

Power, Run, and Manage Your AI Factory

UPC Vektor orchestrates GPU infrastructure, fabric, workloads, tenants, and token consumption through a unified operational layer — full-stack AI operations, from silicon to token.

Orchestrate

Optimize

Govern

Monetize

THE PROBLEM, AND THE SHIFT

OWNING GPUS

- Capacity idle between jobs or stranded by scheduling
- Training jobs starve latency-bound inference
- Operations spread across disconnected tools
- One invoice, no attribution by tenant or model
- Audit evidence assembled by hand



OPERATING A FACTORY

- Idle and stranded capacity surfaced and reclaimed
- Placement by priority, SLA, policy, latency, and cost
- One operational view across every layer
- Cost-to-serve per tenant, workload, model, and token
- Central, inspectable audit trail

WHAT IT DOES — SIX Pillars



Optimize the GPU fleet

Provision and monitor clusters, map rack and network topology, track thermal and power, manage pool capacity.



Optimize token economics

Track token throughput, latency, and cost by model, tenant, and workload; compare private against public APIs.



Multi-tenant service layer

Tenant onboarding, GPU allocation, API access, RBAC, and usage metering from one control plane.



LLMs across runtimes

Abstract inference across vLLM, LiteLLM, TensorRT-LLM and public APIs; port between private and public.



Govern with compliance

RBAC, SSO, MFA, per-tenant policy, centralized audit logs, and hardware-backed confidential computing.



Manage through ChatOps

Monitor fleet health, find idle GPUs, and run runbooks, deployments, and rollbacks with human approval.

WHY IT PAYS

94

Capabilities in one control plane

15

Categories, silicon to token

4

Delivery stages to production

99.999%

UnitedLayer platform availability

THE CAPABILITY SURFACE

Platform Experience 6

Discovery & Inventory 6

Data Center Mgmt 7

Cloud & Containers 5

Infra Observability 7

AI Ops 7

Automation & Config 3

FinOps & Sustainability 8

AI & GPU Platform 12

AI Observability 7

Tenant & Workload 7

Model Management 3

Governance & Security 6

Resilience & Priv. Security 4

Administration & Integration 6

HOW YOU GET THERE

01

Design

GPU topology, tenant architecture, fabric, service catalog, capacity, and token pricing.

02

Build

Deploy the stack, onboard tenants, activate workloads, connect LLM APIs, validate token flows.

03

Operate

Monitor GPU health and token throughput, optimise workloads, manage cost, enforce governance.

04

Monetize

Meter tenant usage, automate chargeback and showback, measure revenue against cost-to-serve.

Turn GPU capacity into AI revenue.

Book a walkthrough of UPC Vektor, or ask us to assess your current GPU estate.

UPC **VEKTOR**[™]
UNITEDLAYER.COM



EMAIL
contact@unitedlayer.com



PHONE
+1-888-853-7733



X
x.com/unitedlayer



LINKEDIN
linkedin.com/company/unitedlayer