XaasIO AI Token Factory

Private Inference, Air-Gapped Model APIs, Sandboxes and GPU Economics

Turn private AI infrastructure into governed model and agent services. XaasIO AI Token Factory brings together approved and custom models, a developer playground, isolated code execution and usage-based commerce. Connect GPU inventory and utilization to model cost and customer margin from one operator-controlled platform.



NVIDIA Rubin platform. Image credit: NVIDIA.

© 2026 XaasIO. All rights reserved.

PLATFORM OVERVIEW

Publish Private Model Services

Serve approved and custom models through OpenAI-compatible APIs. Support chat, streaming and embeddings where the model and runtime allow.

Run Agent Workloads

Connect agents to governed inference and isolated code execution. Apply project keys, resource quotas and usage attribution across the service.

Turn Usage into Services

Meter tokens, endpoint GPU time and sandbox resources. Use built-in commerce for versioned price books, plans, credits, rating and invoices.

DEPLOYMENT ARCHITECTURE

Private and Sovereign Infrastructure

Keep models and service controls on customer infrastructure. Define network paths, storage, resilience and data location around your workload.

Two Execution Runtimes

Run vLLM inference on GPU nodes and Agent Sandbox with Kata Containers on KVM-enabled CPU nodes. Manage both through Kubernetes.

Operator-Controlled Dependencies

Use local identity, S3-compatible storage and a container registry. Align networking, retention and release packaging with deployment requirements.

ONE CONSOLE FOR DEVELOPERS AND OPERATORS

Discover models, run agent code and connect measured consumption to service economics.

CORE TOKEN FACTORY CAPABILITIES

A guided developer experience, governed execution and the controls to operate commercial AI services.

Getting-Started Home and Playground

Guide developers from model discovery and project API keys to a first request. Try supported models in the playground before integration.

Custom Model and LoRA Upload

Upload validated weights and compatible LoRAs to a project-private catalog. Use dedicated endpoints for full models and named adapters for LoRA.

Agent and Code Sandboxes

Run agent tools and code in isolated microVMs. Use templates, persistent storage, terminal/API access, network rules and scoped secrets.

GPU Inventory and Economics

Track GPU inventory, allocation and utilization. Combine hardware, depreciation, power and overhead inputs to review model and customer margin.

Model Catalog and Endpoint Choices

Offer shared APIs, dedicated endpoints, reserved-throughput service profiles and batch jobs. Preserve model versions, licenses and approval records.

Identity, Tenancy and Governance

Use XaasIO IAM + SSO, resource permissions, scoped API keys and project quotas. Integrate configured input, output and tool policies.

Token and Sandbox Metering

Attribute token, GPU-time, CPU, memory and storage usage to tenants and projects. Durable usage events support reporting and billing.

Built-In Commerce

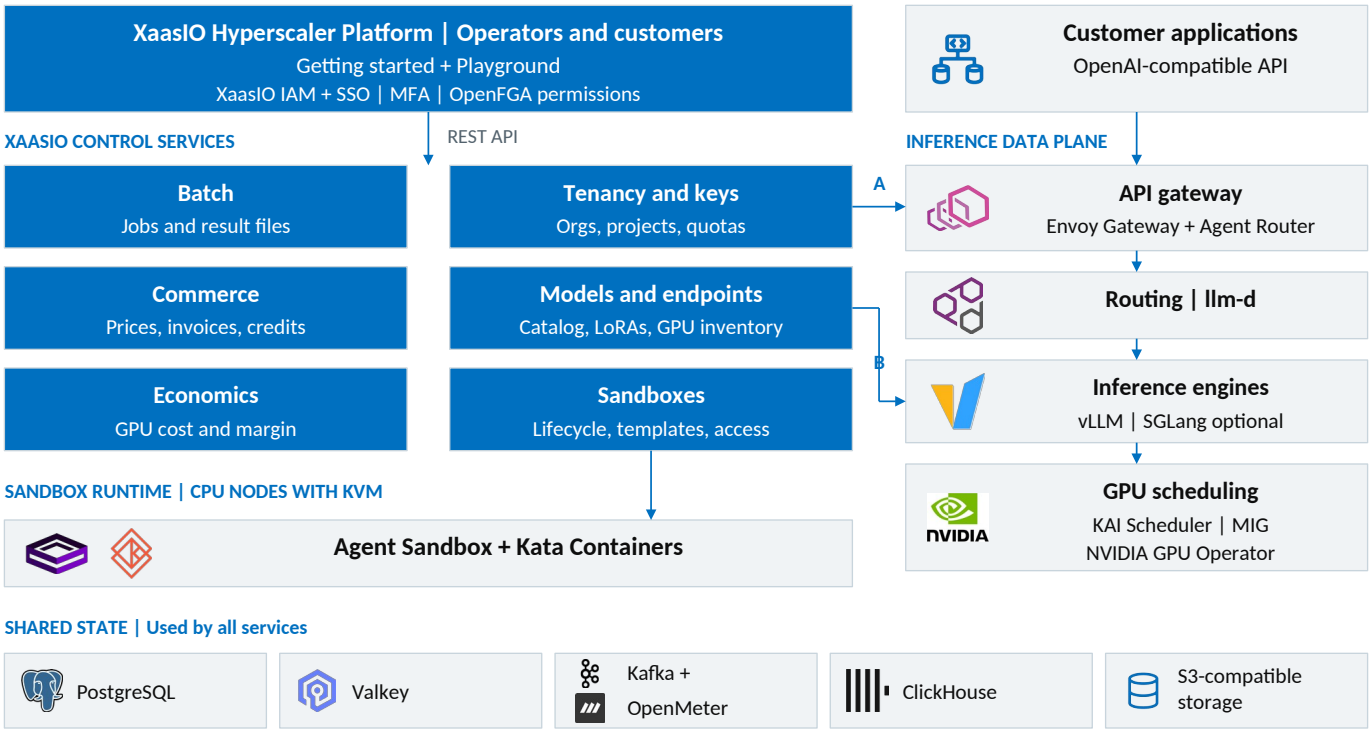
Package services with price books, plans, credits and invoices. Connect payment and tax adapters within the standalone platform.

