Analysis Facilities

Lincoln Bryant
Fengping Hu
University of Chicago

HTC 25
June 5, 2025

Caveat Emptor

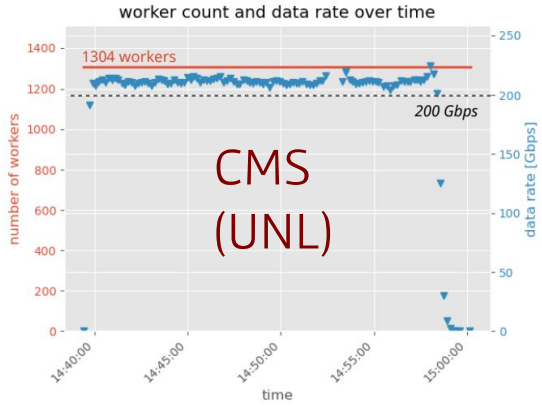
- This is not an exhaustive survey of every analysis facility, or every possible configuration
- These slides are intended to pose some questions to seed discussion for the remainder of the session

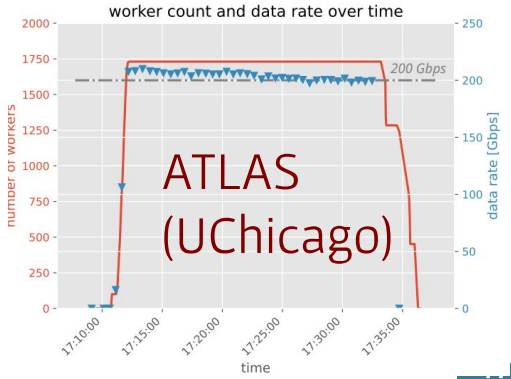
Historical Analysis Facilities

- In the broadest sense, our community has been running Analysis Facilities for a very long time, though we called them different things (Tier 3s)
- These are typified by:
  - Interactive login
  - Some local batch system
  - Some local storage
- HL-LHC efforts are driving the evolution of our facilities, both in terms of software and hardware
  - E.g. last year's 200Gbps challenge within IRIS-HEP

200Gbps Challenge

- In a slide: Analyze 25% of a ~180TB dataset in 30min, representative of what a future HL-LHC analysis may look like
- Participation from Analysis Facilities
  - ATLAS: University of Chicago
  - CMS: University of Nebraska-Lincoln
- XCache, Dask common tools for the challenge
- Various other software and approaches including ServiceX, Taskvine, etc





See: Alex Held's CHEP talk

Hardware evolution

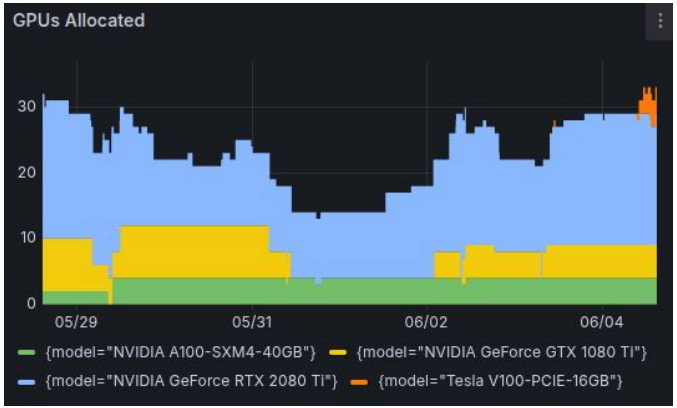
- Looking at the challenge from a hardware-level:
  - Disk:
    - At 250MB/s per disk (ideal, fully sequential workload), need at least 100 spinning disks-worth of performance for a ¼ scale analysis
      - With realistic usage, this easily doubles or triples
  - Network
    - 200Gbps aggregate performance within a switch is generally attainable
    - However: this challenged highlighted weaknesses in switch-to-switch connectivity (at least at UChicago)
- If this sort of analysis becomes typical:
  - **Strong signal to shift to NVMe, and for facilities to consider 100Gbps networking everywhere**
    - Vendors now claim the NVMe $/TB will cross-over with HDD around 2030

Software, infrastructural changes

- For the various software frameworks we deployed for this challenge, Kubernetes (K8S) was an essential platform
  - Coffea Casa
  - Dask
  - ServiceX
  - XCache
- Developers were able to iterate quickly on *services* with K8S, but HTCondor is still king for throughput computing
- Beyond data challenges, **we expect K8S will continue to grow as an important platform for hosting AF services**

GPUs

- Users are asking for GPUs more and more often
- We don't have a lot of GPUs, but we know who does…
- **Can we navigate policy and come up with a technology solution for delivering GPUs to AF users?**
- In US ATLAS, we are exploring an overlay HTCondor pool to connect our GPU allocation at NERSC to AF nodes



Identity, Tokens

- WLCG users essentially have a common identity provider: CERN
- It would be awfully nice to rely on CERN accounts and IAMs for AF identity, fully embracing OAuth2 and OIDC
  - **Let's stop having our own user/password databases, for which we have to manage user lifecycles, rotate passwords, secure against attacks..**
- With X509 going away, we should also think about what end-user tooling will look like for tokens (will there be any?)
  - **We still have a LOT of user *and* sysadmin education to do about token infrastructures**
    - **How many people know the difference between refresh tokens and access tokens?**
    - **How many different OAuth2 flows do you know?**
    - **Do users and admins understand claims, WLCG tokens and SciTokens?**

Federating Analysis Facilities

- The dream: **Log in once, use any Analysis Facility**
- Things that are ~solved:
  - CPU
    - HTCondor glideins, flocking
  - Software delivery
    - Containers, CVMFS
- Still challenging
  - Storage
    - Users love POSIX. Perhaps to the point of being a key differentiator between AF and Grid?
    - What can we do about federating, syncing?
  - Identity
    - See previous slide
  - Networking
    - Negotiating site firewalls, maintaining good performance with disparate storage systems

Shared Analysis Facilities

- Another dimension worth exploring: **Can we construct joint AFs?**
  - HEP collaborations are broadly doing the same sort of work, using similar fundamental tools (HTCondor, XRootD, Tokens, ..)
  - Can we coordinate technology choices and successfully navigate the policy landscape to pull it off?
  - Other experiments could take advantage of this as well
    - For instance: Dune, Belle II

Summary

- AF workloads stress our facilities in different ways than grid workloads
  - Hardware refreshes should take this into consideration, including improving storage, network throughput
- New software frameworks are increasingly using Kubernetes as a platform
  - Are sites prepared to support this? Can we give developers a place to test against e.g. OpenShift?
- Users are knocking at our door for GPUs
  - Can we deliver existing HPC GPU allocations to them?
- There is no single front door for Analysis Facilities
  - What efforts can we undertake to make AFs more uniform, federated, synchronized?
- The work we're doing here is reusable by other collaborations
  - Is it worth exploring the construction of joint, multi-experiment AFs?