

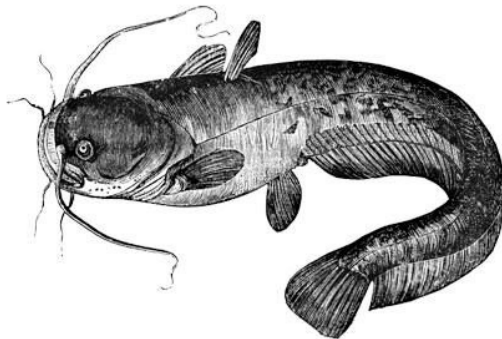


A JupyterHub-based web facility for scientific analysis for Fermilab experiments and scientists (Scientific Computing)

Maria P. Acosta EAF (an SCD project) - ACORN

55th Users Meeting

June 15th 2022



Works on
my machine

The Definitiva Guide

O RLY?

R. William

A few words about modern Scientific data analysis:

- Needs to be fast, reliable (i.e a bash terminal), secure - bonus for replicability and UI/UX features
- Requires persistent and non-persistent data storage
- Fosters collaborative environments, enables distributed teams and multi-disciplinary groups to do science using computing tools
- Work smart, improve where there's room for it.. but don't abandon the old, wise ways

Analysis Facilities for the Future

A working version of an AF definition from the [March HSF AF Forum](#) kick-off meeting:

“The infrastructure and services that provide integrated data, software and computational resources to execute one or more elements of an analysis workflow. These resources are shared among members of a virtual organization and supported by that organization.”

```
NOTICE TO USERS

This is a Federal computer (and/or it is directly connected to a
Fermilab local network system) that is the property of the United
States Government. It is for authorized use only. Users (autho-
rized or unauthorized) have no explicit or implied warranty of
privacy.

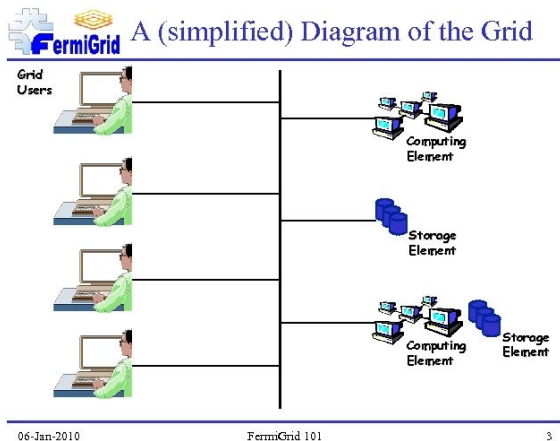
Any or all uses of this system and all data generated by this
system may be intercepted, monitored, recorded, stored, and dis-
closed to authorized site, Department of Energy personnel, as well as auth-
agencies, both domestic and foreign. User consents to such interception, mon-
ing, auditing, inspection, and disclosure of data to authorized site or Department of Energy.

Unauthorized or improper use of this system may result in disciplinary action and civil
liability. By continuing to use this system you indicate your consent to these terms and conditions.
If you do not agree to the conditions, please do not use the system.

Fermilab policy and rules for computer use, may be found at http://www.fnal.gov/

...Powered by CMS-LPC...

Hostname: cmslpc118.fnal.gov      OS Release: 3.10.0-1160.66.1
IP: 131.225.189.99              RAM: 11.57 GiB
Kernel: 3.10.0-1160.66.1         Cores: 8
SSH Logins: 1                    Load: 0.00
For information about computing at the LPC:
08:18:24 macos@cmslpc118:~$
```



Some examples:

- LPC CAF
- Experiment VM nodes
- Custom submit points for Condor jobs (i.e. DES)
- Local bash/sh, IDEs, Jupyter Notebooks

Fermilab's AF -- Fundamental principles:



- Create a user-oriented analysis facility based on our own experience supporting scientists on traditional technologies.
- Explore, deploy and collaborate on industry-level tools and strategies for optimizing data analysis.
- Facilitate the use and access of a pool of large, specialized hardware for all Fermilab users in an Elastic way.
- Foster collaboration with experiments and science groups in order to better understand current and future analysis needs.
- Provide effective, requirement-oriented computing solutions.

