Analysis Facilities Community R&D Status

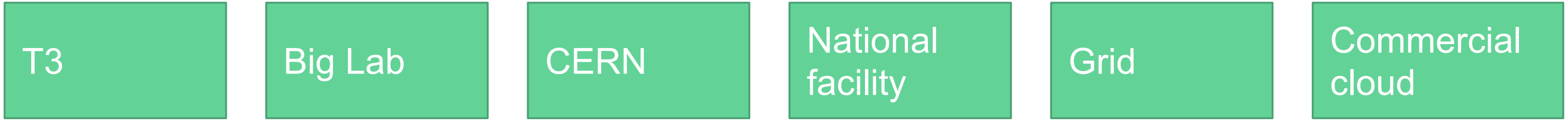
LHCC report

WLCG
Worldwide LHC Computing Grid

HSF
HEP Software Foundation

What is an Analysis Facility (AF)

- A loose definition
  - *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.*
- Do we have analysis facilities at the moment?
  - CERN (all WLCG users), NAF (only German users), FNAL (CMS), BNL (ATLAS),...
    - But also T3s, grid, commercial clouds....
    - GSI and Wigner for Alice are highly specialised sites for analysis trains without interactivity



T3

Big Lab

CERN

National facility

Grid

Commercial cloud

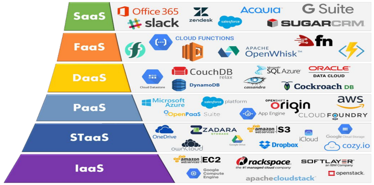
Why are we looking at this?

- Evolution of technologies and analysis techniques
- HL-LHC increased data and interactive computing requirements
    - Much higher trigger rates - many more events for analysis - may need to prototype on larger datasets
- Improve the users’ experience
- Improve site maintainability

Technology Evolution

kubernetes

SCI TOKENS



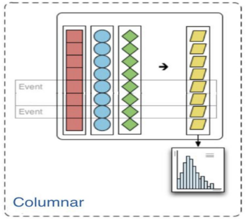
ServiceX

XCache

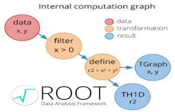
Techniques Evolution

DASK

jupyter



Scikit HEP

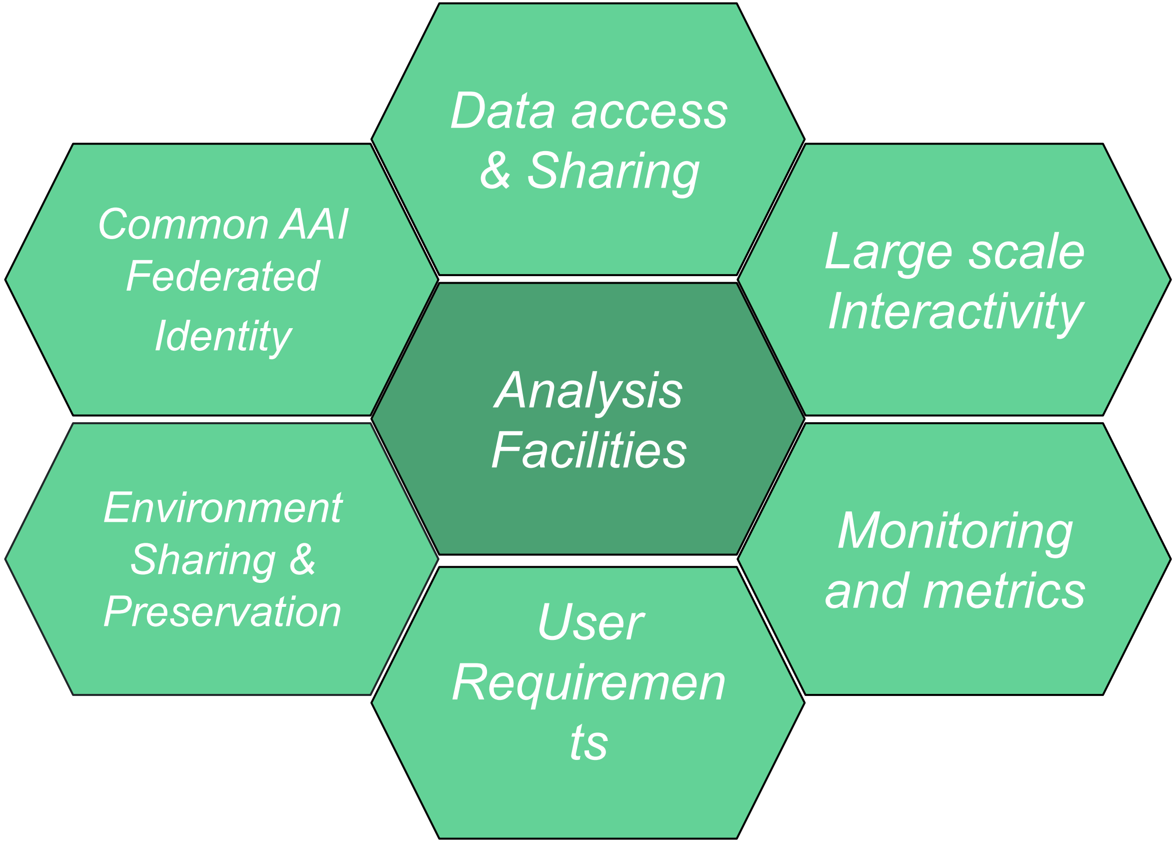


AFs are not isolated sites

A set of R&Ds to improve

Scalability
Ease of use
Data Access
Interoperability
Collaboration

For the end user analysis at the HL-LHC scale


Data access & Sharing
Common AAI Federated Identity
Large scale Interactivity
Analysis Facilities
Environment Sharing & Preservation
Monitoring and metrics
User Requiremen ts

Analysis User Experience: Starting points

- Analysers often leave the global grid early in the analysis cycle to work on local resources
    - T3s, National (NAF) or VO (FNAL) facilities, CERN
- Interactive analysis mostly defined by what fits onto a single machine
    - Developing code, plotting, fitting, playing with cuts on a small sample...
    - Data sample size could go from GB to TB
    - Large-scale interactive analysis not a focus of analysis experience today
- Different access systems between global and local resources
    - Interactive, batch jobs, grid jobs are different workflows
    - Global storage vs local storage have different semantics, accessibility and tools
        - Painstaking manual data preparation and placement
- Sharing outputs, usually via local filesystems
    - Assuming colleagues have access to the resources
    - Main reason for users flocking where everyone has an account.

Evolution of user requirements for HL-LHC

- Interactive fast analysis cycles on large datasets
- Convert interactive to batch-schedulable workloads
- Scale outside of the facility on occasion

Interactive scaling requirements

- Machine learning training models
- ML Inference within an analysis pipeline

ML support

- Efficient access to collaboration data
- Share analysis data artifacts with the collaboration
- Access to user data formats with low latency

Storage access and sharing

- Collaborate in a multi-organizational team on the same AF
- Move Analyses between facilities

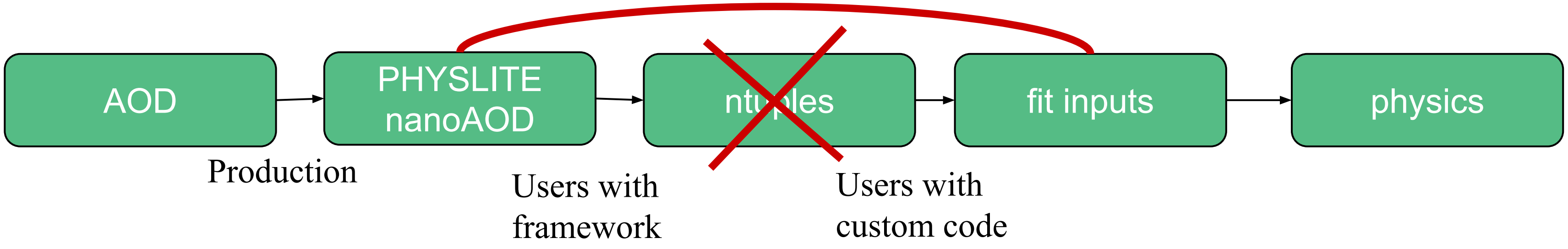
Federated Identity

AAI

- Instantiate desired software stack
- Run legacy analysis
- Share environments with colleagues

Analysis portability

Core use case



AOD

PHYSLITE nanoAOD

~~ntuples~~

fit inputs

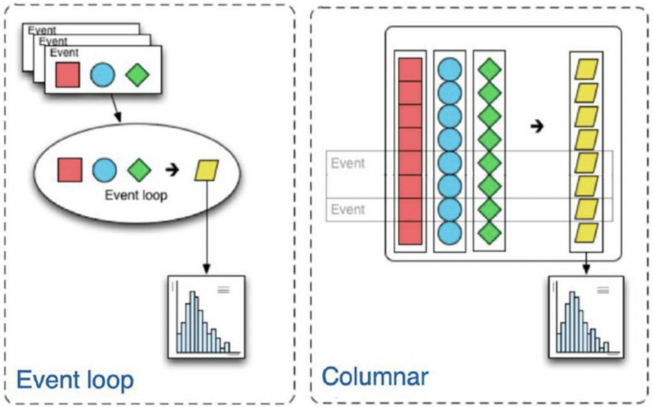
physics

Production

Users with framework

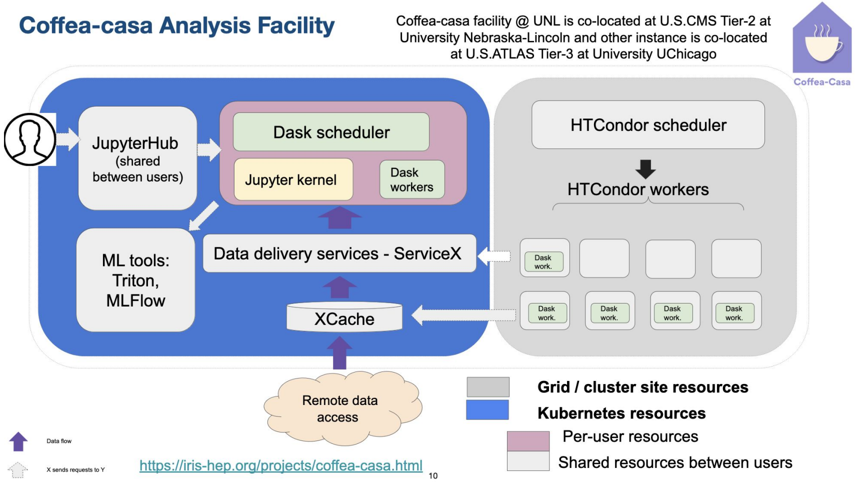
Users with custom code

- Most of the ideas for AF facilities R&D have been developed for Columnar Analysis (CA) with python (uproot or RDataFrame) on reduced formats of pre-calibrated data.
  - Strong ML support.
- Caveats:
  - Reduced formats are still 1-3PB/y
  - Not all analysis will be able to use these formats
    - CMS ~50% can use nanoAOD
    - ATLAS aiming at 80% of analysis to use PHYSLITE



