Analysis Grand Challenge

Alex Held (UW-Madison)
Oksana Shadura (UNL)

IRIS-HEP / Ops Program Analysis Grand Challenge Planning

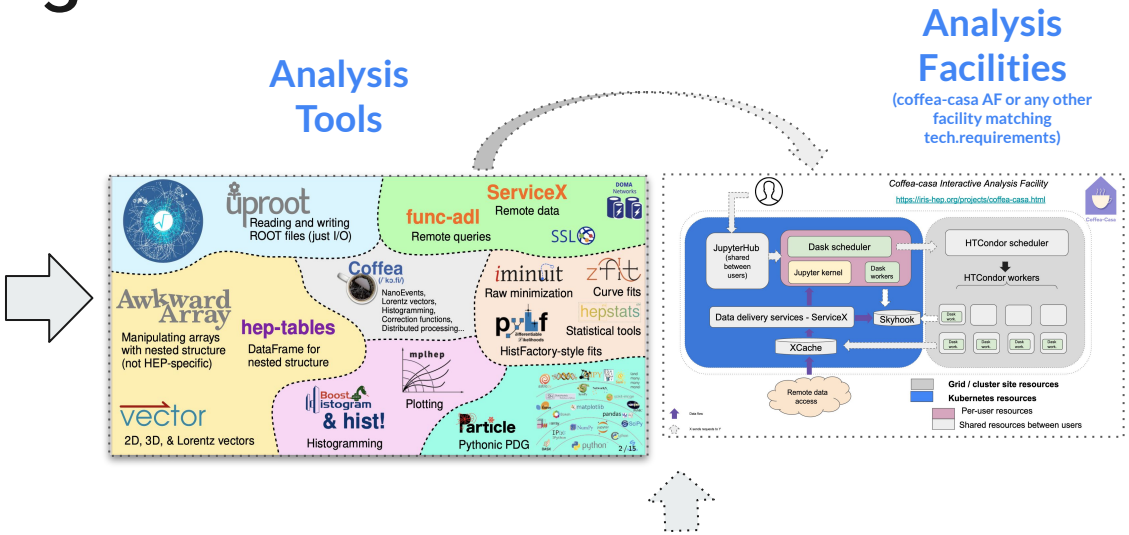
April 5, 2022: https://indico.cern.ch/event/1134876/

This work was supported by the U.S. National Science Foundation (NSF) Cooperative Agreement OAC-1836650 (IRIS-HEP).

Analysis Grand Challenge

Motivation:

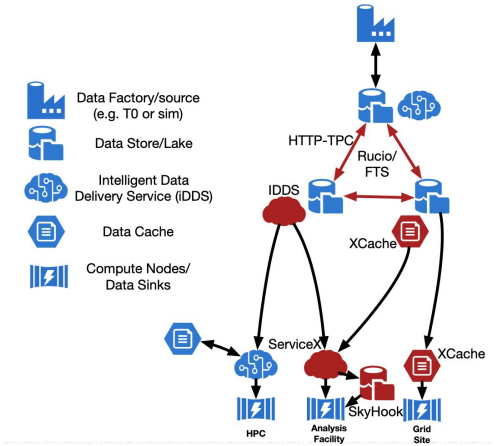
- Allow coping with HL-LHC data sizes by rethinking data pipeline
  - Evaluating the new Python analysis ecosystem and integrating a differentiable analysis pipeline
- Provide flexible, easy-to-use, low latency analysis facilities



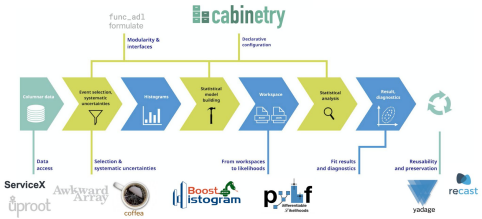
Analysis Grand Challenge will be conducted during **2021–2023**, leaving enough time for tuning software tools and services developed as a part of the IRIS-HEP ecosystem before the start-up of the HL-LHC and *organized together with the US LHC Operations programs, the LHC experiments and other partners.*

AGC is connecting different IRIS-HEP focus areas (partnering with US. CMS/ATLAS Ops programs)

