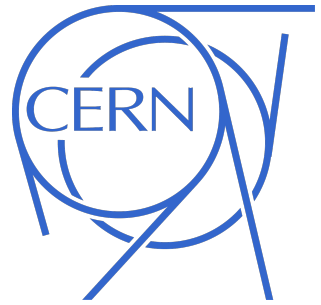
CERN openlab



Accelerating HEP Workloads

Jason Praful Francis Xavier

IT-CM-RPS

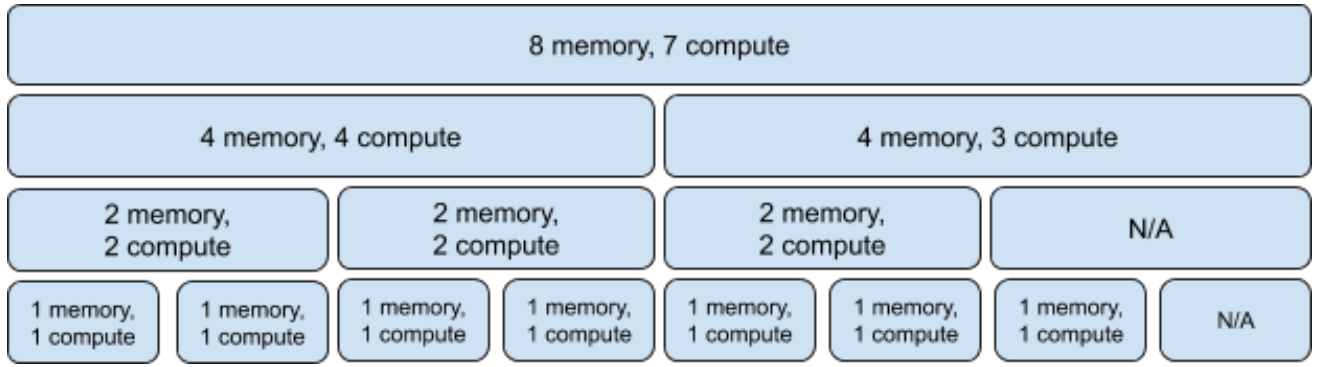
06/09/2021

Motivation

- Increased use of Machine Learning at CERN
- Centralised infrastructure with the ability to use specialized accelerators like TPUs and FPGAs.
- Additional benefits of using a centralised infrastructure.

Testing, Integration and New Features

1. Testing and providing feedback on current examples
2. Reportioning single GPU to be used by multiple services (MIGs)
3. Serving multiple models through a single GPU



