
Meeting ATLAS-IT calcolo : introduzione


(A. Doria, L. Carminati)

- ❑ Benvenuti a tutti !

Operations

WEDNESDAY, 30 NOVEMBER		
14:30 → 14:50	Overview and future perspectives of the ATLAS Computing Speaker: Alessandro Di Girolamo (INFN)	🕒 20m
15:05 → 16:05	Introduction and sites round table Speakers: Alessandra Doria (Istituto Nazionale di Fisica Nucleare) , Leonardo Carminati (Istituto Nazionale di Fisica Nucleare)	🕒 1h
16:15 → 16:35	Migrazione a dCache Speakers: Alessandra Doria (Istituto Nazionale di Fisica Nucleare) , Bernardino Spisso (Istituto Nazionale di Fisica Nucleare)	🕒 20m
16:40 → 17:00	Migrazione a token e possibili upgrades futuri Speaker: Alessandro De Salvo (Istituto Nazionale di Fisica Nucleare)	🕒 20m
17:05 → 17:35	Coffee break	🕒 30m
17:35 → 17:45	News dal CNAF Speaker: Lucia Morganti (Istituto Nazionale di Fisica Nucleare)	🕒 10m
17:55 → 18:15	Tape system del CNAF Speaker: Enrico Fattibene (Istituto Nazionale di Fisica Nucleare)	🕒 20m
18:20 → 18:40	CNC : stato e prospettive future Speaker: Gianpaolo Carlino (Istituto Nazionale di Fisica Nucleare)	🕒 20m

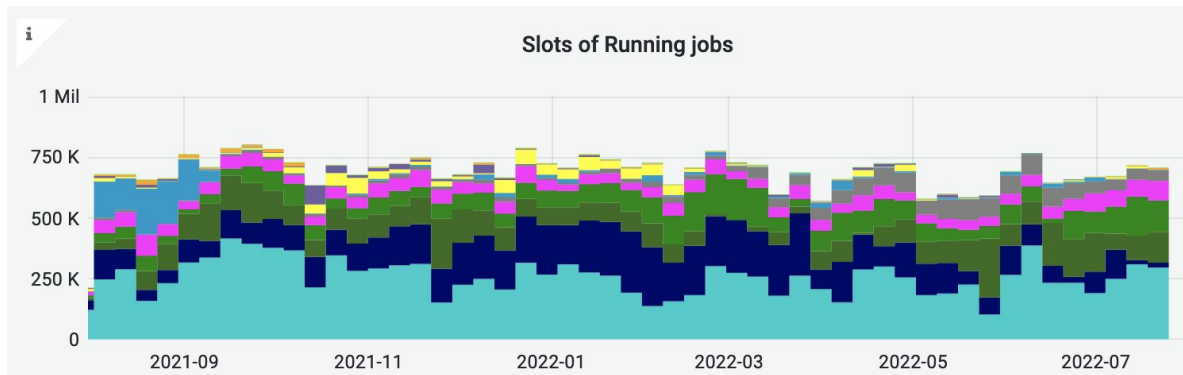
Contributi delle persone della nostra
comunita' al lavoro di ATLAS/ATLAS-Italia

THURSDAY, 1 DECEMBER		
09:05 → 09:25	INFN-Cloud as analysis facility Speaker: Caterina Maria Luigia Marcon (Istituto Nazionale di Fisica Nucleare)	🕒 20m
09:45 → 09:55	Experience on GPU Speaker: Bernardino Spisso (Istituto Nazionale di Fisica Nucleare)	🕒 10m
10:05 → 10:25	Esperienze di calcolo su HPC Speaker: Lorenzo Rinaldi (Istituto Nazionale di Fisica Nucleare)	🕒 20m
10:35 → 11:05	Coffee break	🕒 30m
11:05 → 11:25	Updates on EventIndex et. al Speaker: Dario Barberis (Istituto Nazionale di Fisica Nucleare) 	🕒 20m
11:35 → 11:55	Condition data access e container platform Speaker: Alessandro De Salvo (Istituto Nazionale di Fisica Nucleare)	🕒 20m
12:05 → 12:25	Data compression algorithms Speaker: Caterina Maria Luigia Marcon (Istituto Nazionale di Fisica Nucleare)	🕒 20m

- ❑ Agenda “flessibile” : prendiamoci il tempo necessario per discutere per bene tutti gli argomenti
- ❑ Cena ore 20 (flessibile)

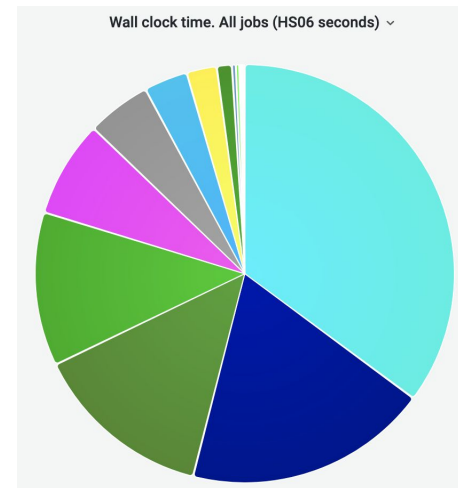
- ❑ Overview delle risorse di computing utilizzate dall'esperimento ATLAS
- ❑ Computing ATLAS svolto sulle risorse Italiane
- ❑ Stato e funzionamento delle nostre risorse : performance, pledge, gare
- ❑ Prossimi passi per garantire il pledge
- ❑ Attività di computing e challenges per il computing ATLAS (vedi Ale Di Girolamo)

- ❑ Circa 690k (average) computing slots ogni giorno per tutto il periodo in esame (pledged and unpledged, all Tiers resources together)
 - ❑ Importante sforzo che ricade su pochi esperti chiave (sia ADC dia DDM)



- ❑ Attività tipica di periodi di analisi (senza data taking): Intensa attività di produzione MC (~70% includendo generazione, simulazione e ricostruzione) e analisi dati (~25% tra derivations e analisi vera e propria)
- ❑ Efficienza (HS06 seconds) ~ 90%

	Value	Percent
MC Simulation Full	91 Tri	36%
MC Reconstruction	49 Tri	19%
MC Event Generation	36 Tri	14%
Group Production	31 Tri	12%
User Analysis	19 Tri	7%
Group Analysis	13 Tri	5%
MC Simulation Fast	9 Tri	3%
Data Processing	6 Tri	2%
MC Resimulation	3 Tri	1%



