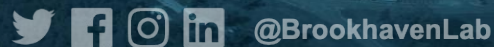




# Analysis Infrastructure for Science at BNL

D. Benjamin, O. Rind

EOSC-Future ESCAPE Science Projects  
Progress Meeting, July 22, 2022



# BNL Scientific Data and Computing Center (SDCC)

Located at Brookhaven National Laboratory (BNL) on Long Island, New York

- Tier-0 computing center for the RHIC experiments.
  - sPHENIX - scheduled to start taking data in 2023
  - Site for the future Electron-ion Collider (EIC)
- US Tier-1 Computing facility for the ATLAS experiment at the LHC
  - Also, one of the US ATLAS shared analysis facilities
- US data center for the Belle II experiment
- Computing facility for the NSLS-II



Shared multi-program facility (Nuclear Physics, High Energy Physics and Photon Science) serving the scientific computing needs of more than 2,500 users from > 20 projects

# What is an Analysis Facility?

There is no common answer in ATLAS (or WLCG for that matter).

- CERN LXplus should be considered one.
  - CERN has several tools for analysis (e.g. SWAN, Kubernetes, etc)
- DESY has the NAF
- US has at least three (BNL, SLAC, University of Chicago)
- Google Cloud project can be considered one
- UK has its own variety

Current ATLAS approach -  
("let many flowers bloom")



# Overview of US 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



## 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

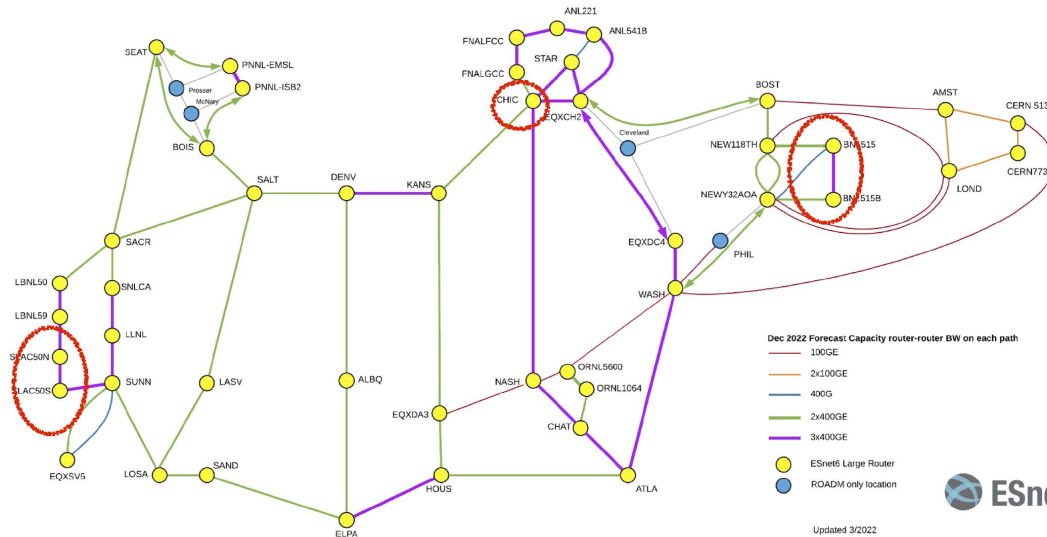
~50 users

# Network connectivity

Strong connectivity among the US Analysis Facilities

(BNL: US Tier 1; UChicago: MWT2 is largest T2 in US, 1.5x others; SLAC: US Nat. Lab)

