ATLAS Simulation: Input for ESPPU

Marilena Bandieramonte (University of Pittsburgh),
Jana Schaarschmidt (University of Washington)

HSF Simulation Community Meeting - 22. November 2024

Introduction & Documents

- HL-LHC to start in 2030. Increase in luminosity leading to much larger event sizes and rates, but our budget for resources doesn't scale accordingly
  → **Changes to computing model crucial to ensure continuous operation**

- Strong R&D program in ATLAS ongoing, with ~30 demonstrators being actively worked on

- First version of some tools intended for Run 4 already deployed in Run 3 to gain feedback

- R&D program is constantly evolving (successful R&D leads to more advanced R&D), roadmap deliverables and milestones tracked every six months

Documents:

- ATLAS HL computing program laid out in the CDR (Fall 2020) and Roadmap document (Summer 2022) with fine-grained milestones and timelines

- Upcoming: HL-LHC computing TDR (~2025) with prioritized milestones and deliverables with effort estimates and resource impact assessments

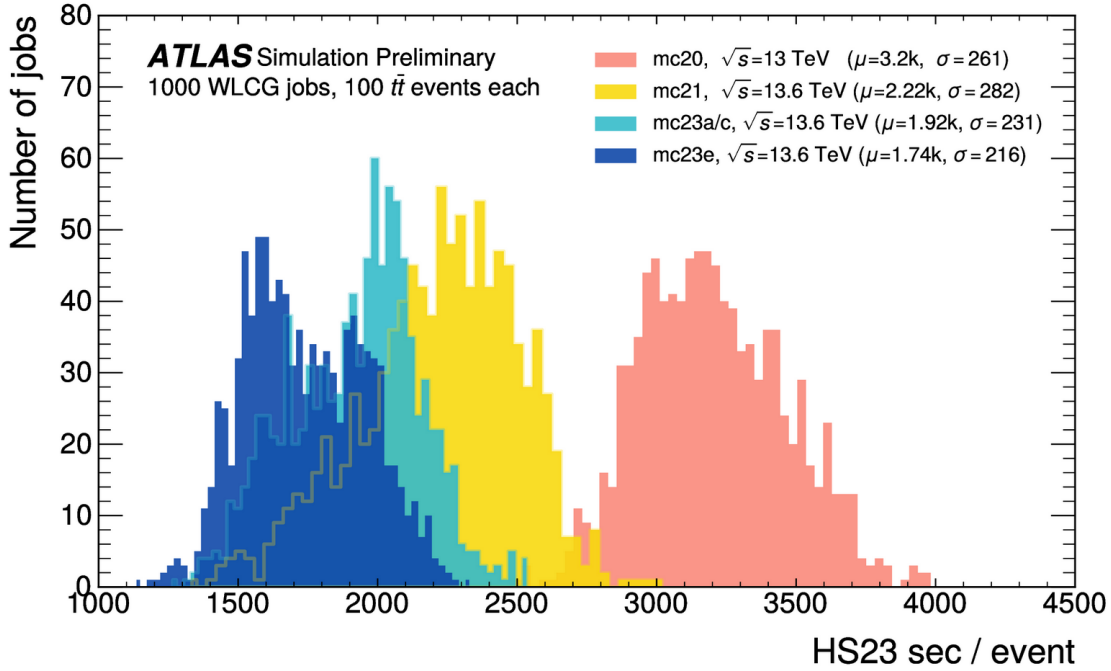
Main Guidelines/Tools for Run 4

- **Optimized simulations** (either detailed simulation with Geant4, or fast simulations), i.e. don't spend CPU on regions in phase space that are not relevant.
  Aim to keep the current data:MC ratio (1:1).
- **Accurate simulations** (best case: Fast sim and full sim can use the same calibrations, to avoid extra delays when those need to be derived – tedious process in ATLAS)
- **Flexible workflows:**
  - Customizable for specific physics needs – being more accurate where needed
  - Workflows that allow to trade for example CPU with tape space, if need arises
- **Utilization of GPU and HPC**
- **Modern code** (concurrency, machine learning support, less dependencies on old software)
- **Support for new detectors** (simulation and digitization)
- **Main tools:**
  - Full Simulation based on Geant4
  - Fast Calorimeter Simulation (AtlFast3)
  - Fast Chain (fast calo simulation + fast ID simulation + fast pile-up modelling)

