200 Gbps Analysis Challenge

Brian Bockelman (Morgridge Institute)
Alexander Held (University of Wisconsin–Madison)
Oksana Shadura (University Nebraska–Lincoln)

IRIS-HEP Steering Board Meeting #21

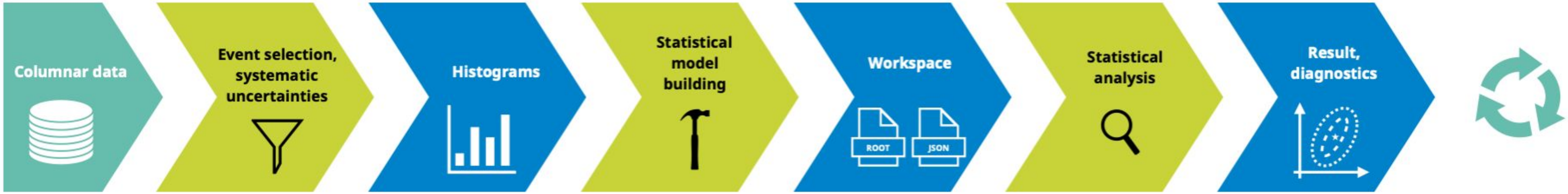
https://indico.cern.ch/event/1388604/

This work was supported by the U.S. National Science Foundation (NSF) cooperative agreements OAC-1836650 and PHY-2323298 (IRIS-HEP).

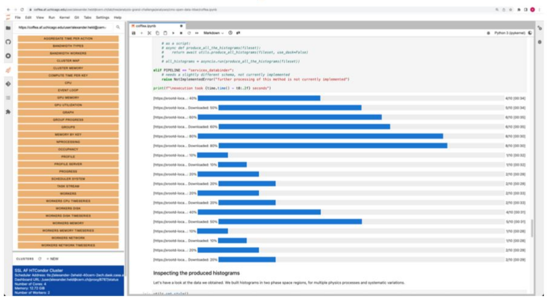
Analysis Grand Challenge (AGC):
execute series of increasingly realistic exercises toward HL-LHC

The AGC is about executing an analysis to test workflows designed for the HL-LHC.
This includes:

- **columnar data extraction** from large datasets,
- **data processing** (event filtering, construction of observables, evaluation of systematic uncertainties) into histograms,
- **statistical model construction and statistical inference**,
- relevant **visualizations** for these steps

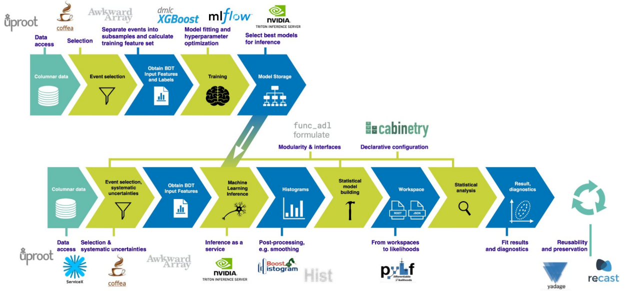


AGC: some previous work (1)



The **AGC project started** properly in the **autumn of 2021**

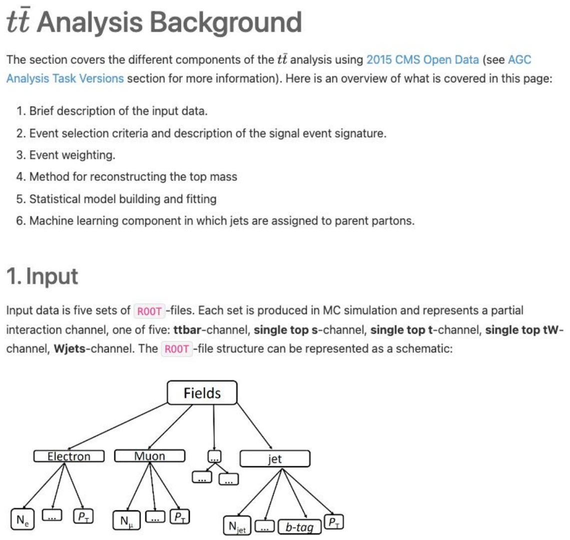
- Physics task definition (multiple versions)
  - Capturing physics analysis requirements matching practical needs of physicists
  - Using CMS Open Data (reformatted to 2 TB of NanoAODs)
- [IRIS-HEP AGC reference pipeline implementation]
  - Analysis implementation based on IRIS-HEP stack of tools
  - Connecting many projects and developers
  - Cycle: iterating with experts and improving implementation



