UPC VEKTOR™

UNITEDLAYER©

THE INTELLIGENT CONTROL PLANE FOR YOUR AI FACTORY

Power, Run, and Manage Your AI Factory.

Optimize GPU performance, reduce cost-to-serve, and streamline AI training and inference with **centralized infrastructure management** — full-stack AI operations, from silicon to token.

Orchestrate | Optimize | Govern | Monetize

THE CHALLENGE

Acquiring GPUs is only the beginning

AI infrastructure is no longer a collection of servers. It is a production environment that must train models, serve inference, support multiple teams, protect sensitive data, meet service-level commitments, and hold to a cost envelope. Traditional infrastructure tools were not designed for that.

GPUs are expensive, and often idle

Capacity sits unused between jobs, becomes stranded by poor scheduling, or is allocated to workloads that never needed it. Investment rises without the utilisation or return to match.

Workloads compete for the same resources

Training needs large clusters for long periods. Inference needs predictable latency. Batch prioritises cost. Without intelligent placement, they collide.

Operations are fragmented

GPUs, fabric, orchestration, runtimes, serving endpoints, identities, and billing each managed separately. No complete view, slow troubleshooting, manual deployment.

AI costs resist measurement

What does this model cost to train? What is the cost per token? Which tenant consumes most? What does idle capacity cost? Without granular data, AI cannot be run as a business.

Governance requirements keep rising

Sensitive data, proprietary models, personal information, and regulated content need access control, audit, and policy — with workloads that may have to stay private.

Workloads need portability

Environments tightly coupled to one cloud, runtime, or API layer make it hard to shift inference, adopt new runtimes, or avoid lock-in.

Questions AI and cloud teams need to answer — and usually cannot:

- What does it cost to train this model?
- What is the cost per inference, or per token?
- Which customers are actually profitable?
- Which business unit uses the most GPU capacity?
- How much does idle capacity cost us?
- Is this workload cheaper privately or in public cloud?

THE GAP

Organisations own powerful AI infrastructure — but do not yet operate an efficient AI factory.

THE PLATFORM

Full-stack AI operations, from silicon to token

UPC Vektor orchestrates GPU infrastructure, fabric, workloads, tenants, and token consumption through a unified operational layer built for scalable AI services.

Optimize your GPU fleet at scale

Provision and monitor clusters, map rack and network topology, track thermal and power utilization, and manage capacity across GPU pools based on workload demand.

Optimize token economics in real time

Track token throughput, latency, and cost by model, tenant, and workload. Compare private inference economics with public APIs to make smarter placement decisions.

Operate a multi-tenant AI service layer

Manage tenant onboarding, GPU allocation, API access, RBAC, and usage metering from one control plane. Align workload placement with SLA, cost, and available capacity.

Orchestrate LLMs across runtimes and clouds

Abstract inference across vLLM, LiteLLM, TensorRT-LLM, and public APIs. Port workloads between private and public environments while supporting PyTorch, TensorFlow, ONNX, and CUDA.

Govern AI operations with control and compliance

Enforce RBAC, SSO, MFA, and per-tenant policies. Centralize audit logs and compliance visibility while securing workloads with hardware-backed confidential computing.

Manage the factory through ChatOps

Monitor fleet health, identify idle GPUs and performance issues, and execute runbooks, deployments, and rollbacks through conversational operations with human approval.

Sovereignty and private inference

Keep models, data, and inference workloads under your control while meeting residency and regulatory requirements — and optimizing the economics of every inference.

- Run LLMs on private GPUs
- Control data residency and compliance
- Protect workloads with confidential computing
- Compare private and public inference costs

The factory analogy

GPUs are the machines. Servers and networking are the production floor. Models and workloads are the products. Data is the raw material. **UPC Vektor is the brain and control system that coordinates the factory** — an AI infrastructure control plane that manages the infrastructure and workloads, not the model's reasoning.

THE CAPABILITY CATALOGUE

94 capabilities, 15 categories

From discovery and data-center management through AIOps, GPU orchestration, AI observability, FinOps, and governance — one platform covering the full operational surface of an AI factory.

PLATFORM EXPERIENCE 6

- Tenant Access and Branding
- Unified Dashboards
- Dashboard Customization
- Activity and Audit Logs
- Support and Feedback
- ITSM Ticket Management

DISCOVERY AND INVENTORY 6

- Collector-Based Onboarding
- Manual and Bulk Inventory Import
- Network Discovery
- Infrastructure Discovery
- Database Discovery
- Asset Intelligence

DATA CENTER MANAGEMENT 7

- Data Center, Rack, and Cabinet Management
- Rack Visualization
- Environmental and Thermal Monitoring
- Power and PDU Management
- Capacity and Growth Management
- Physical Asset Inventory
- Topology and Dependency Mapping

