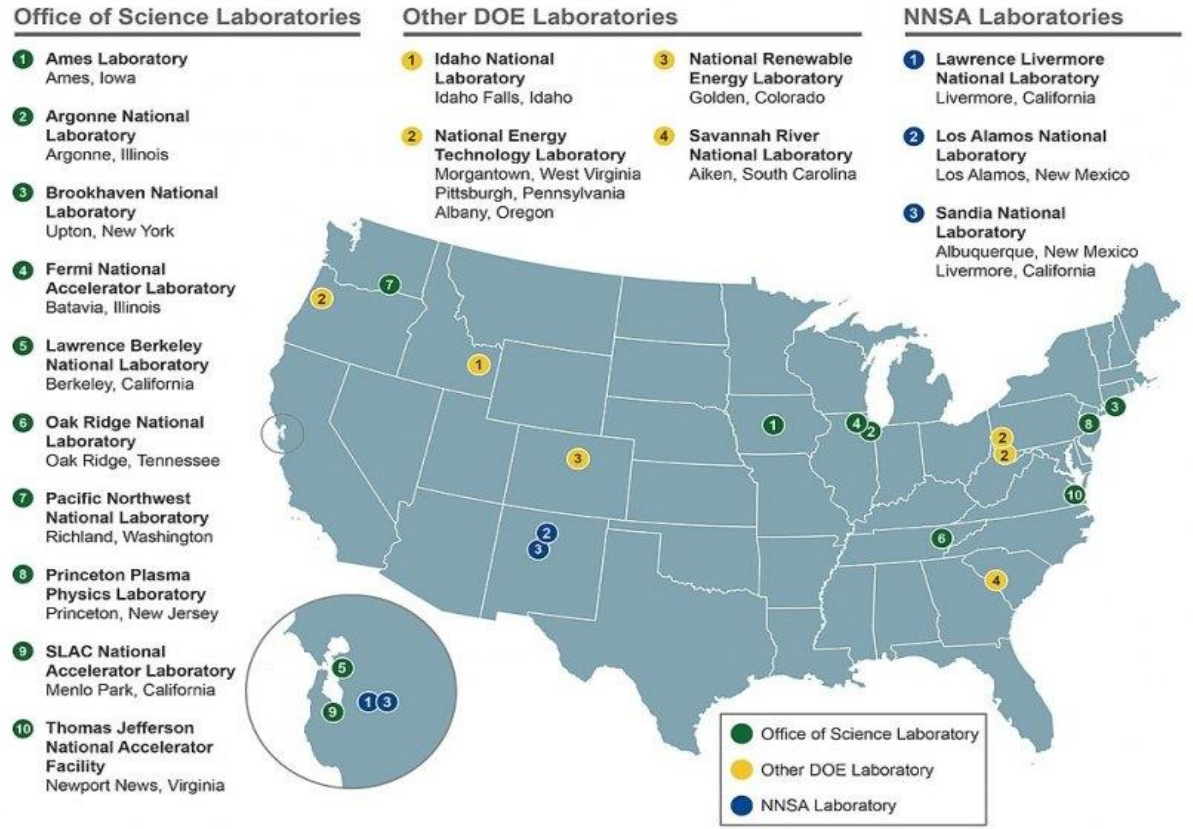
Analysis Facilities activities in (some) DOE multi-purpose computing centers

Doug Benjamin (BNL), **Burt Holzman** (FNAL), Ofer Rind (BNL), Wei Yang (SLAC)

March 25, 2022

In partnership with:

DOE National Laboratories



- 17 national labs
- 4 with large HEP funding: **Fermilab**, **Brookhaven**, **SLAC**, Lawrence Berkeley
- Will highlight work at first 3 today (but we should have LBNL at next forum!)

Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab

Analysis Facilities at National Labs

- Pre-existing computing facilities
  - **Long history** of providing user analysis facilities

RHIC Computing Facility (RCF)

➢ Organizationally established in 1997

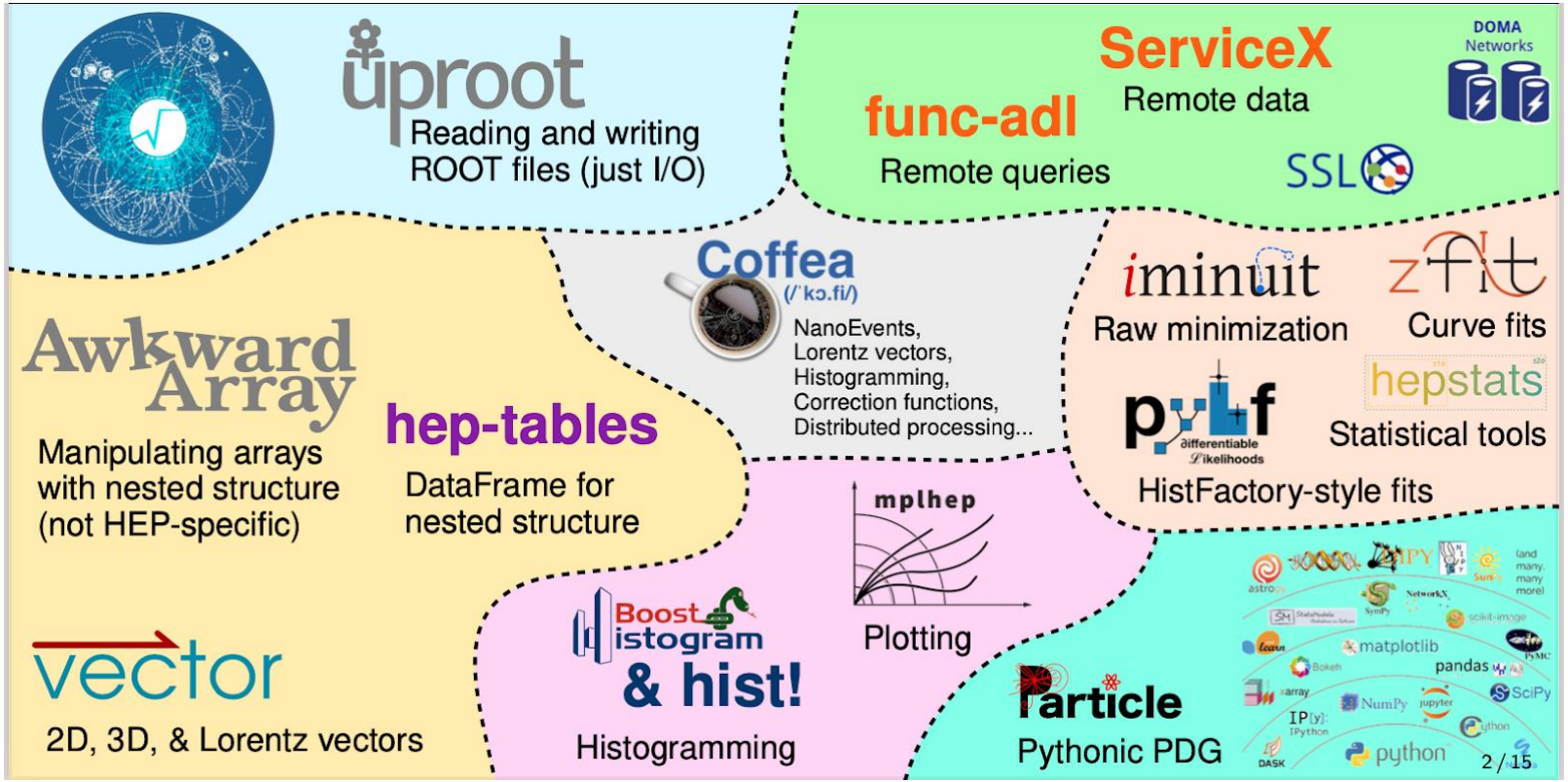
The first scientific non-data computer acquisition by the Laboratory occurred in 1970. About $500K had been allocated for the acquisition of a medium-sized computer to service the bubble-chamber film-measuring and analysis needs generated by FAF. The