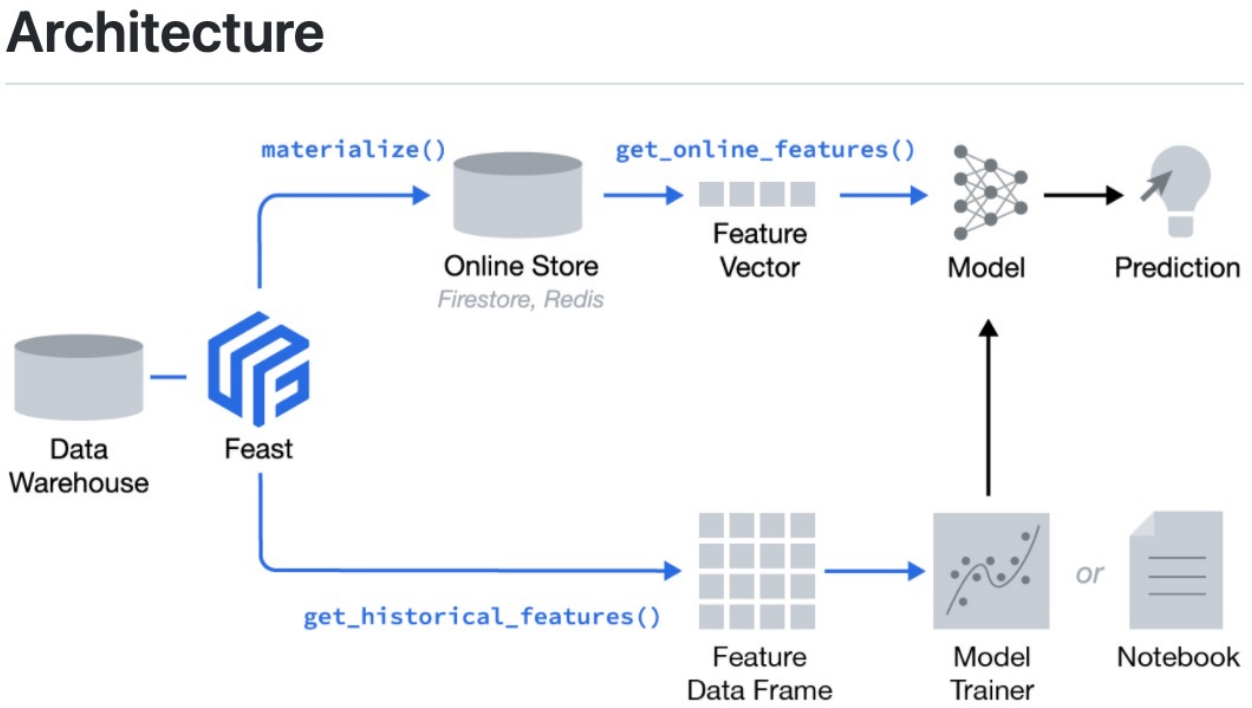
Configuration/Split

Testing, Integration and New Features

4. Multiple additional features such as Feast (**Fea**ture **St**ore), Goofys, Codeserver.

5. Additional Kubeflow components added to improve UX:

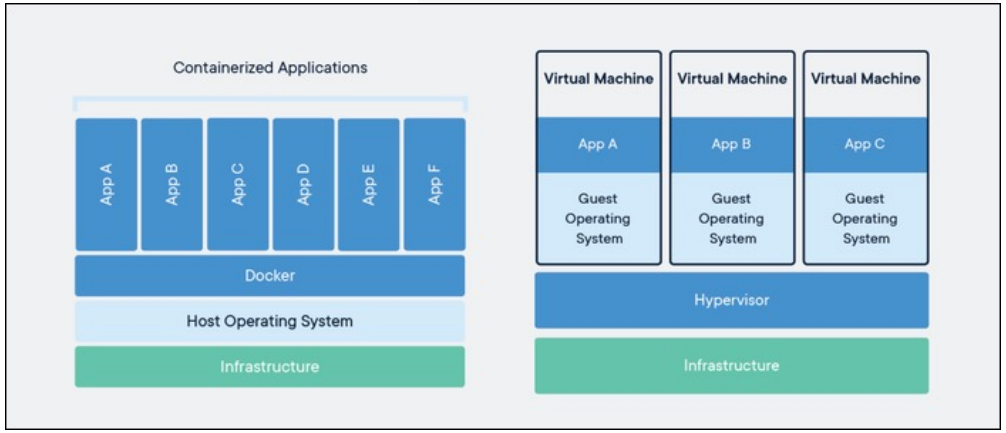
   - vGPU Request
   - GPU Availability Status



User Support

Technical assistance provided to the team

1. Fixing SSL Certificates which forbid users from accessing the service.
2. Fixing EOS mount.
3. Creating docker images with user requested features.



Upgrading Kubeflow to 1.3

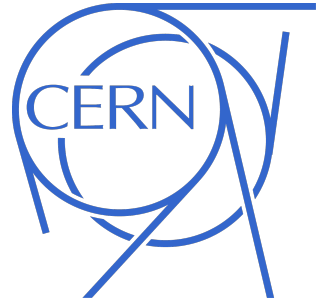
Challenges and Solutions

1. Rebuilding the testbed with requirements (Istio, Knative, etc).
2. Testing compatibility with the current infrastructure.
3. Building and testing scripts to allow easy integration into staging and production server.

Outcome

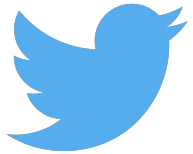
Key takeaways from the Internship

1. Explored Kubeflow, Kubernetes and multiple other new services.
2. Tested and added new examples for users new to ml.cern.ch.
3. Added new components which makes better utilization of GPUs.
4. Built and tested KF1.3 in sandbox as a base for future work.
5. Currently in final stages to release to the wider community



QUESTIONS?

Jason.praful



@jasonpraful

/jasonpraful