Utilizzo risorse ATLAS (luglio 2021 - luglio 2022) : disco e tape

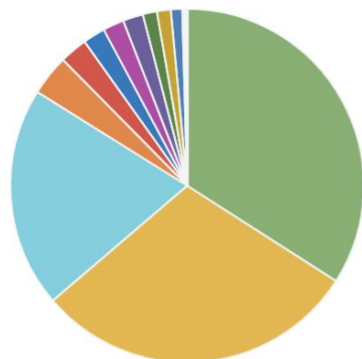
Presentazione referaggio calcolo

- ❑ Tape occupato ~ 360 PB (250M files) su 390 PB: principalmente RAW data, AOD e HITS

- ❑ Disco occupato ~ 280 PB (850M files) su 290 PB: principalmente dati derivati (DAOD) e AOD

TAPE

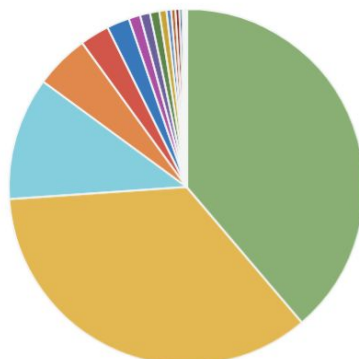
Tape size



	total	percentage
RAW	123 PB	34%
AOD	107 PB	30%
HITS	73 PB	20%
DRAW	13 PB	4%
DAOD	9 PB	3%
NTUP	7 PB	2%
ESD	7 PB	2%
DESD	7 PB	2%
no_name	5 PB	1%
EVNT	5 PB	1%
RDO	4 PB	1%

DISCO

Disk size



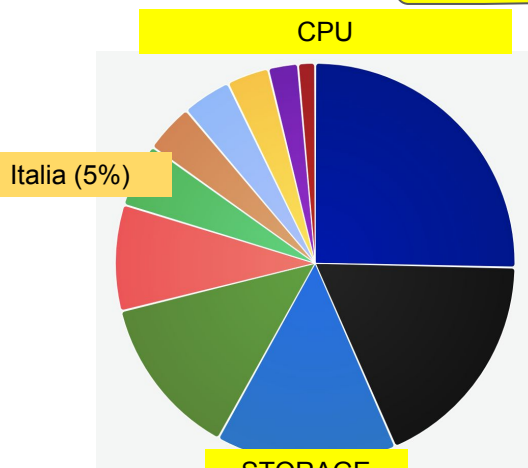
	total	percentage
DAOD	110 PB	39%
AOD	100 PB	35%
HITS	32 PB	11%
EVNT	14 PB	5%
RDO	8 PB	3%
RAW	6 PB	2%
log	3 PB	1%
no_name	3 PB	1%
ESD	2 PB	1%
DESD	2 PB	1%

Contributo italiano computing ATLAS (luglio 2021 - luglio 2022)

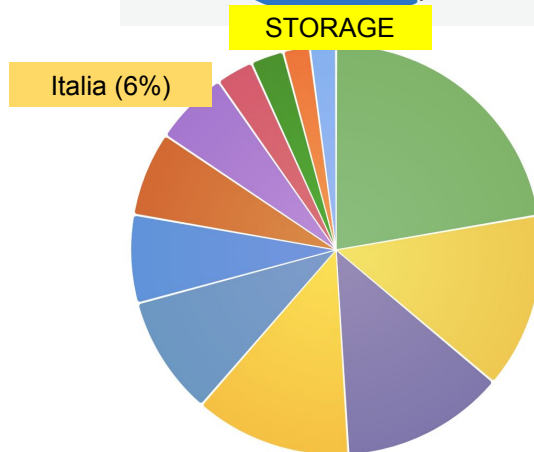
Presentazione referaggio calcolo

Contributo italiano alle attività di computing di ATLAS :

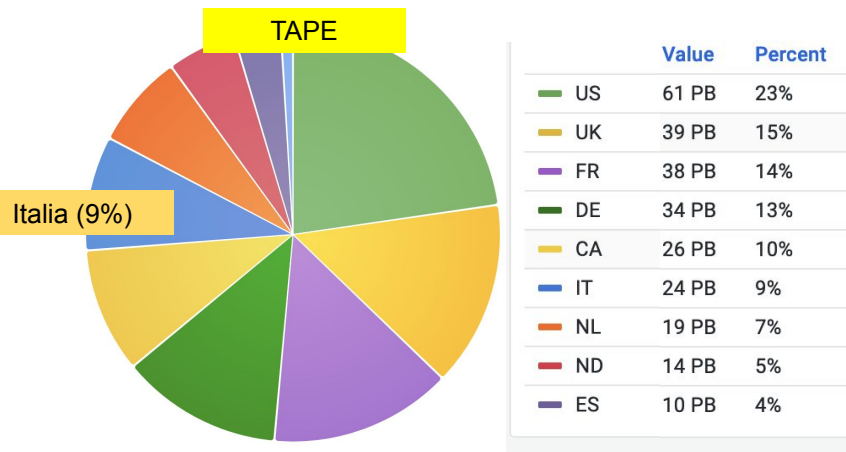
- ❑ CPU : overall ~ 5% wallclock time totale in HS06-sec (T1+T2+T3, solo risorse grid, T0 escluso)
 - ❑ sotto share : consistente overpledge di alcune cloud, sofferenza a inizio anno
- ❑ Disco occupato (datadisk, calibdisk, groupdisk e localgroupdisk) : overall ~6% (escludendo T0)
 - ❑ leggermente sotto share alcune risorse ancora da mettere in linea
- ❑ Tape (T1 only) : ~9%
 - ❑ perfettamente in linea con lo share (9%)



US	31 Tri	25%
DE	22 Tri	18%
FR	18 Tri	15%
UK	16 Tri	13%
CA	11 Tri	9%
IT	6 Tri	5%
NL	5 Tri	4%
ND	5 Tri	4%
ES	4 Tri	3%
RU	3 Tri	2%

