ArgoCon

How to fine tune an LLM with Argo Workflows and Hera

Flaviu Vadan | J.P. Zivalich

ABOUT US

Flaviu Vadan

- Staff Engineer, Dyno Therapeutics
- Watertown, MA, USA

flaviuvadan

in flaviuvadan

J.P. Zivalich

- CTO & Founder, Pipekit
- Atlanta, GA, USA

JPZ13

in jpzivalich

OUTLINE

Motivation

Foundation models & fine tuning

Infrastructure

Walkthrough

MOTIVATION

Show how to do scalable distributed fine tuning for LLMs

Target Audience:

- Individuals, teams, and companies who want to use LLMs, but need additional customization
- Teams interested in distributed model training

FOUNDATION MODELS

- General, open-source models
- Very expensive to train
- Fine tune on your own data
- Good for …
  - Domain-specific training (medical, support, etc.)
  - Training on private/ proprietary data sets

FINE TUNING

- Transfer learning technique
- General guide:
  - Set up infrastructure
  - Take existing model
  - Feed it your own data

INFRASTRUCTURE

Kubernetes cluster with GPUs

Custom Storage Class

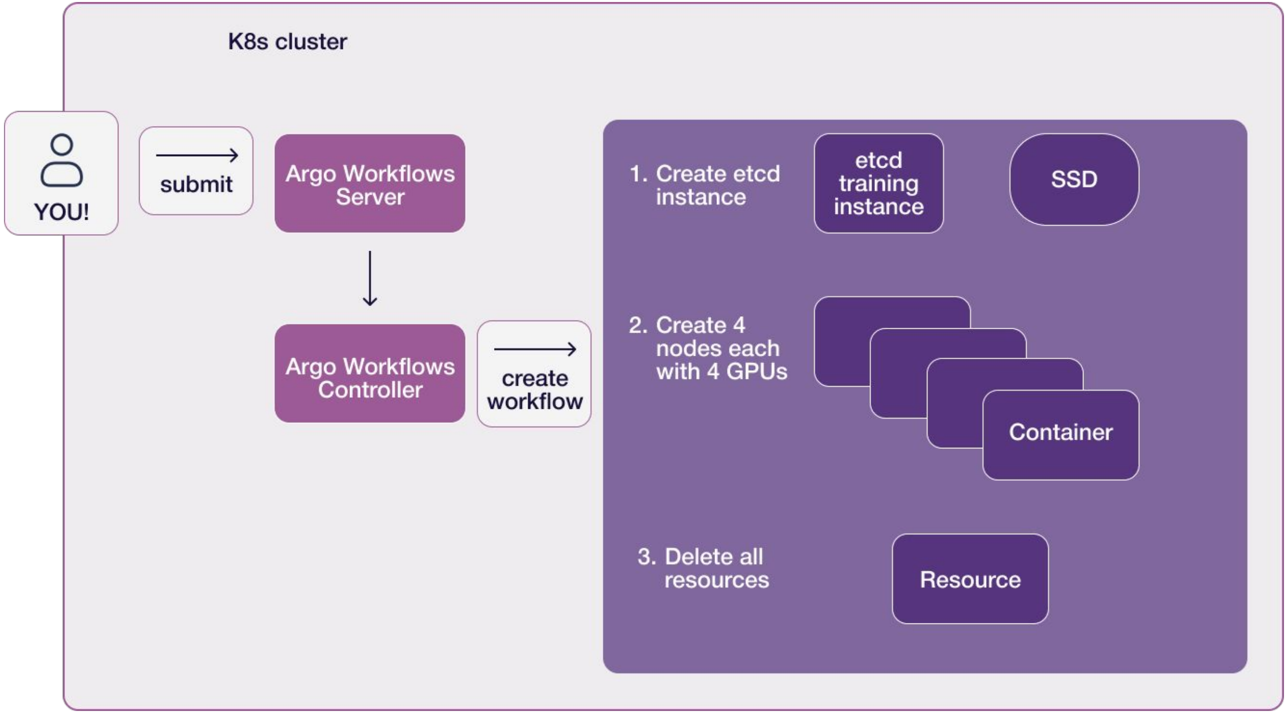
GPUs

Argo Workflows installed

HuggingFace account

Approval from Meta that you can use llama

ARCHITECTURE DIAGRAM



pipekit ArgoCon

DISTRIBUTED KEY-VALUE STORE

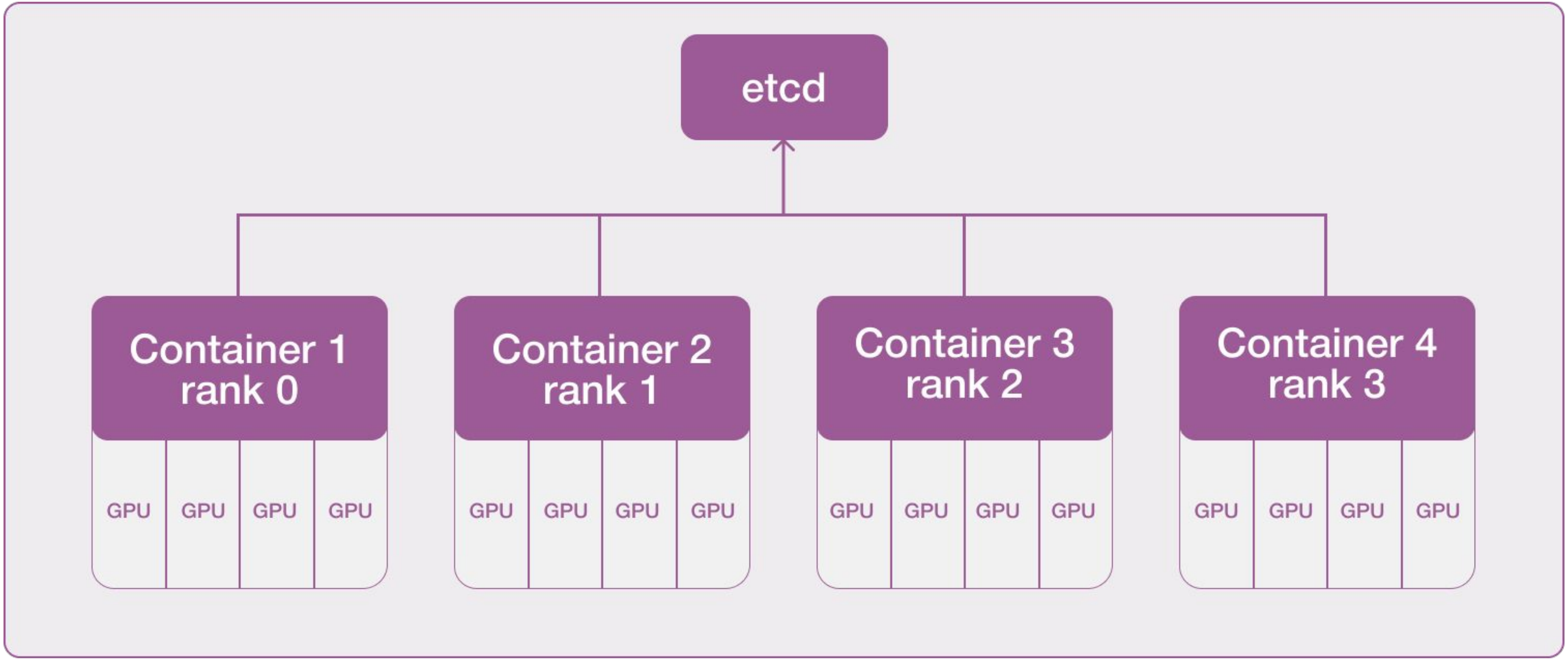
Problem: Track which shards of the model have been trained on which sections of the data set