Coffea-casa: an example of AF

- Designed completely around cloud technologies
- Offers Jupyterhub with a kubernetes backend for interactive resources, an integrated dask scheduler and ML tools.
    - If the jobs scale beyond the capability of the kubernetes resources the system automatically offloads to the local Tier2 batch system. And this is transparent to the user.
- Non storage elements in the blue box are natively integrated.
    - And native support for token based AAI



Coffea-casa Analysis Facility

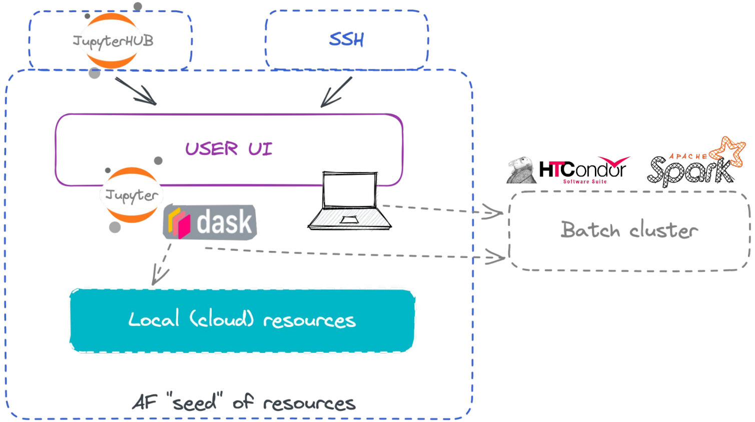
Coffea-casa facility @ UNL is co-located at U.S.CMS Tier-2 at University Nebraska-Lincoln and other instance is co-located at U.S.ATLAS Tier-3 at University UChicago

https://iris-hep.org/projects/coffea-casa.html 10

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

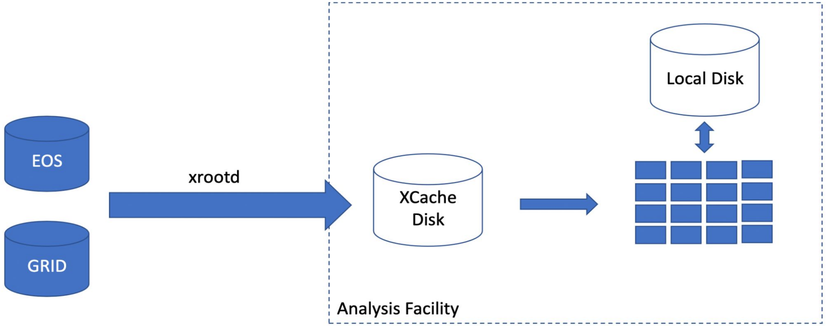
Not a revolution

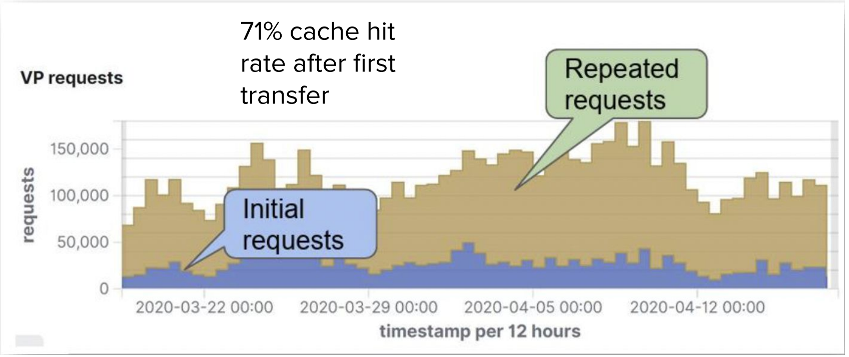
- Does it mean AF facilities with a batch system will have to convert to kubernetes?
  - Jupyterhub can also be integrated on top of an HTCondor cluster NAF did that
- Is it possible to integrate some of this on more distributed resources
  - CMS people at INFN have setup coffea to offload to all their Tier2s
- Does it mean we will not use ssh anymore to access the batch systems?
  - No, most facilities support traditional submission **and** jupyterhubs (CERN, NAF, UChicago, BNL...)
- What about RDataFrame?
  - RDataFrame is also python and is being also tested
  - It is interfaced with dask
- What about the other workflows?
  - Traditional support will not change
  - benefit from a fast cycle for testing and prototyping
- Do we need Jupyter notebooks support at all?
  - They are a useful tool for quickly prototyping code
  - SWAN was built around Jupyter too



DOMA: Input data organisation and access

- Data flows for HL-LHC analysis are **still being defined**
- **Latency** of intense workflows usually reduced by a fast local storage serving the interactive resources
- Caches notable advantages
  - Automatically transfers only accessed data, it doesn't copy the entire datasets
  - Many analyses have a highly repetitive stage, particularly in the development stages when ideas are tested
- Dedicated AFs can still host complete datasets locally
  - Caches also enable smaller centres to remain viable and useful





DOMA: Shared storage

- Local **shared storage** for people share data with colleagues
- Usually a large POSIX-like fs
  - Users repeatedly report EOS as one of the main reasons to use CERN
  - But also NAF (DE only) users expressed the same regarding NAF storage
- Distributed AFs and remote access
  - Some AFs are mounting EOS on their interactive resources
  - CERNbox to share data with remote users - a lot of interest from sites!
- Global Storage could also be used to share between facilities
  - This would mean consistently using DDM tools also for local files
    - When users leave the grid they stop using the DDM tools and rely on the local file system
- Object Stores vs POSIX is a recurrent discussion
  - Users like POSIX-like storage because the applications and the interactive login both work
  - But there is a need to understand where OS can be integrated and what POSIX functionalities are really needed

Analysis Portability & Preservation

- Users want to share with colleagues their setup, code, configuration, small amount of input data....
  - Both at the same facility and between facilities
- Solutions should be easy to setup and should help also preservation
- Containers not straightforward to build from scratch
  - Experiment effort to build base images for different sw bases
- CVMFS distribution and containers can be combined
  - Currently /cvmfs/unpacked.cern.ch has ~3000 images
- Containers help flatten also differences between AFs and are a key ingredient for interoperability

Distributing container images with CernVM-FS

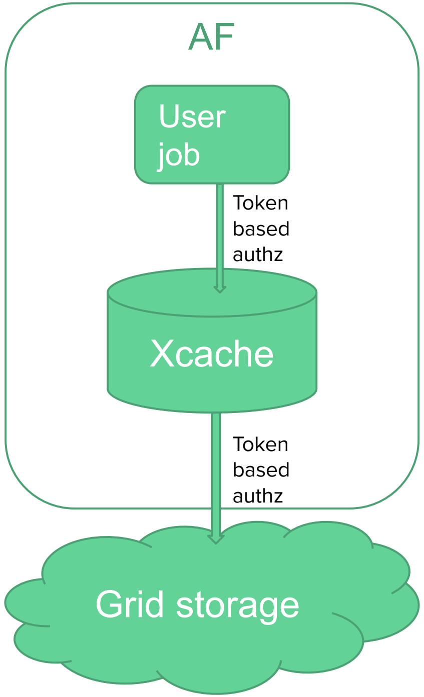
Jakob Blomer (CERN)
HSF Analysis Facilities Forum
22 September 2022

Resource Access

- Large HEP experiments are multi-organisational (global enterprises)
  - Ability to move facilities and get a similar experience is useful
- Grid was successful in democratizing the access to resources.
  - Equitable access to a global pool of resources for any VO member.
- Can this be replicated for interactive resources?
  - Historically security teams didn’t want to give access to interactive resources
  - Many AFs are already giving access to their interactive resources to the whole VO
- Calls for integration with the common AAI
  - Can the grid sites’ definition be extended to include interactive resources? (easier)
  - Can local resources host a physics group? (more difficult, not impossible)

Common AAI and resource access

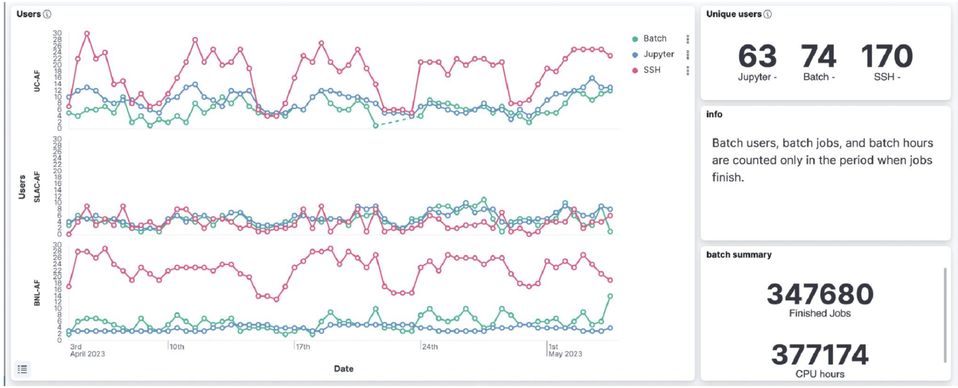
- For AFs there are two reasons to use the same AAI
    - Enable users to access different facilities (resources access requirement)
    - Allow for integration with grid resources in particular with the storage
- Grid slowly moving to a token base AAI by the HL-LHC (T. Dack CHEP2023)
- Integration with cloud technologies can be done much more easily with tokens
    - Coffea has a web based interface and the services between the user and the data use tokens
    - Classic AF (ssh+batch+POSIX FS) will need development work



B. Bockelmann, WLCG/HSF 2023

How do we know all this is useful, works, is optimised?

