JOHN RENKEN

Nebius GPU Cloud

FINETUNING & DEPLOYING QWEN2

6 FEBRUARY, 2025

Agenda

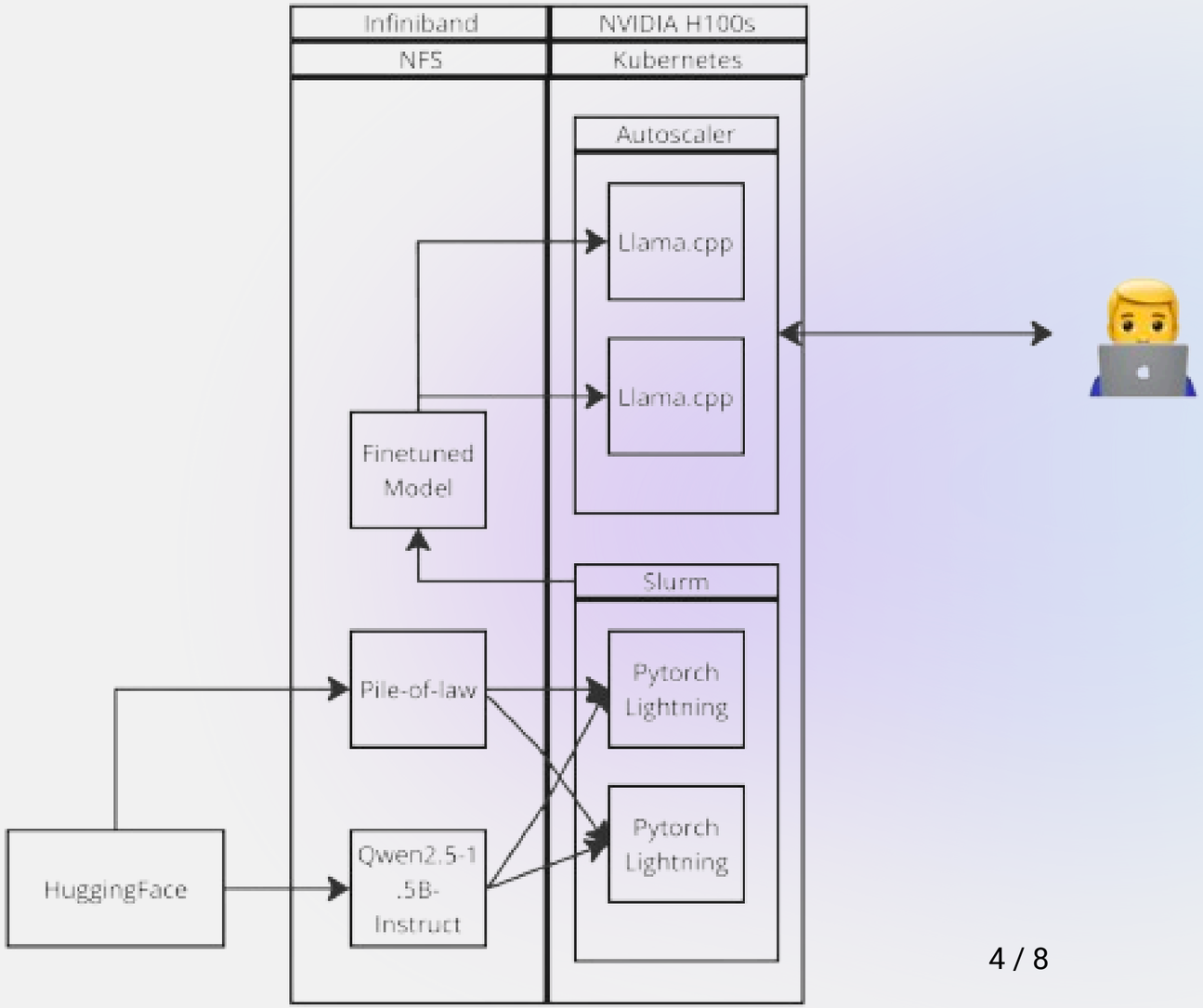
Today we will showcase various features within Nebius GPU Cloud by finetuning & deploying Qwen 2 on a subset of the pile-of-law dataset.

