Status & Plans for CMS Analysis Facilities

D. Hufnagel, et al. on behalf of (US)CMS

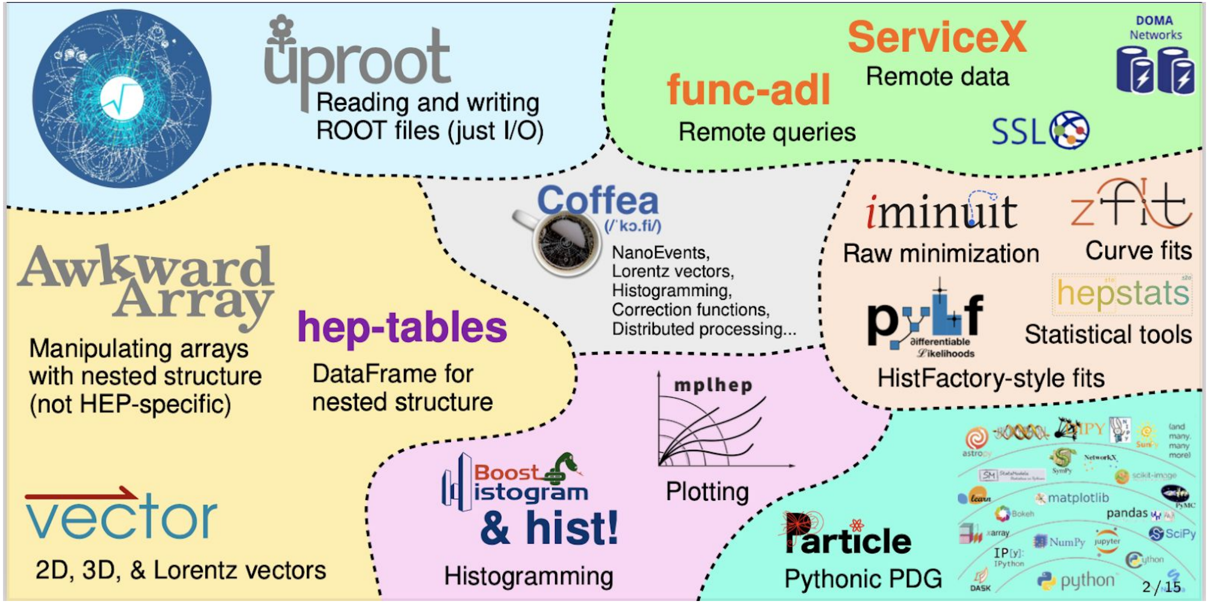
IRIS-HEP Steering Board Meeting 16.Jan 2024

(thanks to Lindsey Gray for providing most of the material)

Analysis Facilities Then and Now

- Physicists can handle an enormous amount of workflow complexity to achieve their goals
  - What's "easy" is incredibly subjective
    - person to person and analysis to analysis
- CMS Analysis facilities circa 2005-2019 have largely been login terminals with batch access
  - Just having that was sufficient to be a facility
  - CMS has published > 1000 papers working this way
- However, it can be easier and less hectic for data analysts
  - Technologies like Dask, Apache Spark, Parsl, and Work Queue encapsulate and abstract physics analysis workflows
  - With this abstraction, administrators are able to determine more optimal resource deployment/usage patterns with a "weaker" binding directly to user code
  - Physicists can focus on physics while also efficiently using clusters
- This talk is a snapshot of current analysis facilities efforts within CMS and their status
  (technical details mostly pushed to backup slides)

New Facilities for New Tools



- Generally: AFs provide a curated substrate upon which to easily deploy these (stacks of) tools at scale
  - The exact way in which this service is provided is currently filled with opinion, but there is a large degree of convergent evolution
  - We'll enumerate all the efforts and the directions under study at present
  - Each effort, while similar, does have different foci
- **While not in the box above - RDataFrame is within the scope of all AFs discussed in this talk**
- **Also a note about alternative statistics tools used by CMS on the next slide**

Public Release of CMS Likelihoods

CMS Collaboration Board approved the publishing of full likelihoods of our analyses from now on!