AGC: some previous work (2)

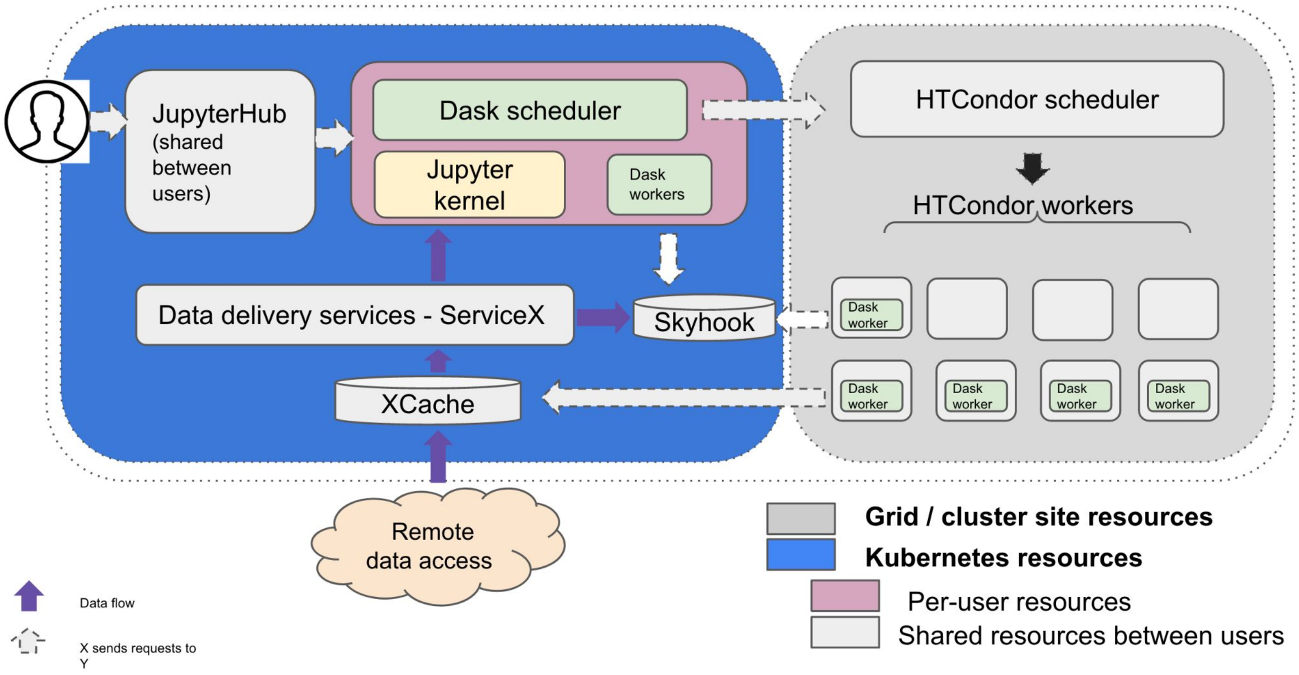
- Developed **website as central resource**: https://agc.readthedocs.io/en/latest/
  - Work based on IRIS-HEP fellow project — AGC hosted and benefited from many great IRIS-HEP fellows

