

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

NEBIUS' NEEDS

ARCHITECTURE

QUESTIONS

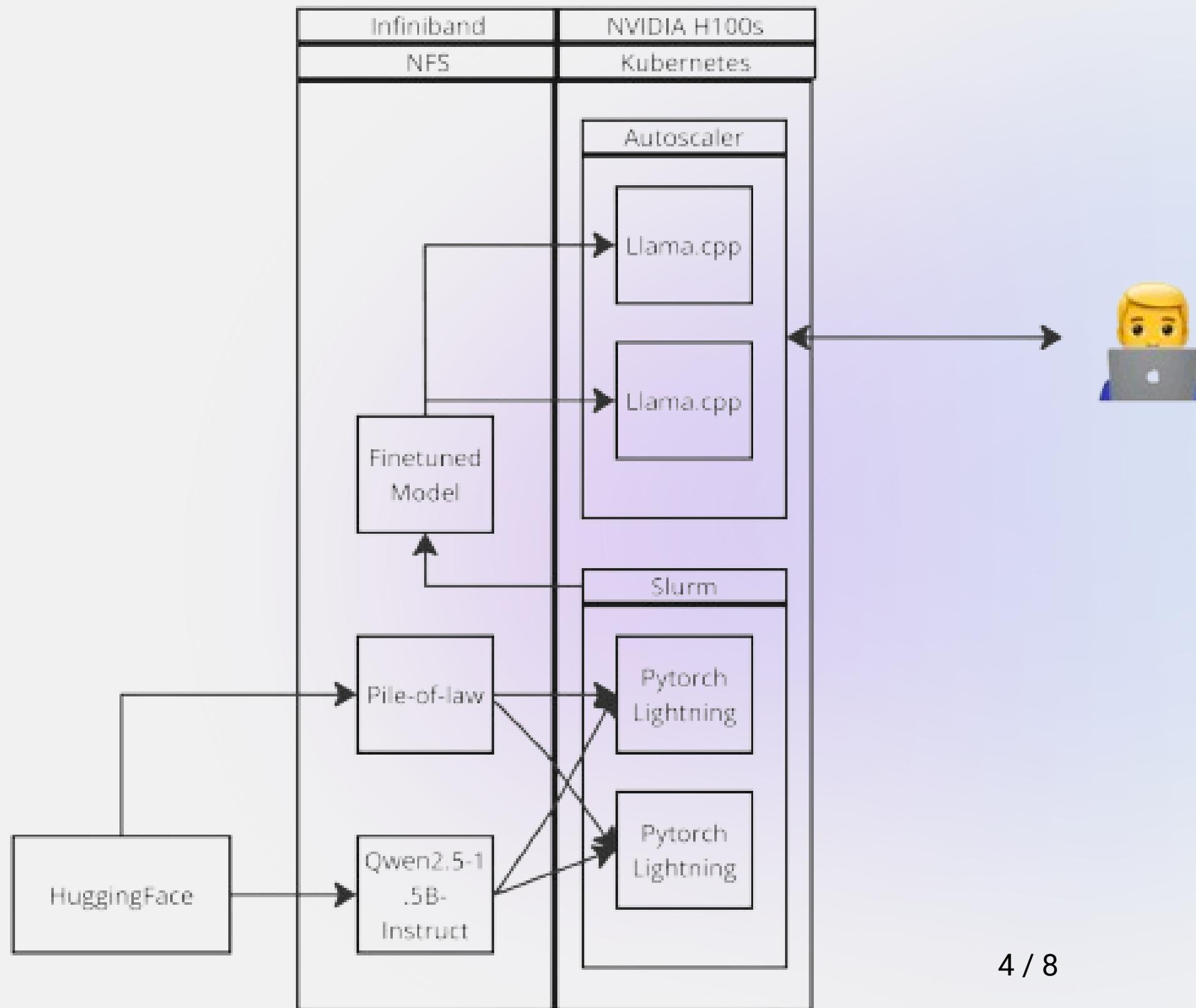
DEMONSTRATION

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 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 profile 'JR'. The main title 'Compute' is centered above the resource count: 'Virtual machines 12 GPU clusters 1 Disks 12 Shared filesystems 3'. A 'Create resource' button is located in the top right. On the left, a sidebar lists 'Services' (Compute, Managed Kubernetes®, Object Storage, Managed MLflow, Managed PostgreSQL®, Managed Spark™, Container Registry, Applications), 'Manage', 'Access', 'Network', 'Billing', and 'Quotas'. The 'Compute' section displays 12 virtual machines in a 3x4 grid. Each card provides details: name, status (Running or Last Seen), creation date, GPU count, vCPU count, memory, and platform. For example, the first four machines have 1 GPU and 16 vCPUs, while the last one has 8 GPUs and 128 vCPUs.

Name	Status	Last Seen	GPUs	CPUs	Memory	Platform
mk8snodegroup-e00z9a...	Running	1 hr ago	1 GPU	16 vCPUs	200 GiB	NVIDIA® H100 NVLink with Intel...
mk8snodegroup-e00z9a...	Running	1 hr ago	1 GPU	16 vCPUs	200 GiB	NVIDIA® H100 NVLink with Intel...
mk8snodegroup-e00z9a...	Running	1 hr ago	1 GPU	16 vCPUs	200 GiB	NVIDIA® H100 NVLink with Intel...
mk8snodegroup-e00cw6...	Running	04 Feb	8 GPUs	128 vCPUs	1,600 GiB	NVIDIA® H100 NVLink with Intel...
mk8snodegroup-e00fad...	Running	04 Feb	8 GPUs	128 vCPUs	1,600 GiB	NVIDIA® H100 NVLink with Intel...
mk8snodegroup-e00c72...	Running	04 Feb	None	8 vCPUs	32 GiB	Non-GPU Intel Ice Lake
mk8snodegroup-e00ad8...	Running	04 Feb	None	4 vCPUs	16 GiB	Non-GPU Intel Ice Lake
mk8snodegroup-e00rwz...	Running	04 Feb	None	16 vCPUs	64 GiB	Non-GPU Intel Ice Lake
mk8snodegroup-e00c72...	Running	04 Feb	None	8 vCPUs	32 GiB	Non-GPU Intel Ice Lake
mk8snodegroup-e00c72...	Running	04 Feb	None	8 vCPUs	32 GiB	Non-GPU Intel Ice Lake
mk8snodegroup-e00h49...	Running	04 Feb	None	32 vCPUs	128 GiB	Non-GPU Intel Ice Lake
mk8snodegroup-e00ad8...	Running	04 Feb	None	4 vCPUs	16 GiB	Non-GPU Intel Ice Lake

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