Secure

Integrated & functional

Multi-VO

DevOps (operational
sustainability)

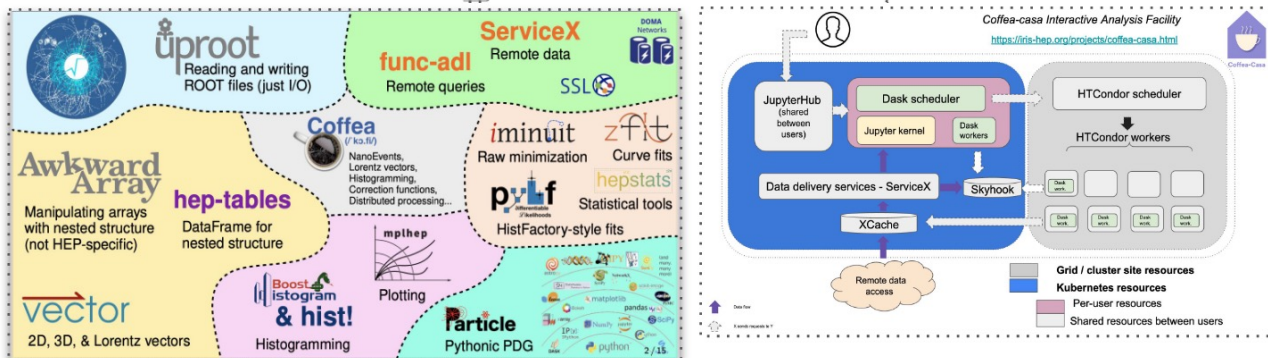
Active collaboration

The ultimate goal:

Analysis Tools

Analysis Facilities

(coffea-casa AF or any other facility matching tech.requirements)



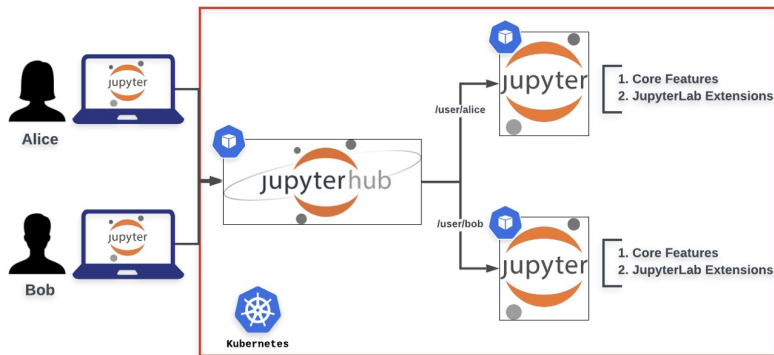
LArSoft

jupyterhub



Fermilab

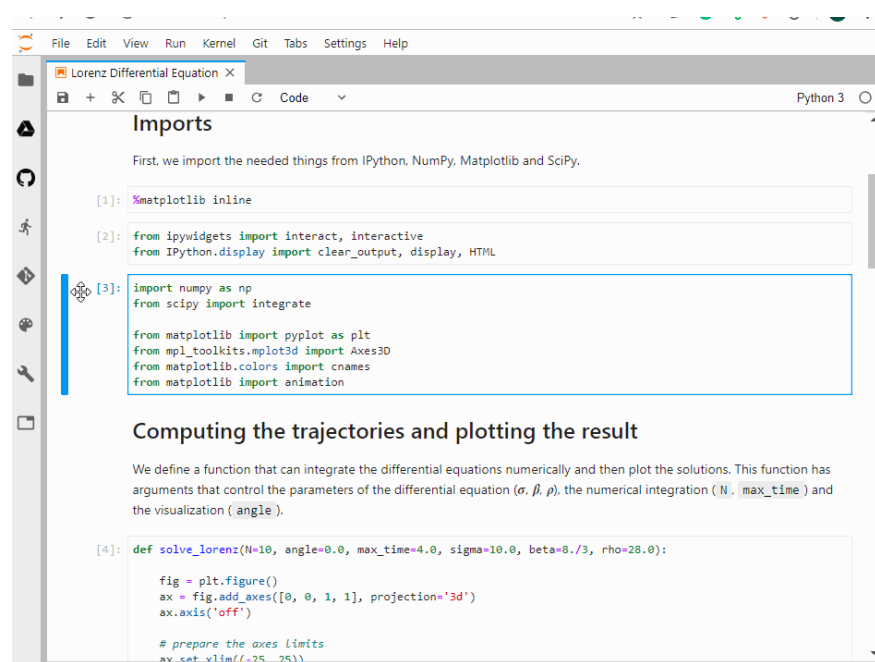
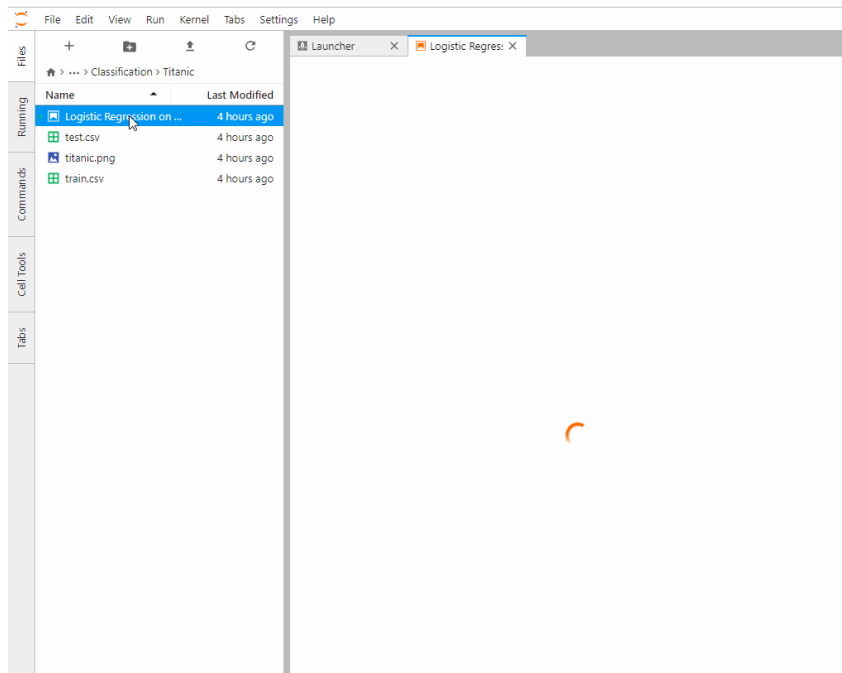
A JupyterHub-based deployment



