Analysis Grand Challenge

Alexander Held (University of Wisconsin–Madison)
Oksana Shadura (University Nebraska–Lincoln)

Jan 24, 2023

IRIS-HEP / Ops Program Analysis Grand Challenge Planning
https://indico.cern.ch/event/1243052/

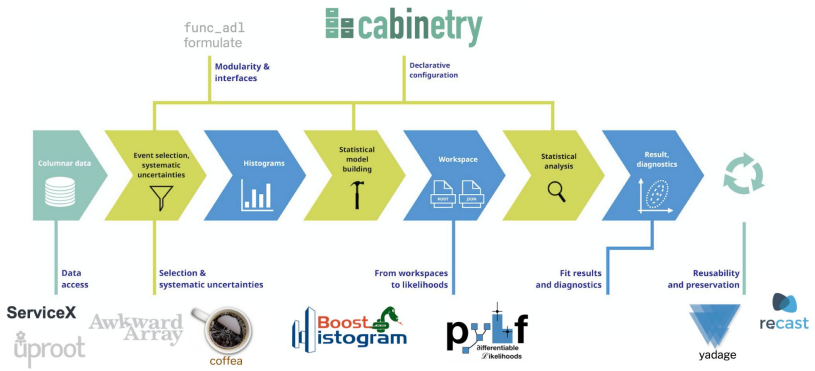
This work was supported by the U.S. National Science Foundation (NSF) Cooperative Agreement OAC-1836650 (IRIS-HEP).

Analysis pipeline

- **Pipeline setup**
  - **ServiceX** delivers columns following declarative **func_adl** request
  - **coffea** orchestrates distributed event processing & histogram production
    - Using **uproot, awkward-array, hist**
  - Visualization with **hist & mplhep**
  - Statistical model construction with **cabinetry** & inference with **pyhf**
- **Everything is openly developed** (IRIS-HEP AGC repository)
  - Including categorization of datasets in terms of role in AGC demonstrator
- Will be executed on various partner facilities: *University Nebraska-Lincoln, UChicago, FNAL, BNL, others*

Other (partial) AGC implementations:

- *ROOT RDF* (Andrii Falko, Enrico Guiraud): andriiknu/RDF/
- *Julia* (Jerry Ling): Moelf/LHC_AGC.jl



An AGC implementation: software stack

Involves large number of packages from IRIS-HEP and partners

uproot

Awkward Array

Func ADL

Coffea

VECTOR

Boost Histogram

mplhep

cabinetry

pyhf differentiable Likelihoods

iminuit

ServiceX

Coffea-Casa

XCache

funcX

Analysis specific frameworks and packages (available in Docker container)

Data delivery service (k8s)

Optional services (k8s)

AGC Plans 2023

AS work items in 2023

Work items

- Testing new **coffea release with awkward-dask**
  - Figuring out new possibilities / workflow / best practices / UX

- **Performance tuning** of AS components

- Extended **analysis task & input size** (more systematics, more histograms, …)
  - including processing implementation improvements (systematics handling, use of *correctionlib*)

ML work items in 2023 (new AGC AS activity)

Work items

- Adding new **AGC pipeline with ML component**
  - This was frequently requested when presenting AGC in the past
  - Including exploring GPU integration at Analysis Facilities into pipeline

- Exploring UX for both ML training and inference
  - MLflow, Triton

SSL work items in 2023

Work items

- Need to **find & resolve performance bottlenecks**
  - Requires close collaboration (e.g. if Dask-related)
- Large scale testing with O(5k+) cores
- Understand pure **I/O throughput** and relate to **hardware specs**

DOMA work items in 2023

Work items

- **Performance tuning** of DOMA related components
  - Understand performance impact of caching
  - Benchmark different data delivery pipelines

- Ensure good integration with different sites

AGC @CHEP 2023

- Preparing **three AGC related talks** (+RDF talk by the ROOT team)

- **Extended, more realistic AGC analysis**
  - ML inference
  - More systematic uncertainties
  - Larger dataset to process (achieved via duplication of inputs)

- **ML training / ML UX** (MLflow, Triton etc.)

- **New developments in coffea-casa AF**
  - Better ServiceX, Triton, MLflow integration

IRIS-HEP Demo Day

AGC Demo Day

Dec 16, 2022

- New “Demo Day” format
  - Short, technical talks
  - Target date for project convergence
  - [Recording on YouTube]

5:00 PM → 5:15 PM **First steps using inference server at coffea-casa facility**
Speaker: Elliott Kauffman (Princeton University (US))
emk_agcdemoday_... | Triton Client Exampl...

