HEP Software Foundation Analysis Facilities Forum

EOSC-Future ESCAPE Science Project progress meeting
July 2022

MANCHESTER
1824
The University of Manchester

erc
BEAUTY2CHARM
European Research Council
Established by the European Commission

What is HSF?

HEP Software Foundation

hepsoftwarefoundation.org/

High Energy Physics has a vast investment in software

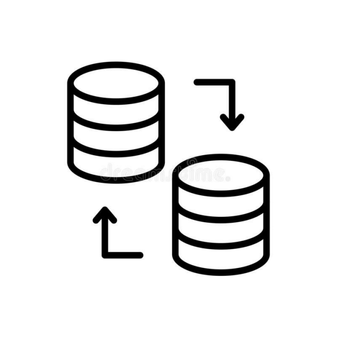
```
class TestClass
{
    int val1;
    int val2;

    TestClass(int x, int y)
    {
        val1 = x;
        val2 = y;
    }
}

class Testing
{
    static public void main (String args[])
    {
        String string1 = new String("Hello world.");
        TestClass foo = new TestClass(2,4);
    }
}
```

50M lines of C++
Worth 500M$

1M CPU cores every hour

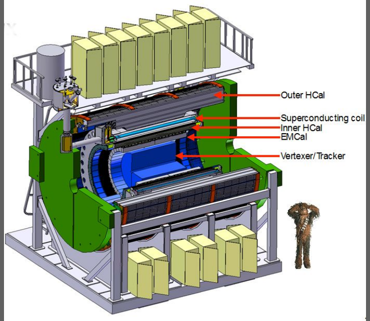


100PB of data
transfers per year

1000PB of data

What is HSF?

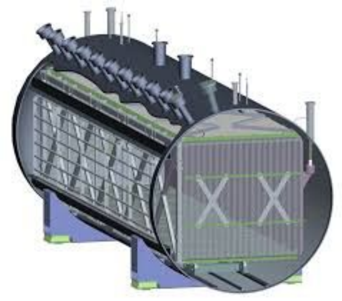
sPhenix



Xenon1T

BelleII

MicroBooNE



LHC and non-LHC experiments face the same software challenges

- Evolve to meet these challenges and overcome limitations
- Exploit expertise inside **and** outside our community
- Cannot afford duplicated efforts

What is HSF?

PyHEP

Data Analysis

Reconstruction and Software Triggers

Training

“The HEP Software Foundation facilitates cooperation and common efforts in High Energy Physics software and computing internationally”

Software Developer tools and packaging

Frameworks

Detector simulation

Physics generators

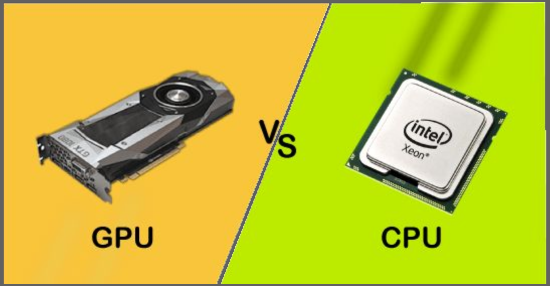
NEW

Analysis facilities forum

What is an Analysis Facility?

Data

Software



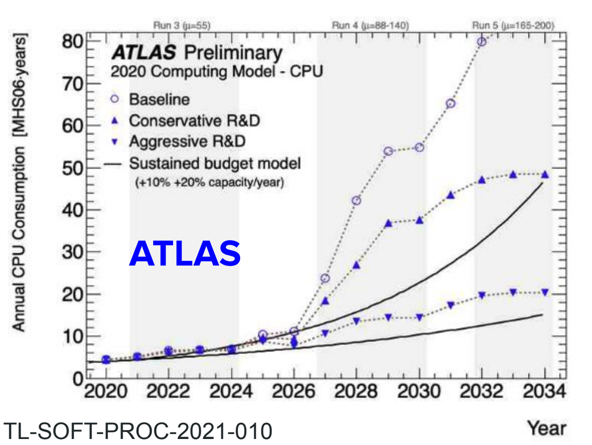
Computing resources

Support

"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."