Analysis task details to allow for re-implementations



Analysis Grand Challenge (AGC): preparing next generation of Analysis Facilities

Coffea-casa Analysis Facility is providing **AGC execution environment** to explore analysis workflows at scale



IRIS-HEP v2: yearly benchmarking exercises

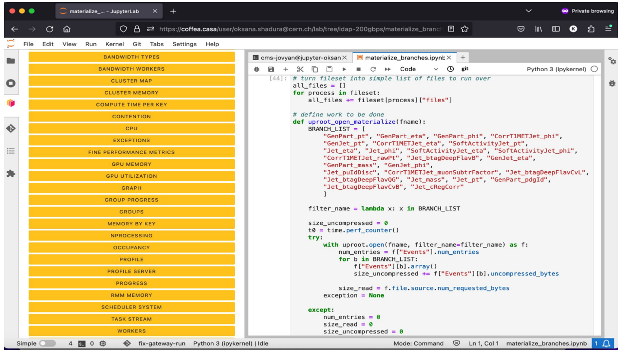
- 2024 - **200 Gbps Challenge**
- End of 2024 - Test **analysis pipeline at scale** with 30 simultaneous users
- 2025-2028 - **benchmark iterative scaling to HL-LHC needs** with AGC

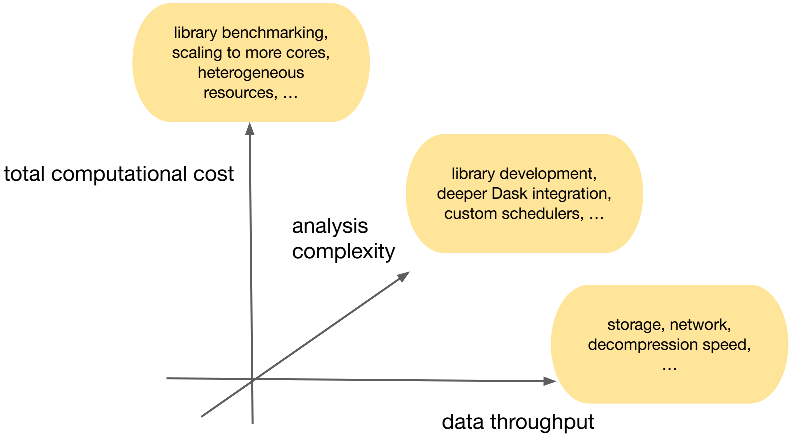
getting ready for HL-LHC

Timeline	Fraction of HL-LHC dataset processed in 1h
2025	20% (40 TB)
2026	50% (100 TB)
2027	75 % (150 TB)
2028	100% (200 TB)

The 200 Gbps setup

- **Simplified analysis setup** compared to what is done in AGC
  - Analysis “straight from NanoAOD / PHYSLITE” with all required computations on-the-fly
  - Goal to gradually add back functionality to match AGC
- **Limited agreement** in the broader field about how HL-LHC analysis will look like
  - Therefore: factorize challenges wherever possible -> **project focused on data throughput**





Defining the task

- **"HL-LHC scale"**: process 25% of 180 TB dataset in 30 min
  - This requires 200 Gbps
  - For a 2 kB event size -> 90 B events, analyze at 50 MHz
  - With 25 kHz / core -> need 2000 cores (12.5 MB/s per core)
- **Two setups**: ATLAS (at UChicago) and CMS (at UNL)
  - CMS: analyze Run-3 NanoAOD
  - ATLAS: analyze Run-2 PHYSLITE
- Very similar task between both setups, but **important differences** when comparing
  - Smaller per-event size in NanoAOD, (currently) different default compression algorithms, different object types
  - Different production facilities

The 200 Gbps NanoAOD setup

Uproot + Coffea notebooks https://github.com/iris-hep/idap-200gbps and using CMS Run2 NanoAOD (~100TB)

- **Read data from XCache on the Coffea-Casa facility at the Nebraska Tier-2 (running in Kubernetes).**
- **Expand scale out into the site HTCondor and Kubernetes cluster.**
- Tasks processed in TaskVine & Dask backends (dask-jobqueue vs dask-gateway).

