AF update

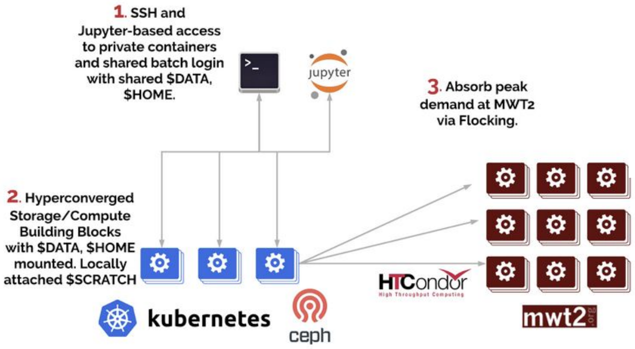
A. Forti

WLCG workshop

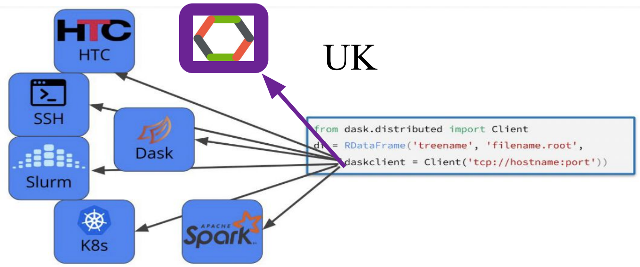
9 November 2022

Analysis Facilities

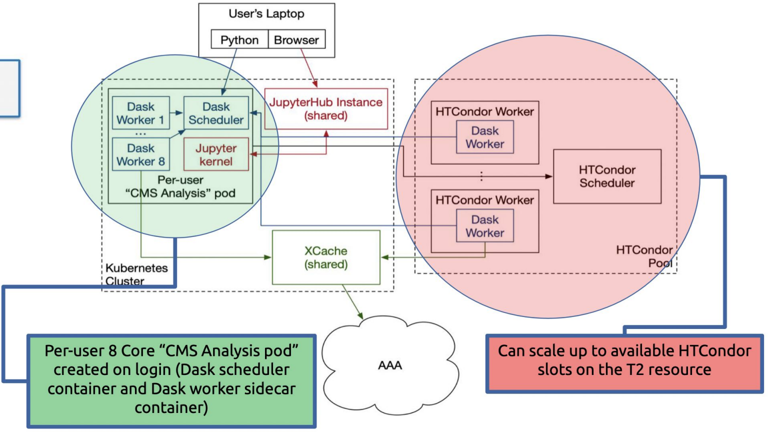
UChicago


1. SSH and Jupyter-based access to private containers and shared batch login with shared $DATA, $HOME.
jupyter
2. Hyperconverged Storage/Compute Building Blocks with $DATA, $HOME mounted. Locally attached $SCRATCH
3. Absorb peak demand at MWT2 via Flocking.
kubernetes
ceph
HTCondor High Throughput Computing
mwt2

UK



```
from dask.distributed import Client
df = RDataFrame('treename', 'filename.root',
    daskclient = Client('tcp://hostname:port'))
```

