Data Delivery for Analysis at the HL-LHC

G. Watts (UW/Seattle)

Thanks to *Ilija Vukotic*, Frank K Wuerthwein, Shawn Mc Kee, Justas Balcas Lukas Heinrich, *Diego Davila, Alessandra Forti, Rob Gardner* and others

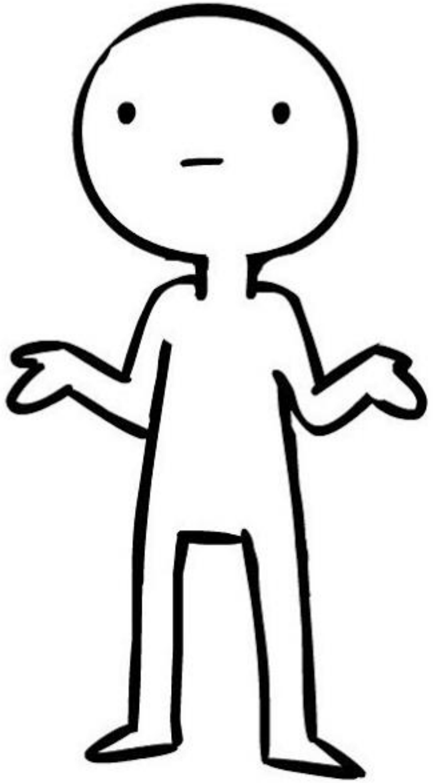
Data Delivery & *Storage* for Analysis at the HL-LHC *At Analysis Facilities*

G. Watts (UW/Seattle)

Outline

- Framing Slides
- Discussion Topics:
    - Local Storage – Large user files
    - Caching of External Resources
    - Local Storage – Smaller user files

What is an Analysis Facility?



And how should we think about data delivery and storage at the facility?

HSF AF Whitepaper - Google Docs

Traits of an Analysis Facility (Draft)

- Ability to perform **fast research iterations on large datasets**
- Ability to convert interactive to **batch-schedulable workloads**
- Ability to **scale outside of the facility** on occasion
- Ability to **efficiently train** machine learning models for HEP

How one might access the local data

- Ability to reproducibly instantiate desired software stack
- Ability to collaborate in a multi-organizational team on a single resource
- Ability to efficiently **access collaboration data** as well as make **intermediate data products** available to the team

What External Data One Might Access

- Ability to move Analyses to new Facilities
- Ability to express interdependent distributed computations at small and large scales
- Ability to run legacy analysis on infrastructures

Data Access at an AF

jupyter

BASH
THE BOURNE-AGAIN SHELL

Interactive:

- Very fast turn around expected
- Very spikey load profile
- Variable data sizes

BATCH PROCESSING

Delayed/Batch:

- Large Data sizes
- Load is more uniform

Event data, not conditions data, not source code…

Getting At **External** Data

Sources for External Data

- EOS (from CERN)
- GRID Datasets (RUCIO)
- Other AF’s

Does CERNBox belong in this list?

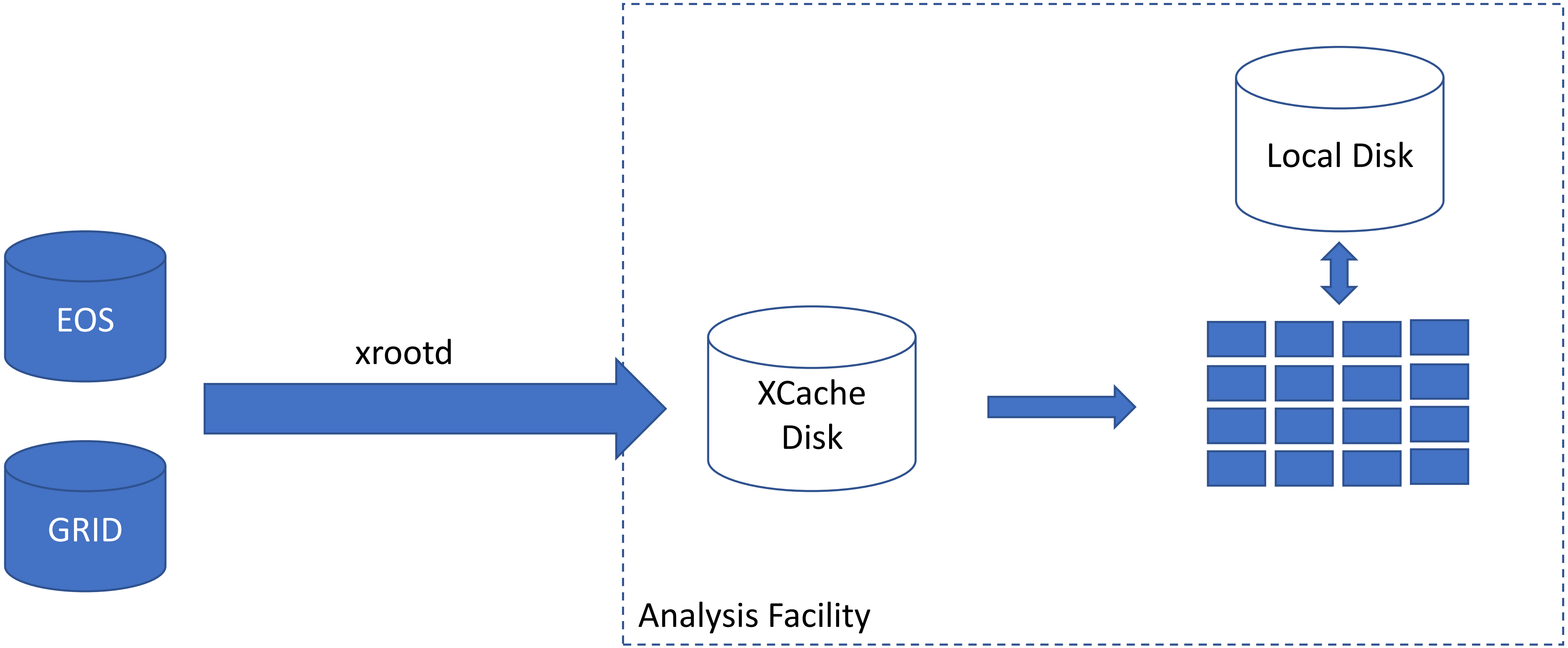
Is it safe to assume they will be available by:

- https ?
- xrootd

And authentication will integrate seamlessly with the AF’s security model

- User **doesn’t need to think** about this at all

Making External Data Appear Local



EOS

GRID

xrootd

XCache Disk

Local Disk

Analysis Facility

First Time Can Be Expensive…

“first time could be hours”

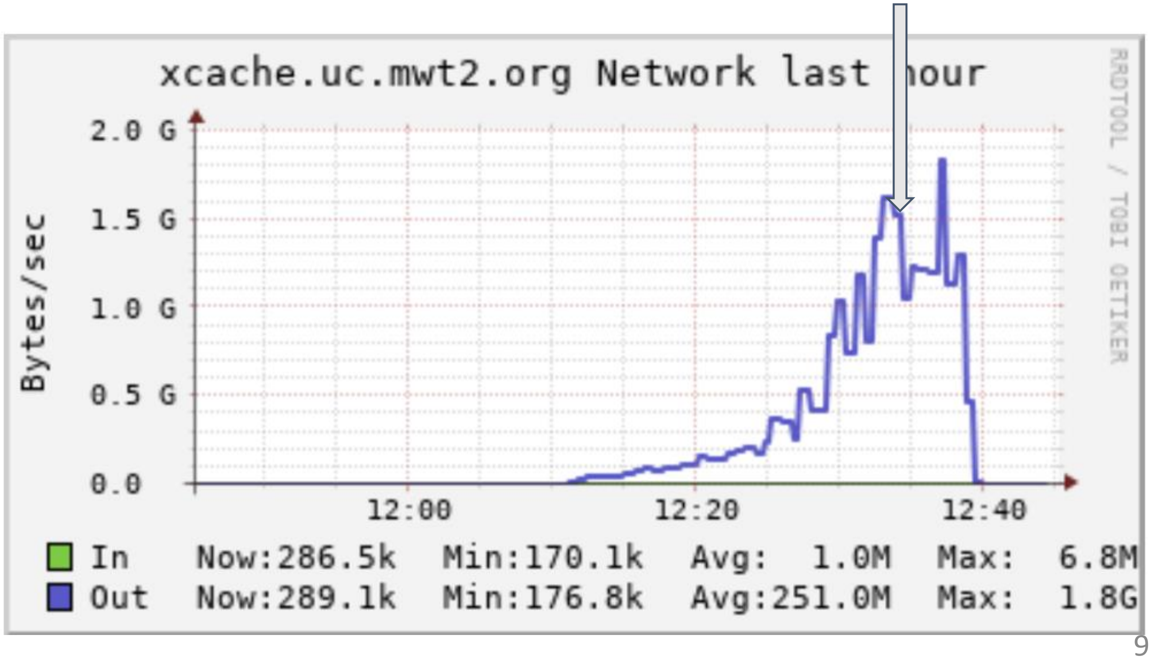
From tests by Ilija Vukotic on 10K files totalling 1TB

But then it is very fast

Second time:

- Running on 30K files
- Totaling 30 TB
- Could go even better with further optimization of consumers

Plato at 750 transformers reached.



```
■ In    Now:286.5k   Min:170.1k   Avg:  1.0M   Max:   6.8M
■ Out   Now:289.1k   Min:176.8k   Avg:251.0M   Max:   1.8G
```