FullSim Optimizations

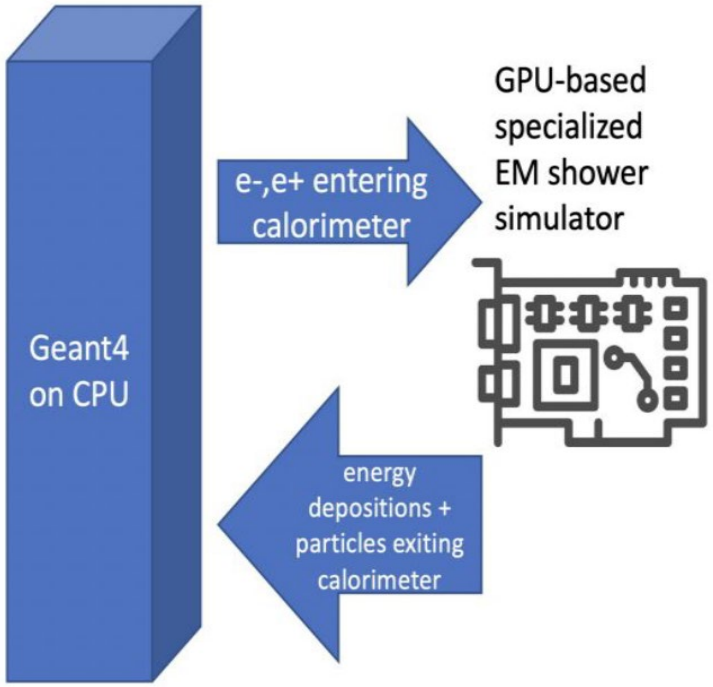
- Aggressive scenarios foresee 90% of all simulation will be fast sim in Run 4
- Expect that FullSim optimization work will decrease eventually in Run 4, but FullSim still much needed: - For training the fast sim models
  - For special physics cases that can't use fast sim
- In Run 3, we have observed some reluctance of analysers to use fast sim
  - Resources currently not limited
  - Delay/unavailability of object recommendations (calibrations, uncertainties)

  -> Fear that switch to "only fast sim" won't happen easily/quickly



- CPU optimizations deployed for Run 3 detailed in Run 3 S&C paper
- Many small improvements led to ~48% CPU reduction in total, without compromising accuracy
- Still new ideas being worked on, and also efforts to keep up with G4 developments and automatizing deployment (calculating sampling fractions, tuning, validations, …)

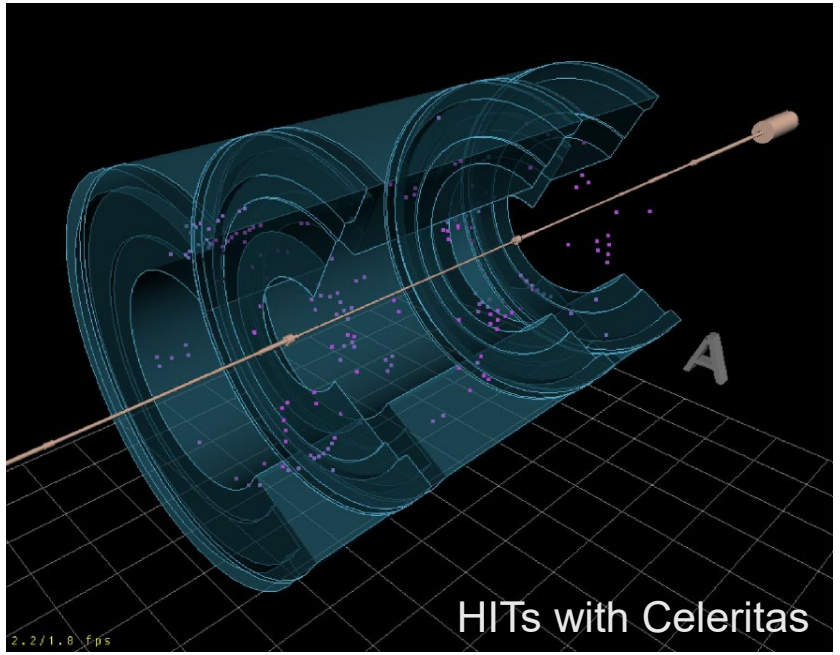
GPU Usage



- Goal is to offload at least some part of the full simulation to GPU (parts that are very CPU intensive, e.g. EMEC)
- For FullSim, two approaches actively worked on:
  AdePT (CERN/SFT)
  Celeritas (ORNL,ANL,LBL,FNAL)
- Prototype version for calo simulation that is integrated with Athena is worked on now, first results available
- No end-to-end comparisons to standard grid jobs available yet