## December 2022 Target Backbone Capacity

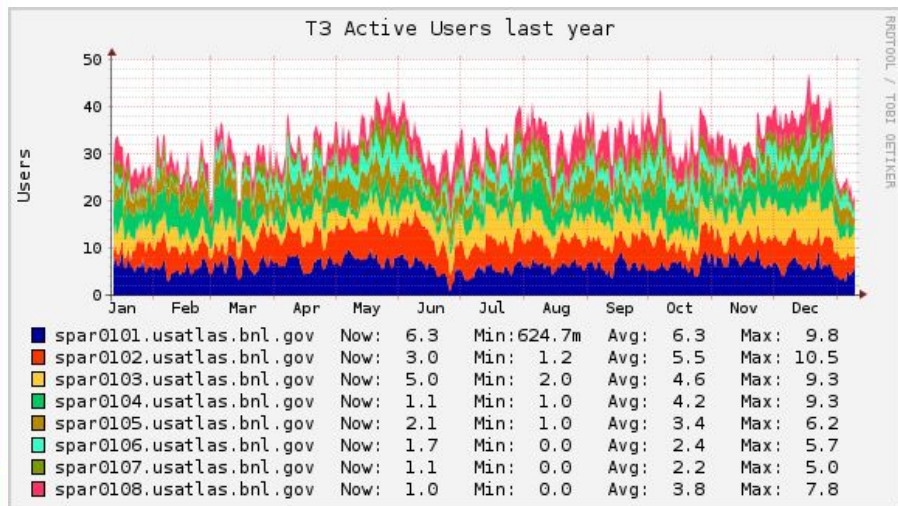


US AF attached at points in the red circles

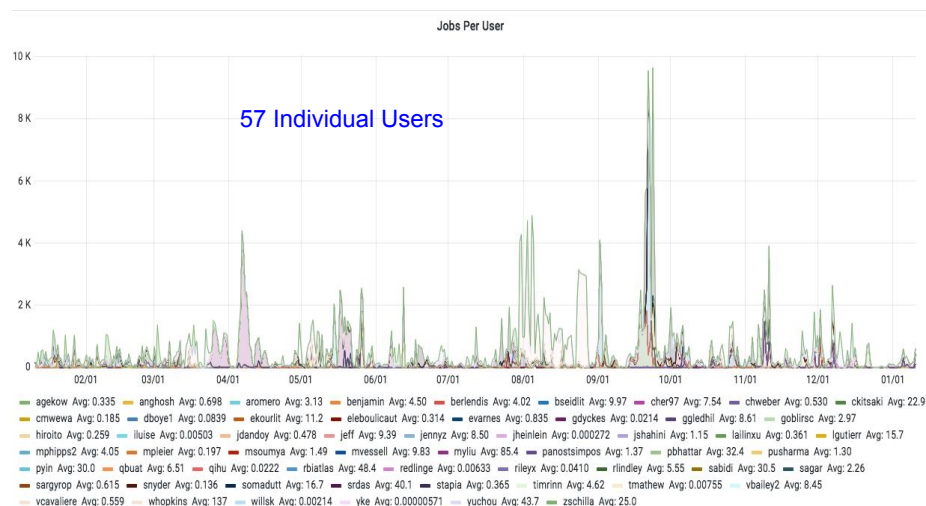
Green – 2x400Gbe  
Purple – 3x400 Gbe

# “Traditional” BNL Tier-3 Usage

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



Number of Running HTCondor Jobs Per Tier-3 User



# Evolving the Facilities

- New data challenges...New data formats...New architectures...New techniques...New tools....
- Tier-3 model – bulk analysis on grid, interactive ssh login, analyze final ntuples – needed to evolve as well
- First foray into this ~4 years ago at BNL with GPU-based Institutional Cluster and development of Jupyterhub portal
  - Philosophy of leveraging existing resources, batch interfaces (HTCondor for HTC, Slurm for HPC)
- Activity ramped up steadily with support from US ATLAS and IRIS-HEP...ML Platform at UC, Jupyterhub at SLAC, Analysis Facility at UC, and other updates

# SDCC Jupyterhub Portal



## SDCC JupyterHub

The SDCC offers multiple JupyterHub instance and back-end combinations for different users and accounts. Choose the appropriate option from the instances displayed below.

[More information](#)

[Questions and support](#)



[Run Jupyter](#)

[Manage Sessions](#)

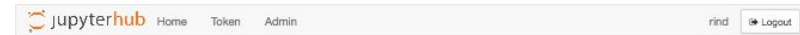
Click "Run" above to navigate to jupyter and choose to launch a notebook via HTC, HPC or directly on our condor pool.  
Click on the "Manage Sessions" button to be able to start and stop an already running session or manage named sessions

Jupyterhub interface to multiple resources backends:

- Dedicated set of Jupyterhub nodes
  - Dask scheduling onto HTCondor pools
- Notebook spawning onto HTCondor shared pools
- Notebook spawning onto HTC cluster via Slurm
  - Access to dedicated GPU resources

