute is scaled.
- **Storage tiering & cost optimisation**: The storage module can manage tiering across NVMe flash, QLC SSDs and object storage to reduce cost while meeting performance goals. Hot data (e.g., training datasets) sits on NVMe; warm data (e.g., inference models) sits on SSD; cold data (e.g., archives) resides on object storage.Compression, deduplication and erasure coding further reduce storage cost.

Implementation tip: Underpowered storage can seriously affect your system performance despite using the best-in-class compute. Consider using parallel storage systems or ultra-high performance object storage (MiniO) for the highest level of storage performance especially for training workloads where performance, resilience and consistency are a key requirement for robust checkpointing.

5.3 Scheduler & orchestrator

This is the core intelligence layer that turns raw compute into an efficient service. It acts as the central engine for the Automation and Orchestration Engine (AOE).

- **GPU awareness**: Unlike stock schedulers (Kubernetes or Slurm), the Ori Scheduler is natively GPU-aware, providing primitives like bin-packing, NUMA optimization and secure MIG (Multi-Instance GPU) segmentation.
- **Resource allocation**: It supports flexible allocation types including full node, NVLink-level segments and fractional sharing. This allows the platform to assign exactly the right capacity per job, cutting idle spend and securely running up to seven isolated workloads per GPU.
- **Multi-domain orchestration**: It is capable of orchestrating complex workflows across multiple environments: bare-metal (Supercomputers), virtualized (VM Instances) and containerized (Serverless Kubernetes). This allows the platform to run training and serving workloads concurrently on the same hardware fleet.

5.4 Multi-tenancy

Ori's Secure Multi-Tenancy framework enables multiple teams or customers to share AI clusters without compromising isolation, performance, or governance. Ori AI Fabric provides three tenancy modes: Soft, Strict, and Private each balancing resource efficiency, performance consistency, and isolation:

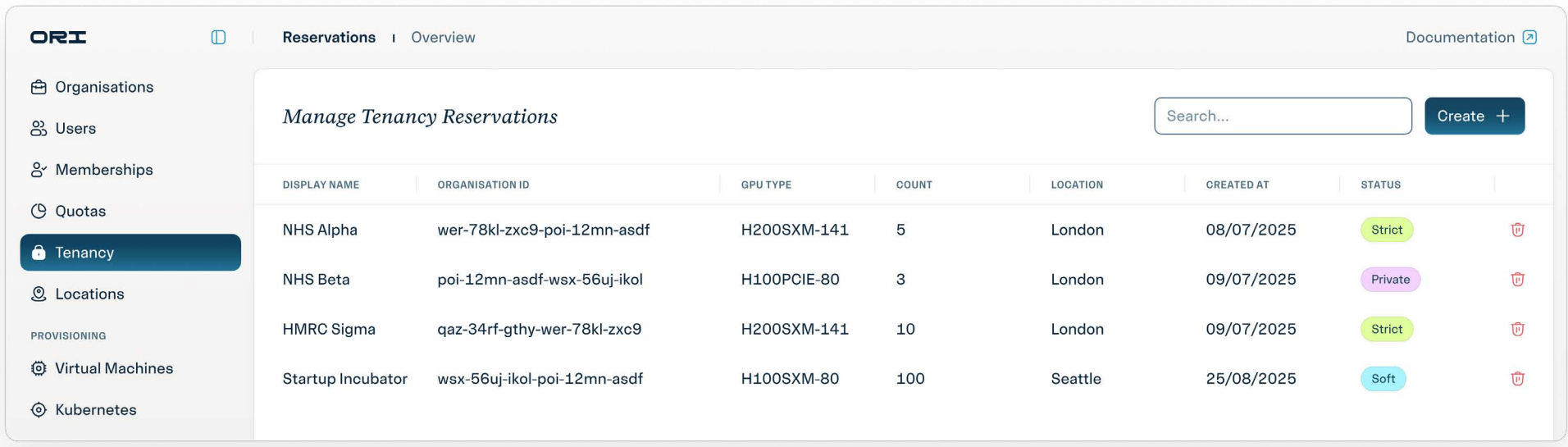
Mode	Description	Benefits	Ideal for
SOFT	GPU nodes are shared between tenants, with namespace separation for isolation	Maximizes GPU cluster utilization for platform operators, while giving end-customers the ability to granularly control their compute usage	Serving different teams or projects within a single organization where cost-effectiveness is a priority
STRICT	Entire GPU node(s) are reserved for a tenant with ring-fenced compute, storage and NVIDIA Infiniband networking	No workload sharing across tenants on nodes and consistent performance without the "noisy neighbor" effect [1]	Serving distinct external customers or housing sensitive, business-critical workloads
PRIVATE	Dedicated, single-tenant deployment along with full platform management capabilities	Ultimate level of isolation and data privacy with control over the underlying management of the private location	For organizations with strict data residency or air-gapped regulatory requirements

