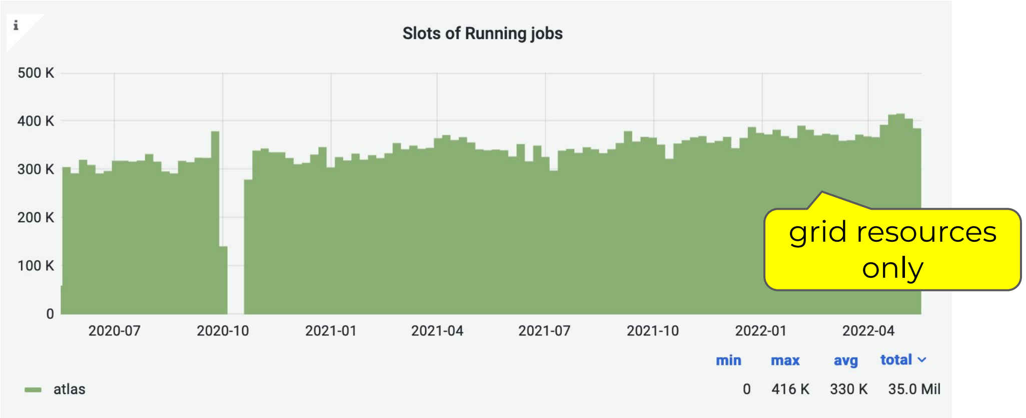
Highlights of the ATLAS experiment computing developments

CCR Workshop - Paestum

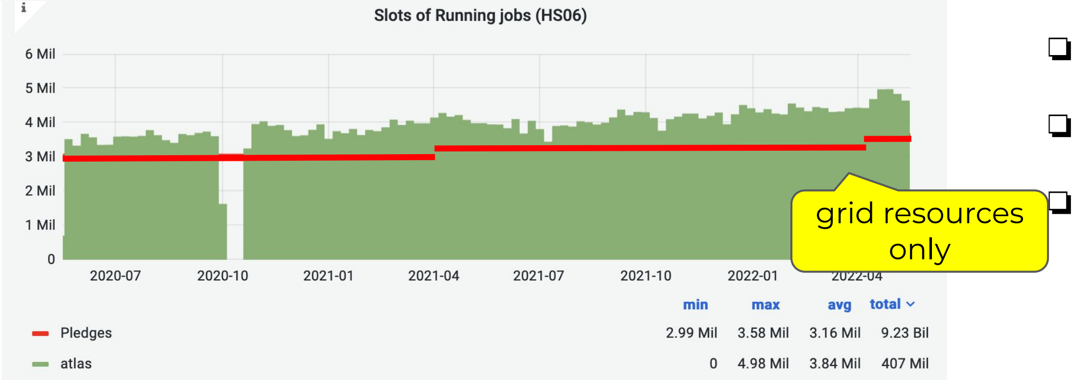
L. Carminati (UNIMI), A. Doria (INFN Napoli) for the ATLAS Italia Computing Team

ATLAS Computing performance : standard “grid” resources

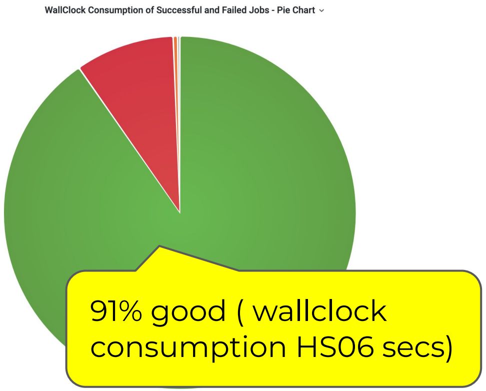
ATLAS experiment has been running jobs full speed over the last two years (1.5.2020-1.5.2022) despite the pandemic emergency constraints :



grid resources only



grid resources only

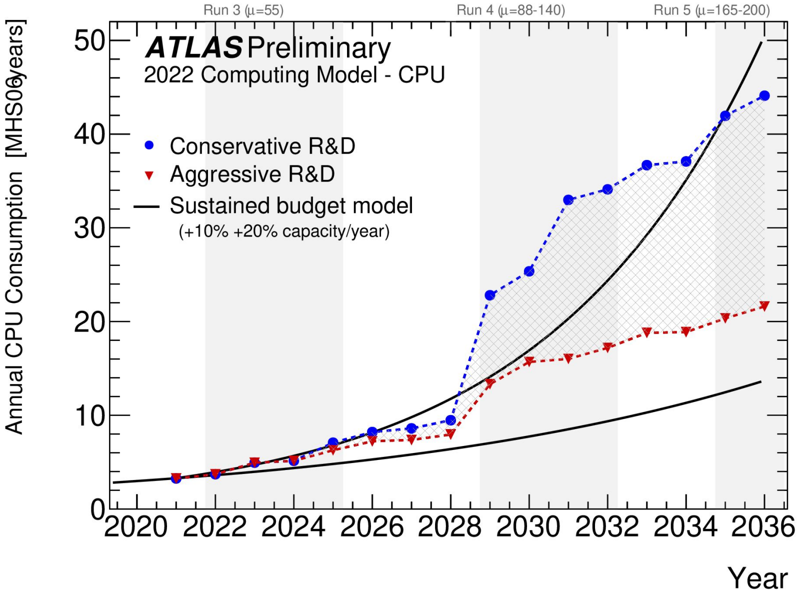


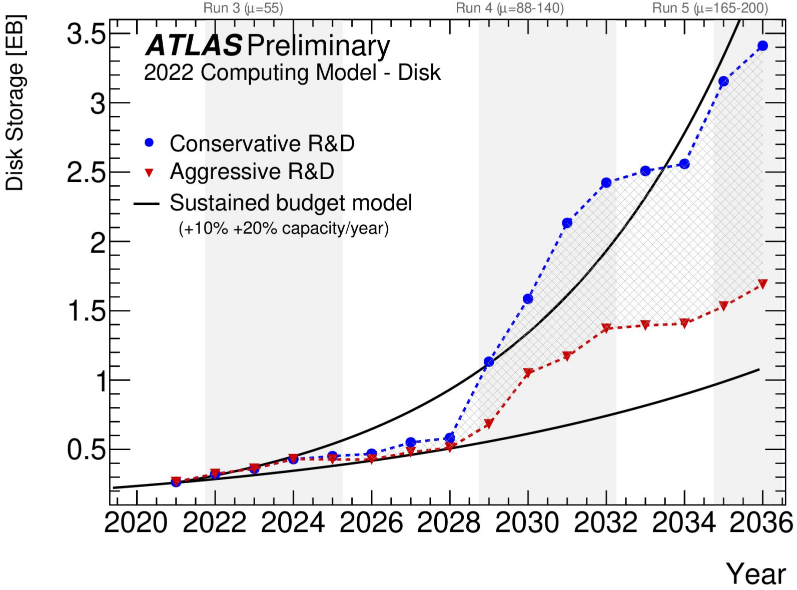
91% good (wallclock consumption HS06 secs)

- ❑ ATLAS has fully exploited the standard (grid) computing pledged resources
- ❑ On average 330 k jobs running in parallel
- ❑ Overall efficiency ~ 91% (a bit higher if we exclude single users analysis jobs)

ATLAS Computing requirements for HL-LHC

ATLAS HL-LHC Computing Conceptual Design Report : projections of ATLAS computing requirements for Run3 and HL-LHC to fully exploit the machine physics potential is quite scaring !

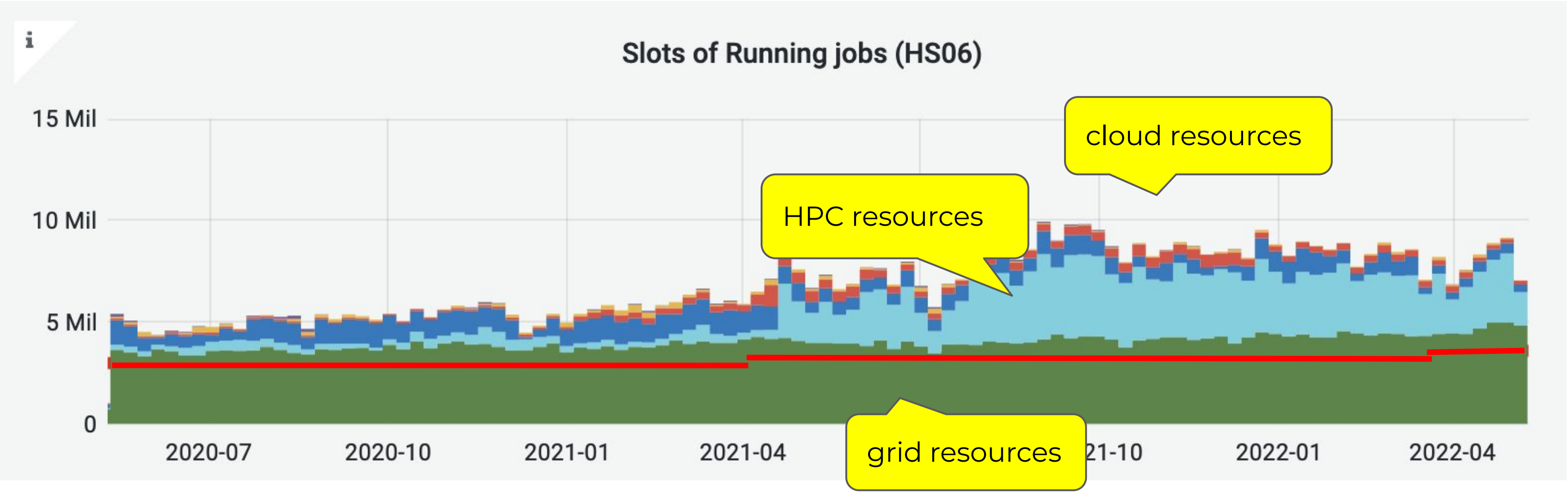




Discussion started on possible strategies to meet the demanding requirements of HL-LHC