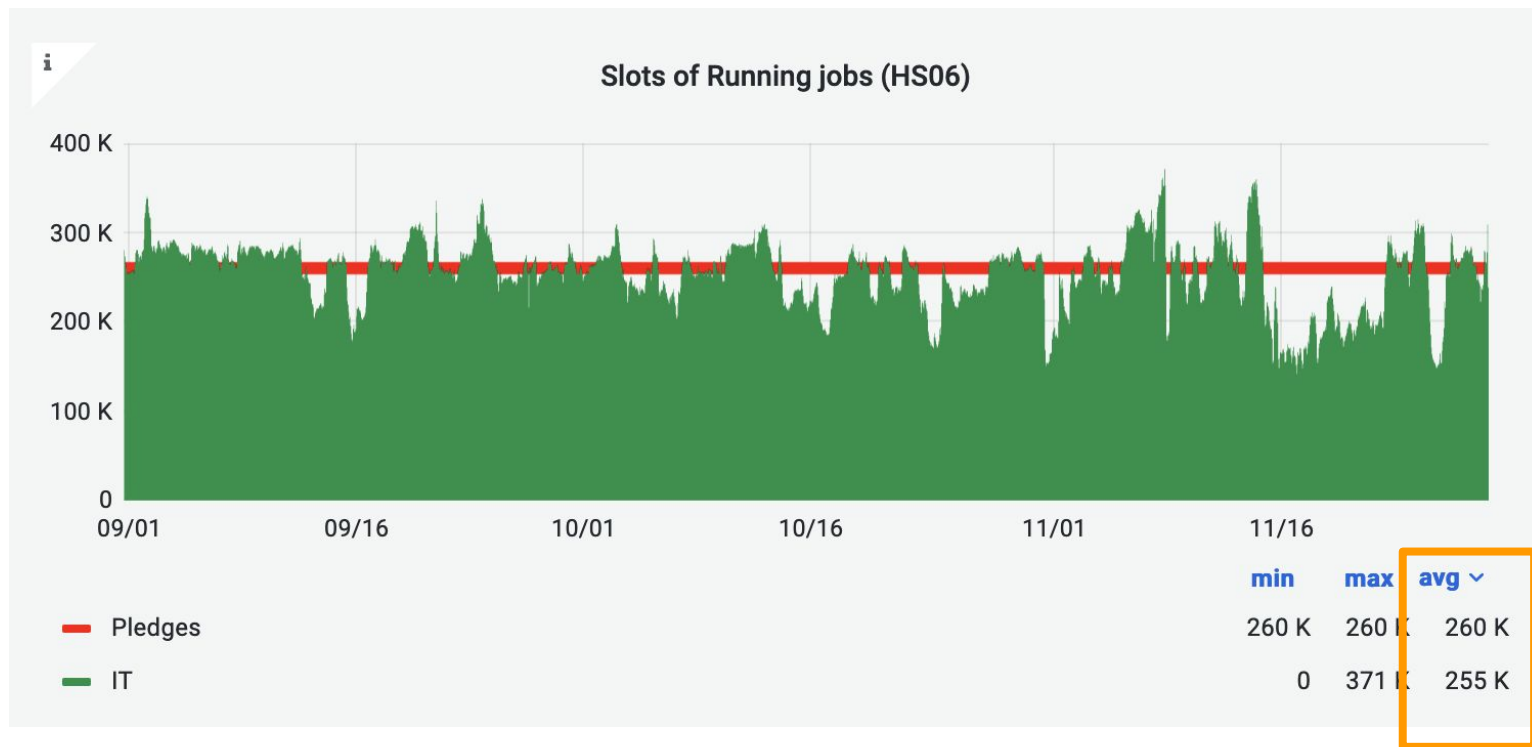


DE	41 PB	14%
FR	38 PB	13%
UK	37 PB	12%
CERN	28 PB	9%
ND	21 PB	7%
CA	20 PB	7%
IT	17 PB	6%
NL	9 PB	3%
ES	8 PB	3%
RU	6 PB	2%



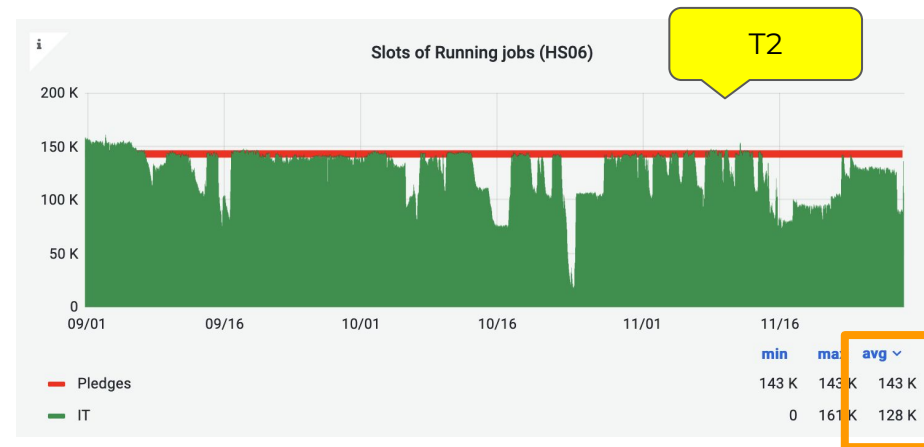
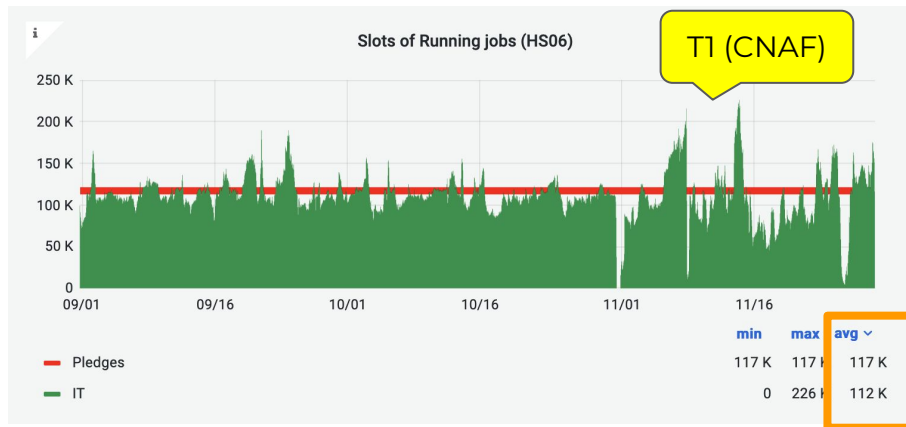
US	61 PB	23%
UK	39 PB	15%
FR	38 PB	14%
DE	34 PB	13%
CA	26 PB	10%
IT	24 PB	9%
NL	19 PB	7%
ND	14 PB	5%
ES	10 PB	4%