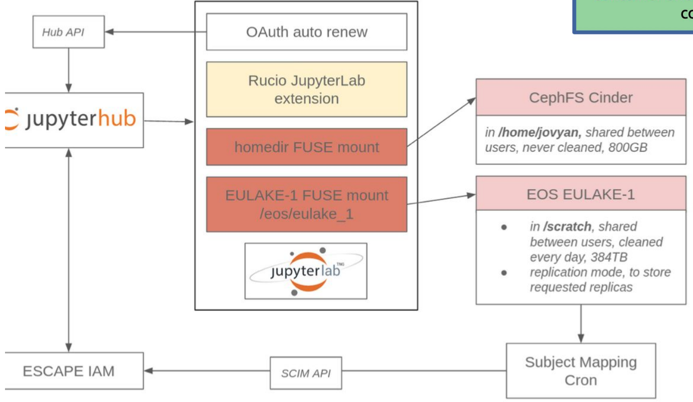


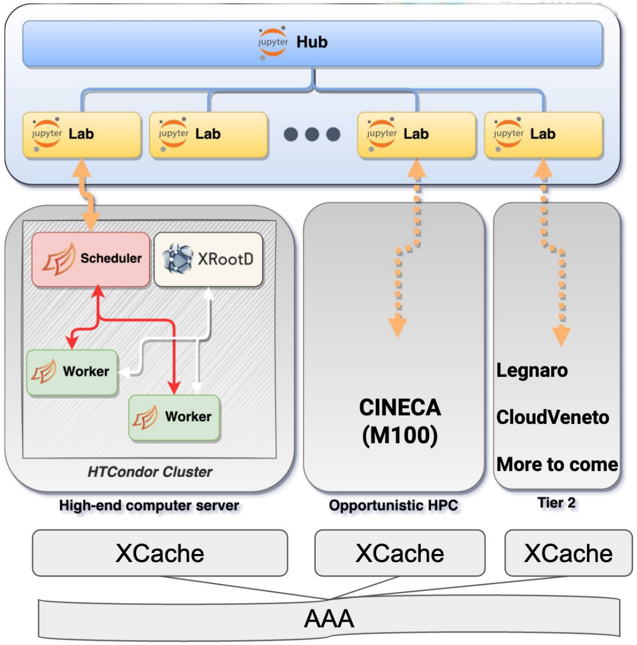
Per-user 8 Core "CMS Analysis pod" created on login (Dask scheduler container and Dask worker sidecar container)

Can scale up to available HTCondor slots on the T2 resource

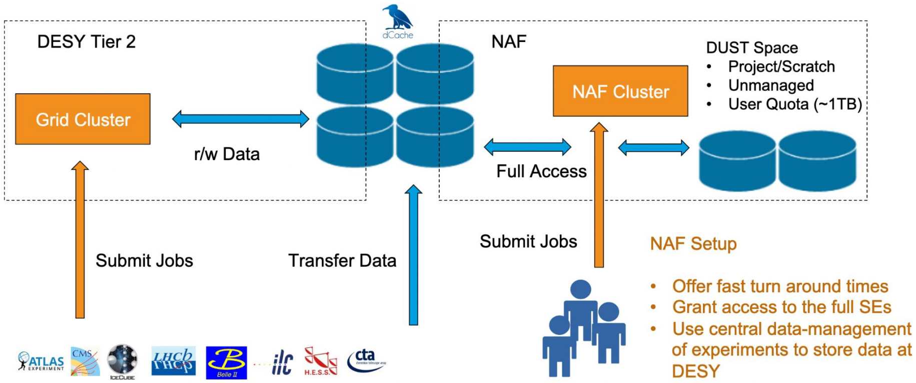
Coffea

ESCAPE DLaaS


Hub API
OAuth auto renew
Rucio JupyterLab extension
homedir FUSE mount
EULAKE-1 FUSE mount /eos/eulake_1
jupyterlab
jupyterhub
CephFS Cinder
in /home/jovyan, shared between users, never cleaned, 800GB
EOS EULAKE-1
in /scratch, shared between users, cleaned every day, 384TB
replication mode, to store requested replicas
Subject Mapping Cron
SCIM API
ESCAPE IAM



INFN



NAF Setup

- Offer fast turn around times
- Grant access to the full SEs
- Use central data-management of experiments to store data at DESY

NAF

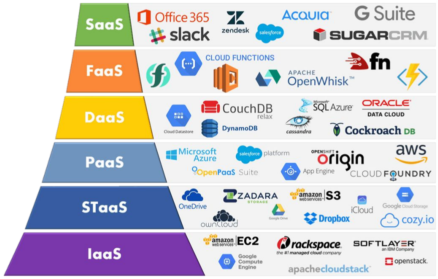
Analysis Facilities

- AF definition in HSF is still
  - “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.”
- Whatever shape an AF takes the important is being able to integrate the technologies to support the analysis evolution.
  - Some of thee technologies might be adopted by T3s like CVMFS in the past

Technology Evolution

kubernetes

SCI TOKENS



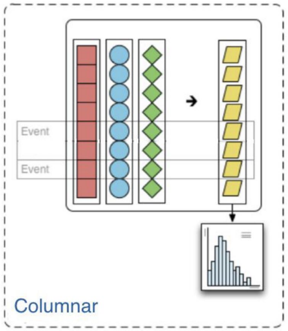
ServiceX

XCache

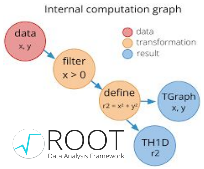
Techniques Evolution

DASK

jupyter



Scikit HEP



Topics

- Out of the Analysis Ecosystem workshop report and other places the need to concentrate on building blocks rather than specific architectures
- Some topics
  - AAI
  - DOMA topics
    - Grid storage access and xcaches
    - Shared storage
    - Object stores
  - Environment sharing: containers
  - Tracking analysis performance
  - Resource sharing and services deployment (k8s)
  - IRIS-HEP AGC as a concrete goal to pull various threads
- Need to discuss today HOW to follow this up
  - Preference would be to have groups of people willing to put sometime for each topic to follow also in other forum and then create a coherent picture from AF point of view
  - Some topics needs tight cooperation with other projects WLCG/IRIS-HEP/EOSC/… or HSF groups