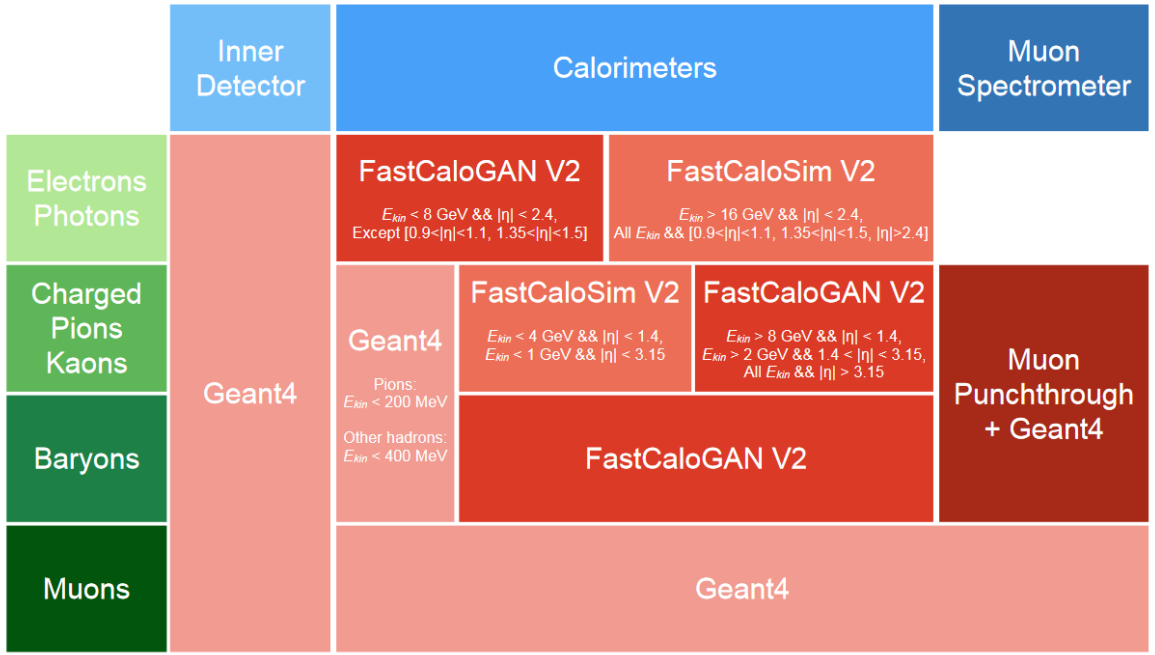
Fast Sim:

- Also work ongoing for fast sim inference (using CUDA), see paper
- GPUs also useful for training ML models used for fast sim -> even more so in the future



Fast Calorimeter Simulation: Overview

- Calorimeter simulation is the slowest part of simulation, takes ~70% of all simulation CPU
- Fast calorimeter simulation has long history in ATLAS, has been used since Run 1
- **AtlFast3 (AF3)** is the latest & greatest tool (paper), actually a combination of tools, mainly:
  - Calorimeter: FastCaloSim (parametrized) or FastCaloGAN (ML-based)
  - Inner detector, muon system: G4
- Speeds up simulation by factor 3-15 depending on the process (cool plots)



- In Run 2 fast sim was 50%, in Run 3 it is 40% (calibrations missing), in **Run 4 should be 90%**
- Trying to get more physics analysers to use AF3 instead of FullSim for Run 3
- Expect AF3 and future versions to be of critical importance and priority to ATLAS

Fast Calorimeter Simulation: R&D Studies

Lots of R&D ongoing to improve AF3:

- Voxelisation optimisations (granularity of the training inputs) -> for better accuracy (trade-off with memory). Promising!
- Moving to more sophisticated ML models, e.g. normalizing flows, diffusion models, VAEs, …
- Many such ideas were tried out for the CaloChallenge →
- Effort ongoing to make AF3 experiment-independent in the future: CHEP talk.
  As part of that, FCS code is now open source.
- Plan to test (perhaps optimize) AF3 also for long-lived particles (currently not supported, one of the few physics cases that can't use AF3)