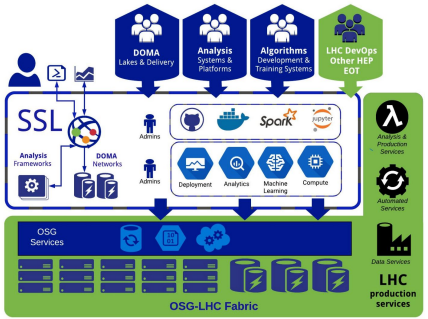
DOMA: *Data delivery*



AS: *tools*



SSL: *deployment techniques and resources*



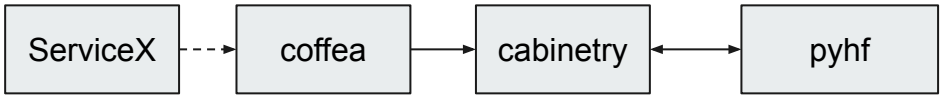
Towards a benchmark analysis

Towards a benchmark analysis: datasets

- Main AGC analysis example will be based on **Run-2 CMS Open Data**
  - Currently using PhysObjectExtractorTool to convert existing miniAOD datasets into ntuple format
    - Found and resolved an issue during validation that broke b-tagging output variables
  - In close contact with CMS to make conversion to nanoAOD format possible in the medium term
    - Will switch to nanoAODs when available to more closely mirror PHYSLITE / nanoAOD workflows
- Prepared image for **converting miniAOD -> custom ntuple**
  - Expect to start large-scale conversion at UNL T3 in the next days, will use ntuples for next AGC workshop
- **Categorizing datasets** in terms of role in AGC demonstrator (AGC repository)
  - In contact with CMS here as well, as this may be of interest more broadly for people doing physics analyses

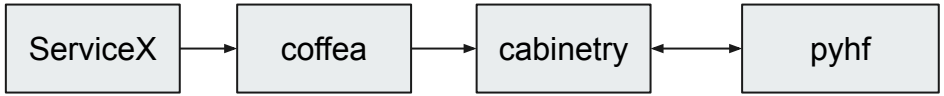
How to participate in AGC

Analysis Grand challenge pipelines



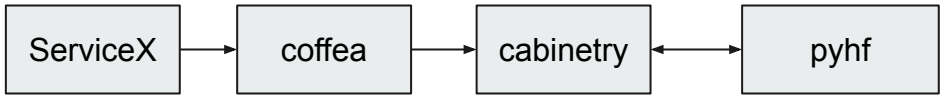
Columns from NanoAOD or flat ntuples from ServiceX or directly through coffea

Generic analysis pipeline based on *Open Data dataset* allowing to easily port AGC to other analysis frameworks



Columns from NanoAOD
and request column
from MiniAOD (only ServiceX)

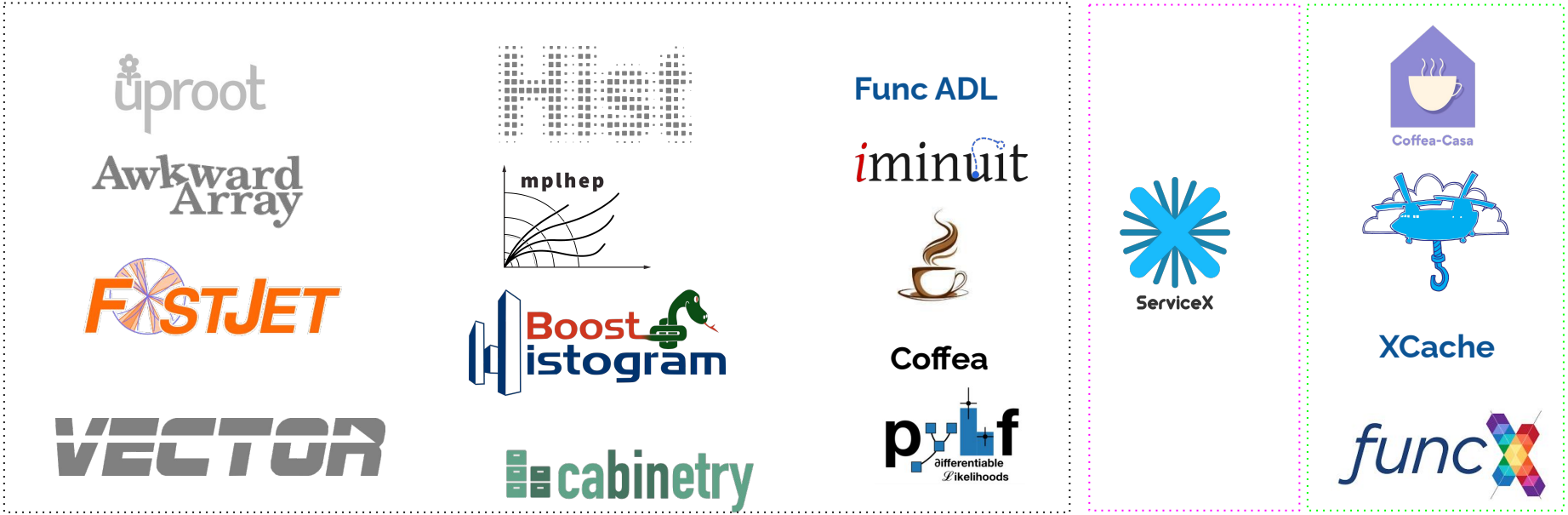
CMS specific analysis pipeline based on *Open Data datasets* and *CMS datasets*