AAI

- Technologies current analysis facilities are exploring are better suited to tokens than x509
  - AF R&D and more flexible and adaptable to change
  - AF users are testers and are more adaptable
- AF better built around this from the beginning
- Token infrastructure not production ready but things can be followed in WLCG authz and DOMA BDT
- The question of integrating AF R&D in the testbed needs proactive engagement with above people
- Separate discussion should be held about the federated identities and access policies.
  - Final decisions on this depend on funding and experiments models and should be independent from creating the token infrastructure

Priority (tokens): could help the current effort !!

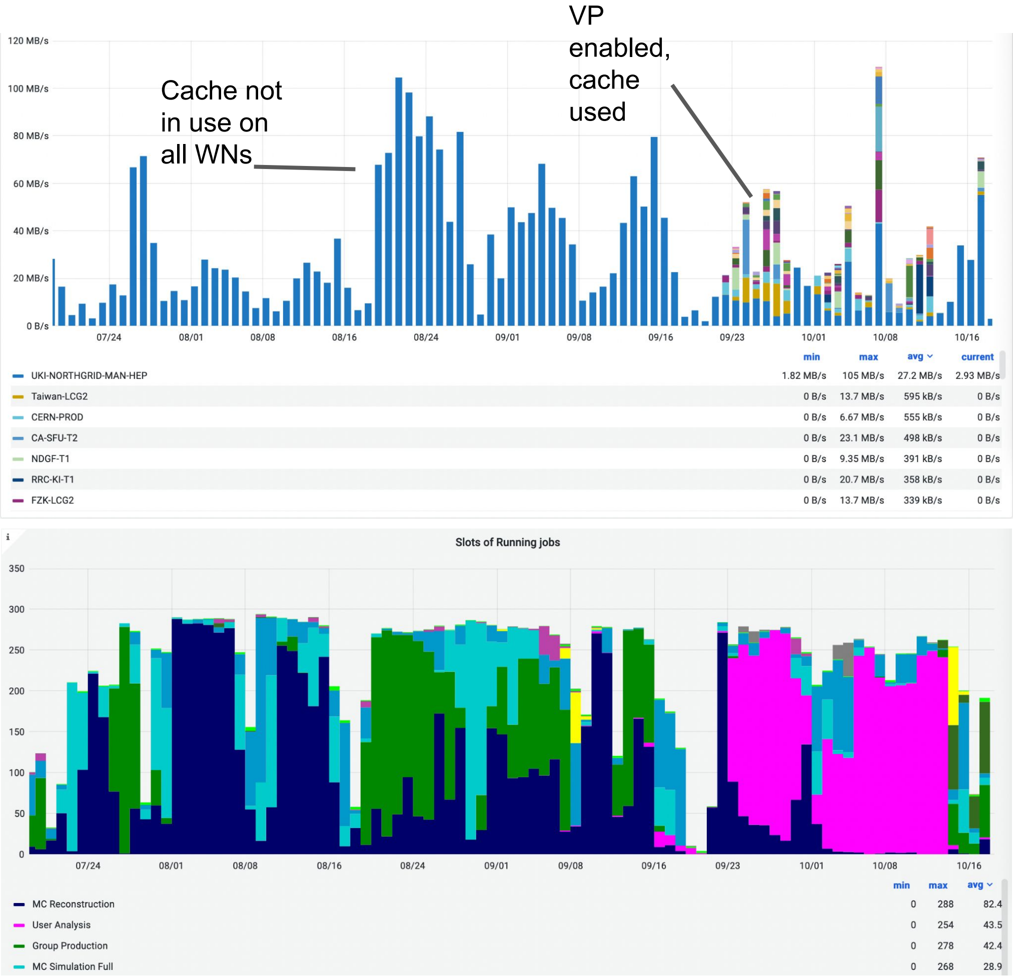
DOMA (1)
Bulk Storage And XCaches

- Accessing, integrating and caching input from the grid bulk storage (aka Data Lake)
  - Many ideas already in DOMA Access
  - Ongoing testing in different places
  - In ATLAS use of XCaches at AF and storageless sites
    - Integration with rucio using Virtual Placement
- Hardware specs
  - Range of cache sizes and performance but there is no recommendation on the actual specs
- Xcaches monitored w/ custom scripts, there is no agreed solution
    - Like for Xrootd streamed data most problematic
    - WLCG xrootd monitoring is a way forward towards consolidating also xcaches monitoring but need experiments agreement
    - Other sciences might be interested

Priority: understanding hardware and having monitoring !!!