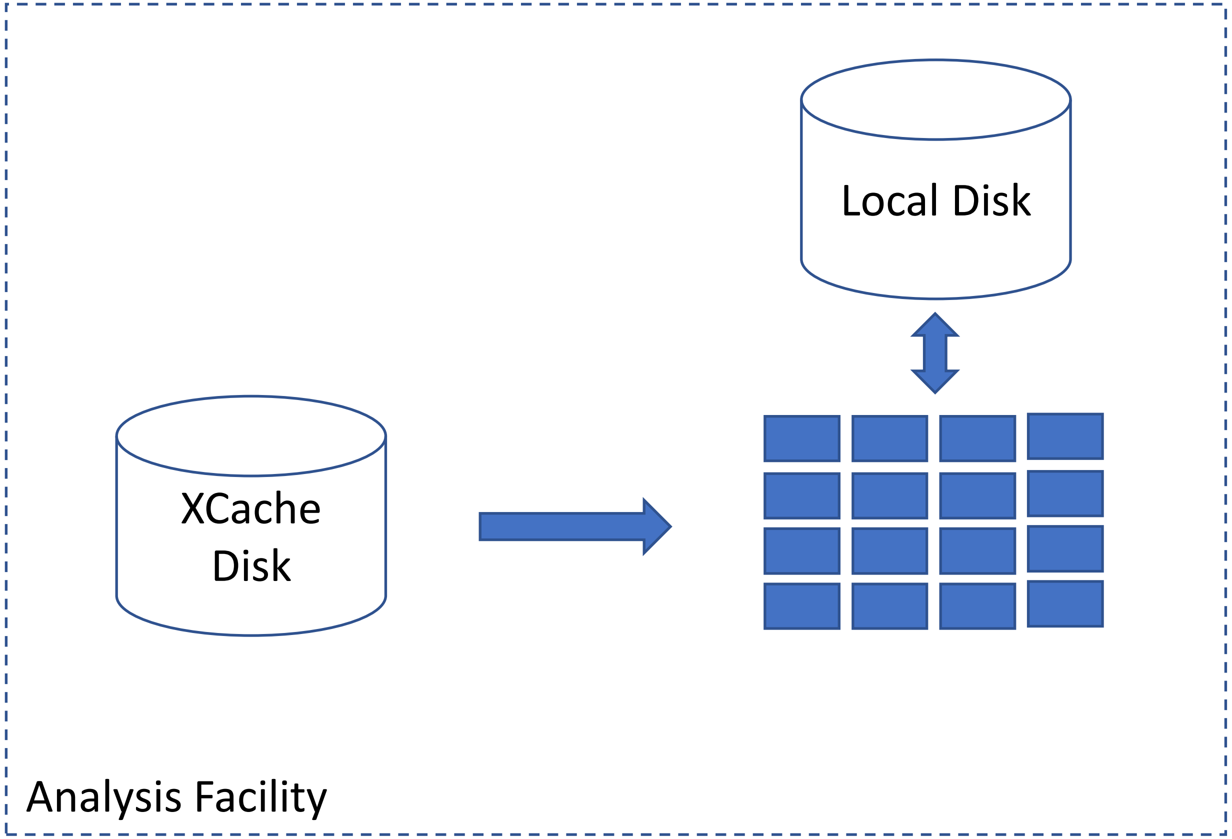
Local Data – Generating & Sharing Work

Two types of local disk

- Storage suited to home directories, small files, compile jobs – 10-100 GB’s?
- Large File Storage – >= 10 TB’s

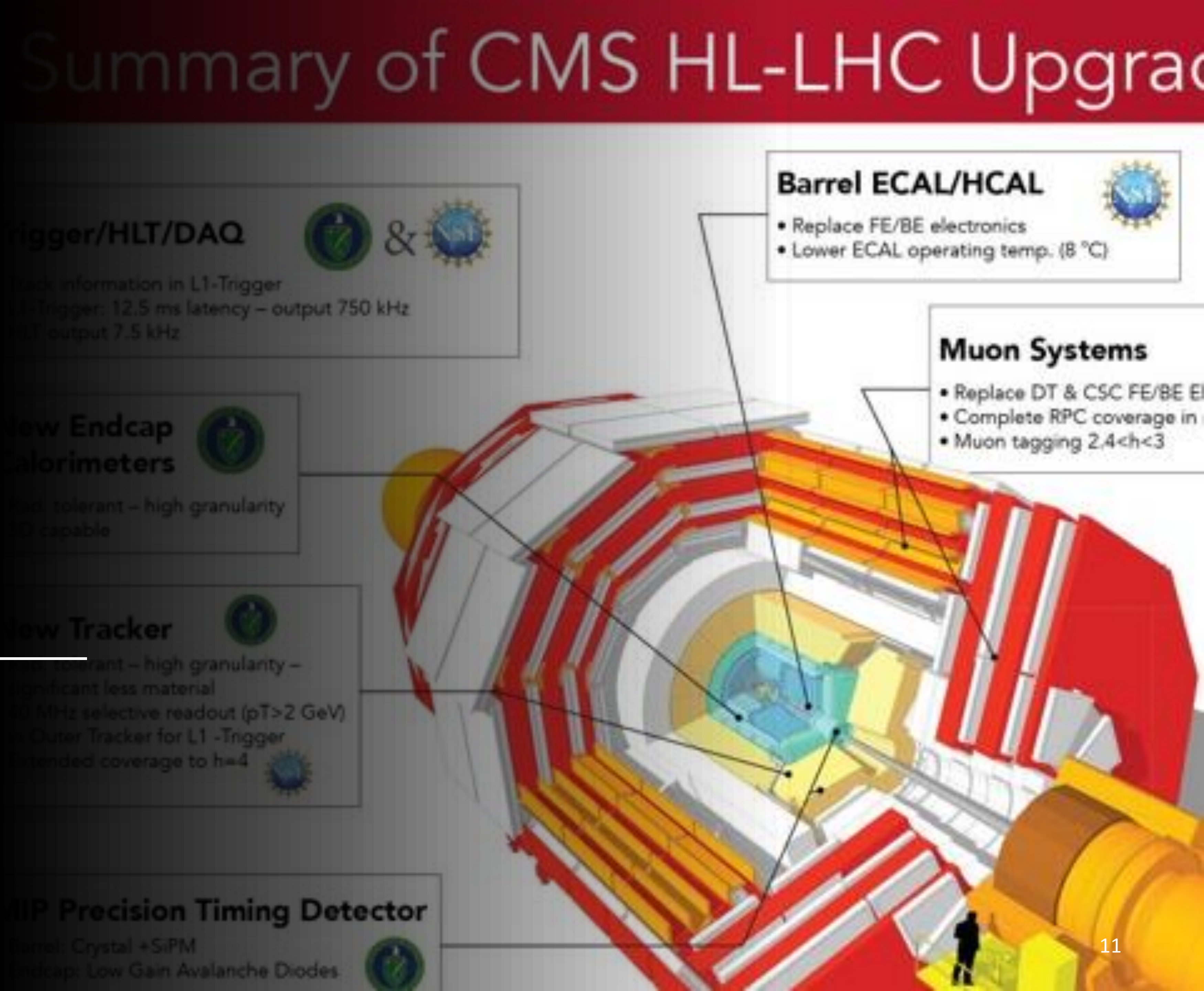
Sharing Options:

- Context: Local Batch, login-shell, Jupyter Notebooks, GRID, CERN
- Users: Analysis Groups, one or two other users, etc.