Performance cloud (CPU) (ultimi tre mesi)

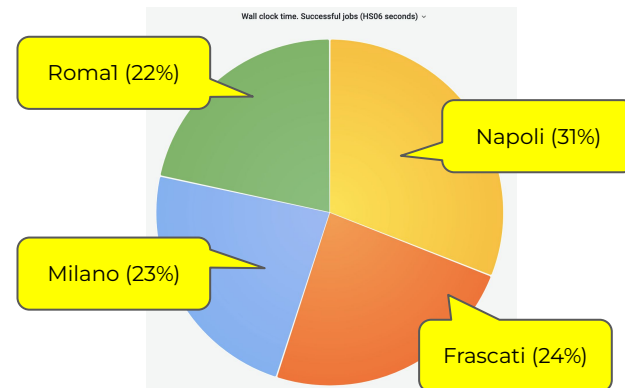


- ❑ A livello di cloud integrato negli ultimi tre mesi siamo praticamente a pledge

Performance cloud (CPU) (ultimi tre mesi)



- ❑ CNAF e' stabilmente a pledge !
- ❑ Contributo dei T2 in leggera sofferenza (non c'e' un sito con dei problemi evidenti)
- ❑ T3 : ~ 15k HS06 forniti che compensano l'underpledge dei T2



Situazione e disco e TAPE

	Disco occupato (PB) (dati ottobre 2022)	Disco allocato (PB) (dati ottobre 2022)
Frascati	2.48	3.69
Milano	1.45	1.68
Napoli	3.89	4.26
Roma	1.53	1.74
tot.	9.35	11.37

Non ancora validati dai siti !

Importante intervento di IBISCO
per crisi disco ATLAS

- Pledge 2022 : 9.94 PB
- Pledge 2023 : 11.76 PB

	Disco occupato (PB) (dati ottobre 2022)	Disco allocato (PB) (dati ottobre 2022)	Tape (PB) (dati ottobre 2022)
CNAF	8.91	9.57	25.6

- Pledge 2022 (disco): 10.44 PB
- Pledge 2022 (tape): 24.48

- ❑ Storage:
 - ❑ La saga della gara unica da 2.2 PB (2020 !) conclusa : risorse online ?
 - ❑ Gara comune disco 2022 : richiesta inviata in giunta, aggiornamenti ?

- ❑ CPU :
 - ❑ Acquisti 2021 : ok a Roma1 in convenzione, Milano problemi con la convenzione (acquisto diretto)
 - ❑ Acquisti 2022 (Milano e Roma1) : non ancora finalizzati

Assegnazioni 2023

solo acquisto CPU, no disco

ATLAS	CPU crescita (HS06)	CPU dism. (HS06)	DISK crescita (TB-N)	DISK dism. (TB-N)	CPU (HS06)	DISK (TB-N)	CPU (k€)	DISK (k€)	Server (k€)	Totale (k€)
Frascati	3000	16866	0	0	19866	0	298,0	0,0	20,9	318,8
Milano	3000	8426	0	0	11426	0	171,4	0,0	12,0	183,4
Napoli	3000	0	0	0	3000	0	45,0	0,0	3,2	48,2
Roma1	3000	14521	0	0	17521	0	262,8	0,0	18,4	281,2
Totale	12000	39813	0	0	51813	0	777,2	0,0	54,4	831,6

News and todo list

- ❑ Studio del passaggio a DCache : lavoro e' in corso a Napoli
- ❑ Token
- ❑ Acquisti e nuovo hardware
 - ❑ Risorse CPU 2022
 - ❑ Preparazione gara disco FraRoMi 2022 (Roma1)
- ❑ Utilizzo risorse cloud per analisi : Caterina ha cominciato a lavorarci
- ❑ Utilizzo risorse HPC Italiane per ATLAS: Leonardo e Marconi100
 - ❑ Utilizzo Marconi100 per analisi (Lorenzo)
- ❑ Assegno CNAF per supporto ATLAS :
 - ❑ sensibilizzare possibili candidati presso le nostre sedi (neo-laureati o post-doc)
- ❑ Sito di documentazione ATLAS Italia computing
- ❑ Centro Nazionale Calcolo

Schedule

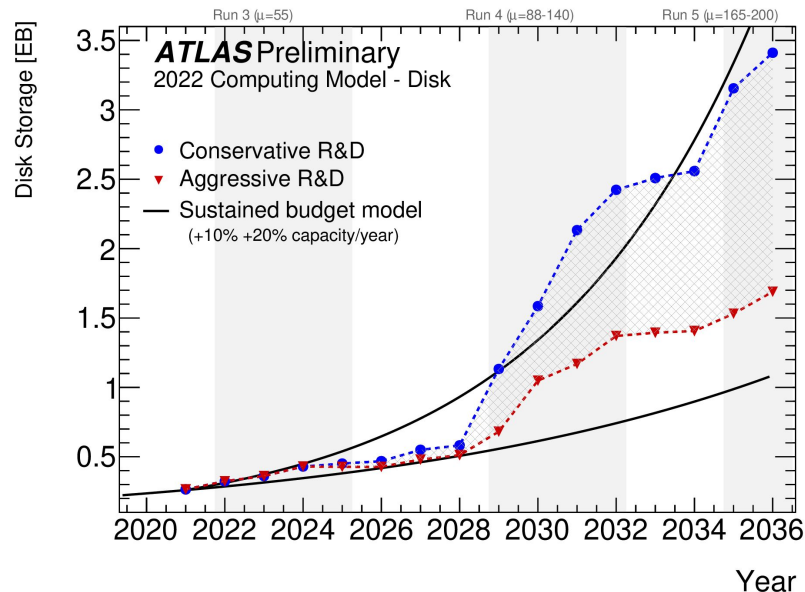
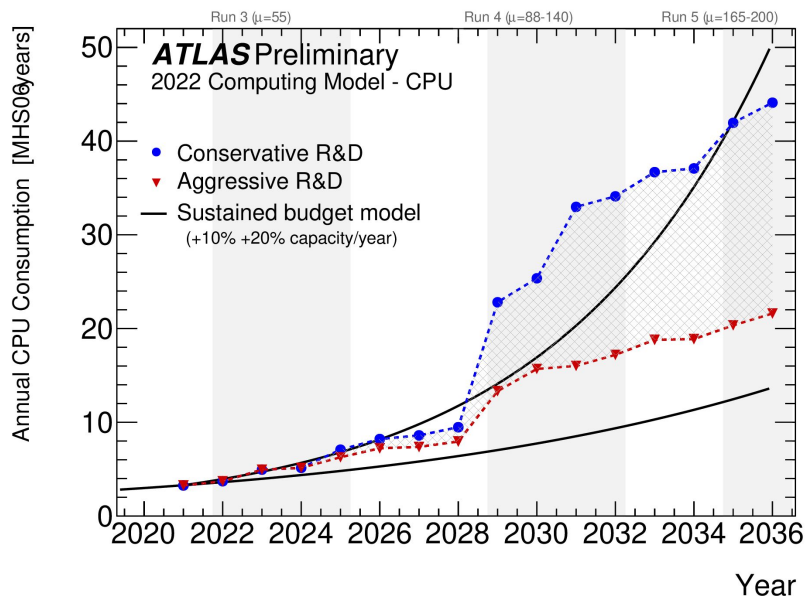
Intenso piano di lavoro per la comunita' computing per il 2022/2023