XCache in UK example

- Bham ~300 cores
  - XCache points to Mcr
  - VP allows more than one source
- Before mid September XCache was not enabled on all the WNs and part of the traffic is direct access job -> Manchester SE
  - 800 Mb/s peaks are large-ish for 300 cores
- After mid September there is a clear correlation between enabling XCache+VP and the amount of Analysis jobs.



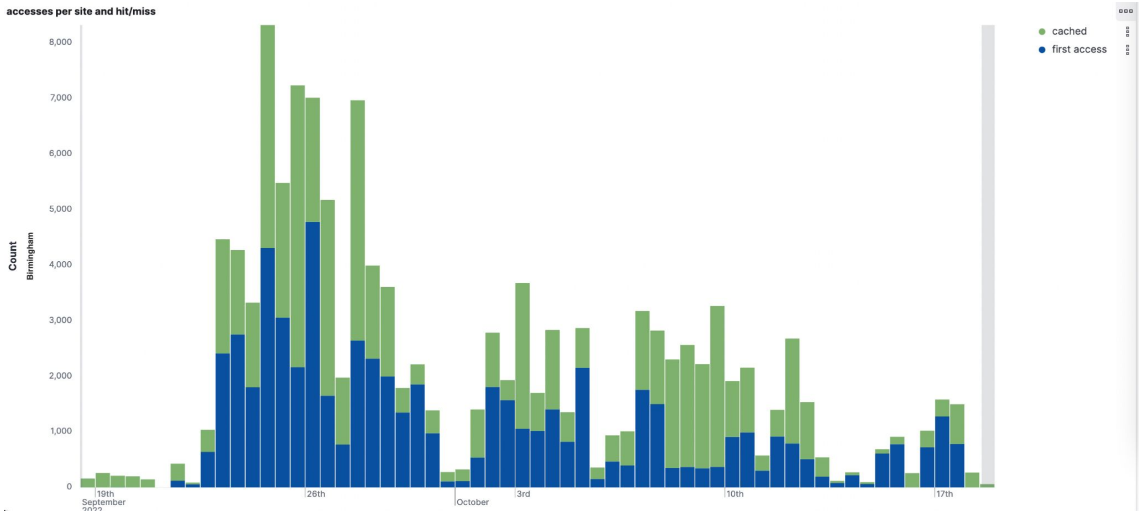
	min	max	avg	current
UKI-NORTHGRID-MAN-HEP	1.82 MB/s	105 MB/s	27.2 MB/s	2.93 MB/s
Taiwan-LCG2	0 B/s	13.7 MB/s	595 kB/s	0 B/s
CERN-PROD	0 B/s	6.67 MB/s	555 kB/s	0 B/s
CA-SFU-T2	0 B/s	23.1 MB/s	498 kB/s	0 B/s
NDGF-T1	0 B/s	9.35 MB/s	391 kB/s	0 B/s
RRC-KI-T1	0 B/s	20.7 MB/s	358 kB/s	0 B/s
FZK-LCG2	0 B/s	13.7 MB/s	339 kB/s	0 B/s

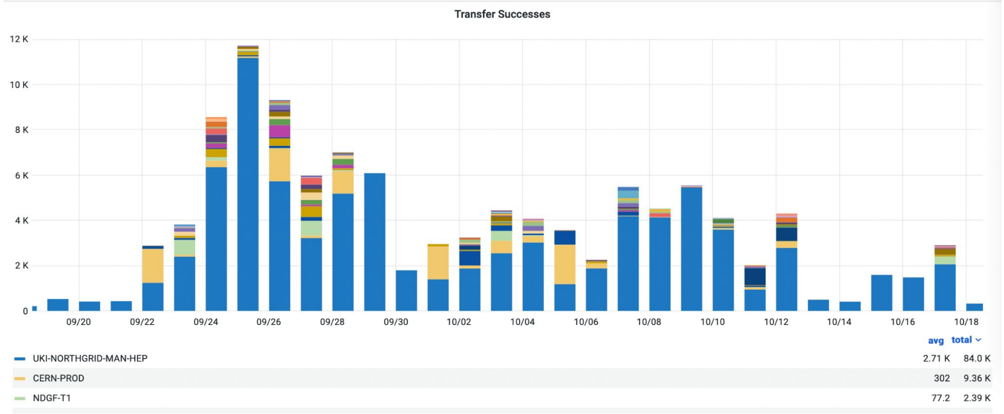
	min	max	avg
MC Reconstruction	0	288	82.4
User Analysis	0	254	43.5
Group Production	0	278	42.4
MC Simulation Full	0	268	28.9