- Monitoring, tracking, benchmarking…
- Analysis is chaotic and often “dark”
    - Users go off to their local resources and we lose track of them
- Analysis Grand Challenge (AGC) extensively used
    - Set of analysis workflows distributed with OpenData to use as a benchmark
- Legitimate reasons to choose carefully metrics of “success”
    - We routinely monitor resources and trends ➔ results require interpretation



63 Jupyter -
74 Batch -
170 SSH -

info

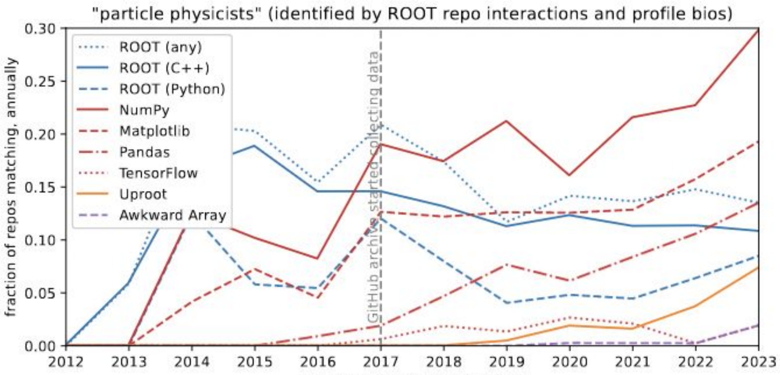
Batch users, batch jobs, and batch hours are counted only in the period when jobs finish.

batch summary

347680 Finished Jobs

377174 CPU hours

ssh , batch, Jupyter usage
USATLAS federated T3



Tracking of CMS software trends

Conclusions

- Analysis will continue to be supported at sites of different sizes
    - The aim is to make it easier to integrate building blocks necessary to run workflows
- The R&D on Analysis Facilities will continue guided by users and facilities
    - The experiment feedback would be useful
- Columnar Analysis core use case should cover the majority of analysis but….
    - Experiments should look more in depth at how the analysis at the HL-LHC will look
- Computing around us keeps evolving
    - Integrating new technologies takes several years
    - AFs are freer to experiment

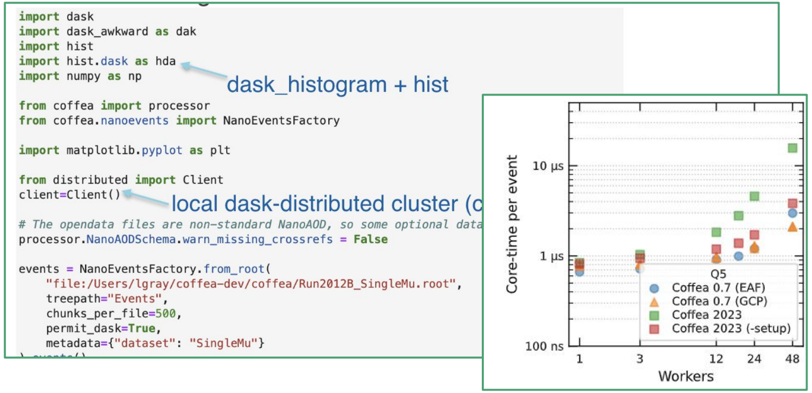
Backup/Further reading

User requirements

L. Heinrich WLCG/HSF WS 2023

Interactive Analysis beyond single-node

- Analysis work lives across a spectrum of (new R&D ↔ crank-turning later on) with varying user expectations on access, availability & turn-around
- Maintaining the ability to creatively explore data at the HL-LHC is still important and may not fit "single-node".
- Software is being prepared for this across the board
  ➔ if we want to use these modes, facilities need to support it



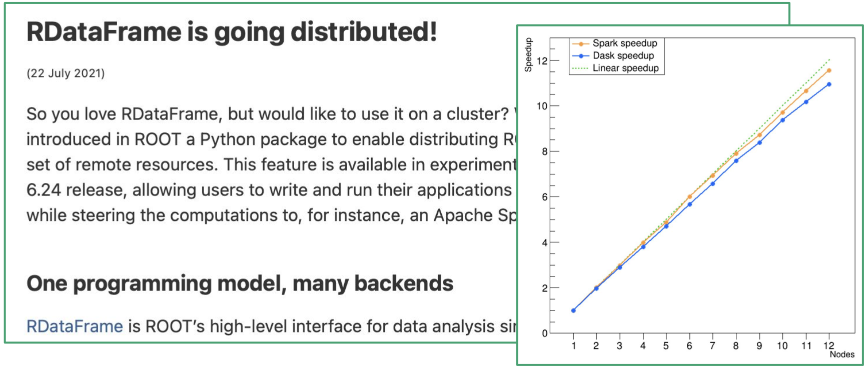
```
import dask
import dask_awkward as dak
import hist
import hist.dask as hda
import numpy as np

from coffea import processor
from coffea.nanoevents import NanoEventsFactory

import matplotlib.pyplot as plt

from distributed import Client
client=Client()

# The opendata files are non-standard NanoAOD, so some optional data
processor.NanoAODSchema.warn_missing_crossrefs = False

events = NanoEventsFactory.from_root(
    "file:/Users/lgray/coffea-dev/coffea/Run2012B_SingleMu.root",
    treepath="Events",
    chunks_per_file=500,
    permit_dask=True,
    metadata={"dataset": "SingleMu"}
```

dask_histogram + hist

local dask-distributed cluster (c

L. Gray CHEP 2023

https://root.cern/blog/distributed-rdataframe/



RDataFrame is going distributed!

(22 July 2021)

So you love RDataFrame, but would like to use it on a cluster?
introduced in ROOT a Python package to enable distributing R
set of remote resources. This feature is available in experiment
6.24 release, allowing users to write and run their applications
while steering the computations to, for instance, an Apache Sp

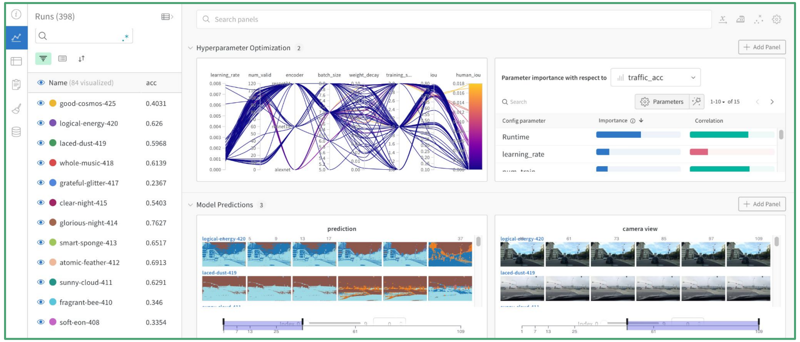
One programming model, many backends

RDataFrame is ROOT's high-level interface for data analysis si

Machine Learning and Heterogeneous Resources

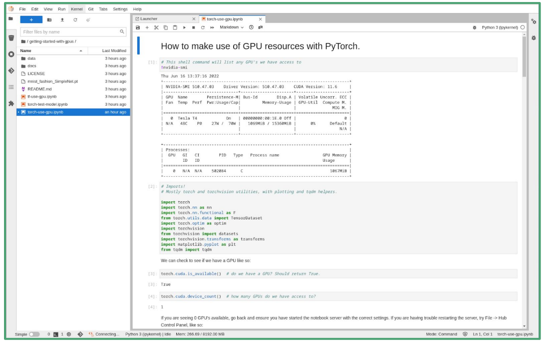
ML is and has been growing in analysis. As a user one expects to be able to exercise the full ML analysis lifecycle on an AF

- Data Exploration & Preparation, Interactive R&D and training
- Large-scale non-interactive training and hyperparameter optimization
  - Requires large GPU resources currently not available at standard WLCG sites
- ML Inference within an analysis pipeline



Runs (398)

Name (84 visualized)	acc
good-cosmos-425	0.4031
logical-energy-420	0.626
laced-dust-419	0.5968
whole-music-418	0.6139
grateful-glitter-417	0.2367
clear-night-415	0.5403
glorious-night-414	0.7627
smart-sponge-413	0.6517
atomic-feather-412	0.6913
sunny-cloud-411	0.6291
fragrant-bee-410	0.346
soft-eon-408	0.3354



How to make use of GPU resources with PyTorch.

```
[1]: # This shell command will list any GPU's we have access to
     !nvidia-smi

Thu Jun 16 13:37:16 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.47.03    Driver Version: 510.47.03    CUDA Version: 11.6     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            On   | 00000000:00:1E.0 Off |                    0 |
| N/A   48C    P0    27W /  70W |   1069MiB / 15360MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A    502084      C                                    1067MiB |
+-----------------------------------------------------------------------------+
```

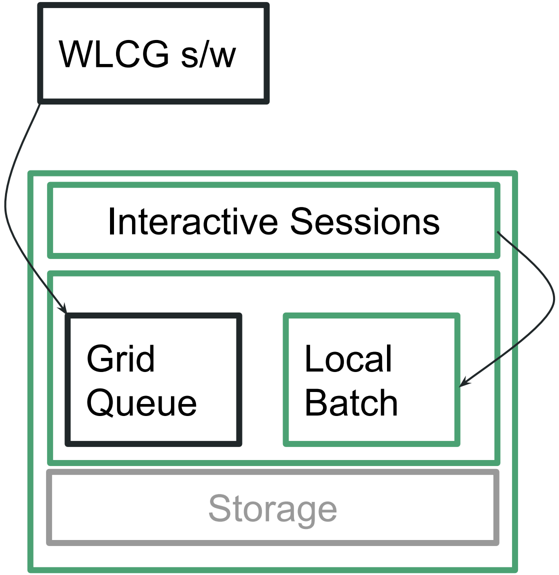
```
[2]: # Imports!
     # Mostly torch and torchvision utilities, with plotting and tqdm helpers.

     import torch
     import torch.nn as nn
     import torch.nn.functional as F
     from torch.utils.data import TensorDataset
     import torch.optim as optim
     import torchvision
     from torchvision import datasets
     import torchvision.transforms as transforms
     import matplotlib.pyplot as plt
     from tqdm import tqdm
```

We can check to see if we have a GPU like so:

```
[3]: torch.cuda.is_available()  # do we have a GPU? Should return True.
[3]: True
[4]: torch.cuda.device_count()  # how many GPUs do we have access to?
[4]: 1
```