Data sample	Activity	2021				2022				2023			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Run 2 Data	DAOD Production and User Analysis												
	DAOD_PHYS Production												
	Reprocessing in Release 22												
Run 2 MC	New Production for Ongoing Analyses												
	DAOD_PHYS Production												
	Reprocessing in Release 22												
Run 3 Data	Tier-0 Reconstruction												
	Partial Reprocessing												
	Full Reprocessing												
	Delayed stream Reconstruction												
	DAOD Production and User Analysis												
Run 3 MC	Generation												
	Simulation												
	Reconstruction												
	Re-Reconstruction												
	DAOD Production and User Analysis												
Upgrade MC	Production and Analysis												
Heavy Ions	First pass Reconstruction												

 Full steam

 Partial

[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

- ❑ Exploitation of HPC resources : 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)
 - ❑ HPC jobs overall efficiency ~ 90%, very close to more 'standard' grid resource
 - ❑ VEGA very peculiar: able to run almost all workflows
- ❑ Data carousel in phase IV : use it for many workflows in parallel respecting computing share per workflow. Run Data Carousel jointly for more than one experiment
- ❑ Simulation optimisation :
 - ❑ Optimise G4 full simulation : expect up to 30% reduction
 - ❑ Pushing on fast calo simulation (ATLFAST III) : GAN based e/o parametrized CaloSim
 - ❑ Fast simulation for the ATLAS Tracker, FATRAS (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
- ❑ Reduce the impact of the analysis model on the storage : Introduced 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.

- ❑ I siti italiani hanno continuato a lavorare in modo efficiente per tutto il periodo luglio 2021-luglio 2022. Siti affidabili (elevata reliability e availability) e con efficienza paragonabile ai migliori centri ATLAS
- ❑ Varie attività in corso dentro l'esperimento per ottimizzare l'uso delle risorse (G4 optimisation, fast simulation, nuovo analysis model, tape carousel, AthenaMT, integrazione di risorse HPC)
- ❑ Principali attività in corso nella comunità italiana :
 - ❑ Studi per la migrazione dello storage (Frascati, Napoli, Roma) a DCache : preparazione di un testbed entro fine anno, migrazione effettiva il prossimo anno
 - ❑ Esplorazione attività di analisi su risorse INFN-Cloud recentemente partita (C. Marcon)
- ❑ Presentata la situazione delle risorse disponibili (gare) e le richieste per il 2023
 - ❑ Mantenuti gli stessi share delle scorse richieste
 - ❑ Nessuna azione immediata per quanto riguarda la crisi russa ma la situazione è monitorata costantemente

Proiezione obsolescenze

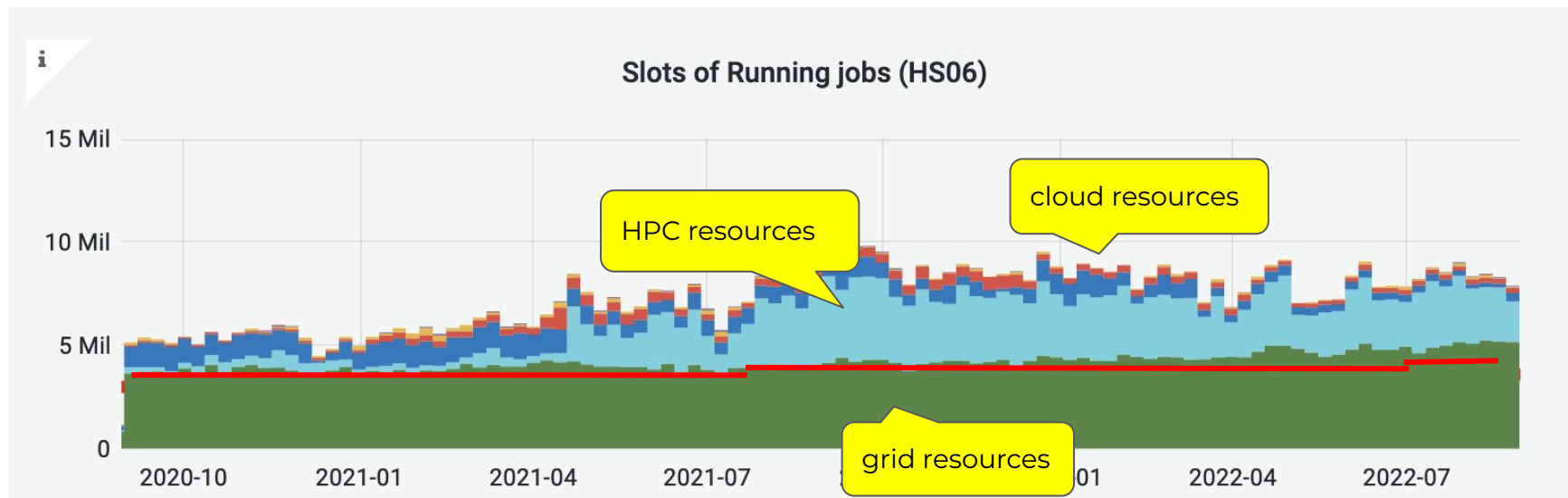
Proiezione obsolescenze CPU (HS06)

