Building Scalable Analysis Infrastructure for ATLAS

Lincoln Bryant
University of Chicago

CHEP 2024
October 19 - 24, 2024

Robert W. Gardner[1], Farnaz Golnaraghi[1], Eric Christian Lancon[2], Fengping Hu[1], David Jordan[1], Judith Lorraine Stephen[1], Ryan Paul Taylor[3], Ilija Vukotic[1]

on behalf of the ATLAS Computing Activity

[1]University of Chicago [2]Brookhaven National Laboratory [3]University of Victoria

Analysis Facilities

O. Rind, CHEP '23 (talk)

Broad Snowmass Report definition:

> *"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."*

In the HL-LHC era, scale becomes an issue, impacting the line between interactive/batch analysis and access to datasets. Increasingly complex workflows and heterogeneous architectures will also play a role.

> *"What elevates a resource to the level of an analysis facility is official support as a shared resource within an organization of people with shared interests"*

Overview of US Analysis Facilities for ATLAS

O. Rind, CHEP '23 (talk)

US ATLAS has three shared analysis facilities providing software & computing

- Resources that fill gaps between grid jobs and interactive analysis on local computers
- All leveraging substantial local batch and storage resources
- Interactive ssh login, with access to Rucio and PanDA resources and notebook servers
- GPU resources

Brookhaven National Laboratory

BNL Facility

~2000 cores, part of a larger shared pool, opportunistic access up to 40k cores

User quota: 500GB GPFS plus 10TB Lustre

~200 users

SLAC National Lab Shared Scientific Data Facility (SDF)

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

ANALYSIS FACILITY at UCHICAGO

UChicago Facility

~3000 cores, co–located with MWT2, opportunistic access up to 16k cores

User quota: 100GB home, 5–10TB for data

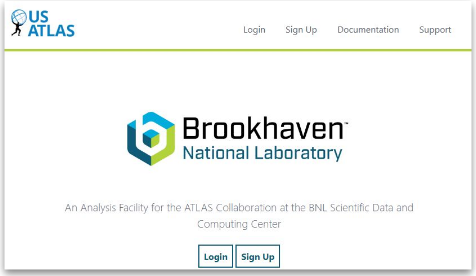
~400 users

Towards Federated Analysis Facilities

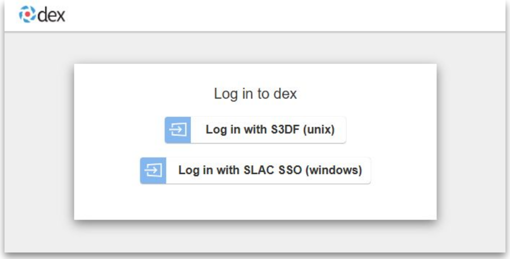
- We are working on a federated analysis platform prototype that unifies resources from multiple facilities
  - We have broken the problem down to five key areas: ***Policy, Identity, Network, Data & Compute***
- Goal is to provide the following benefits to ATLAS physicists:
  - **Seamless User Experience** By unifying access across multiple sites, users can log in and work as if they were on a single, centralized system, reducing friction and improving productivity.
  - **Harmonized Policies and Security** Aligning IT policies and trust across national labs and universities ensures compliance while enabling broader and more secure access to resources.
  - **Improved Data Access and Management** Leveraging advanced data delivery tools and caching systems like ServiceX, XCache, Rucio and EOS will accelerate data access and improve performance during iterative analysis sessions, which is crucial for effective Run-3 analysis.
  - **Enhanced Resource Utilization** The ability to marshal additional CPU resources beyond pledged Tier-1 and Tier-2 center capacities allows for better handling of low-latency, interactive analysis workloads, optimizing the use of computing power across facilities.
  - **Scalability and Flexibility** The adoption of modern tools from the cloud native ecosystem will provide a more scalable and flexible infrastructure, capable of adapting to the diverse needs of the ATLAS physicists for Run-3 and in future HL-LHC.