1. Techtarget, What is noisy neighbor (cloud computing performance)?

Maximize resource utilization securely

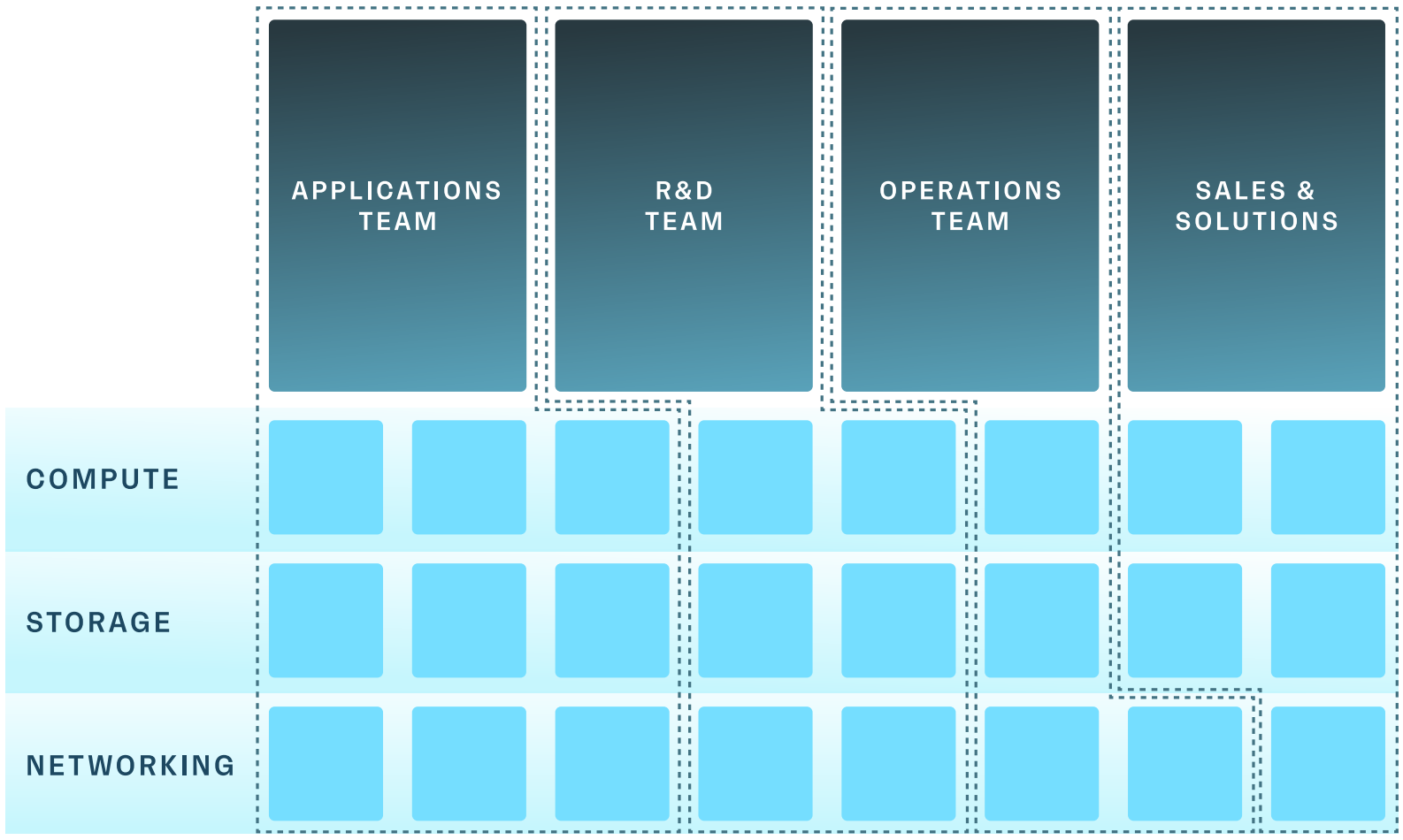
At its core, each tenancy mode is designed to enable customers to choose the right level of isolation for their workloads. Strict and Private tenancy provide advanced controls for data protection and resource isolation:

- **Full-stack isolation**: Strict and Private tenancy modes provide complete isolation across compute, storage, and networking for each tenant.
- **Restricted access**: Resource allocations in the strict tenancy mode are restricted to only certain team(s)/tenant(s) so that their data, apps and models are theirs alone.



- **Enhanced data sovereignty**: Locations under Private Tenancy are visible only to a single tenant and have a dedicated cloud operating system that manages scheduling and resource management for that location.
- **Flexible cluster partitioning**: Enterprise AI initiatives are often distributed across teams and geographies, increasing the need for sharing of infrastructure, but in a secure way so that every team can keep their models, apps and data private. The diagram below shows how Private Tenancy allows organizations to partition or re-partition their AI clusters among teams or use it for different purposes (training, inference) etc, so they can make the most of the cluster.

Resource distribution in Private Multi-tenancy



5.5 Cluster Utilization

This component is the business logic engine that measures and actively optimizes the cluster's efficiency. Designed to achieve maximum utilization, the module is entirely focused on eliminating idle time and fragmentation.

- **Resource reclamation**: The platform supports suspend and resume capability for VMs and Serverless Kubernetes clusters with the goal of delivering elastic scale without waste. This allows them to scale down to zero when idle and reclaim resources during inactive phases.
- **Fragmentation reduction**: Fractional sharing (MIG) and node-level bin-packing minimize idle stranded capacity.
- **Performance metrics**: The module ensures that every action is metered and traceable, providing the payment APIs and usage telemetry needed to connect with external billing systems and accurately charge back customers based on minute-level or token-based usage.