- ❏ optimisation (both speed and flexibility) of the experiment (e.g. reconstruction, simulation) and non-experiment (e.g. generation) software
- ❏ optimisation of the available hardware infrastructure usage

ATLAS Computing per type of resources

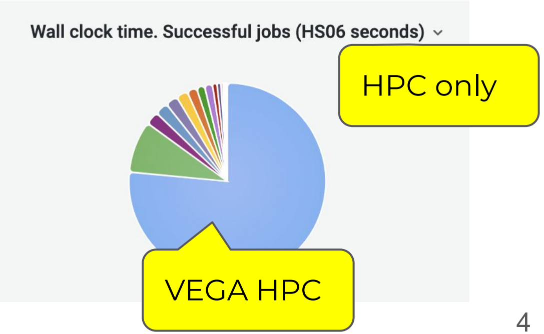


cloud resources

HPC resources

grid resources

- ❑ When including ALL resources, on average 576k jobs running in parallel (spikes ~900k jobs running) overall efficiency ~ 90%
- ❑ ATLAS has used up to 2.5 the computing pledge resources last year
- ❑ Impressive contribution from HPC resources, mainly (~75% of the full HPC resources used by ATLAS) from opportunistic access to VEGA (Maribor, Slovenia within the EuroHPC program)
- ❑ Possibility to exploit HPC resources needs to be investigated carefully !

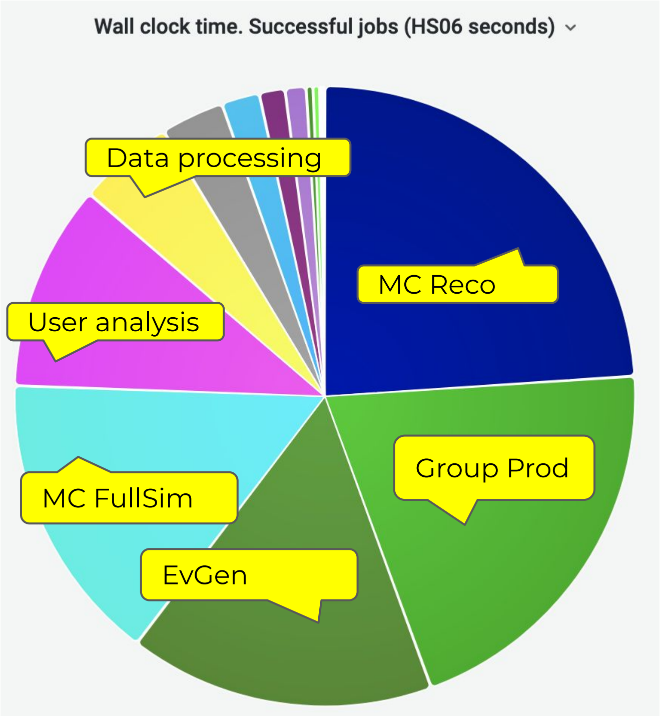


HPC only

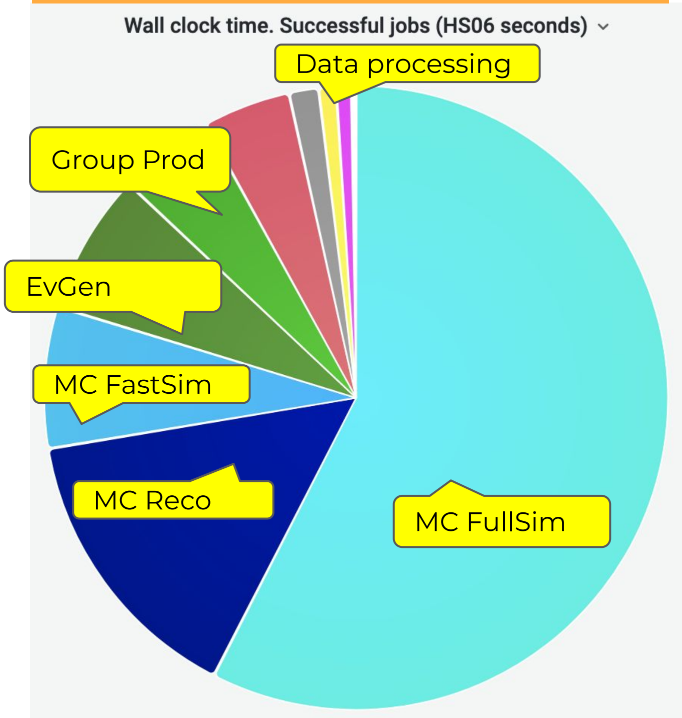
VEGA HPC

ATLAS Computing per type of resources

GRID RESOURCES



VEGA(HPC)



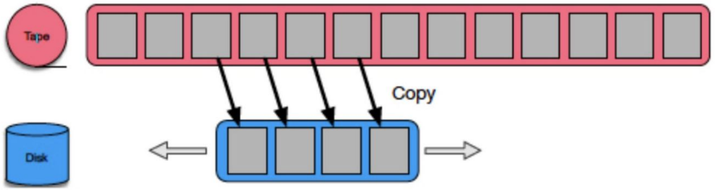
- ❏ Typical share of shutdown period (MC, analysis and reprocessing)
- ❏ VEGA (HPC) able to run all ATLAS workflows
- ❏ 1/16/64 threads jobs, 1GB/thread (4GB/thread queue available)
- ❏ CAVEATS (I): Opportunistic usage at the startup of the cluster, not guaranteed in the next years. Sharing with other users might introduce inefficiencies
- ❏ CAVEATS (II): lot of tunings (size, number of events etc) needed and still not optimal usage of the hardware.

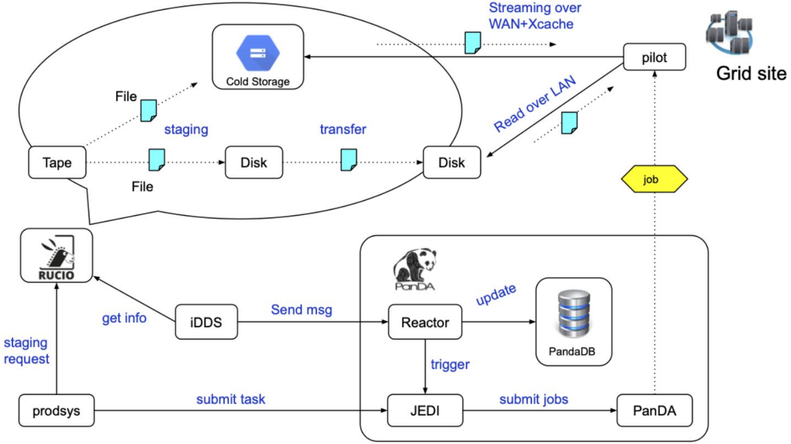
The Tape Carousel model

ATL-SOFT-PROC-2021-012

The Data Carousel is the orchestration between the workflow management systems ProdSys2 and PanDA, the distributed data management system Rucio, and the tape services. It enables a bulk production campaign, with input data resident on tape, to be executed by staging and promptly processing a sliding window of a fraction of the input onto buffer disk such that only a percentage of the data are pinned on disk at any one time

- □ Phase I: Tape system performance evaluation at CERN and the WLCG Tier-1 centers.
- □ Phase II: Deeper integration between workflow, workload and data management systems (ProdSys2/PanDA/Rucio), and Identify missing software components
- □ Phase III: Run Data Carousel at scale in production for the selected workflows with an ultimate goal to have it operational before LHC Run 3 in 2022.
  - □ reprocessing of run2 data/MC
  - □ production of derived data ✓
- □ Phase IV : use data carousel for many workflows in parallel respecting computing share per workflow. Run Data Carousel jointly for more than one experiment