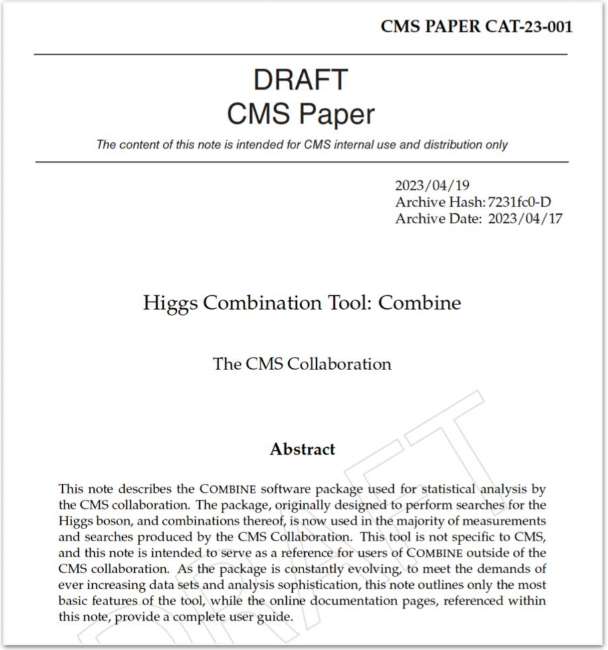
CMS will provide a standalone software image for this purpose.

Datacard common format efforts should be supported.

The Combine tool is used by ~90% of CMS analysis.

Like zfit it supports unbinned likelihoods, which is considered essential for many CMS analysis.

A binned fitting tool like pyhf could cover at most 50% of CMS use cases.


CMS PAPER CAT-23-001

DRAFT
CMS Paper

The content of this note is intended for CMS internal use and distribution only

2023/04/19
Archive Hash: 7231fc0-D
Archive Date: 2023/04/17

Higgs Combination Tool: Combine

The CMS Collaboration

Abstract

This note describes the COMBINE software package used for statistical analysis by the CMS collaboration. The package, originally designed to perform searches for the Higgs boson, and combinations thereof, is now used in the majority of measurements and searches produced by the CMS Collaboration. This tool is not specific to CMS, and this note is intended to serve as a reference for users of COMBINE outside of the CMS collaboration. As the package is constantly evolving, to meet the demands of ever increasing data sets and analysis sophistication, this note outlines only the most basic features of the tool, while the online documentation pages, referenced within this note, provide a complete user guide.

Considering further publications on analysis tools

AF efforts within CMS

- Several AF efforts within CMS heavily focused on deploying modern workflows
  - Coffea Casa @ UNL, AF @ (MIT,Purdue,INFN,Ciemat), ElasticAF @ FNAL
  - Providing the usual terminal access as well as notebooks predominantly through JupyterHub
  - Coffea-casa, INFN, ElasticAF well-advanced on interface & access
  - More details in backup slides
- Each effort focusing on different aspects of eventual common goal
  - Healthy (and natural!) split between software infrastructure, hardware infrastructure, and multi-cluster overlay
- Expect scale-up, benchmarking, (more) publications using these facilities over the course of 2024
  - Exciting times ahead, and best-practices will emerge!

AF efforts within CMS

	UNL	MIT	Purdue	FNAL	INFN	Ciemat
Multi-VO	No	Yes?	No	Yes	Yes?	No
Authentication	CMS VO	Local but investigating VO	?	Local	?	?
RDataFrame/ Coffea (to start)	Coffea	Both	Coffea	Coffea	RDataFrame	?
Burst out	yes	yes	yes	yes	yes	?
Published Analysis	yes	In process	?	yes	benchmarks	?
Scale	O(10) analysis	multi-group	O(3)	multi-expt	?	?
Portable Software	Shared with FNAL	?	Similar to FNAL/UNL	Shared with UNL	?	?

Evolution and what's missing

- We want to develop an Analysis Facility Capability (being able to offer facilities that provide a set of features to users, where the users have a choice which of these features they want to use)
  - Each feature requires certain hardware/software and backend services
  - Currently the different AF don't all offer the same features, so a user has to pick the "right" AF based on what is needed for a given analysis. While the AF instances might never be completely identical, in the long term the feature sets they provide should be broad/common enough that analyses can be moved between them.
- Missing: Scalable access to GPU from AF
- Missing: Easy/dynamic way to provide access to data in the AF (without preloading it)
  - Provide access to selected columns from lower level data tiers
  - Create input ntuple on demand