Last 3 months

XCache in UK example

- A good 35-40% of the data is accessed from the cache
  - The rest is still accessed via XCache (buffer cache in = out)
- Monitoring might not be universal needs to be looked at.
  - For streamed analysis differences with monit.





	avg	total
UKI-NORTHGRID-MAN-HEP	2.71 K	84.0 K
CERN-PROD	302	9.36 K
NDGF-T1	77.2	2.39 K

Atlas data Last 1 month

Doma(2)
Shared Storage

- Recurring topic: Local shared storage for people to seamlessly run from different resources and share with colleagues
  - Discussed during the SWAN/ScienceBox, EOSC and the INFN multi-site AF presentations.
    - First two use EOS at CERN, the latter has a local cephFS storage
  - Users repeatedly report it as a main feature at CERN
- As soon as the AF is distributed this becomes problematic
- If access is from remote or with xrootd gateways it's still not as straightforward and can cause problems to the site
  - Typical example users want to access/mount EOS
- For different solutions monitoring of traffic and access needs to be harmonised
- Needs discussion and recommendations

Priority: any user strong requirement !!!

Doma(3) Object Stores

- Another topic covered is the possibility to use object stores
  - Several grid sites installed ceph with cephFS for POSIX access
  - Object stores advantage is scalability because they don't need any metadata db (true or false?)
- AF workflows may have different requirements: which?
  - IRIS-HEP is introducing object stores in their testing
  - ServiceX backend storage is an object store recently moved to use S3
    - Users don't access S3 directly
- Evaluating object stores without posix access for standard experiment software is a large body of work
  - There is a CMS R&D proposal
  - Experiments requirements on this need to reviewed and initial recommendations if it is worth pursuing written

Priority: !

Environment sharing

- Users want to share with colleagues their setup,code, configuration, small amount of input data….
  - Shared storage is the traditional way but not the only way
    - Conda (LHCb), Containers (CMS, ATLAS)
- Solutions easy to setup and should help also preservation
- Containers
  - We still don't have an official way to distribute images or a supported registry
  - Need to look at building and supporting base images
  - CVMFS unpacked.cern.ch has now 3000 images
    - Pulls images from any public register
    - Uses harbor.cern.ch as a proxy to sanitize images
  - Users using harbor directly could open various development roads
    - Images life cycle management
    - Possibly solving private images problem at least for containerd

Priority: demand is growing but scale not yet clear !

Tracking Analysis Performance

- Need an agreed upon list of metrics
  - 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
  - ……
- Need also an infrastructure where to make them accessible
  - Centralised monitoring may take a long time to develop
  - Some request to instrument jobs like ATLAS does on the grid
- Other types of site monitoring in the same situation (networking, tape, xrootd, benchmarking…)
  - May draw from that experience
  - Some of the metrics may as well be the same

Priority: monitoring has always high priority !!!

Shared Resources and kubernetes

- Recommendation is to colocate AFs with existing sites to maximise resource sharing
- k8s is the proposed method of deployment for services
    - As well as an alternative backend to run users jobs
- It has been embraced by a number of large sites
    - But not everyone agrees it will solve problems
- In WLCG it is recurringly talked about
    - Last time at June pre-GDB (agreed another pre-GDB mid next year)
    - Ryan's talk later
- There aren't recommendations on deployment or for k8s service developers and users
    - Short document for app devs to get things to work both on okd and vanilla k8s
- CERN has a really robust documentation but it looks CERN specific
- So there is still no coordinated effort for this
    - Ricardo Rocha gave a comprehensive list of forums people can attend also report from kubecon on batch system features development

Priority: in US it is a requirement in Europe up to the sites !!! !! !

IRIS-HEP AGC

- AGC status at the last week
- AGC designed to measure new techniques and new services
- Official challenge with Dask and OpenData
- RDataFrame and non public data are ✓
- Non IRIS-HEP sites can also participate
  - INFN and UK plan to participate with their infrastructures and workflows (even non-LHC)
  - Monthly ops meetings (next tomorrow)
  - analysis-grand-challenge google group for announcements and discussion
- Need to couple sites and users to be more productive

Priority: !!