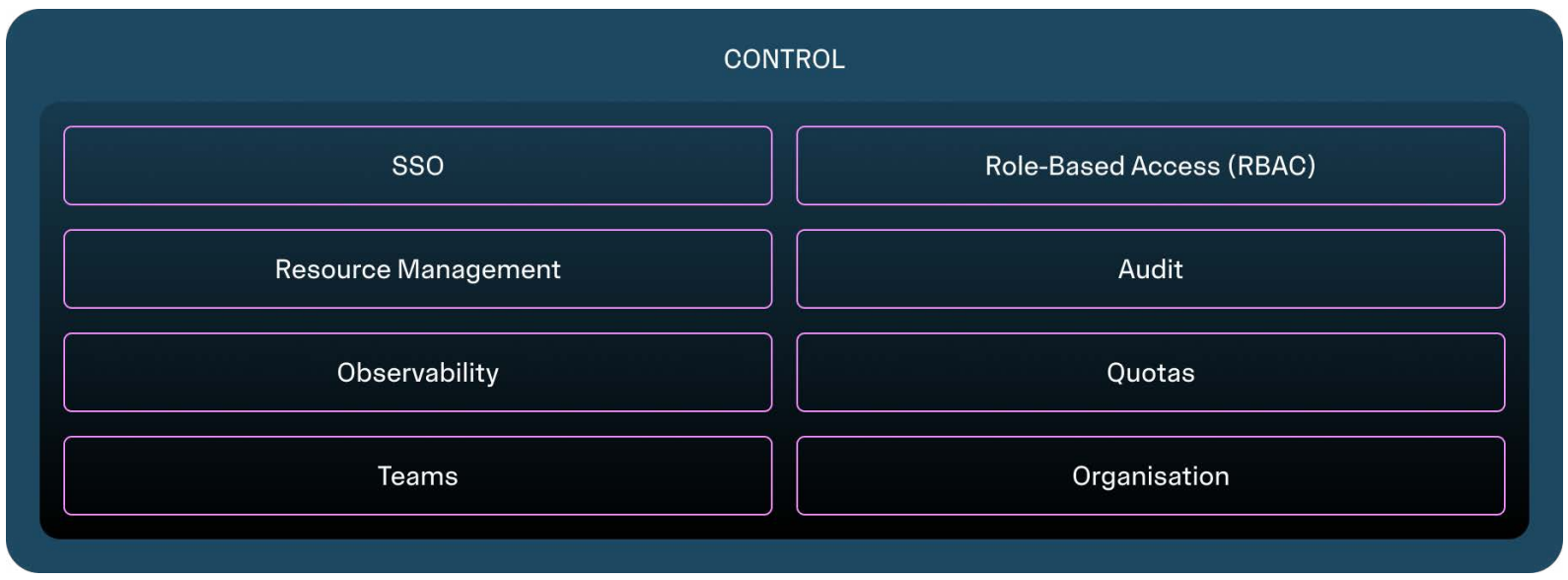
5.6 Inference Delivery Network

This is the architectural function responsible for routing, distributing and caching model artifacts for consumption. It is required for any global, low-latency AI-as-a-Service.

- **Architecture**: It manages a multi-region Inference Delivery Network (IDN) that caches and routes models for fast compliant and location-aware inference across federated sites.
- **Model distribution**: The system pulls models directly from the Model Registry, automatically distributing them to the point of compute—whether in the public cloud or a customer's private infrastructure—for fast compliant and location-aware inference.
- **Latency optimization**: It supports high-throughput low-latency serving via both Dedicated Endpoints (for predictable performance) and Serverless Endpoints (with autoscaling to zero for cost efficiency). The architecture is proven to achieve cold starts from zero in just a few seconds.

See section 3.4 Inference Delivery: Dedicated and Serverless Endpoints above for additional detail.

6. Control module



The Control layer is the primary authority for ensuring security, compliance, financial accountability, and fair resource distribution across all tenants and infrastructure segments. It manages identity, policy and usage telemetry for the entire platform.

6.1 Single Sign-On (SSO)

This module manages the federated identity lifecycle for all users and services across the Ori platform. It provides a single, secure authentication gateway, typically based on Keycloak, which is the core of the Ori IAM & Federation Hub. It supports external identity providers via SAML and OIDC standards. The SSO module integrates directly with enterprise directories (e.g., Azure AD, Okta LDAP) to enforce authentication policies like Multi-Factor Authentication (MFA) and passwordless access.

This approach centralizes security enforcement, guaranteeing all access is authenticated before being routed to the Role-Based Access Control (RBAC) engine.

6.2 Role-Based Access Control (RBAC)

The RBAC module defines and enforces authorization policies across all layers of the AI Fabric.

- **Role structure**: The system supports three built-in roles: Owner (full access to all resources and billing), Editor (manage resources but not billing) and Viewer (read-only access to resources and billing). These roles can be supplemented with custom roles that include granular permissions.
- **Authorization scope**: The module applies authorization policies down to the individual resource level (e.g., specific GPU pools model registries inference endpoints). This ensures every user and service account adheres to the principle of least privilege.
- **Integration and security**: Policies are enforced by the Control Plane and tied directly to the identity asserted by the SSO module. Service accounts and API tokens inherit these roles, have limited scopes and are subject to regular rotation to maintain secure auditable access.

6.3 Resource management

This component provides the operational tools necessary for provisioning and controlling all infrastructure assets used by the platform. The resource manager manages the lifecycle (creation configuration scaling) of bare-metal VM and containerized compute resources. It utilizes templates for GPU-backed instances and training clusters. It integrates with Infrastructure-as-Code (IaC) tools like Terraform and Ansible for declarative deployments.

The resource manager interfaces directly with the Automation and Orchestration Engine (AOE) to translate portal or API requests into physical or virtual resource allocation and configuration workflows.

This enables self-service provisioning for customers while giving administrators the granular control needed to manage capacity and maintain consistent configuration baselines across the federated estate.

6.4 Audit Module

The Audit module maintains the immutable record of all activity and state changes within the Ori platform, serving as the foundation for security and governance. It logs every action performed by users API tokens and system agents into a dedicated, immutable PostgreSQL-backed event store. Each log is enriched with contextual metadata (timestamp resource affected user/token).

Logs are streamed in real-time to the Ori Observability Framework and are made available via API for integration with external SIEM (Security Information and Event Management) and SOAR systems (e.g., Splunk Elastic).