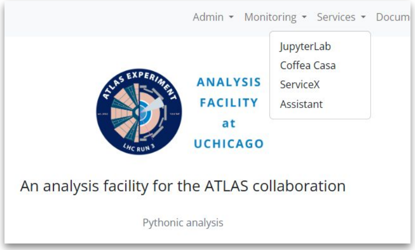
Considerations for Federation

- The US ATLAS AFs are quite diverse in functionality, and each offers unique resources, e.g.:
  - BNL has a large shared pool for opportunistic overflow and nearby Tier 1 storage
  - UChicago has Coffea Casa, ServiceX instances and nearby Midwest Tier 2 LOCALGROUPDISK
  - SLAC has a dedicated GPU cluster with a large number of NVIDIA A100 devices
- Acquiring and maintaining access to these three resources can be a challenge for users
- As the AFs grow in popularity, it's important to spread users among them to use resources in the most efficient manner possible and keep queue times reasonably low in a way that is transparent to users



An Analysis Facility for the ATLAS Collaboration at the BNL Scientific Data and Computing Center





An analysis facility for the ATLAS collaboration

Pythonic analysis

Building a *distributed* Analysis Facility

- Taking lessons learned from building the UChicago Analysis Facility, a Kubernetes-based site, we have set out to develop tools and approaches for federating resources
- The goal:
  - Better resource utilization across the US computing fabric
  - A more streamlined user experience
- Efforts to date have been focused on building services that run on a stretched Kubernetes platform under the umbrella of Facility R&D in the US
  - We use Kubernetes as a vehicle for demonstrating the general ideas without necessarily prescribing this technology everywhere

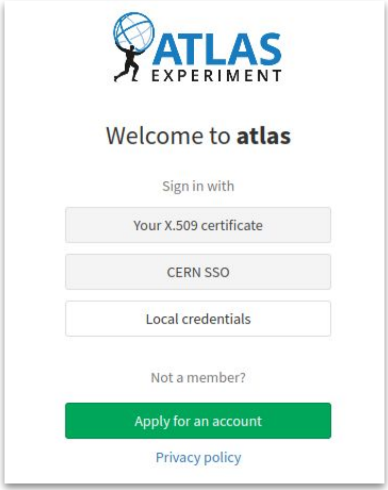
The 5 Key Areas for a Unified Experience

- **Identity**
    - Leveraging existing, experiment-specific identity for authentication and authorization
- **Network**
    - Utilizing modern network overlay technologies to provide seamless connectivity between sites
- **Data**
    - Broad deployment of data caching infrastructure
- **Compute**
    - Embracing notebook-centric technologies and Pythonic frameworks, leveraging advances in Identity, Network and Data
- **Policy**
    - Developing policy framework(s) that provide an easy on-ramp for experiment end-users to use resources at all AFs

(Simplifying) Identity (& Authorization)

- Over the last several years, OAuth2, OIDC, and JSON Web Tokens have become ubiquitous, cornerstone technologies to authenticate and communicate metadata about users
- Indigo-IAM (INFN) is a great tool in our field which lets us create experiment-specific OAuth authentication clients
  - e.g., https://atlas-auth.web.cern.ch/login (as well as cms-auth, lhcb-auth, etc)
- We have combined this with Keycloak to:
  - Allow any ATLAS user to login to our Federated Facility
  - Be implicitly authorized to use the facility vis-à-vis membership in ATLAS
  - Create and store extra metadata (claims) about users as they pass through the system
    - including POSIX identity information (uid, gid, username etc)

KEYCLOAK



ATLAS EXPERIMENT

Welcome to **atlas**

Sign in with

Your X.509 certificate

CERN SSO

Local credentials

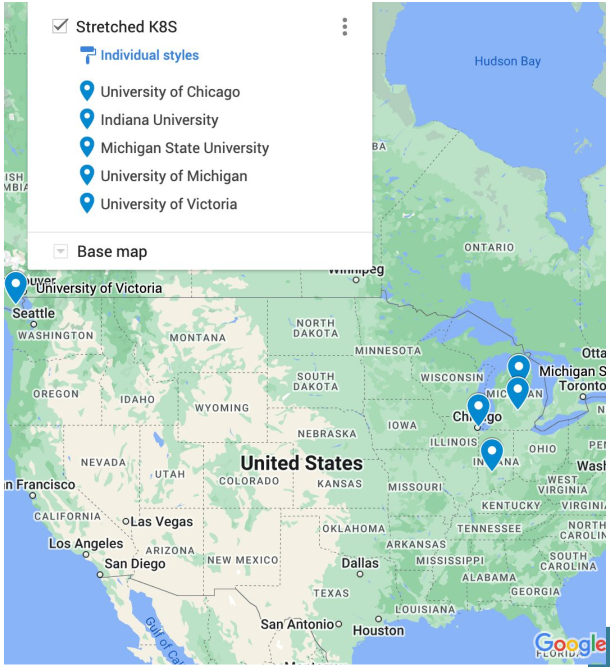
Not a member?