Columns from PHYSLITE
and request column
from PHYS (only ServiceX)

ATLAS specific analysis pipeline based on *ATLAS datasets*

Software and services requirements



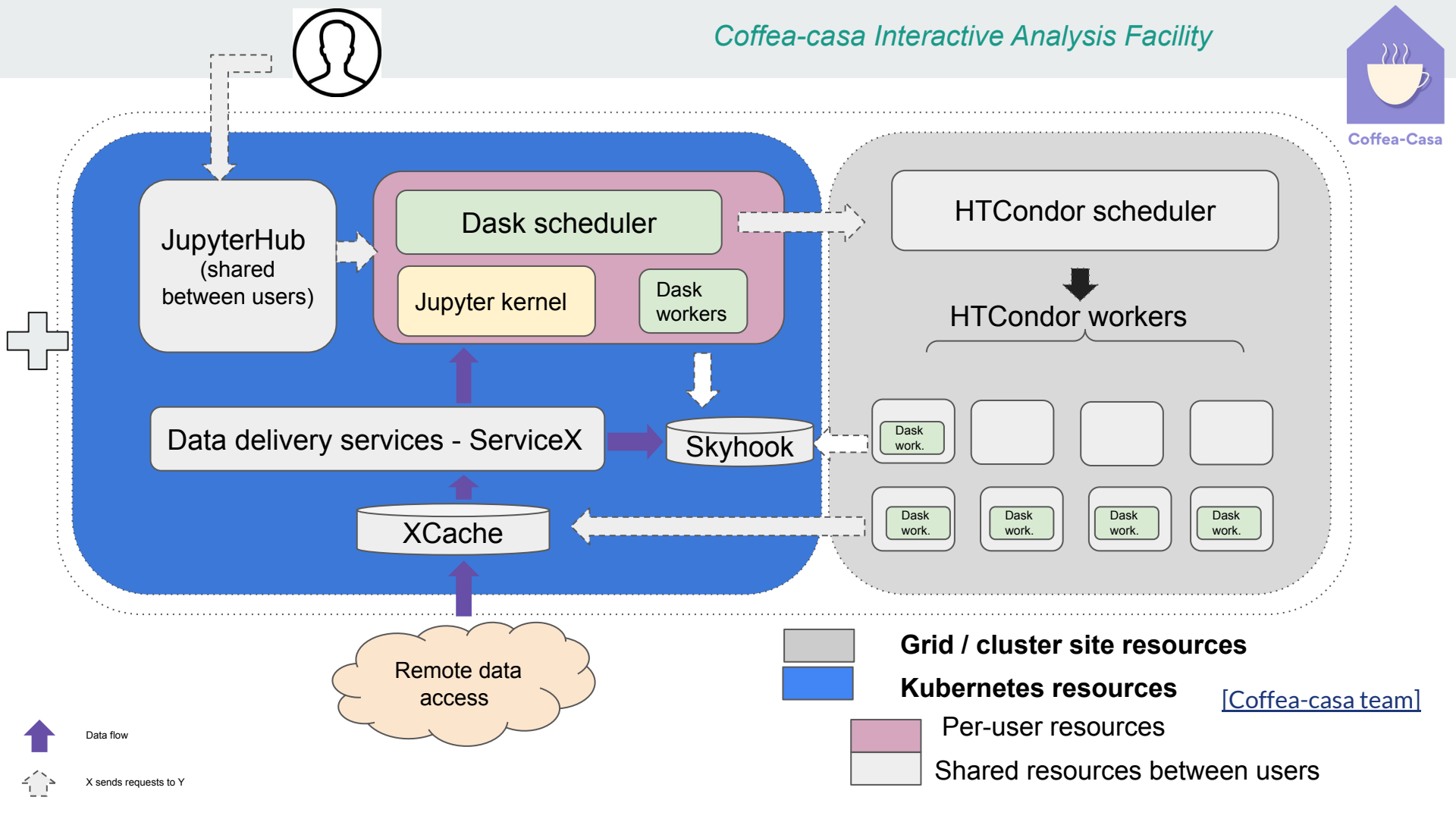
Analysis specific frameworks and packages (available in Docker container)

Data delivery service (k8s)

Optional services (k8s)

Coffea-casa Interactive Analysis Facility

Coffea-Casa



[Coffea-casa team]

How you can use coffea-casa AF experience?