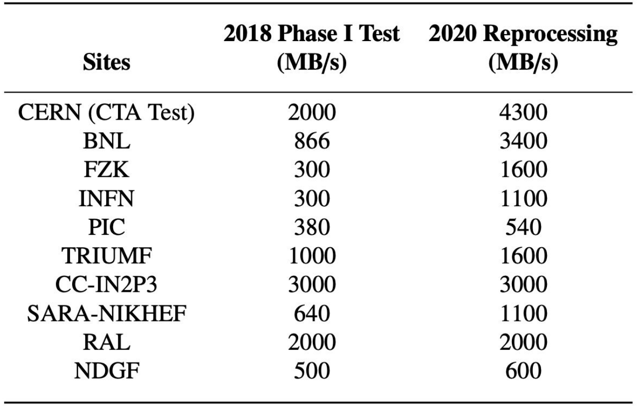




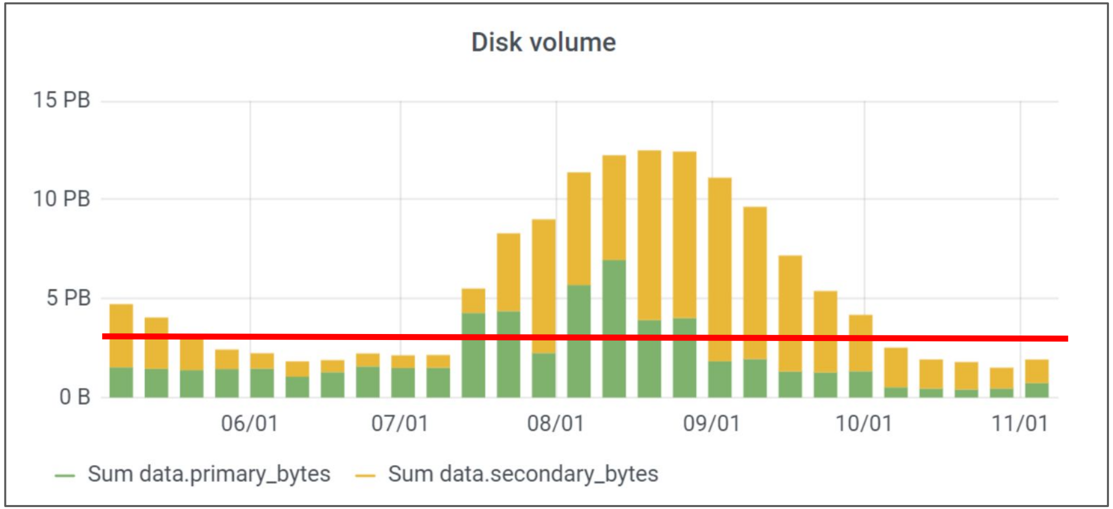
The Tape Carousel model

ATL-SOFT-PROC-2021-012

- ❑ Data Carousel for the reprocessing of all data collected by ATLAS in 2015-2018.
- ❑ The total data volume was close to 18.5 PB.
- ❑ Impressive improvement of the tape performance at T1 thanks to the work of local experts
- ❑ Target to keep on average 3 PB of data on disk (red line in the plot), generally achieved
- ❑ Several issue found and solved :
  - ❑ tuning of the algorithm of data replication in rucio
  - ❑ fixes in data pinning
  - ❑ introduced iDDS : allows JEDI to incrementally release tasks so tha tasks can start processing even if input data are only partially staged-in.



Sites	2018 Phase I Test (MB/s)	2020 Reprocessing (MB/s)
CERN (CTA Test)	2000	4300
BNL	866	3400
FZK	300	1600
INFN	300	1100
PIC	380	540
TRIUMF	1000	1600
CC-IN2P3	3000	3000
SARA-NIKHEF	640	1100
RAL	2000	2000
NDGF	500	600



Multi-threaded reconstruction software (AthenaMT)

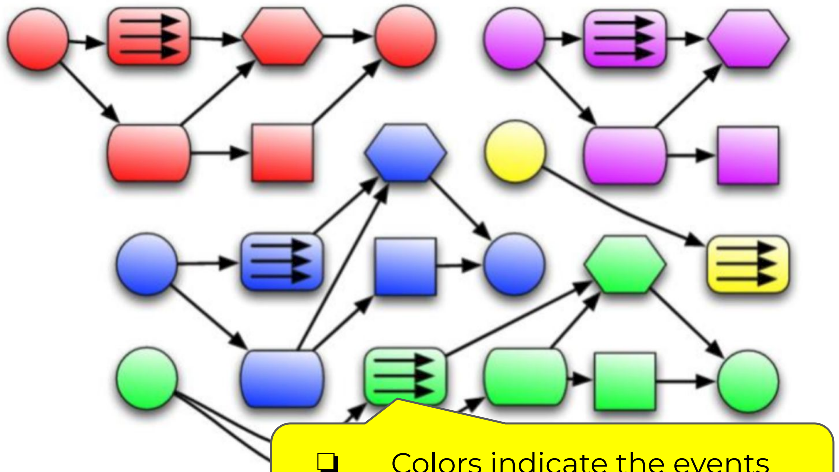
ATL-SOFT-PUB-2021-002

The new Athena release 22 (used since 2021 for reprocessing of Run 2 MC and data, as well as for Run3 data taking and MC simulations) is able to offer both multi-process and multi-thread parallelism.

- In multi-process (MP) parallelism, workers are forked from the primary process at a pre-configured stage during execution (e.g. before the first event is processed). Each worker also has its own unique memory space and produces its own outputs, which need to be merged via a post-processing step.

- In multi-thread (MT) threads are spawned and assigned some work (e.g. execute an algorithm). Single pool of heap memory shared across all threads. Various difficulties must be overcome:
  - multiple threads cannot write to the same memory at the same time;
  - threads must not attempt to read memory that is actively being written
  - algorithms must be scheduled such that all input is fully available before they run.

  However, the performance benefit from using a single pool of memory for all threads can be significant.



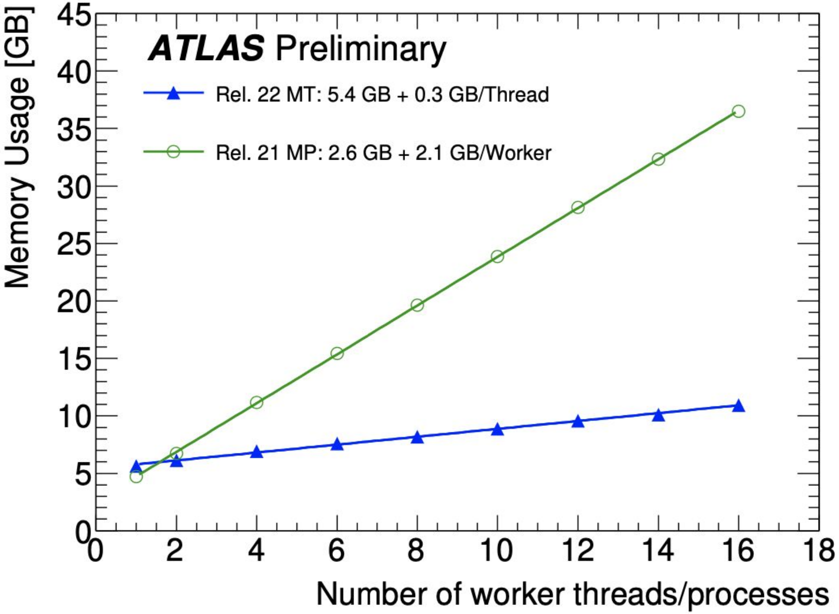
- Colors indicate the events
- Shapes indicate algorithms

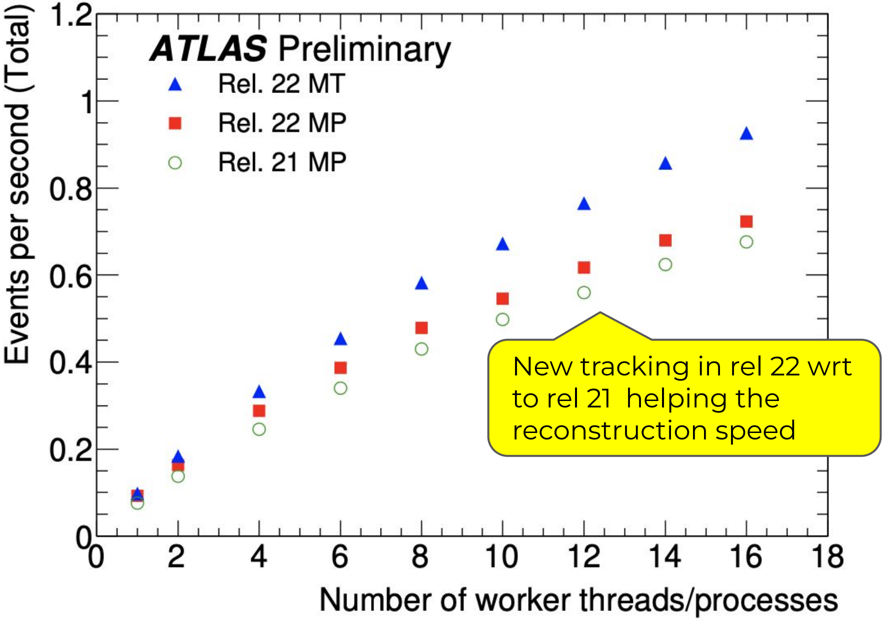
- Multi-threaded reconstruction software allows a better exploitation of opportunistic resources (eg HPC)

Multi-threaded reconstruction software (AthenaMT)

ATL-SOFT-PUB-2021-002

- ❑ The benchmark jobs use real data from run 357750 taken during 2018, with 250 events per worker process or thread.
- ❑ The data events have an average number of interactions per bunch crossing ⟨μ⟩ = 50, which is approximately that expected for the luminosity-leveling period during Run 3.
- ❑ Tests on an Intel®Xeon®CPU E5-2630 v3 at 2.40 GHz (16 cores no SMT) machine + 126 GB of memory.