Outlook

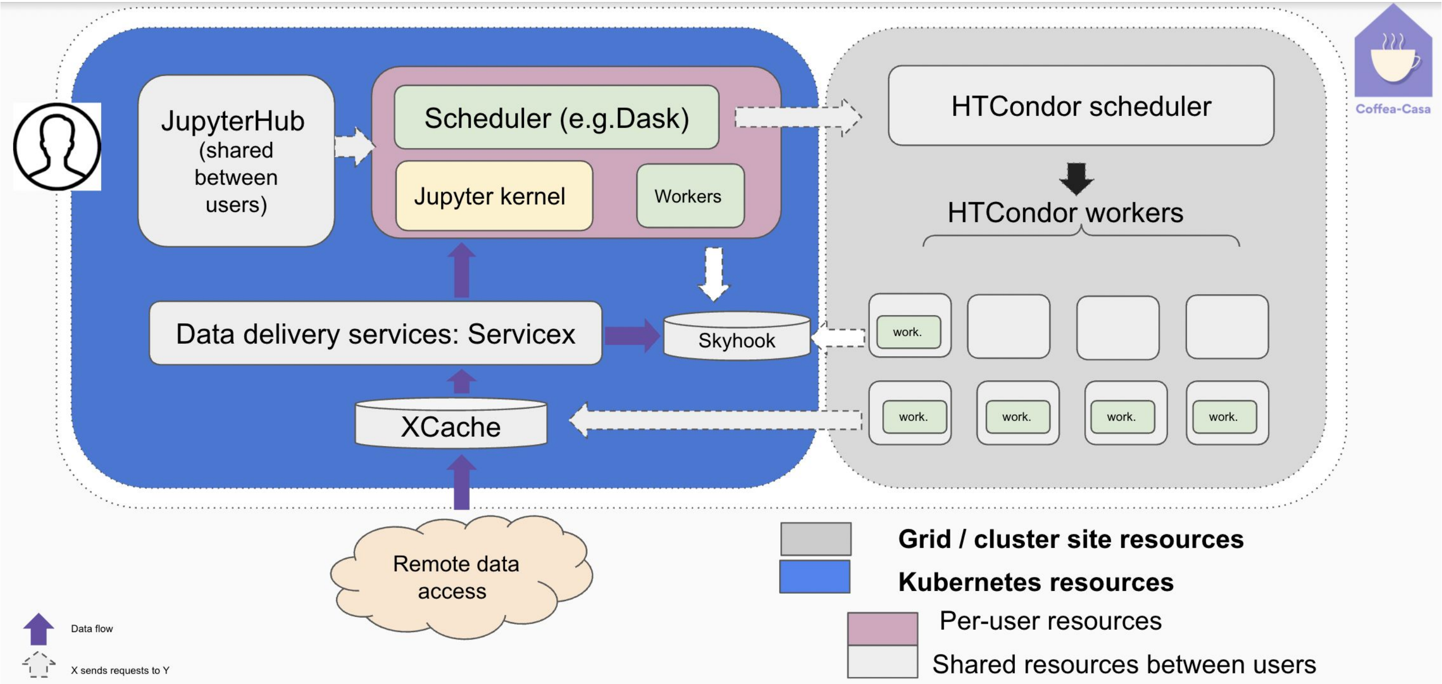
- Lots of AF activities in CMS and we are starting to move from pure R&D to actual use for published analysis
- Building an AF community isn't just software/hardware/tools/features
  - Need users to accept and use the AF, need support, need training
  - With increased use will automatically also come a settling on the list of common features we need, as users can vote with their feet and decide what is useful (or not) for them

BACKUP

CHEP Talks

- There were a few talks on AF activities during CHEP 2023
  - AF @ INFN
    - https://indico.jlab.org/event/459/contributions/11593/
  - AF @ Ciemat
    - https://indico.jlab.org/event/459/contributions/11627/
  - Coffea-Casa (focus on UNL)
    - https://indico.jlab.org/event/459/contributions/11610/
  - Coffea and Dask
    - https://indico.jlab.org/event/459/contributions/11533/

