OcientAIQ **for Factory Operations**

Cross-layer GPU-fleet telemetry correlated into utilization recovery.

The Problem

Sovereign AI factory and GPU-cloud operators face a margin problem that is structural, not operational. The capital is committed: the GPU hardware is bought, installed, and drawing power. The only lever left is utilization: how many GPU hours the fleet actually sells versus how many it loses to degradation, idle time, or invisible underperformance.

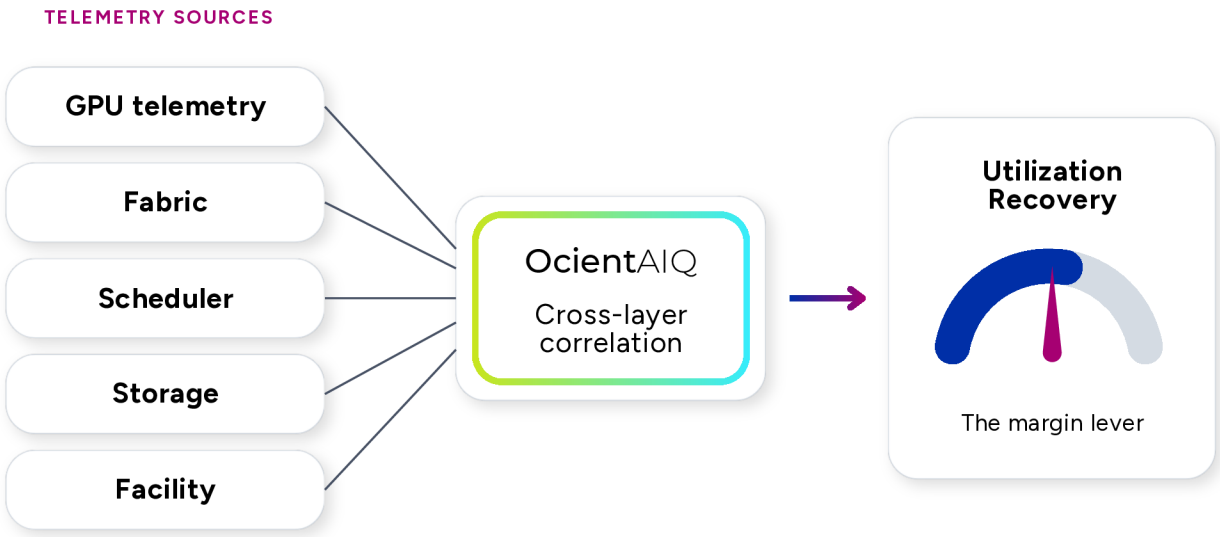
The per-domain monitoring stack covers each layer well — GPU telemetry, network fabric, job scheduling, storage, and facilities — but there's a gap at the seams. When a training job fails with an NCCL timeout, the root cause may live in GPU memory, fabric congestion, a PSU fluctuation, or a storage latency spike, and resolving it means pulling evidence manually from four or five separate tools. That process runs for hours on infrastructure where every hour is billable capacity the operator is not recovering.

The Solution

OcientAIQ™ provides what per-domain tools cannot: a single joined view across GPU, fabric, scheduler, storage, and facilities. Queries run against the full multi-year history of the fleet — not a 90-day window or a sampled dashboard. With OcientAIQ, an investigation that currently requires multiple tools, three shift handoffs, and a morning of manual normalization can run as a governed query and return a traceable evidence package in seconds, not hours.

OcientAIQ makes this level of correlation possible by building a knowledge base from the fleet itself: operational history, topology, playbooks from resolved incidents, utilization and thermal baselines, and business context — tenant priorities, SLA tiers, and billing. OcientAIQ ranks the incident queue by SLA exposure and revenue impact, with every finding backed by a query result, not an assertion.

Five layers in. One query out.



How OcientAIQ Recovers GPU Fleet Utilization

Cross-layer observability. Use a single SQL expression to query the full multi-year history of the fleet — not a sampled dashboard or 90-day window.

Capacity planning. Anchor fleet-wide health baselines, workload-placement recommendations, and capacity forecasts to actual GPU utilization and thermal profiles.

Predictive failure detection. Identify GPU failures and thermal breaches 48-72 hours before XID errors impair a billable job.

Root-cause analysis. Every evidence package comes with lineage back to the source records, ranked by SLA exposure and revenue impact.

SLA assurance and chargeback. Deliver full-fidelity SLA evidence per tenant workload, with incident queues ranked by revenue impact.

Differentiated service tiers. Define, enforce, and report against tiered SLAs backed by evidence — not assertions.

What Makes OcientAIQ Different

Per-domain tools stay in place. OcientAIQ is not a rip-and-replace. It sits underneath tools already running, preserving existing investments while delivering the joined view existing tools cannot achieve individually.

Predictable economics. No query tax. Infrastructure TCO reduction of 50 - 90% versus legacy stacks. Each additional solution area lands at a fraction of the marginal cost of the first.

Connects to the AI ecosystems you already run. OcientAIQ builds and maintains the structured context your agents need. Bring Claude, GPT, Gemini, or Llama and put them to work on production analytics, not demos.

Sovereign by design. Runs on prem, hybrid, or air-gapped. SAML/OIDC zero trust. No data movement.

Value at Every Stage

Monitor
Unify telemetry from multiple sources on one platform.

→

Correlate
Join signals across layers to surface root cause in seconds.

→

Automate
Close the loop between investigation and action through existing control planes.

→

Optimize
Improve fleet utilization, capacity placement, & SLA performance over time.

→

Monetize
Define and enforce differentiated service tiers — and turn the data into a revenue line.

Find out how much utilization your fleet is leaving on the table.

sales@ocient.com ocient.com