HL-LHC Computing R&D in US ATLAS

Torre Wenaus (BNL)

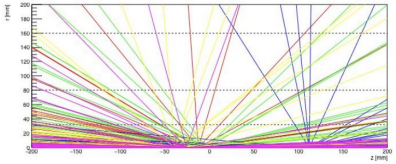
IRIS-HEP Workshop on HL-LHC Computing R&D
Catholic University of America
October 23 2019

BROOKHAVEN

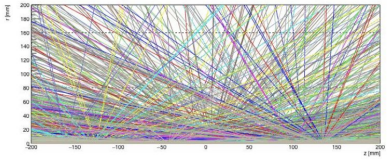
High Luminosity LHC (HL-LHC)

Starting in 2026, the LHC enters a new era

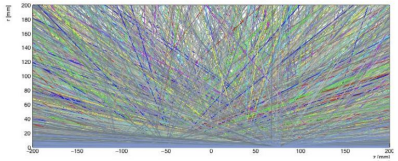
- 5-7x LHC original design luminosity, pileup up to 200 interactions/crossing
- Extensively upgraded, more complex detectors: 4-5x increase in event size
- Upgraded trigger system: up to 10x increase in event rate
- **A 10+ year program of precision and discovery physics with a ten-fold increase in integrated luminosity**



2010, <μ>=5



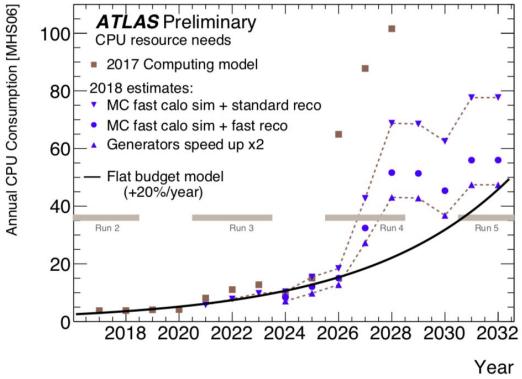
2018, <μ>=40

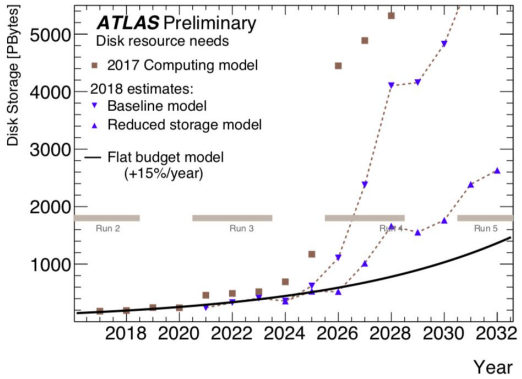


2026, <mu>=200

S&C challenges for the HL-LHC

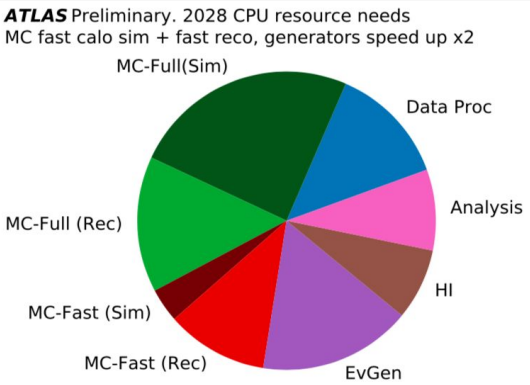
- CPU and disk fall well short of estimated requirements
  - Still true assuming successful computing model evolution that itself requires substantial development work
- Estimates of available resources, and the throughput we gain from them, also carry substantial uncertainty
  - Is 15-20% flat-budget capacity growth realistic?
  - Will our payloads perform efficiently on new architectures?
- To succeed we need substantial investments in creative, long-view software R&D
  - Building on and complementing computing evolution work already in the pipeline
- Forward looking projects like IRIS-HEP -- and IRIS-HEP in particular -- are essential for this
  - **Thank you for your support!**