- **Deploy coffea-casa AF at your facility**

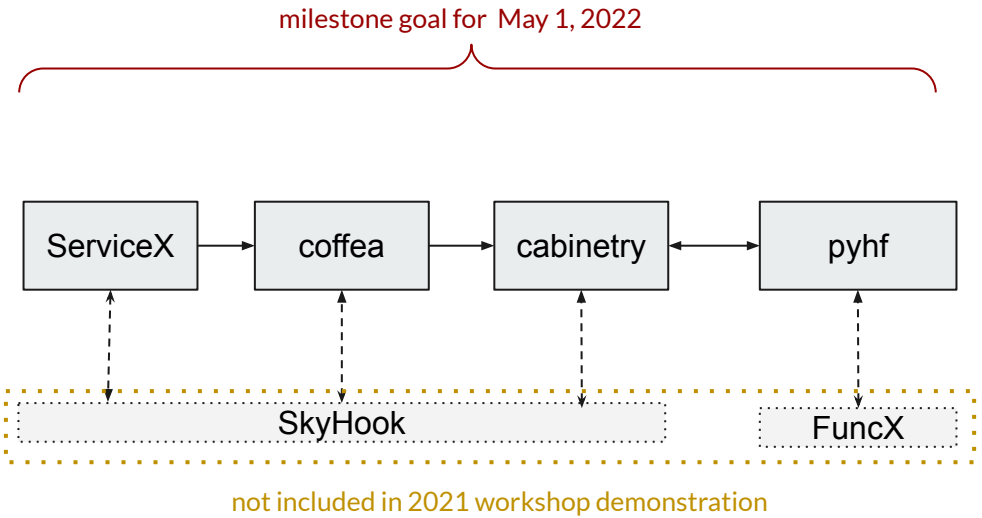
  (you can easily deploy bridging next to your existing facility as it is done @UChicago)

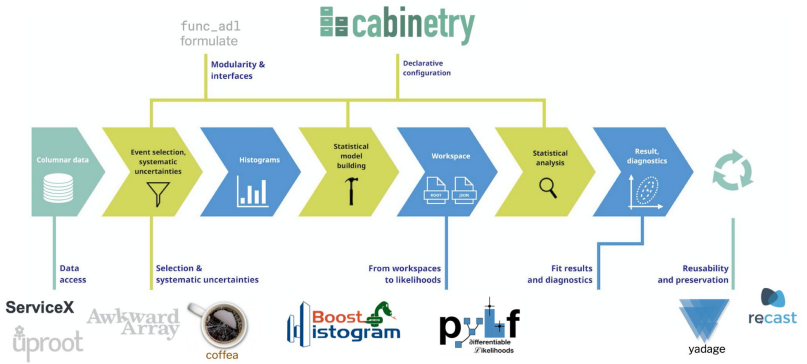
or

- Use coffea-casa docker images
- Condor helm charts (in progress by UChicago)
- XCache helm charts (in progress by FNAL/OSG)
- Authentication configuration recipes (ATLAS / CMS Auth, CILogon)
- Ready Jupyter notebooks, testing various software components and services
- Many other interesting features (just ask us!)

Expanding existing analysis pipeline

- Demonstration of **ServiceX ->Skyhook-> coffea -> cabinetry -> pyhf pipeline on CMS Open Data**





Analysis Facilities participating in AGC (for now)

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”

New facility with ATLAS IAM, setting this up generated valuable feedback for future coffea-casa developments.

