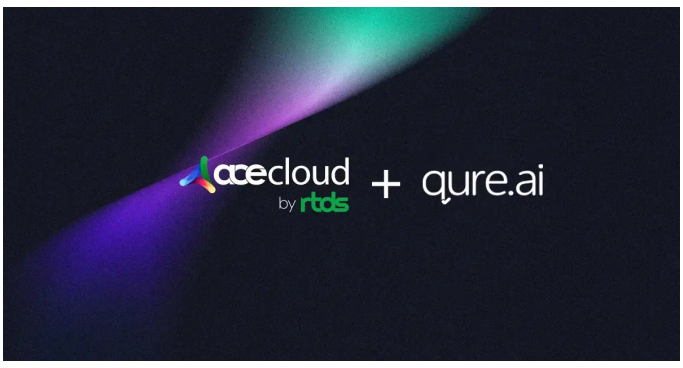

Qure.ai Locks In Predictable GPU Costs and a Zero-Downtime Path to Production AI

Started from a month-long GPU benchmark to 10x infrastructure scale-up plan, Qure.ai is building its LLM training and diagnostic AI stack on AceCloud.

INDUSTRY
Health-Tech - AI/ML

LOCATION
Mumbai, Maharashtra, India

USE CASE
GPU Compute for LLM Training & Diagnostic AI Workloads

Key Outcomes

Zero
Downtime during UAT/Dev migration

1-Month
POC to validated production readiness

5 GPU
Types benchmarked (L4, L40S, H200, RTX PRO 6000, A30)

Overview

Qure.ai needed a GPU cloud partner that could prove price-to-performance economics and platform stability before it would commit production healthcare workloads to the migration.

> **Moving our LLM training to AceCloud was one of the better infrastructure decisions we made. The stability has directly improved our delivery timelines.**
> – Suchit Kotiyan, Infra Head. Qure.ai

Use Case

- AI-Powered Diagnostic Imaging
  Builds AI software that helps radiologists and clinicians detect diseases from X-rays, CT scans, and other medical imaging. Impact on 45M+ lives across 105+ countries.
- LLM Training & Diagnostic Inference
  Runs GPU-intensive workloads for both diagnostic model inference and large language model training, demanding consistent performance across workload types.
- Multi-Provider Infrastructure, Consolidating to AceCloud
  Operates across multiple providers. UAT and Dev earlier on E2E, production currently on AWS.

Challenges

- Cost Unpredictability Stood in the Way of Scaling Confidently
  As a fast-growing health-tech company, Qure.ai expects its infrastructure usage to scale significantly in the coming period. But unpredictable cloud billing made it hard to plan that growth with confidence. Cost was one of the clearest blockers to committing further spend to any single provider.
- GPU Performance Had to Hold Up Under Clinical-Grade Stakes
  Training LLMs and running diagnostic models for healthcare use cases demands consistent GPU performance across workload types, not a one-time benchmark, but performance that holds as workloads scale. And because these systems touch clinical workflows, any migration had to happen without service disruption.

Solutions

- Broad GPU Fleet, On-Demand
  On-demand access to a broad GPU Fleet (L4, L40S, H200, RTX PRO 6000, and A30 GPUs) let Qure.ai benchmark real performance across its actual workload mix, not a single reference configuration.
- Zero-Downtime Migration
  UAT and Dev environments moved from E2E (or previous vendor) with zero downtime. De-risking the case for moving production workloads next.
- Predictable, Price-Performance-Led Cost Model
  A predictable infrastructure cost structure replaced the billing uncertainty that had been blocking Qure.ai's scale-up plans.
- Validated in a One-Month POC
  A month-long proof of concept benchmarked seamless performance across multiple GPU types alongside platform stability. It gave Qure.ai the evidence it needed to move forward.

Impact Highlights

Qure.ai validated AceCloud's GPU performance and platform stability in a month-long POC, then moved LLM training into production with zero downtime. This improved delivery timelines with full production migration from AWS planned next, contingent on sustained 99.95% uptime.