

# Analysis Facility Introduction: Facilities

WLCG-HSF Pre-CHEP Workshop 2023

[Whitepaper](#)  
[Session notes to contribute to](#)



# Foreword (repetita iuvant)

- The objective of this talk is to **provide an overview of what has been raised from the several contributors to the HSF AF forum activity**
  - Activities are in R&D and that this discussion is a checkpoint of what has been done so far, and that we are interested in as many opinions as possible on how to move forward
    - Users, analysis experts, system administrators,...
- Many of the topics come from the [white paper](#) draft that we already circulated
  - Thanks to everyone for the comments, we are looking forward to more coming after today's discussions
  - Some topics have dedicated talks and will only be touched upon

# Introduction

- HL-LHC Analysis Facilities R&D appears mainly focused on **extending the current models** toward a few set of open items:
  - Interactivity and introduction of more heterogeneous resources
  - Interoperability
  - Integration of cloud software (to satisfy broader set of requirements beyond batch).
  - Support for a large fraction of users (if not whole VO)
- It is important to look at **reusable building blocks** making AF and other resources interoperable
  - E.g. You can have a full blown AF at BNL or Fermilab but we can still deploy some blocks at Tier3s or adapt them to the grid.
- Uniformity of tools has made the grid capable of **supporting many communities**
  - Not same implementation but well defined requirements

# Integrating computing resources

The focus of the computing resources integration for analysis workflow is mostly related to the following two macro areas:

- **What the user sees:** framework interfaces for offloading payloads
  - prototyping my code and then being able to scale out over distributed machines
- **What the site sees:** resources management
  - Grid vs cloud vs ...

These are NOT some new problems, it is just the evaluation of new ways to do it that has been reported as something people are starting to look at.

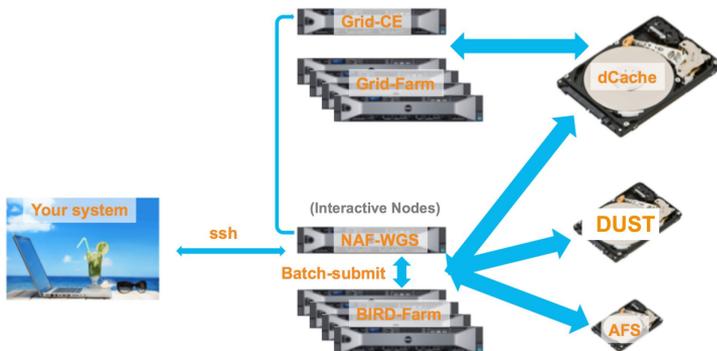
What follows is an attempt to have a (non-exhaustive) list of what are the **pros and cons of the presented R&D activities.**

# 1. HTCondor from a remote UI

- SSH login to a remote UI is not going away
- **Access to batch cluster** is a first class citizen in the current infrastructures dedicated to analysis (e.g. NAF and Ixplus)
  - Particularly **handy for reducing/skimming data to be shared by many people/groups**
    - Independently from the progress of the experiments in producing “already-reduced” data formats, this is extremely unlikely to disappear.
- The **turnaround for an interactive analysis is going to be a limiting factor in the future**

The main focus of all the R&D activity talks has been around extending the user experience with tools capable of enabling a more “interactive” way of performing the most-reused part of their analysis workflows.

# CERN and NAF examples



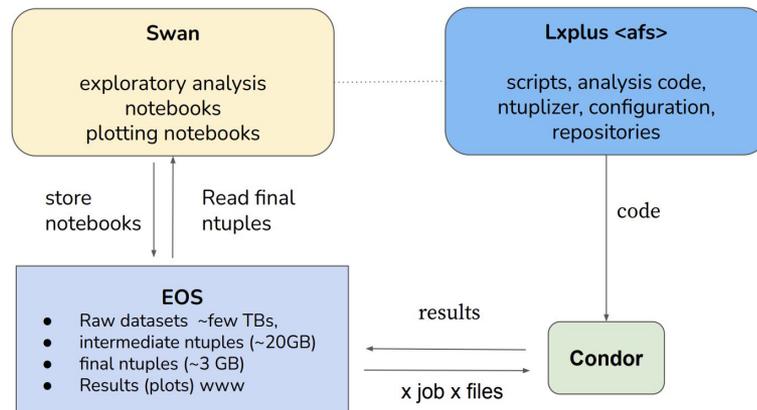
- NAF is considered an analysis facility for ATLAS.
- One of the main benefits of NAF is large and accessible storage.
  - Ease of sharing of the data between analysers inside DESY and in Germany.
- Many workflows supported so everything can be done in one location.