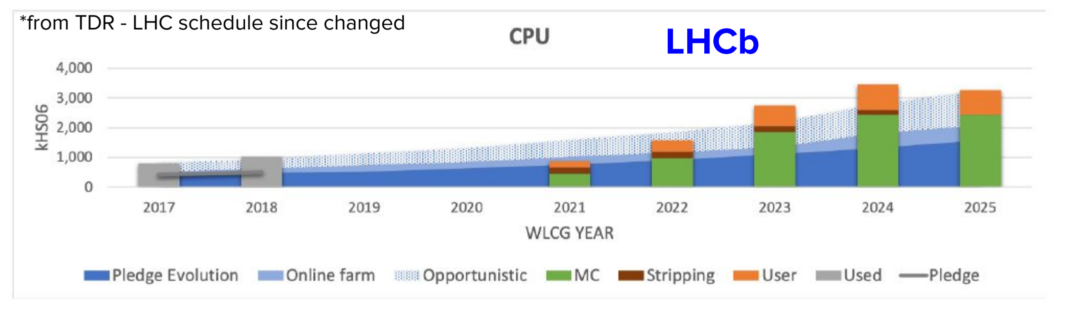
Why is this important now?

HL-LHC will see orders of magnitude more data - unprecedented scientific data volume at multi-exabyte scale



ATLAS

TL-SOFT-PROC-2021-010



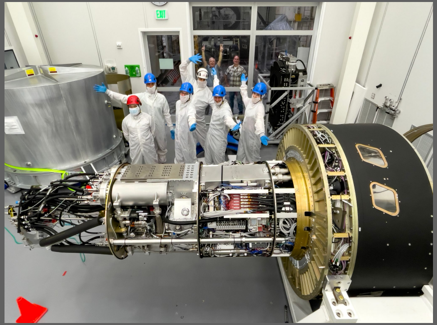
*from TDR - LHC schedule since changed

LHCb

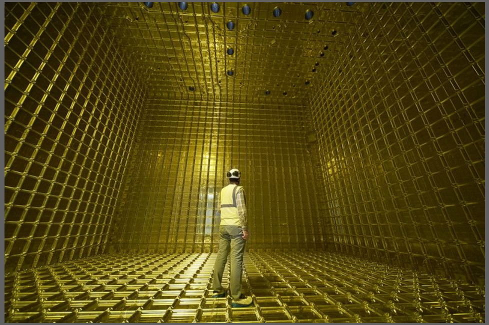
Current LHC computing model will not provide the required data processing capabilities even with hardware evolution

Why is this important now?

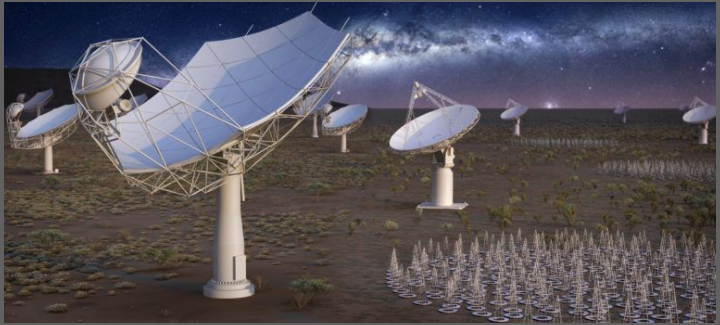
Not just the LHC...



LSST



DUNE



SKA

- Current "local" end-user data analysis methods and tools will not scale - common solutions?
- Sharing and optimising use of specialized infrastructure will become more and more important

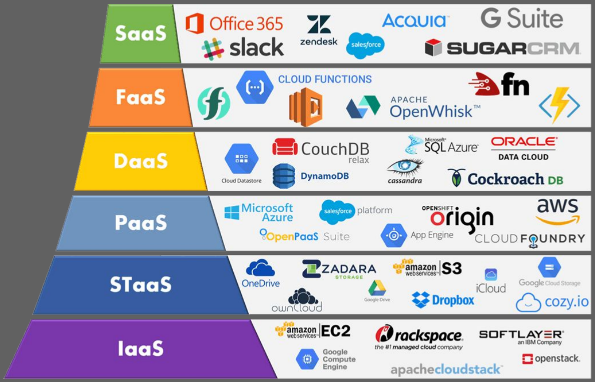
Why is this important now?