Custom interface with (MFA) access control based on SDCC login account group

Common software environment shared with SLAC AF on CVMFS



## SDCC Jupyter Launcher

[HTC / Standard](#) [HTCondor Pool](#) [IC / HPC Systems](#)

Run a notebook on a standard interactive HTCondor submit-node

Select JupyterLab Environment

- Default
- Default HPC
- USATLAS

Singularity Container

- None
- Custom

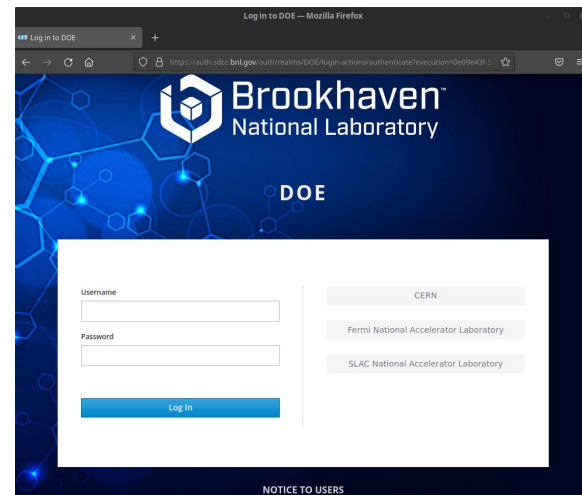
[Start](#)



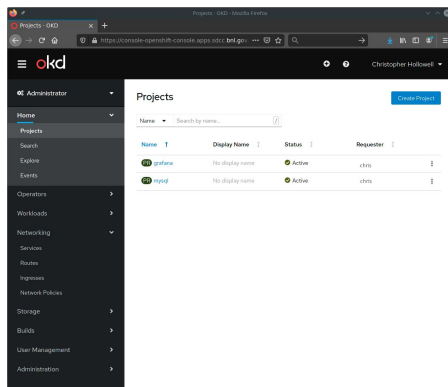
# Federated JupyterHub

- **Technical implementation of a federated Jupyter instance for ATLAS complete**
  - Allows users to login via BNL/SDCC, CERN, FNAL or SLAC credentials
    - Users without existing SDCC accounts are given “lightweight accounts” where their only access at the facility is Jupyter (i.e. cannot login to SSH gateways, etc.)
  - Implementation requires MFA to satisfy DOE requirements
  - Users must fill out a form and be approved for initial access after their ATLAS affiliation is confirmed
  - Account activation turn-around time goal of 1-2 business days
- **EPPN OIDC token attribute mapping used in LDAP for authorization and to tie federated ID to a local UNIX account**
  - Reverse proxies doing the authentication
  - Modifications to the Jupyter [jhub\\_remote\\_user\\_plugin](#) to support the mapping
  - Users have access to all SDCC ATLAS network filesystems
- **Being tested by a number of users**

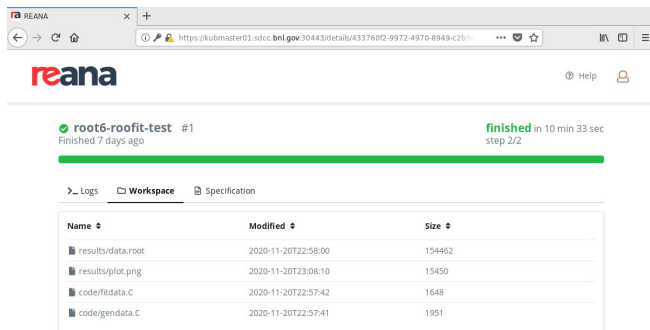
Federated JupyterHub Login Screen



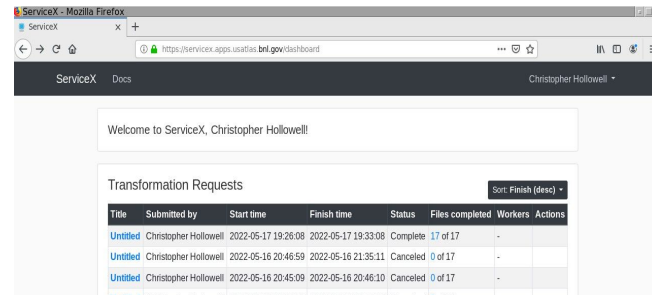
# New Tools and Services Under Development