Apply for an account

Privacy policy

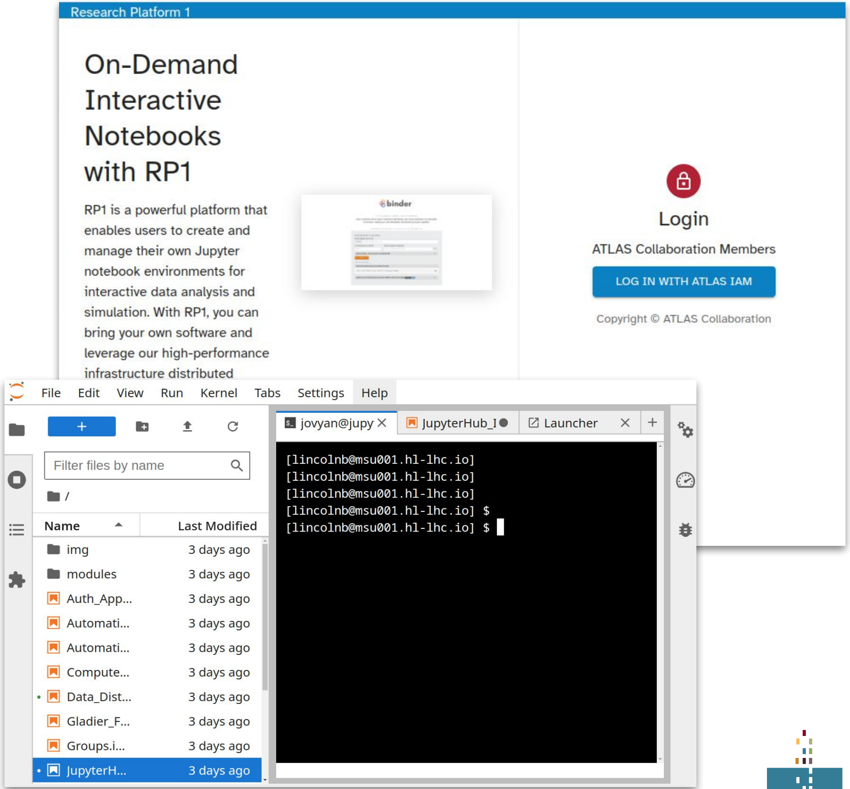
(A Secure) Network

- Modern overlay network technologies provide a way to create LAN-like functionality across a wide-area network
- We have leveraged WireGuard (Encrypted VPN mesh) technology (via an open source tool, NetBird) to build a stretched Kubernetes environment across 5 North American ATLAS institutions:
  - University of Chicago
  - Indiana University
  - Michigan State University
  - University of Michigan
  - University of Victoria (ATLAS Canada)
- For applications running across this mesh, the network appears to be a LAN environment



(Accessing, transparently) Compute

- Many of the analysis frameworks being built today, such as Coffea & Dask, are largely focused on Jupyter Notebook technology and surrounding tooling in the Pythonic HEP ecosystem
    - including Awkard Array, Uproot, ServiceX, etc
- Leveraging the stretched Kubernetes platform and ATLAS IAM integration, we have created a Jupyter-based platform that any ATLAS user can log in to and automatically get a notebook



Research Platform 1

On-Demand Interactive Notebooks with RP1

RP1 is a powerful platform that enables users to create and manage their own Jupyter notebook environments for interactive data analysis and simulation. With RP1, you can bring your own software and leverage our high-performance infrastructure distributed