For examples see the German NAF

CMS Caching



Summary of CMS HL-LHC Upgrad

rigger/HLT/DAQ

rack information in L1-Trigger
1-Trigger: 12.5 ms latency – output 750 kHz
HLT output 7.5 kHz

Barrel ECAL/HCAL

- Replace FE/BE electronics
- Lower ECAL operating temp. (8 °C)

Muon Systems

- Replace DT & CSC FE/BE El
- Complete RPC coverage in
- Muon tagging 2.4<h<3

ew Endcap Calorimeters

Rad. tolerant – high granularity
3D capable

ew Tracker

tolerant – high granularity –
significant less material
40 MHz selective readout (pT>2 GeV)
in Outer Tracker for L1 -Trigger
Extended coverage to h=4

IP Precision Timing Detector

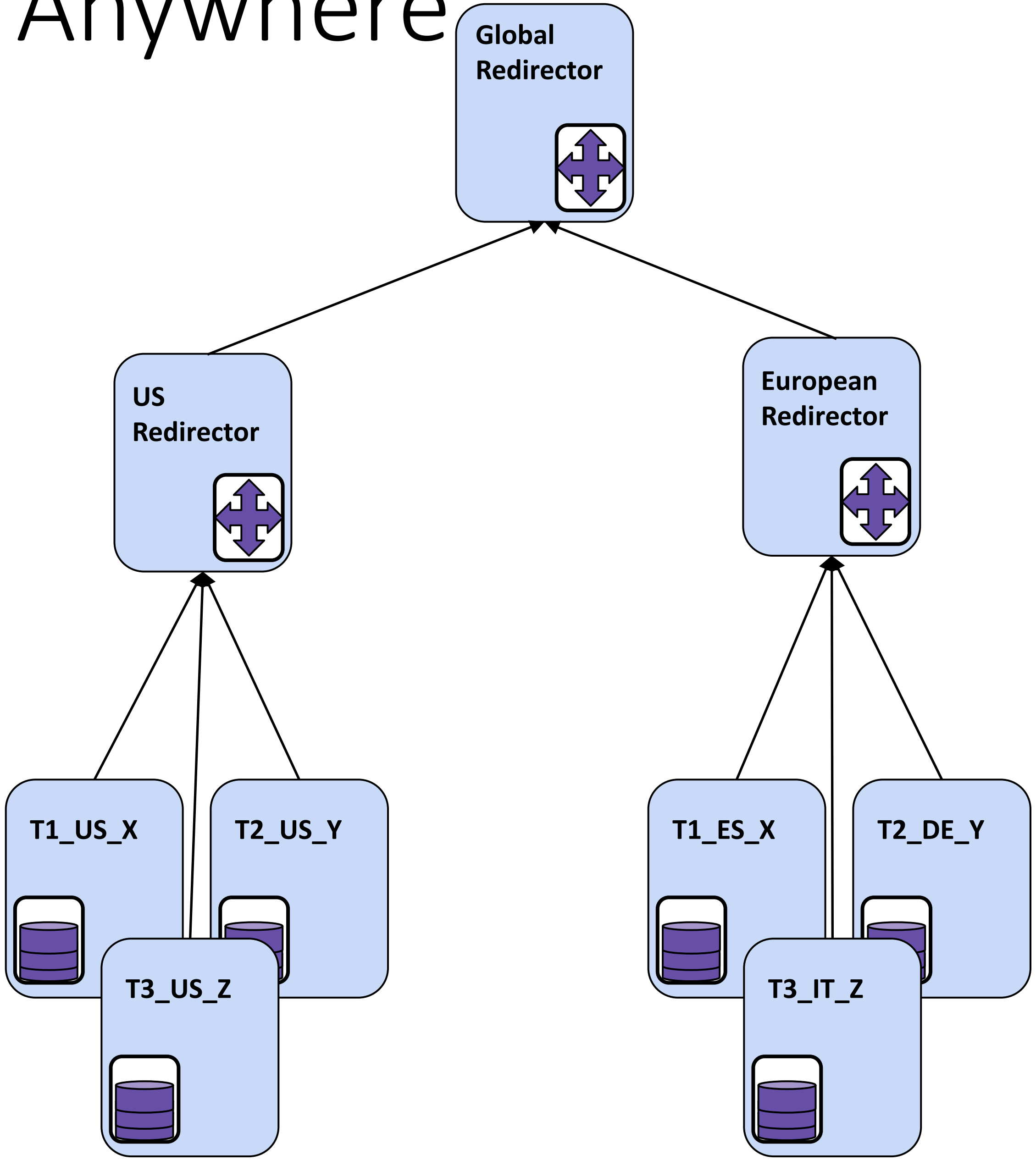
Barrel: Crystal +SiPM
Endcap: Low Gain Avalanche Diodes

CMS: Any Data, Any Time, Anywhere

A **tree shaped structure** based on the **XRootD framework** which allows any user to read any file within the CMS namespace.