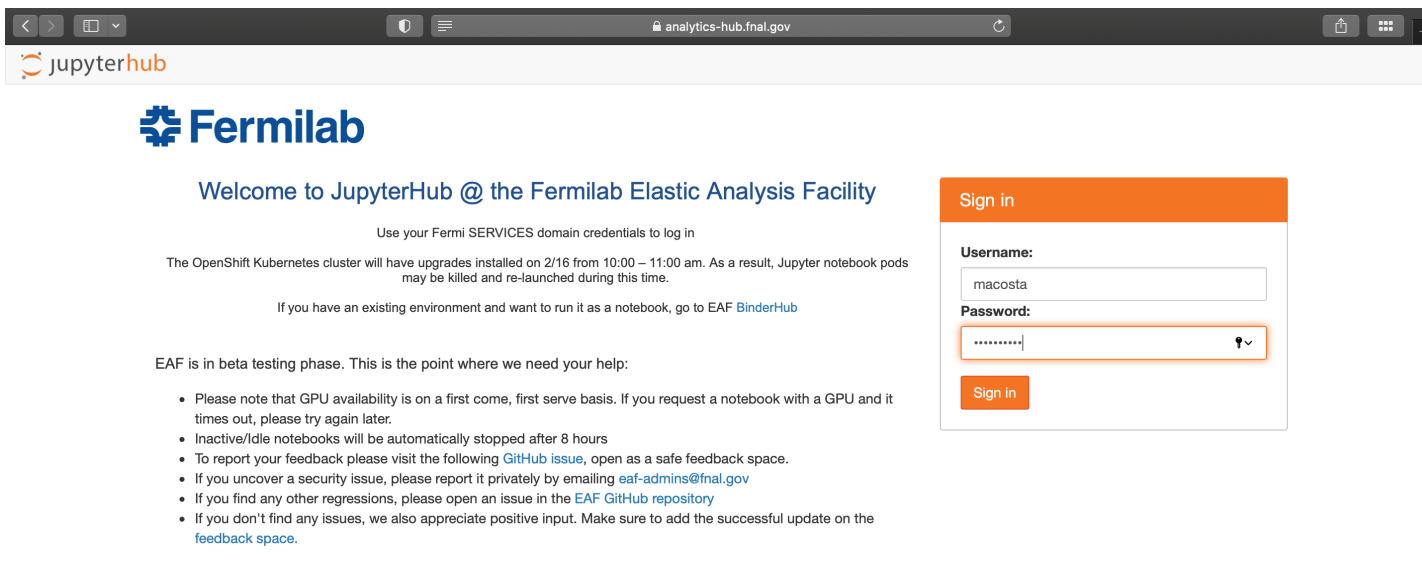
- Originally standalone Jupyter Notebooks.
- Evolved to a self-hosted, multi-user platform for hosting multiple notebooks, kernels and **highly customizable** environments.
- Can be deployed in multiple platforms including Cloud, on prem and Kubernetes.