Notes on realism:

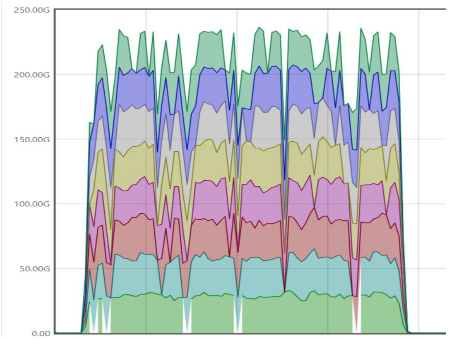
- Real XCache setup (works now in production at facility). Token-based auth using the IAM service at CERN.
- LZMA decompression dominates analysis time (~70%). To hit our target 25KHz-per-core processing rate, we recompressed the NANOAOD using ZSTD. About 20% larger than the original dataset, ~2.5x faster.
- We scale-out to HTCondor with pre-created workers, no autoscaling.

Uproot results, NanoAOD

From the statistics in the notebook:

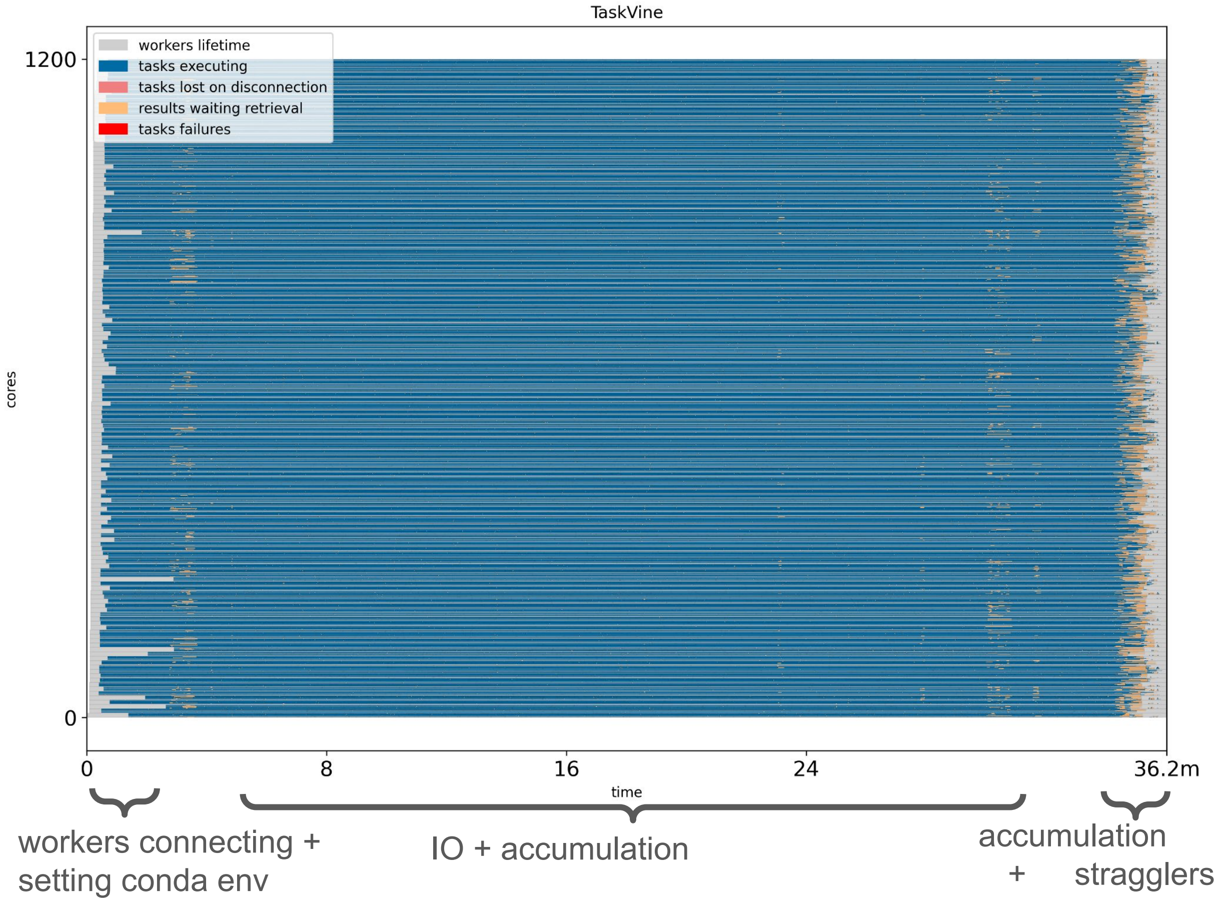
- Data read (compressed): 58.33TB
- Average data rate: 221 Gbps
- Peak data rate: 240 Gbps
- Processed 40,276,003,047 events total
- Files processed: 63,762 (17 failed)

