CUSTOMER SUCCESS STORY

Three business lines, one GPU estate they can govern.

A global bank had the GPUs. What it did not have was a way to decide who got them, prove who used them, or show finance what an inference actually cost. **UPC Vektor** became the control plane over the whole estate.

62–68% **Financial services sits among the highest AI adopters**, with most large institutions running AI in production. At that stage the binding constraint stops being access to GPUs and becomes governance: who may use capacity, under which policy, at what cost, with what audit trail.

THE SITUATION

Capacity was there. Control was not.

Fraud, the relationship-manager copilot, and risk modelling all ran on the same private GPU estate. The friction was operational, not technical:

- **Training crowded out inference.** Long risk-model training runs competed with fraud scoring that had a latency commitment attached.
- **Capacity was requested by ticket.** Allocation was manual, slow, and impossible to reconcile against entitlements.
- **No cost attribution.** Finance saw one infrastructure bill and could not tell which line of business or model drove it.
- **Audit questions were hard to answer.** Who accessed which model, under what policy, was assembled by hand from several tools.

THE DECISION

Put a control plane over the estate

Rather than buy more GPUs to end the contention, the bank made the existing estate governable — placement by policy, tenancy by business line, and cost visible per model.

> ***We did not have a capacity problem. We had an allocation problem, and we could not prove to the regulator or the CFO who was using what.***
>
> — HEAD OF AI PLATFORM ENGINEERING

DEPLOYMENT SNAPSHOT

INDUSTRY
Banking & financial services

ESTATE
Private GPU estate across two sites plus DR

TENANTS
Fraud · Retail Banking · Risk

WORKLOADS
Real-time fraud scoring · RM copilot · Risk model training

PRIORITY MODEL
SLA-backed inference above batch and training

GOVERNANCE
RBAC, SSO, MFA, audit logging, tenant policy

ECONOMICS
Chargeback by business line and model

WHAT UPC VEKTOR CONTROLS

Orchestrate	GPU pools inventoried centrally, with **workload-aware placement** keeping latency-bound fraud scoring clear of long training runs.
Govern	Tenant onboarding per business line, RBAC and SSO, entitlement-based GPU allocation, and **audit logging** the risk function can inspect directly.
Measure	GPU consumption, token throughput, and inference volume attributed per tenant and model, driving **chargeback** to each line of business.
Operate	Idle-capacity identification, guided remediation, and runbook execution, so platform engineers stop investigating layer by layer.

WHAT CHANGED

BEFORE UPC VEKTOR

- Capacity allocated by **manual ticket**, days to fulfil
- Risk training runs degrading fraud-scoring latency
- One infrastructure bill, no line-of-business attribution
- Access evidence assembled by hand across several tools
- Idle GPUs invisible until someone went looking

WITH UPC VEKTOR

- Entitlement-based **self-service** within policy
- SLA-aware placement protects latency-bound inference
- Cost per model and per business line, reported monthly
- Central **audit trail** across tenants, access, and workloads
- Idle and stranded capacity surfaced automatically

WHERE THE GAINS CAME FROM

Placement by policy

Fraud inference carries an SLA; training does not. Vektor enforces that difference rather than leaving it to whoever submitted first.

Tenancy per business line

Fraud, Retail, and Risk hold their own entitlements, so one team's experimentation cannot consume another's committed capacity.

Chargeback that lands

Attributing GPU-hours and tokens per model turned an unallocated central cost into budgets each business line owns.

WHY THIS MATTERS IN FINANCIAL SERVICES

Capabilities from the UPC Vektor catalogue doing the work in this deployment:

- GPU Requests and Allocations
- Quota and Entitlement Management
- SLA Policy and Reporting
- Activity and Audit Logs
- Revenue and Margin Analytics
- Users, Groups, and RBAC
- Policies and Guardrails
- Model Risk and Attestation

BOTTOM LINE

UPC Vektor turned a contended GPU estate into a governed, measurable platform three business lines can share.

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

Customer anonymised at their request; the profile and quotation shown are representative of a global-bank deployment. Outcomes described are qualitative and reflect the operating changes UPC Vektor enables; they are not measured benchmarks, and results vary by estate, workload mix, and configuration. Adoption ranges reflect published industry surveys, which vary by scope and methodology. UnitedLayer®, United Private Cloud®, and UPC Vektor are marks of UnitedLayer.

© 2026 UNITEDLAYER. ALL RIGHTS RESERVED.