HSF AF Forum

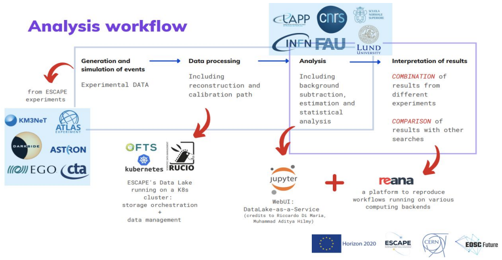
HSF Analysis Facilities Forum Kick-off Meeting

Mark Neubauer

University of Illinois at Urbana-Champaign

March 25, 2022

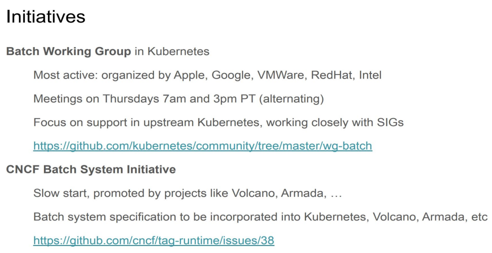
Kick-off meeting



Analysis workflow

Generation and simulation of events
Experimental DATA
from ESCAPE experiments
Data processing
Including reconstruction and calibration path
Analysis
Including background subtraction, estimation and statistical analysis
Interpretation of results
COMBINATION of results from different experiments
COMPARISON of results with other searches
KM3NeT
ATLAS
DARKSIDE
ASTRON
EGO
cta
FTS
kubernetes
RUCIO
ESCAPE's Data Lake running on a K8s cluster: storage orchestration + data management
jupyter
WebUI: DataLake-as-a-Service
(credits to Riccardo Di Maria, Muhammad Aditya Hilmy)
reana
a platform to reproduce workflows running on various computing backends
LAPP
cnrs
SCUOLA NORMALE SUPERIORE
INFN
FAU
LUND UNIVERSITY
Horizon 2020
ESCAPE
CERN
EOSC Future

EOSC update



Initiatives

Batch Working Group in Kubernetes

Most active: organized by Apple, Google, VMWare, RedHat, Intel

Meetings on Thursdays 7am and 3pm PT (alternating)

Focus on support in upstream Kubernetes, working closely with SIGs

https://github.com/kubernetes/community/tree/master/wg-batch

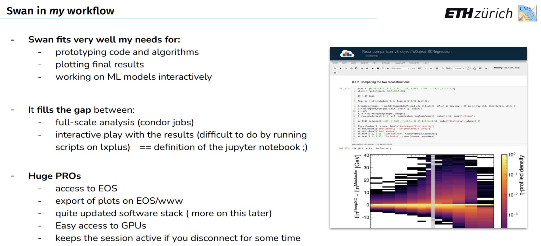
CNCF Batch System Initiative

Slow start, promoted by projects like Volcano, Armada, …

Batch system specification to be incorporated into Kubernetes, Volcano, Armada, etc

https://github.com/cncf/tag-runtime/issues/38

Kubernetes

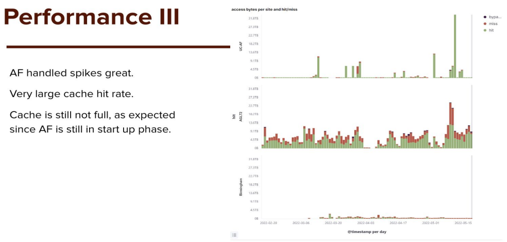


Swan in *my* workflow

- **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 ;)
- **Huge PROs**
  - access to EOS
  - export of plots on EOS/www
  - quite updated software stack (more on this later)
  - Easy access to GPUs
  - keeps the session active if you disconnect for some time

ETH zürich
CMS
Reco_comparison_v8_objectToObject_SCRegression
0.1.2 Comparing the two reconstructions
$En^{DeepSC} - En^{Mustache}$ [GeV]
η-profiled density

UX SWAN&coffea



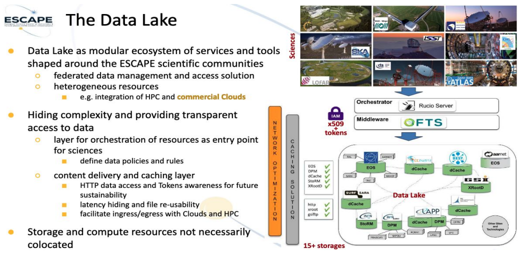
Performance III

AF handled spikes great.

Very large cache hit rate.

Cache is still not full, as expected since AF is still in start up phase.