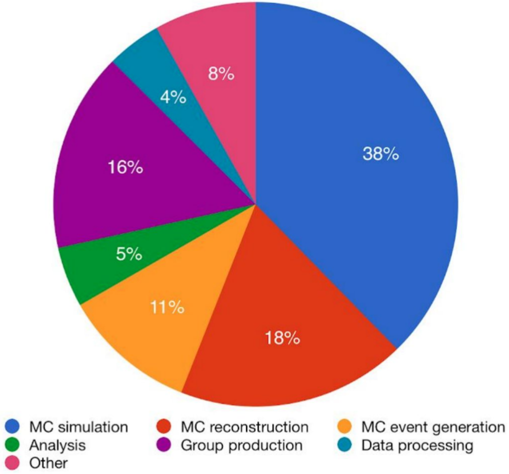
New tracking in rel 22 wrt to rel 21 helping the reconstruction speed

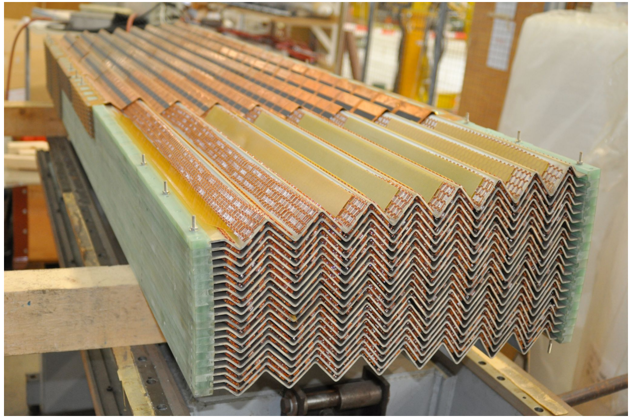
MC events production

ATL-SOFT-PROC-2021-017

Multipurpose experiments cover a wide ranging physics program from precision measurements to searches for new physics

- Monte Carlo events (both hard scatter and pile up) are functional to this process
- Typically the number of simulated MC events is ~ 2.5 the number of data events !
- Most of ATLAS CPU time used for MC detector simulation and ~80-90 of detector simulation time spent on calorimeters (complex geometries)



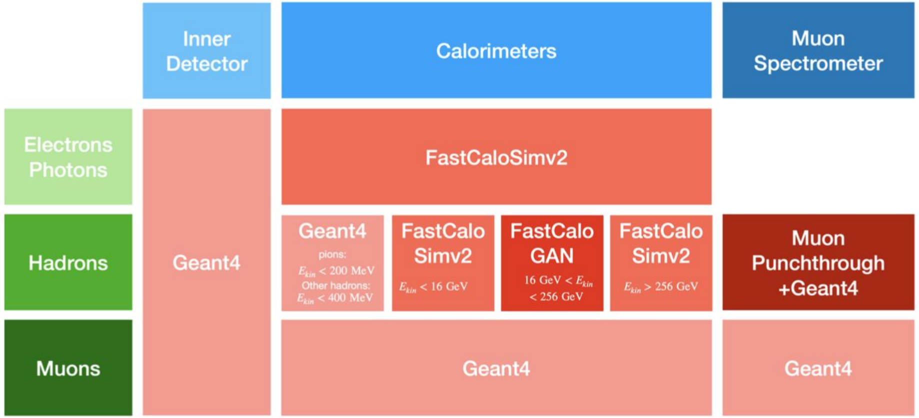


To reduce the impact of the preparation of MC events :

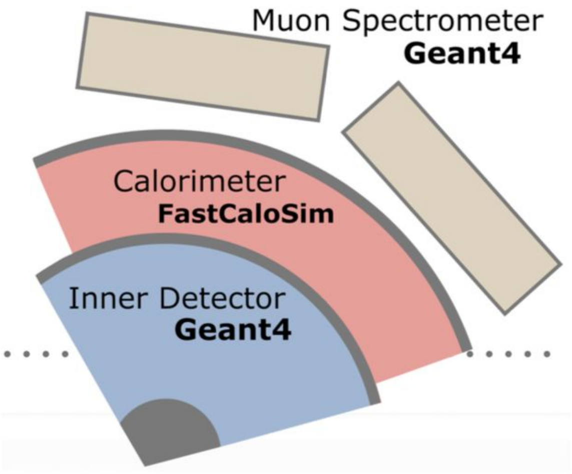
- Optimise G4 full simulation (see Caterina’s talk)
- Pushing on fast (calo) simulation
  - Reduce simulation time keeping as much accuracy as possible + memory efficiency
  - Increase the number of analyses using FastSim : Run 3: >50% events with fast simulation, Run 4: >75% events with fast simulation
- Part of the full-simulation on accelerators (e.g. GPUs)

Fast simulation : Atlfast-III

Fast Simulation : instead of simulating interactions of particle travelling through detector parametrise the detector response of single particles (Atlfast) : use electrons and photons for electromagnetic showers and pions for hadronic showers



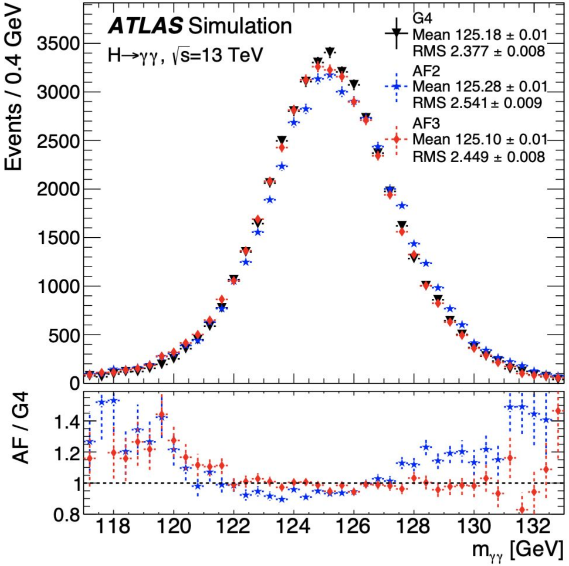
	Inner Detector	Calorimeters				Muon Spectrometer
Electrons Photons	Geant4	FastCaloSimv2				
Hadrons		Geant4 pions: $E_{kin} < 200$ MeV Other hadrons: $E_{kin} < 400$ MeV	FastCalo Simv2 $E_{kin} < 16$ GeV	FastCalo GAN 16 GeV $< E_{kin}$ < 256 GeV	FastCalo Simv2 $E_{kin} > 256$ GeV	Muon Punchthrough +Geant4
Muons		Geant4				Geant4

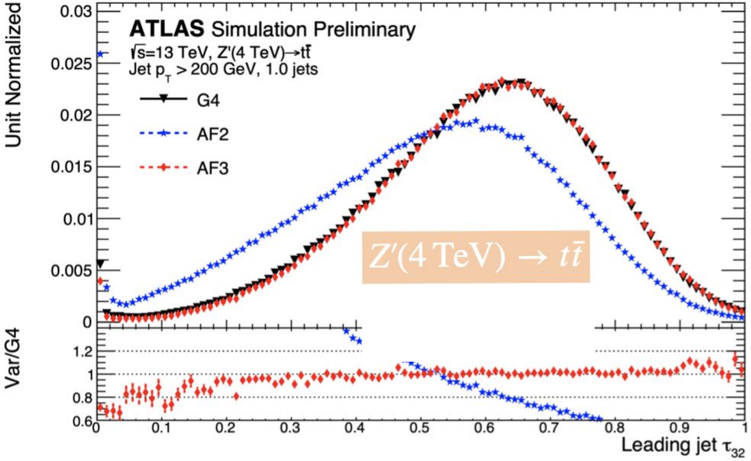


- ❑ Atlfast-III (AF3) is the successor of the Atlfast-II (AF2) simulator
- ❑ Full simulation for tracking (ID + muons) and parameterized simulation of the calorimeter
- ❑ AF3 implements two distinct approaches of shower generation:
  - ❑ FastCaloSimV2: parameterized modelling (separate parameterisations of longitudinal and lateral shower developments)
  - ❑ FastCaloGAN: Generative Adversarial Network: GAN trained to reproduce voxels and energies in the layer as well as total energy in one single step
- ❑ Dedicated parameterization for punch through particles

Fast simulation : Atlfast-III

Encompassing complex parameterized and deep learning algorithms, AF3 is the state of the art fast simulation in ATLAS and able to simulate a broad range of physics processes with high precision





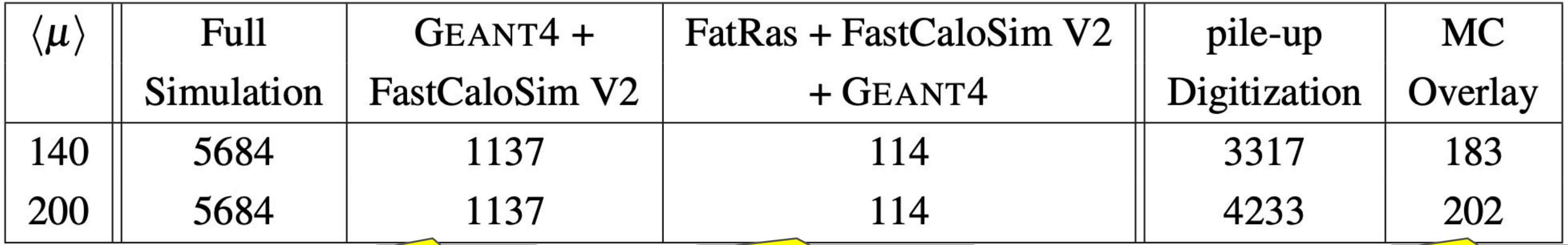
- ❏ AF3 provides significant improvements in physics performance compared to AF2 while giving a speedup of $\mathcal{O}$ (5-10) compared to Geant4
- ❏ Improvements include better modelling of jet masses, constituents and substructure, better *e/γ* simulation and more
- ❏ AF3 was used for the re-processing of ~7 billion Run 2 events • Many more improvements expected for Run 3 and beyond