CLOUD AND CONTAINERS 5

- Private Cloud Management
- Public Cloud Management
- Cloud Resource Monitoring
- Kubernetes Management
- Docker and Container Management

INFRASTRUCTURE OBSERVABILITY 7

- Infrastructure Monitoring
- Network Monitoring
- Database Monitoring
- Virtualization Monitoring
- Device and Interface Monitoring
- Metrics and Performance Analytics
- Health, SLA, and Site Analytics

AIOPS 7

- Event and Alert Management
- Event Correlation and Impact Analysis
- Noise Reduction
- Maintenance Management
- Incident Management and RCA
- AIOps Analytics and Reporting
- Third-Party Event Integrations

AUTOMATION AND CONFIGURATION 3

- Network Configuration Management
- Workflow and Runbook Automation
- Service and Provisioning Catalog

FINOPS AND SUSTAINABILITY 8

- Cloud Cost Analytics
- AI and GPU Cost Analytics
- Budget Planning and Alerts
- Pricing and Rate Management
- Cost Estimation
- Billing, Invoices, and Payments
- Revenue and Margin Analytics
- Sustainability and GreenOps

AI AND GPU PLATFORM 12

- GPU Fleet Inventory and Analytics
- GPU Requests and Allocations
- Bare-Metal GPU Provisioning
- Full GPU Virtual Machines
- Fractional and Sliced GPUs
- Container Environments
- Serverless AI Endpoints
- Endpoint Lifecycle Management
- AI Job Management
- GPU Node Lifecycle
- GPU Capacity and Fleet Partitioning
- Fabric Topology and Inspection

AI OBSERVABILITY 7

- Model and LLM Monitoring
- Prompt and Trace Monitoring
- Token and Latency Analytics
- AI Cost and FinOps Monitoring
- Drift and Quality Monitoring
- AI Safety and Guardrail Monitoring
- Agent, RAG, Tool, and Vector Monitoring

TENANT AND WORKLOAD MANAGEMENT 7

- Tenant Lifecycle Management
- Tenant Portfolio Analytics
- Workspace Management
- Project Management
- Quota and Entitlement Management
- SLA Policy and Reporting
- VDC Management

MODEL MANAGEMENT 3

- Model Catalog
- Model Marketplace and Skills
- Model Risk and Attestation

GOVERNANCE AND SECURITY 6

- Policies and Guardrails
- PII Governance
- Feature and Plan Controls
- Users, Groups, and RBAC
- Device-Level Permissions
- API Keys and SDK Lifecycle

RESILIENCE AND PRIVILEGED SECURITY 4

- Disaster Recovery as a Service
- Cybervault — Immutable Backup
- Vault — Secrets Management
- PAM Integration

ADMINISTRATION AND INTEGRATION 6

- Credential Management
- Directory and User Import
- Enterprise Integrations
- Notifications and SMTP
- Connection and Attribute Management
- Reports and Scheduling

WHY UPC VEKTOR

Turn AI infrastructure into measurable ROI

Optimize GPU utilization, power, cooling, and workload placement while gaining control over token economics, isolation, and portability across private and public inference environments.

01
Maximize fleet utilization
Optimize placement using real-time cost per million tokens, factoring in GPU power, PUE, and interconnect economics.

02
Make AI revenue measurable
Meter token consumption and workload costs by tenant to support transparent chargeback, showback, and pricing.

03
Keep workloads portable
Orchestrate across private and public LLM environments without vendor lock-in as models and providers evolve.

04
Scale without losing control
Enforce tenant-level policies, maintain auditability, and protect workloads with confidential computing.

05
Accelerate production readiness
Move from factory design to operational handover through a structured, repeatable implementation path.

DELIVERY MODEL

From infrastructure to production-ready AI

UPC Vektor structures AI factory operations across four delivery stages — from infrastructure design to operational handover — giving clients a production-ready factory with full autonomy post-delivery.

01
Design
Plan GPU topology, tenant architecture, network fabric, service catalog, capacity, and token pricing.

02
Build
Deploy the infrastructure stack, onboard tenants, activate workloads, connect LLM APIs, and validate token flows.

03
Operate
Monitor GPU health and token throughput, optimize workloads, manage costs, and enforce governance in production.

04
Monetize
Meter tenant usage, automate chargeback and showback, and measure revenue against real-time cost-to-serve.

BOTTOM LINE

Turn GPU capacity into AI revenue — run the entire factory with intelligence.

Ready to run your AI factory with intelligence?

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

Product positioning and capability descriptions reflect UnitedLayer published materials for UPC Vektor. The 94-capability catalogue reflects the current consolidated feature set; framework interoperability, deployment models, and feature availability may vary by release, plan, and configuration. UnitedLayer®, United Private Cloud®, UnitedSecure™, UnitedConnect®, UnitedEdge®, and UPC Vektor are marks of UnitedLayer.

© 2026 UNITEDLAYER. ALL RIGHTS RESERVED.