Solution: Use a distributed key-value store

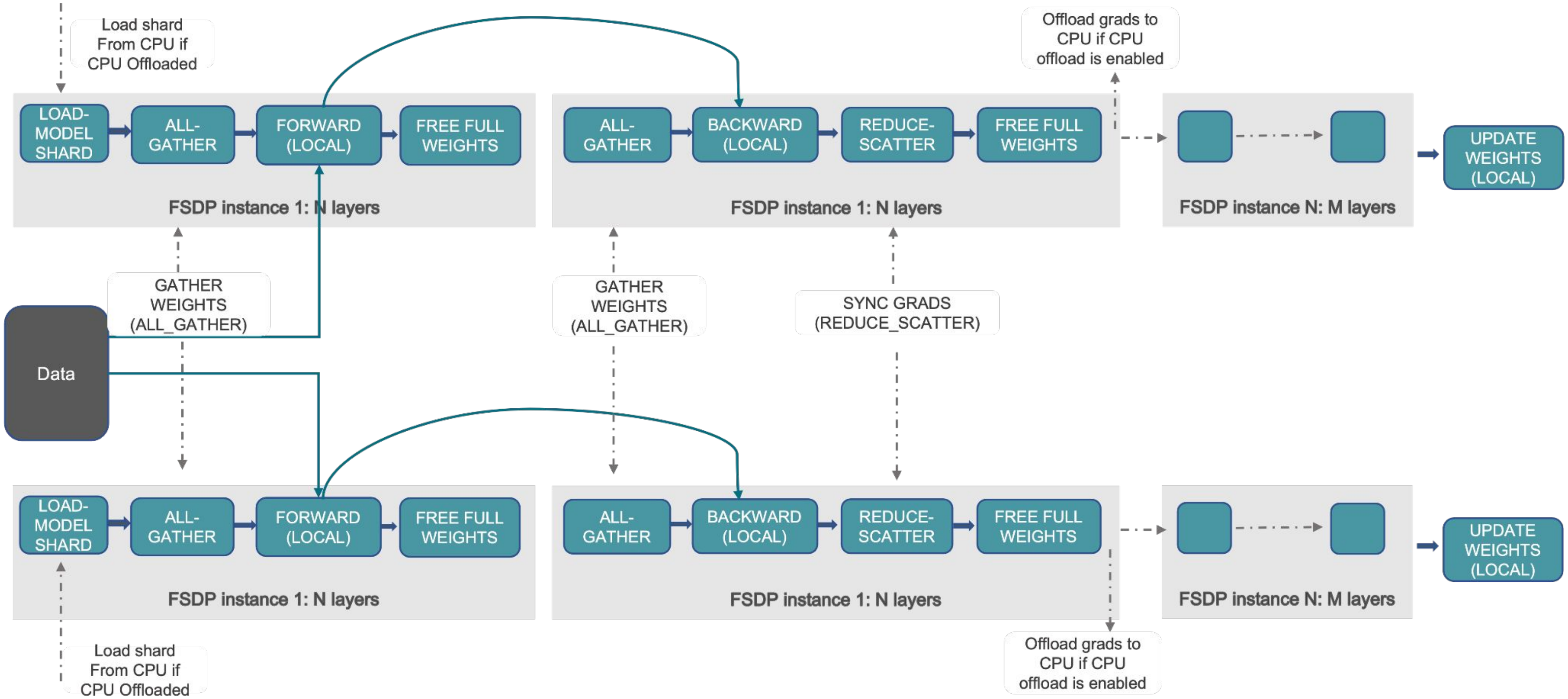
We chose to provision a replicated etcd instance

This is separate from the existing etcd instance that Kubernetes uses

Workflow steps for distributed training



PyTorch Fully Sharded Data Parallel (FSDP)



TEAR DOWN

A
The training etcd instance is torn down at the end of the workflow run using an Exit Handler in Hera/Argo Workflows

B
The cluster autoscaler tears down the GPUs, as they are no longer needed

C
This allows us to ensure tear down regardless of success or failure of the workflow run itself

D
As a general rule, workflow runs should be as ephemeral as possible

pipekit ArgoCon

Walkthrough of Hera code

https://github.com/flaviuvadan/kubecon_na_23_llama2_finetune

Acknowledgements / Resources

- https://hera.rtfd.io/
- https://argoproj.github.io/workflows/
- https://github.com/etcd-io/etcd
- https://pytorch.org/tutorials/intermediate/FSDP_tutorial.html
- https://github.com/facebookresearch/llama-recipes
- https://huggingface.co/meta-llama/Llama-2-7b-hf

Share your feedback and check out the code

https://s.pipekit.io/argo-llm

Chat more with us about Argo & LLMs

https://s.pipekit.io/chat-argo-llm