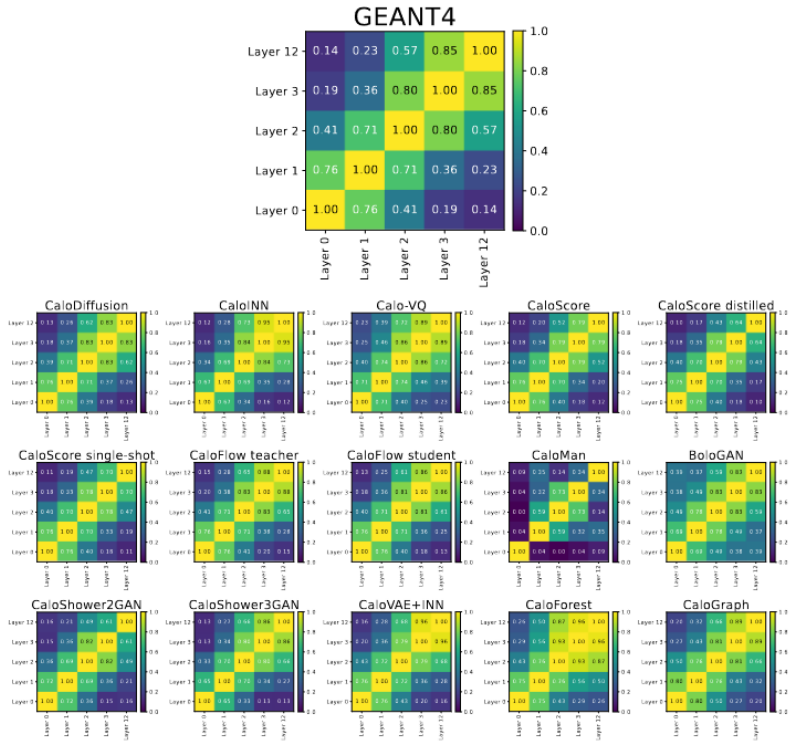
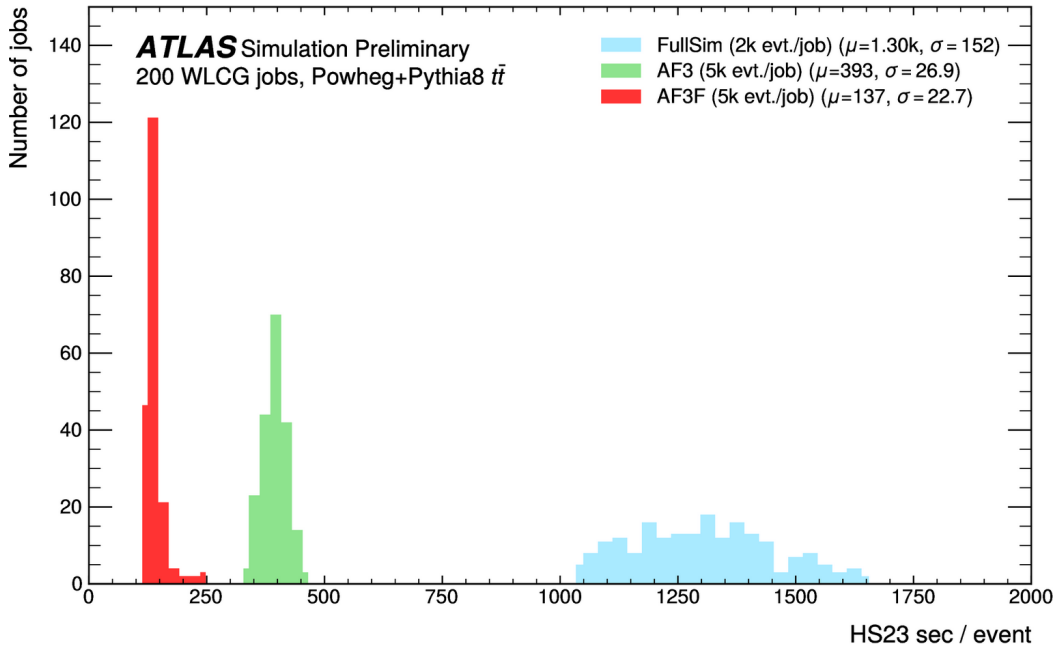