Future of fast simulation : FastChain

For Run 4(5), ATLFAST-III will not be ‘fast’ enough to keep up with increased data statistic

- The next step will be a fast simulation for the ATLAS Tracker, FATRAS (within ACT : FATRAS + FastCaloSim is ~50 times faster than pure Geant4).
- More in general : build a chain of fast simulation tools (FastChain) for fast simulation, digitization and reconstruction, to be used interchangeably depending on the specific analyses need

HS06 x seconds



$\langle\mu\rangle$	Full Simulation	GEANT4 + FastCaloSim V2	FatRas + FastCaloSim V2 + GEANT4	pile-up Digitization	MC Overlay
140	5684	1137	114	3317	183
200	5684	1137	114	4233	202

Fast Calo simulation

Fast Calo + tracking simulation

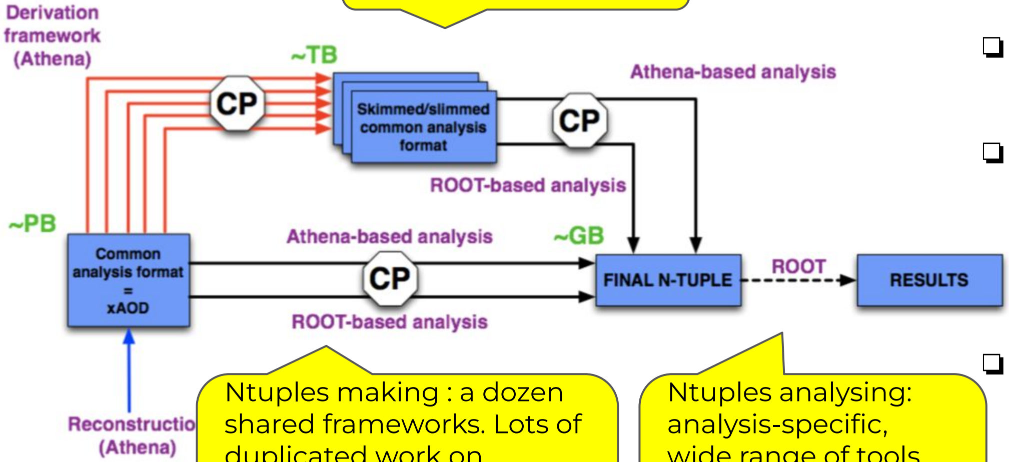
RDO overlay

- Might also stop saving simulation output (HITS) as an intermediate format and go straight from EVNT to AOD in a single production step on the grid (save storage) .
- Aiming for production-readiness before the end of Run 3.

A new analysis model

The ATLAS Run-2 analysis model has been highly successful in the view of the productivity of ATLAS, but it has been expensive in terms of resource usage. The ATLAS Analysis Model Study Group for Run-3 (AMSG-R3) setup at the end of Run-2 was tasked to analyse the efficiency and suitability of the current model and to propose significant improvements.

Run2 Analysis model



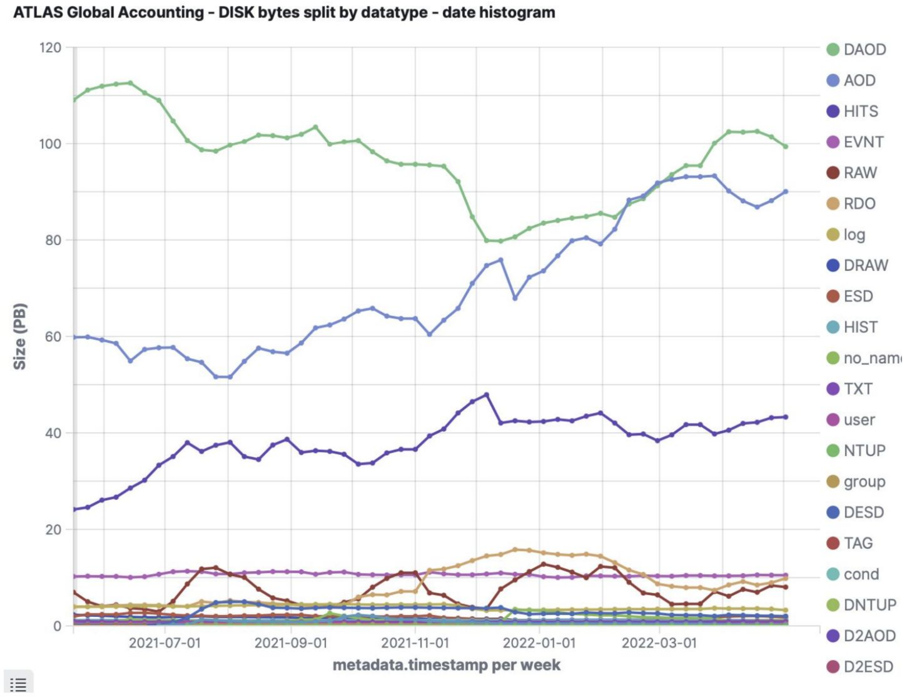
Ntuples making : a dozen shared frameworks. Lots of duplicated work on implementation of calibrations + systematic uncertainties

Ntuples analysing: analysis-specific, wide range of tools (ROOT macros to RDataFrame to uproot …)

- The output of the data and MC reconstruction is stored in Analysis Object Data (AOD) files and grouped in datasets on the various Grid sites.
- These datasets are processed in the derivation framework which produces about 80 different derived AOD (DAOD) formats that contain a subset of events and reduced reconstruction information tailored for specific physics analysis and performance groups.
- These DAOD types are processed by many individual analysers in a random manner who produce very condensed individual ntuples for further processing or final physics results

A new analysis model

The ATLAS Run-2 analysis model has been highly successful in the view of the productivity of ATLAS, but it has been expensive in terms of resource usage. The ATLAS Analysis Model Study Group for Run-3 (AMSG-R3) setup at the end of Run-2 was tasked to analyse the efficiency and suitability of the current model and to propose significant improvements.


ATLAS Global Accounting - DISK bytes split by datatype - date histogram

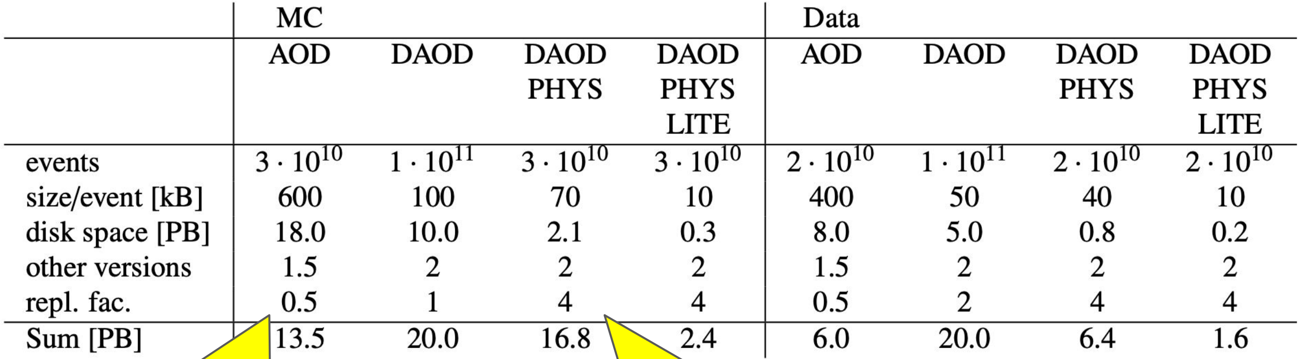
- The size/event for the AOD is about 600 kB
- Around 80 DAOD formats, size in the range of 40-450 kB depending on the type of the physics selection and the information retained.
- only 1-2 replicas of each dataset and campaign can be kept on disk.
- AODs and DAODs which are the two formats taking more than 70% of the disk space today
- As a rough Run-2 input parameter an initial sum of 132 PB of disk space used for AOD and DAOD format
- For the HL-LHC the projections of the ATLAS needs are significantly over the yearly flat budget increase. ATLAS is therefore investing significantly in methods to reduce the disk space needs in several areas

A new analysis model

❑ Introduce instead a new single DAOD_PHYS targeted for all (>~80%) physics analysis (~50 kB/event).

❑ In addition a new smaller DAOD_PHYSLITE format (10 kB/event) will be introduced that contains already calibrated physics objects and will be centrally produced with frequent updates, typically every few months. A larger fraction of the AODs will be removed from disk and staged-in back from tape storage on demand in a so called data carousel mode of operation.

❑ Allow exceptions for performance groups, B-physics (separate stream), long lived particle searches….