Network rates from XCache storage:

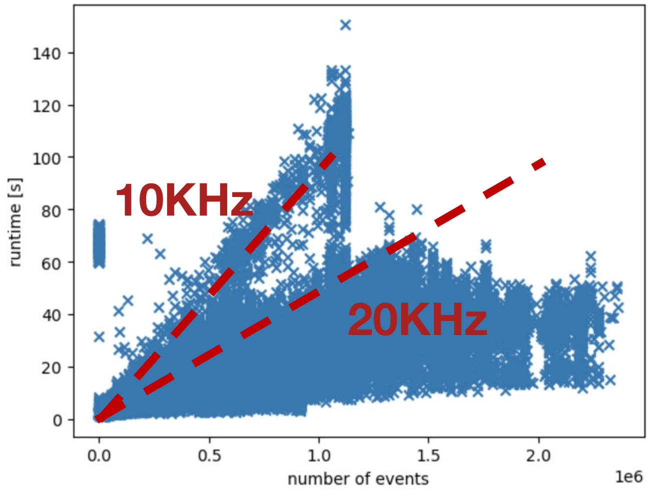


Rates from different, but representative run)

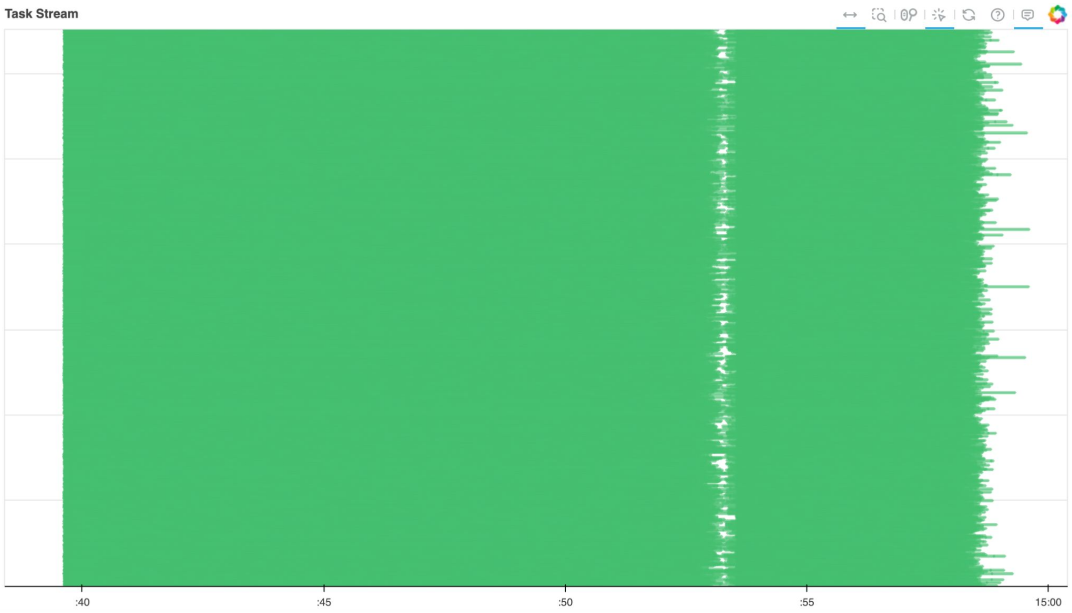
1200 cores across 150 8-core workers

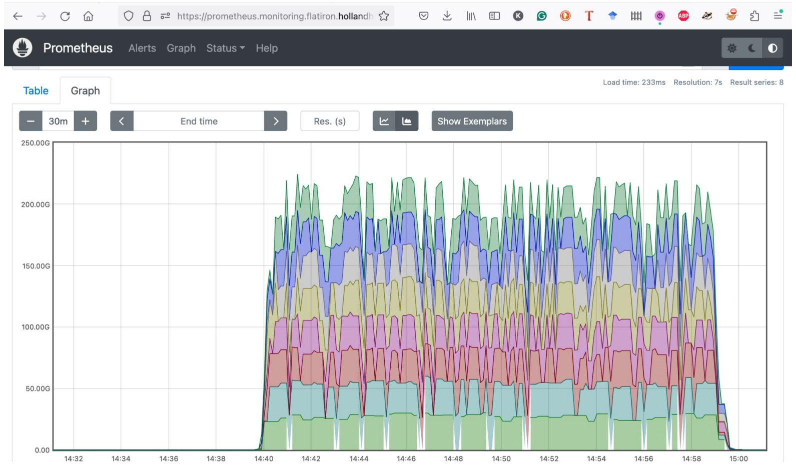


Runtime vs # Events as seen by xcache



Dask task stream and xcache stats over the same run





More results coming soon for upcoming CHEP 2024 conference

The 200 Gbps PHYSLITE setup

Uproot + Coffea notebooks https://github.com/iris-hep/idap-200gbps-atlas and using PHYSLITE (~190TB)

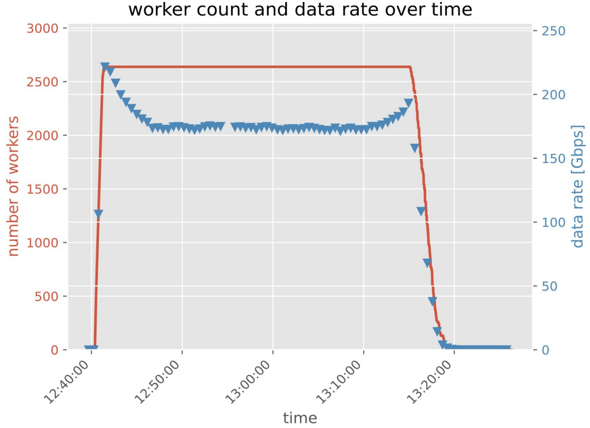
- At UChicago, also processed ATLAS PHYSLITE files directly in Python.
- 218k files, 190TB data, 23B events, ~8kHz/core
- Goal was using coffea 2024, dask-awkward, uproot; ended up using direct processing in uproot.

Highlights:

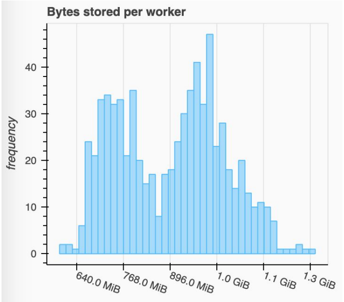
- Scaled Dask up to around 2.5k cores
- 200Gbps throughput sustained in network monitoring; slightly less in ‘effective bytes’ into Dask.

Biggest challenge has been understanding memory usage; significant difference between “uproot only” and the full Coffea 2024 (the same situation was observed at CMS setup).