5:15 PM → 5:30 PM **ServiceX: ROOT files from uproot transformer**
Speaker: Tal van Daalen (University of Washington (US))
AGC HZZ OpenData...

5:30 PM → 5:45 PM **Data management of HEP data (Apache Iceberg)**
Speaker: Jayjeet Chakraborty
iceberg-spark-demo...

5:45 PM → 6:00 PM **Integrating AGC pipeline at BNL facility**
Speaker: Matthew Feickert (University of Wisconsin Madison (US))
feickert_2022-12-16... | talk: Integrating AG...

6:00 PM → 6:15 PM **Using JWT tokens for XCache at coffea-casa facility**
Speaker: Andrew Wightman (University of Nebraska Lincoln (US))

6:15 PM → 7:15 PM **Discussion**

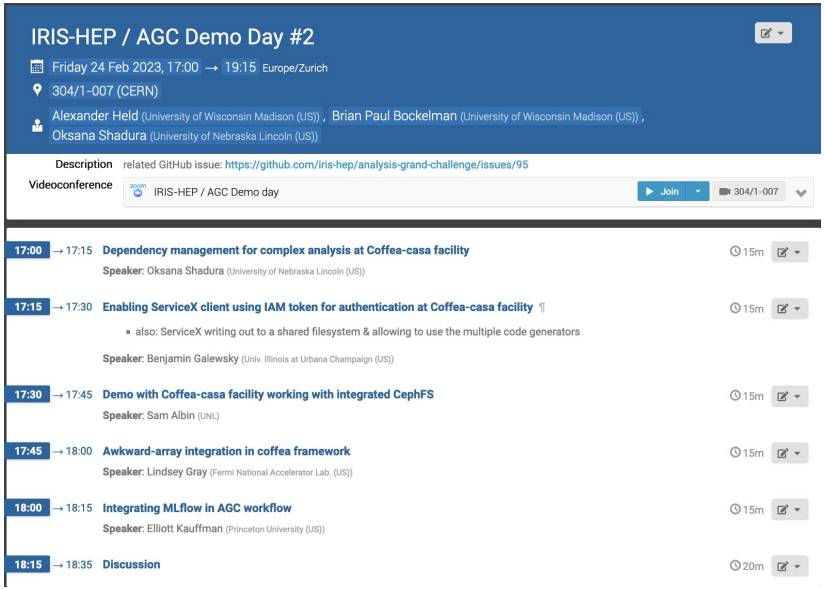
- Variety of topics covered
  - Opportunity to showcase latest developments -> open to contributions!
- Will repeat “Demo Day” format every 2 months

AGC events during IRIS-HEP year 5

Next IRIS-HEP Demo Day

24 Feb 2023, 17:00 CET / 10:00 Central

- New IRIS-HEP demo day is scheduled!
- See [agenda] & [GitHub issue]
- Open meeting, you are **welcome to join**!
  - 304/1-007 booked at CERN



AGC in-person workshop

Timing: around CHEP (early May?)

- Planning a **2/3-day in-person workshop at UW-Madison**

- **Format**
  - Extended “demo day” with longer contributions / discussions
    - Survey AGC deployments

  - Make detailed work plan towards AGC execution event
    - Identify remaining bottlenecks & plan to address them

  - Possibly tutorial-like contributions / community outreach

AGC execution event

- ***AGC Execution Workshop*** in September
- Inviting **everyone** who is interested to **share their setup and to present the results**
  - Interesting combinations of hardware, network site configurations
  - Any type of “combinatorics” of AGC analysis implementation / components setup
  - Performance measurements
  - The chance to publicize your computing resources to physics analysis community :-)
- Not planned as the end of the AGC project

Strategic plan for a 2nd phase of IRIS-HEP

- **Strategic plan v0.95** sent out earlier today to SB/EB -> arXiv soon!

- Includes section with **AGC plans**
  - Expand to two flagship analyses (high volume, high complexity)
    - Further increase scale & complexity (+ ML)
  - Continue annual workshops
  - Demonstrate AOD column joining, differentiable analysis pipeline
  - Many connections to IRIS-HEP focus areas

- **Experiment-specific (ATLAS/CMS) implementations**

Summary

- Outlined **work items for 2023** and events on the way towards **“AGC execution”**
  - IRIS-HEP / AGC Demo Days
  - CHEP 2023 & AGC workshop @ UW-Madison
  - AGC execution event

- Stay in touch: analysis-grand-challenge@iris-hep.org (sign up: google group link), and please also feel free to contact us if you’d like to get involved or have any questions!