- ✓ Implements authentication, login pages and token-based roles
- ✓ Tracks activity and does effective resource management
- ✓ Proxying is done behind the scenes

A JupyterHub-based deployment



A JupyterHub-based deployment - Login and Auth



The screenshot shows a web browser window with the URL `analytics-hub.fnal.gov`. The page header includes the JupyterHub logo and the Fermilab logo. The main content area has the following text:

Welcome to JupyterHub @ the Fermilab Elastic Analysis Facility

Use your Fermi SERVICES domain credentials to log in

The OpenShift Kubernetes cluster will have upgrades installed on 2/16 from 10:00 – 11:00 am. As a result, Jupyter notebook pods may be killed and re-launched during this time.

If you have an existing environment and want to run it as a notebook, go to [EAF BinderHub](#)

EAF is in beta testing phase. This is the point where we need your help:

- Please note that GPU availability is on a first come, first serve basis. If you request a notebook with a GPU and it times out, please try again later.
- Inactive/Idle notebooks will be automatically stopped after 8 hours
- To report your feedback please visit the following [GitHub issue](#), open as a safe feedback space.
- If you uncover a security issue, please report it privately by emailing eaf-admins@fnal.gov
- If you find any other regressions, please open an issue in the [EAF GitHub repository](#)
- If you don't find any issues, we also appreciate positive input. Make sure to add the successful update on the [feedback space](#).

On the right side, there is a "Sign in" form with the following fields:

Sign in

Username:

Password:

- Accessible from the Lab network or via VPN
- Login with SERVICES account
- UID/GID will be propagated to the notebook in order to preserve permissions

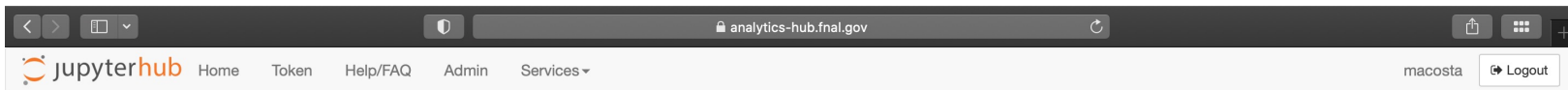
A JupyterHub-based deployment - Current Catalog

The screenshot shows the JupyterHub interface for the analytics-hub.fnal.gov instance. The page is titled "Server Options" and displays a grid of server configurations for different projects. Each project has a list of available server types, each with a radio button for selection. The projects and their options are:

- ACCEL-AI** (Teal buttons):
 - L-CAPE CPU Only SL7 Interactive
 - L-CAPE GPU SL7 Interactive (NVIDIA Tesla K40)
 - L-CAPE GPU SL7 Interactive (NVIDIA Tesla T4)
 - READS CPU Only SL7 Interactive
 - READS GPU SL7 Interactive (NVIDIA Tesla K40)
 - READS GPU SL7 Interactive (NVIDIA Tesla T4)
- DES/LSST/ASTRO** (Yellow buttons):
 - SL7 Interactive General Purpose Notebook
 - GPU SL7 Interactive (NVIDIA Tesla K40m)
 - GPU SL7 Interactive (NVIDIA Tesla T4)
- CMSLPC** (Dark red buttons):
 - SL7 Interactive
 - COFFEA-DASK SL7 Interactive
 - GPU SL7 Interactive (NVIDIA Tesla K40m)
 - GPU SL7 Interactive (NVIDIA Tesla T4)
 - GPUaaS - Boosted Decision Trees SL7 Interactive (NVIDIA Tesla T4)
- LBNF/DUNE/ProtoDUNE** (Orange buttons):
 - SL7 Interactive General Purpose Notebook
 - GPU SL7 Interactive (NVIDIA Tesla T4)
 - GPU SL7 Interactive (NVIDIA Tesla K40m)
- FIFE/Neutrinos** (Green buttons):
 - SL7 Interactive General Purpose Notebook
 - GPU SL7 Interactive (NVIDIA Tesla K40m)
 - GPU SL7 Interactive (NVIDIA Tesla T4)
- Fermi generic SL7/CC8** (Dark blue buttons):
 - Basic SL7 Interactive
 - Basic CC8 Interactive

At the bottom of the page is a large orange "Start" button.