Technologies evolution

kubernetes

SCITOKENS

ServiceX

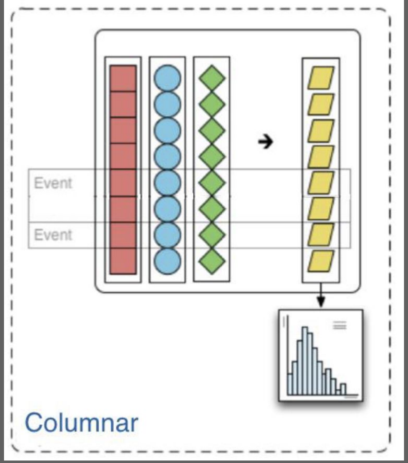


XCache

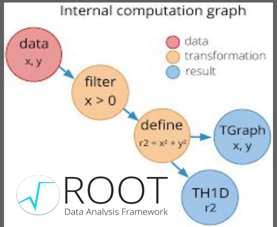
New analysis techniques

DASK

jupyter



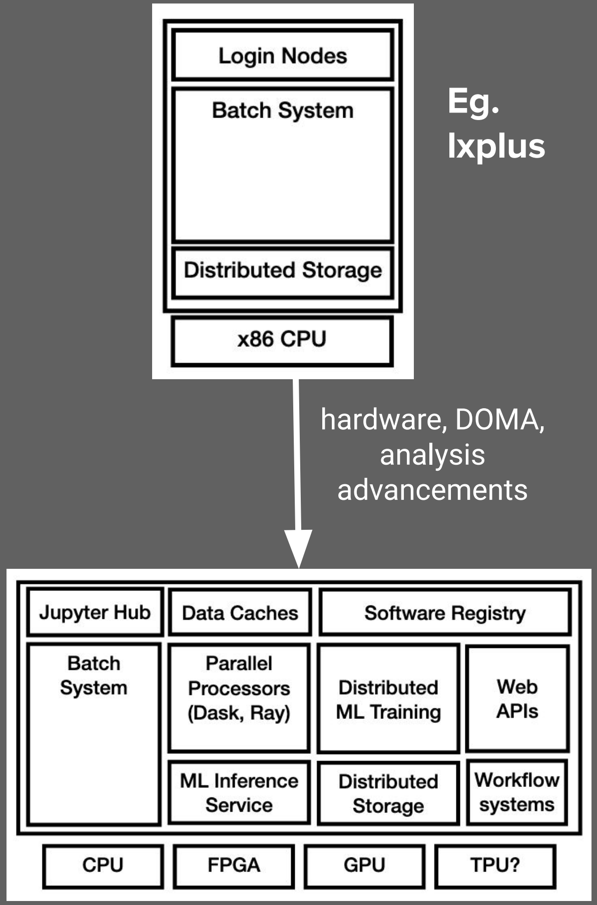
Scikit HEP



Analysis Facility requirements

Essential components are now considered to be

- ☑ Interactive ssh machines
- ☑ Classic batch system (HTcondor or slurm)
- ☑ Jupyter hub either integrated with HTCondor or k8s
- ☑ Heterogenous resources available for the users, not only CPU
- ☑ Local storage with POSIX interface and possibly object store access



Eg. lxplus

hardware, DOMA, analysis advancements

What is the current status of analysis facilities?

US Analysis Facilities

- Coffea-casa @ UNL and UChicago
- Elastic Analysis Facility at FNAL
- AF @ Purdue
- AF @ MIT
- AF @ UChicago
- DOE facilities

EU Analysis facilities

- Distributed Dask-based national facility @INFN
- National Analysis Facility (NAF) @ DESY
- SWAN facility @ CERN
- AF @ PIC (WIP)

Healthy variety of options emerging with a different focus!