Coffea Casa @ UNL



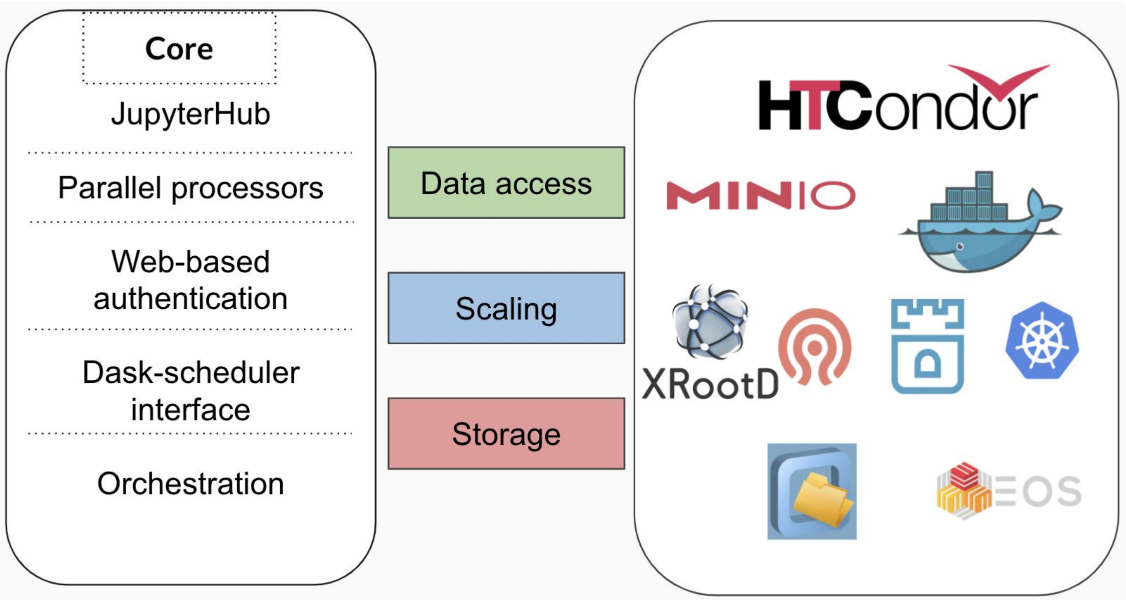
- JupyterHub + kubernetes deployment with spillover to HTCondor for large workloads
  - Supports running tasks for anyone with a valid CMS VO registration, supports O(10) analyses, with room for more
  - Major interest in developing a sharable infrastructure as software, exploiting cloud-native deployment patterns

Coffea Casa @ UNL

CMSAF @T2 Nebraska
“Coffea-casa”
https://coffea.casa

OpenData AF @T2 Nebraska
“Coffea-casa”
https://coffea-opendata.casa

ATLAS AF @Scalable System Lab (UChicago)
“Coffea-casa”



AF @ MIT - Initial Setup

- Login
  - Key based with MIT account (sponsored guest accounts?)
  - CMS data access authentication x509 for now
- Work environment
  - Load balanced JupyterHub access, Coffea type of analysis
  - Dask sitting on top of MIT Tier-3/Tier-2 centers
  - HTCondor and Slurm as batch managers
- Data access optimization
  - Tiered storage seems an obvious candidate for ‘sophisticated’ optimization of storage… work in progress:
    - 50 TB of NVMe should function as a hot cache for most accessed data
    - Tape is ideal candidate for rarely used data or just as safety net to recover from disaster

AF @ MIT - Infrastructure

- Computing for login
  - Order of ten beefy machines including, large memory, O(500) CPU cores and O(10) big GPUs (NVidia T100/T4)
- Network
  - 100 Gb/s for all machines, RDMA enabled
- Storage
  - Tiered Storage:
    - Tape storage from MIT Tape Pilot project (being commissioned)
    - Spinning disks: T2 (10 PB) at 100 Gb/s, T3 (300 TB) at 2x10 Gb/s
    - NVMe sticks: Local (50 TB) at 2x100 Gb (waiting for delivery)
  - XCache is planned
- Behind the scenes
  - HTCondor: Tier-2, Tier-3, global pool, OSG
  - Slurm: local HPC resources (old lattice QCD cluster)