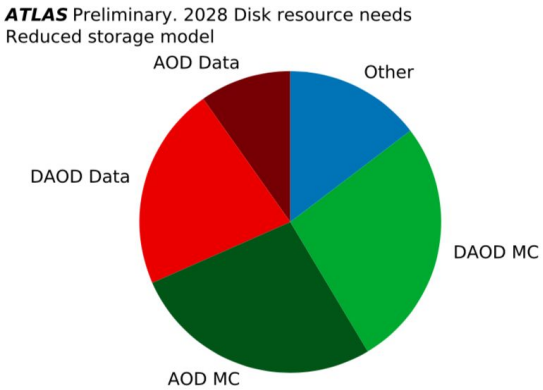




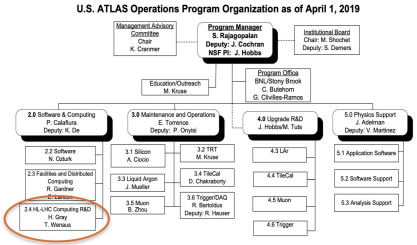
S&C challenges for the HL-LHC

- CPU challenges
  - Full Geant4 simulation remains a substantial part of the CPU
  - Assumption of 2x generator speedup, while generator precision must increase for the high precision physics of HL-LHC
  - None of our production workloads today utilize accelerators
- Disk challenges
  - 2028 disk usage already assumes very efficient usage: disk dominated by ‘hot’ compactified analysis formats
  - Opportunistic storage essentially doesn’t exist, unlike for CPU
  - We will have to make greater use of cheap(er) cold storage
  - We will rely on new dynamic network usage modes to mate distributed processing resources with short lived replicas

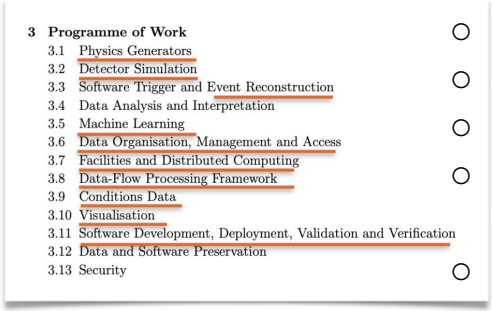




US ATLAS HL-LHC S&C R&D approach



- US ATLAS has had **three software focus areas**
  - Core offline **framework**, event **I/O** and **distributed** software
  - Selected for impact on US ATLAS analysis and alignment with US expertise
- Focus areas **frame our strategy** towards HL-LHC
- Many of the greatest challenges of HL-LHC computing are in our focus areas
  - Aim to be major contributors to solving them within and beyond ATLAS
  - Long time US ATLAS strength in leveraging HPCs is an important asset
- Established strategy and program as follows:
  - Strong participants in the **HSF community roadmap**
  - **Drew on roadmap conclusions** relevant to our focus areas
  - Factored in other elements: **DOE HPC strategy**, NSF strategy, **IRIS-HEP, ...**
  - Developed **prioritized work program** for HL-LHC computing R&D
    - Integrated with and complementing existing work in our focus areas
  - Have been assigning/hiring accordingly
    - Hiring emphasizes new blood from universities, almost 5 FTEs
      - 50/50 support split between S&C/analysis leverages university contributions and couples R&D efforts to real-world analysis work



3 **Programme of Work**

3.1 Physics Generators
3.2 Detector Simulation
3.3 Software Trigger and Event Reconstruction
3.4 Data Analysis and Interpretation
3.5 Machine Learning
3.6 Data Organisation, Management and Access
3.7 Facilities and Distributed Computing
3.8 Data-Flow Processing Framework
3.9 Conditions Data
3.10 Visualisation
3.11 Software Development, Deployment, Validation and Verification
3.12 Data and Software Preservation
3.13 Security

US ATLAS members were key members of the development of the HSF roadmap; contribute to most areas

