Analysis Grand Challenge

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

July 11, 2023

IRIS-HEP / Ops Program Analysis Grand Challenge Planning
https://indico.cern.ch/event/1303818/

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

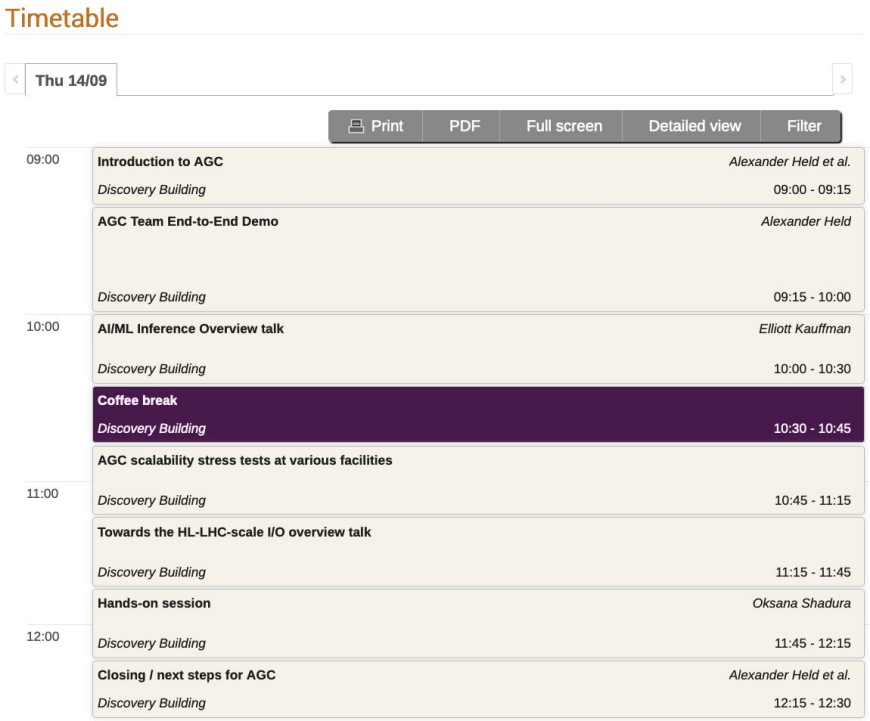
AGC demonstration event

- We are organizing an **AGC demonstration event** in September
  - Hybrid event: in-person and over Zoom - September 14th @ University Wisconsin-Madison

- **AGC demonstration** will be a half-day event (morning US time)
  - **Inviting everyone (the facilities and also other AGC implementations) to record their demo setups and we will upload them to Indico event web-page**
    - Interesting combinations of hardware, network site configurations and other settings
    - Any type of “combinatorics” of AGC implementations / components setup
    - Can include performance measurements
  - Chance to showcase your computing resources to physics analysis community :-)

AGC demonstration event

- **AGC demonstration event**
  - 2 large demo and 1 combined slides decks *with the results from the different facilities*
  - Bonus: hands-on session, held on one of analysis facilities



AGC versions

Description of versioning scheme: documentation

- The **AGC analysis task** evolves via **major versions**
  - **v0:** custom ntuple inputs -> superseded (do not use this anymore)
  - **v1:** NanoAOD inputs -> baseline to use
  - **v2:** machine learning, more systematic uncertainties -> heavier CPU & I/O requirements (not yet fully finalized)
- Implementations of the AGC task can tag improvements via minor / patch versions
  - The reference implementation of AGC v1 is v1.4.0 in our repository

We recommend to use v1.4.0 tag for the measurements

AGC pipeline configuration for execution event

What we would like to see in contributions

- **Baseline:** full AGC pipeline with distribution via **Dask** (`USE_DASK = True`)
  - Can also be ROOT version with distributed RDF
- **Advanced:** demonstrate pipeline with **ServiceX** (optional)
  - USE_SERVICEX = True
  - Employ your XCache if available and compare performance
- **Advanced:** include **additional ML functionality** (optional, AGC v2)
  - Training: run `jetassignment_training` & reproduce models, more advanced: `USE_MLFLOW = TRUE`
  - Inference: `USE_TRITON = TRUE`

Options on this slide refer to the ttbar_analysis_pipeline.ipynb implementation.

Advanced performance studies

Additional aspects available for studies

- Execution event target for facilities: demonstrate baseline setup

- **Additional functionality** provided for more studies
  - Variations in I/O requirements for benchmarking (`IO_FILE_PERCENT`)
  - Turn on/off ML inference & columnar calculations (`USE_INFERENCE`, `DISABLE_PROCESSING`)
    - inference requires current HEAD or upcoming v2 tag

AGC execution event

Metrics that might be of interest

- **Goal** of execution event: **showcase functionality**, but welcome to use existing setups for more beyond that!
- **Standard metrics** (in the many configurations outlined previously)
  - Data volume processed (per time and core)
  - Event processing rate per core
  - Scheduling efficiency à la David Koch's slides, page 12
- **Data pipeline comparisons**: ratio of ServiceX+coffea and coffea (directly reading original input) runtimes
  - Assumption: input data sitting in XCache
  - Goals: no substantial slowdown of initial execution of ServiceX+coffea setup, demonstrate significant speedup in repeated runs (hitting ServiceX cache)
- **Additional points of interest**
  - Capture multi-user setups: run multiple AGC pipelines in parallel
  - Evaluate UX: how much manual intervention is needed (e.g. copying & settings tokens)

Summer fellow projects

IRIS-HEP and US CMS PURSUE fellows

- **AGC with RDF:** Andrii Falko, co-supervised by Enrico Guiraud and Alex
  - Provide implementation for v1 task, extend to include ML aspects
  - https://github.com/root-project/analysis-grand-challenge

- **AGC in Julia:** Atell Krasnopolski, co-supervised by Jerry Ling and Alex
  - Develop pure Julia implementation of v1 task

- **AGC on CMS Run-2 data:** Christina Mondelli, supervised by Andrew Wightman
  - Scale up implementation to use (internal) CMS Run-2 data

Summary

- **AGC demonstration event** happening in September
- Please prepare material: demos, documentation, measurements
    - Anything that helps showcase your facility in the AGC context
- If you have any questions, please feel free to get in contact directly or via analysis-grand-challenge@iris-hep.org (sign up: google group link)