  - Today we will focus on the **AFs in development** (support fast columnar analyses) that complement our existing AFs
- Security
  - As .gov sites, labs are generally subject to increased scrutiny and oversight
  - Certification of software / path to FedRAMP certification is helpful
- Multi-tenancy
  - Serve broad communities, not single experiments (and not necessarily just HEP)

Fundamental principles:

- Create a user-oriented analysis facility based on our own experiences supporting scientists on traditional grid technologies.
- Explore, deploy and collaborate on industry-level technologies and strategies for optimizing data analysis partly in preparation for HL-LHC and upcoming experiments with large data demands such as DUNE.
- Foster collaboration with HEP experiments in order to better understand science analysis needs and provide computing solutions accordingly.

Analysis Systems Ecosystem



Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab

Brookhaven National Laboratory

Brookhaven National Laboratory

Overview of US ATLAS Analysis Facilities

US ATLAS has three shared Tier 3 analysis facilities providing software & computing

- Resources that fill gaps between grid jobs and interactive analysis on local computers
- Interactive ssh login, local batch, non-grid storage, LOCALGROUPDISK, PanDA
- GPU resources available
- Documentation: https://usatlas.readthedocs.io/projects/af-docs/en/latest/

DOE

BNL Facility
~2000 cores, part of a larger shared pool, opportunistic access up to 40k cores
User quota: 500GB GPFS plus 5TB dCache
~200 users

SLAC Facility
~1200 cores, part of larger shared pool, opportunistic access up to 15k cores
User quota: 100GB home, 2-10TB for data
~100 users

Launched Oct 2021

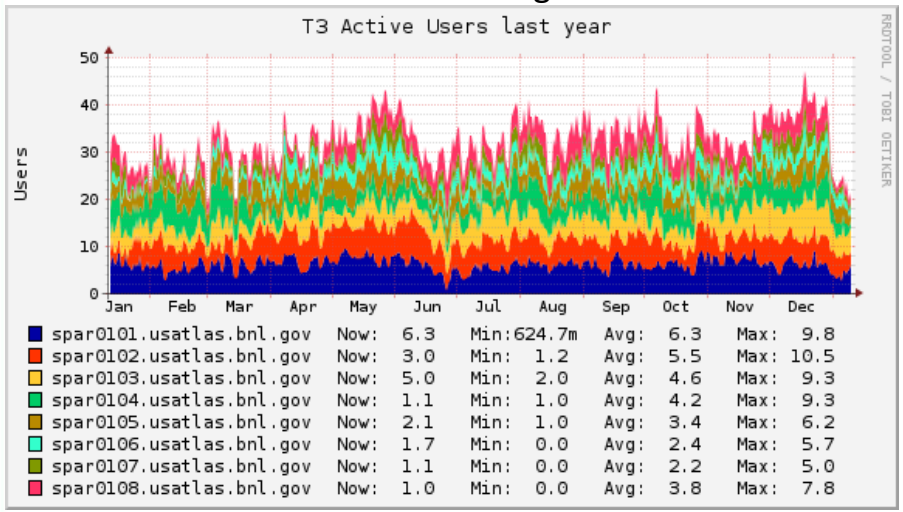
U Chicago Facility
~1000 cores, co-located and close integration with MWT2
User quota: 100GB home, 10TB for data
~50 users

Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab

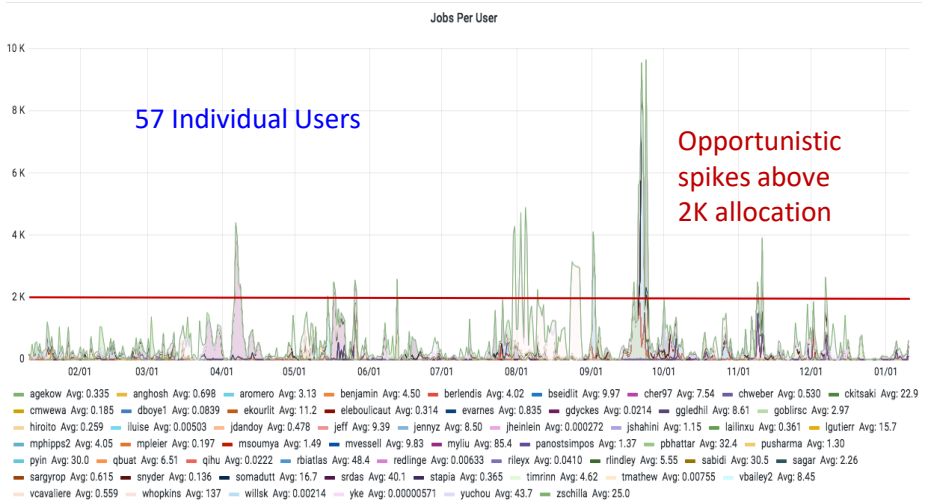
BNL Tier-3

"Traditional" ssh + interactive/batch usage

Total Number of Interactive Users
across BNL Tier-3 Login Hosts

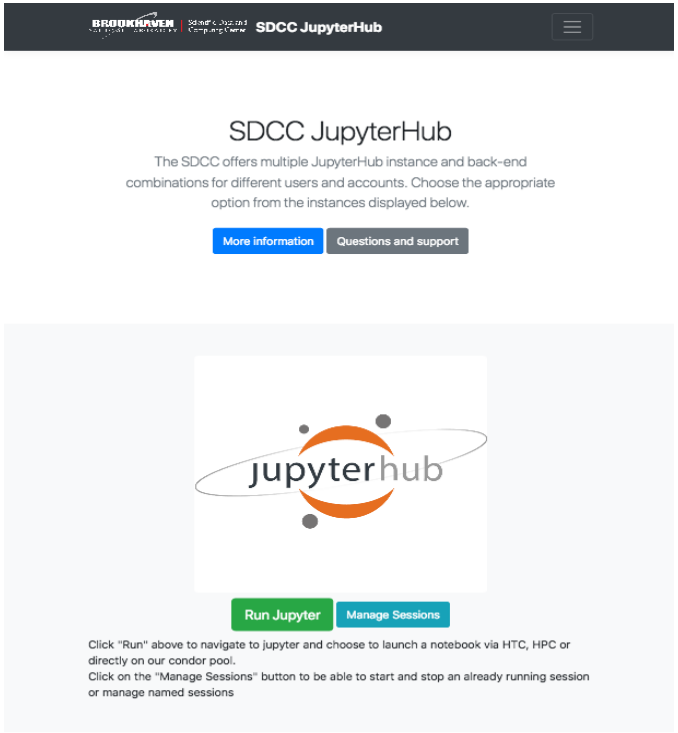


Number of Running Condor
Jobs Per Tier-3 User



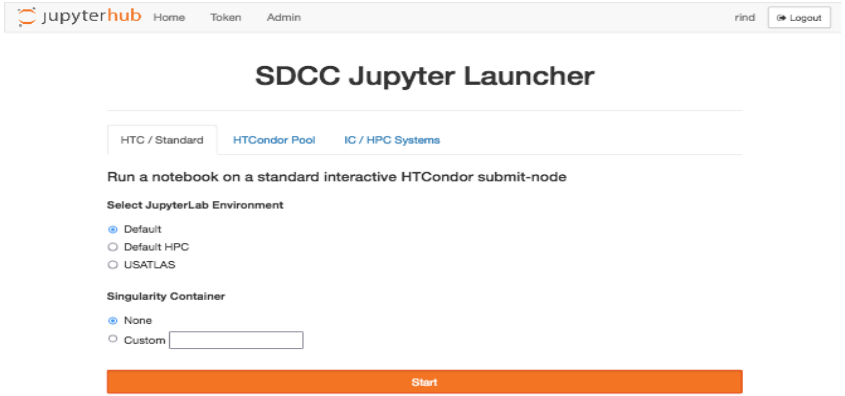
Brookhaven National Laboratory
SLAC NATIONAL ACCELERATOR LABORATORY
Fermilab

BNL SDCC Jupyterhub Portal



Custom Jupyterhub interface to multiple resources backends with account-based access control:

- Dedicated set of Jupyterhub nodes
  - Dask scheduling onto HTCondor pools
- Notebook spawning onto HTCondor pools
- Notebook spawning onto HPC cluster via Slurm
- Shared environment with SLAC on BNL cvmfs server

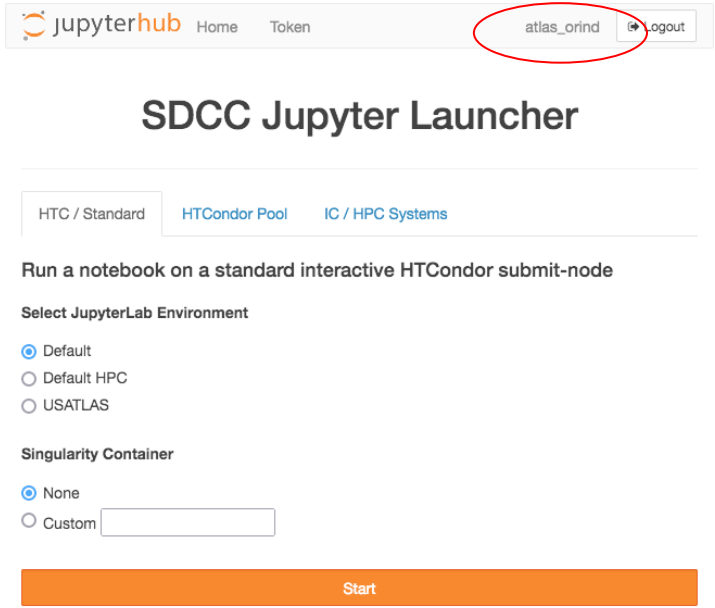


Brookhaven National Laboratory
SLAC NATIONAL ACCELERATOR LABORATORY
Fermilab

Federated Access

- Preproduction Jupyterhub with federated login in testing at BNL (accepts CERN, SLAC, FNAL logins)
  - Access for all ATLAS users collaborating with a US institution
  - Simple form for creating lightweight AF account
  - DOE lab policy requires multi-factor auth for interactive access