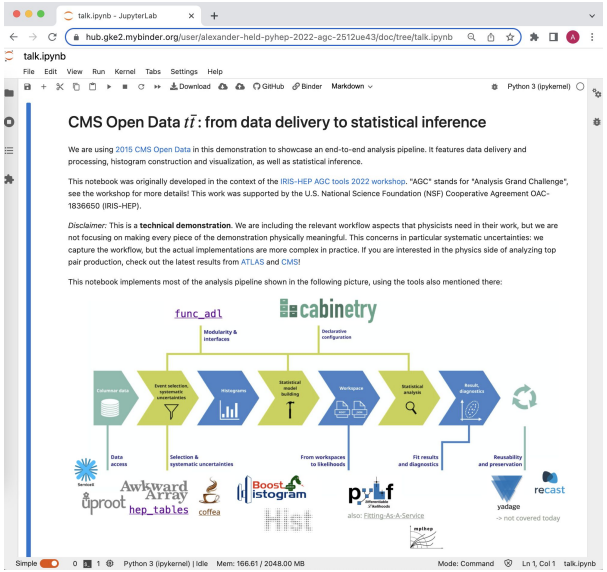
Backup

AGC: give it a try!

We are making it easy for you to try out our setup

- **One click** to get PyHEP notebook in Binder environment
  - Try it out today!

- You can also use the UNL Open Data coffea-casa
  - Or SSL (ATLAS members), or your favorite facility
  - This allows you to scale up (limited on Binder)
  - Everything is available in the AGC repository



AGC: two components

The IRIS-HEP Analysis Grand Challenge (AGC) has two components:

- Defining a **physics analysis task** of realistic HL-LHC scope & scale

- Developing an **analysis pipeline** that implements this task
  - Finding & addressing performance bottlenecks & usability concerns

You can (for example) take take an analysis task and develop a different implementation, take a pipeline and try it with a new analysis task, or adopt task & implementation and run it on your favorite facility.

AGC: how we envisioned it initially

An “integration exercise” for IRIS-HEP

- Demonstrate method for **handling HL-LHC data pipeline requirements**
  - Large data volumes + bookkeeping
  - Handling of different types of systematic uncertainties
  - Use of reduced data formats (PHYSLITE / NanoAOD), aligned with LHC experiments

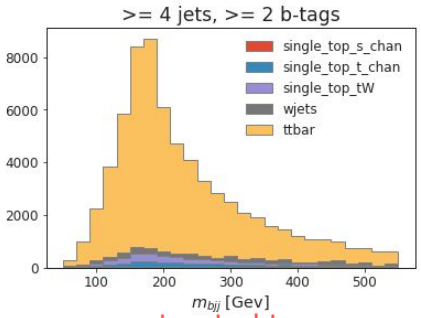
- Aiming for **“interactive analysis”**: turnaround time of ~minutes or less
  - Made possible by highly parallel execution in short bursts, low latency & heavy use of caching

- **Specify all analysis details** to allow for **re-implementations** and re-use for benchmarking

- Execution on **Analysis Facilities**

AGC: analysis task

Community benchmark



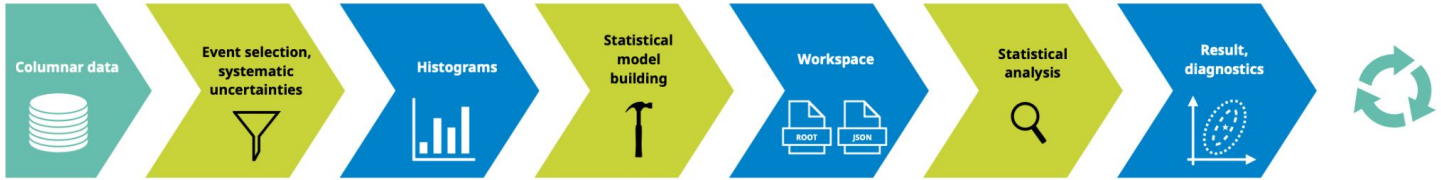
reconstructed top mass

- Analysis task: **ttbar cross-section measurement** in single lepton channel
  - Includes simple top reconstruction
  - Captures relevant workflow aspects and can easily be extended
    - E.g. conversion into a BSM search
  - Analysis task prominently features handling of systematic uncertainties
- Analysis is based on **Run-2 CMS Open Data** (~400 TB of MiniAOD available)
  - Open Data is crucial: everyone can participate
  - Currently using 4 TB of ntuple inputs (pre-converted, ~1B events before cuts)
- Goal of setup is showing **functionality**, not discovering new physics
  - Want to capture workflow; use made-up tools for calibrations & systematic uncertainties

AGC: what we mean by “analysis”