A JupyterHub-based deployment - Named servers



Stop My Server

My Server

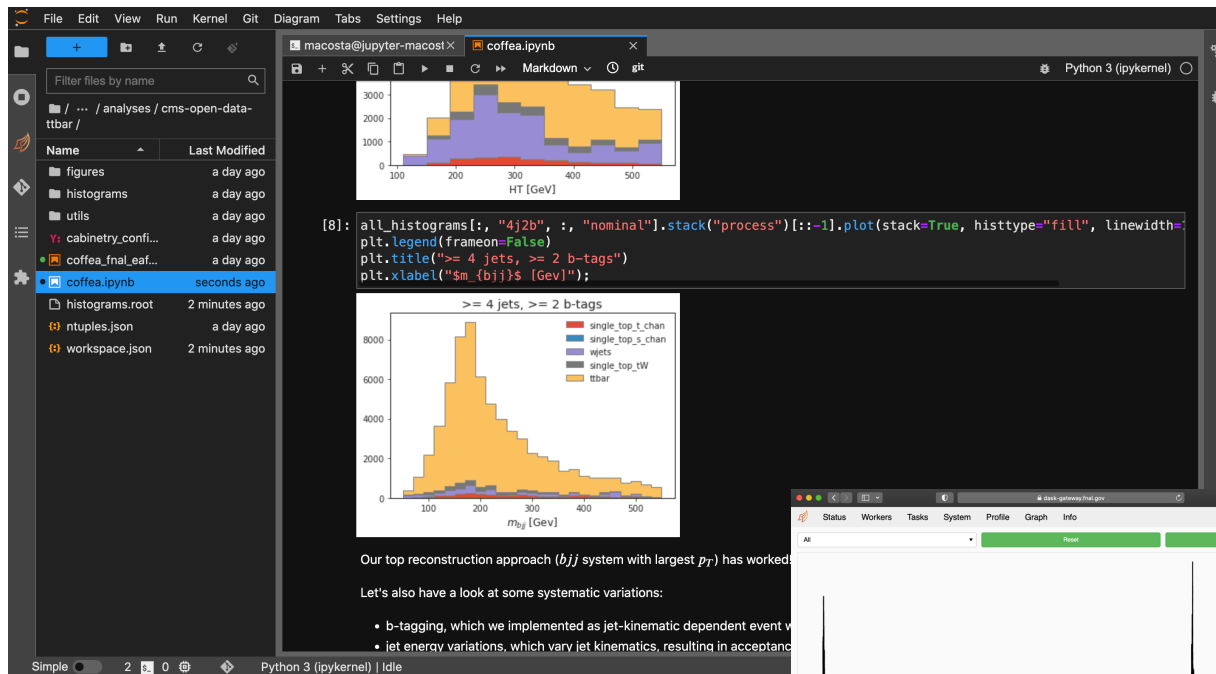
Named Servers

In addition to your default server, you may have additional 5 server(s) with names. This allows you to have more than one server running at the same time.

Server name	URL	Last activity	Actions
<input type="text" value="Name your server"/>	Add New Server		
ad		21 days ago	start delete
dask	/user/macosta/dask	a day ago	stop
dune		a month ago	start delete
fife		3 months ago	start delete
lpc		5 days ago	start delete

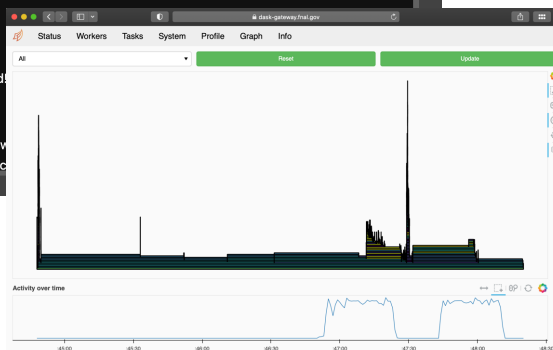
- Up to five, independent, isolated environments with shared persistent storage
- Activity monitoring and Application Token page
- CVMFS, Home areas and other specialized software will be included in the notebook