US ATLAS HL-LHC R&D work breakdown

Tightly integrated with near term work, lots of overlap in people

Covers all areas in which we expect US ATLAS effort over time

Aligned with HSF roadmap, ATLAS strategy, US strengths/interests

Areas with current effort are underlined

Collaborations leverage additional effort and build a coherent HEP program

2.4.1 Software reengineering and algorithm development

- 2.4.1.2 Event generation
- 2.4.1.3 Simulation
- 2.4.1.4 Reconstruction
- 2.4.1.5 Analysis
- 2.4.1.6 Frameworks & Services
- 2.4.1.7 Event I/O & Persistency

2.4.2 Workflow porting to new platforms

- 2.4.2.2 Common Platform Support Infrastructure
- 2.4.2.3 HPC and Exascale Platforms
- 2.4.2.4 Cloud, commercial and other platforms

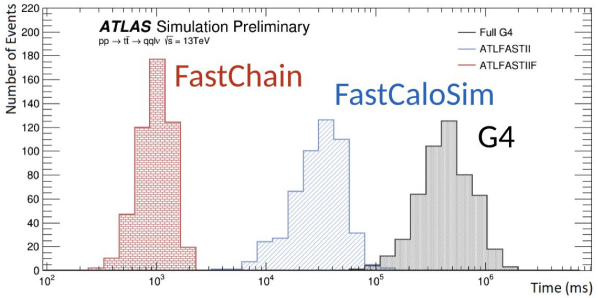
2.4.3 Distributed computing development

- 2.4.3.2 DOMA
- 2.4.3.3 Workload and workflow management
- 2.4.3.4 Analysis services
- 2.4.3.5 Common infrastructure

USATLAS funded
Collaborators funded

SW reengineering & algorithm development

- About 3 US ATLAS supported FTEs, plus ~1 FTE leveraged through collaborators. Current activities:
- **Fast simulation & fast chain** (including digi/reco)
  - Strong ATLAS emphasis on fast simulation, a critical assumption in the HL-LHC computing model
  - US ATLAS investment targeted towards FastChain on HPCs and accelerators
    - Three new university hires ramping up
    - Also useful to evaluate LCF use for high I/O intensity
  - Exploring production FastCaloSim on accelerators (BNL CSI/CCE)
  - Promising R&D underway on next-gen GAN based version
    - The one thing we *know* we can do on LCF HPCs is machine learning
    - Test case for distributed training on HPCs
- **Accelerator-enabled offline code**
  - Prerequisite MT framework will be in production for Run-3
    - This is the present focus of ATLAS work
  - Early days in exploring C++ tool stack, structuring workflows to benefit from accelerator-enabled algorithms
  - Some work on porting key algorithms to GPUs, e.g. track seeding on GPUs (IRIS-HEP)
- **Data layout, concurrent I/O, compression**
  - Data layout transformation needed to efficiently exploit accelerators
    - Synergy with optimizing layouts for efficient transport/streaming
  - Shared I/O components must be extended to support high HPC core counts
    - e.g. multiple instances of SharedWriter
- **Event generation**
  - US ATLAS involved in efforts to reengineer Sherpa and MadGraph



2.4.1 Software reengineering and algorithm development
- 2.4.1.2 Event generation
- 2.4.1.3 Simulation
- 2.4.1.4 Reconstruction
- 2.4.1.5 Analysis
- 2.4.1.6 Frameworks & Services
- 2.4.1.7 Event I/O & Persistency

Workflow porting to new platforms (HPCs)

- Currently about half an FTE of US ATLAS supported effort. Why so small?
  - With one exception, we have no workflows compatible with using new/imminent LCF HPCs (ie able to use accelerators)
  - After having put 10-15 FTE-years of effort into HPCs, our current ERCAP allocation is equivalent to less than 1/2 of one Tier 2, and we were told to expect cuts…
- Our present focus is on the exception: ML
  - Scaling up ML training and optimization on HPCs - more on the next slide
  - A joint effort of the porting and distributed software activities
- Next focus will be on the most promising future workflow: FastChain
  - When the software is ready to begin HPC porting
- US ATLAS ops is investing existing and new effort in both areas
- A welcome recent development: direct OLCF involvement
  - OLCF folks have joined our ML services meeting and will help
  - Very useful as our primary current target is Summit