- Testing ServiceX and REANA instances on a staff kubernetes cluster, and on our production ATLAS [OKD](#) cluster
- [ServiceX](#) - Columnar data delivery service
- [REANA](#) - framework for reusable analysis



REANA Web interface



ServiceX Web interface

# Local Storage Resources

- Lustre
  - 3 data servers (MDT) - 3 PB useable - 2x 50 Gb network,
    - BNLBox storage and Globus Connect Servers (GCS v5) storage
      - Globus - can be used by users to move data between GCS endpoints (typically US HPC centers, put other institutions also)
- XCache
  - Single host, 57 TB NVME cache, dual-homed 2 x 40Gb network
  - Part of a network of XCache servers for VP queues that exist at all US sites, including two of the US AF's.
- EOS
  - Kerberos authenticated fuse mounts of CERN EOS ATLAS/USER volumes available on interactive hosts

# The Challenge of Data Access at AFs

- AF data caches requirements are different from regular storage:
  - Depends on the reliability input data due to XCache requirements of origin data source
  - Need to handle large and very bursty IO
  - Handle many small files
  - Responsive low latency
  - Won't need much space due to a significant data reuse. SSDs are the best option. No replication.
  - AF storage caches are more sensitive to changes in data popularity
- Group data repositories vs. user directories (notebooks, personal ntuples, etc.)
  - Can we provide easy access and common work environments across a wide distribution of AFs?
- AAI integration - Role of storage tokens?

Goal: information rich content to users.

# User Output

How do we handle the user outputs?

- Sync and share – ala world-wide CERN Box or similar technology for small files
- Larger output files – Rucio upload?

Structured output – Rucio

- Rucio Upload - to what RSE? **Users need some data space** – Is EOS enough? Should this space be distributed across data centers?

“Unstructured” output (ie - directories with user defined file names)

- What is the interoperability amongst these sync-share systems (ie how to avoid vendor lock in)?
- See work of Science Mesh (<https://sciencemesh.io/>)

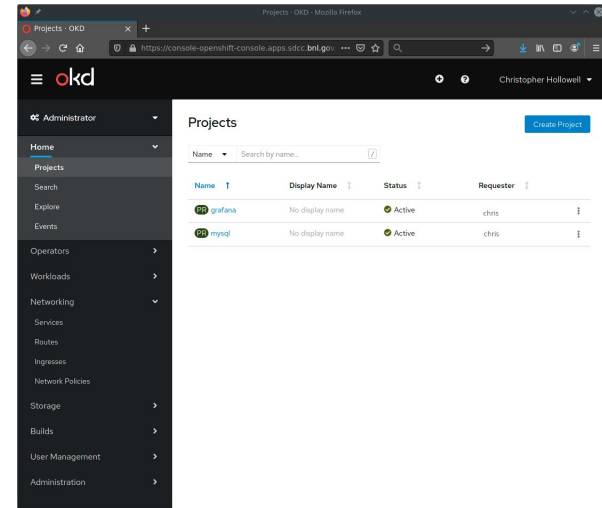
# Conclusion

- Moving toward a new era of large scale interactive LHC analysis support
- Actively building up infrastructure and developing computational tools *while listening to the users*
  - User feedback and input is vital to the success of the project – need to provide what users will actually use
- Distributed storage access is still a limitation...thus our interest in applying, and perhaps building upon, ESCAPE DLAAS concepts