	2023	2024	2025	2026	2027	2028
Frascati	16866	10004		8440		
Milano	8003	5877			14000	8400
Napoli	0		2667			36200
Roma	8003			14000		8400
totale	32872	15881	2667	22440	14000	53000

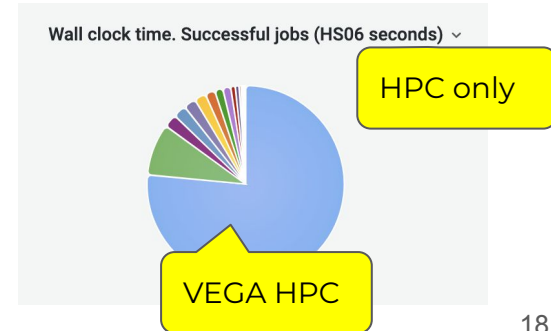
Proiezione obsolescenze disco (TBn)

	2023	2024	2025	2026	2027	2028
Frascati	0			561		1800
Milano	0	800		850		
Napoli	0			500		3500
Roma	0	800				
totale	0	1600	0	1911	0	5300

ATLAS Computing per type of resources

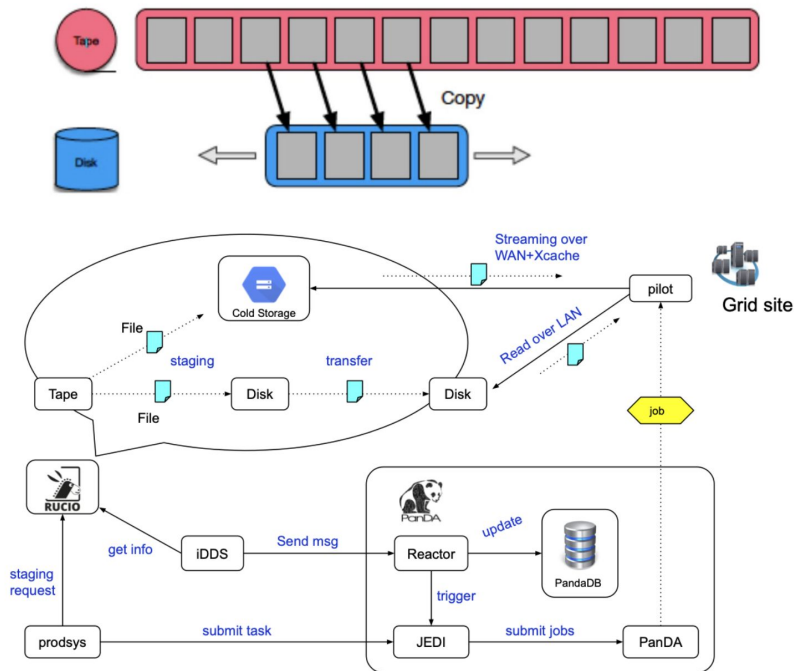


- ❑ ATLAS has used up to 2.5 the computing pledge resources last year
- ❑ Impressive contribution from HPC resources, mainly (~75% of the full HPC resources used by ATLAS) from opportunistic access to VEGA (Maribor, Slovenia within the EuroHPC program)
- ❑ HPC jobs overall efficiency ~ 90%, very close to more 'standard' grid resource
- ❑ VEGA very peculiar: able to run almost all workflows



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



Future of fast simulation : FastChain

- ❑ Optimise G4 full simulation : expect up to 30% reduction
- ❑ Pushing on fast (calo) simulation ATLFAST III : GAN based e/o parametrized CaloSim
- ❑ 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

- ❑ AODs and DAODs which are the two formats taking more than 70% of the disk space today
- ❑ 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 PHYS	DAOD PHYS LITE	AOD	DAOD	DAOD PHYS	DAOD PHYS 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

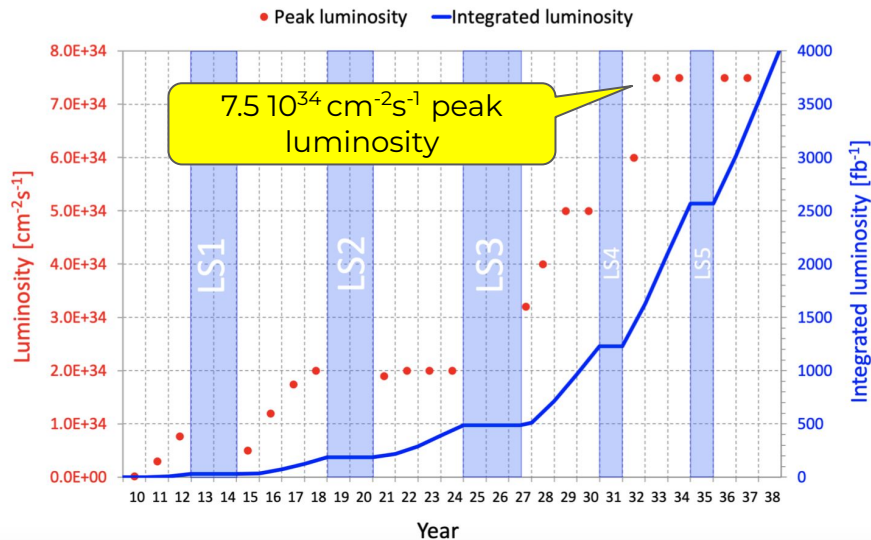
- ❑ The new model opens to the possibility of the creation of Analysis Facilities (few PB of disk space)

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;
- ❑ Aggressive R&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