Automatic account creation based on CERN login

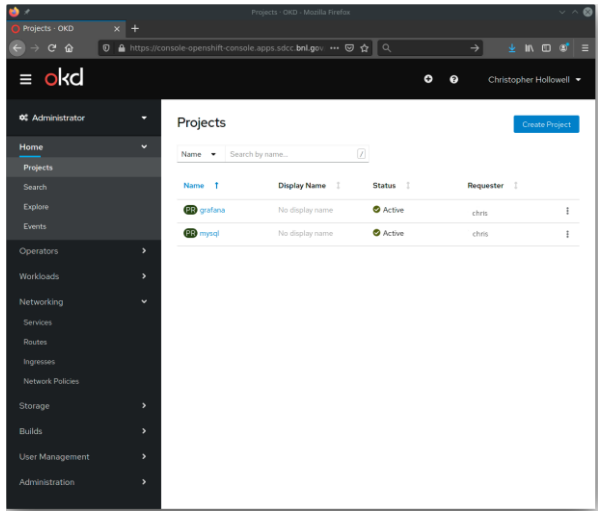


Brookhaven National Laboratory
SLAC NATIONAL ACCELERATOR LABORATORY
Fermilab

OKD at BNL SDCC

OKD provides a platform for container orchestration, similar to Kubernetes (k8s)

- Community-supported release of Openshift
- Allows for simplified deployment of services via helm charts and Openshift templates
- Contains numerous security enhancements out of the box vs k8s
  - Users are never root by default



OKD 4.7 cluster online at SDCC

- 4 user nodes - 20C/40T, 128 GB RAM

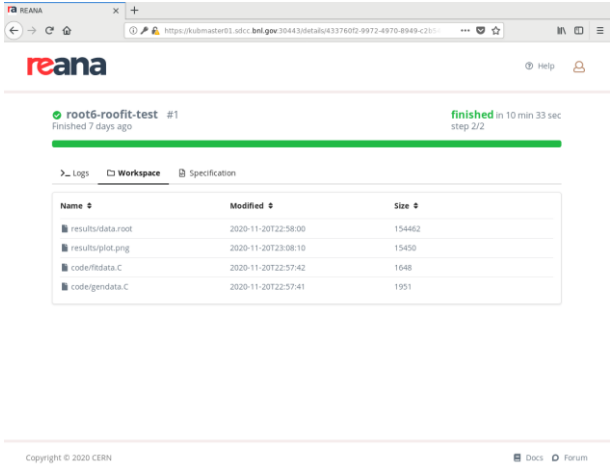
- 15 TB NMVE storage (NetAPP 250)
- Console authentication tied to our keycloak IDP
- Currently only accessible from inside BNL

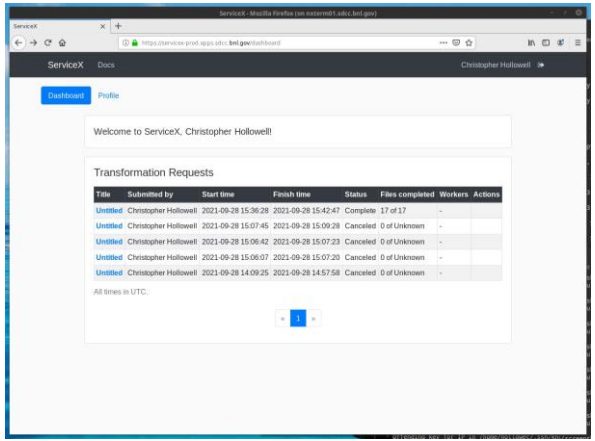
Brookhaven National Laboratory
SLAC NATIONAL ACCELERATOR LABORATORY
Fermilab

AF at BNL SDCC on OKD

REANA testbed

- User testing in progress
  - Web interface currently accessible via SSH tunnel/SOCKS proxy
- Can interface and submit container jobs to SLURM on the HPC cluster





ServiceX test deployment

- Authentication using SDCC IDP
- Currently only available from within the SDCC network
- Working with developers on integrating needed changes to containers and helm charts to support OKD upstream
  - Currently only in our modified deployment

Brookhaven National Laboratory
SLAC NATIONAL ACCELERATOR LABORATORY
Fermilab

SLAC National Accelerator Laboratory

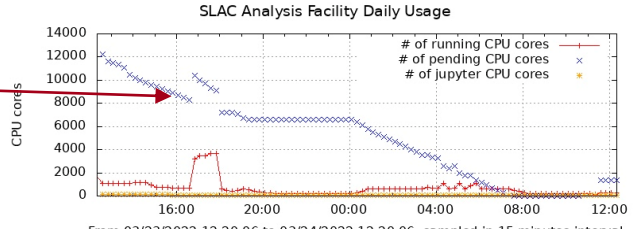
SLAC Analysis Facility

- SLAC AF established in 2017
  - Serving US ATLAS users and their international collaborators
  - Initially inherit hardware from US ATLAS Western Tier 2. Most of them are retired now.
  - Added new storage (800TB), CPU (512cores)
    - adding a GPU node
  - 40-70 active users (depend on how/what to count)
- Services:
  - Traditional login/batch: ssh, SLURM, user home and private data space
  - Grid services: RSE & Xcache (for users to access official ATLAS data products)
  - Jupyter/DASK: a user expandable Jupyter environment, back by SLURM
    - Kernels to access ATLAS release (needs updates), Tensorflow/Keras, Rapids
    - These kernels are shared with BNL via CVMFS.
  - Opportunistic access to a large pool of CPUs and GPUs
- Support:
  - SLAC specific document site and email/ticket
  - US ATLAS AF document site in github/readthedocs.io
  - Mattermost channel/e-mail support, to be moved to a central US ATLAS Discourse service