What is the current status of analysis facilities?

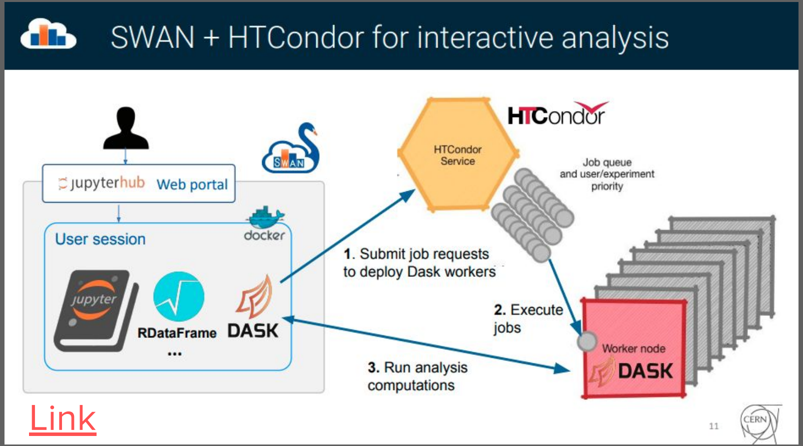
SWAN

Interactive Data Analysis, in the Cloud.

SWAN: Service for Web-based Analysis

- Interface: Jupyter notebook
- Storage: EOS/CERNbox
- Resources: Spark clusters + HTCondor pools + GPUs
- For: LHC experiment agnostic
- Runs: RDataFrame/Coffea with DASK

Already deployed outside CERN (ScienceBox)!



SWAN + HTCondor for interactive analysis

jupyterhub Web portal

User session

docker

jupyter

RDataFrame

...

DASK

HTCondor Service

HTCondor

Job queue and user/experiment priority

1. Submit job requests to deploy Dask workers

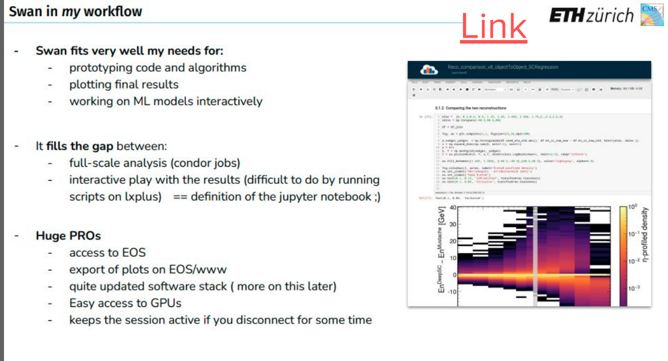
2. Execute jobs

3. Run analysis computations

Worker node DASK

Link

11



Swan in *my* workflow

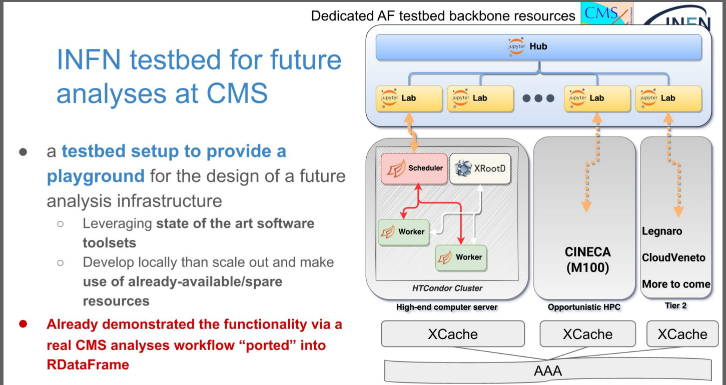
Link

ETH zürich

- **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

What is the current status of analysis facilities?