2.4.2 Workflow porting to new platforms

- 2.4.2.2 Common Platform Support Infrastructure
- 2.4.2.3 HPC and Exascale Platforms
- 2.4.2.4 Cloud, commercial and other platforms

Distributed computing development

- About 3 US ATLAS supported FTEs, plus ~1 FTE leveraged through collaborators. Current activities:
- DOMA - Data (coupled to workflow) management
  - **Data carousel**
    - Target is Run-3, but also key part of Run-4 strategy
    - Storage-dataflow-workflow orchestration to dynamically provide a sliding window of disk-resident data from a tape-resident sample
  - **Fine grained event processing**
    - Next phase of the in-production event service: event streaming service
    - Being developed jointly with IRIS-HEP as experiment-agnostic “intelligent data delivery service” (iDDS)
    - First ATLAS application target: dynamic distributed data carousel
- Workload and workflow management: **ML services**
  - Scaled-up training and optimization via PanDA on grid and HPCs
  - Bootstrapped by the analysis community themselves: built a PanDA-based distributed hyperparameter optimization capability on the grid
  - Our ops task: extending this to HPCs and LCFs as supported, scalable ML service suite
  - Currently, training/optimization turnaround times are days to weeks
    - Direct impact on analyst productivity, and constraint on complexity & creativity
  - Allow training/optimization workloads to spike into large resources for fast turnaround
  - Make more sophisticated, processor-intensive networks possible: unshackle scientific creativity

2.4.3 Distributed computing development
- 2.4.3.2 DOMA
- 2.4.3.3 Workload and workflow management
- 2.4.3.4 Analysis services
- 2.4.3.5 Common infrastructure

Conclusion

- US ATLAS strategy towards HL-LHC continues our focus on three areas: core framework, event I/O, distributed computing
  - They include many of the big challenges of HPC computing
  - Dedicated HL-LHC computing activity area established June 2018
    - Substantial investments: now in a 5 FTE, 10 person ramp at universities
- Leveraging longstanding US ATLAS expertise with HPCs
- Strong engagement in common planning & collaborative projects: HSF, roadmap, WLCG, IRIS-HEP, CCE, Google, …
- Biggest contributor by far to HL-LHC computing R&D in US ATLAS is the ops program
- Next biggest is IRIS-HEP - thank you!
- Biggest support anomaly: essentially no support outside the ops program for our most promising early target for exascale: scaled-up ML training and optimization
- US ATLAS is in the vanguard of many R&D areas
  - Leveraging HPC/LCF/Exascale, GAN based fast calo simu, fine grained processing, data carousel, caching
- US ATLAS is a major contributor within ATLAS and beyond via common projects

Thank you

Thank you Heather, Paolo, Kaushik, and many others!

Supplementary

Details of the resource needs plots

- Disk
  - Estimated total disk resources (in PBytes) needed for the years 2018 to 2032 for both data and simulation processing. The plot updates the projection made in 2017 which was based on the Run-2 computing model and with updated LHC expected running conditions. The brown points are estimates made in 2017 and based on the current event sizes and using the ATLAS computing model parameters from 2017. The blue points show the improvements possible in two different scenarios, **which require significant development work**: (1) top curve, **reduction of AOD and DAOD size of 30%** compared to the Run-2 trend, bottom curve further reduction with the inclusion of a **common DAOD format** to be used by most analysis, removal of previous year AODs from disk and the **storage of only one DAOD version**. The solid line shows the amount of resources expected to be available if a flat funding scenario is assumed, which implies an **increase of 15% per year**, based on the current technology trends.
- CPU
  - Estimated CPU resources (in MHS06) needed for the years 2018 to 2032 for both data and simulation processing. The plot updates the projection made in 2017 (which was based on the Run-2 computing model) with updated LHC running conditions and revised scenarios for future computing models. The brown points are estimates made in 2017, based on the current software performance estimates and using the ATLAS computing model parameters from 2017. The blue points show the improvements possible in three different scenarios, **which require significant development work**: (1) top curve with fast calo sim used for 75% of the Monte Carlo simulation; (2) middle curve using in addition a faster version of reconstruction, which is seeded by the event generator information; (3) bottom curve, where the time spent in event generation is halved, either by software improvements or by re-using some of the events. The solid line shows the amount of resources expected to be available if a flat funding scenario is assumed, which implies an **increase of 20% per year**, based on the current technology trends.