	MC				Data			
	AOD	DAOD	DAOD	DAOD	AOD	DAOD	DAOD	DAOD
			PHYS	PHYS			PHYS	PHYS
				LITE				LITE
events	$3 \cdot 10^{10}$	$1 \cdot 10^{11}$	$3 \cdot 10^{10}$	$3 \cdot 10^{10}$	$2 \cdot 10^{10}$	$1 \cdot 10^{11}$	$2 \cdot 10^{10}$	$2 \cdot 10^{10}$
size/event [kB]	600	100	70	10	400	50	40	10
disk space [PB]	18.0	10.0	2.1	0.3	8.0	5.0	0.8	0.2
other versions	1.5	2	2	2	1.5	2	2	2
repl. fac.	0.5	1	4	4	0.5	2	4	4
Sum [PB]	13.5	20.0	16.8	2.4	6.0	20.0	6.4	1.6

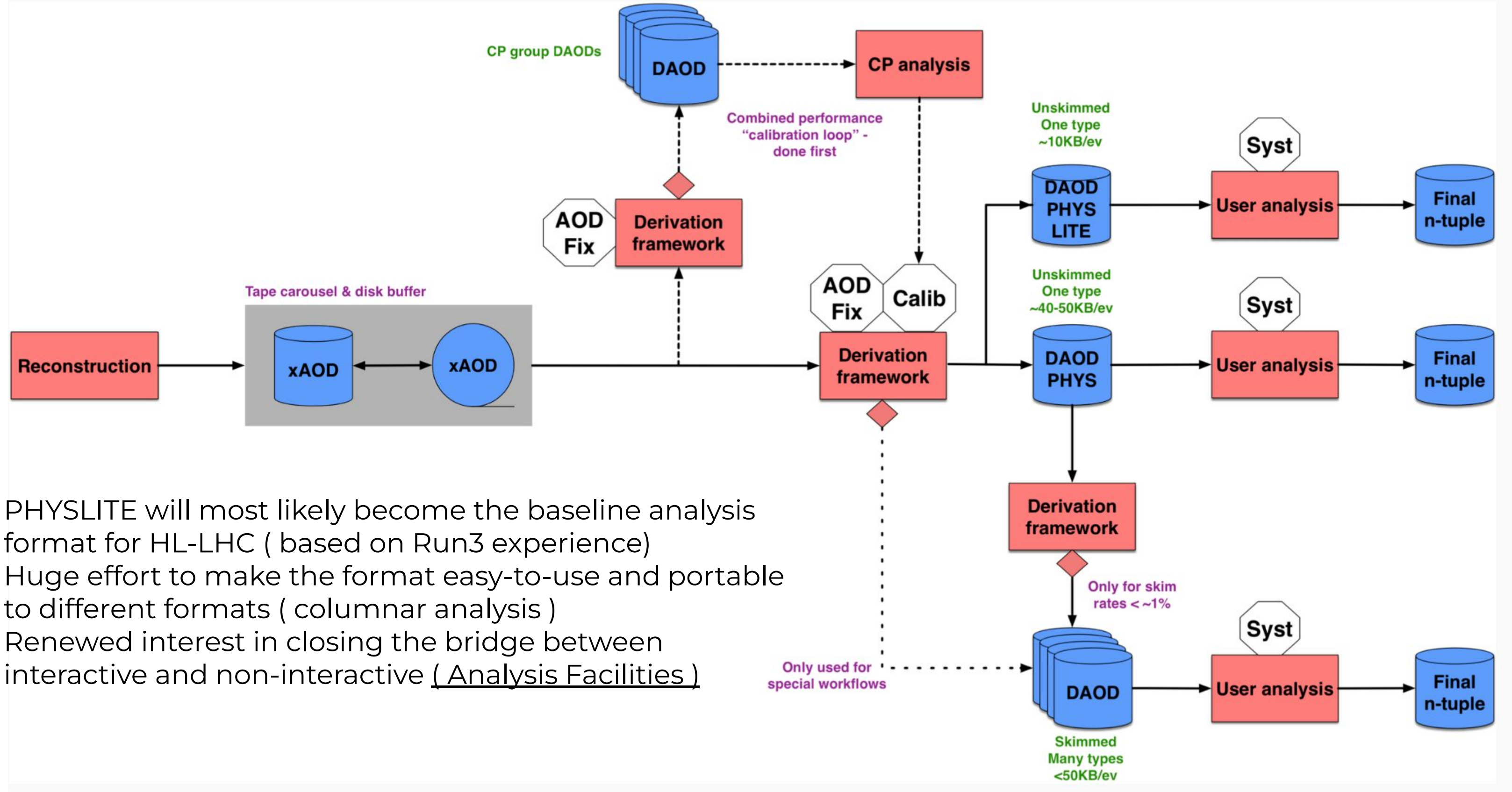
50% of AOD on tape

4 replicas of derived data formats, 2 versions kept

❑ Run2 AM requires 132 PB
❑ Run3 AM would require ~85 PB

❑ The new model opens to the possibility of the creation of Analysis Facilities (few PB of disk space)

A new analysis model



- PHYSLITE will most likely become the baseline analysis format for HL-LHC (based on Run3 experience)
- Huge effort to make the format easy-to-use and portable to different formats (columnar analysis)
- Renewed interest in closing the bridge between interactive and non-interactive (Analysis Facilities)

Conclusions

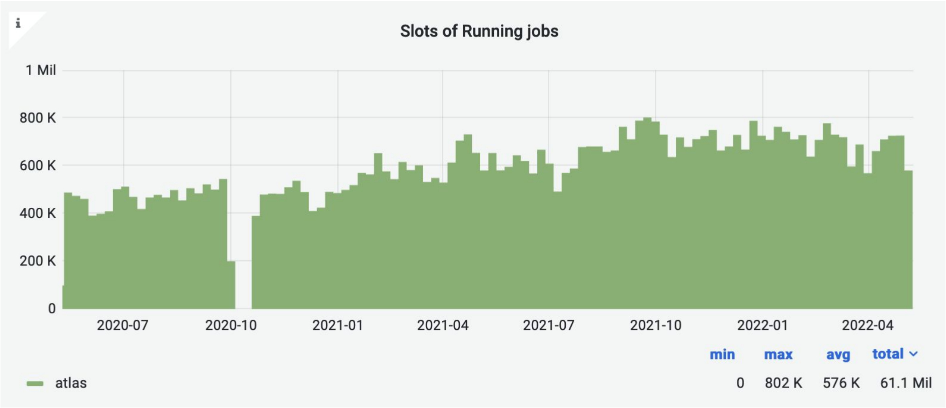
- ❑ ATLAS collaboration has been running smoothly its computing tasks (despite the difficult working conditions imposed by the pandemic emergency) thanks to the dedication of a relatively small group of persons

- ❑ Huge effort to prepare the roadmap to match the demanding computing requirements for HL-LHC : tackling the problem from several different perspectives :
  - ❑ access to opportunistic computing resources like HPC, cloud, accelerators (e.g. GPU). No opportunistic disk !
  - ❑ optimisation of the available resources (taper carousel)
  - ❑ software improvements, portability
  - ❑ reconstruction/simulation improvements, more efficient analysis model

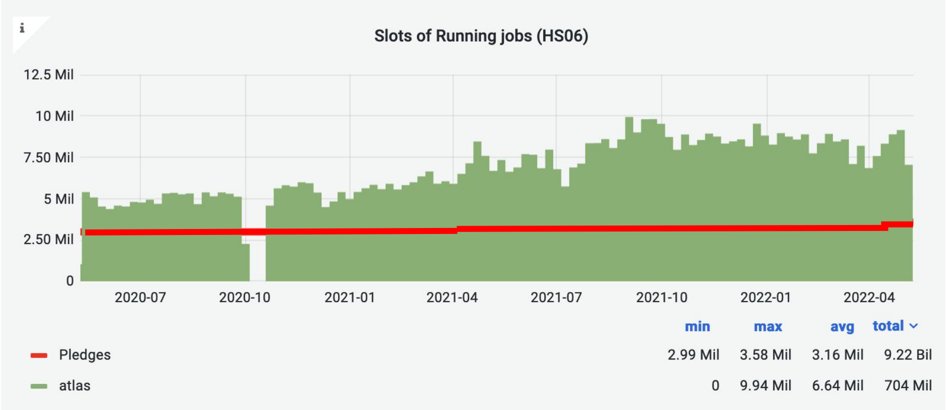
- ❑ ATLAS is continuously evolving the computing model and software to allow the full exploitation of the physics potential of the HL-LHC looking for the opportunities offered by the new hardware (portability) and software solutions on the market
  - ❑ Often the manpower availability is an issue in several critical areas

ATLAS Computing performance : all resources

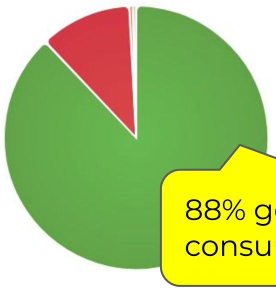
ATLAS experiment has been running jobs full speed over the last two years (1.5.2020-1.5.2022) :



	min	max	avg	total ⌄
🟩 atlas	0	802 K	576 K	61.1 Mil



	min	max	avg	total ⌄
🟥 Pledges	2.99 Mil	3.58 Mil	3.16 Mil	9.22 Bil
🟩 atlas	0	9.94 Mil	6.64 Mil	704 Mil



88% good (wallclock consumption)

- ❑ On average 576k jobs running in parallel (spikes > 1 M jobs running)
- ❑ overall efficiency ~ 90% (a bit higher if we exclude single users analysis jobs
- ❑ ATLAS has used up to 2.5 the computing pledge resources last year