- NAF is vital for German CMS analyzers
  - for many, grid jobs are not even necessary

NAF: <https://indico.cern.ch/event/1214418>

- **Swan fits very well my needs for:**
  - prototyping code and algorithms
  - plotting final results
  - working on ML models interactively

- **It fills the gap between:**
  - full-scale analysis (condor jobs)
  - interactive play with the results (difficult to do by running scripts on lxplus) == definition of the jupyter notebook ;)

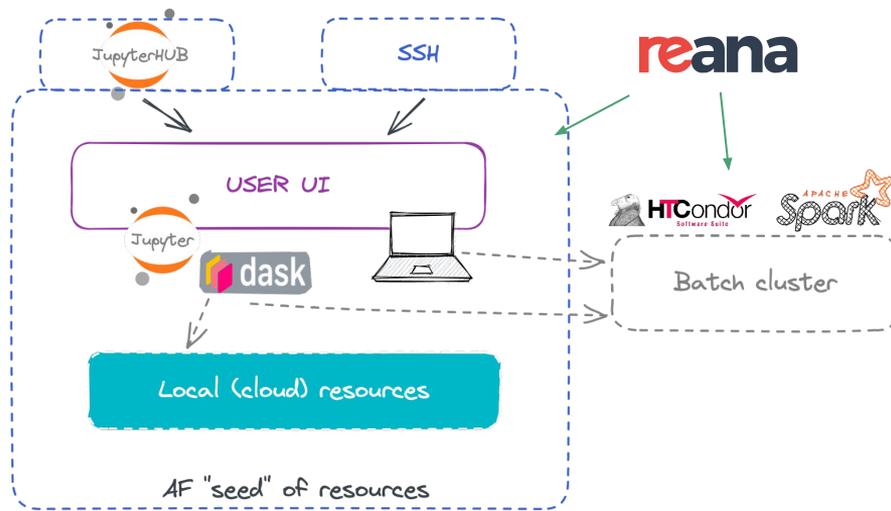
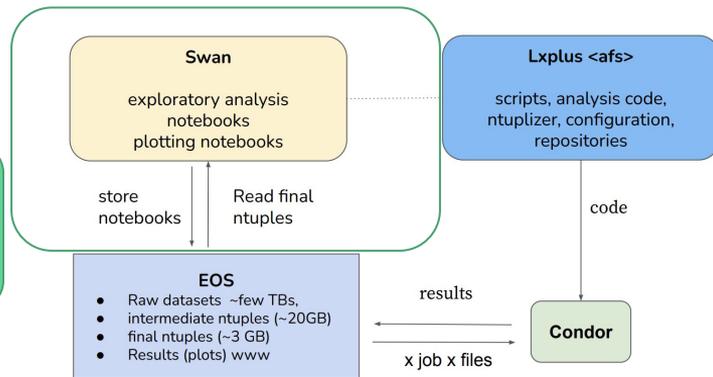


SWAN: <https://indico.cern.ch/event/1180396/>

## 2. AF R&Ds: Common traits

### Extending HTC resources consumptions with interactive (web) UIs

- Most of the presented R&D are **focused on pythonic environments**
  - Led by the data analysis of reduced formats (e.g. NanoAOD/PHYSLITE) frameworks
  - the need for a workflow management for generic environments it's on the radar → [REANA](#)
- **Containerized** UIs
  - This can be transparent to non-expert users that will use base images provided centrally
  - Opening for flexibility both at infrastructure and expert user level
- **Declarative** analysis
  - Hide code optimization in the framework, expose ~only physics
- **Offload** from local to distributed
  - With interoperability in mind, via an abstraction that will hide the resource manager interactions underneath



Let's see a summary of the presented initiatives →

# The HUB abstraction R&D

## There is a pattern...

A raising pattern of interest is surely around the **integration of JupyterLab experience with DASK parallelization framework capabilities**

Re-use of the batch resources to allow for scaling out python notebook execution.

The presented activities **differs mainly in the integration pattern** b/w the local/cloud resource seed where the UI runs and the batch cluster resources for scaling out.