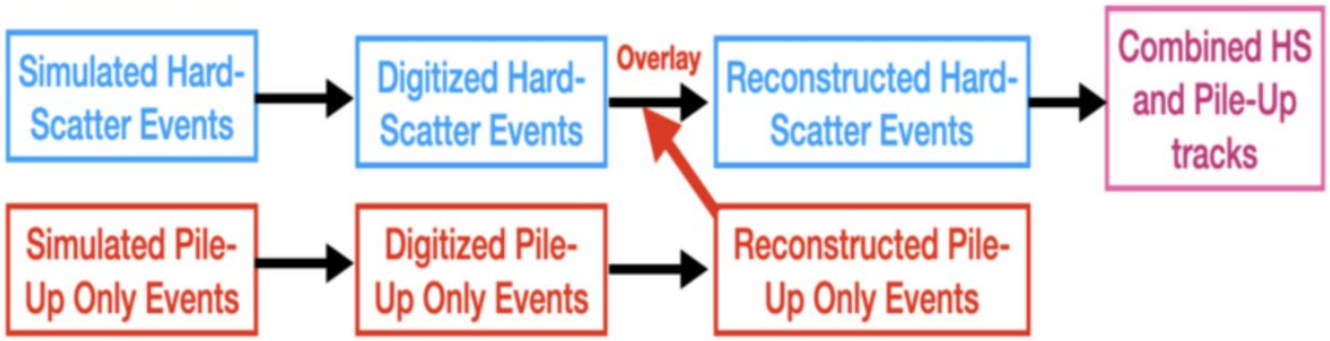
Figure 38: Pearson correlation coefficients of layer energies in ds $1-\gamma$, with threshold at 1 MeV.

Fast Inner Tracking Simulation

- After the calorimeter, the inner detector is the next CPU-intensive part of simulation
- FATRAS is the ATLAS tool for Fast Tracking simulation (note), in development since long time
  -> Not ML-based. Using simplified geometry, Bethe-Bloch, Bethe-Heitler formalisms and other approximations.
- FATRAS used only for e/gamma, no hadronic models (not good enough) – use G4 instead
- Not deployed yet, because physics performance not good enough
- But R&D continues, very worth the effort, will bring down sim CPU by another factor 3 (ttbar)
- Once ready for physics, will play a major part in ATLAS as well
- Run 4 FATRAS tool will be integrated in ACTS and MT-ready
- Future research: Using ML for ID simulation (inspired by success for fast calo sim)