ATLAS Computing performance

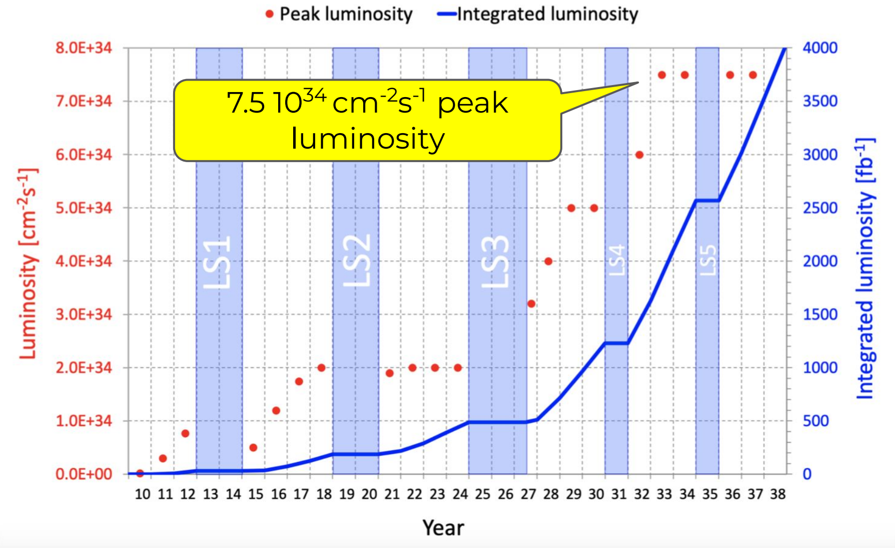
- ❏ Baseline: ATLAS implements the new data formats foreseen by the Run 3 analysis model, the multi-threaded software framework AthenaMT, and updates to the tracking code, but otherwise continues in largely the same way as in Run 2. In particular the CPU time per event for event generation, detector simulation and reconstruction is assumed to remain at the level currently achieved by applying the current software to the Phase-II detector simulation, and the mixture of generators and simulation remains the same;

- ❏ Conservative R&D: the research and development activities currently under way for Run 3 are assumed to be successful, including the data carousel, fast track reconstruction, lossy compression, and most of the detector simulation is done with fast simulation;

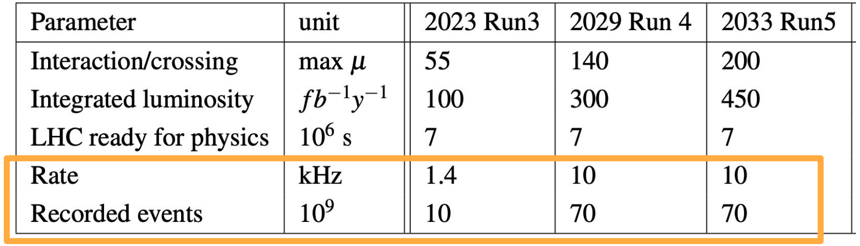
- ❏ AggressiveR&D: ATLAS implements new developments that very significantly improve the speed or storage volumes of workflows that currently are heavy consumers of resources, for example, porting of high-precision generators to GPUs, sharing events with CMS, or speeding up the full simulation either by software efficiencies or porting parts of the code to GPUs. Almost universal adoption by the physics groups of DAOD_PHYSLITE and development of very high quality fast simulation that could replace full simulation in almost all cases, would also fall into this category.

ATLAS Computing requirements for run3/4/5

- The physics potential of HL-LHC is enormous, in 2034 expect 5 times the total statistic collected up to now in previous runs
- The amount of data and experimental conditions will pose severe challenges to the computing model



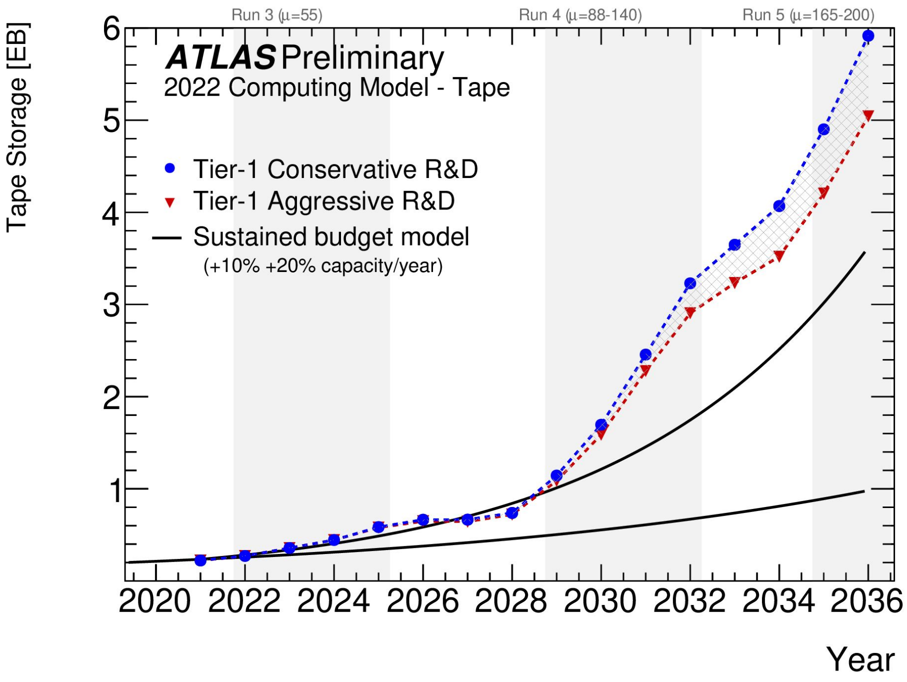
7.5 10^{34} cm^{-2}s^{-1} peak luminosity



Parameter	unit	2023 Run3	2029 Run 4	2033 Run5
Interaction/crossing	max μ	55	140	200
Integrated luminosity	$fb^{-1}y^{-1}$	100	300	450
LHC ready for physics	10^6 s	7	7	7
Rate	kHz	1.4	10	10
Recorded events	10^9	10	70	70

- ATLAS HL-LHC Computing Conceptual Design Report : projections of ATLAS computing requirements for Run3 and HL-LHC to fully exploit the machine physics potential is quite scaring !

ATLAS Computing requirements



The re-simulation workflow

A new workflow, MonteCarlo ReSimulation, was developed to minimize the resources needed to apply physics improvements to already generated FullSim HITS: the resources used by this workflow are 5-10% of the ones which would have been needed if we should have re-run the FullSim completely.

- ❏ Quasi-stable particles (b-hadrons, τ) not propagated correctly in Geant4 → impact on performance
- ❏ Resimulation of events with long living high-pT particles. Only a fraction of the events (varying for different samples) is processed
- ❏ Current status: total production: ~16 B events passed through re-simulation

More on VEGA setup

❑ Vega site, 3 PQs, aCT push mode:
    ❑ Vega (1GB/thread, 16-core)
    ❑ Vega_largemem(4GB/thread, 16-core)
    ❑ VEGA_MCORE (simul only, 64-core), testing 16-core as well

❑ NDGF-T1 storage endpoint
    ❑ + CERN-PROD_DATADISK for simul inputs

❑ 2 ARC-CEs, 6 ARC data-delivery, 6 squids
    ❑ Arex optimized (6.13 coming) for memory usage and transfer throughput

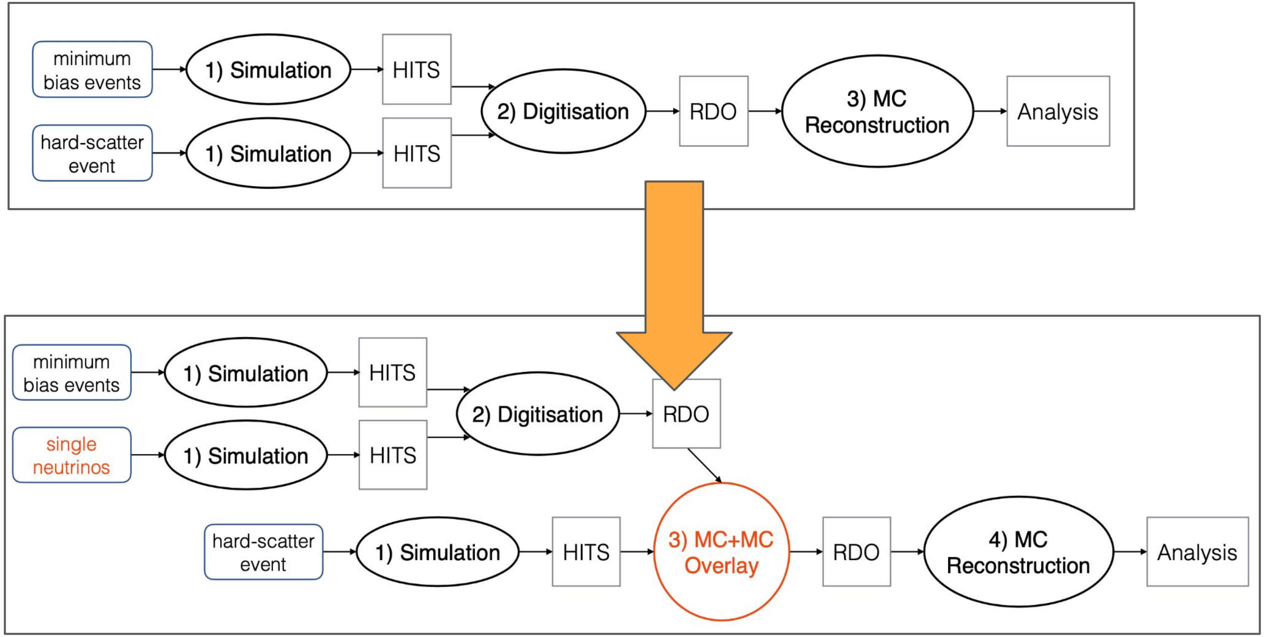
❑ Node outbound through 100 Gb/s NAT (ipv4, ipv6)

❑ Nodes: cvmfs + local nvme, 50GB file swap added for stability

New pileup modelling

The ATLAS detectors readout is sensitive to up to 39 LHC bunch-crossings (BCs) around the trigger BC.