INFN
Istituto Nazionale di Fisica Nucleare



INFN testbed for future analyses at CMS

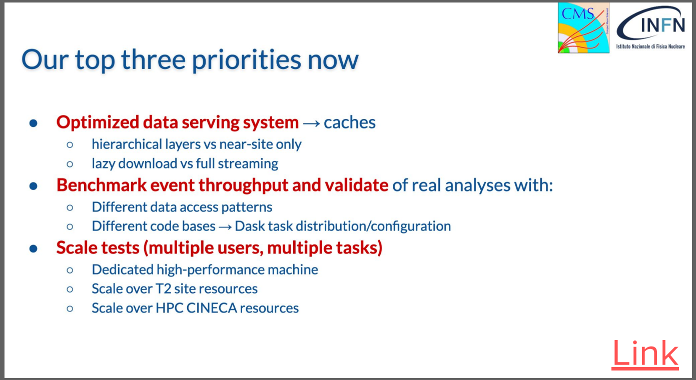
- a **testbed setup to provide a playground** for the design of a future analysis infrastructure
  - Leveraging **state of the art software toolsets**
  - Develop locally than scale out and make **use of already-available/spare resources**
- **Already demonstrated the functionality via a real CMS analyses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

Dedicated AF testbed backbone resources

INFN analysis facility

- Interface: Jupyter notebook
- Storage: Local AF area (CEPH)
- Resources: T2 sites + CINECA (HPC)
- For: CMS
- Runs: Benchmarking with RDataFrame with DASK, starting validation with Coffea
- Services: XCache

Expanding to other experiments!



Our top three priorities now

- **Optimized data serving system** → caches
  - hierarchical layers vs near-site only
  - lazy download vs full streaming
- **Benchmark event throughput and validate** of real analyses with:
  - Different data access patterns
  - Different code bases → Dask task distribution/configuration
- **Scale tests (multiple users, multiple tasks)**
  - Dedicated high-performance machine
  - Scale over T2 site resources
  - Scale over HPC CINECA resources

Link

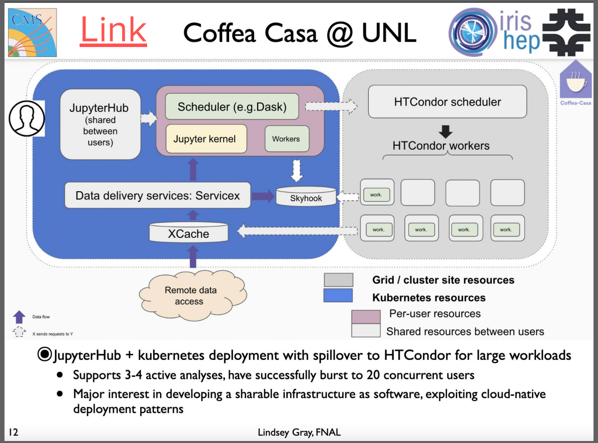
What is the current status of analysis facilities?

Coffea-Casa

Coffea-casa AF - services for rapid processing of data in a column-wise fashion

- Interface: Jupyter notebook
- Storage: Local AF area (NVMe CEPH)
- Resources: K8s colocated with T2 sites resources
- For: CMS/ATLAS
- Runs: Coffea analysis framework with DASK/HTCondor
- Services: XCache and ServiceX (Skyhook in progress)

Deployed at multiple sites (UNL, UChicago)



Link Coffea Casa @ UNL

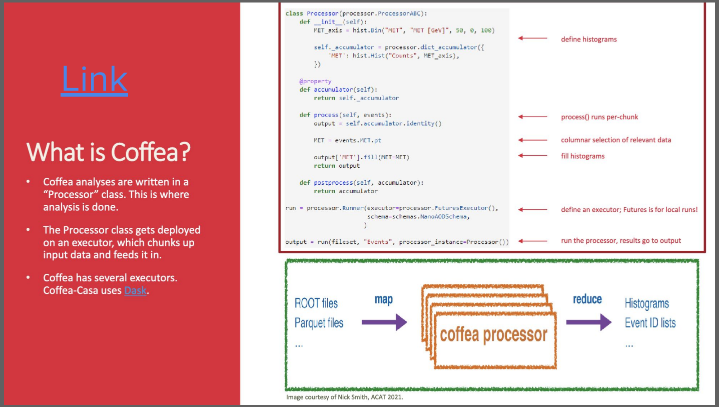
▨	**Grid / cluster site resources**
🟦	**Kubernetes resources**
▨	Per-user resources
□	Shared resources between users

◉ JupyterHub + kubernetes deployment with spillover to HTCondor for large workloads

- Supports 3-4 active analyses, have successfully burst to 20 concurrent users
- Major interest in developing a sharable infrastructure as software, exploiting cloud-native deployment patterns

12

Lindsey Gray, FNAL



Link

What is Coffea?