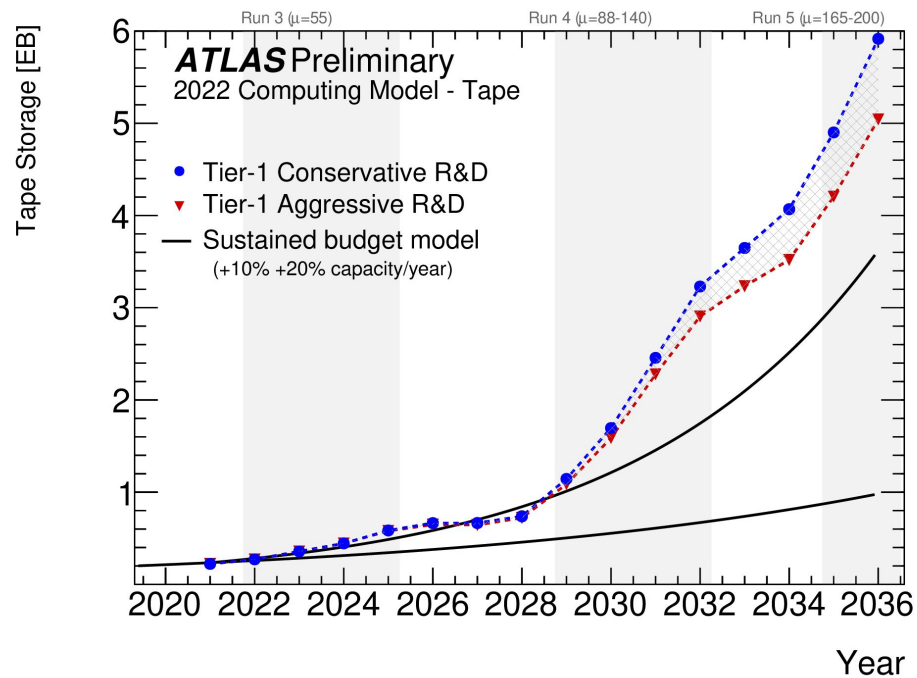
- ❑ The physics potential of HL-LHC is enormous, in 2034 expect 5 times the total statistic collected up to now in previous runs
- ❑ The amount of data and experimental conditions will pose severe challenges to the computing model



Parameter	unit	2023 Run3	2029 Run 4	2033 Run5
Interaction/crossing	$\text{max } \mu$	55	140	200
Integrated luminosity	$\text{fb}^{-1}\text{y}^{-1}$	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