Login

ATLAS Collaboration Members

LOG IN WITH ATLAS IAM

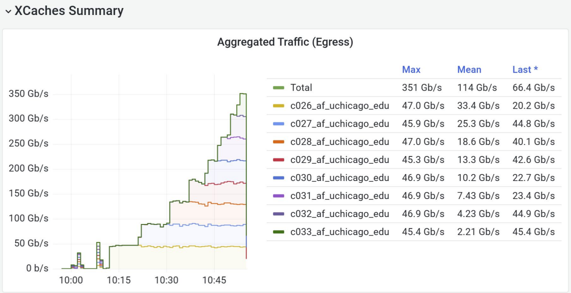
Copyright © ATLAS Collaboration

Name	Last Modified
img	3 days ago
modules	3 days ago
Auth_App...	3 days ago
Automati...	3 days ago
Automati...	3 days ago
Compute...	3 days ago
Data_Dist...	3 days ago
Gladier_F...	3 days ago
Groups.i...	3 days ago
JupyterH...	3 days ago

```
[lincolnb@msu001.hl-lhc.io]
[lincolnb@msu001.hl-lhc.io]
[lincolnb@msu001.hl-lhc.io]
[lincolnb@msu001.hl-lhc.io] $
[lincolnb@msu001.hl-lhc.io] $
```

Data

- When building a distributed environment with a unified interface, we must be careful with file access methods:
  - Users like POSIX file access patterns, but they become challenging especially with small file I/Os over wide-area links
  - The latency involved is larger by an order of magnitude or more (<1ms on a LAN, 10-100ms typical on WAN)
- XCache, a caching layer built on top of XRootD, provides considerable value to sites
  - Simply put the local XCache address in front of xrdcp or http access
    - i.e., "xrdcp root://<cache>/root://<file URL>"
  - Placed near compute, we are using 100s TB of XCache today to accelerate access to popular data
- In both LAN and distributed environments, effective caching can provide a lot of value to users



XCaches Summary

Aggregated Traffic (Egress)

	Max	Mean	Last *
🟩 Total	351 Gb/s	114 Gb/s	66.4 Gb/s
🟨 c026_af_uchicago_edu	47.0 Gb/s	33.4 Gb/s	20.2 Gb/s
🟦 c027_af_uchicago_edu	45.9 Gb/s	25.3 Gb/s	44.8 Gb/s
🟧 c028_af_uchicago_edu	47.0 Gb/s	18.6 Gb/s	40.1 Gb/s
🟥 c029_af_uchicago_edu	45.3 Gb/s	13.3 Gb/s	42.6 Gb/s
🟦 c030_af_uchicago_edu	46.9 Gb/s	10.2 Gb/s	22.7 Gb/s
🟪 c031_af_uchicago_edu	46.9 Gb/s	7.43 Gb/s	23.4 Gb/s
🟪 c032_af_uchicago_edu	46.9 Gb/s	4.23 Gb/s	44.9 Gb/s
🟩 c033_af_uchicago_edu	45.4 Gb/s	2.21 Gb/s	45.4 Gb/s

Policy

- Reasonable, targeted carve-outs for specific federated workloads, or general policy frameworks, have to be developed for unified Analysis Facilities *just as it was done for Grid technology 20 years ago*
- These challenges are not unique to our colleagues at national laboratories, they simply see them first in many instances
  - Much of what we may view as very restrictive policy is considered best practice in broader security contexts
- We must start with what is possible today and take very small steps toward federation
  - e.g., at Brookhaven, users can request "Lightweight ATLAS Accounts" that allow users to create notebooks with CERN identity
- In this particular instance, we are working to understand the smallest incremental step between lightweight accounts and connecting to a distributed resource pool

Summary & Conclusions

- In US ATLAS we are prototyping Federated Analysis Platforms
  - To aggregate our analysis resources and simplify the user experience
- Innovating in Key Technical Areas
  - To create a scalable, efficient infrastructure that supports high-throughput analysis workflows, emphasizing user-friendly interfaces and flexible policy models.
- Building Distributed Cloud-Native Systems
  - R&D is leveraging Kubernetes to unify diverse resources into a stretched, distributed analysis platform, exploring the use of secure, VPN-based network overlays to simulate LAN-like conditions across geographically dispersed facilities.

Thank you!

SOTERIA

FAB

FABRIC Across Borders

iris hep

National Science Foundation

This work was supported in part by these NSF grants:

- OAC-2115148 CICI:UCSS:Securing an Open and Trustworthy Ecosystem for Research Infrastructure and Applications (SOTERIA)
- OAC-2029176 Collaborative Research: IRNC: Testbed: FAB: FABRIC Across Borders
- OAC-1836650 Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)
- PHY-2120747, U.S. ATLAS Operations: Discovery and Measurement at the Energy Frontier