This provides full, non-repudiable traceability required to meet stringent compliance standards (ISO 27001 SOC 2 GDPR, SOC 2 HIPPA) and facilitates rapid root-cause analysis for security and operational incidents.

6.5 Observability

This component aggregates and visualizes health performance and usage telemetry across the entire AI Fabric platform. It collects real-time metrics (CPU, GPU, network, I/O) via standard agents (Prometheus and OpenTelemetry) and centralizes structured logs. It provides Grafana dashboards for visualizing system health, model performance (latency loss functions) and cost metrics.

Telemetry is gathered from the Infrastructure Layer (via vendors like NVIDIA UFM) the Operating Layer (VMs containers) and the AI Software Layer (MLOps pipelines). The data is then used to power the Alerting and Reporting modules.

This facilitates proactive management by giving SREs and customers end-to-end visibility into service-level objectives (SLOs) and quickly diagnosing distributed system failures.

6.6 Quotas

The Quotas module ensures fair sharing of finite GPU and storage resources across all tenants and projects. It defines and enforces hard or soft resource limits at multiple levels (user project organization). Limits are applied to compute (GPU hours, CPU vRAM) and storage capacity.

The Quota Manager is tied to the billing engine which tracks consumption on a per minute basis. When a soft limit is approached an alert is triggered; when a hard limit is hit, the Automation and Orchestration Engine (AOE) is signaled to suspend or deny further resource provisioning for that entity.

This guarantees service fairness and prevents resource saturation, providing essential cost governance by enforcing planned capacity allocation.

6.7 Teams

This component manages the collaborative and hierarchical organization of users accessing the platform. It defines and manages multi-user Organizations and Projects, creating isolated infrastructure workspaces, providing the mechanism for Organization Owners to delegate and manage access for users and groups within their scope.

The Team structure forms the top level of the RBAC and Quota hierarchy. All policies are applied per team or per organization before being enforced at the resource level.

This streamlines multi-tenant operations, ensuring strict data and resource isolation between customer organizations and providing a clear framework for chargeback and internal governance.

6.8 Organizations

The Organisation module provides the administrative tools for high-level customer and vendor management. The module manages vendor-level and large enterprise contracts, service catalogs and jurisdictional policy enforcement (e.g., Sovereign Control Suite), which enables platform administrators to customize SKUs availability and pricing per tenant.

The module integrates with external ERP/Billing systems (via APIs) with the platform's internal Metering Engine to align usage with financial contracts. It also integrates with external IAM systems for customer identity federation.

This module acts as the strategic control point for commercial relationships, legal compliance and large-scale federated deployments, ensuring infrastructure services align with business objectives and regulatory requirements.

7. FinOps, billing & chargeback

The financial framework of the Ori AI Fabric is a core differentiator, moving beyond simple hourly metering to provide granular, auditable and multi-modal consumption tracking essential for both internal financial accountability and external customer billing. This system ensures that infrastructure cost directly aligns with business value delivered, optimizing the Total Cost of Ownership (TCO) for expensive GPU assets.

7.1 Granular metering and consumption models

The system achieves financial precision by tracking usage across every dimension of the AI lifecycle, enabling flexible pricing strategies that support multiple business models:

- **Fine-grained metering**: The platform utilizes the Ori Metering Engine to collect usage data at a fine granularity, primarily minute-level for most resources. For complex services like Serverless Kubernetes, metering is broken down per GPU, CPU, memory and storage unit.
- **Multi-modal billing**: Consumption is tracked across all service modalities to support diverse customer needs:
  - **GPU-as-a-Service (GPUaaS)**: Tracks bare-metal, virtual machine (VM) and container compute usage based on GPU hours and CPU cycles.
  - **Token-as-a-Service (TokenaaS)**: For LLM inference, usage is billed per token or per request, ensuring customers only pay for the immediate value consumed.
  - **Job-based accounting**: Metering is tied to execution, capturing the resource allocation, duration and even model-specific metrics (e.g., loss function progression, training epochs) for complex batch workloads.
- **Elasticity and cost control**: The framework directly supports resource reclamation features like the suspend and resume capability for VMs and Serverless Kubernetes clusters. This ensures billing stops immediately when resources scale down to zero, directly transforming utilization efficiency into predictable economics.

7.2 Quota enforcement and financial governance

The financial tools are tightly integrated with the Control Layer's governance primitives to ensure fair resource distribution and prevent costly runaway workloads.

- **Quotas and limits**: The Ori Quota Manager defines and enforces hard or soft resource limits at multiple levels (user project organization). Limits are applied to compute (GPU hours CPU vRAM) and storage capacity.
- **Cost attribution and chargeback**: Every billing event is tagged with rich attribution metadata (tenant, project, user, API token). This ensures that usage is auditable and can be accurately charged back to specific business units or customers, providing the financial transparency required for enterprise-grade operations.
- **Flexible billing integration**: The Ori Billing Manager aggregates metered usage, applies customer-specific pricing models (flat-rate, tiered, consumption-based) and supports multi-currency operations and tax logic. It exposes REST APIs and event streams to seamlessly connect with external enterprise systems (ERP, BI, CRM) for final invoice generation and financial reporting. You can leverage Ori's default integration with Stripe or plug in your own billing system for seamless integration with the platforms your organization uses.

7.3 Observability and auditability

Financial data is treated as a critical security and observability asset, ensuring compliance and operational transparency.

- **Immutable audit trail**: All usage and billing events are logged immutably in a dedicated event store, ensuring full auditability required for regulatory compliance. Role-Based Access Control (RBAC) is enforced over financial data.
- **Integrated reporting**: The Ori Reporting Hub provides self-service customer dashboards with real-time usage, historical consumption trends and cost forecasts. This aligns financial accountability with technical observability, feeding financial metrics directly into the Grafana-based monitoring stack.
- **AI-enabled FinOps**: The system provides advanced AI-specific financial insights, including per-model cost tracking, token usage heatmaps for LLM workloads, and planned AI-driven anomaly detection to flag unusual billing activity or potential runaway jobs. This capability optimizes operations by correlating technical metrics with financial risk.