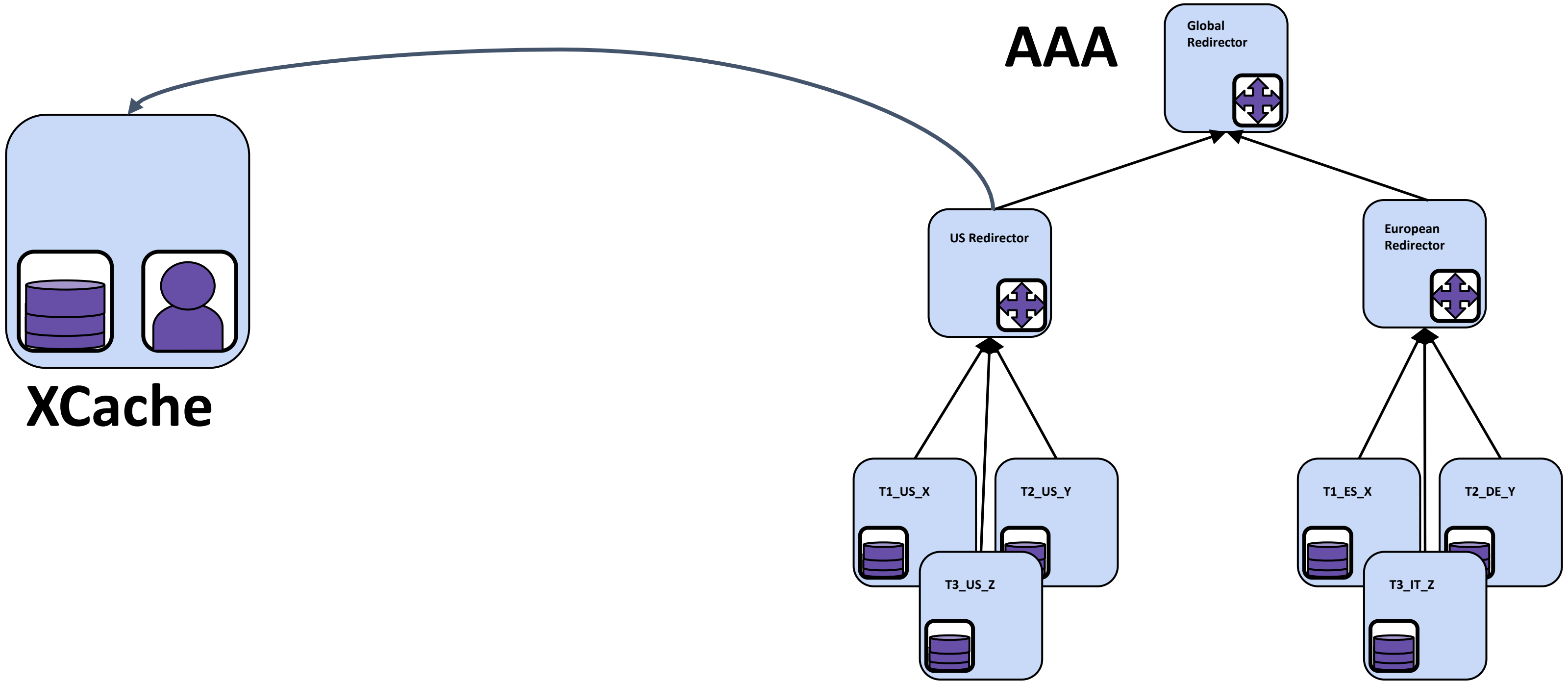
Every CMS site has a XRootD endpoint that connects to this tree.

There are **2 regional trees**: one in the US(FNAL) and one in Europe(INFN) **interconnected by a global redirector**(CERN), this tries to prevent a user in the US to read remotely from Europe and vice versa.