- **Co-located clusters:** low latency, no network segregation
- **Federated infrastructure:** with existing distributed Tier2s

The two approaches cover complementary phase spaces and both are following a **co-design process with voluntary group of analyzers.**

Quite interesting to see unrelated activities “naturally” converging to a common ground → collaboration?

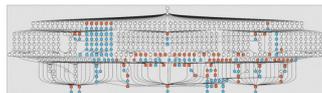
[SWAN](#)

[Coffea Casa](#)

[INFN](#)

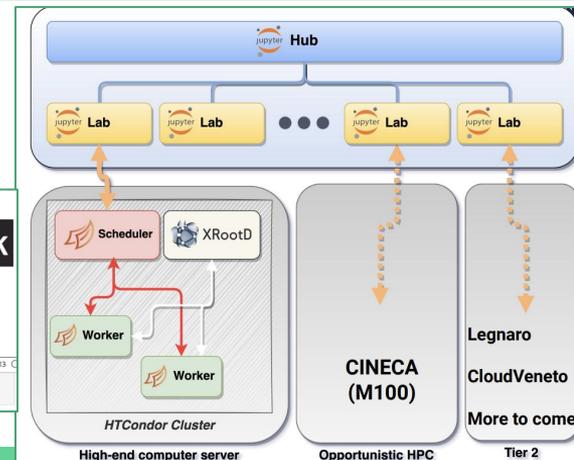
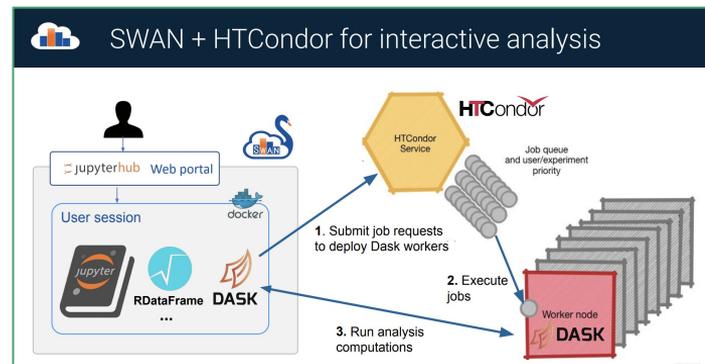
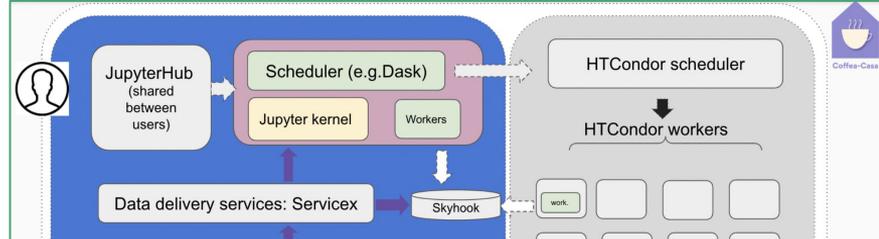
[NAF](#)

### dask-jobqueue



- attempt to make dask-jobqueue conveniently usable @NAF
- more interactive support of columnar analysis workflows
- spawn workers in **existing HTCondor infrastructure**

```
[10]: from dask.distributed import Client
      client = Client('tcp://131.159.168.86:4677')
```



# DOMA: Input Data organisation and access

- Data flows for HL-LHC analysis are **in definition**
  - Full reduced datasets (PHYSLITE/nanoAOD) are supposed to be only few PB but expect copies, different versions and derivatives to access and manage
  - **Not all analysis will be able to use PHYSLITE/nanoAOD**
  - AF should/could/would/? support all workflows
- **Latency** also a factor for input data for intense workflows
  - Usually achieved by a fast local storage serving the interactive resources
- **Caches** have some notable advantages
  - They can host a subset of data there is no need for the users to copy entire datasets
  - Analysis can have a highly repetitive access patterns particularly in the development stages when ideas are tested so once the data is cached it is reused

• Just replace the remote root filepath's redirector with xcache:

• `root://xrootd.unl.edu//...` → `root://xcache//...`

• xcache integration means your files will be cached and future analysis runs retrieve data faster

- Caches don't stop having AF centres specifically dedicated to analysis with all the users data, and don't cut off all the other resources either.