If you are seeing 0 GPU's available, go back and ensure you have started the notebook server with the correct settings. If you are having trouble restarting the server, try File -> Hub Control Panel, like so:

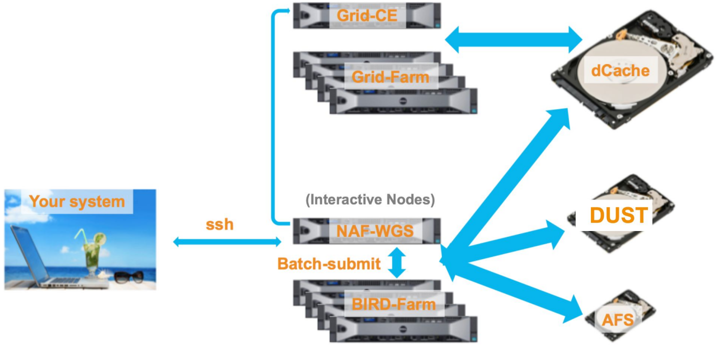
Integration into federated infrastructure



- From current analysis facilities it’s possible to interact with the global infrastructure
  - send jobs, receive data
  - if grid-site co-located: also receive jobs
  - at the same time AF currently are very distinct from grid
- Frequently stated: facility should be able to do full analysis lifecycle, but this should not lead to sealed facilities.
- Evolve what it means to be a grid-site instead of replacing them
  - e.g. add interactive options

L. Heinrich WLCG/HSF WS 2023

NAF and SWAN Users



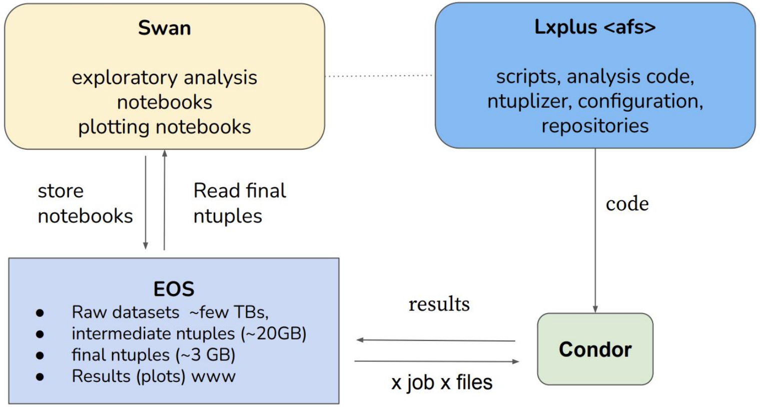
- NAF is considered an analysis facility for ATLAS.
- One of the main benefits of NAF is large and accessible storage.
  - Ease of sharing of the data between analysers inside DESY and in Germany.
- Many workflows supported so everything can be done in one location.

- 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

exploratory analysis notebooks
plotting notebooks

Lxplus <afs>

scripts, analysis code, ntuplizer, configuration, repositories

store notebooks

Read final ntuples

code

EOS

- Raw datasets ~few TBs,
- intermediate ntuples (~20GB)
- final ntuples (~3 GB)
- Results (plots) www

results

x job x files

Condor

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

CMS User for general AF

- Plans for several CMS analysis facilities with services, software, hardware for analysis and dedicated support team
  - Reliable platform to plug in technologies and enable efficient analysis
- Services:
  - Access to experimental data products ← Access to global storage
  - Storage space for per-group or per-user data (often ntuples) ← Shared storage
  - User support
  - Physics software: ROOT and the growing Python-based ecosystem ← same ecosystems
  - Computing hardware: CPUs and disks (in future GPUs) ← Heterogeneous h/w to go with the storage
- Data Access desiderata:
  - Flexibility: Remote running (xrootD remote reads) vs local mass storage (eos) vs cache (content aware) vs local disk ← seamless access to different types of storage (Data Lake?)
  - [.....]
- ML workflows support
  - [......]

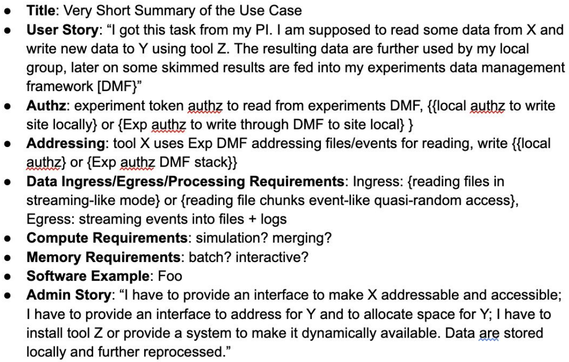
M. D’alfonso Analysis mini-workshop (2021)

How do we know this is important for all (most?) users?

- To design a new infrastructure
  - A model is written by expert users, computing trends and solutions that may satisfy the requirements of the model are looked at and proposed, a test infrastructure is setup and tried
    - This is what is happening with coffea-casa AF types of facilities
- Average users use the path of least resistance or what they are offered
  - Right now they are offered mostly ssh+batch and moving to something else is either not documented or not fully supported or too much effort
- System administrators and site managers would like us to provide well defined requirements.
  - Requirements vs implementations
    - “I like EOS” vs “I need this I/O and this AAI to interact with my data”
    - Users don’t know these details

Tracking users

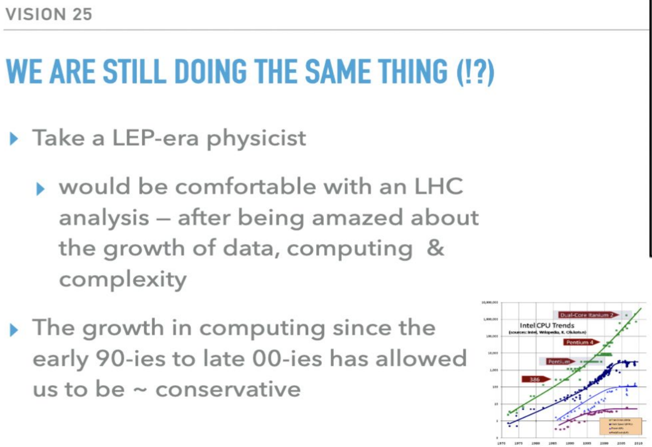
- A possibility to solve this is to track what users do, automatically where possible.
  - *The "analysis" step is the only one in the pipeline for which we don't even know*
  - *who all the users are*
    - [H. Schreiner CHEP 2023]
  - This will come back in later slides when talking about the metrics
- Another possibility is try to make surveys interesting
  - "I have a story" initiative proposed in the HSF forum for example
    - Pair a user with a system administrator and try to understand the computing requirements behind a use case
      - [T. Hartmann survey]
      - Community initiative we need to engage the users



- **Title**: Very Short Summary of the Use Case
- **User Story**: "I got this task from my PI. I am supposed to read some data from X and write new data to Y using tool Z. The resulting data are further used by my local group, later on some skimmed results are fed into my experiments data management framework [DMF}"
- **Authz**: experiment token authz to read from experiments DMF, {{local authz to write site locally} or {Exp authz to write through DMF to site local} }
- **Addressing**: tool X uses Exp DMF addressing files/events for reading, write {{local authz} or {Exp authz DMF stack}}
- **Data Ingress/Egress/Processing Requirements**: Ingress: {reading files in streaming-like mode} or {reading file chunks event-like quasi-random access}, Egress: streaming events into files + logs
- **Compute Requirements**: simulation? merging?
- **Memory Requirements**: batch? interactive?
- **Software Example**: Foo
- **Admin Story**: "I have to provide an interface to make X addressable and accessible; I have to provide an interface to address for Y and to allocate space for Y; I have to install tool Z or provide a system to make it dynamically available. Data are stored locally and further reprocessed."

HL-LHC analysis workshops

- Analysis Ecosystem I (2017) ← Seeds
- WLCG/HSF in Adelaide (2019) Analysis systems
- HL-LHC Analysis mini workshop (2021)
- HL-LHC Analysis Ecosystem II (2022)
- WLCG/HSF in Norfolk (2023)
- CHEP ‘23 (2023)
- IRIS-HEP



VISION 25

WE ARE STILL DOING THE SAME THING (!?)

- Take a LEP-era physicist
  - would be comfortable with an LHC analysis – after being amazed about the growth of data, computing & complexity
- The growth in computing since the early 90-ies to late 00-ies has allowed us to be ~ conservative

I'VE TRAVELED HERE FROM THE YEAR 1983 TO SAY THIS:
ARE THERE ANY BAGELS LEFT?
WHILE IT'S TECHNICALLY TRUE, I WISH SHE'D STOP PREFACING EVERY SENTENCE WITH THAT.

G. Raven Analysis Ecosystem workshop 2017

HSF Analysis Facility Forum (Whitepaper)

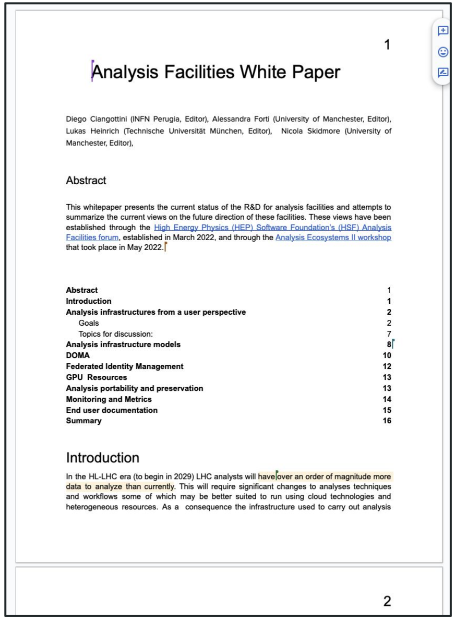
Forum started in March 2022 with Analysis Facility Kick-off [indico]

- roughly monthly meetings [last: April ‘23]

Mandate:

- Provide Forum to discuss efforts across the community
- Collect main ideas in a Whitepaper
  - Drafted by coordinators to provide basis for discussion
  - Goal: collect broad community views ➜ HSF authorship/endorsement
    HSF AE II Workshop report
- Paper is still a first draft and currently has 360 open comments+replies from people of different backgrounds
  - Debates in the comments will be summarized and become part of the paper
- The agenda of the Analysis Facility day at the WLCG/HSF workshop was built to reflect the paper
- The BoF on CERNbox (attended by 30 people) as a sharing service for AF in the morning was also a spin off of one of the HSF AF meetings (will need to follow up)

Link: [Google Docs]



1

Analysis Facilities White Paper

Diego Ciangottini (INFN Perugia, Editor), Alessandra Forti (University of Manchester, Editor), Lukas Heinrich (Technische Universität München, Editor), Nicola Skidmore (University of Manchester, Editor),

Abstract

This whitepaper presents the current status of the R&D for analysis facilities and attempts to summarize the current views on the future direction of these facilities. These views have been established through the High Energy Physics (HEP) Software Foundation's (HSF) Analysis Facilities forum, established in March 2022, and through the Analysis Ecosystems II workshop that took place in May 2022.

