

JOHN RENKEN

Nebius GPU Cloud

FINETUNING & DEPLOYING QWEN2

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

ARCHITECTURE

DEMONSTRATION

NEBIUS' NEEDS

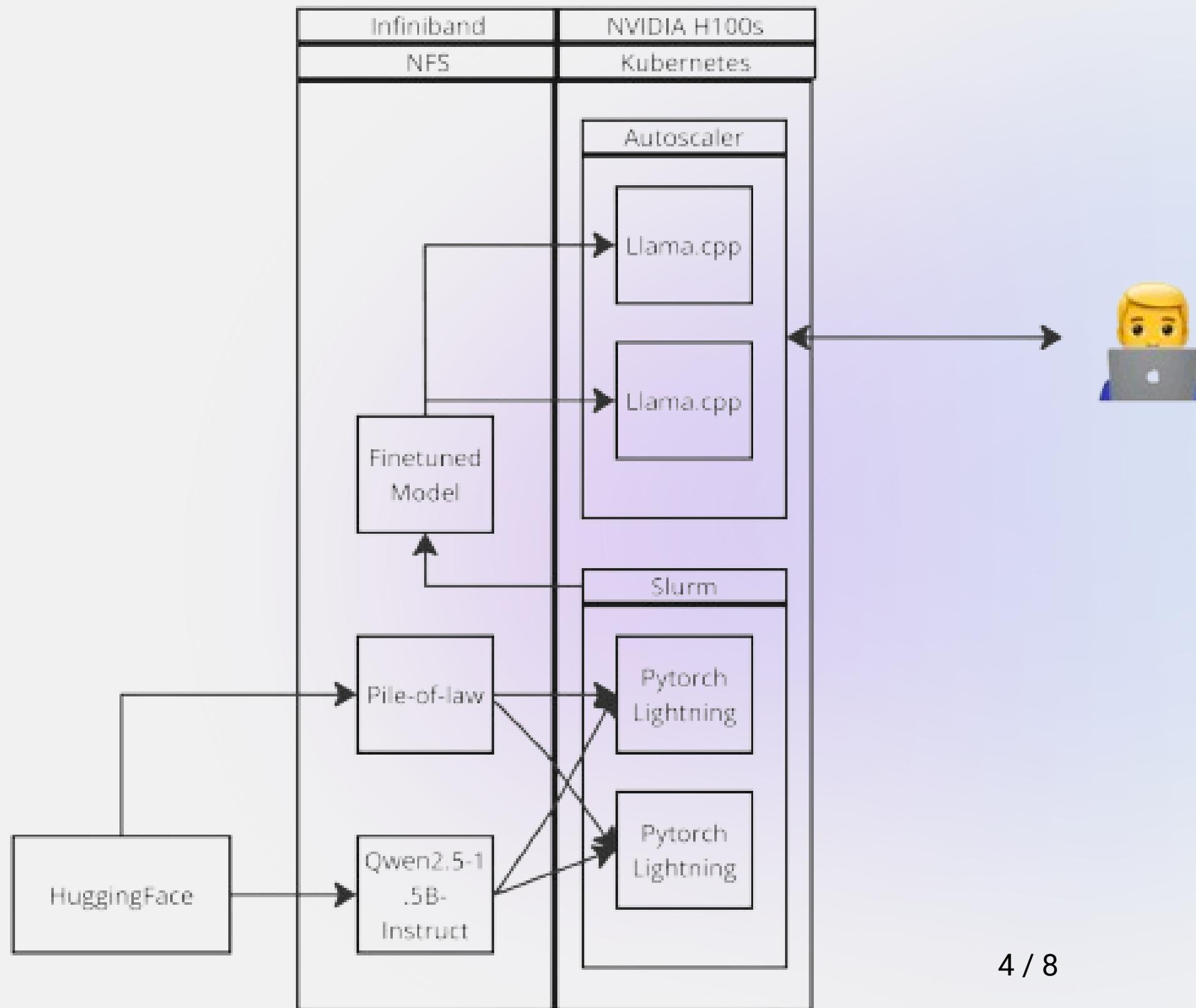
QUESTIONS

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.

The screenshot shows the NEBIUS Compute interface. The top navigation bar includes a project dropdown for 'csa-hiring-sandbox2' and 'default-project eu-north1', along with 'Support', 'Documentation', and user 'JR'. The left sidebar has sections for 'Services' (Compute, Managed Kubernetes®, Object Storage, Managed MLflow, Managed PostgreSQL®, Managed Spark™, Container Registry, Applications), 'Manage', and 'Access' (Network, Billing, Quotas). The main 'Compute' section displays 12 virtual machines. The first row contains four VMs: 'mk8snodegroup-e00z9a...' (1 GPU, 16 vCPUs, 200 GiB memory, NVIDIA® H100 Platform, NVLink with Intel...), 'mk8snodegroup-e00z9a...' (1 GPU, 16 vCPUs, 200 GiB memory, NVIDIA® H100 Platform, NVLink with Intel...), 'mk8snodegroup-e00z9a...' (1 GPU, 16 vCPUs, 200 GiB memory, NVIDIA® H100 Platform, NVLink with Intel...), and 'mk8snodegroup-e00cw6...' (8 GPUs, 128 vCPUs, 1,600 GiB memory, NVIDIA® H100 Platform, NVLink with Intel...). The second row contains four VMs: 'mk8snodegroup-e00fad...' (8 GPUs, 128 vCPUs, 1,600 GiB memory, NVIDIA® H100 Platform, NVLink with Intel...), 'mk8snodegroup-e00c72...' (None, 8 vCPUs, 32 GiB memory, Non-GPU Intel Ice Lake Platform), 'mk8snodegroup-e00ad8...' (None, 4 vCPUs, 16 GiB memory, Non-GPU Intel Ice Lake Platform), and 'mk8snodegroup-e00rwz...' (None, 16 vCPUs, 64 GiB memory, Non-GPU Intel Ice Lake Platform). The third row contains four VMs: 'mk8snodegroup-e00c72...' (None, 8 vCPUs, 32 GiB memory, Non-GPU Intel Ice Lake Platform), 'mk8snodegroup-e00c72...' (None, 8 vCPUs, 32 GiB memory, Non-GPU Intel Ice Lake Platform), 'mk8snodegroup-e00h49...' (None, 32 vCPUs, 128 GiB memory, Non-GPU Intel Ice Lake Platform), and 'mk8snodegroup-e00ad8...' (None, 4 vCPUs, 16 GiB memory, Non-GPU Intel Ice Lake Platform). Each VM card includes a status icon (green for 'Running'), creation date, and a more options menu.

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

JOHN RENKEN

Questions