AF @ Purdue Status and Plans

- The Purdue CMS T2 provides interactive AF capabilities for distributed physics data analysis since 2020, utilizing both interactive SSH sessions and JupyterHub to scale DASK/Spark clusters on HPC systems to over 1000 cores in parallel.
  - In that configuration the AF at Purdue was used in a CMS publication
    - DOI: 10.1007/JHEP01(2021)148) and in multiple ongoing analyses
  - MuonHLT upgrades, H→μμ Snowmass, Z'→ll, Top quark spin correlation

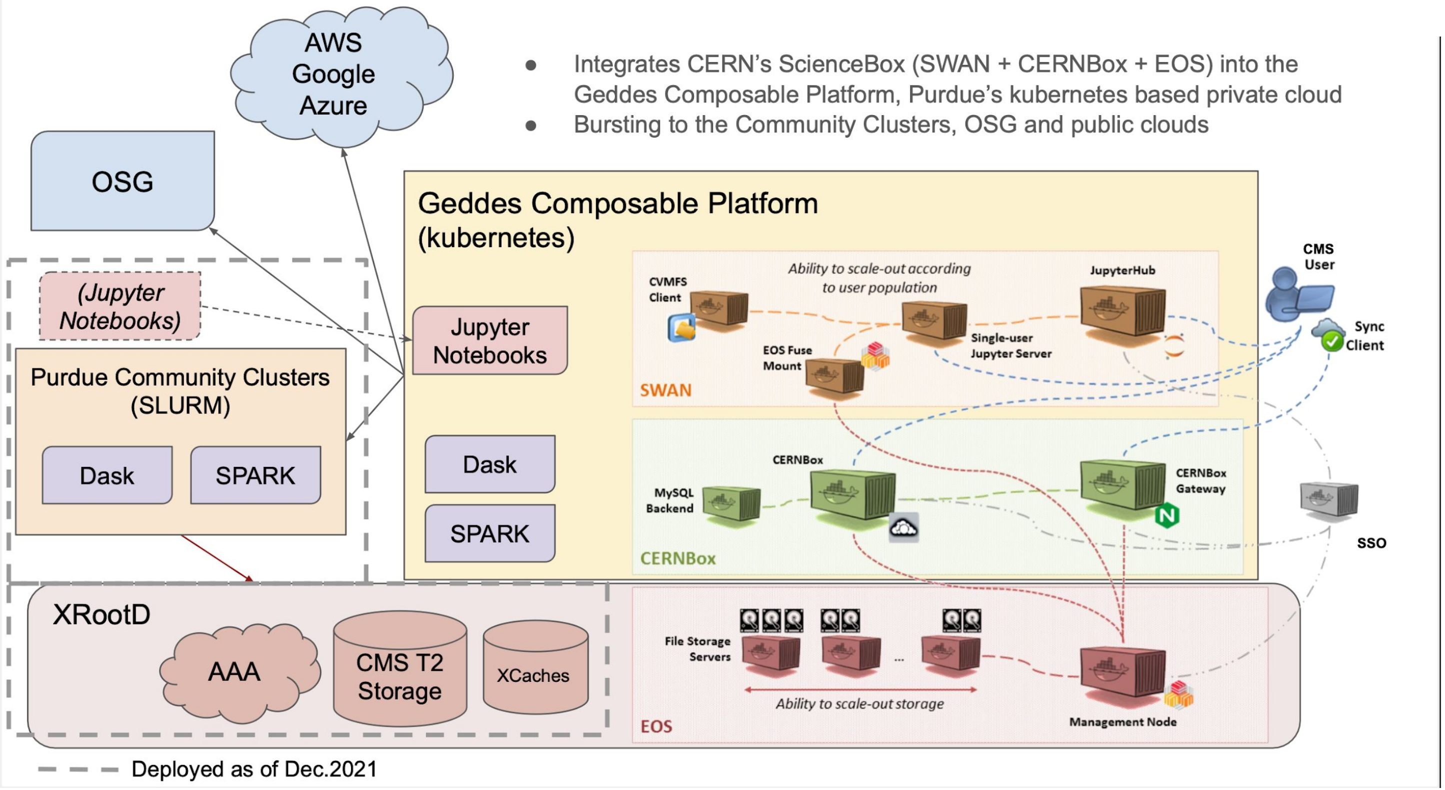
- In 2021 Purdue received USCMS funding for dedicated AF hardware, and our design evolved to include new AF capabilities based on CERN's ScienceBox (EOS, CERNBox, SWAN) running in Kubernetes, and leverage the new Geddes Composable Platform, a Kubernetes-based "Community Cloud" resource at Purdue.
  - Provides user-defined virtual clusters via DASK and Spark, for massively parallel user analyses based on coffea framework.
  - Integrates with Purdue's Kubernetes-based private cloud 'Geddes', and Purdue's Community Clusters.
  - Investigating OSG and public cloud integration in the future.