Introduction

In the HL-LHC era (to begin in 2029) LHC analysts will have over an order of magnitude more data to analyze than currently. This will require significant changes to analyses techniques and workflows some of which may be better suited to run using cloud technologies and heterogeneous resources. As a consequence the infrastructure used to carry out analysis

2

Analysis Facilities

Direction

- HL-LHC Analysis Facilities R&D is mainly focused on **extending the current models** toward a few set of open items:
  - Scaled up interactivity on more heterogeneous resources
  - Integration of cloud technologies (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

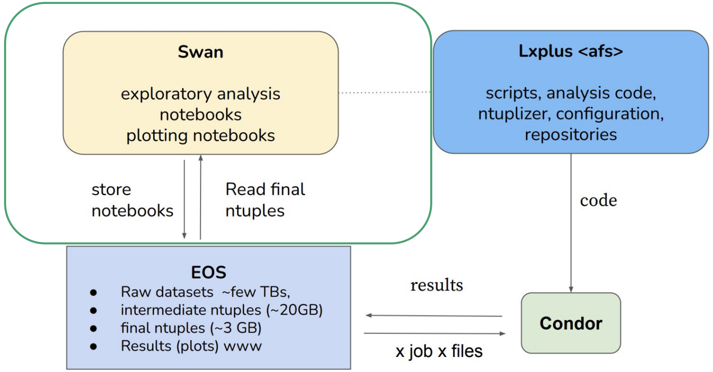
1. SSH login to a remote UI

- 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 lxplus)
- The **turnaround for an interactive analysis is going to be an important factor in the future**
  - Interactivity limited to 1 node on current facilities
  - Batch system are most of the time a backend of interactive resources

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.

2. AF R&Ds: Common traits

Extending HTC resources consumptions with interactive (web) UIs



Swan

exploratory analysis notebooks
plotting notebooks

Lxplus <afs>

scripts, analysis code, ntuplizer, configuration, repositories

store notebooks

Read final ntuples

code

EOS

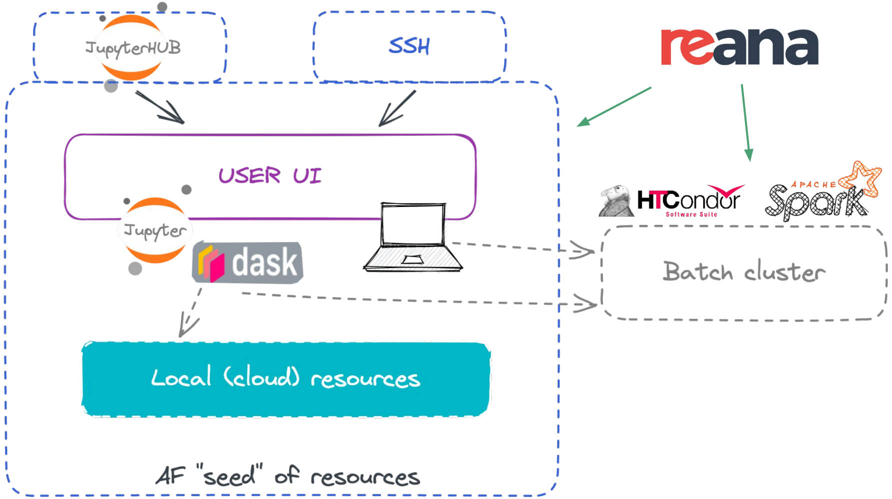
- Raw datasets ~few TBs,
- intermediate ntuples (~20GB)
- final ntuples (~3 GB)
- Results (plots) www

results

x job x files

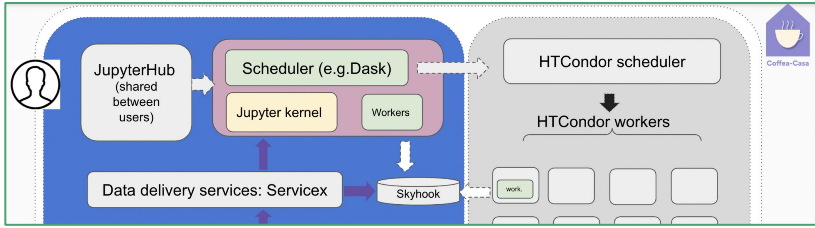
Condor

- Most of the presented R&Ds are focused on scaled-out python applications and RDataFrame
  - Led by the data analysis of reduced formats (e.g. NanoAOD/PHYSLITE) frameworks
- Containerized UIs
  - Supporting different levels of image building expertise
- 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

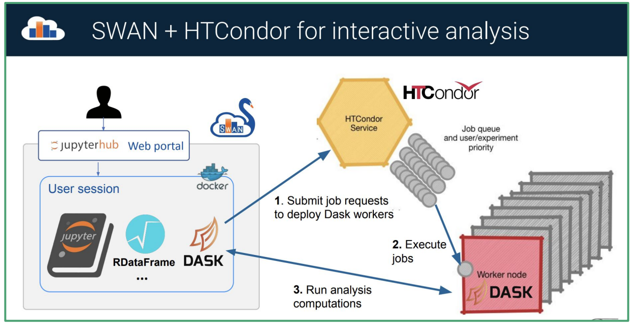


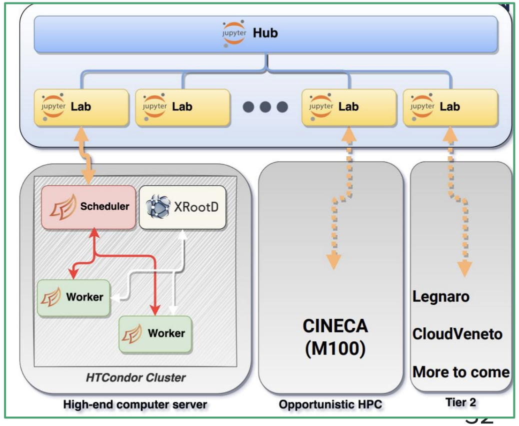
Let's see a summary of the presented initiatives ➜

The HUB abstraction R&D
There is a pattern…



- A growing 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 notebook execution.
- Main difference is in the integration pattern b/w the seed resources 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


SWAN + HTCondor for interactive analysis
jupyterhub Web portal
SWAN
User session
docker
jupyter
RDataFrame
...
DASK
HTCondor Service
HTCondor
Job queue and user/experiment priority
1. Submit job requests to deploy Dask workers
2. Execute jobs
3. Run analysis computations
Worker node
DASK