- The average number of interactions that must be included is ~ 1560 (assuming 40 average interactions per bunch-crossing) : simulating this many extra interactions for each hard-scatter event would be prohibitive

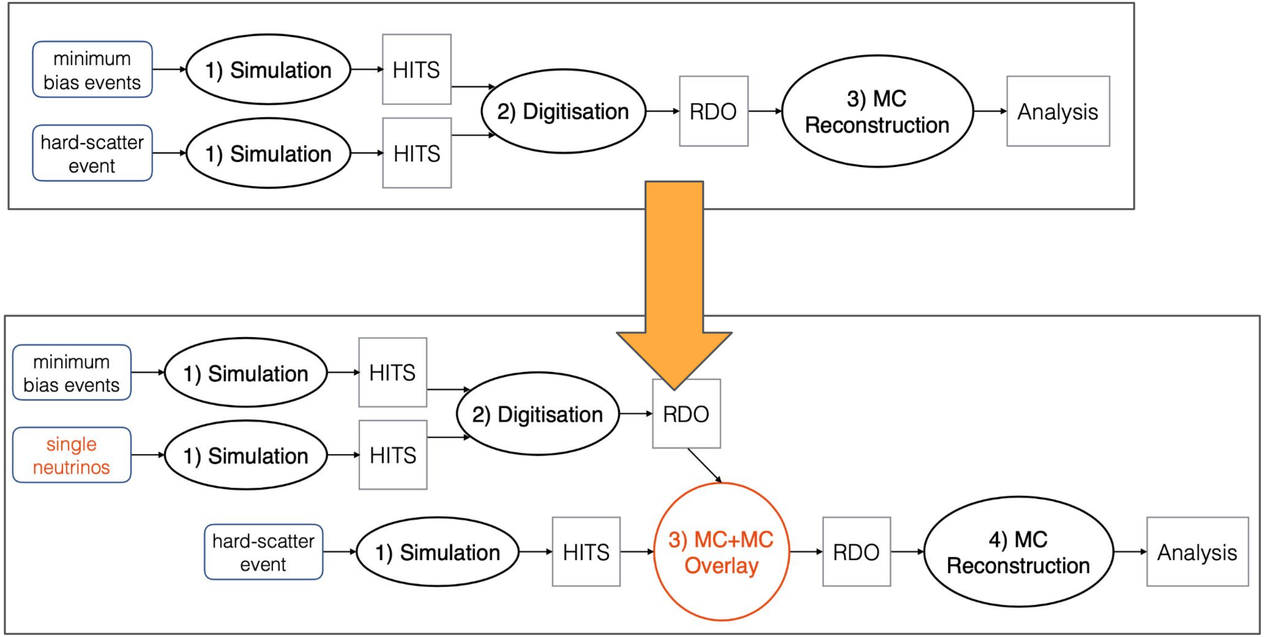


- Simulate hard-scatter and minimum bias evts with G4
- Presampling: a large sample (1B) of combined pile-up events is produced from simulated minimum bias events during a separate digitisation step.
- Each simulated hard-scatter event is digitised and combined with an event sampled from these pileup datasets.
- CPU and I/O requirements of the digitisation are significantly lower and have almost no dependence on μ.
- Pre-mixed pile-up events can be reused for different hard-scatter samples

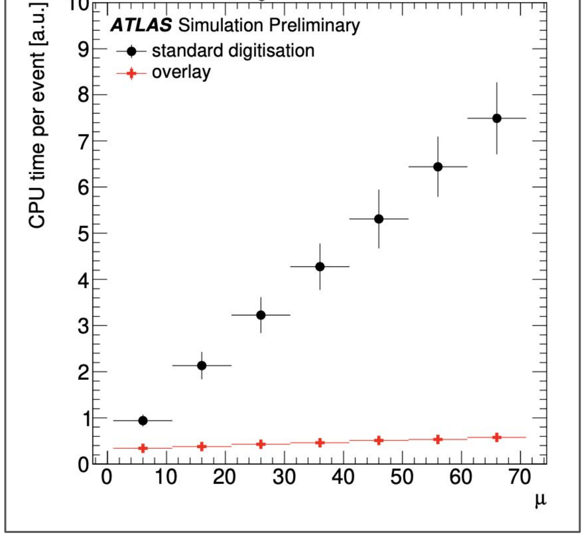
New pileup modelling

The ATLAS detectors readout is sensitive to up to 39 LHC bunch-crossings (BCs) around the trigger BC.

- ❏ The average number of interactions that must be included is ~ 1560 (assuming 40 average interactions per bunch-crossing) : simulating this many extra interactions for each hard-scatter event would be prohibitive



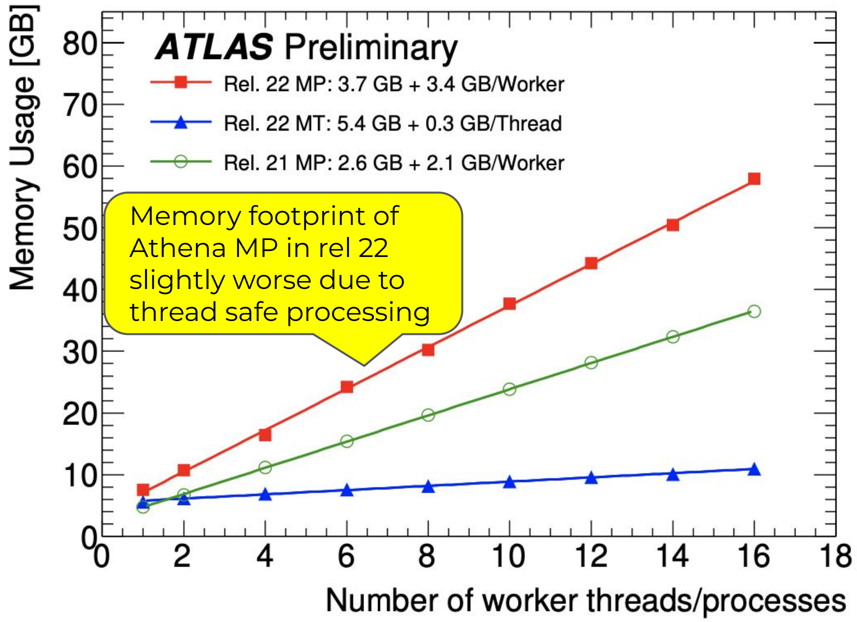
CPU time flat in μ !



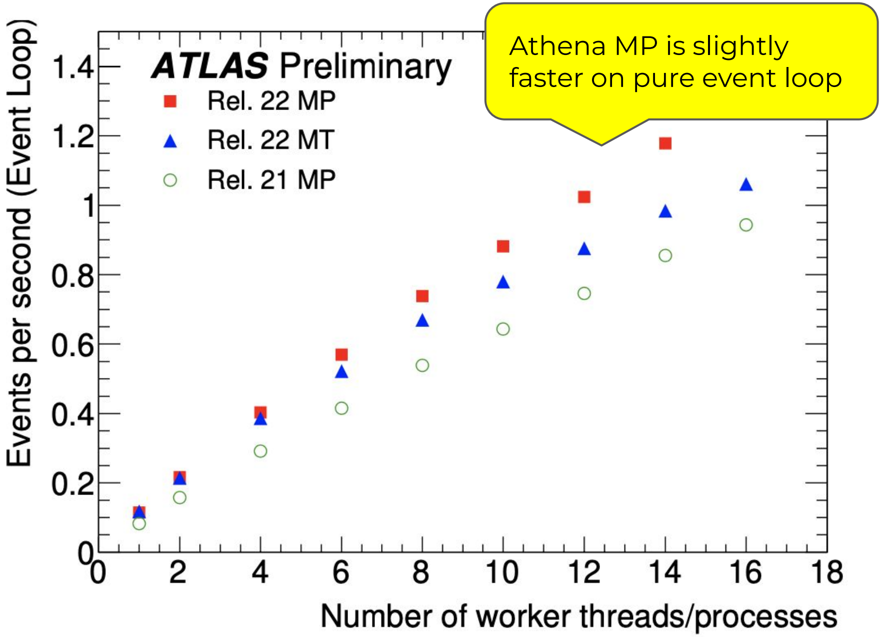
Multi-threaded reconstruction software (AthenaMT)

ATL-SOFT-PUB-2021-002

- ❑ The benchmark jobs use real data from run 357750 taken during 2018, with 250 events per worker process or thread.
- ❑ The data events have an average number of interactions per bunch crossing ⟨μ⟩ = 50, which is approximately that expected for the luminosity-leveling period during Run 3.
- ❑ Tests on an Intel®Xeon®CPU E5-2630 v3 at 2.40 GHz (16 cores no SMT) machine + 126 GB of memory.



Memory footprint of Athena MP in rel 22 slightly worse due to thread safe processing



Athena MP is slightly faster on pure event loop