Elastic AF @ Fermilab (U.S. CMS)

FNAL team is participating in AGC, helping to test coffea-casa AF and re-using its components at the FNAL facility

U.S. ATLAS facilities
BNL / SLAC are participating in AGC
(we are evaluating the possibility to use coffea-casa experience)

How to participate

- Hoping to run technical tests at all interested sites to evaluate compatibility with AGC plans
  - Keep on collecting various examples testing pieces of the ecosystem: https://github.com/CoffeaTeam/coffea-casa-tutorials
- Proposed first step: evaluate ServiceX setup via the func_adl + ServiceX example notebooks provided
  - Currently uses (public) CMS files, ATLAS to be added as well
  - Happy to help resolve issues, any feedback related to ease of setup would be great as well
    - Any particular things required that would simplify setup? Happy to iterate with you + ServiceX team
- Coffea without and with ServiceX, and the various executors (including dask)
  - Examples also provided in repository
  - Expect that after this stage, remaining required pieces are going to be comparatively simple

Related activities and upcoming events

HSF AF kick-off event

- Kick-off event on March 25 https://indico.cern.ch/event/1132360/
- Idea: introduce activity area & context, very briefly show aspects of developments occurring
- 67 participants
- Recordings are available (link)!
- Hoping that this activity area can become the place for AF-related discussions where the whole community can be involved
    - **Expecting strong participation from IRIS-HEP**

IRIS-HEP AGC Tools Workshop: April 25–26th 2022

- Workshop showing **AGC toolchain at *AF instances (coffea-casa and EAF)***, aimed at PhD / postdoc level
  - https://indico.cern.ch/e/agc-tools-2
  - 2 afternoons CERN time (15:30 - 19:30) on April 25/26
  - Brief introductions to individual packages, notebook talks focusing on interfaces between tools
  - Using **Open Data** examples, also including **ATLAS / CMS** - specific tracks
    - **Interested in additional tracks? Let us know!**

Previous workshop 2021
https://indico.cern.ch/e/agc-tools

- **102 registered** participants
  - *Closed registration because we were not sure if available AF resources would be able to host more participants*
- **81 people connected** to Zoom on first day
- Event recorded & shared on Youtube

Upcoming events

HSF analysis ecosystem workshop: https://indico.cern.ch/event/1125222/

- May 23-25 in Paris (in person)
- Oksana & Alex co-convening **AF track** and **analysis user experience / declarative language track**

Please join us in Paris!

Next IRIS-HEP / Ops program AGC meeting

- **Proposal:**
  - **Skip next month**
    - would be in between AGC workshop (April 25/26) and IRIS-HEP 42 month review (May 16/17)
    - followed by Analysis Ecosystem Workshop (May 23-25)
  - **Meet again on June 7, 2022**, same time slot (9:00 PT / 11:00 CT / 12:00 ET / 18:00 CERN)

Related IRIS-HEP Fellow proposals (see more here)

- Selection process is ongoing, in contact with candidates for all projects

- **Enabling support for MiniAOD Transformer for ServiceX Data Delivery Service**: ServiceX is a distributed, cloud-native application that extracts columnar data from HEP event data and delivers it to an analyst. The func_adl data query language is used to tell ServiceX how to extract the data (the columns, simple cuts, etc.). The func_adl data query language has two backends that are currently part of ServiceX - one based on C++ for ATLAS data and CMS data, and one based on columnar processing using uproot and awkward arrays. The C++ backend currently runs only on the ATLAS binary format (xAOD) and CMS binary format (CMS AOD). This project will modify the C++ backend to also run on CMS MiniAOD binary files (available publicly as a part of Run 2 CMS Opendata release). The MiniAOD transformer is an important ingredient for a physics analysis workflow envisioned in the Analysis Grand Challenge. (Contact(s): Gordon Watts Ben Galewsky Oksana Shadura Alexander Held)

- **Benchmarking of prototype analysis system components**: The Analysis Grand Challenge of IRIS-HEP focuses on performing a high energy physics analysis at scale, including all relevant features encountered by analyzers in this context. It is performed using tools and technologies developed within both IRIS-HEP and the broader community, making use of the Python ecosystem and the required cyberinfrastructure to run at scale. This project will happen after a first preliminary benchmarking has been performed, and it will build on that: the prospective fellow will use pieces of an example physics analysis to study the performance of different system components in more detail. Fellows are expected to have prior Python experience and interest in working with a diverse stack of analysis tools available in the ecosystem. (Contact(s): Oksana Shadura Alexander Held)

- **Metrics to define user activities and engagement on the various coffea-casa Analysis Facility deployments**: coffea-casa is a prototype of analysis facility (AF), which provides services for "low latency columnar analysis", enabling rapid processing of data in a column-wise fashion. These services, based on Dask and Jupyter notebooks, aim to dramatically lower time for analysis and provide an easily-scalable and user-friendly computational environment that will simplify, facilitate, and accelerate the delivery of HEP results. The goal of the project is to define a set of various user engagement metrics, collected from Jupytehub and other AF services, as well from underlying infrastructure (e.g. Kubernetes) and available through Elasticsearch. Expected results are the development of the various metrics, a data collection infrastructure for them, and possibly visualization dashboards. (Contact(s): Brian Bockelman Ken Bloom Oksana Shadura)