A JupyterHub-based deployment (on Beta) <https://analytics-hub.fnal.gov>

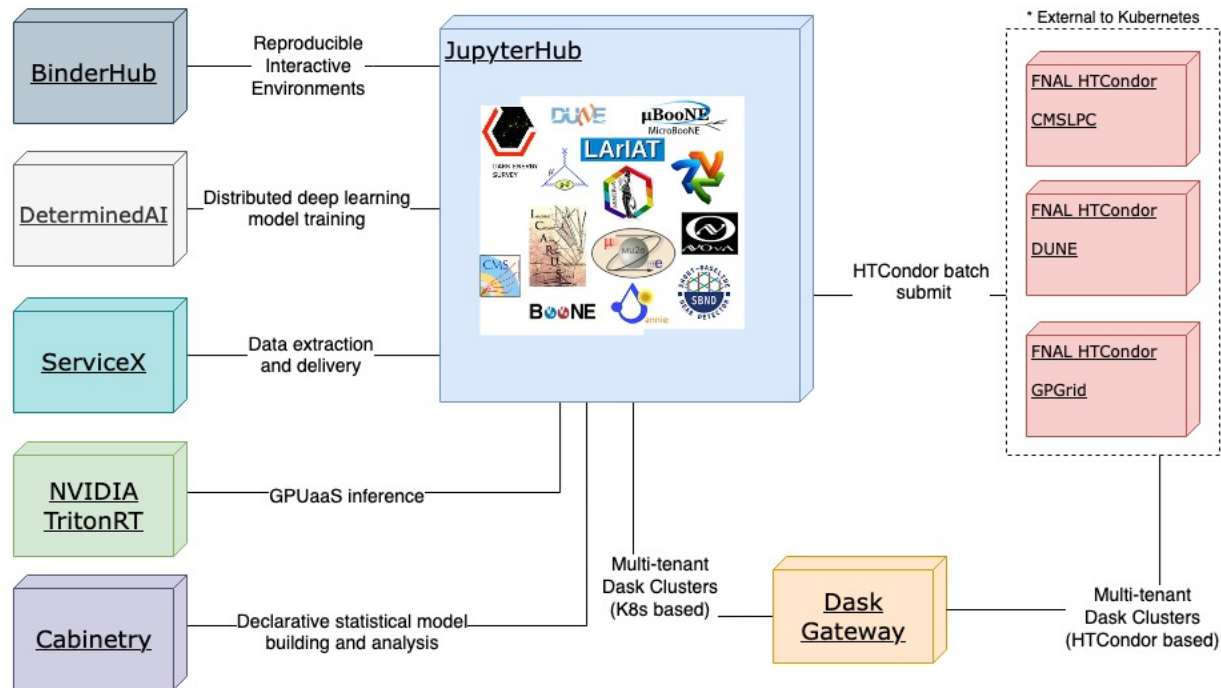


CMSLPC notebook running AGC COFFEA analysis

- 43 Beta users (thank you!)
- 22 Notebook flavors
- 1.2 Tb Ceph persistent storage allocated (of 45TB)

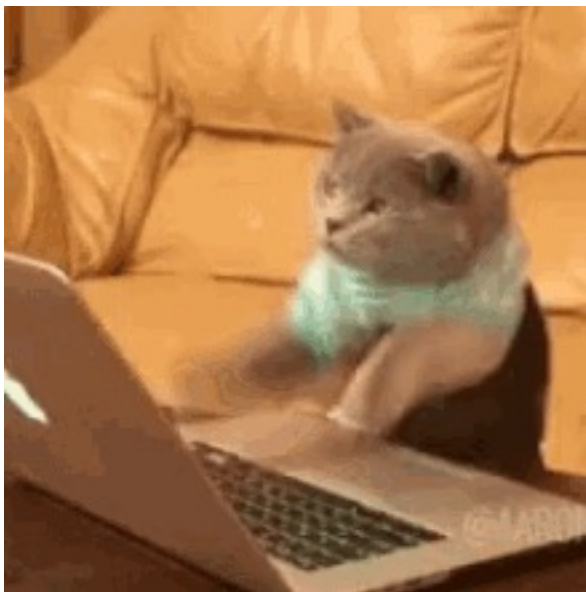


Current applications Ecosystem



We need your help!

- Have a project that could benefit from JupyterHub?
- Are you tired of running local jupyter notebooks?
- Is there a computing need or requirement that fits the AF model?



Contact us!

- Email me (macosta@fnal.gov) and Burt Holzman (burt@fnal.gov) with your thoughts!
- If you uncover a security issue, please report it privately by emailing eaf-admins@fnal.gov.
- If you find any other regressions, please open an issue in the [EAF GitHub repository](#).
- If you don't find any issues, we also appreciate positive input. Make sure to add the successful update on the [feedback space](#).