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

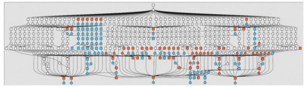
SWAN

Coffea Casa

INFN

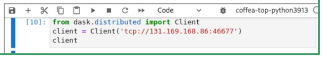
NAF

dask-jobqueue



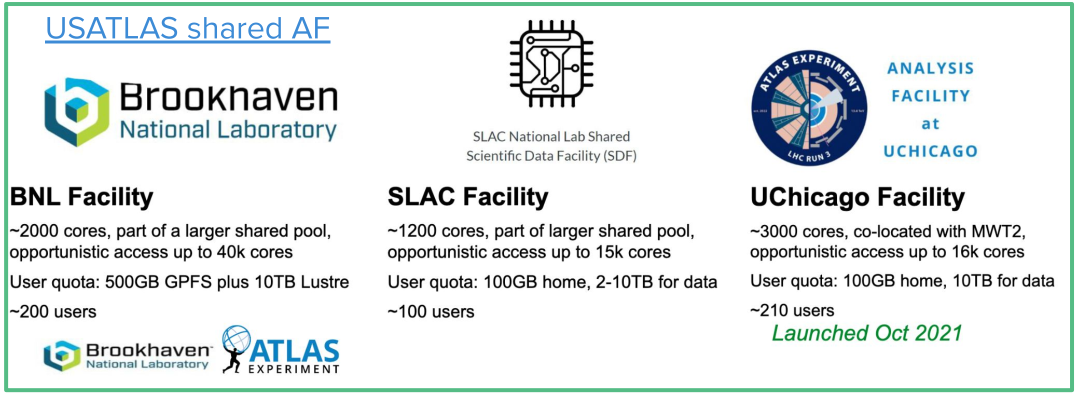
dask

- 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://131.169.168.86:46677')
      client
```

Other facilities presented at CHEP



USATLAS shared AF

Brookhaven National Laboratory

SLAC National Lab Shared Scientific Data Facility (SDF)

ANALYSIS FACILITY at UCHICAGO

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 Facility

~1200 cores, part of larger shared pool, opportunistic access up to 15k cores

User quota: 100GB home, 2-10TB for data

~100 users

UChicago Facility

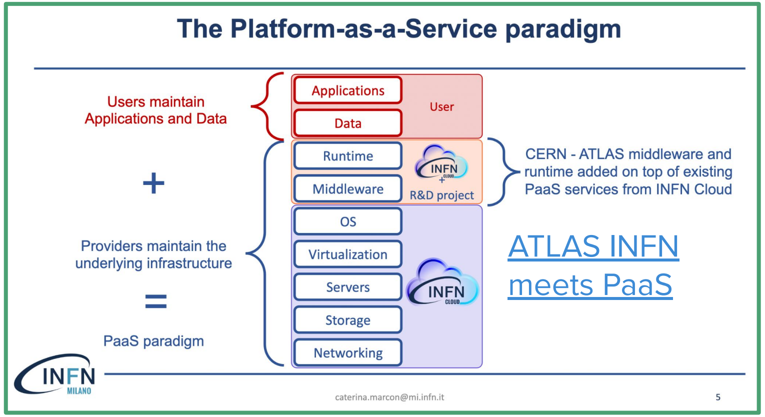
~3000 cores, co-located with MWT2, opportunistic access up to 16k cores

User quota: 100GB home, 10TB for data

~210 users

Launched Oct 2021

Brookhaven National Laboratory ATLAS EXPERIMENT



The Platform-as-a-Service paradigm

Users maintain Applications and Data

+

Providers maintain the underlying infrastructure

=

PaaS paradigm

Applications

Data

User

Runtime

Middleware

INFN CLOUD + R&D project

OS

Virtualization

Servers

Storage

Networking

INFN CLOUD

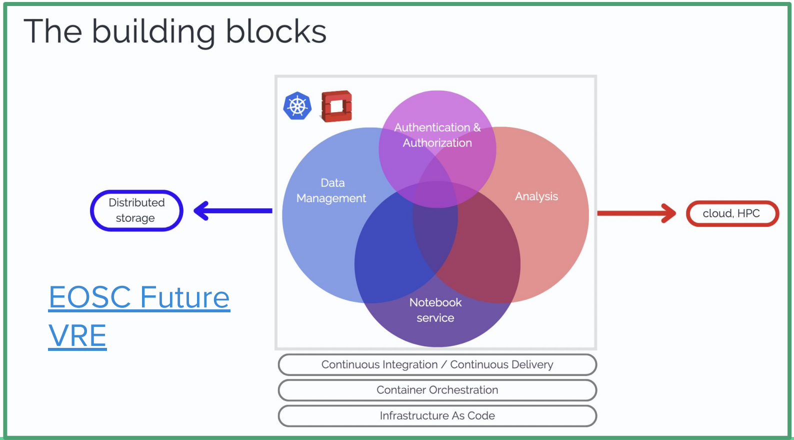
CERN - ATLAS middleware and runtime added on top of existing PaaS services from INFN Cloud

ATLAS INFN meets PaaS

INFN MILANO

caterina.marcon@mi.infn.it

5



The building blocks

Distributed storage

Data Management

Authentication & Authorization

Analysis

Notebook service

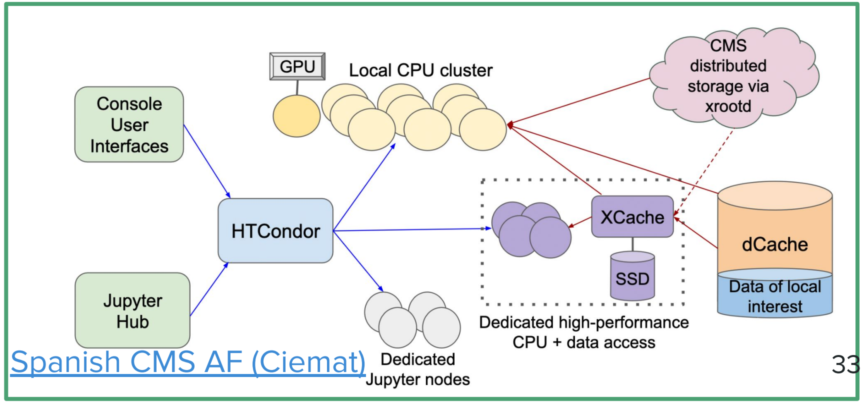
cloud, HPC

Continuous Integration / Continuous Delivery

Container Orchestration

Infrastructure As Code

EOSC Future VRE



Console User Interfaces

Jupyter Hub

HTCondor

GPU

Local CPU cluster

CMS distributed storage via xrootd

XCache

SSD

dCache

Data of local interest

Dedicated high-performance CPU + data access

Dedicated Jupyter nodes

Spanish CMS AF (Ciemat)

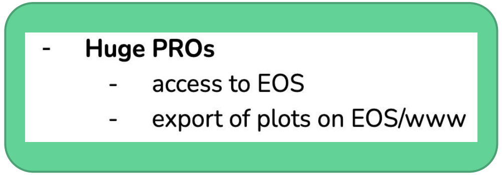
Storage and AAI

DOMA: Input Data organisation and access

- Data flows for HL-LHC analysis are **still being defined**
    - 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 analyses will be able to use PHYSLITE/nanoAOD**
    - AF should/could/would/? support all workflows
- **Latency** is a factor for input data for intense workflows
    - Usually reduced 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
    - Many analyses have a highly repetitive stage, when the cache hit rate is high particularly in the development stages when ideas are tested so once the data is cached it is reused
- Some dedicated AFs can still host complete datasets locally, but caches enable smaller centres to remain viable and useful

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
  - But also NAF users expressed the same regarding NAF storage
- If we want to express this in terms of functionalities
  - POSIX-like 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...)



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

Swan user CERN

DOMA: Shared Storage problems and evolution (?)

- 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 is how to federate instances and data ownership
  - Mapping of general authorization with linux ACLs at different AFs (CERNbox BoF discussion)
  - Could this be solved by Auth2.0? (next slide)
- Object Stores vs POSIX is a recurrent discussion
  - Users like POSIX-like storage because the applications and the interactive login both work
  - More and more sites are installing OS integration in the applications will make possible user interaction
    - Xrootd-S3 (W. Yang CHEP2023), RDataFrame-S3 (G. Lazzari CHEP2023)
  - A notebook infrastructure may handle more easily

Federated Identity

- 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 (T. Dack CHEP2023)
- Integration with cloud technologies being proposed for AF can be 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

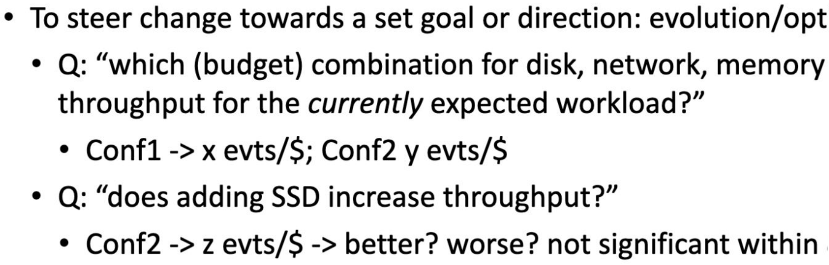
coffea-casa user Nebraska

- coffea works with tokens because the services between the user and the data use tokens
  - Classic AF (ssh+batch+POSIX FS) will need development work (F. Fornari CHEP2023)
  - Object stores would work better for this too.

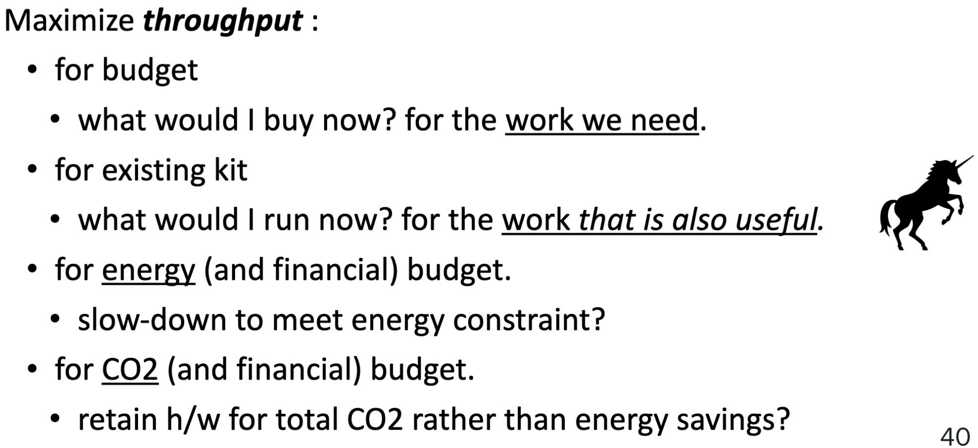
Monitoring and metrics

Metrics and monitoring

- Important points about the trade offs of choosing metrics respect to competing target.
  - To really decide what to do about the metrics particularly for funding probably need a focus group



- To steer change towards a set goal or direction: evolution/optimisation
  - Q: “which (budget) combination for disk, network, memory and CPU investment maximizes throughput for the *currently* expected workload?”
    - Conf1 -> x evts/$; Conf2 y evts/$
  - Q: “does adding SSD increase throughput?”
    - Conf2 -> z evts/$ -> better? worse? not significant within



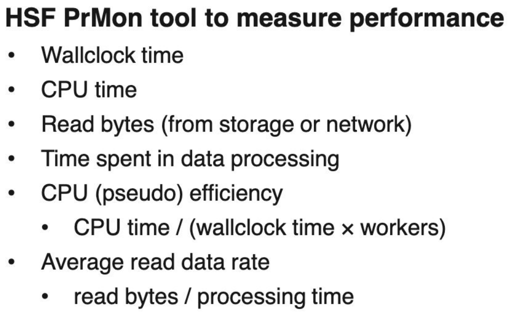
Maximize ***throughput*** :

- for budget
  - what would I buy now? for the work we need.
- for existing kit
  - what would I run now? for the work *that is also useful*.
- for energy (and financial) budget.
  - slow-down to meet energy constraint?
- for CO2 (and financial) budget.
  - retain h/w for total CO2 rather than energy savings?

D. Duelmann WLCG/HSF WS

Instrumenting user jobs

- *A lot is analysis dependent and we need to understand the situation quantitatively as it will change how much evolution is required. ML situation is unexplored - if analysis area is main driver for GPUs we need to know as we cannot leverage these resources from other areas (S. Campana)*
- We need ways of actually measure performance
  - Applications are diverse
  - prMon mentioned in 5 presentations
    - Has also some basic GPU monitoring



HSF PrMon tool to measure performance