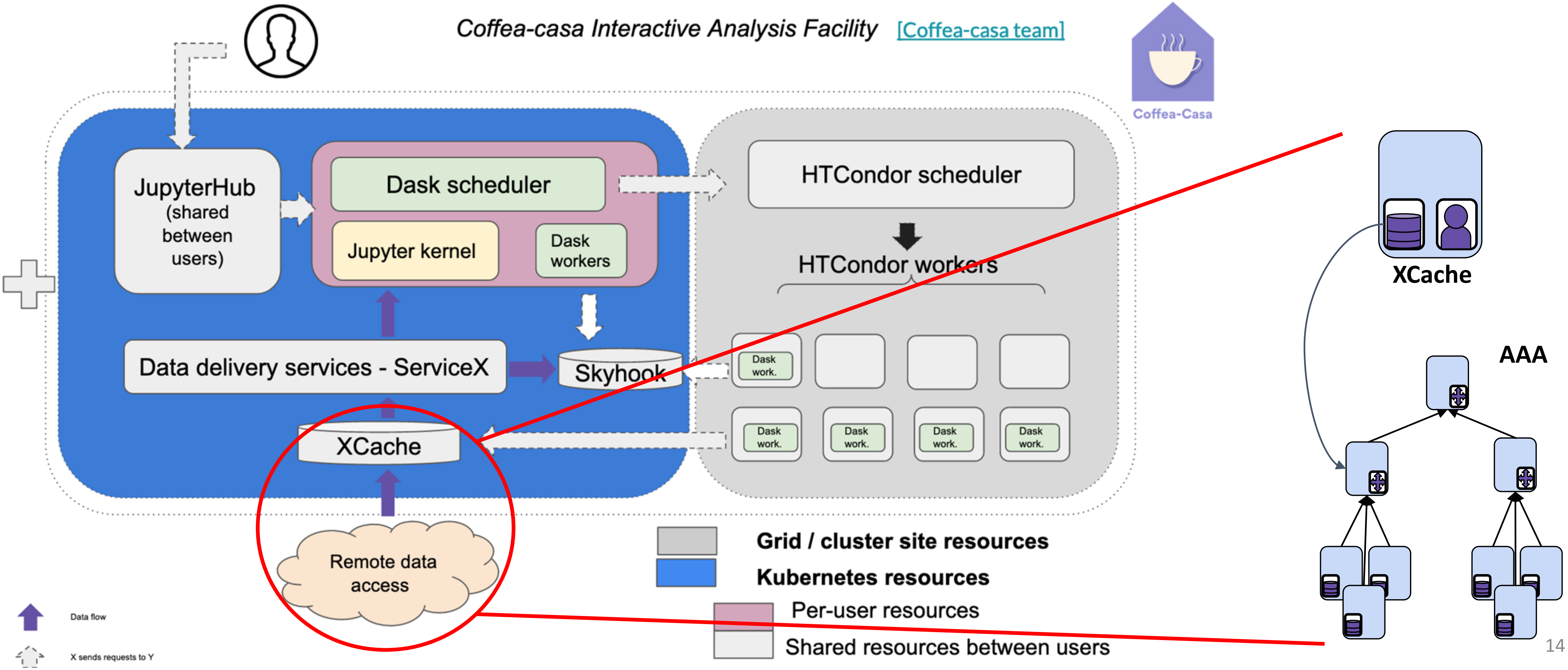
XCache and AAA

- CMS caches use AAA as source.
- Jobs submitted via CRAB use the Site's Trivial File Catalog (TFC) to automatically decide where to read for: a cache(if exists), locally or remotely
- Home-made scripts need to provide their own logic to decide where to read from



Coffea-Casa and XCache

Coffea-casa Interactive Analysis Facility [Coffea-casa team]



Coffea-Casa

JupyterHub (shared between users)

Dask scheduler

Jupyter kernel

Dask workers

HTCondor scheduler

HTCondor workers

Dask work.

Data delivery services - ServiceX

Skyhook

XCache

Remote data access

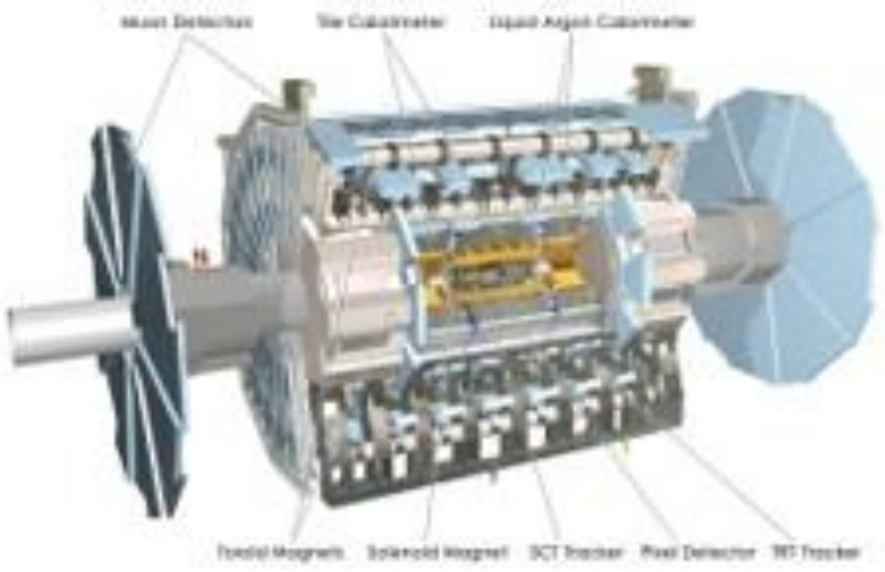
XCache

AAA

- ▨ Grid / cluster site resources
- 🟦 Kubernetes resources
- ▨ Per-user resources
- ☐ Shared resources between users