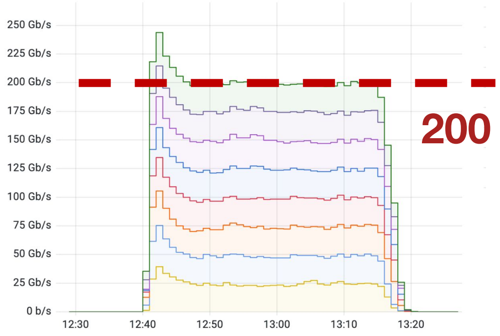
Scaling to HL-LHC: 200 Gbps PHYSLITE setup



memory profile across workers



Network monitoring

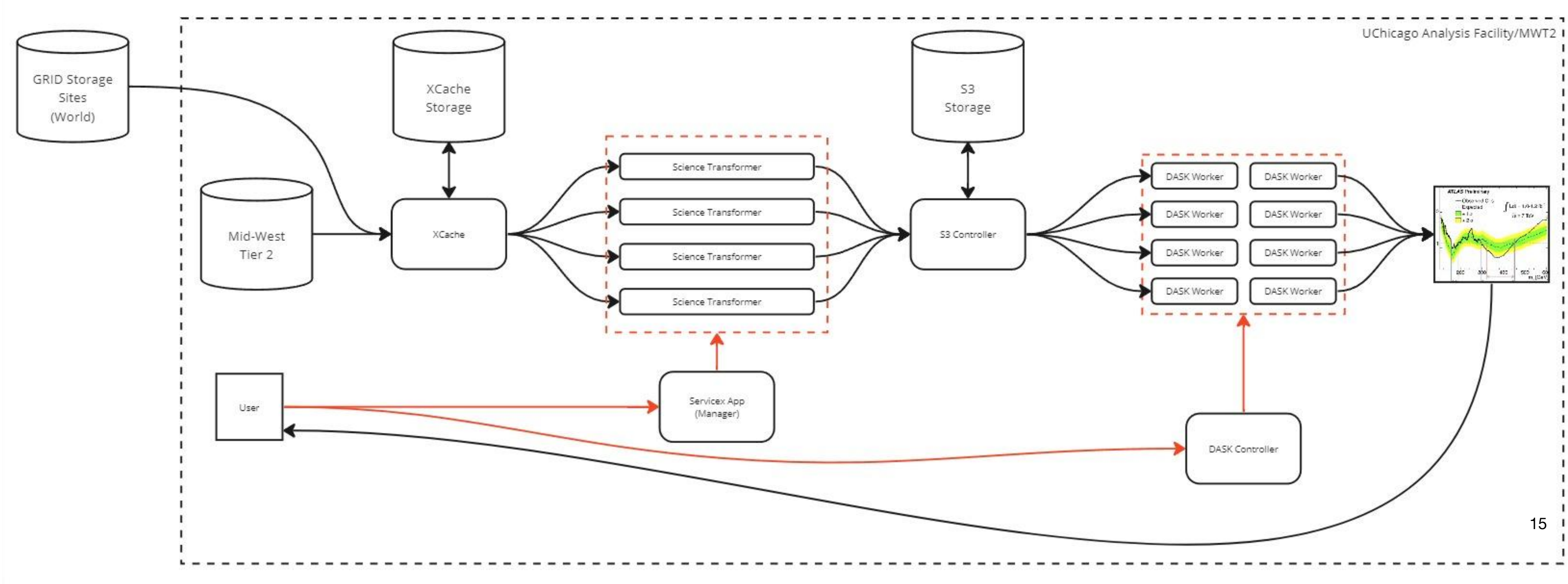


200 Gbps

More results coming soon for upcoming CHEP 2024 conference

Using ServiceX: data flow

- Two-stage process, various unique features and performance-relevant aspects to consider
  - e.g. data written out to S3 must be compressed (can be CPU-intensive)