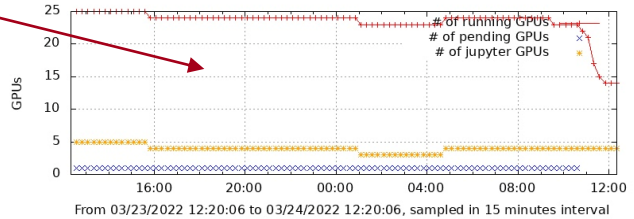
SLAC Analysis Facility

- Heavy Jupyter/DASK usage
  - More Jupyter/DASK users than batch users in 2021
- Users are utilizing SLAC's large pool of GPUs

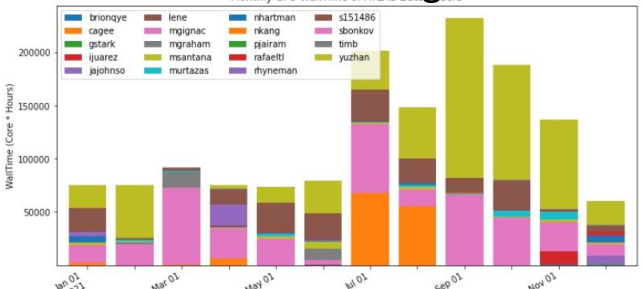
DASK jobs

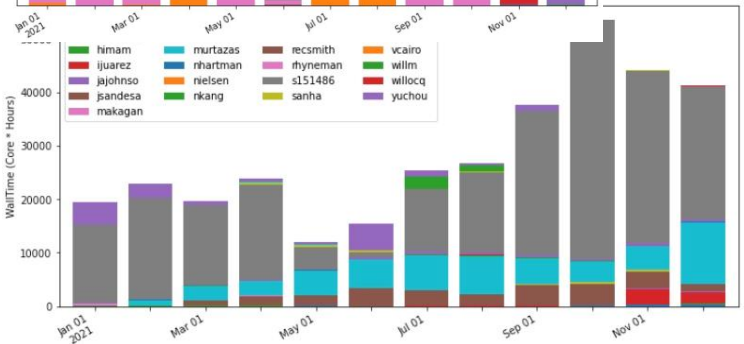


GPU jobs

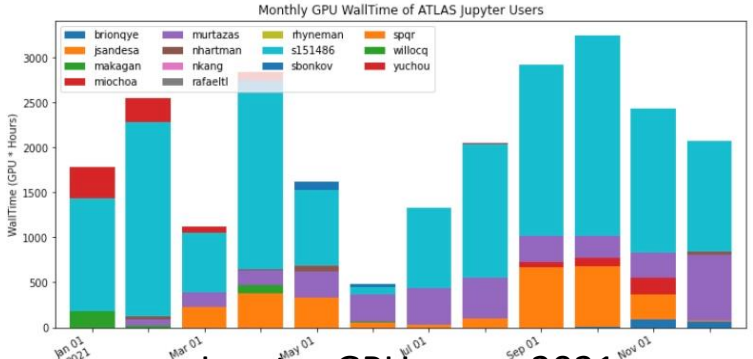


Batch CPU usage 2021





Jupyter CPU usage 2021



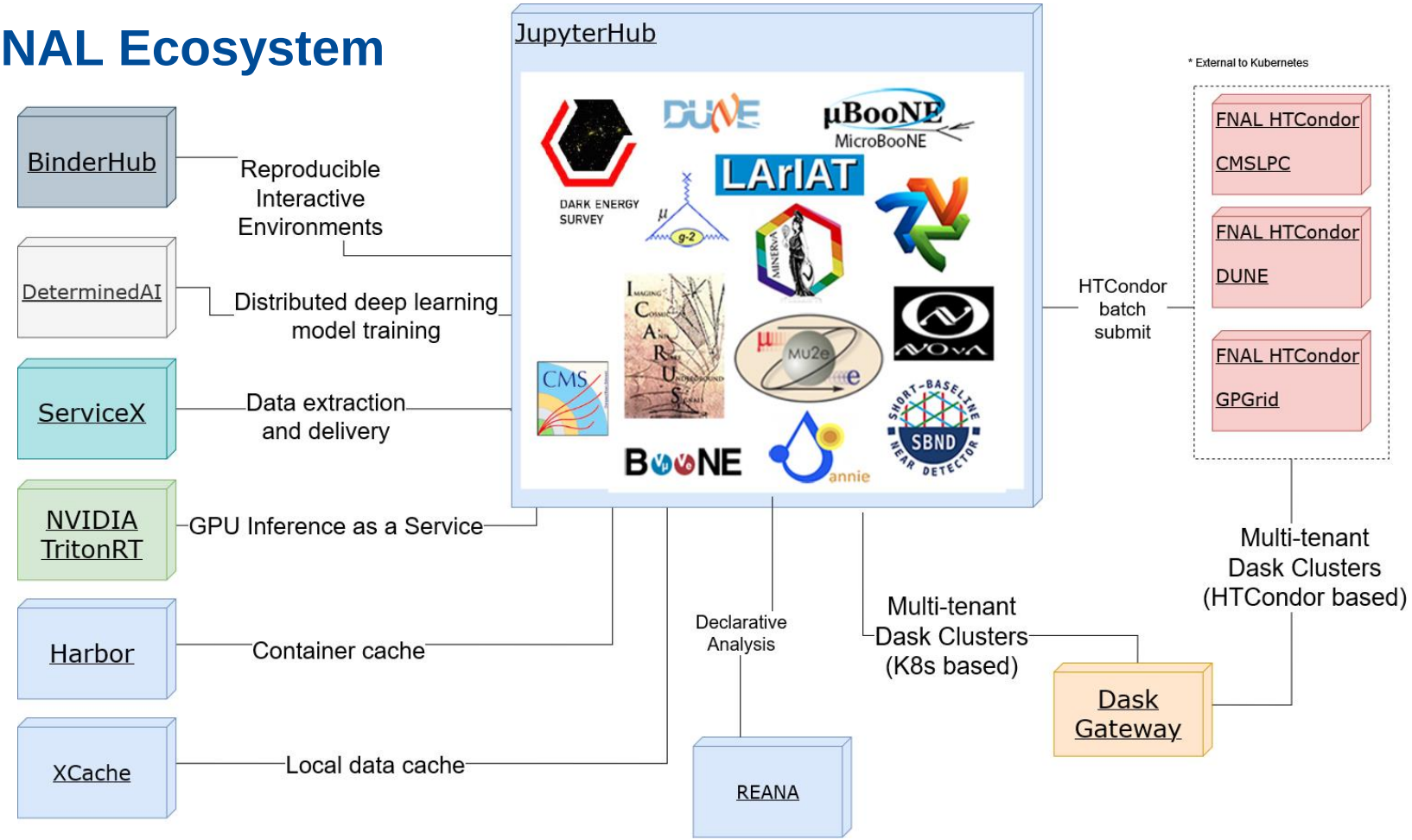
Jupyter GPU usage 2021

SLAC Analysis Facility

- Modernizing the SLAC overall computing infrastructure
  - SLAC AF is embedded in the SLAC computing infrastructure. So SLAC AF will benefits
- Storage:
  - Moving from a mixed of AFS/GPFS/Lustre for $HOME to Weka SSD based home.
  - Have very positive experience with Weka posix home at LCLS
    - Hopefully this will significantly boost the start up speed for Python
  - Not completely settled for data space.
    - Object store is preferred by Rubin. But Weka may provide a tiered posix with hidden S3 backend.
- Jupyter environment:
  - Currently using OpenOndemand to spawn Jupyter on SLURM nodes (as SLURM jobs).
    - Good: utilize SLURM scheduling
    - Bad: exclusive allocation of GPUs is expensive (even for a 400+ GPU cluster)
  - Maybe looking at traditional methods: allowing several users to freely login to a few shared GPU nodes.
- Authentication/Authorization:
  - Looking at Federated Access. A grey area with a policy side and a technical side. On tech side:
    - Easy for web based access but hard for SSH
    - Experimenting ways to integrate SSH
- K8s for future service
  - Like will use k8s to host most of the services. VM are still needed in some cases
  - This is mostly driven by Rubin but ATLAS and other will benefits from the experience/expertise.