Data flow

X sends requests to Y

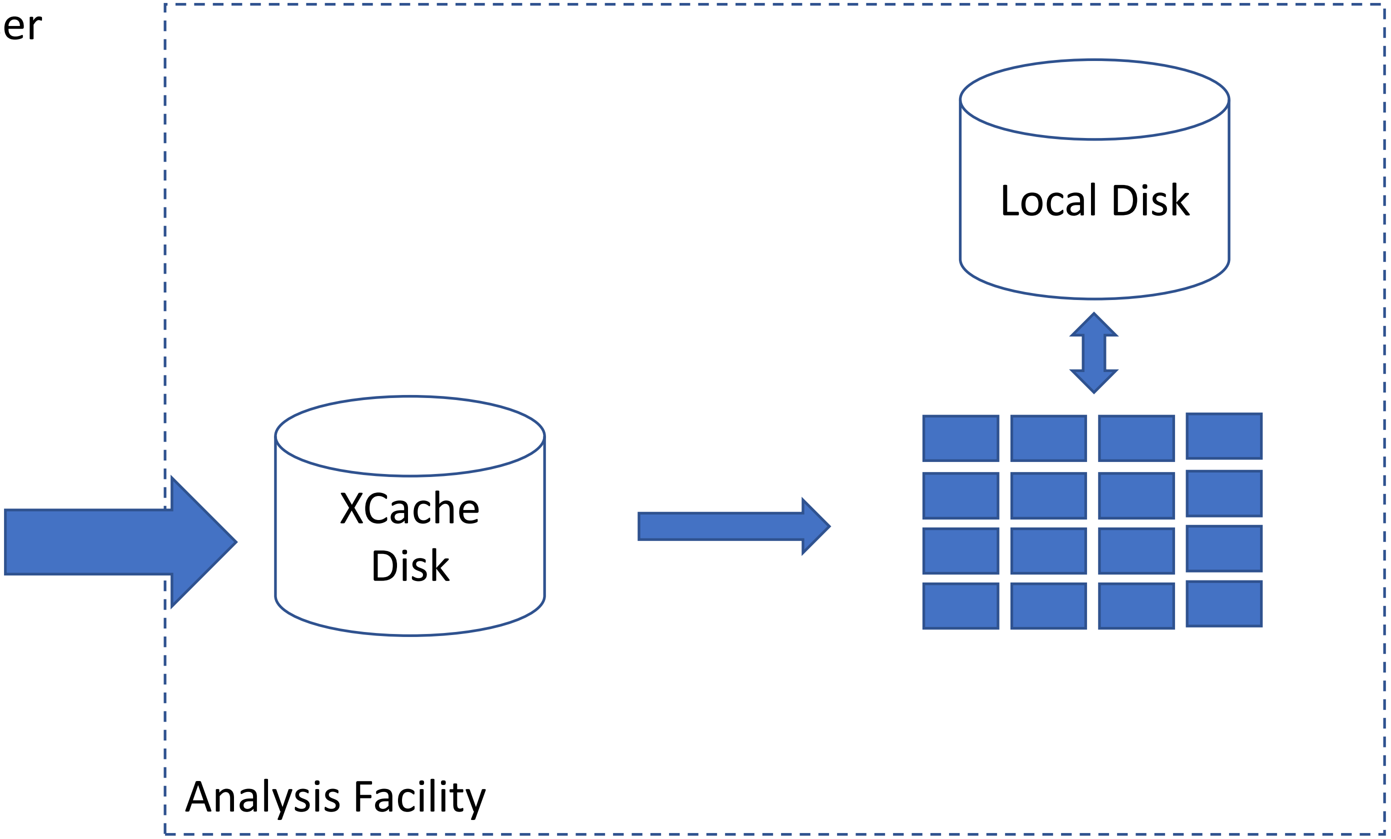


ATLAS Caching

Chicago AF

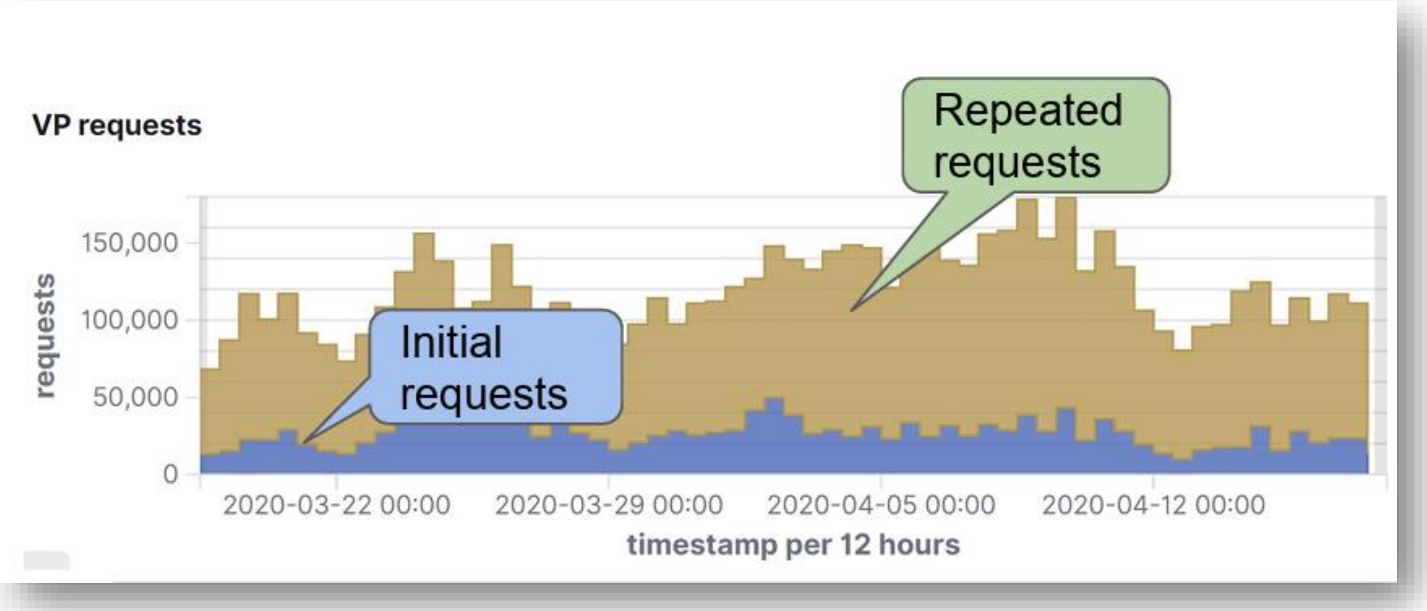
1. External Reads are triggered by user requests
2. Rucio generates the replica list to read
3. The xrootd URI's are rewritten to use the cache

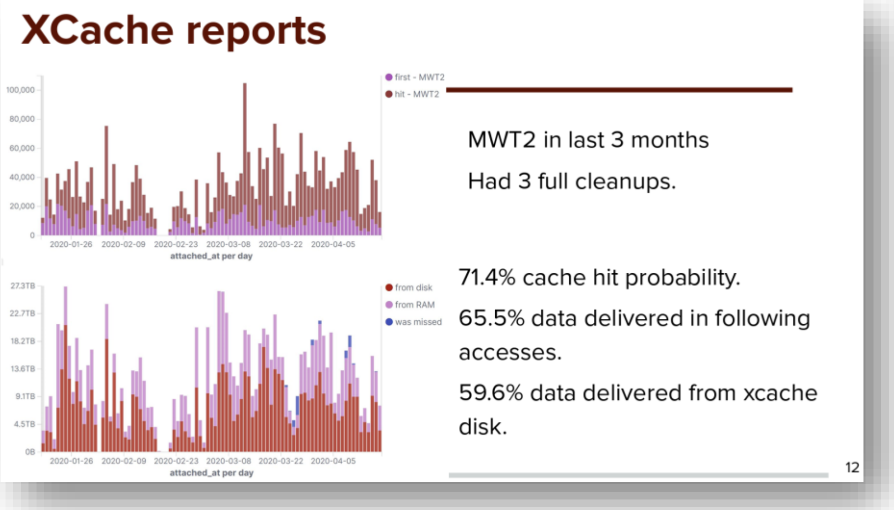
Requires every process running to locally know about the cache!