8. Networking layer

For large AI projects, networking becomes as critical as compute. An AI-optimized network must deliver hundreds of gigabits per second of throughput, micro-second latency and non-blocking traffic. Distributed training requires frequent all-reduce operations - without low latency, GPUs idle waiting for parameters to synchronise. Inference at scale requires a network fabric that can handle thousands of concurrent requests without congestion.

8.1 Multi-fabric network design and performance

Ori separates networking into specialized, non-blocking fabrics that adhere to NVIDIA's DGX SuperPOD reference architecture, ensuring high throughput and microsecond latency for all workloads.

- **Compute Fabric (GPU Interconnect)**: This dedicated InfiniBand or RoCE Ethernet leaf-and-spine network isolates latency-sensitive peer-to-peer traffic (e.g., all-reduce operations) on its own non-blocking fabric. Ori implements this using NVIDIA Quantum-series switches and ConnectX NICs to deliver up to 3.2 Tbit/s of RDMA-accelerated traffic, maximizing distributed training performance that scales linearly with added nodes.
- **Storage Fabric**: A separate, high-speed network connects compute nodes to parallel file systems and object stores. This segregation prevents dataset transfers and checkpoint I/O from creating congestion with inter-GPU communications, enabling tuned throughput for terabyte-scale data transfers.
- **Management Fabrics (In-Band and Out-of-Band)**: Two distinct Ethernet networks manage administrative flows. The In-Band Management Network carries job scheduling monitoring registry access and orchestration traffic. The Out-of-Band Management Network is physically isolated (typically 1 GbE) to link server BMCs and switch service ports for power control, firmware updates and health monitoring, ensuring uninterrupted hardware access even under full AI-training load.

Interoperability and governance

Ori clusters follow validated DGX SuperPOD blueprints, which guarantees predictable performance and reliability at production scale. The platform also natively bridges to standard Ethernet (including RoCE) at cluster edges through gateways or dual-connected nodes, combining InfiniBand's ultra-high performance with Ethernet's broad compatibility. This allows the seamless integration of an Ori cluster without overhauling the datacenter fabric.

Networking is fully integrated with the compute and orchestration layers. The scheduler job orchestrator and data-management services are topology-aware, optimizing workload placement and delivering unified dashboards that expose network throughput and latency alongside cluster utilization.

Implementation tip: Virtual Private Clouds (VPCs) provide isolated network environments that allow cloud deployments to securely communicate with each other and the internet. They enable the creation of segmented, secure networks within the Ori Cloud Platform, giving organizations precise control over traffic flow, resource access and data protection, critical for teams operating in cloud environments with strict security and compliance requirements.

9. Support

Ori considers its support to be a differentiating feature of the product and we treat it with the same seriousness as we do the software. This support mentality was born from the public cloud product and we have built the organization to deliver 24/7/365 support to ensure the secure and uninterrupted operation of the deployed platform. Support is available via email, Slack and our helpdesk portal.

Under our standard Shared Responsibility Matrix, Ori is responsible for:

- Availability and maintenance of the Ori Cloud Platform
- Security patching and software upgrades
- Platform monitoring and incident response
- Role-based access control enforcement
- Backup configuration for metadata and critical configs
- Optional: advisory services for physical infrastructure for customer-owned infrastructure

If a different operational model is preferred, Ori is happy to work with the Customer to define a tailored approach.

9.1 SLAs

For fully managed deployments, Ori provides service-level agreements (SLAs) covering availability, performance and support response times.

- Ori commits to a minimum 99.9% service availability for the Ori Cloud Platform, calculated on a monthly basis.
- When we have to interrupt the availability of the Ori Cloud Platform to undertake scheduled maintenance the customer receives advance notice of seven days and we strive to provide the best possible notice for emergency maintenance.

10. Implementation considerations

10.1 Hardware selection

Selecting appropriate hardware depends on workload characteristics, cost constraints and scalability goals. Considerations include GPU type (Blackwell vs Hopper), GPU memory capacity, NVLink connectivity, CPU configuration (x86 vs ARM with integrated GPUs), memory bandwidth, local NVMe capacity and network adapter speeds (400 Gb/s vs 200 Gb/s). Interoperability between vendor ecosystems (NVIDIA vs AMD vs other accelerators) should be validated.

10.2 Site requirements

AI clusters consume significant power and cooling. Plan for sufficient power density (e.g., 30–60 kW per rack), water or air cooling (direct liquid cooling may be required for dense clusters), and floor space. The network design should support the number of network ports and cabling for multiple fabrics. Redundant power feeds and backup are essential for high availability. In hybrid deployments, plan for high-capacity cross-site links (100 Gb/s or higher) to handle bursts of traffic.

10.3 Security hardening

Baseline hardening includes disabling unused services, applying secure configuration templates, enforcing key rotation, scanning container images for vulnerabilities, enabling secure boot and BIOS passwords, monitoring firmware integrity and implementing network security groups. Regular penetration tests and vulnerability scans should be performed. Audit logs should be reviewed regularly.

10.4 Operational expertise

Running an AI factory requires expertise in GPU clusters, high-performance storage, RDMA networks, Kubernetes and MLOps. Consider training internal teams or partnering with Ori's support and professional services. Build cross-functional teams involving platform engineers, data scientists, security engineers and financial managers. Document runbooks for incident response.