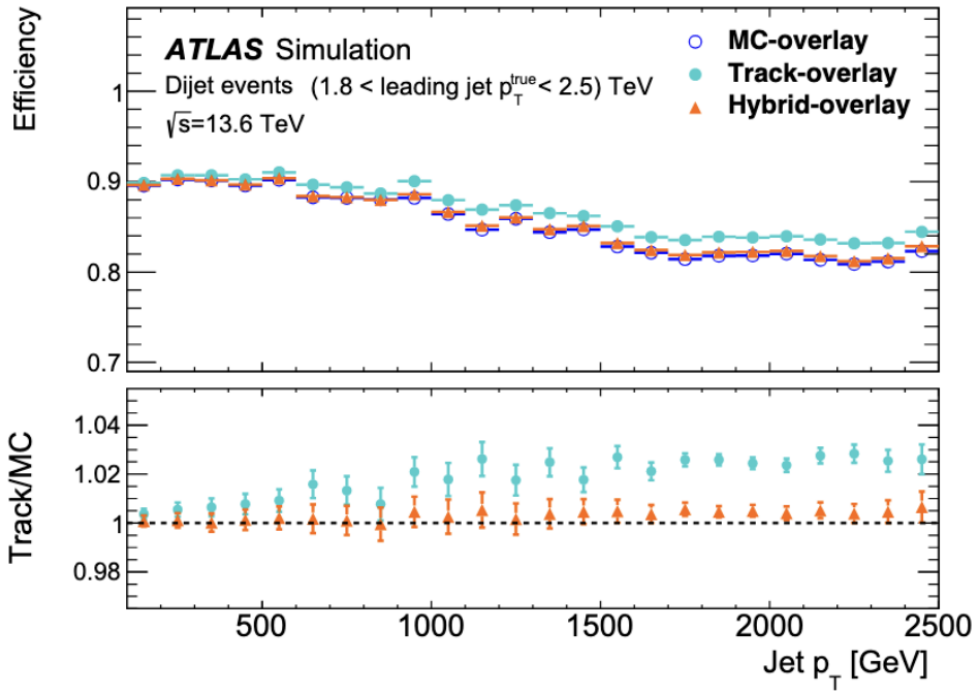
Track Overlay for Fast Reconstruction



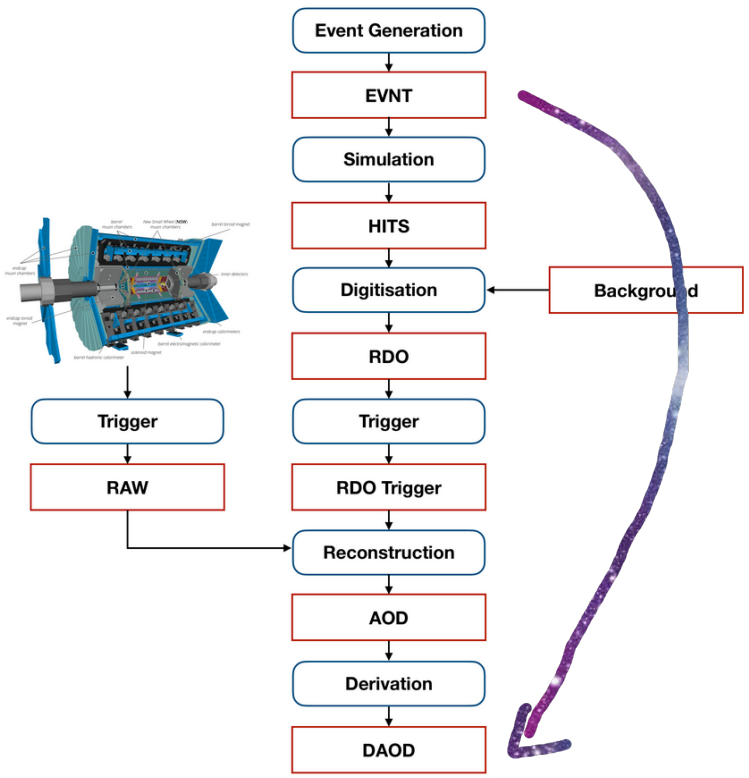
Idea is to **reconstruct pile-up events** before overlaying with hard scatter (HS)

-> Saves 50% reconstruction time (poster, talk)

- Approach does not work well in dense environments -> ML to the rescue!
- ML used to decide per event whether to merge PU and HS before or after track reconstruction
- Will deploy still in Run 3 -> get experience
- Expect this technique to play major role during HL
- Previous approach to speed-up reconstruction by using truth information obsoleted
- Also, fast digitization not currently pursued

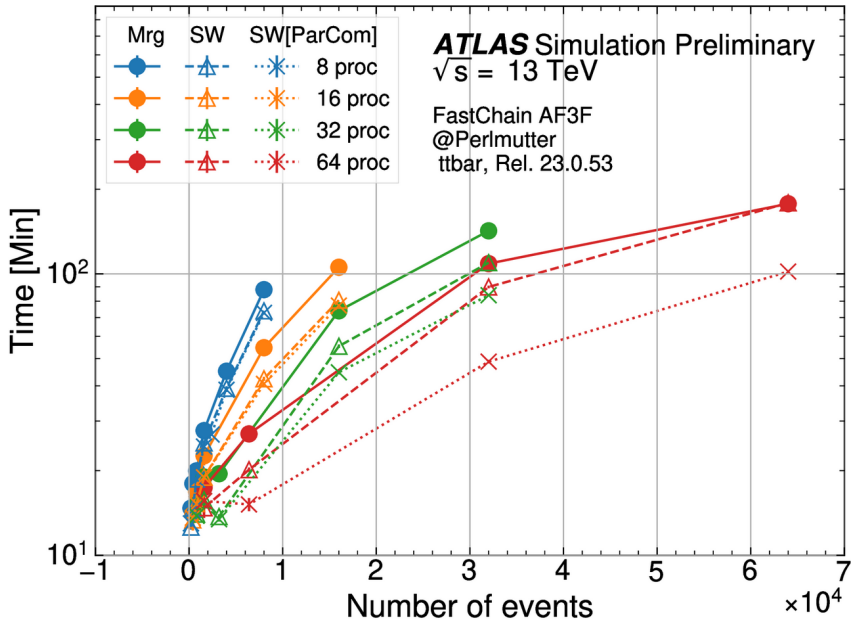


Fast Chain: Workflows and HPCs



- Fast Chain is flexible! Can combine tools as needed (Fast Calo Sim, FATRAS, track overlay)
- Can also **skip storing intermediate formats** (HITs) and only store derivations (analysis-level)
  -> Saves 200 PB tape per year, but need to rerun simulation for annual reprocessing campaign (see proceedings)
  -> This is *not* the same idea as for example CMS flash sim – uses ML to go directly from generator objects to NANO-AOD