- Wallclock time
- CPU time
- Read bytes (from storage or network)
- Time spent in data processing
- CPU (pseudo) efficiency
  - CPU time / (wallclock time × workers)
- Average read data rate
  - read bytes / processing time

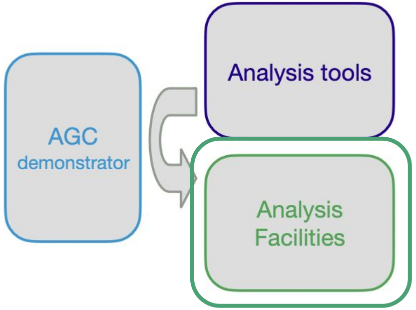
IRIS-HEP Analysis Grand Challenge

- AGC designed to measure new techniques and new services
  - Pre-prepared workflows working on OpenData
- It is a good tool to use to understand if the implementation/deployment of specific analysis tools, workflows, techniques

- The AGC has **two aspects**

1. define a physics analysis task of realistic scope & scale
2. develop analysis pipelines that implements the task

   - find & address performance bottlenecks & usability concerns

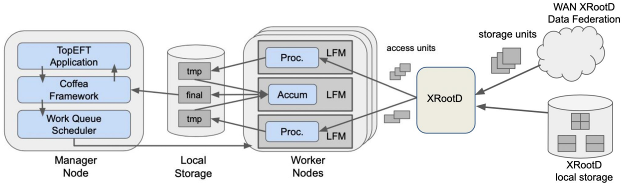


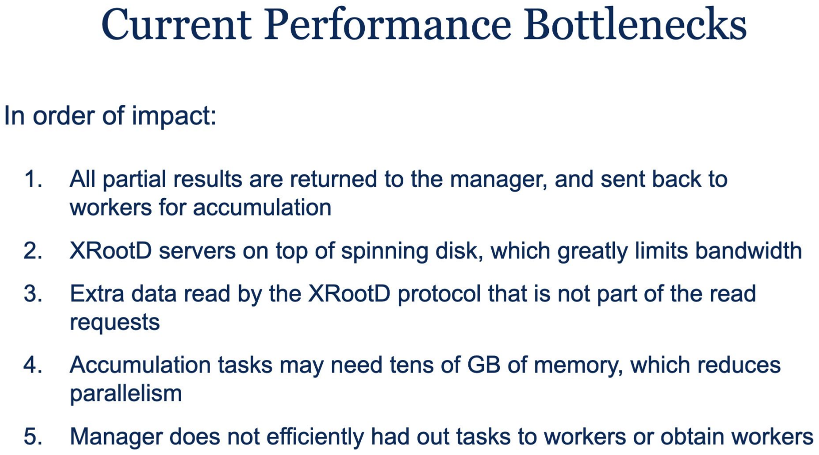
A Held CHEP 2023

- Can also be used to test AF setup

AGC to test AF setup: Notre Dame

- Goal: 10 mins analysis ➜ Achieved 100 mins
    - Identified bottlenecks
    - List of possible improvements to the infrastructure

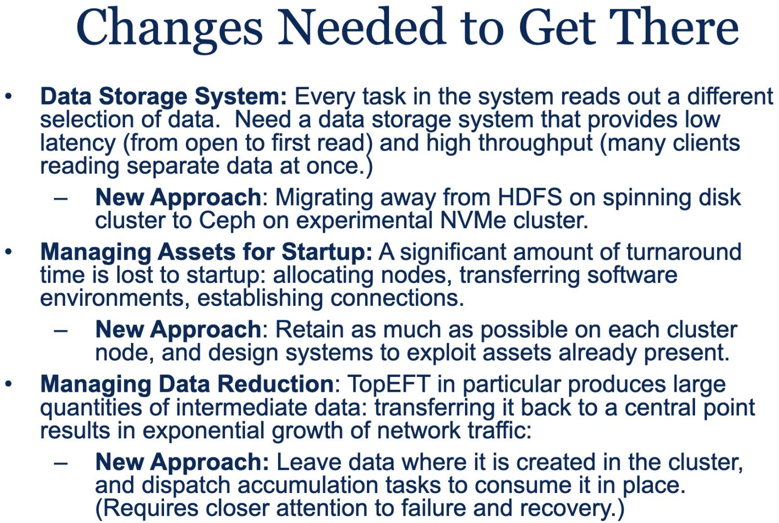




Current Performance Bottlenecks

In order of impact:

1. All partial results are returned to the manager, and sent back to workers for accumulation
2. XRootD servers on top of spinning disk, which greatly limits bandwidth
3. Extra data read by the XRootD protocol that is not part of the read requests
4. Accumulation tasks may need tens of GB of memory, which reduces parallelism
5. Manager does not efficiently had out tasks to workers or obtain workers



Changes Needed to Get There

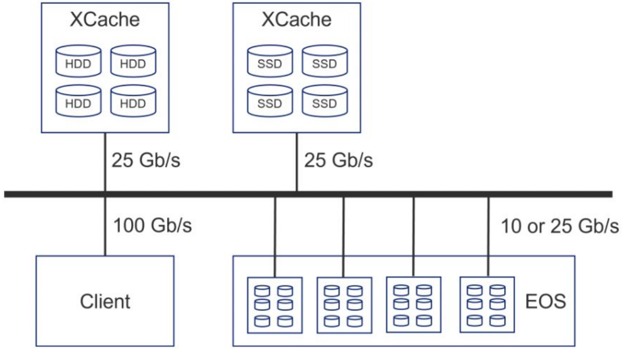
- **Data Storage System:** Every task in the system reads out a different selection of data. Need a data storage system that provides low latency (from open to first read) and high throughput (many clients reading separate data at once.)
    - **New Approach**: Migrating away from HDFS on spinning disk cluster to Ceph on experimental NVMe cluster.
- **Managing Assets for Startup:** A significant amount of turnaround time is lost to startup: allocating nodes, transferring software environments, establishing connections.
    - **New Approach**: Retain as much as possible on each cluster node, and design systems to exploit assets already present.
- **Managing Data Reduction**: TopEFT in particular produces large quantities of intermediate data: transferring it back to a central point results in exponential growth of network traffic:
    - **New Approach:** Leave data where it is created in the cluster, and dispatch accumulation tasks to consume it in place. (Requires closer attention to failure and recovery.)

J Lawrence CHEP 2023

AGC to test AF setup: CERN I/O performance studies

- Goal is to collect analysis workloads and tools to measure I/O performance in different storage configurations and levels of parallelism
  - The final objective is to optimize resource allocation

- **Several workloads available now or soon**
  - ROOT’s *rootreadspeed* I/O benchmark
  - ROOT’s RDataFrame benchmark
  - IRIS-HEP Analysis Grand Challenge, both Coffea and RDataFrame implementations
  - A real CMS analysis using Coffea by A. Novak
  - A real CMS analysis using RDF by T. Tedeschi



XCache

HDD HDD HDD HDD

XCache

SSD SSD SSD SSD

25 Gb/s

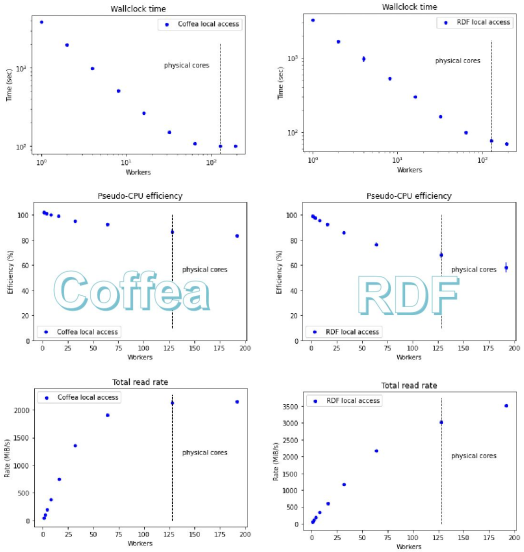
25 Gb/s

100 Gb/s

10 or 25 Gb/s

Client

EOS

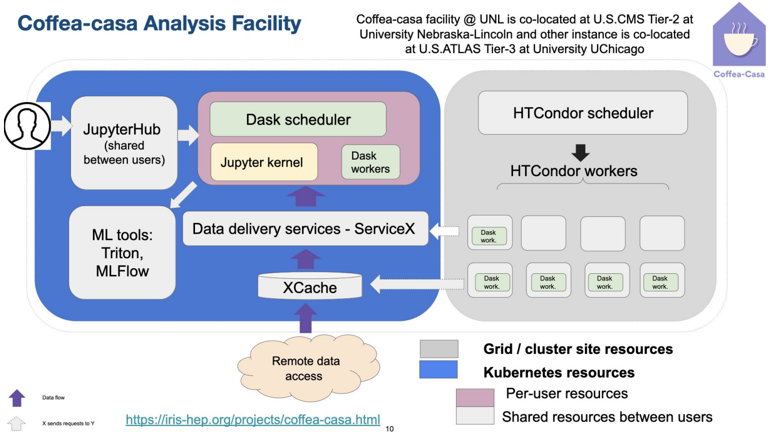


A. Sciabà CHEP 2023

Experiments

CMS

- CMS is fully engaged in this activity
  - (O. Gutsche CHEP 2023)
- National CMS communities are investigating AFs on dedicated resources.
- They already have a widely adopted reduced format
    - Almost 5 years of experience (A. Rizzi, S&C roundtable 2021)
- CMS OpenData is facilitating the AGC using “CMS data”



Coffea-casa Analysis Facility

Coffea-casa facility @ UNL is co-located at U.S.CMS Tier-2 at University Nebraska-Lincoln and other instance is co-located at U.S.ATLAS Tier-3 at University UChicago

O. Shadura CHEP 2023

ATLAS

- Similar Analysis model are expressed in the CDR and in the more recent Roadmap to HL-LHC document
- There is an AGC based demonstrator for the TDR (not public yet)
- ATLAS doesn’t have large scale OpenData for this exercises (yet)
- PHYSLITE first real production this year and physics groups are asked to give it a try
  - Working on columnar analysis and systematics on the fly (N. Krumnack CHEP 2023)

Run4 PHYSLITE: ~2.6 PB/y (data+mc)

J.Elmsheuser CHEP2019

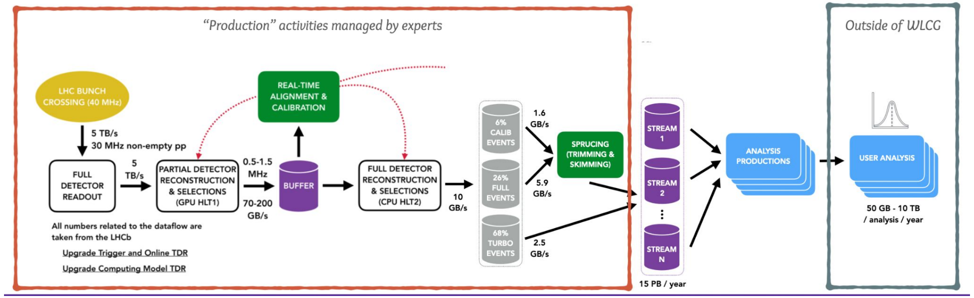
DAOD_PHYSLITE
One format
Written for all events
10-12KB/event
xAOD EDM
Targets ~80% of analyses

Can be directly analysed without need for n-tuples

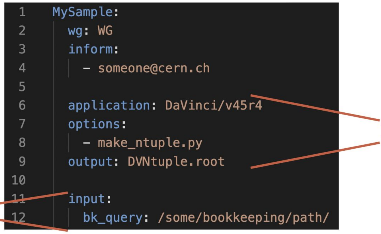
J. Schaarschmidt CHEP2023

LHCb: AF for Ntuple production

- The Analysis Productions infrastructure allows a user-friendly, declarative approach to ntupling
    - Historically analysts were responsible for running O(10,000) grid jobs to produce ROOT files
    - Centralised production ensures e.g. better validation hence more efficient use of resources
    - Integration of testing and monitoring using gitlab CI/CD - web based monitoring



"Production" activities managed by experts
Outside of WLCG
LHC BUNCH CROSSING (40 MHz)
5 TB/s
30 MHz non-empty pp
FULL DETECTOR READOUT
5 TB/s
PARTIAL DETECTOR RECONSTRUCTION & SELECTIONS (GPU HLT1)
0.5-1.5 MHz
70-200 GB/s
REAL-TIME ALIGNMENT & CALIBRATION
BUFFER
FULL DETECTOR RECONSTRUCTION & SELECTIONS (CPU HLT2)
10 GB/s
6% CALIB EVENTS
26% FULL EVENTS
68% TURBO EVENTS
1.6 GB/s
5.9 GB/s
2.5 GB/s
SPRUCING (TRIMMING & SKIMMING)
STREAM 1
STREAM 2
STREAM N
15 PB / year
ANALYSIS PRODUCTIONS
USER ANALYSIS
50 GB - 10 TB / analysis / year
All numbers related to the dataflow are taken from the LHCb
Upgrade Trigger and Online TDR
Upgrade Computing Model TDR

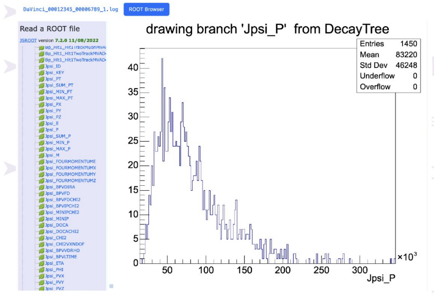


