

CUSTOMER SUCCESS STORY

Every customer on one estate, every cost accounted for.

A B2B SaaS platform served AI features to thousands of tenants from a shared GPU estate — with no way to tell which customers were profitable. **UPC Vektor** gave them per-tenant metering, isolation, and a serving layer they could price against.

78%

Technology and software lead every AI adoption survey, with the large majority running AI in production. For a SaaS vendor that maturity shifts the question from capability to unit economics: what does each tenant cost to serve, and which ones make money.

THE SITUATION

Serving thousands of tenants, blind to the economics

The AI tier was live across the customer base. The platform team could see aggregate GPU spend but nothing underneath it:

- **No per-tenant cost.** Heavy users and light users were indistinguishable in the bill, so pricing was guesswork.
- **Noisy tenants degraded everyone.** One customer's bulk workload could slow interactive features for the rest.
- **Enterprise buyers wanted isolation** in writing, which a flat shared estate could not evidence.
- **Runtime lock-in.** Serving was coupled to one stack, making it costly to adopt faster runtimes as they appeared.

THE DECISION

Meter the estate, then price it

The team put a control plane between the product and the GPUs — isolating tenants, routing inference across runtimes, and metering consumption precisely enough to build a price book on.

“We were selling an AI tier without knowing what it cost us to serve any individual customer. Metering per tenant changed the pricing conversation entirely.”

— VP OF PLATFORM ENGINEERING

DEPLOYMENT SNAPSHOT

INDUSTRY

B2B SaaS, multi-tenant platform

ESTATE

Shared private GPU estate, two regions

TENANTS

Shared tier and isolated enterprise tier

WORKLOADS

Interactive copilot · Semantic search · Batch enrichment

PRIORITY MODEL

Interactive above batch, per-tenant entitlements

PORTABILITY

vLLM, TensorRT-LLM, LiteLLM behind one API

ECONOMICS

Per-tenant metering, showback and chargeback

WHAT UPC VEKTOR CONTROLS

Orchestrate

Interactive inference prioritised over bulk enrichment, with **per-tenant GPU entitlements** preventing any single customer from consuming the pool.

Port

Runtime abstraction across vLLM, TensorRT-LLM, and LiteLLM, so the serving stack can change **without touching the product**.

Measure

Token consumption, inference volume, and GPU time metered per tenant, producing **cost-to-serve** that feeds directly into pricing.

Govern

Tenant isolation, RBAC, SSO, API access management, and audit logs — the evidence enterprise procurement asks for.

WHAT CHANGED

BEFORE UPC VEKTOR

- Aggregate GPU spend, **no per-customer visibility**
- Bulk workloads slowing interactive features
- Isolation claims that could not be evidenced
- Serving coupled to a single runtime
- Pricing set by estimate rather than measurement

WITH UPC VEKTOR

- **Cost-to-serve per tenant**, per feature, per token
- Interactive traffic protected by priority placement
- Isolated enterprise tier with documented controls
- Runtimes swapped behind a stable API
- Price book built on measured consumption

WHERE THE GAINS CAME FROM

Per-tenant metering

Token and GPU consumption attributed to each customer turned cost-to-serve from an estimate into a number the pricing team could work with.

Priority placement

Interactive inference outranks batch enrichment by policy, so one tenant's bulk job no longer shows up as everyone's latency.

Runtime abstraction

Adopting a faster serving runtime became a platform decision rather than a product rewrite, protecting margin as the stack evolves.

WHY THIS MATTERS FOR A SAAS BUSINESS

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

Tenant Lifecycle Management

Tenant Portfolio Analytics

Token and Latency Analytics

AI Cost and FinOps Monitoring

Endpoint Lifecycle Management

Serverless AI Endpoints

Billing, Invoices, and Payments

API Keys and SDK Lifecycle

BOTTOM LINE

UPC Vektor turned a shared GPU estate into a metered, isolated serving platform the business can price and defend.

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