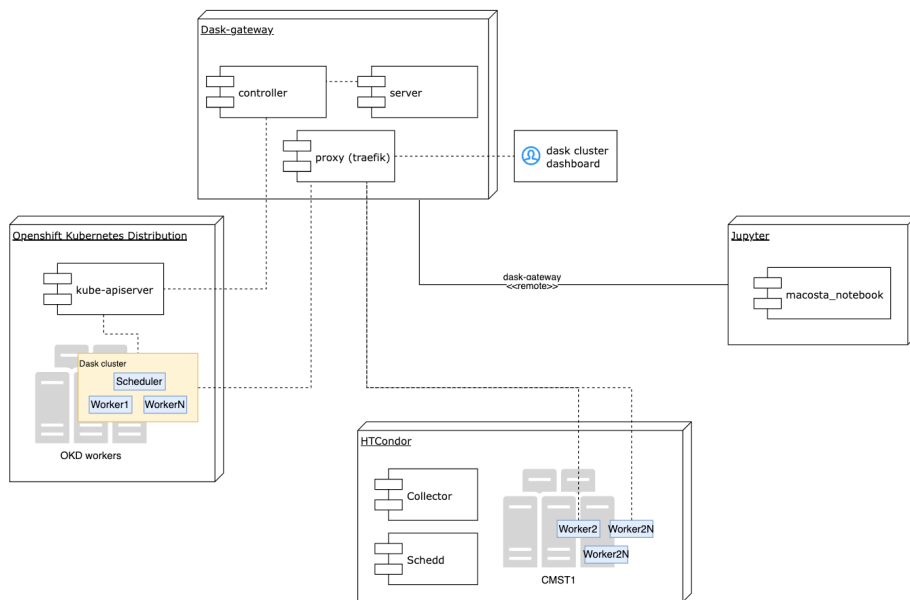
Closing thoughts

- Enabling data analysis tools and platforms and optimizing current CMS Software stack in preparation for **HL-LHC is a huge motivator, but not the only one.**
- DUNE, mu2e, cosmic frontier and **other areas of the lab will benefit** from a flexible, web-based terminal with Python engines, centralized authentication and authorization, shared home mounts and persistent storage.
- Opportunities for innovation and broader collaboration with industry and inter-division groups.
- Computing needs to understand needs of science to properly support it, this is a joint effort, and everyone is a part of it.
- Fermilab (USCMS) is leading effort on R&D and supporting other institutions on the implementation of Afs. Guidelines COFFEA-CASA at UNL and the Fermilab EAF, pioneer projects.

Thanks 😊 Questions?

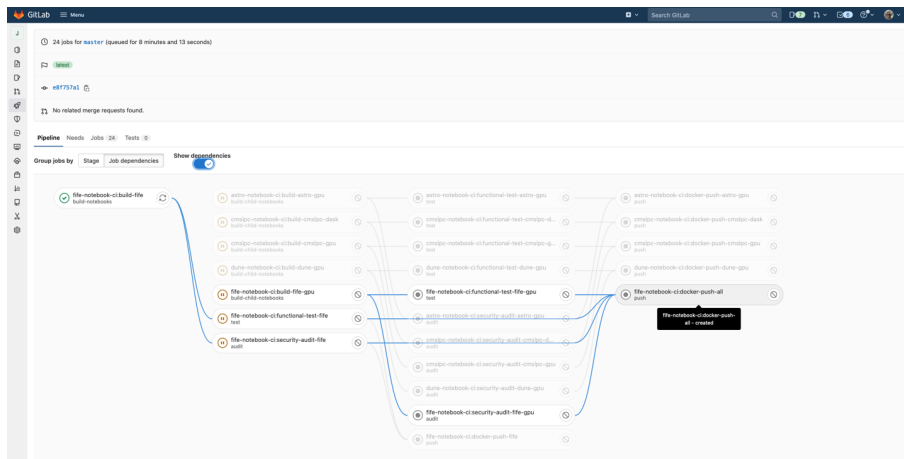
Maria Acosta - SCS/CSI
macosta@fnal.gov
@macosta on Slack

Backup - Hybrid Dask Clusters



- Ipcdaskgateway is a client extension for Dask Gateway which enables CMS users to dynamically obtain Dask compute resources from the LPC pool in the form of containerized dask worker jobs and from Kubernetes if they need/prefer.
- Latest version is installed by default and deployed to the COFFEA-DASK notebook on EAF.
- We are working on contributing multiple patches upstream as a result of this R&D work

Backup - GitLab analytics for CI/CD

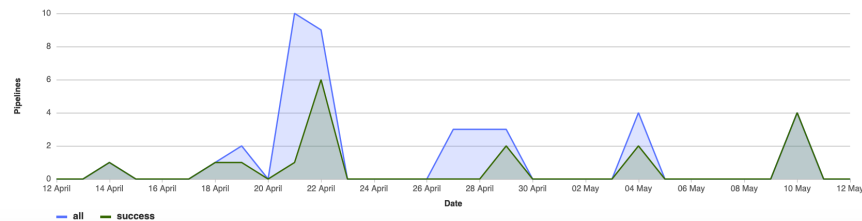


A partial pipeline with dependencies (This thing has automated ~90 hours of my time)