US ATLAS external project partial list

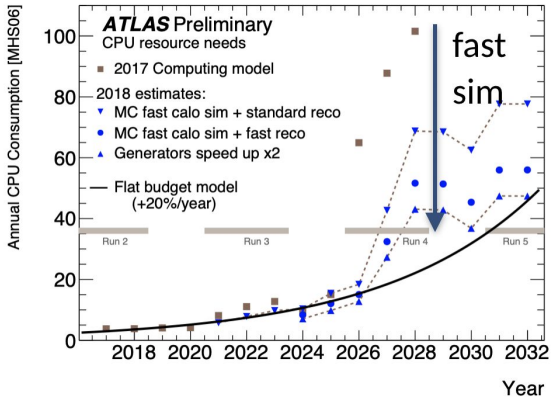
- BNL-CSI/CCE (ATLAS)
- Distributed ML, FastCaloSim on GPUs
- Geant ECP/CCE (HEP)
- Identify GEANT physics models suitable to run on ECP architectures. Quantify gains
- Exalearn ECP (HEP):
- Graph Neural Networks for scientific data processing.
- IRIS-HEP (NSF): effort tightly integrated with WBS 2.4 New Tracking, Analysis Model, Data Management
- Tracking CCE (ATLAS/CMS)
- BigPanDA (DOE-ASCR) workflow management, OLCF usage by ATLAS
- Parallelization opportunities for tracking algorithms
- HEP.TrkX HEP (ATLAS/CMS/DUNE)
- ML tracking algorithms
- NESAP for Data (HEP)
- Parallel distributed ROOT I/O (with ALCF/FNAL)
- Aurora Early Science
- Simulating and Learning in the ATLAS Detector at the Exascale
- Early access to A21 architecture/ software platform
- HEP.QPR (HEP)
- HEP Quantum Pattern Recognition

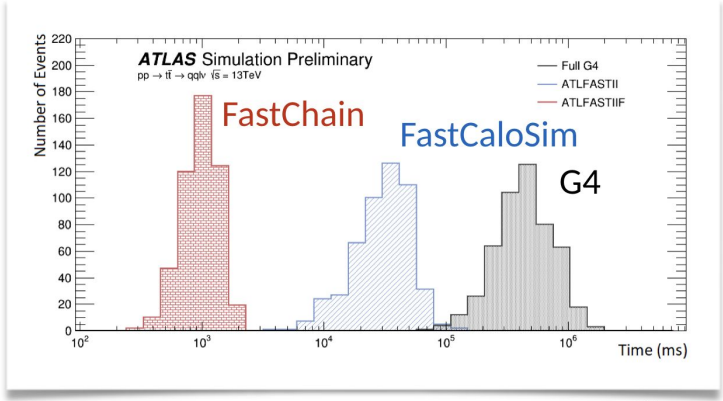
Ongoing R&D: Simulation

2.4.1
2.4.3

ANL
BNL UMass
UCI
UW

- Moving physics analyses from full to fast simulation is a critical assumption in ATLAS' computing model for HL-LHC
  - FastCaloSim (parametrised calo response) gains an order of magnitude over G4
  - FastChain (fast sim + fast reco) gains a further order of magnitude
  - Project has been struggling for manpower for many years
- Recent USATLAS investment in fast simulation
  - Targeted towards deployment on HPCs and accelerators