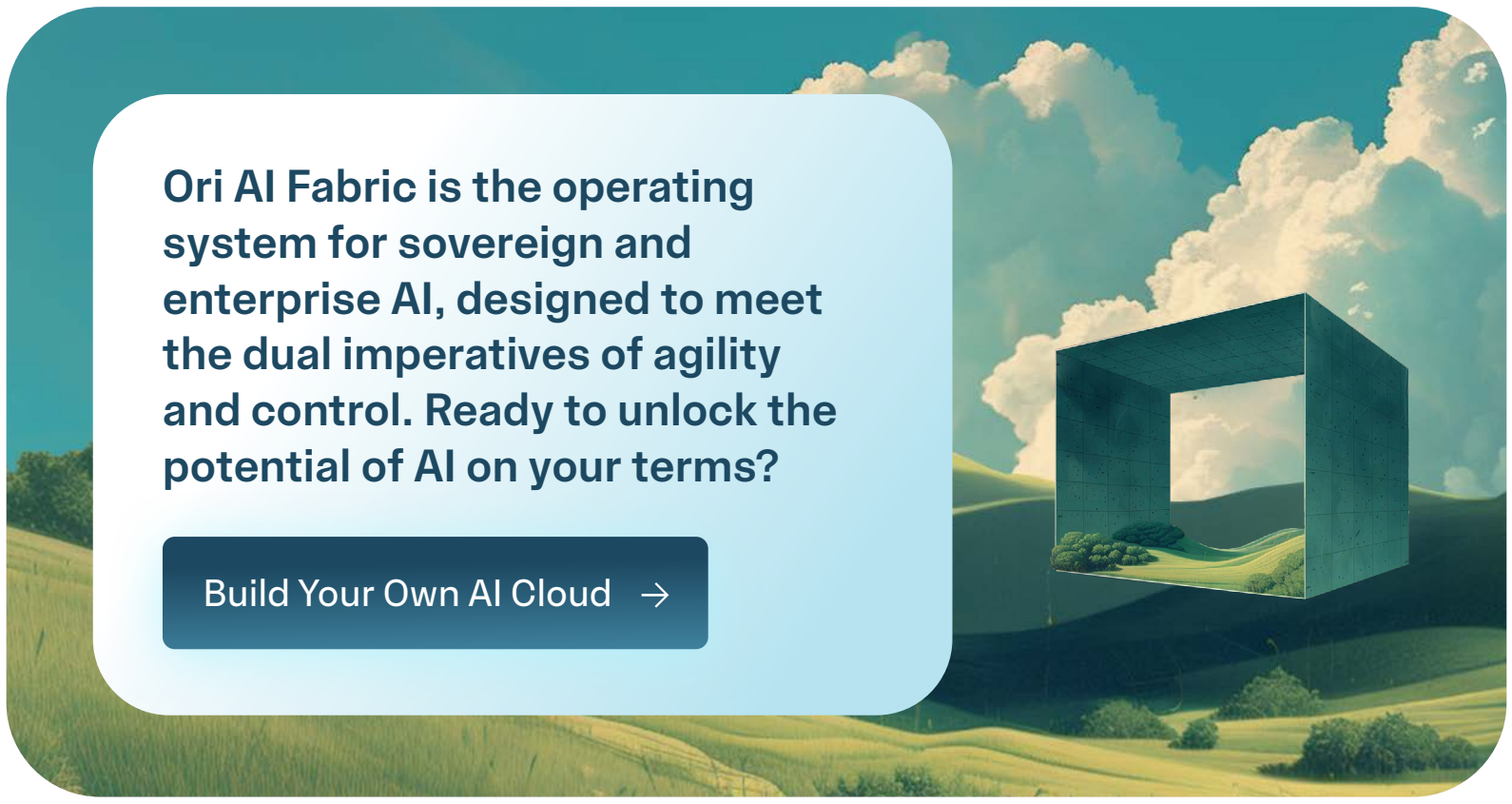
10.5 Branding customization

When customers license Ori AI Fabric, they can fully tailor the platform's look and feel to match their brand identity. From uploading their own logo to applying custom colors, the platform can be styled to reflect their brand.

11. Conclusion

Ori's AI Fabric platform delivers an unparalleled convergence of cutting-edge compute, storage, networking, orchestration and MLOps tooling under a single, unified control plane. This reference architecture has demonstrated how enterprises can:

- **Achieve true sovereignty** by deploying air-gapped or fully on-prem clusters with hard multi-tenancy and region-aware scheduling, ensuring data residency, encryption key control and audit-grade isolation meet the strictest regulatory requirements.
- **Scale without compromise** via modular "Scalable Units" that blend bare-metal GPU clusters, virtual supercomputers, and serverless Kubernetes, all backed by ultra-low-latency NVMe storage and non-blocking InfiniBand/RoCE fabrics. Whether you own the hardware or burst into Ori's managed cloud, performance grows linearly—so your largest AI workloads never outpace the infrastructure.
- **Govern every layer** with a control plane that enforces SSO/2FA, hierarchical RBAC, resource quotas, policy-driven autoscaling and full audit logging. Operators gain real-time FinOps visibility, cost-allocation reports and automated policy enforcement, eliminating shadow IT and runaway cloud bills.
- **Accelerate AI innovation** through integrated MLOps services—Model Registry, Fine-Tuning Studio, Dedicated and Serverless Inference Endpoints—so ML teams can iterate from prototype to production in hours, not weeks.
- **Bridge hybrid & sovereign islands** with a consistent operational model that spans customer-owned clusters, Ori's global regions and air-gapped governmental deployments. Secure VPN links stitch sites together, enabling seamless workload bursting, data synchronization and unified governance.

