

POWERED BY

UPC **AI Factory**

— ENTERPRISE PRIVATE CLOUD WITH PRIVATE AI

From AI Pilots to a **Private AI Factory**

How a global bank moved **fraud, copilot, and risk models** onto a private, GPU-first cloud — cutting cost, keeping data in-house, and scaling with confidence.

↓ **42%**

LOWER TCO FOR
AI WORKLOADS

< **100ms**

P95 INFERENCE
LATENCY

↑ **40%**

FASTER AI
TRAINING

100%

DATA INSIDE
TENANT BOUNDARY

01 THE SPARK

AI that worked in the lab — but not in the bank

The bank's AI team had real momentum. In the sandbox, the models were impressive: fraud detection flagged anomalies as they happened, a relationship-manager copilot answered policy questions in seconds, and new risk models beat the legacy scorecards.

Then came production. Data could not legally leave the bank's boundary. Cloud bills swung month to month. Latency spiked without warning. And nobody could answer the CFO's simple question:

“What does one million tokens actually cost us — and what happens to that number at scale?”

— THE QUESTION NO ONE COULD ANSWER

02 THE PROBLEM

Four walls blocking enterprise AI

As the bank's ambitions grew, so did the friction. Every route to production hit the same four barriers.



Data control risk

Prompts, embeddings, fine-tuned models, and training data all sat on shared hyperscaler infrastructure — unacceptable to legal and risk for customer PII and transaction records.



Costs that would not sit still

GPU hours, egress, vector database, and API charges made every invoice a surprise. Finance had no way to model TCO as AI spread across regions and business lines.



Latency you could feel

Data shuttled between storage, vector DB, and GPU clusters across regions. Noisy neighbours on shared GPUs made real-time inference inconsistent when it mattered most.



Too many moving parts

Separate teams stitched together Kubernetes, GPU operators, vector databases, model serving, and observability — with no single view of utilisation or cost per model.

They did not need more GPUs. They needed a **private AI platform** where data, models, and inference stayed inside the trust boundary — with predictable cost and enterprise-grade operations.

03 THE TURNING POINT

Building a factory, not renting one

“If AI is going to be core to our business, it cannot run on borrowed infrastructure. We need our own factory — where we control the raw materials, the production line, and the finished goods.”

— THE BANK'S CIO

That factory became **UPC AI Factory**, UnitedLayer's enterprise private cloud with private AI. Rather than bolt AI onto a public cloud, the bank chose a **GPU-first, fully managed private cloud** built for three things: data that never leaves the tenant boundary, a fixed and predictable cost base, and one operated platform covering infrastructure, security, observability, and AI services.

04 THE BUILD

Standing up the private AI estate

Twelve weeks, three phases, delivered end to end as a managed service.

PHASE 1

Discovery & Assessment

Weeks 1–2

- Mapped the AI workloads — **fraud, copilot, risk** — and every data flow behind them
- Captured hyperscaler baselines for latency, throughput, and cost
- Agreed data-control and compliance guardrails with legal, risk, and security

PHASE 2

Parallel Deployment

Weeks 3–8

- Stood up **UPC AI Factory** across two primary data centres, plus a third for disaster recovery
- Created isolated **tenant vDCs** per business line: Fraud, Retail Banking, and Risk
- Deployed GPU pools with MIG partitioning, managed Kubernetes with the NVIDIA GPU Operator, and fast storage
- Turned on **UPC AI Services** — Router, Model Registry, managed Vector DB, and Workbench
- Re-hosted models and RAG pipelines, then ran both environments side by side to prove latency and cost

PHASE 3

Cutover & Optimisation

Weeks 9–12

- Moved production traffic across and decommissioned the hyperscaler GPU clusters
- Tuned GPU scheduling, vector indexing, and model-serving configurations
- Switched on **FinOps** dashboards: GPU-hours, cost per million tokens, cost to serve per model
- Wired observability and AIOps into the bank's existing ITSM and security tooling

Delivered as a managed service under a single license and a fixed monthly recurring charge, with updates included for the term.

UnitedLayer® · United Private Cloud®

05 THE NEW NORMAL

How AI runs now



Data stays inside

Prompts, embeddings, fine-tuned models, and training data all live in tenant vDCs. Any external routing is opt-in and policy-controlled.



One operated stack

Compute, GPU, storage, networking, backup, DR, observability, and AI services on a single UnitedLayer-operated platform.



AI economics in plain view

Dashboards show GPU utilisation, latency, token throughput, and cost per million tokens by model and business line.



Real time, for real

Co-located GPU, storage, and vector DB cut end-to-end latency. Dedicated resources remove noisy-neighbour effects entirely.

06 THE RESULTS

What changed in numbers

Within the first year on UPC AI Factory, the bank saw results that track UnitedLayer's published outcomes for enterprise AI on private cloud.

↓ 42%

Lower TCO for AI workloads

No egress or cross-region charges, fixed-cost GPU capacity, higher utilisation.

↑ 40%

Faster AI training

Single-tenant, RDMA-native architecture tuned for GPU workloads (35–40%).

< 100ms

p95 inference latency

Down from an unpredictable 150–250ms on shared infrastructure.

Predictable

A single monthly charge for AI infrastructure

Clear cost per million tokens and cost to serve per model, making multi-year AI budgeting possible for the first time.

100%

Data and models inside the tenant boundary

Satisfying internal risk, external auditors, and data-residency requirements in every region.

~60%

Less operational toil for the platform team

One operated stack instead of stitched-together tools, unified dashboards, and integrated backup, DR, and cyber vault for critical AI assets.

07 THE TAKEAWAY

A clear story for your buyers

Three beats take a prospect from familiar pain to a controlled, predictable, fully owned AI estate.

The three-beat pitch

01 / OPEN

Start with the pain

AI that shines in a demo but stalls in production because of data, cost, latency, and operations.

02 / POSITION

The private AI factory

UPC AI Factory as the place to scale AI safely across business lines, on a cost base you control.

03 / ANCHOR

Close on numbers

Lower TCO, faster training, predictable cost per million tokens, data kept in-house, simpler operations.

Ready for enterprise AI without compromise?

Talk to us about a private AI factory of your own, or book a walkthrough of UPC AI Factory.

UPC AI FACTORY

EMAIL
US**contact@unitedlayer.com**CALL
US**+1-888-853-7733**

Figures reflect UnitedLayer's published outcomes for enterprise AI on private cloud. The ~42% lower TCO, ~35–40% faster training, and sub-100ms p95 latency shown here are illustrative first-year results for a global-bank deployment profile and will vary by workload, region, and configuration. UnitedLayer®, United Private Cloud®, and UPC AI Factory are marks of UnitedLayer.