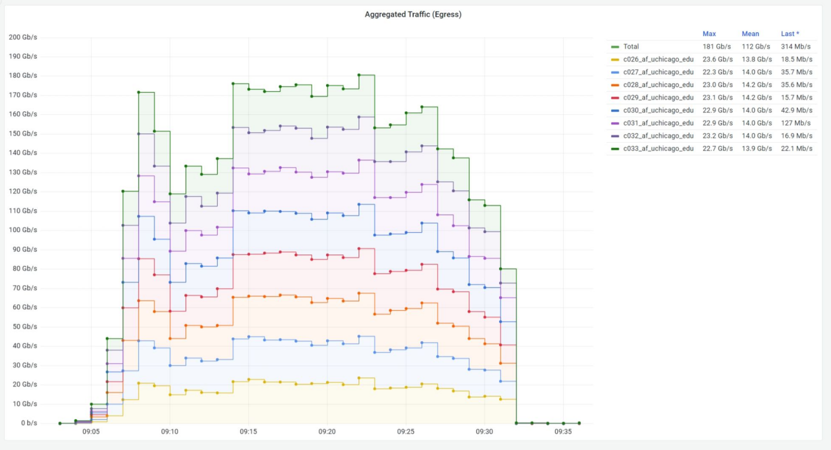


ServiceX Results

From Brian Bockelman talk “IRIS-HEP 200Gbps challenge”
HSF/WLCG workshop

Using ServiceX data extraction and delivery delivery service as part of pipeline:

- To reduce the overhead of small datasets, we ran on a subset that consisted of the bulk of the data.
- Highlight run:
  - 4 Datasets
  - 146TB total
  - 19,074,862,754 Events
  - 170Gbps
  - Limited to 1,000 pods.
  - Time: 32:28
  - Event Rate: 9,787 kHz



(Some of the) lessons learned

- Very **successful exercise format**: huge amount of progress and activity within 7 weeks
- **Started slightly too big** in scope with more complex task graph using coffea 2024
  - Faced some challenges with memory use and scaling to all available resources: following up to improve this now
  - Instead went back to a simpler uproot-based setup for this challenge
- **PHYSLITE:** rate-per-branch can vary a lot (cost of decompression, interpretation), some branches are not (yet) readable with uproot
- **NanoAOD:** very large effect of compression algorithm: switching from LZMA to ZSTD brought 2.5x rate improvement
- **Scaling Dask to 2k+ workers** generally works fine, need more testing combining large numbers of workers and very complex graphs
- Good performance observed also with **TaskVine** as alternative scheduler for graphs
- Scale of challenge allowed to identify **new bottlenecks** (many of which have already been fixed), e.g. object store needing to scale to ServiceX output

Summary

- The 200 Gbps project brought together a broad community with a shared ambitious goal
- It was a success! Demonstrated feasibility of the desired data rate
- Many lessons learned and follow-up work identified
- This is a checkpoint on our way towards HL-LHC

More information: WLCG / HSF workshop talk, iris-hep/idap-200gbps, iris-hep/idap-200gbps-atlas

200 Gbps related slides summarize a large body of work across IRIS-HEP and USCMS/USATLAS. Thank you to everyone for your contributions!

- Fermilab: Lindsey Gray, Nick Smith
- Morgridge: Brian Bockelman
- Notre Dame: Ben Tovar
- Princeton: Jim Pivarski, David Lange
- UChicago: Lincoln Bryant , Rob Gardner, Fengping Hu, David Jordan, Judith Stephen , Ilija Vukotic
- National Center for Supercomputing Applications: Ben Galewsky
- U. Nebraska: Sam Albin, Garhan Attebury, Carl Lundstedt, Ken Bloom, Oksana Shadura, John Thiltges, Derek Weitzel, Andrew Wightman
- UT-Austin: KyungEon Choi, Peter Onyisi
- U. Washington: Gordon Watts
- U. Wisconsin: Alex Held, Matthew Feickert

Thank you for your attention!

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)