Some Extensive Testing

Usage shows who much a cache could help





XCache reports

MWT2 in last 3 months

Had 3 full cleanups.

71.4% cache hit probability.

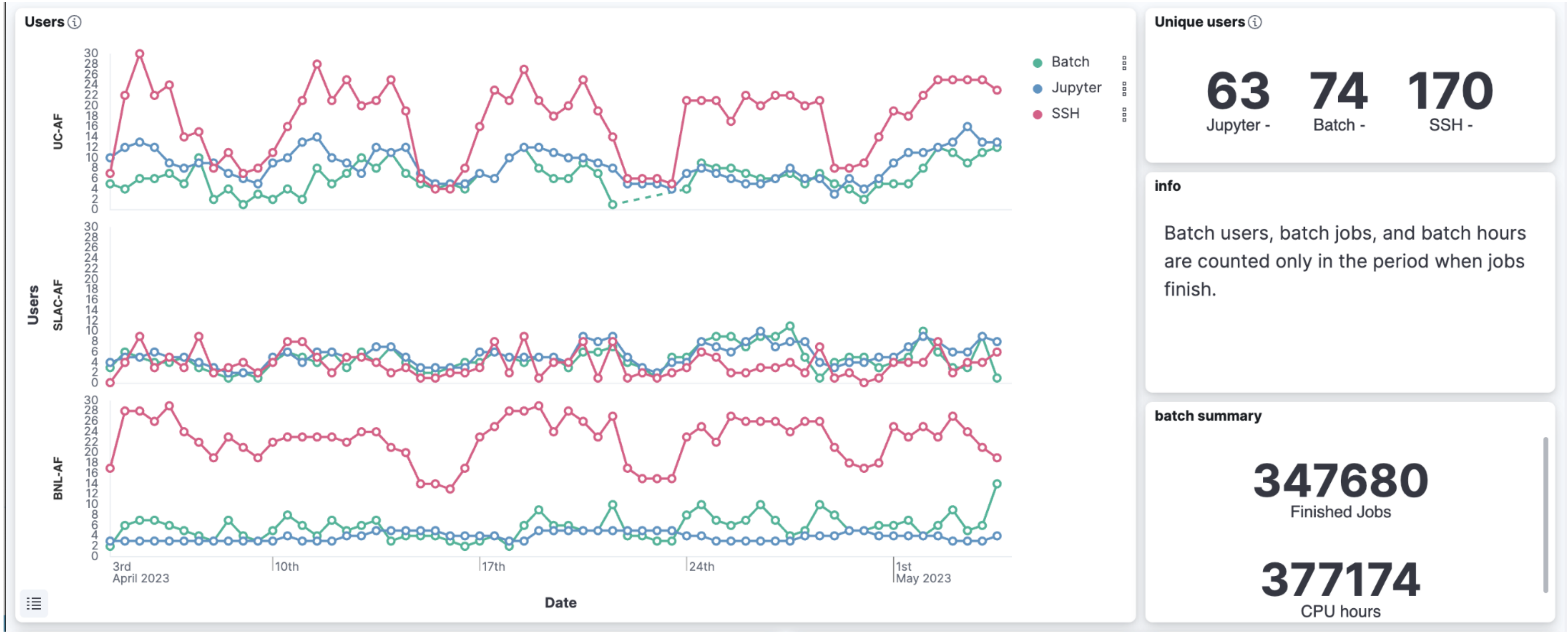
65.5% data delivered in following accesses.

59.6% data delivered from xcache disk.

Logical question with this sort of usage pattern:

- Should an AF without XCache be considered at all?

Chicago AF



Unique users

63 Jupyter -

74 Batch -

170 SSH -

info

Batch users, batch jobs, and batch hours are counted only in the period when jobs finish.

batch summary

347680 Finished Jobs

377174 CPU hours

From earlier this week (note dates)…

Discussion

Following slides are meant to start conversations – I will do my best to keep notes up here on the slides

Local Storage – Small User Files

- Used for: scripts, git checkouts, compiles, etc.
- ~GB’s (perhaps 100 GB to hold a few Athena builds?)
- Questions:
  - Technology: AFS? Posix semantics?
  - Linked with CERN account?

Local Storage – Data Files

- Context: Local Batch, login-shell, Jupyter Notebooks
  - Should it be accessible at CERN or the GIRD? Implications for scale-out?
- Sharing with another Analysis Facility
  - Easy to move/copy vs Can work on any analysis anywhere anytime, all at once
- Sharing: Local Analysis Groups, one or two other users, etc.
  - What about with other Analysis Facilities

Caching

- What size do we need to support N users?
  - What do we need to get to the point we can calculate this number?
- Accessible for all local jobs (batch, notebooks, terminal windows, etc.)
  - Technically easy?
- The DOMA Data Challenge is looking for an Analysis Facility related data challenge
  - What would make sense for ~1.5 years from now ("DC24")?
- XCache?