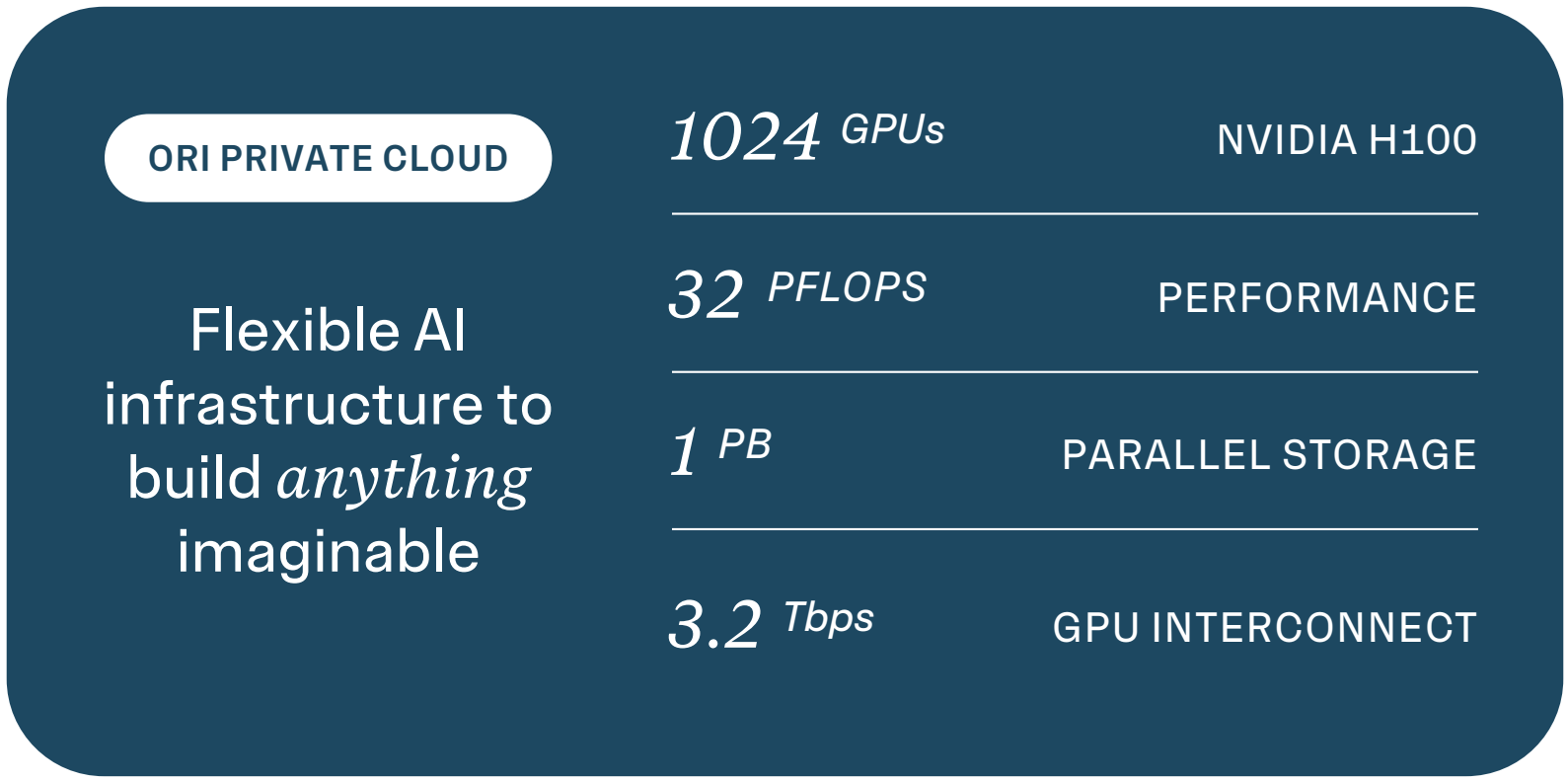


Appendix: Ori Global Cloud

A proven production environment

Ori Cloud is the operating environment that validates every aspect of the Ori AI Fabric platform. It is not a separate system but rather the global, at-scale deployment of the very same software stack that licensed customers adopt. This dual role is important: Ori Cloud demonstrates the resilience, performance, and operational maturity of the Ori AI Fabric in real time, while also providing the benchmark by which customers can measure their own deployments.

To ground the architecture in a concrete example, the following are published details of a 1024 H100 GPU cluster built for a large AI lab.



The cluster includes:

- 128 HGX nodes with 8 NVIDIA H100 SXM GPUs each.
- 1 Petabyte of Weka parallel storage delivering an aggregate of 9 million read IOPS and 3 million write IOPS.
- 3.2 Tbps NVIDIA InfiniBand GPU interconnect and 400 Gb/s Ethernet storage fabric.
- High-speed external connectivity and data mobility between sites.

The cluster achieved 32 PFLOPS sustained performance over several days (≈95 % of theoretical peak) and now powers large-language model training and inference. This implementation demonstrates Ori's ability to assemble supercomputers-class private clouds using modular building blocks. By following the reference architecture presented in this document, enterprises can replicate similar designs tailored to their workloads.

