

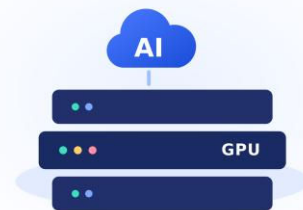
AI FACTORY ENGINE

Manage & Scale AI Factory with VEKTOR

AI factory is the novel unit of computing, and the token is its unit of value. VEKTOR is the full-stack engine that orchestrates GPU and AI workloads on one platform, so every operator delivers intelligence with **peak utilization** and **lowest cost per token**.

FLEET UTILIZATION
73%

COST / 1M TOKENS
\$4.20



IDLE GPUS RECLAIMED
42 GPUs

LUMI
[Ask the fleet...](#)



**Higher GPU
utilization**



**Lower cost per
token**



**Faster AI workload
delivery**



Neoclusters & GPU-as-a-Service



Colocation & DC operators



Enterprise AI platforms



High Performance Compute



AI model hosting



Sovereign & regulated AI

CAPABILITIES

Unified AI Factory Operations at Scale

VEKTOR unifies GPU, workload, token, and tenant operations into a single AI Factory platform – eliminating the sprawl of schedulers, marketplaces, spreadsheets, and disconnected dashboards.



GPU Farm Management

Run GPU farm at max output. Smartly orchestrate GPU, improve utilization and reduce capacity waste.



AI Workload Orchestration

Schedule training, inference, and HPC across pools, clusters, and models.



Token Farm/ AI Compute Marketplace

Meter the token. Monetize the compute. One platform turns GPU time into billable capacity.



GPU Observability & AIOps

Optimize performance, utilization, and cost across the estate for more tokens per watt and lower cost.



Revenue & Monetization

Turn utilization into revenue. Track usage, pricing, and margin across the estate in real time.



Multi-Tenant GPU Cloud Platform

Built for multi-tenant operation. Deliver isolated, governed, self-service GPU clouds at scale.



LUMI, the AI Copilot

Ask the fleet anything, then act on the answer. LUMI investigates incidents, surfaces the right levers, and executes the fix under human-in-the-loop approval.

PLATFORM

Built For The AI Era

Operationalize AI infrastructure at scale, with the controls, interoperability, and elasticity required to run GPU estates as reliable, monetizable business platforms.

Operator console

Fleet snapshot · Jun 11, 2026, 12:16 AM

5 of 5 sites healthy SLA 99.99% · 0 breaches 3 active alerts

GPU UTILIZATION

72.8%

1,118 of 1,536 GPUs in use

MEAN MFU

61%

Model-FLOPs utilization · target 60%

ACTIVE TENANTS

13

1,204 GPUs allocated

REVENUE · 30D

\$1664.9K

47.8% gross margin

POWER DRAW

542 kW

PUE 1.21 · WUE 0.18 L/kWh

Fleet topology · eastus2 · UnitedLayer-LAS01



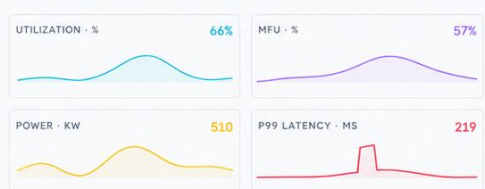
Cost-to-serve · live · Per 1M tokens (inference average)



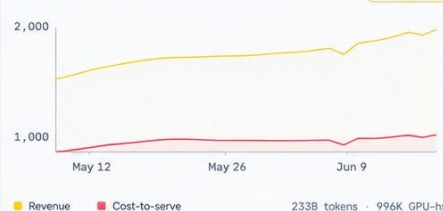
Active alerts · 5 items · last 2h

- ⚠ Rack B-12 · GPU thermal warnings on 3 nodes
2m ago · topology-heat-investigator
- ⚠ NCCL all-reduce regression on tenant aurora-research
14m ago · nccl-hang-triage
- ⚠ IB fabric port flap · spine-2 → leaf-7
38m ago · fabric-debug
- 📌 Idle GPU reclaim available · 42 GPUs · ~\$1,820/day savings
1h ago · idle-reclaim
- ⚠ GPU dgx-07-04 node-3 · Xid 79 (NVLink fatal)
2h ago · xid-error-decoder

Fleet · 24h trend · Utilization · MFU · power · P99



Revenue · 30d · cost-to-serve overlay · margin 50.7%



Workload mix · GPU-hours · 30d



Secure & Compliant

Guardrails, model security, RBAC, and MFA; endpoint security, compliance, and audit logs across every tenant.



Sustainable Ops

GreenOps built in. Sustainability metrics and smart optimization across the fleet.



Multi-Tenant & Interoperable

GPU provider federation, API-first control plane, deployable flexibly as on-prem software or SaaS.



Orchestrated & Governed

AI factory operations that adapt to your estate, deployed on-prem or as SaaS.

A GPU Farm Produces Compute. Vektor Makes It Produce A Business.

VEKTOR unifies GPU orchestration, AI workload management, token metering, monetization, and governance in one engine, so operators can scale AI infrastructure with control, utilization, and margin.

- Lower cost to serve
- Scale AI factory with control
- Deliver AI workloads faster
- Run secure multi-tenant AI infra

100+ INTEGRATIONS



Request a demo →

vektor.unityone.ai

contact@unityone.ai