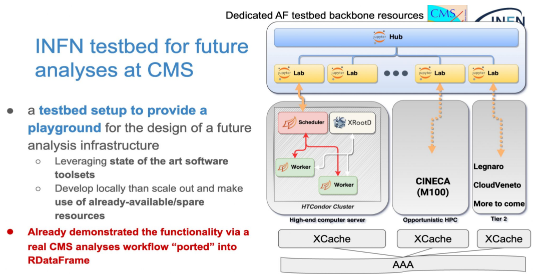
XCache



ESCAPE The Data Lake

- Data Lake as modular ecosystem of services and tools shaped around the ESCAPE scientific communities
  - federated data management and access solution
  - heterogeneous resources
    - e.g. integration of HPC and commercial Clouds
- Hiding complexity and providing transparent access to data
  - layer for orchestration of resources as entry point for sciences
    - define data policies and rules
  - content delivery and caching layer
    - HTTP data access and Tokens awareness for future sustainability
    - latency hiding and file re-usability
    - facilitate ingress/egress with Clouds and HPC
- Storage and compute resources not necessarily colocated

Escape

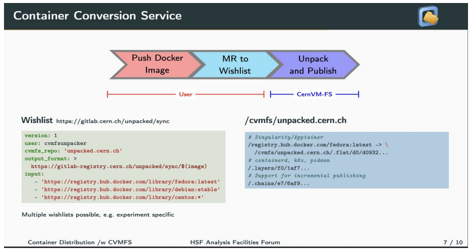


Dedicated AF testbed backbone resources

INFN testbed for future analyses at CMS

- a **testbed setup to provide a playground** for the design of a future analysis infrastructure
  - Leveraging **state of the art software toolsets**
  - Develop locally than scale out and make **use of already-available/spare resources**
- **Already demonstrated the functionality via a real CMS analyses workflow "ported" into RDataFrame**

Multi site AF model at INFN



Container Conversion Service

Wishlist https://gitlab.cern.ch/unpacked/sync

```
version: 1
user: cvmfsunpacker
cvmfs_repo: 'unpacked.cern.ch'
output_format: >
  https://gitlab-registry.cern.ch/unpacked/sync/$(image)
input:
   - 'https://registry.hub.docker.com/library/fedora:latest'
   - 'https://registry.hub.docker.com/library/debian:stable'
   - 'https://registry.hub.docker.com/library/centos:*'
```

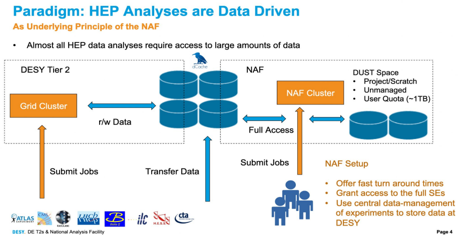
/cvmfs/unpacked.cern.ch

```
# Singularity/Apptainer
/registry.hub.docker.com/fedora:latest -> \
  /cvmfs/unpacked.cern.ch/.flat/d0/d0932...
# containerd, k8s, podman
/.layers/f0/1af7...
# Support for incremental publishing
/.chains/e7/6af9...
```

Multiple wishlists possible, e.g. experiment specific

Container Distribution /w CVMFS HSF Analysis Facilities Forum 7 / 10

Containers in CVMFS



Paradigm: HEP Analyses are Data Driven

As Underlying Principle of the NAF

- Almost all HEP data analyses require access to large amounts of data

NAF Setup

- Offer fast turn around times
- Grant access to the full SEs
- Use central data-management of experiments to store data at DESY