- Coffea analyses are written in a "Processor" class. This is where analysis is done.
- The Processor class gets deployed on an executor, which chunks up input data and feeds it in.
- Coffea has several executors. Coffea-Casa uses Dask.

```
class Processor(processor.ProcessorABC):
    def __init__(self):
        MET_axis = hist.Bin("MET", "MET [GeV]", 50, 0, 100)

        self._accumulator = processor.dict_accumulator({
            'MET': hist.Hist("Counts", MET_axis),
        })

    @property
    def accumulator(self):
        return self._accumulator

    def process(self, events):
        output = self.accumulator.identity()

        MET = events.MET.pt

        output['MET'].fill(MET=MET)
        return output

    def postprocess(self, accumulator):
        return accumulator

run = processor.Runner(executor=processor.FuturesExecutor(),
                       schema=schemas.NanoAODSchema,
                      )

output = run(fileset, "Events", processor_instance=Processor())
```

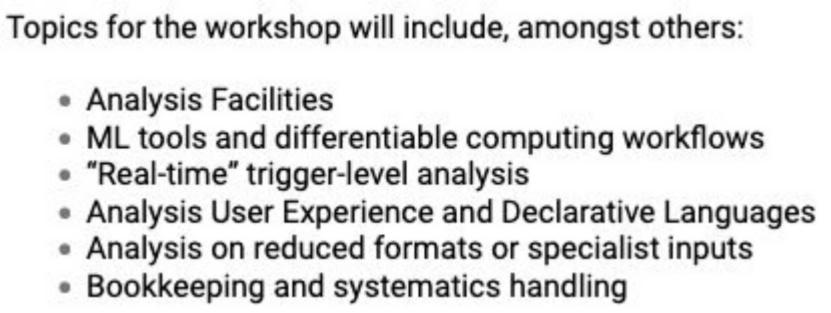
Image courtesy of Nick Smith, ACAT 2021.

Next steps with analysis facilities?

Meeting link

Analysis Ecosystems Workshop II

23–25 May 2022
IJCLab
Europe/Zurich timezone



Topics for the workshop will include, amongst others:

- Analysis Facilities
- ML tools and differentiable computing workflows
- “Real-time” trigger-level analysis
- Analysis User Experience and Declarative Languages
- Analysis on reduced formats or specialist inputs
- Bookkeeping and systematics handling

Report being prepared - key areas identified

Interoperability
Resource sharing
Benchmarking (AGC)

Identity management
Sharing environments

DOMA
Surveying analysts

The HSF analysis facilities forum

- Bring together invested parties for dedicated, technical discussions **on a bi-weekly basis**
- Every 2 months session dedicated to **user experiences**
- Build/foster community and **“bridge” various involved parties**: experiments, software stakeholders, data centres, WLCG, IRIS-HEP and HSF
- Evaluate proposed solutions with **white paper** after one year outlining community vision for future AFs designed for HL-LHC scale analysis

Activities so far

Indico link

All recorded

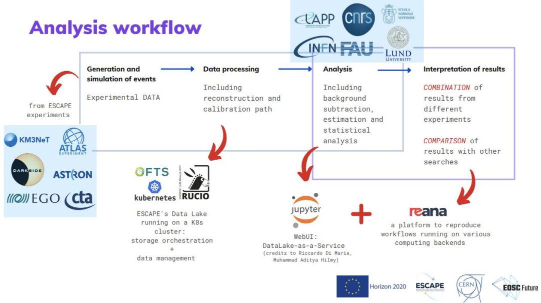
HSF Analysis Facilities Forum Kick-off Meeting