- Geddes Composable Platform
  - Purdue Research Computing has just built the Geddes Composable Platform - a private cloud resource based on Rancher and Kubernetes. This "Community Cloud" resource is a platform for flexible, scalable and reproducible scientific data analysis.
  - In June 2021, Purdue received NSF funding to build out a private campus cloud focused on data analytics and machine learning. Synergies with AF effort funded by USCMS

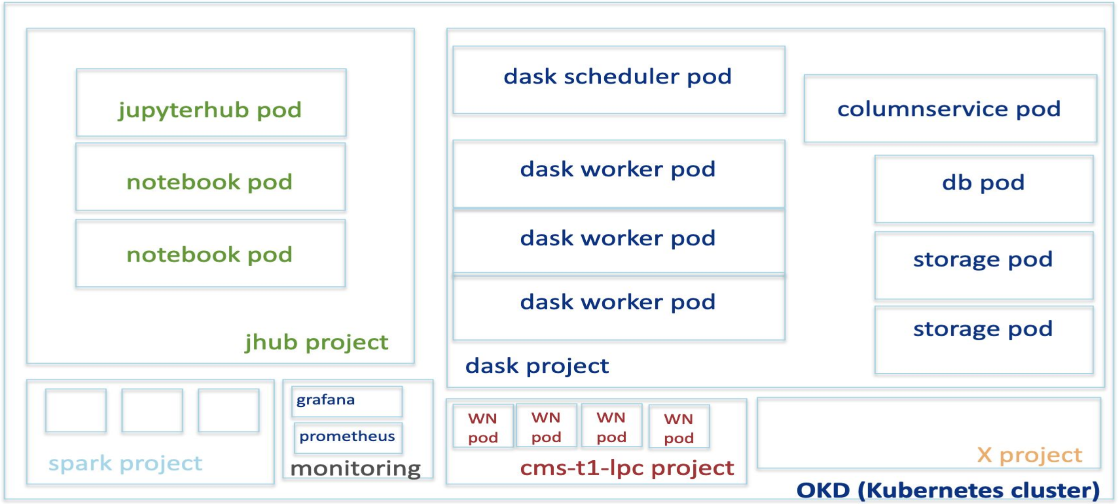
- The new hardware has already been received, and the upgrade took place over the course of 2022 in close collaboration with USCMS Operations Program.

AF @ Purdue Conceptual Layout

- Integrates CERN's ScienceBox (SWAN + CERNBox + EOS) into the Geddes Composable Platform, Purdue's kubernetes based private cloud
- Bursting to the Community Clusters, OSG and public clouds



Elastic Analysis Facility @ FNAL



From Burt's slides at OSG AHM: https://indico.fnal.gov/event/22127/contributions/194934/attachments/133990/165498/Elastic_AF_-_OSG_USLHC.pdf

Secure

- LDAP and VPN login, Kerberos. Docker image audits, and mitigation strategies put in place for data preservation and least privilege guarantee.

Integrated

- Ferry, Htcondor, dask-gateway, spark, triton

Multi-vo

- user management, centralized authorization, specialized environments, large-ish cvmfs infrastructure in place via NFS auto-scalable pods

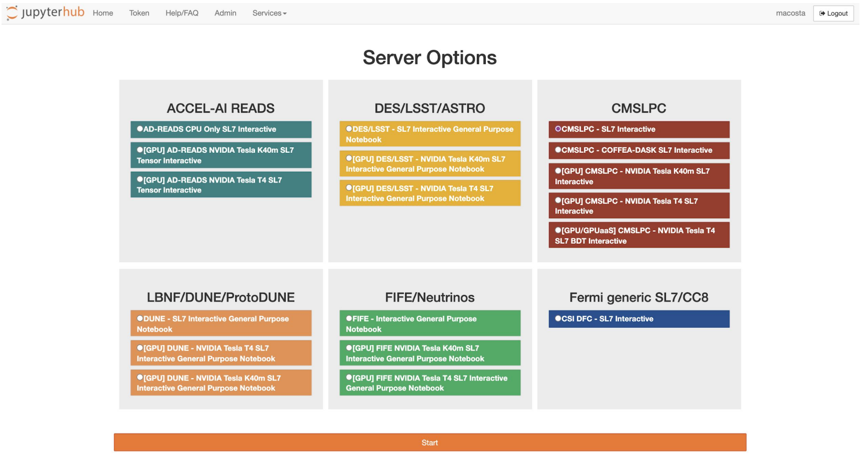
DevOps:

- CI/CD pipelines for all environments, CPU and GPU flavors

Other horizons:

- Now supporting Fermilab's Accelerator division Edge AI. Hoping to foster effort from SCD and AD on designing analysis facilities beyond SCD
- Collaborating with the Dask team on developing a plugin to integrate Dask Gateway with HTCondor, coming soon

Elastic AF: A Multi-Experiment AF

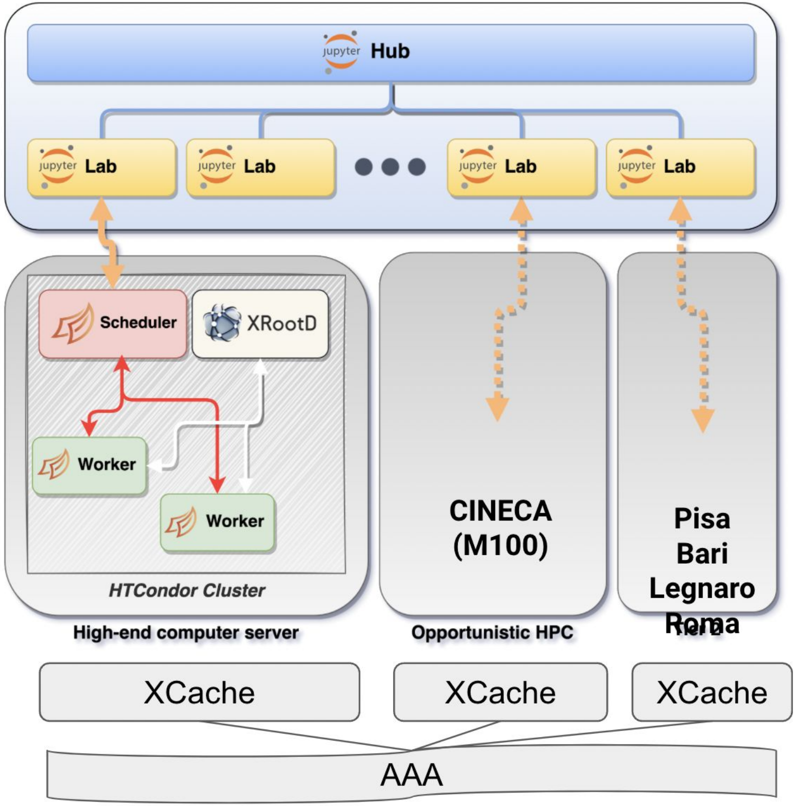


- Started as a USCMS project but has grown to be a multi-experiment project providing all services to multiple FNAL experiments.
  - EAF heard from and are actively collaborating with YorkU/Compute Canada for a prototype EAF for DUNE.
  - EAF developed more than 15 environments for experiments with dedicated CVMFS mounts, shared storage and specific scientific software, all in compliance with DOE cybersecurity requirements
  - EAF started collaborating with Fermilab's Accelerator Divison and designed an environment for the READS project Accelerator Real-time Edge AI for Distributed Systems (READS)
- Working examples of deploying inference as a service for analysis, spill over to LPC batch

AF @ INFN

Building on similar choices to other efforts

- Interfaces
  - JupyterHub as user entrypoint
  - JupyterLab to manage the user-facing interface
  - Direct access to HTCondor
  - User interface (either from JLab or old fashoned UI)
- DASK to introduce the scaling over a batch system
  - Multiple clusters per user → DASK cluster as atomic unit of work
- HTCondor as the batch system of choice
  - User prioritization and in general configuration tuning is under study
- XRootD as data access protocol toward AAA:
  - To understand how to best implement xcache
- IAM@CMS "token native" authentication
- Focus on modular infrastructure