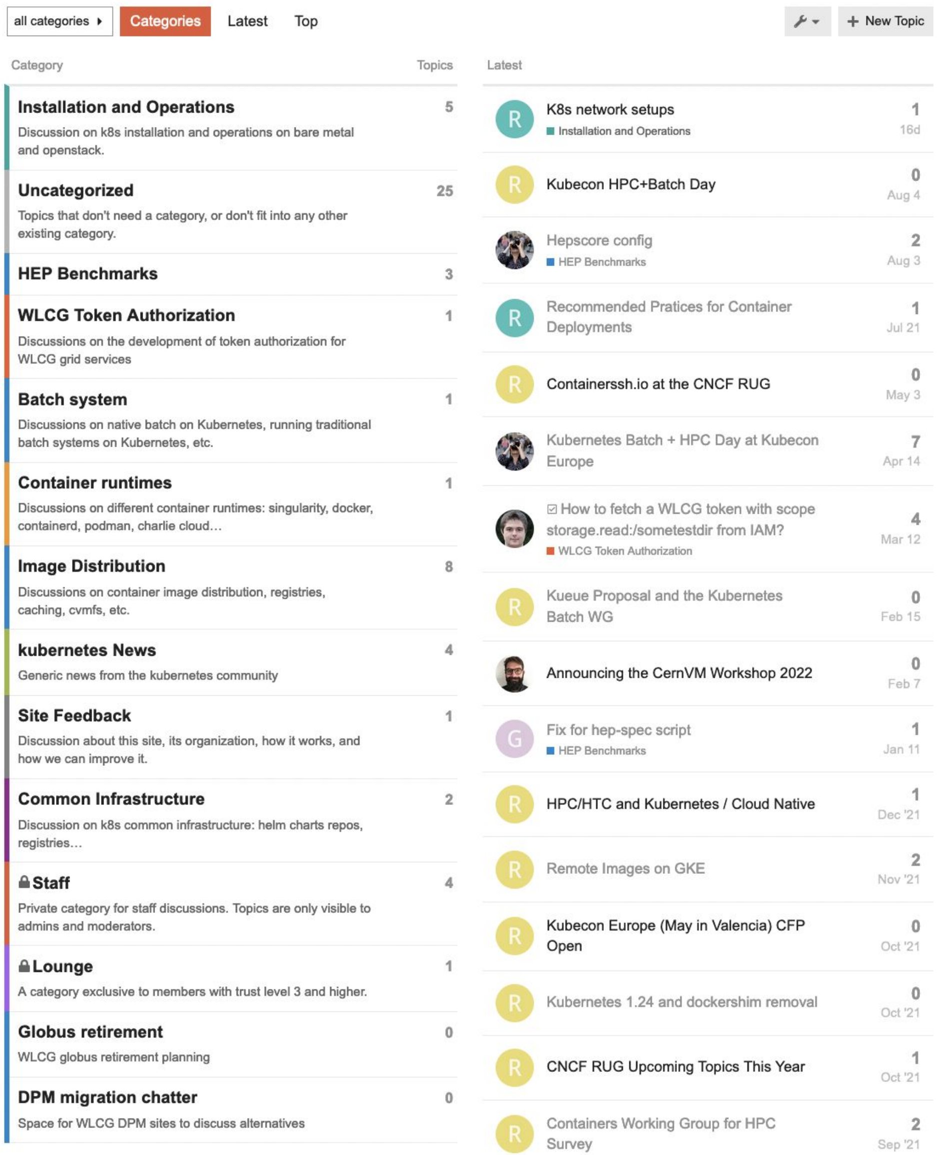
DESY. DE T2s & National Analysis Facility Page 4

UX DESY NAF

- HSF AF page
- Indico Category

WLCG Discourse

- Proposal to use wlcg-discourse.web.cern.ch for discussion since there is no other general forum for sites starting or needing an answer.
- Pros/cons
  - Pros: All in one place and easy to search.
  - Cons: Not well known enough to generate useful discussions



Category	Topics
Installation and Operations Discussion on k8s installation and operations on bare metal and openstack.	5
Uncategorized Topics that don't need a category, or don't fit into any other existing category.	25
HEP Benchmarks	3
WLCG Token Authorization Discussions on the development of token authorization for WLCG grid services	1
Batch system Discussions on native batch on Kubernetes, running traditional batch systems on Kubernetes, etc.	1
Container runtimes Discussions on different container runtimes: singularity, docker, containerd, podman, charlie cloud…	1
Image Distribution Discussions on container image distribution, registries, caching, cvmfs, etc.	8
kubernetes News Generic news from the kubernetes community	4
Site Feedback Discussion about this site, its organization, how it works, and how we can improve it.	1
Common Infrastructure Discussion on k8s common infrastructure: helm charts repos, registries…	2
🔒Staff Private category for staff discussions. Topics are only visible to admins and moderators.	4
🔒Lounge A category exclusive to members with trust level 3 and higher.	1
Globus retirement WLCG globus retirement planning	0
DPM migration chatter Space for WLCG DPM sites to discuss alternatives	0

Latest		
K8s network setups Installation and Operations	1	16d
Kubecon HPC+Batch Day	0	Aug 4
Hepscore config HEP Benchmarks	2	Aug 3
Recommended Pratices for Container Deployments	1	Jul 21
Containerssh.io at the CNCF RUG	0	May 3
Kubernetes Batch + HPC Day at Kubecon Europe	7	Apr 14
☑ How to fetch a WLCG token with scope storage.read:/sometestdir from IAM? WLCG Token Authorization	4	Mar 12
Kueue Proposal and the Kubernetes Batch WG	0	Feb 15
Announcing the CernVM Workshop 2022	0	Feb 7
Fix for hep-spec script HEP Benchmarks	1	Jan 11
HPC/HTC and Kubernetes / Cloud Native	1	Dec '21
Remote Images on GKE	2	Nov '21
Kubecon Europe (May in Valencia) CFP Open	0	Oct '21
Kubernetes 1.24 and dockershim removal	0	Oct '21
CNCF RUG Upcoming Topics This Year	1	Oct '21
Containers Working Group for HPC Survey	2	Sep '21