Mark Neubauer

University of Illinois at Urbana-Champaign

March 25, 2022

Kick-off meeting



Analysis workflow

Generation and simulation of events → **Data processing** → **Analysis** → **Interpretation of results**

from ESCAPE experiments

Experimental DATA

Including reconstruction and calibration path

Including background subtraction, estimation and statistical analysis

COMBINATION of results from different experiments

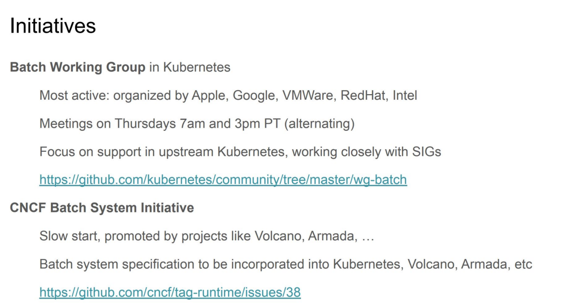
COMPARISON of results with other searches

ESCAPE's Data Lake running on a K8s cluster:
storage orchestration
+
data management

WebUI:
DataLake-as-a-Service
(credits to Riccardo Di Maria, Muhammad Aditya Hilmy)

a platform to reproduce workflows running on various computing backends

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

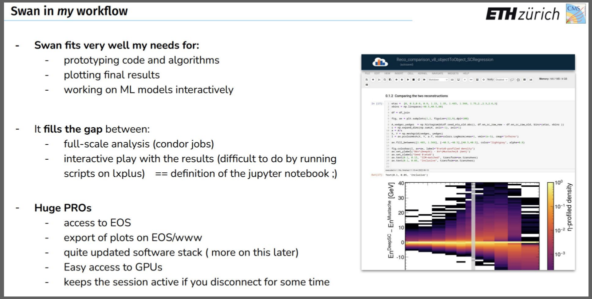
CNCF Batch System Initiative

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

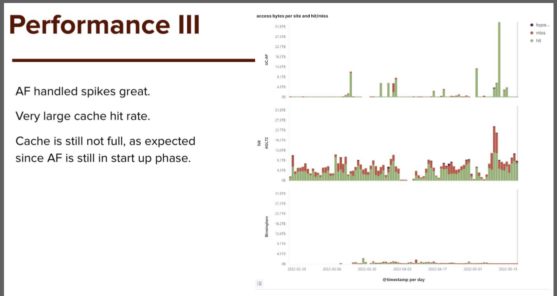


Swan in *my* workflow

- 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
η-profiled density

User experience



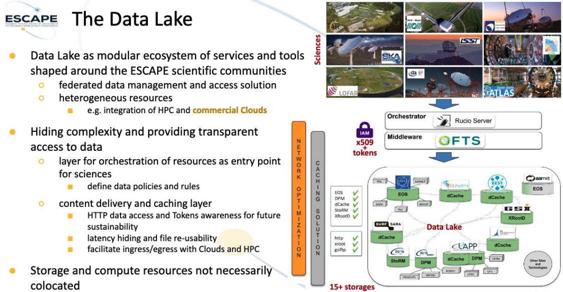
Performance III

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

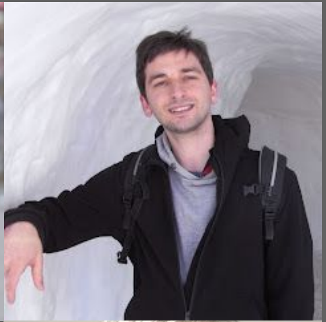


The Data Lake

- 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 DLaaS

Practical information



Conveners:

Alessandra Forti (ATLAS)

Lukas Heinrich (ATLAS)

Diego Ciangottini (CMS)

Nicole Skidmore (LHCb)