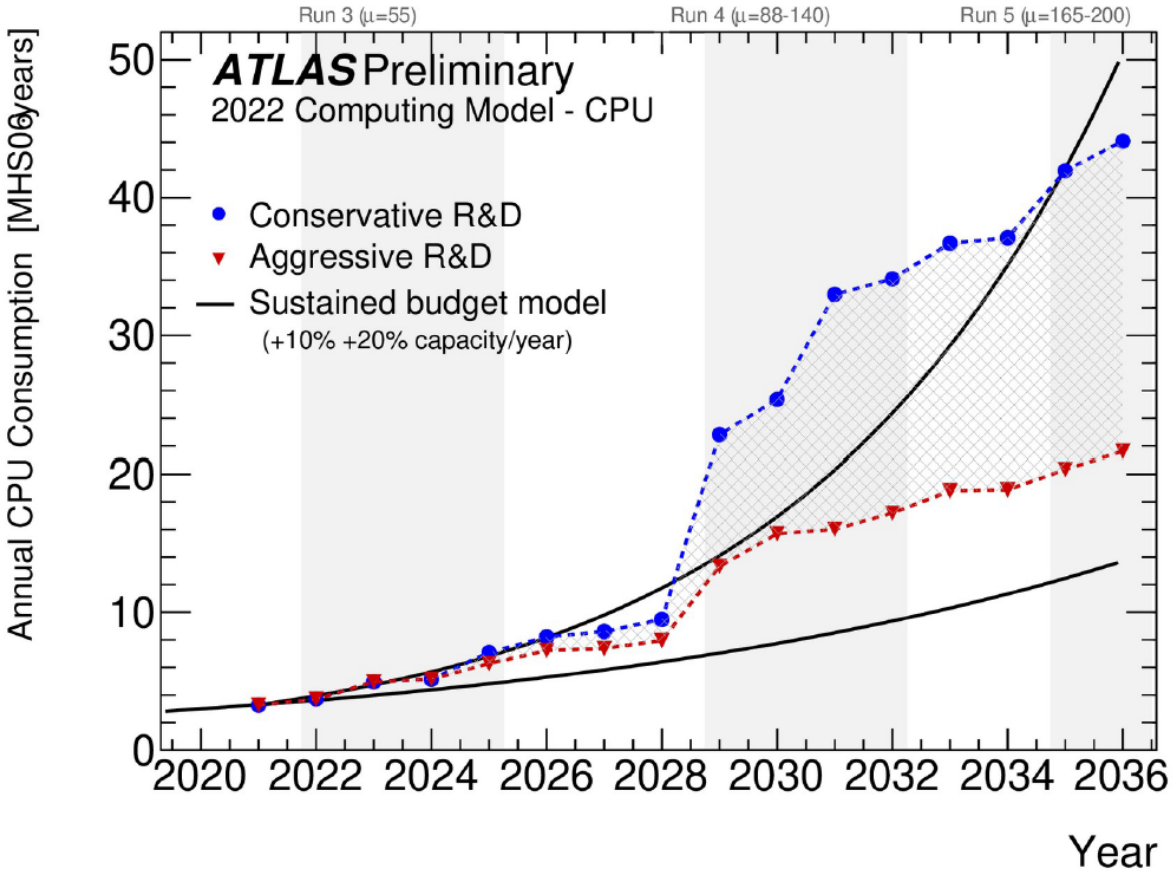
- FastChain workflow AF3+FATRAS was also tested on HPCs (Perlmutter)
- Several job configurations were tested to find one that scales best
- More cool plots