Summary

- Made progress with technical aspects of handling CMS Open Data
  - Have a way forward towards benchmarking milestone this summer
  - Now working towards shaping this into a benchmark analysis
- Hope to be able to run technical tests at all available sites
  - Described first step to evaluate compatibility of setups with AGC requirements
- Range of interesting upcoming events
  - AGC workshop & HSF analysis ecosystem workshop

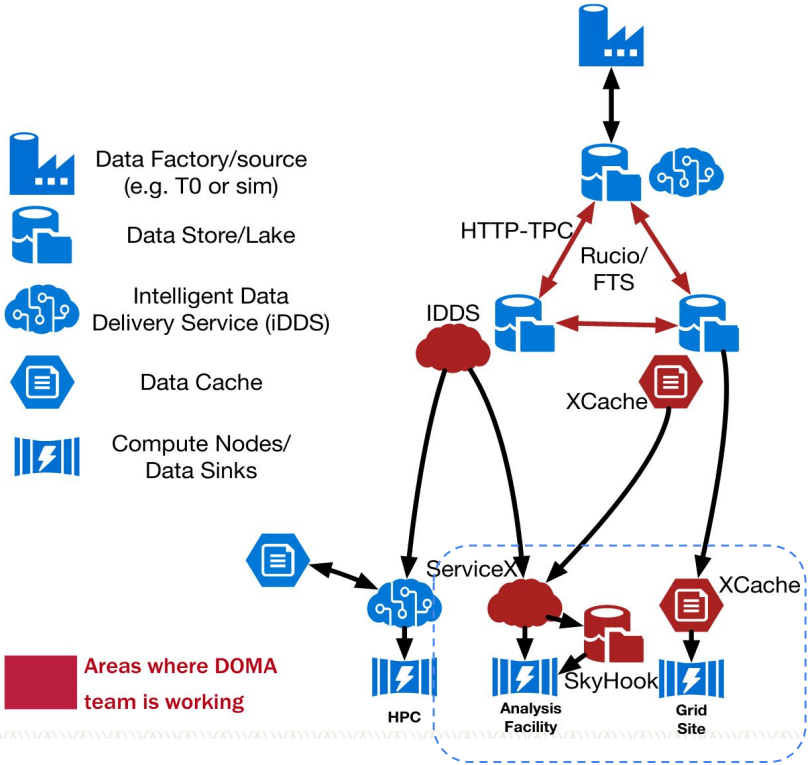
Backup slides

Analysis selection

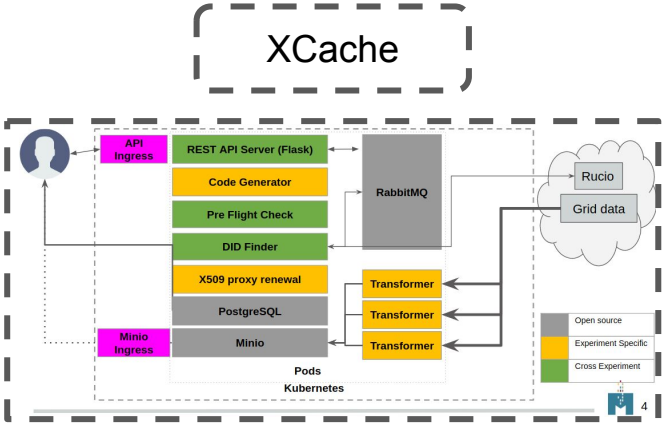
- New CMS Open Data release provides a lot of flexibility
  - All major MC samples available for many analyses
  - Will do a generic search in a ttbar phase space, likely modeled after existing public analysis
    - More familiarity with relevant objects / phase space / systematics / techniques
    - Possible synergies and collaboration with Swift-HEP / University of Manchester
- Developing an analysis from scratch gives us flexibility
  - Can e.g. easily showcase columnar kinematic reconstruction or MVAs, and all other relevant aspects
  - It also allows us to proceed step by step: some aspects of this kind of analysis are quite generic, so can implement overall structure now and follow up with details later when they matter