AN INTEGRATED XAASIO PLATFORM EXPERIENCE

Use XaasIO Hyperscaler Platform for developer and operator access. Connect local identity, tenancy, usage and commerce services with upstream model pipelines and XaasIO AI Lake retrieval workflows.

© 2026 XaasIO. All rights reserved.

REFERENCE ARCHITECTURE



A: Local key and quota checks at the gateway. B: Model service deploys inference workloads. Usage flows through Kafka and OpenMeter into billing; OpenTelemetry, VictoriaMetrics and VictoriaLogs provide telemetry.

GPU INVENTORY AND ECONOMICS

Fleet and Endpoint Visibility

See GPU inventory and pool allocation alongside utilization, memory and hardware health. Correlate endpoint queues, token throughput and latency with capacity demand.

Cost and Margin by Model and Customer

Combine hardware cost, depreciation, power and overhead inputs with measured usage. Relate service cost and commercial revenue to model and customer margin.

OPERATIONS AND GOVERNANCE

Controlled Access and Execution

Use XaasIO IAM + SSO for sign-in and OpenFGA for permissions. Apply scoped keys and quotas; isolate code execution with Kata microVMs, network rules and secrets.

Lifecycle and Operational Evidence

Use the installer operator for deployment and upgrades. Observe service health, preserve usage and audit records, and validate model changes before rollout.

SERVICE PROFILES AND METERING

Profile	Typical use	Metering basis
Shared endpoint	Pooled model APIs for teams and apps	Input, output and cached tokens
Dedicated endpoint	Private replicas and custom models	Allocated GPU time; plan-based rates
Reserved throughput	Capacity for predictable demand	Reservation and excess per plan
Batch inference	Queued document or data processing	Tokens under the batch price book
Agent/code sandbox	Agent tools and isolated code execution	CPU time, memory time and storage

Profiles and pricing depend on the selected model, release, deployment and service plan.

© 2026 XaasIO. All rights reserved.

DESIGNED FOR ENTERPRISE AI SERVICES

01 Enterprise Models and Agent Development

Give teams a getting-started home, playground and governed model APIs. Upload approved custom models and LoRAs, then run agent tools and code with project limits and clear usage attribution.

02 NeoCloud and Sovereign AI Services

Offer shared APIs, dedicated endpoints, reserved capacity and sandboxes. Connect built-in commerce with GPU inventory, utilization, cost and margin to guide service pricing and capacity decisions.

03 Private AI for Sensitive Environments

Keep model stores, identity, inference and usage accounting within the required operating boundary. Apply permissions, network policies and controlled execution to local AI and agent workloads.

XAASIO AI TOKEN FACTORY

Connect GPU capacity to AI services, customer value and margin.

Start with your models, agent workloads, GPU fleet and data boundaries. XaasIO brings together guided developer access, governed inference, sandbox execution and service economics to shape your private AI platform.

Explore the Platform at xaasio.com/platform/ai-factory/

Book an AI platform workshop at xaasio.com/contact/

Learn more at xaasio.com

AMERICAS
+1 650 523 6628

EMEA
+44 20 3031 1074

APAC
+91 93423 61010

© 2026 XaasIO Group. All rights reserved. Information is subject to change. | xaasio.com

Third-party names and logos are trademarks of their respective owners.