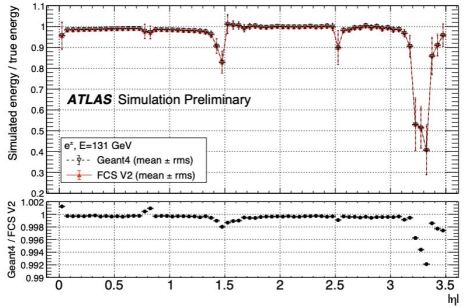


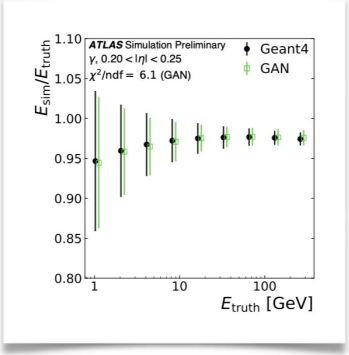
Ongoing R&D: Simulation

2.4.1
2.4.3

ANL BNL
UMass UCI
UW

- FastCaloSim on HPCs and accelerators
  - ATLAS-BNL CSI project underway to start exploring this
  - Standalone binary code developed and delivered by ATLAS
  - Optimisation and accelerator deployment is underway
- Ongoing ATLAS effort for calorimeter simulation with GANs starts to show promising performance
  - Test case for distributed training in 2.4.3
- Three new university hires in 2.4 for FastChain development and deployment on HPCs and accelerators
  - Currently ramping up, looking at detailed profiling and strategy developments
  - Workflow also useful to evaluate LCF use for high I/O intensity





Ongoing R&D: Accelerators

- Early explorations of developing GPU-enabled code in C++
  - Tool stack seems immature
- ATLAS core reconstruction efforts are currently focussed on the migration towards athena MT
  - Strategy: port key individual components to GPUs
    - e.g IRIS-HEP: tracking on GPUs
- Event data model and data layout need to be transformed to fully exploit accelerators
  - Synergy with optimizing layouts/formats for transport/streaming on the wire, data transformations in moving from disk->tape, how layout relates to caching hierarchy, integrating (de)compression in workflows to hide latency, ..

2.4.1
2.4.2

Berkeley
LBNL
Stanford

track seeding on GPUs

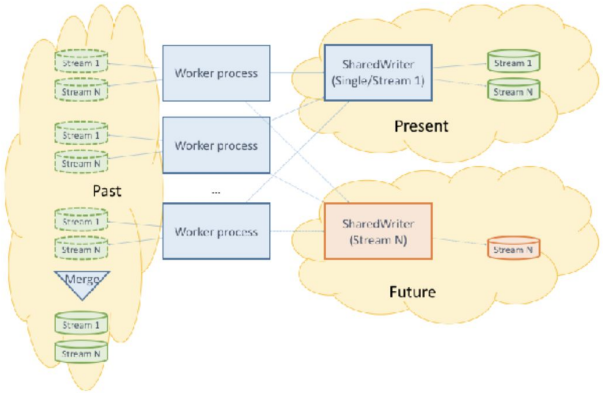
X. Ai, UC Berkeley

Ongoing R&D: Shared I/O

2.4.1

ANL
BNL

- High HPC core count make shared I/O components highly beneficial
- Measurements for sim (reco) show that these scale to over 32 (16) processes w/o significant slow down by a single SharedWriter
- For more I/O intensive processing, e.g. derivations, single SharedWriter can be overburdened by number of worker processes
- New developments for multiple SharedWriters: Each SharedWriter writes all the events from all workers, but different output streams/ files use separate instances of the SharedWriter