Expanding analysis pipeline: software components

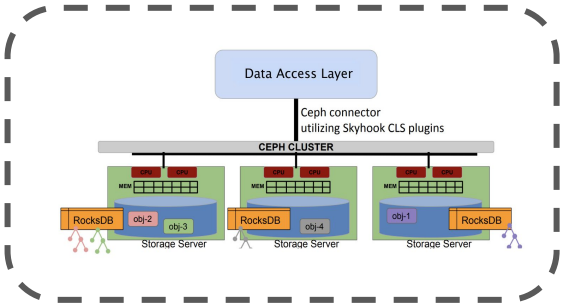
Analysis Facility and Distributed Ecosystem (Data Lakes)



Coffea-casa AF



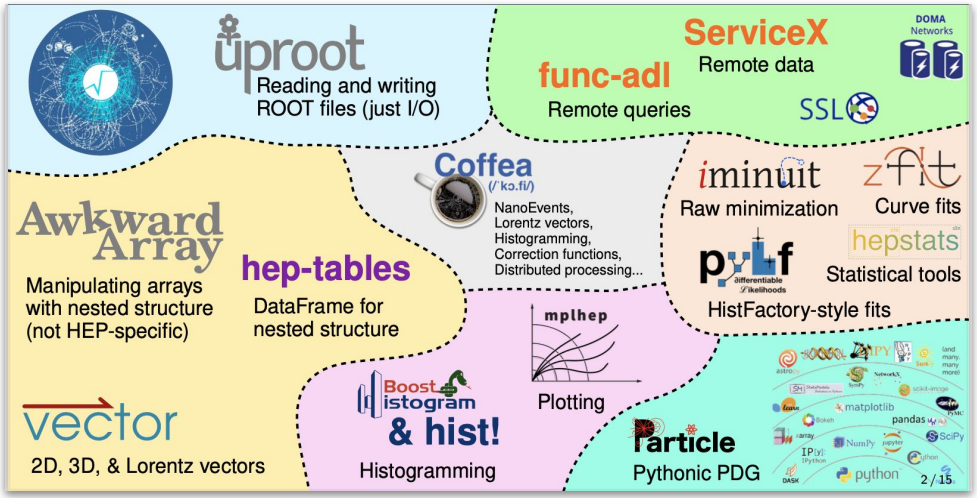
Servicex



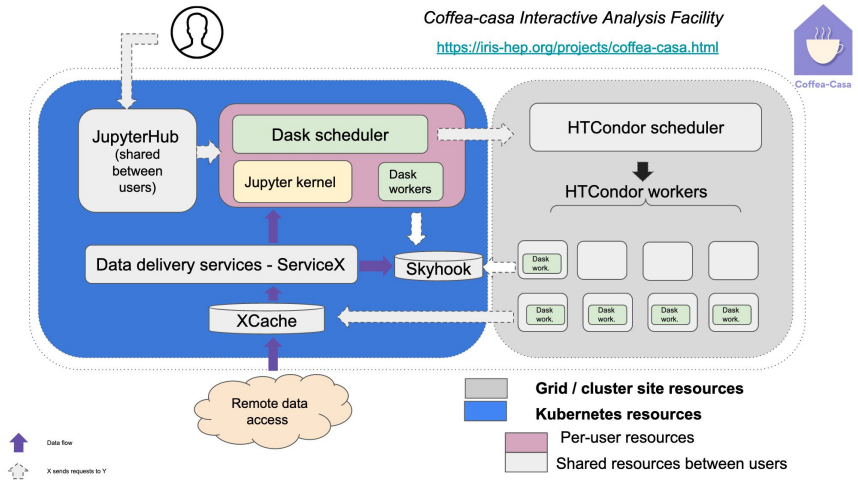
Skyhook

Building blocks used for designing AFs

Analysis Tools



Analysis Facilities

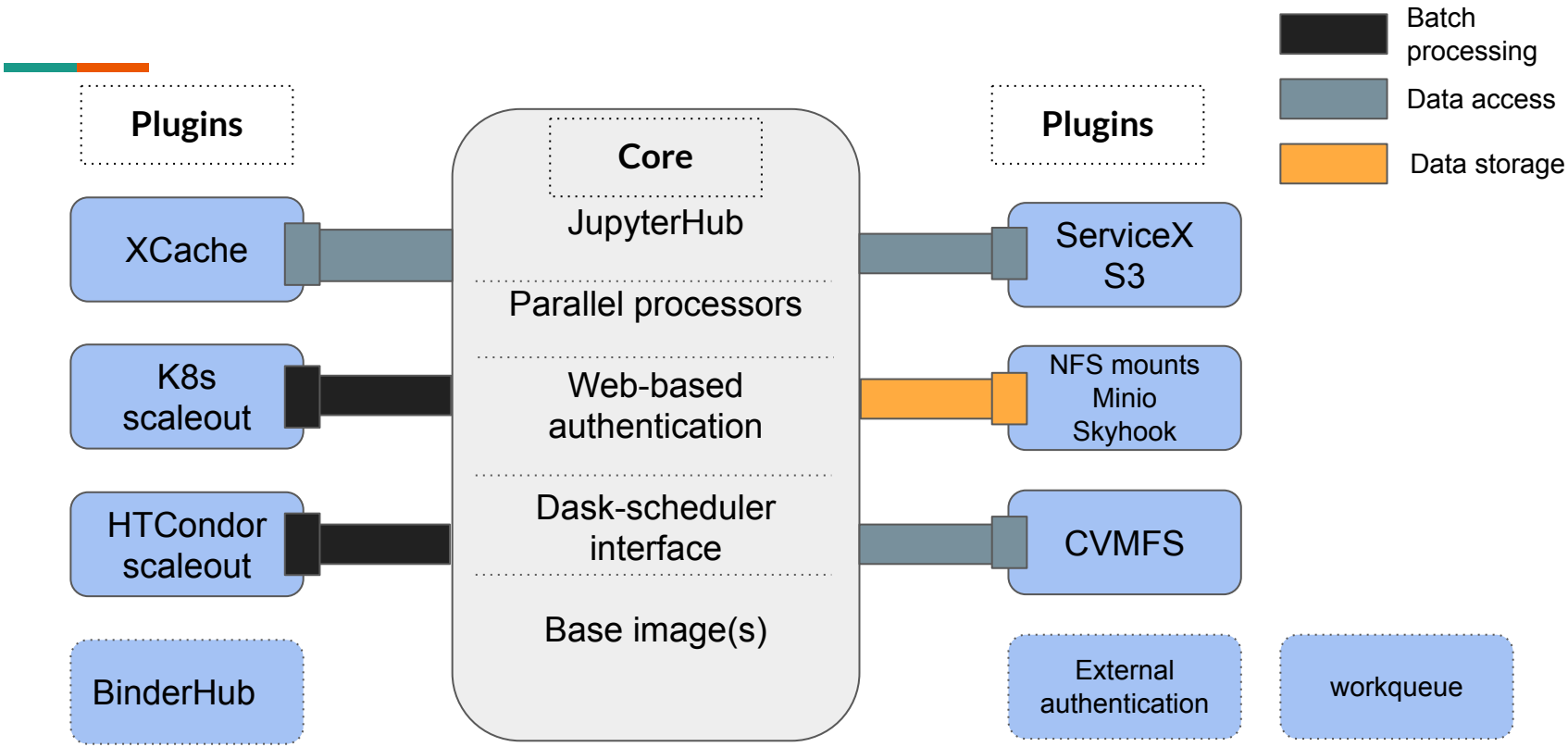


Requirements for AFs

Modern authentication (AIM/OIDC), tokens, macaroons, scitokens	Ongoing R&D on moving to use scitokens natively for AF (write/read)
Efficient data delivery and data management technologies	Ongoing work on integration ServiceX/Skyhook data delivery services
Columnar analysis and support new pythonic ecosystem	
Modern deployment and integration techniques	
Support for object storage	
Efficient data caching solutions	Integrating XCache in analysis pipeline
Easy integration with existing HPC resources	Looking to add support for other batch systems and task managements frameworks

[Coffea-casa team]

Designing AF: components of Coffea-casa Analysis Facility



[Coffea-casa team]

Feedback from SB meeting

- Very useful feedback at SB meeting two weeks ago (https://indico.cern.ch/event/985528/), thank you!
  - AGC efforts seem generally aligned with CMS interests
  - Envisioned workflow is slightly more disconnected from reality of current ATLAS analysis patterns
  - Also looking forward to LHCb / HSF feedback in next SB meeting

Kaushik De's slides

- **Analysis Grand Challenge**
  - Need to work with experiments to bring real users into AGC
    - With well debugged end-to-end tools
    - Goals and plans defined in collaboration with ATLAS
- **Need algorithms and tools (and facilities) to work with common ATLAS data formats - PHYS and PHYSLITE**
  - Need to engage physicists doing analysis with hot data - while ATLAS putting a lot of effort on "Open Data," it has lower physics priority
  - Run 3 data is good for testing "aggressive" scenarios