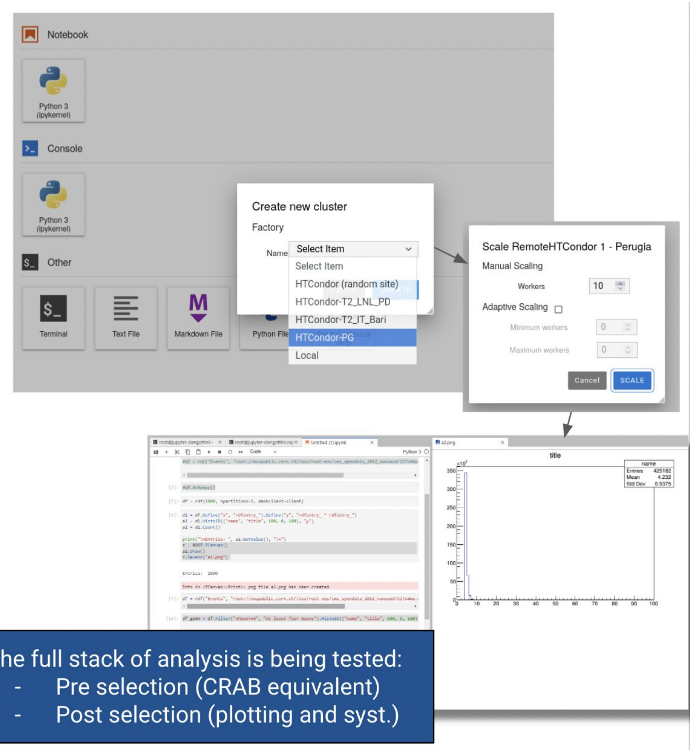
AF @ INFN - Status

INFN developed a testbed system which is now available:
https://cms-it-hub.cloud.cnaf.infn.it

Focusing on RDataFrames, using a VBS analysis to benchmark the system ⇒ to do a first validation of the whole interactive workflow of data. o(10TB) 2017-18

INFN now focusing on:

- Optimized data serving system → caches
  - hierarchical layers vs near-site only
    - Setup in place, just started the benchmarking
  - Studying 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



The full stack of analysis is being tested:
- Pre selection (CRAB equivalent)
- Post selection (plotting and syst.)

AF @ CIEMAT

Background

- CIEMAT contributes to CMS Computing operating the Spanish Tier-1 at PIC (Barcelona), and a Tier-2 site at CIEMAT HQ (Madrid)
    - 15 years of experience, deeply involved in CMS Computing and WLCG
- Considering the data analysis challenges posed by future LHC scenarios on the CIEMAT CMS physicists, we presented an AF@CIEMAT 3-year project proposal to the Spanish Ministry of Science funding call.
    - Funding approved: Sept 2021

Goals for the AF@CIEMAT: Enable CMS scientists exploiting the maximum scientific potential of the data

- Ensure full local access to CMS data (NanoAOD and Ntuples): include locally produced final analysis datasets (Ntuples) by locally slim/skim/expand from NanoAOD samples
- Enable low-latency, high-rate access and interactive exploration of data
- Expand the analysis software palette
- Elasticity of the AF in order to expand when needed, Absorb peaks in demand
    - Expand AF capacity to HPC and Cloud resources (e.g. BSC via HTCondor)

Designing the projected AF at CIEMAT

Key features and resources for the AF:

- Access to data:
  - Improved network capabilities (100 Gbps in LHCONE) to ensure performant data lake access
  - Full copy of the Run3 NanoAOD sample for local access + locally derived data: additional storage capacity HDD (300 TB)
  - Streaming and caching capacity for remote data access (xcache), latency hiding with Data Lake and massively parallel local access, including SSD (200 TB) capacity
- Local processing capacity
  - Deployment of dedicated processing resources, e.g. bulk production of Ntuples with additional CPU (500 CPU cores)
  - HTCondor as manager to AF compute capacity for maximum use efficiency, integrating AF with the overall T1+T2 resources
  - Data reduction, interactive analysis and Machine Learning
    - GPUs for efficient training of ML-based analysis tools (e.g. DNNs for supervised classification)
- Lower usage barrier for CMS physicists at CIEMAT
  - Modern architecture of services and user interfaces
    - Updated authentication procedures
  - Jupyter hub deployment