# Backup

# OKD @ SDCC

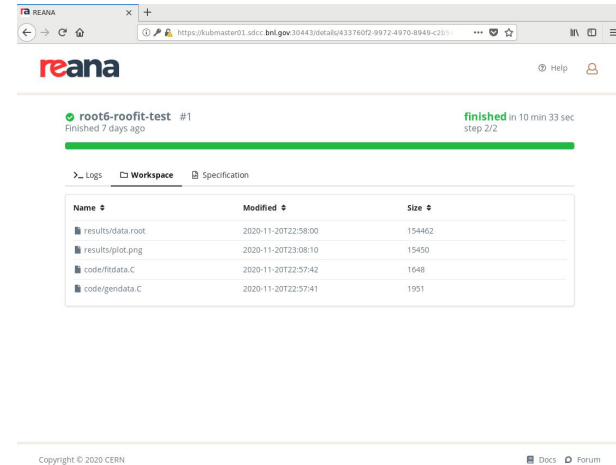
- **OKD provides a platform for container orchestration, similar to Kubernetes (k8s)**
  - Community-supported release of Openshift
  - Allows for simplified deployment of services via helm charts and Openshift templates
  - Contains numerous security enhancements out of the box vs k8s
    - Users are never root by default
- **Two OKD 4.7 clusters online at SDCC**
  - One for sPHENIX experiment at BNL (Nuclear Physics)
  - One for ATLAS experiment (CERN) (High Energy Physics)
  - Kept separate intentionally for isolation.
  - Consists of 26 TB NetApp nvme storage and 4 compute nodes ( 2x 20 HT cores) each cluster.
  - Console authentication tied to our keycloak IDP
  - Currently only accessible from inside BNL
  - Users run `oc/kubect` commands to manage projects and service deployments from our interactive compute nodes
  - Token obtained from web interface





# REANA Test Cluster

- **Deployed REANA testbed at SDCC**
  - A platform for reproducible scientific data analysis
    - <https://www.reanahub.io/>
    - Primarily being developed by CERN
      - Still in a pre-release phase
  - Being used by a number of test users
    - Web interface currently accessible via SSH tunnel/SOCKS proxy
    - Working on tying accounts/auth to our LDAP/K5/IDP
    - Can interface and submit container jobs to SLURM on the IC



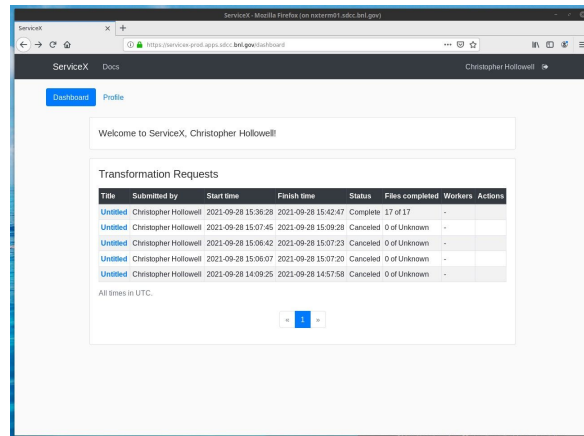
The screenshot shows the REANA web interface in a browser window. The address bar displays the URL: `https://kubemaster01.sdcc.bnl.gov:30443/portal/4337602-9972-4970-8949-c251...`. The page header includes the REANA logo and a 'Help' link. The main content area shows a job titled 'root6-rootfit-test #1' with a status of 'finished' in 10 min 33 sec, completed 7 days ago. Below this, there are tabs for 'Logs', 'Workspace', and 'Specification'. The 'Workspace' tab is active, displaying a table of files:

Name	Modified	Size
results/data.root	2020-11-20T22:58:00	154462
results/plot.png	2020-11-20T23:08:10	15450
code/fitdata.C	2020-11-20T22:57:42	1648
code/gendata.C	2020-11-20T22:57:41	1951

At the bottom of the page, there is a copyright notice: 'Copyright © 2020 CERN' and links for 'Docs' and 'Forum'.

# ServiceX Test Deployment

- **Testing ServiceX deployment at SDCC**
  - Columnar data delivery and pre-processing service
  - Deployed in our test OKD cluster
    - 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



```
#!/usr/bin/env python3
from func_adl_servicex import ServiceXSourceXAOD

dataset_name = "mc15_13TeV:mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee\
.merge.DAOD_STDM3.e3601_s2576_s2132_r6630_r6264_p2363_tid05630052_00"
src = ServiceXSourceXAOD(dataset_name)
df = src \
    .SelectMany('lambda e: e.Jets("AntiKt4EMTopoJets")') \
    .Select('lambda j: j.pt()/1000.0') \
    .AsPandasDF('JetPt') \
    .value()
print(df)
```

```
JetPt
0      36.319766
1      34.331914
2      16.590844
3      11.389335
4       9.441805
...
857133  6.211655
857134  47.653145
857135  32.738951
857136  6.260789
857137  5.394783
```

```
[11355980 rows x 1 columns]
```

# Listening to the users

Need to provide the tools and services that make analyzers more efficient and productive.

- No “Field of Dreams” – “If you build it, he will come”
- User feedback and input vital for successful project