Typical steps in an analysis workflow

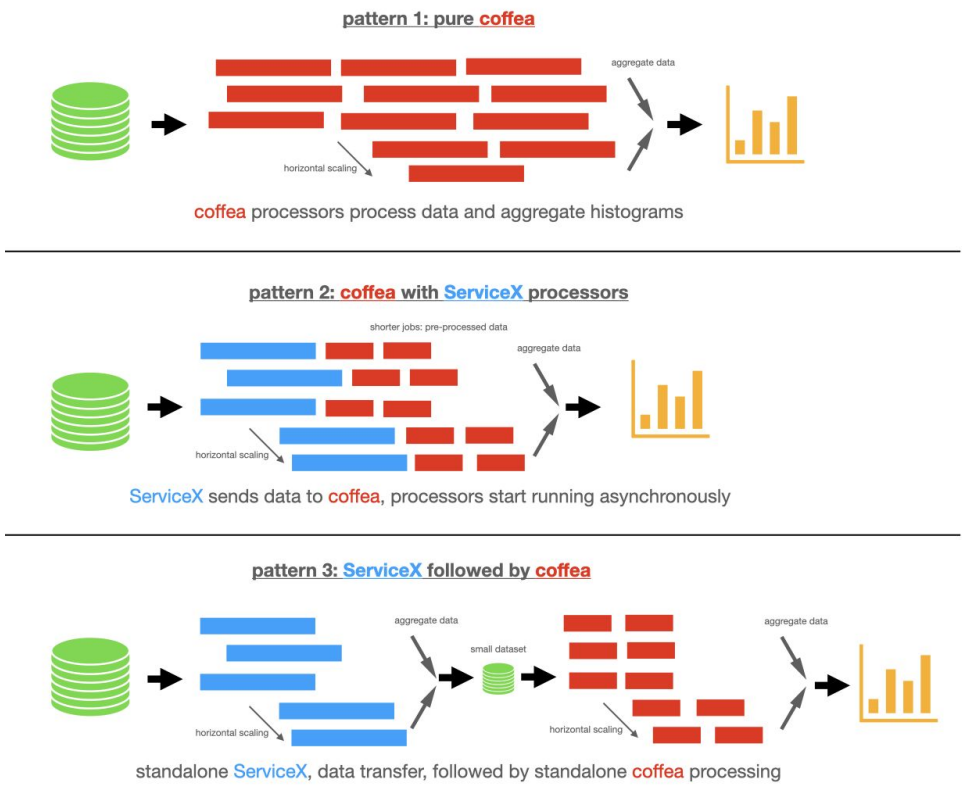
- Start from centrally produced **common data samples**
- Perform all subsequent steps (in a reproducible way)
  - Extract relevant data
  - (Re-) calibrate objects & calculate systematic variations
  - Filter events & calculate observables
  - Histogramming (for binned analyses)
  - Construct statistical model & perform statistical inference
  - Visualize results & provide all relevant information to study analysis details



Adding ServiceX to the mix

Benefits of caching

- Investigating different **data pipelines**
- Data delivered by ServiceX can be **filtered** and is **cached locally**
  - Subsequent runs can hit **(filtered) cache for significant speedup**



What currently runs where?

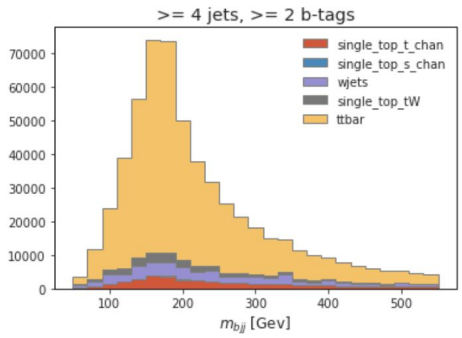
(please help us update the gaps)

	BNL	FNAL	SLAC	UNL	UChicago
basic `coffea` (e.g. `IterativeExecutor`) -> notebook with `USE_DASK = False`	✓	✓	✓	✓	✓
`coffea` scaling (e.g. with Dask) -> notebook with default settings*		✓	✓	✓ (using HTCondor @ Tier2, planning to switch to k8s)	✓
standalone `ServiceX` -> notebook (no configuration)	✓	✓		✓	✓
`ServiceX`+`coffea`+scaling -> notebook with `PIPELINE = "servicex_processor"`				✓	✓
`XCache` support	✓	✓ (some performance caveats, to be understood)	✓	✓	✓

* may need site-dependent Dask cluster configuration, see implementation, please get in touch in case of questions

AGC implementations

Community effort



- *coffea*: iris-hep/analysis-grand-challenge/
- *ROOT RDF* (Andrii Falko, Enrico Guiraud): andriiknu/RDF/
- *Julia* (Jerry Ling): Moelf/LHC_AGC.jl