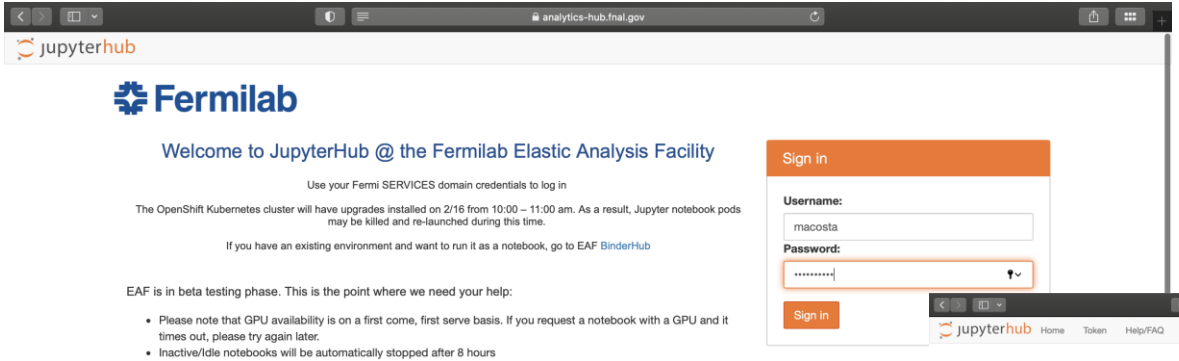
Fermi National Accelerator Laboratory

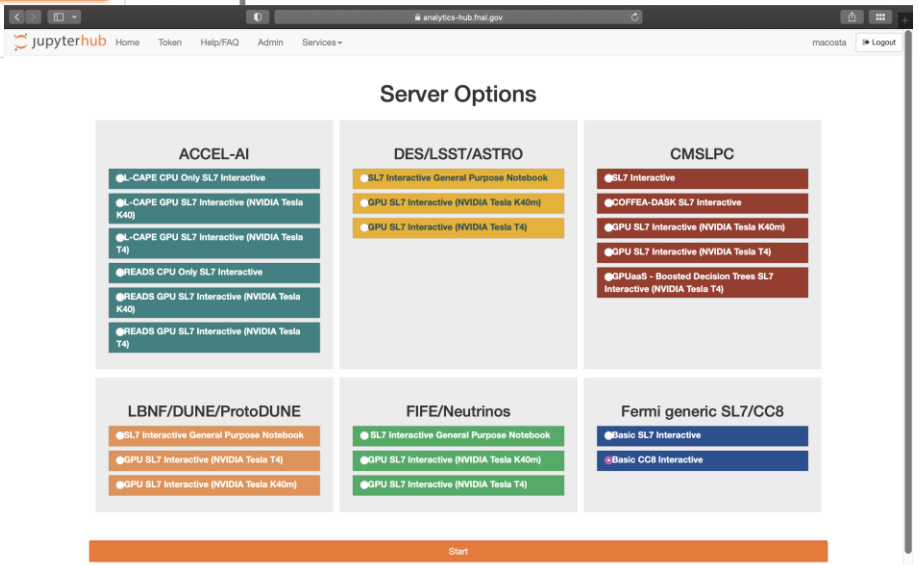
FNAL Ecosystem



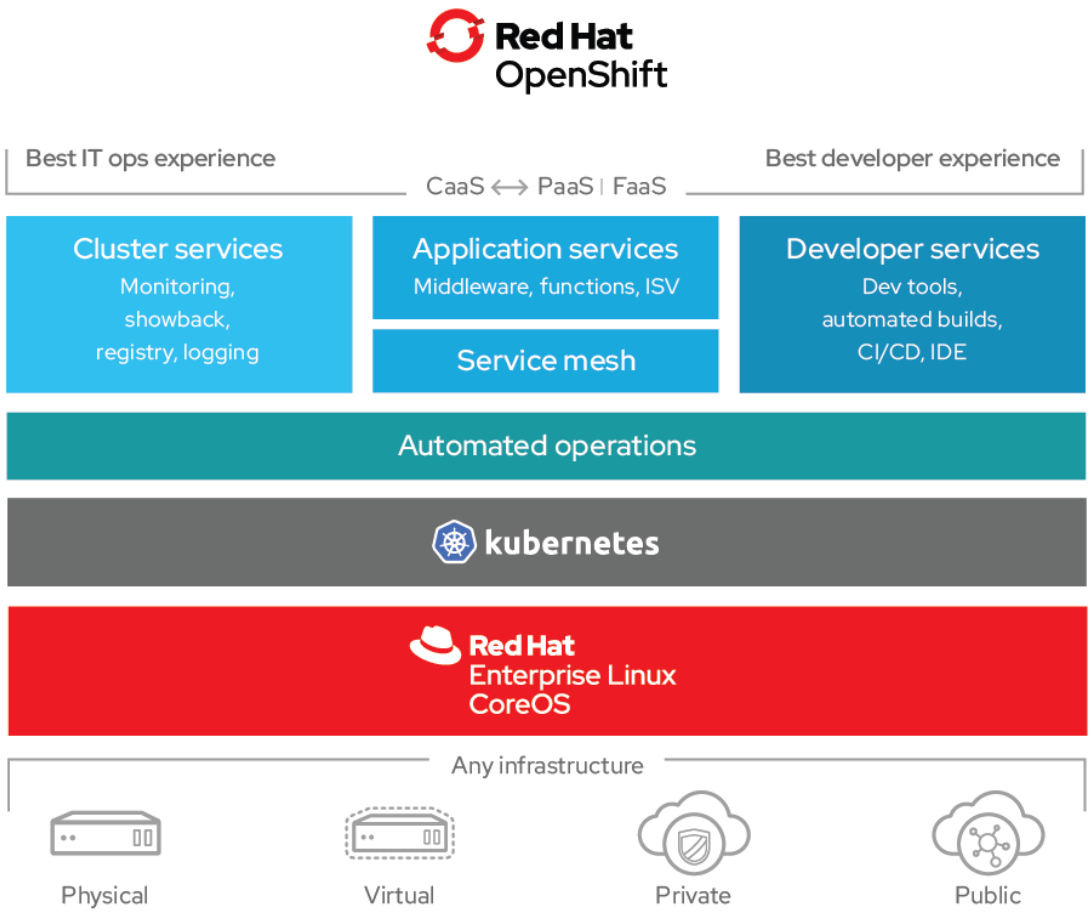
Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab

Canonical Jupyterhub Screenshots





Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab



okd
The Community Distribution of Kubernetes that powers Red Hat OpenShift

OKD: open-source OpenShift

Provides:

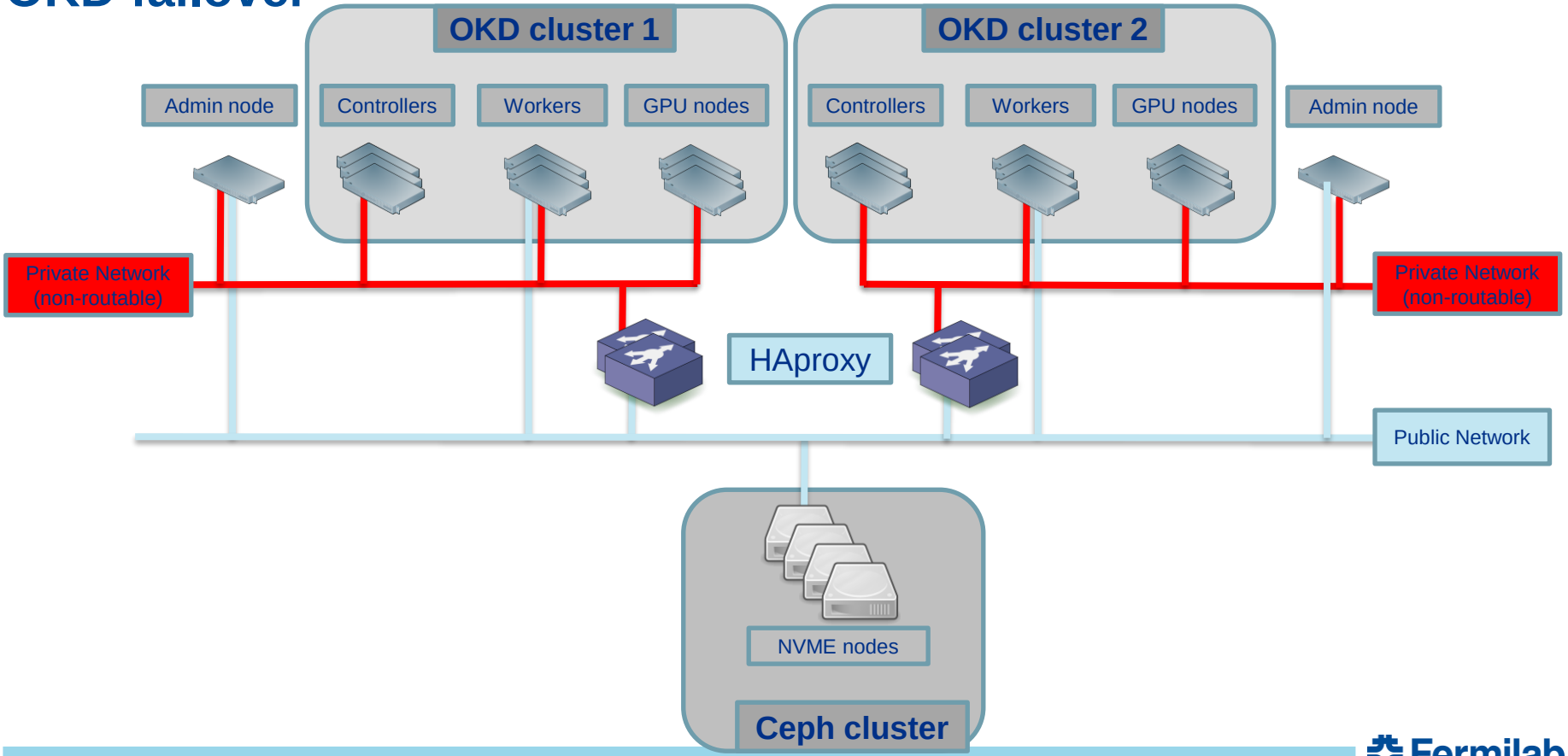
- Namespace isolation
- Good default security policies (e.g. containers are unprivileged and non-root)
- Monitoring/logging primitives

Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab

OKD deployment overview

- OKD 4.x on thin VM on bare metal
- Installation fully scripted
  - Configuration files generated by Puppet
  - Additional configs (chrony, yum, RPM's Ceph keys) done through OKD `MachineConfig` objects
  - One single command to install OKD and other components (OPA, ceph-CSI)
- Using isolated network, users can access through HAproxy using LDAP authentication
- Redundant cluster strategy
  - Second cluster used as a cold spare and for testing upgrade or troubleshooting
  - Permanent storage on external Ceph cluster so it can be accessed from both clusters

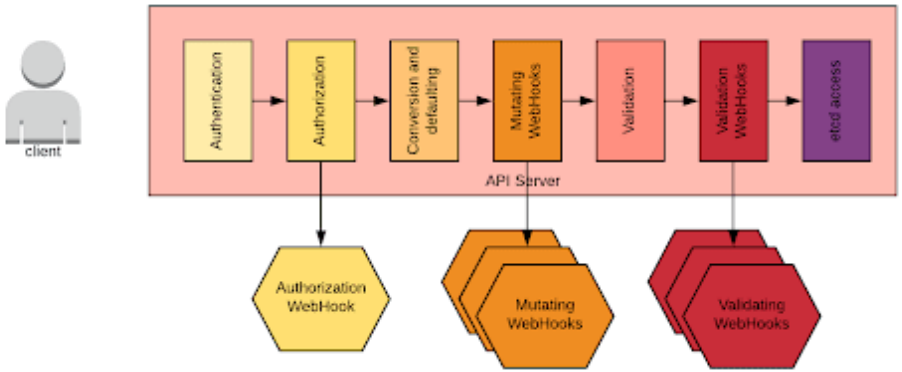
OKD failover



Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab

Open Policy Agent (OPA)

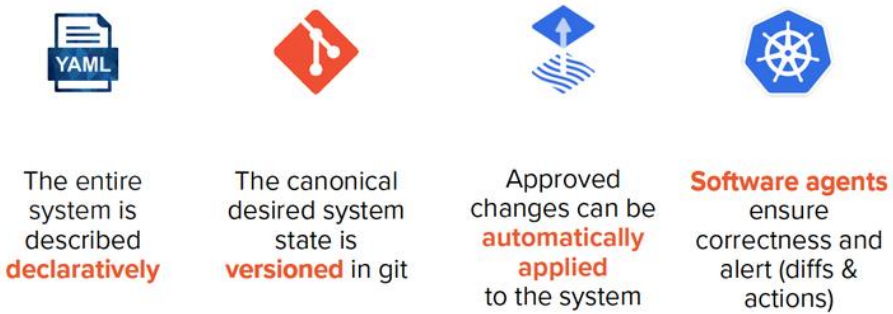
- OPA on OKD
  - OKD allows creation of webhooks in the validation stage
  - OPA can inject homemade policies through the webhook to allow/reject objects
- OPA Policies:
  - Rego language
  - OKD agnostic, policies just parse yaml files
  - Policies can be updated live from Puppet Hiera
- Currently used for:
  - NFS mounts
  - External IPs



Brookhaven National Laboratory SLAC NATIONAL ACCELERATOR LABORATORY Fermilab

Operating a multi-VO analysis facility

- EAF and its components have evolved over a couple of years of prototyping
- Applications, integration strategies, infrastructure and storage provisioning has been designed following **Cloud Native** recommendations and best practices.
- Among other reasons like **operational efficiency** and **faster development cycles**, cloud native allows us to take advantage of our on-prem OKD resources and scaling capabilities – hence the 'Elastic' on the name



Principles of GitOps: https://gitops-community.github.io/kit/

Operating a multi-VO analysis facility

GitOps: Why?

- Takes advantage of a declarative system to manage the configuration and operations of every element of the platform, from the infrastructure through to the applications.
- Provides observability and control - ensuring that the platform is reliable and operable.

GitOps: How?

- All our production applications are deployed and managed via Helm charts
- Application/IaaS code is version controlled, securely hosted in lab's internal GitLab instance
- Published OKD templates to deploy specific applications for hosting CVMFS as well as containerized HTCondor worker nodes for Kubernetes
- CI/CD pipelines implemented for build, test, audit and push to image repository
- Orchestration and deployment via Argo CD



Operating a multi-VO analysis facility – Lessons learned

- UNIX User management must be consistent lab wide: Not trivial to achieve in JupyterHub as LDAP authenticator does not support it.
- Built multiple custom hooks in KubeSpawner which handles UNIX user creation via FERRY (lab source-of-truth for user account information) for Pod customization, CVMFS mounts, access control, NFS areas, CephFS persistent volumes and initialization containers.
- Layered container builds allow for code recycling while providing flexibility for diverse custom environments.
- BinderHub provides an extra layer of flexibility for notebook customization by users – Requirements are wide as we are not only supporting VOs but specific research groups within experiments.
- Some of the open-source applications deployed are specifically designed for vanilla installations of Kubernetes. Notably, there is a general absence of Service Accounts, poorly defined and delimited security contexts and wide role definition.
- A considerable R&D effort has been destined to patching applications, submitting pull-requests to code owners and making recommendations in open, collaborative workspaces within the open-source community.

Summary and Common Themes

- HEP Analysis Facilities at US National Labs are deployed and in use at Brookhaven, Fermilab, SLAC
- Use of declarative installation and infrastructure tools gives us agility to meet evolving analysis needs
- Elasticity provided to already-existing batch software systems