- ABOUT ME
- NEBIUS' NEEDS
- ARCHITECTURE
- QUESTIONS
- DEMONSTRATION

John Renken

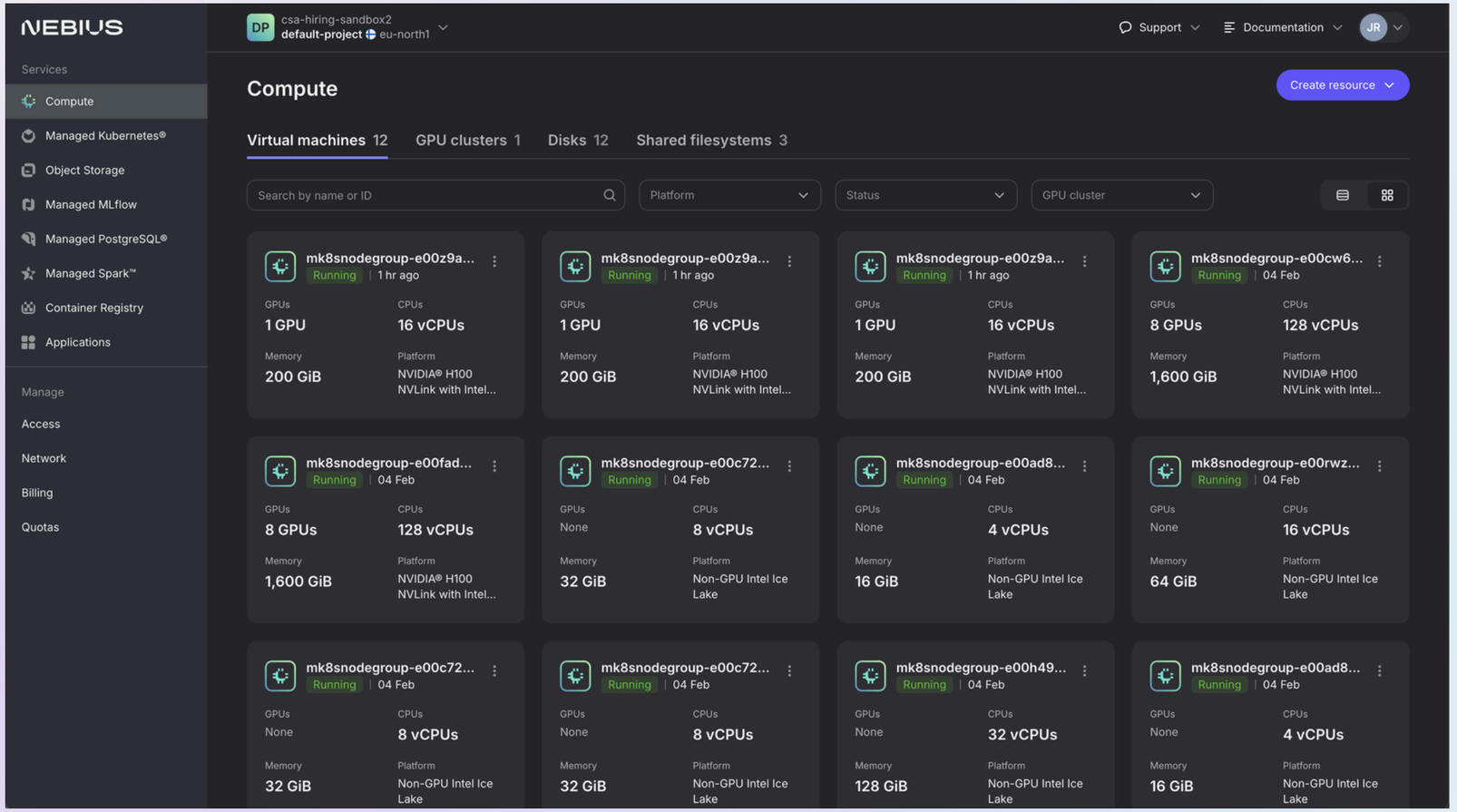
ML ARCHITECT

I’m a machine learning engineer with previous experience selling technical products at IBM and Goldman Sachs. Languages and technology have long been passions of mine, in particular helping people solve important problems with those tools.



Architecture

We leverage Nebius' strategic advantages in hardware, networking, and system design to deliver hyper-performant training and inference.



Demonstration

6 FEBRUARY, 2025

Adaptability

SLURM

Picked up SLURM and the soperator accelerator quickly

Experience

LLAMA.CPP

Deployed the finetuned model using GPU-enabled Llama.cpp

Scale

LIGHTNING

Scaled training to all available hardware with Pytorch Lightning

Client Focus

NFS

Leveraged the shared filesystem and Infiniband features

Adaptability

SLURM

Picked up SLURM and the soperator accelerator quickly

Experience

LLAMA.CPP

Deployed the finetuned model using GPU-enabled Llama.cpp

Scale

LIGHTNING

Scaled training to all available hardware with Pytorch Lightning

Client Focus

NFS

Leveraged the shared filesystem and Infiniband features

Questions