

## Case Study

# How ITQ reduced AI infrastructure spend by 40% while processing 141B monthly tokens on Neysa Velocis

## Neysa Setup

### Service

Velocis Managed  
Inference Engine

### Scale

141 Billion Tokens  
Processed Monthly

### Model

Qwen2.5  
7B-Instruct

## Outcome

**< 2-Sec**

P99 Latency

**40%**

Cost Reduction

**99.5%**

Output Accuracy

## Overview

General-purpose cloud infrastructure wasn't built for production AI at scale. Self-hosting yielded poor unit economics and erratic performance; multi-tenant environments compromised latency and IP security.

ITQ partnered with Neysa to deploy a dedicated managed inference environment - optimized with vLLM and tensor parallelism - delivering sub-two-second P99 latency at 2,500 tokens per second, a 40% cost reduction, and full model IP sovereignty.

## Client Profile

ITQ Technologies is a leading travel technology provider, and the exclusive regional Travelport partner across India, Sri Lanka, Maldives, and Bhutan. It powers high-volume booking operations by connecting agencies directly to global airline inventory.

## Client Use Case

ITQ deployed AI to automate flight change and cancellation penalty calculations. The system parses complex airline policies to provide real-time, precise fee estimates, preventing agency revenue loss and ensuring pricing transparency.



Our goal was to revolutionize travel retailing with AI, but processing billions of tokens monthly on shared infrastructure created unsustainable cost structures and prohibitive performance latency barriers.

**Jitendra Shrivastava**

Data center &amp; Technology Lead, ITQ

## Challenges

Production AI requires deterministic latency and sustainable token economics. For ITQ, processing hundreds of billions of tokens monthly, every traditional architectural choice introduced a critical trade-off.

- **Tokenomics at Scale**

Serverless frontier models became economically unsustainable at ITQ's usage, threatening the economic viability of the project.

- **Token Latency & TTFT**

Early deployments on legacy clouds produced latency spikes exceeding 14s - far too slow for real-time production workflows.

- **Retaining Model IPR**

Multi-tenant cloud environments exposed ITQ's proprietary models to noisy neighbor risks, with no guarantee that intellectual property would remain sovereign and protected.



Neysa enabled us to bridge the gap between ambition and execution. By engineering a dedicated, high-velocity inference engine for us, they delivered the deterministic LLM performance we required while securing our model and reducing our total infrastructure spend by 40%.

**Anil Porter**

Chief Technology Officer, ITQ

## Transformation

ITQ moved to a private, managed inference environment on Neysa Velocis - dedicated NVIDIA GPUs, deterministic QoS, sovereign IP protection, and predictable unit economics.



### Serving stack optimization

Neysa implemented vLLM with tensor parallelism across dedicated NVIDIA GPU clusters, eliminating compute bottlenecks at ITQ's massive token volumes.



### Deterministic latency tuning

Fine-tuned KV cache utilization and TTFT slashed P99 latency from 14s to under 2s while maintaining 2,500 tokens/sec throughput.



### Sovereign performance engineering

The dedicated environment isolated ITQ from noisy-neighbor disruptions, ensuring SLA-backed reliability across billions of monthly tokens.



## Start Your Production AI Deployment With Neysa

Get in touch with our experts. Request a demo.

[hello@neysa.ai](mailto:hello@neysa.ai)

**neysa**