Metrics to define user-engagement activities on Coffea-Casa AF Deployments

- Durbar Chakraborty, NIT Durgapur

Mentors: Oksana Shadura, Ken Bloom

About the project:

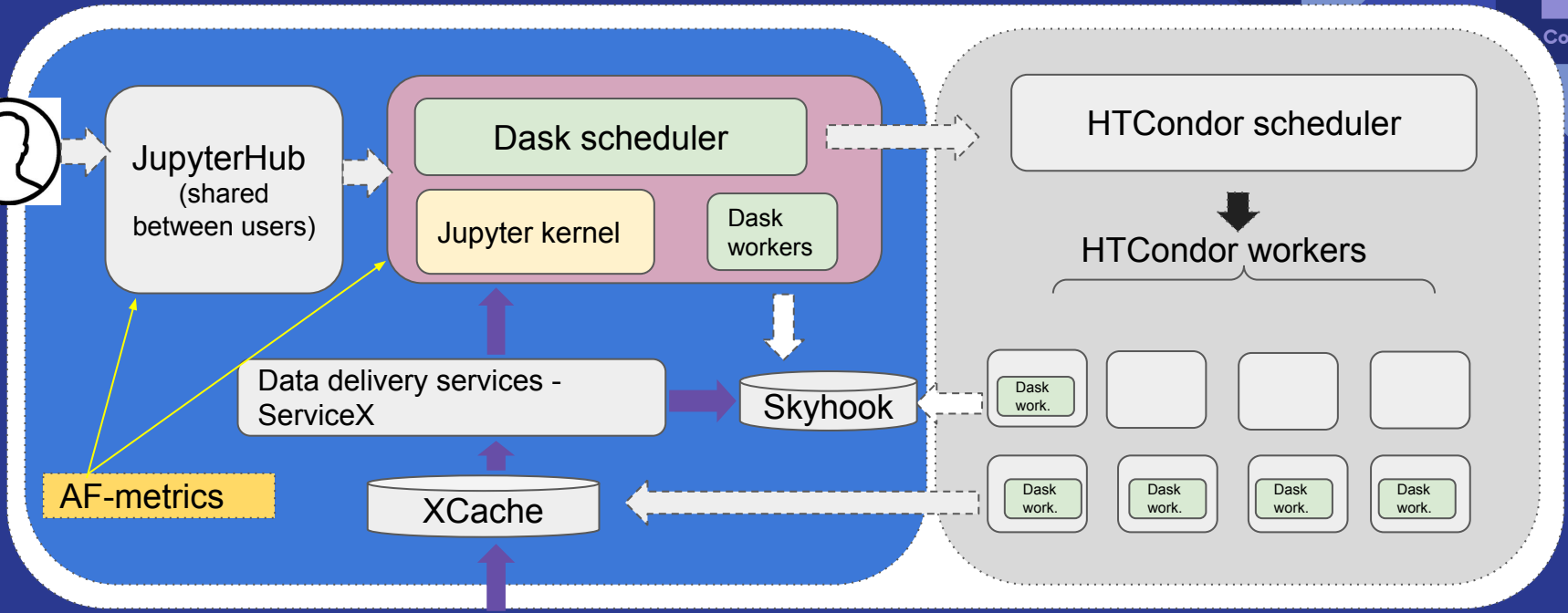
Coffea, one of the major Python packages developed by the IRIS HEP Analysis Systems domain, is a columnar object framework for analysis to enhance Data Analysis functionalities on Python for HEP event data. Coffea can be deployed and used at Coffea-Casa Analysis Facility.

This project aims to define a set of various user-engagement metrics based on the data collected from various platforms including Jupyterhub, Dask, Kubernetes clusters and etc. for Coffea-casa Facility.

We are trying to define a package “af-metrics” which will automate the monitoring and data collection process for this purpose.

Proposed Infrastructure:

Coffea-Casa



Grid / cluster site resources

Kubernetes resources

Per-user resources

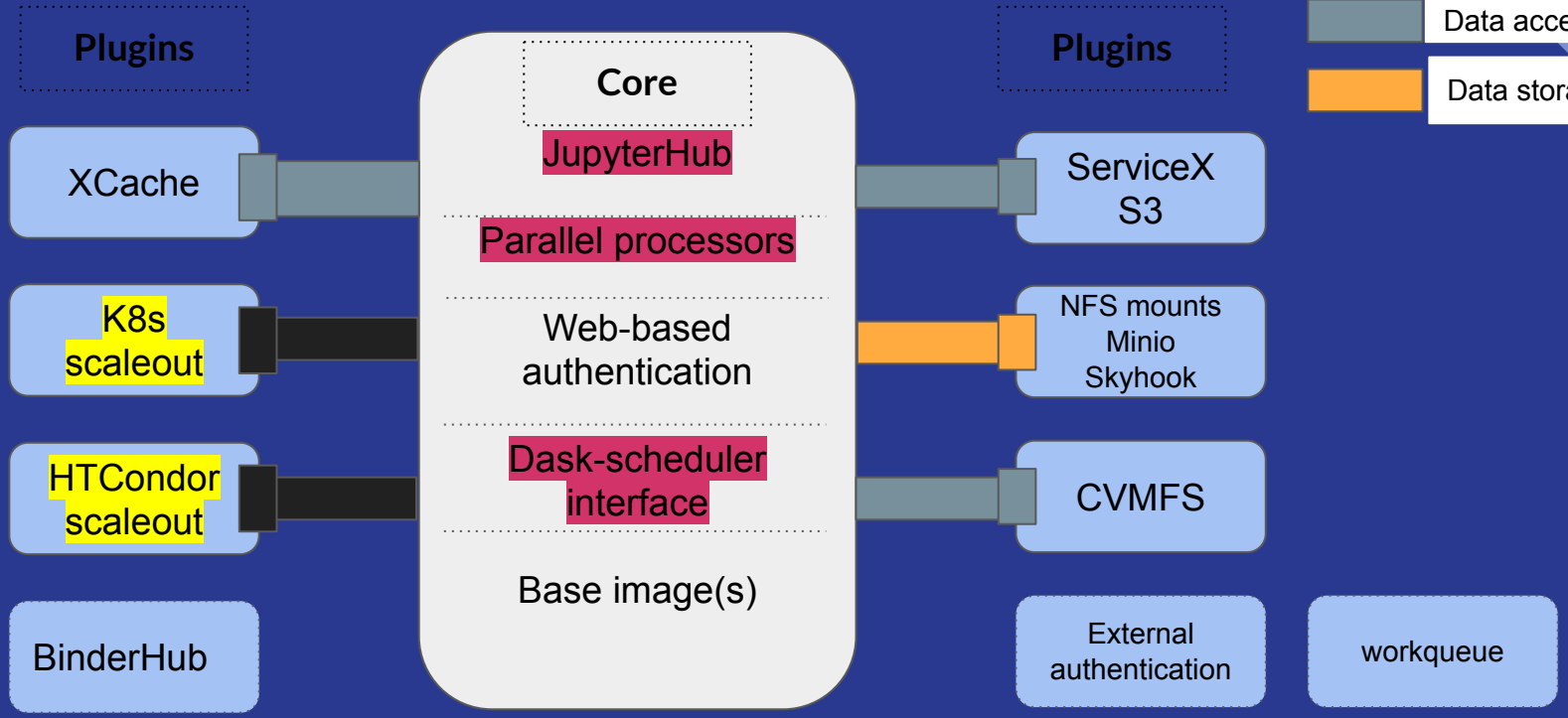
Shared resources between users

Data flow

X sends requests to Y

Matous Adamec, Garhan Attebury, Brian Bockelman, Ken Bloom, Carl Lundstedt, John Thiltges, Oksana Shadura, Andrew Wightman

Components of Coffea-casa Analysis Facility:



[Coffea-casa team]

What we have achieved so far:

- We have been able to locally host Jupyterhub and Dask and connect them to Prometheus monitoring tool and visualise the results of the telemetry operations using Grafana dashboards.
- We have been able to define custom metrics (those which are not inherently provided in the corresponding metrics endpoint) for a custom ML Model and monitor the user-defined metrics.
- We have built a python script with user-defined functions and dockerized the script with prometheus and grafana such that upon running the entire container, the monitoring process gets automated.

Metrics we have analysed:

(we will have much more…)

Jupyterhub Metrics:

- jupyterhub_server_spawn_duration_seconds: time taken for server spawning operation.
- jupyterhub_running_servers: the number of user servers currently running.
- jupyterhub_request_duration_seconds: request duration for all HTTP requests.
- jupyterhub_total_users: total number of users.
- jupyterhub_running_servers: the number of user servers currently running.

Dask Metrics:

- dask_worker_latency_seconds: Latency of worker connection.
- dask_worker_task_duration_median_seconds: Median task runtime at worker.
- dask_worker_transfer_bandwidth_median_bytes: Bandwidth for transfer at worker in Bytes

What we aim to achieve in future:

- Package the entire script such that it can be run on any facility without having to locally set up all the tools individually.
- Implement and test and host the package setup on the UNL infrastructure.
- Test user experience on Coffea-Casa Analysis Facility(s).
- Make the setup scalable so that it manages to function even in case of large dataset which is to be monitored.

References:

- https://github.com/durbar2003/Coffea-Casa-Test-Results
- https://github.com/durbar2003/af-metrics
- https://github.com/durbar2003/prometheus-monitoring-on-kubernetes-clusters