CASE REFERENCE · HOSPITALITY & PHYSICAL ASSETS

International Marketing Group

FDX - Apex (GPU)

International Marketing Group	Hospitality & Physical Assets	Design · Develop	
CLIENT	INDUSTRY	DELIVERED AS	

- NVIDIA CUDA
- PyTorch
- Open-source LLM
- Private GPU Cluster

PRIVATE AI INFERENCE

Private inference, local currency, global impact

We moved an international marketing group's generative-AI inference off public cloud onto our private NVIDIA GPU cluster, so their data and token generation stay local, they pay in local currency, and they run cheaper after hours.

01

THE CHALLENGE

The group leans on generative AI to produce campaign content at volume for its hospitality and travel clients, and the cloud bill was climbing with every prompt. Inference on the public cloud is priced per token and billed in US dollars, so cost scaled directly with usage and carried exchange-rate exposure on top. There was also the question of where the data and the generated content actually lived, which for a marketing group handling client material is not a small thing. They wanted the AI without the runaway, foreign-currency bill.

02

THE APPROACH

We moved their inference off public cloud and onto our private GPU compute cluster. The cluster runs NVIDIA GPUs on the CUDA stack with PyTorch-based serving, hosting an open-source LLM, so the group gets the same kind of generative capability they had in the cloud, but on infrastructure we run locally. Their data and every token generated stay inside the local environment rather than crossing into a hyperscaler's region. Because the workload is predictable, we could schedule the heavier generation runs into off-peak hours, where our inference pricing is lower.

03

THE OUTCOME

The group kept its generative-AI capability but got off the dollar-denominated cloud meter. Inference now runs on our private NVIDIA cluster, the data and the tokens stay local, and the bill comes in local currency with no exchange-rate surprises. By pushing the heavier, batchable generation into off-peak windows they bring the cost down again, and they have a predictable basis for scaling usage up rather than watching a cloud invoice climb.

CASE REFERENCE · HOSPITALITY & PHYSICAL ASSETS

International Marketing Group

FDX - Apex (GPU)

International Marketing Group	Hospitality & Physical Assets	Design · Develop	
CLIENT	INDUSTRY	DELIVERED AS	• NVIDIA CUDA • PyTorch • Open-source LLM • Private GPU Cluster

04 SUMMARY & BENEFITS

- ✓ Generative-AI inference moved off public cloud onto our private NVIDIA GPU cluster.
- ✓ Reduced cost, billed in local currency with no US-dollar exchange-rate exposure.
- ✓ Locally hosted data and token generation, nothing leaves the local environment.
- ✓ Lower after-hours inference pricing for batchable generation runs.

Want a result like this?
Talk to us about your own challenge — info@firsttech.digital or +27 10 501 0800.

Let's Talk →