CASE REFERENCE · TECHNOLOGY & INNOVATION

First Technology

FDX - Apex (GPU)

First Technology	Technology & Innovation	Design · Develop	
CLIENT	INDUSTRY	DELIVERED AS	• NVIDIA CUDA • PyTorch • Open-source LLM • Private GPU Cluster

PRIVATE AI INFERENCE

Private token, no per-seat cost, company wide impact!

We gave First Technology staff local AI chat and agentic desktop agents running on our private NVIDIA GPU cluster, so the data and token generation stay local, the bill is in local currency, and responses come back fast.

01

THE CHALLENGE

First Technology wanted to put AI in the hands of its own people, a local AI chat assistant and agentic desktop agents that could actually do work on the desktop, not just answer questions. The blocker was the same one everyone hits: the cost of cloud-based inference. Priced per token and billed in US dollars, an organisation-wide rollout to staff would have meant an unpredictable, foreign-currency bill that grew with every conversation. For an interactive assistant, the round trip out to a cloud region also added latency that staff would feel on every request.

02

THE APPROACH

We stood the inference up on our own private GPU compute cluster instead of the public cloud. The cluster runs NVIDIA GPUs on the CUDA stack with PyTorch-based serving, hosting an open-source LLM, and the chat and the desktop agents call it directly. Because it is our own infrastructure running locally, the data and every token generated stay inside the environment, and requests do not have to travel out to a hyperscaler and back, which keeps responses quick enough for an interactive assistant and the agents that act on the desktop.

03

THE OUTCOME

Staff got a local AI chat assistant and desktop agents that feel responsive, because the inference runs close by on our own cluster rather than out in a cloud region. The data and the tokens stay local, the cost comes in local currency instead of dollars, and the low latency makes the assistant something people keep using rather than abandon. First Technology can roll AI out across its people on a cost base it controls.

04

SUMMARY & BENEFITS

- ✓ Local AI chat and agentic desktop agents for staff, served from our private NVIDIA GPU cluster.
- ✓ Reduced cost, billed in local currency rather than US dollars.
- ✓ Locally hosted data and token generation, nothing leaves the environment.
- ✓ Low latency, inference runs locally, so the assistant and agents feel responsive.

CASE REFERENCE · TECHNOLOGY & INNOVATION

First Technology

FDX - Apex (GPU)

First Technology	Technology & Innovation	Design · Develop	
CLIENT	INDUSTRY	DELIVERED AS	

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

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

Let's Talk →