```yaml
MySample:
  wg: WG
  inform:
    - someone@cern.ch

  application: DaVinci/v45r4
  options:
    - make_ntuple.py
  output: DVNtuple.root

  input:
    bk_query: /some/bookkeeping/path/
```

1. What to do

2. What input data

Also support chaining productions, e.g.
Run HLT1 then
use as input for HLT2 then
as input for ntupling



DaVinci_00012345_00006789_1.log

Read a ROOT file

JSROOT version 7.2.0 11/08/2022

Bp_Hlt1_Hlt1TrackMuonMVA
Bp_Hlt1_Hlt1TwoTrackMVAD
Bp_Hlt1_Hlt1TwoTrackMVAD
Jpsi_ID
Jpsi_KEY
Jpsi_PT
Jpsi_SUM_PT
Jpsi_MIN_PT
Jpsi_MAX_PT
Jpsi_PX
Jpsi_PY
Jpsi_PZ
Jpsi_E
Jpsi_P
Jpsi_SUM_P
Jpsi_MIN_P
Jpsi_MAX_P
Jpsi_M
Jpsi_FOURMOMENTUME
Jpsi_FOURMOMENTUMX
Jpsi_FOURMOMENTUMY
Jpsi_FOURMOMENTUMZ
Jpsi_BPVDIRA
Jpsi_BPVFD
Jpsi_BPVFDCHI2
Jpsi_BPVIPCHI2
Jpsi_MINIPCHI2
Jpsi_MINIP
Jpsi_DOCA
Jpsi_DOCACHI2
Jpsi_CHI2
Jpsi_CHI2VXNDOF
Jpsi_BPVVDRHO
Jpsi_BPVLTIME
Jpsi_ETA
Jpsi_PHI
Jpsi_PVX
Jpsi_PVY
Jpsi_PVZ

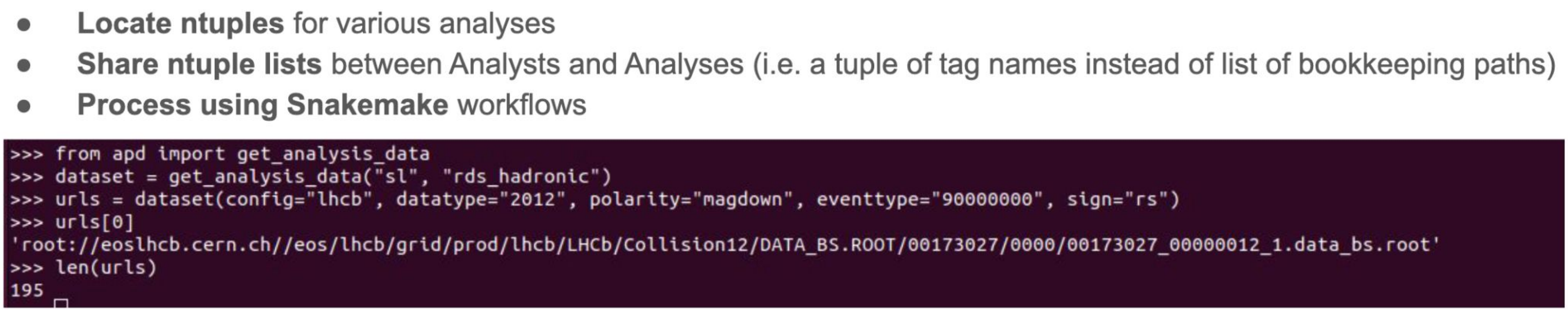
drawing branch 'Jpsi_P' from DecayTree

Entries	1450
Mean	83220
Std Dev	46248
Underflow	0
Overflow	0

[C.Burr CHEP 2023](#)

LHCb: Analysis data

- Analysis productions data (apd) tool
  - Integration of JWT/EOS tokens to access data
  - Provides provenance between the grid and local worlds
  - Simple interface which can provide local caching, authentication, long-term reproducibility
  - Well suited to analysis facilities
- Snakemake
  - Tracking of analysis artefacts and tagging of data to allow for reproducibility



- **Locate ntuples** for various analyses
- **Share ntuple lists** between Analysts and Analyses (i.e. a tuple of tag names instead of list of bookkeeping paths)
- **Process using Snakemake** workflows

```
>>> from apd import get_analysis_data
>>> dataset = get_analysis_data("sl", "rds_hadronic")
>>> urls = dataset(config="lhcb", datatype="2012", polarity="magdown", eventtype="90000000", sign="rs")
>>> urls[0]
'root://eoslhcb.cern.ch//eos/lhcb/grid/prod/lhcb/LHCb/Collision12/DATA_BS.ROOT/00173027/0000/00173027_00000012_1.data_bs.root'
>>> len(urls)
195
```

B. Coutourier CHEP 2023

ALICE O2: analysis framework in Run 3

- **Runs 1 and 2**: C++ object-based analysis framework (arrays of structures)
- **Run 3+**: continuous readout and higher interaction rates ➔ revision required!

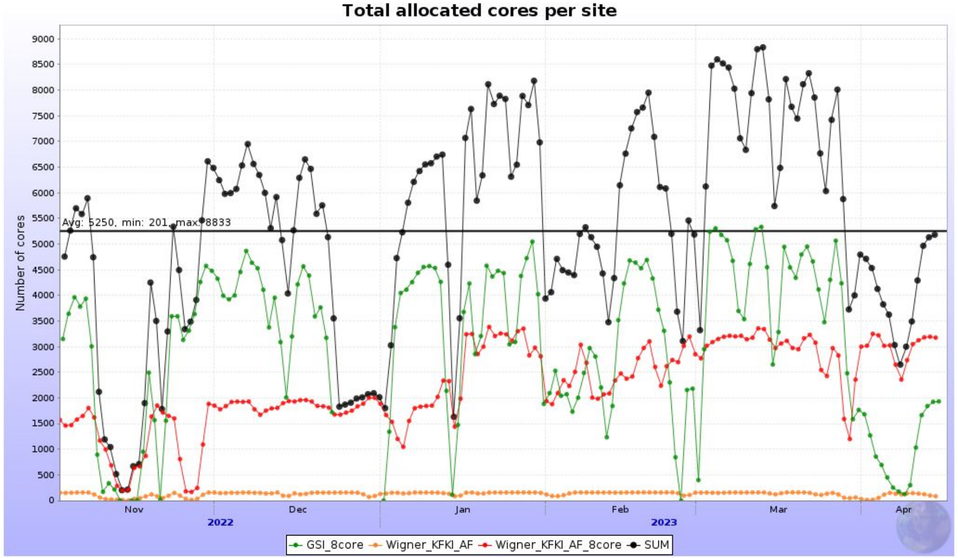
- New ALICE analysis framework:
  - **Revised data model** using structures of arrays: flat memory space for fast operations
  - **Declarative programming** for bulk of data processing
  - **Highly scalable**: multiprocessing capable
  - **Multiple target architectures**: CPU and GPU technically viable

- Successfully deployed and in use!
  - **≥10x faster per event** compared to Run 1/2 framework
  - Large-scale running (multi-petabyte-scale) successfully on the grid via **'hyperloop' analysis train system**
  - Very flexible data format allows for smaller subsets of data ('skimming' and software triggering)

- Final optimizations well underway
  - Necessary for continued analysis of all Run 3+ data

ALICE Analysis Facilities in Run 3

- ALICE Run3 computing model foresees AFs to achieve a fast turnaround cycle for analysis validation and cut optimisation before submitting **trains** to the GRID for full statistics analysis
- Goal: **fast turnaround** for cut tuning and task validation
  - About 10% of AO2D will be made available
  - 2023: serve 7,500 job slots at aggregate throughput of 100 GB/s to digest 4 PB of Run 3 data in 12h
  - Planned growth to 20K job slots to process 8-10 PB/day in 2026



GSI Analysis Facility (Darmstadt, Germany)

- 2023 pledges: 63 kHS06 (~6.0K job slots) and 6.1 PB
- Worker nodes have direct mount of the cluster file system Lustre
- Optimised I/O throughput: custom XRootD Plugin provides direct access to data (expected 115 GB/s)

Wigner Analysis Facility (Budapest, Hungary)

- 29 kHS06 (3.6K job slots) and ~1.1PB storage EOS FS
- Legacy hardware from the decommissioned Wigner Tier 0
- Electricity and operations (2 FTE) provided by Wigner Research
- Centre (in addition to pledged grid resources)