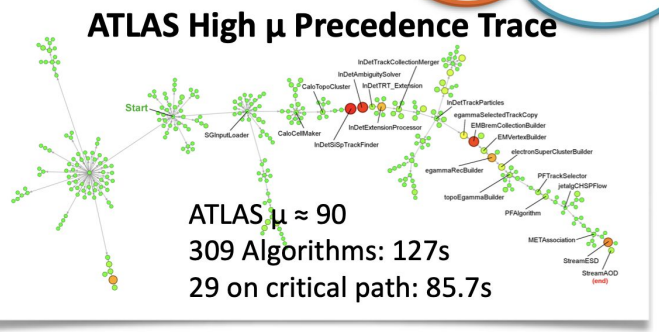


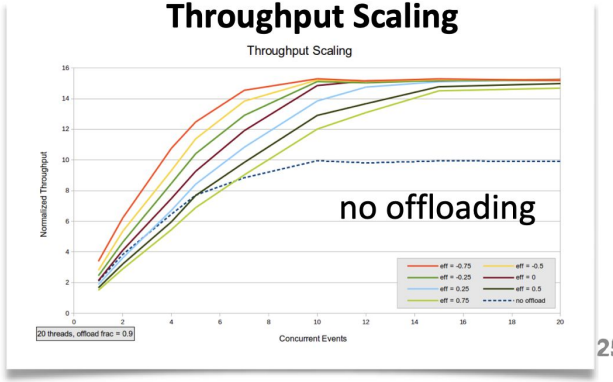
Ongoing R&D: Data&Workflow

2.4.1
2.4.3

BNL
Wisconsin

- Event Streaming Service (ESS) = Fine-grained event processing
  - Principal event service developer is now lead ESS developer
  - Joint effort with ATLAS & IRIS-HEP to develop an experiment-agnostic intelligent data delivery service (iDDS), with ESS the ATLAS specialization of it
    - Design and prototyping is underway
  - DOMA new hire will contribute here
- Infrastructure for simulating diverse workflows on heterogeneous architectures, applicable across experiments with ATLAS, CMS and LHCb participating, has been developed with results reported at ACAT 2019



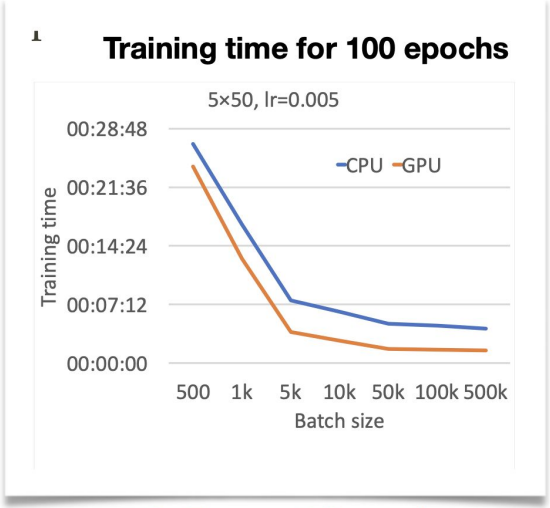


Ongoing R&D: Machine Learning

2.4.3

BNL
UT Arlington
Wisconsin

- Work began on a strategy and workplan for interfacing machine learning frameworks with Athena, with TensorFlow being the first target (new hire)
- Demonstrator for PanDa-based hyperparameter optimisation has been run on the GRID (discussed in many ATLAS meetings)
  - Broadening deployment to GPU-equipped sites, Summit, etc
  - Developing from demonstrator to generally usable service
- PanDA-based distributed training service is another joint ATLAS-BNL CSI project
  - evaluating distributed training tools (e.g. GANs for FastCaloSim)
  - building the service itself



R. Zhang, Wisconsin

Ongoing R&D: Analysis Systems

- **RECAST/REANA**: Analysis Reuse and Preservation
  - ATLAS uses this for all Run 2 published exotics analysis
  - Workflow and data management and archiving system
  - Preservation
  - Simplifies determining sensitivity of analysis to new signal
  - Connections to SCAILFIN
- **Columnar Analysis**: Towards a compelling analysis solution in the Python ecosystem
  - Coffea tool from CMS
  - Members of ATLAS and CMS and LHCb
  - Profit from wealth of tools outside HEP
  - Different approach to writing an analysis; more compatible with GPU, AVX architecture.
- Other Projects: Declarative Analysis Languages, parallel fitting, B decay language

2.4.1

NYU, UW, UIUC