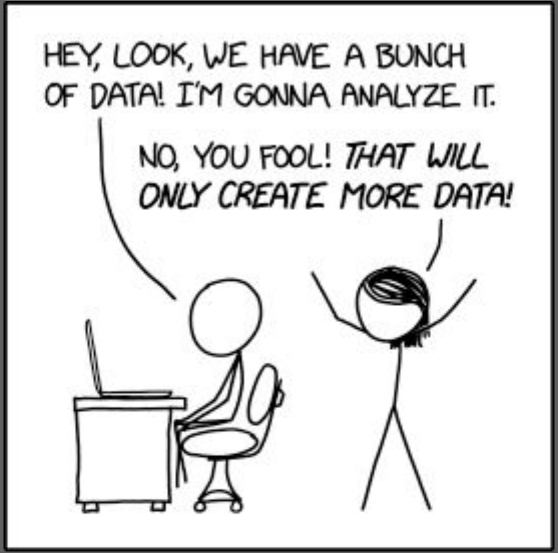
Mailing lists:

[HSFAFFORUM]

Mattermost:

[hsf-af-forum]

Thank you


HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! THAT WILL ONLY CREATE MORE DATA!

Glossary

- CEPH: open source software-defined storage solution for block, file and object storage
- DOMA: Data Organization, Management and Access
- WLCG: Worldwide LHC Computing Grid
- HTCondor:
- HTTP: Hypertext Transfer Protocol. HTTP is the protocol used to transfer data over the web. A typical flow over HTTP involves a client machine making a request to a server, which then sends a response message.
- HTTPS: Hypertext Transfer Protocol Secure - the secure version of HTTP used for secure communication over a network
- SciTokens: The SciTokens project builds a federated ecosystem for authorization on distributed scientific computing infrastructures.
- IAM: Identity and Access Management
- OIDC: OpenID Connect. An authentication protocol which verifies user identity when trying to access a protected HTTPs end point.
- aaS: “as a Service”. Eg. PaaS = Platforms as a Service, SaaS = Software as a Service
- Kubernetes: (k8s) is an open source container orchestration platform that automates many of the manual processes involved in deploying, managing, and scaling containerized applications.
- Apache Spark: Apache Spark is an open-source unified analytics engine for large-scale data processing.
- Dask: flexible library for parallel computing in Python. Similar to Apache Spark but integrates with existing Python tools.
- Ray: Ray is a high-performance distributed execution framework targeted at large-scale machine learning and reinforcement learning applications

Glossary

nVME: NVMe is the latest and greatest storage interface for laptops and desktops, and it offers much faster read and write speeds than older interfaces.

- FTS: a low level data movement service, responsible for reliable bulk transfer of files between storages. It’ s responsible for globally distributing the majority of the LHC data across the WLCG infrastructure
- REANA: reusable and reproducible research data analysis platform
- Rucio: provides services and associated libraries for allowing scientific collaborations to manage large volumes of data spread across facilities at multiple institutions and organisations (ATLAS uses this). LHCb has DIRAC for this.
- HSF: HEP Software Foundation
- SLATE
- ServiceX: ServiceX is a data extraction and delivery delivery service
- XCache: cache-based data approaches to increase efficiency of CPU use (via reduced latency) and network (reduce WAN traffic)
- Data lake: storage service geographically distributed across large data centers connected by fast network with low latency. Alternative to running jobs at site where files are located.
- Object store access: Discrete data units - complex hierarchies as in a file-based system. Each object is a simple, self-contained repository that includes the data, metadata and ID number (instead of a file name and file path). Scales well.
- POSIX: Portable Operating System Interface - standards for maintaining compatibility between operating systems. Defines system- and user-level API, with command line shells and utility interfaces for software compatibility (portability) with variants of Unix and other operating systems.
- Skyhook: service to recognize the layout of files and “push down” structured queries from client to server, taking advantage of the computational capacity in the storage hardware and reducing data movement significantly. Its an extension of the Ceph open source distributed storage system

Glossary

- API: Application Programming Interface
- HTCondor: open-source high-throughput computing software framework for distributed parallelization of computationally intensive tasks - used to manage workload on computing clusters. Formally known as Condor
- Federated ID management: linking a person's electronic identity and attributes, stored across multiple distinct identity management systems - related to single sign-on (SSO), in which a user's single authentication ticket, or token, is trusted across multiple systems/organizations