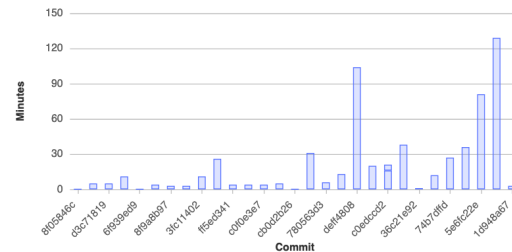
Pipelines charts

Last week Last month Last year

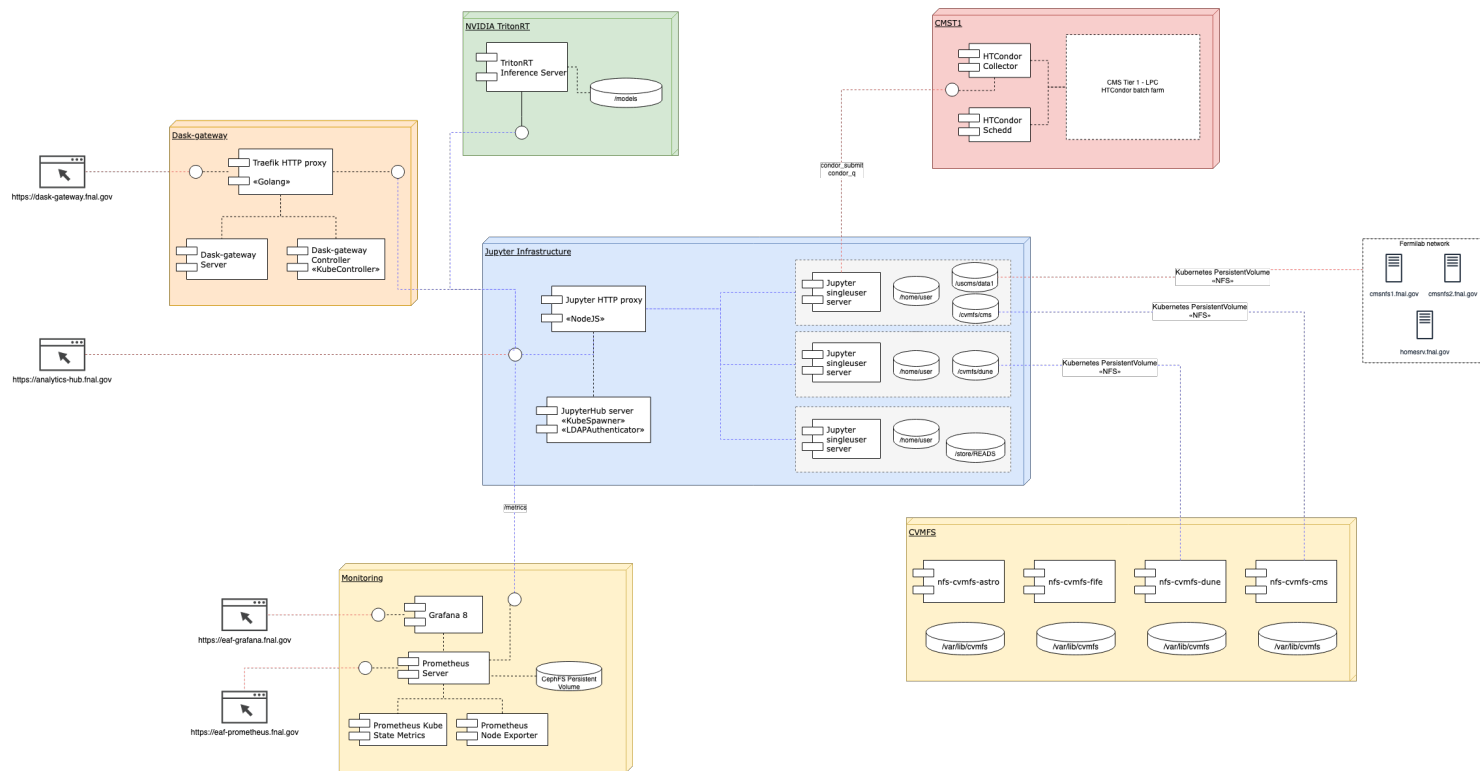
Date range: 11 Apr - 12 May



Pipeline durations for the last 30 commits



Backup – detailed component diagram



Backup - Dask cluster burst – the ‘Elastic’ side of the facility