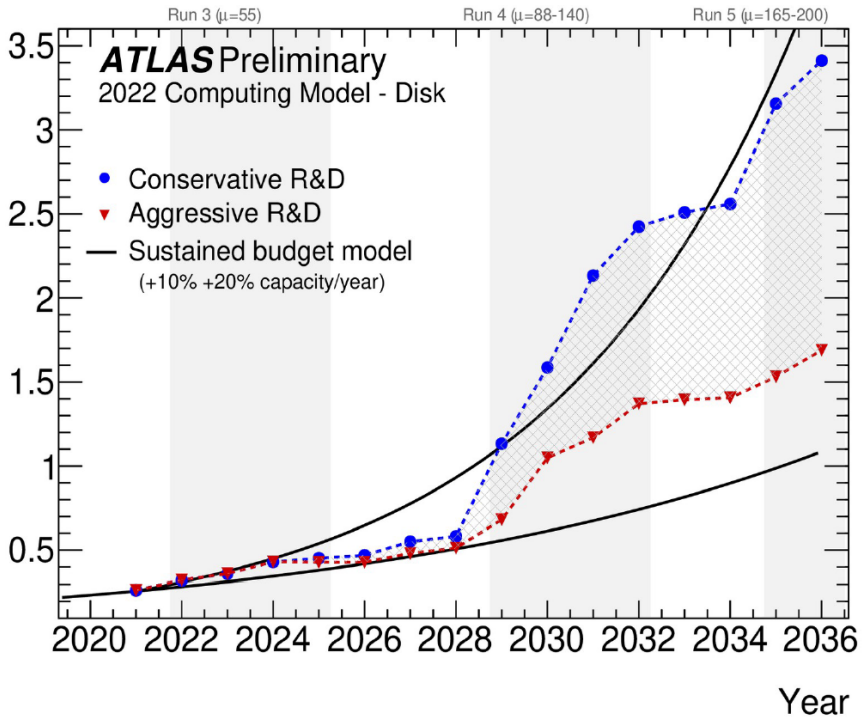


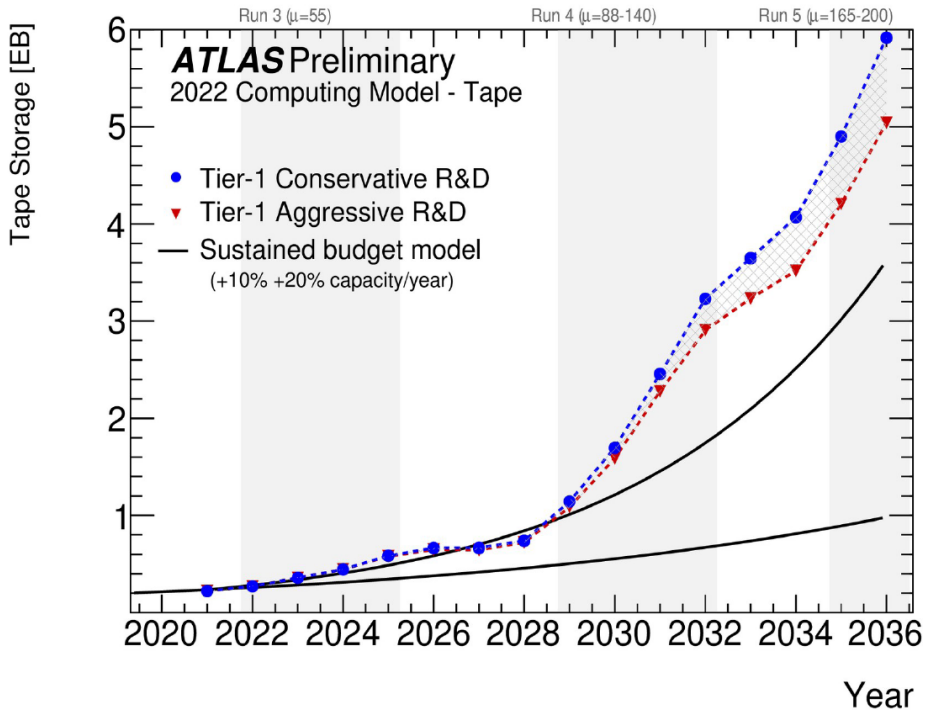
Conclusions

- Strong R&D program in ATLAS towards HL, simulation is a major part
- Expect FastSim to become default simulator, but perhaps not immediately
- FullSim CPU optimizations will become less critical later during Run 4
- GPU usage will become more relevant in the future, when software integration is ready
- ML extremely relevant and useful for simulation and also reconstruction
- Beneficial if tools can be deployed in Run 3 already, to get feedback from users, can further optimize methods

Backup


Run 3 (μ=55)
Run 4 (μ=88-140)
Run 5 (μ=165-200)
ATLAS Preliminary
2022 Computing Model - CPU
Conservative R&D
Aggressive R&D
Sustained budget model
(+10% +20% capacity/year)
Annual CPU Consumption [MHS06 years]
Year


Run 3 (μ=55)
Run 4 (μ=88-140)
Run 5 (μ=165-200)
ATLAS Preliminary
2022 Computing Model - Disk
Conservative R&D
Aggressive R&D
Sustained budget model
(+10% +20% capacity/year)
Year


Run 3 (μ=55)
Run 4 (μ=88-140)
Run 5 (μ=165-200)
ATLAS Preliminary
2022 Computing Model - Tape
Tier-1 Conservative R&D
Tier-1 Aggressive R&D
Sustained budget model
(+10% +20% capacity/year)
Tape Storage [EB]
Year

ATLAS Simulation Preliminary
Subdetectors CPU Fraction, mc20
100 $t\bar{t}$ events