# DOMA: Shared storage

- Recurring topic: Local **shared storage** for people to seamlessly run from different resources and share with colleagues
  - Users repeatedly report EOS+CERNBox as one of the main reasons to use CERN
- If we want to express this in terms of functionalities
  - POSIX semantics
  - Common name space
  - Accessible by interactive nodes, cloud resources, batch system, and grid
  - Different protocols and services to interact with the resources (CERNbox, xrootd gateways, fuse mount)
  - Integrated with DDM (rucio, dirac...)
- Rucio/xrd/dav don't have any posix semantics but can offer the protocols
  - Some of these aspects could be object of R&D [rucio fuse-posix integration](#)
- What happens when we have **distributed facilities**?
  - **How can facilities share the same storage (AFS spoiled a lot of people)**
- Big issue to discuss is how to federate instances and data ownership
  - Mapping of general authorization with linux ACLs at different AFs ([CERNbox BoF discussion](#))

- **Huge PROs**
  - access to EOS
  - export of plots on EOS/www

Swan user CERN

# Federated Identity

- A big differences between an AF and a local resource is federated ID and VO access support more than services supported
  - i.e. a local resource may install dask, coffea and jupyterhub
- One of the biggest successes of the grid has been to **democratize access to computing resources.**
  - At the core of this is a Federated Identity Management based on X509 certificates
- Grid will move on to a token base AAI by the HL-LHC
- Integration with cloud technologies being proposed for AF is done much more easily with tokens
- AF tools should be built around tokens from the beginning

- Tokens give access to CMS data without certificate set-up

# Analysis Portability



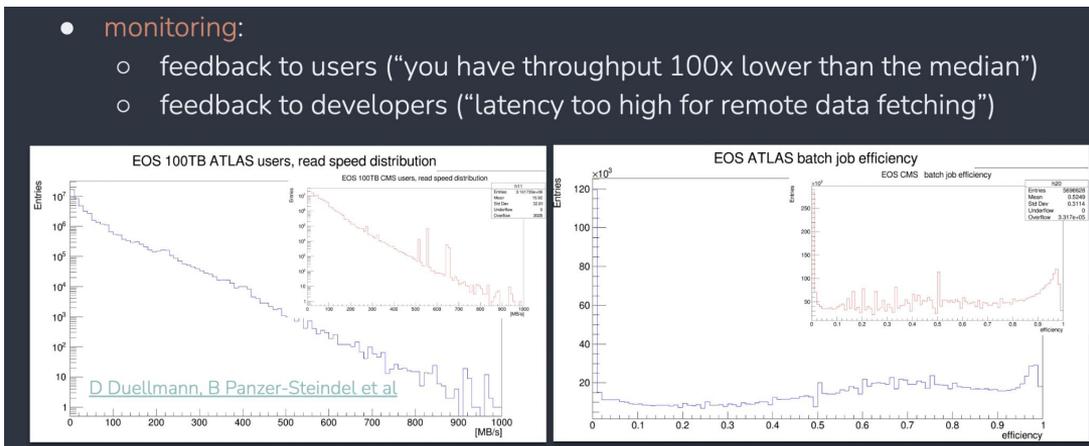
- Users want to **share with colleagues their setup, code, configuration, small amount of input data...**
- Shared storage is the traditional way of doing this but not the only way
  - Conda (LHCb), Containers (CMS, ATLAS)
- Solutions should be easy to setup and should help also preservation
- Typically when we talk about preservation we talk more about containers than conda
  - But even containers have their limitations for this aspect
- [CVMFS distribution and containers can be combined](#)
  - Currently /cvmfs/unpacked.cern.ch has now 3000 apptainer images

Distributing container images with CernVM-FS

Jakob Blomer (CERN)  
HSF Analysis Facilities Forum  
22 September 2022

# Monitoring and Metrics

- Need an agreed upon list of metrics and how to use them
  - Workflow ID, CPU, RAM, swap, I/O (local storage and network), Software stack, Job failure rate, Time To Completion (TTC), Data source local or cached from a Data Lake, Formats used on input (PHYS, PHYSLITE, DAOD, NTuple, etc..), Formats written (columns), ratios.....
- This is needed by users, developers AND AF resource providers



# Way forward

- Need to continue discussion between users, developers, providers...
- White paper should be a start not an end