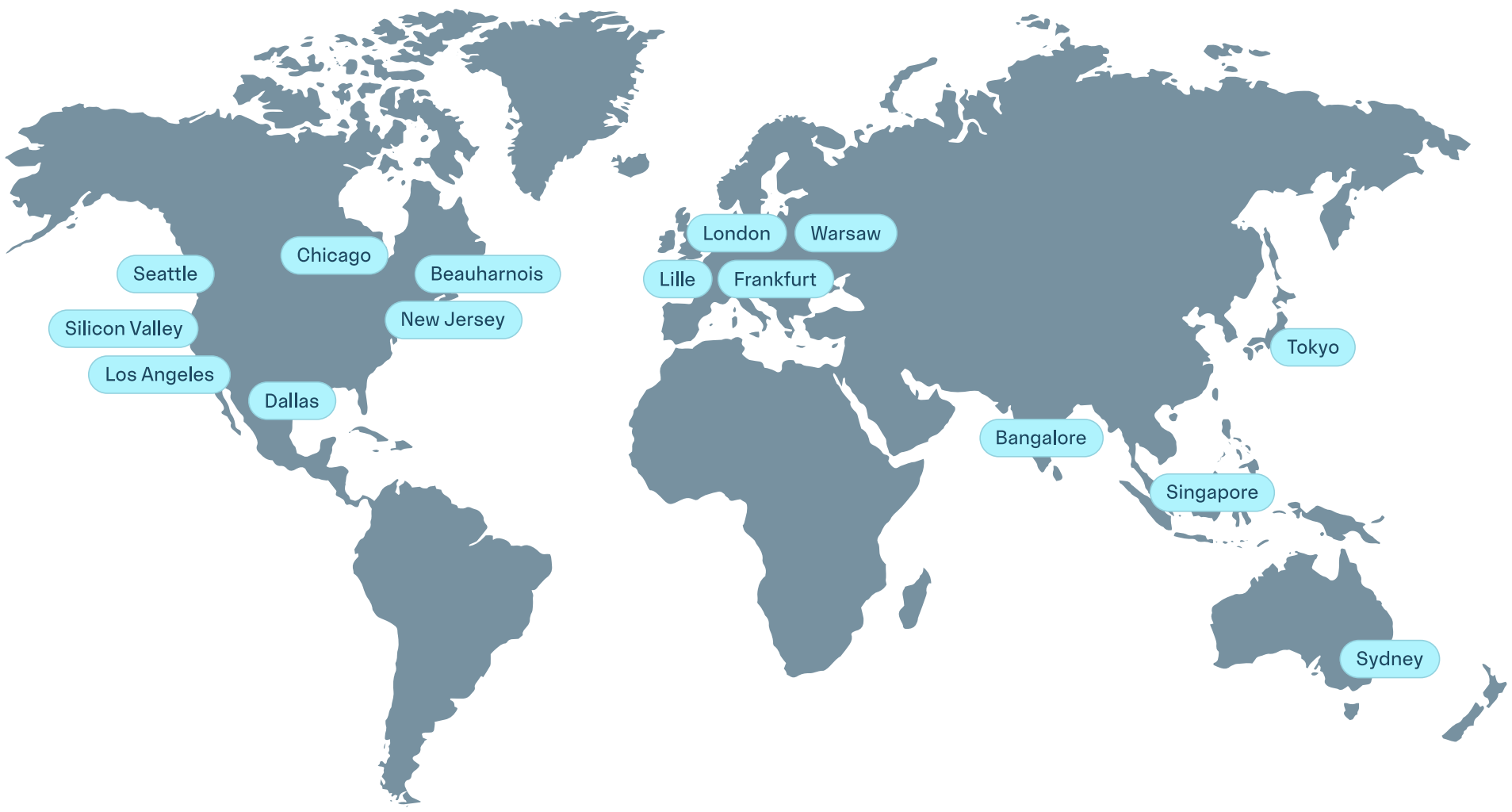
Simplicity without compromise

At its core, Ori Cloud is designed to remove complexity for AI practitioners. Bare-metal GPU performance is delivered without the friction of hypervisor layers or fragmented scheduling systems. Users interact with a coherent stack—compute, storage, and orchestration integrated by the AI Fabric—that translates into predictable performance and a rapid path from idea to deployment. What customers experience in Ori Cloud is the distilled essence of features outlined in the reference architecture but operated at global scale.

Global reach and elasticity

AI workloads are increasingly distributed. Ori Cloud addresses this by operating across a global footprint of data centers, ensuring that capacity is available where customers need it. The platform supports elastic consumption models: from burstable GPU clusters for experimentation, to long-running, distributed training jobs that demand consistency across thousands of accelerators. This elasticity is made possible by the AI Fabric's scheduling and orchestration logic, proven in daily operation within Ori Cloud.

Ori's Global Footprint



Ori has a network of 15 regions across the globe, with more being added soon. If your choice of region is not covered currently, we can consider deploying a new region based on the scale of infrastructure needed.

Depth of Services

Beyond raw compute, Ori Cloud offers a complete spectrum of AI-as-a-Service capabilities built natively on the AI Fabric:

- **Inference-as-a-Service** with low-latency edge distribution
- **Training-as-a-Service** optimized for multi-node, multi-GPU scaling
- **MLOps-as-a-Service** for lifecycle management, monitoring, and retraining
- **Data Services** including versioning, storage optimization, and bandwidth-aware caching

Each service is a direct extension of the reference architecture, hardened through constant use and iteration in Ori Cloud.

Completeness and support

Customers of Ori Cloud benefit from a full-stack offering: GPUs, storage, networking, orchestration, and monitoring delivered as an integrated service with enterprise support. This end-to-end approach is not only about convenience—it reinforces the credibility of the Ori AI Fabric platform. Licensed deployments can trust that the same software powering their environment has been stress-tested continuously in production.

Why this matters for licensed deployments

Ori Cloud is more than an offering; it is the evidence base for Ori AI Fabric's maturity. Every design choice in the reference architecture has been validated under real-world load, diverse customer demands, and global scale operations. Licensed customers inherit this maturity: a platform that is not theoretical, but proven daily in the crucible of cloud operations.

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 →

ORI

ori.co