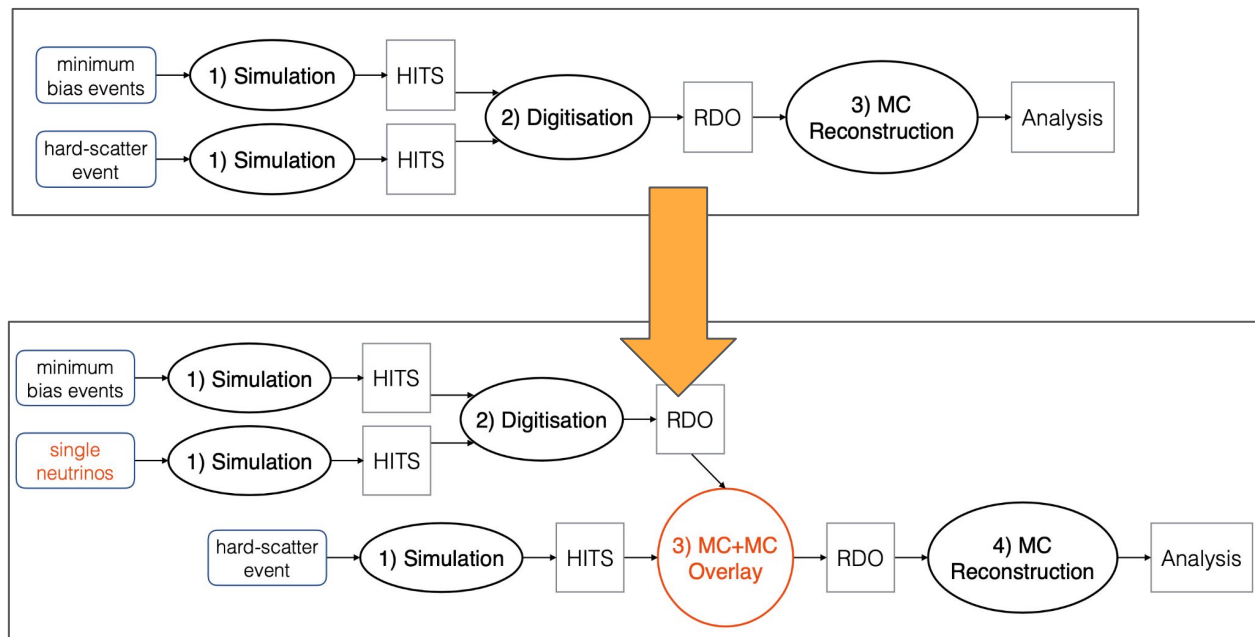
More on VEGA setup

- ❑ 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

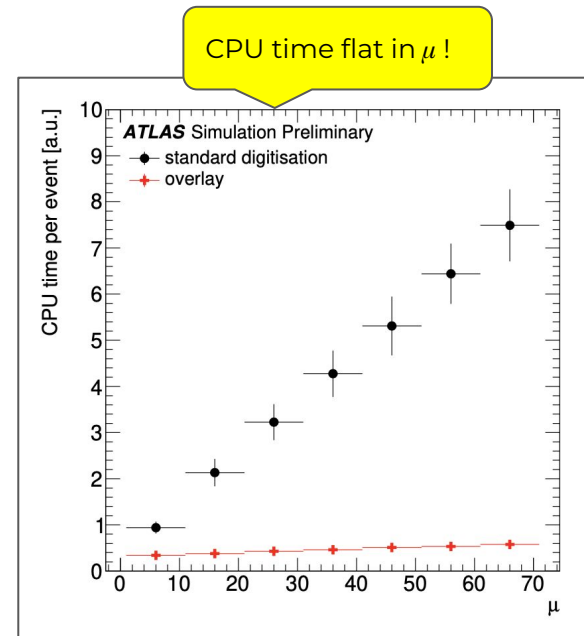
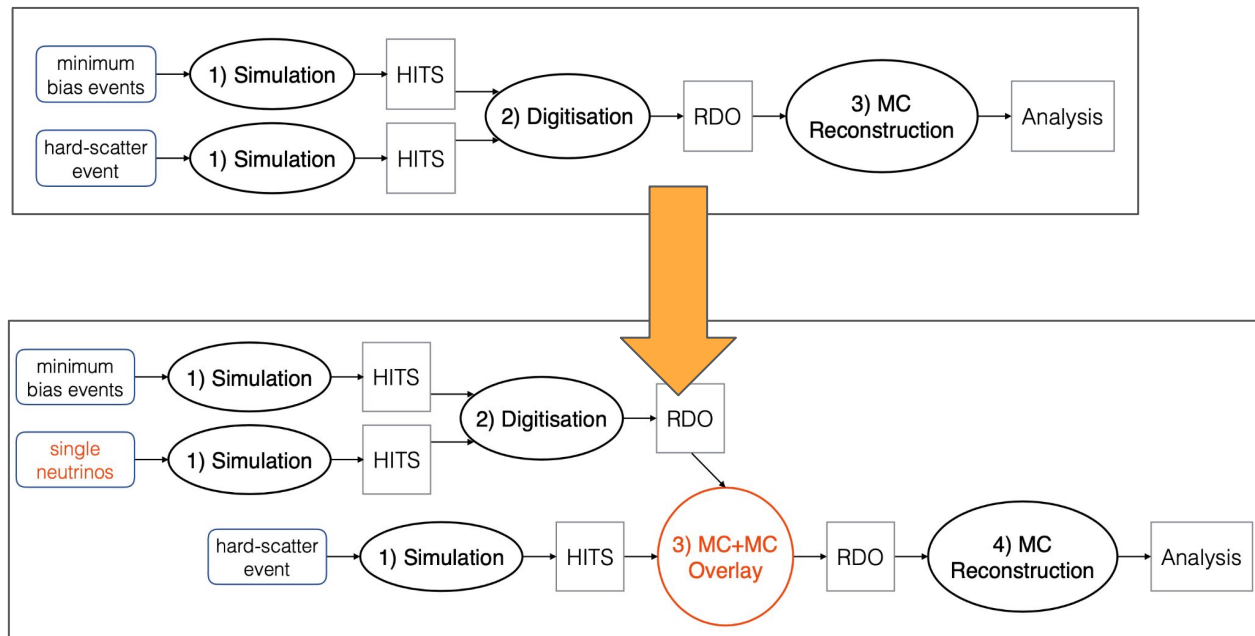


- ❑ Simulate hard-scatter and minimum bias evts with G4
- ❑ Presampling: a large sample (1B) of combined pile-up events is produced from simulated minimum bias events during a separate digitisation step.
- ❑ Each simulated hard-scatter event is digitised and combined with an event sampled from these 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

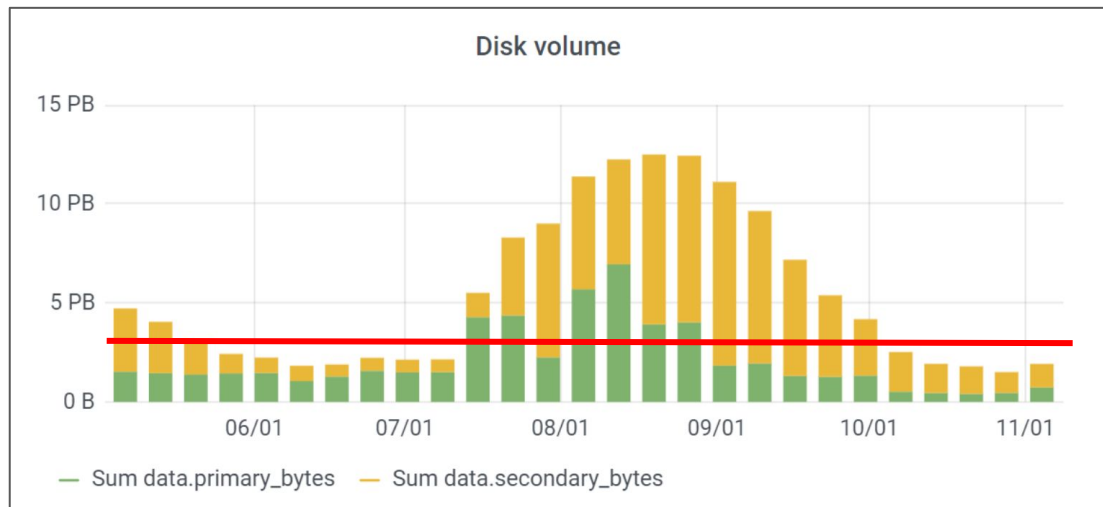
The ATLAS detectors readout is sensitive to up to 39 LHC bunch-crossings (BCs) around the trigger BC.

- ❑ 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



- ❑ 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

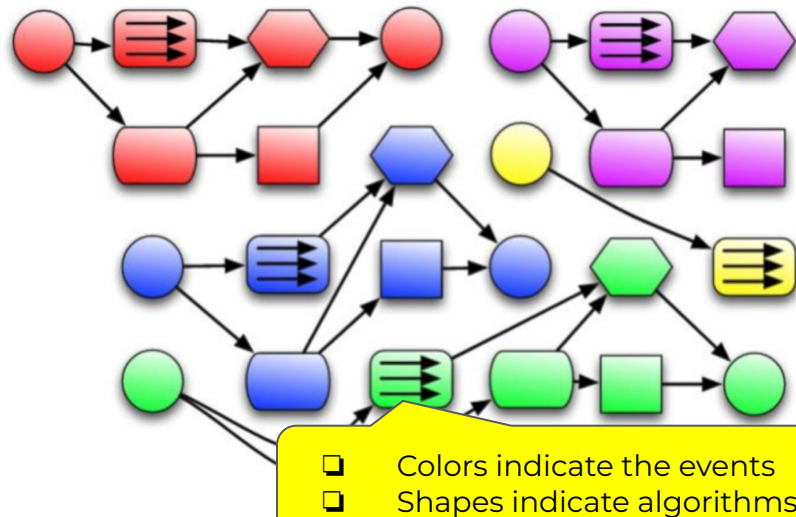


The new Athena release 22 (used since 2021 for reprocessing of Run 2 MC and data, as well as for Run3 data taking and MC simulations) is able to offer both multi-process and multi-thread parallelism.

- ❑ 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:

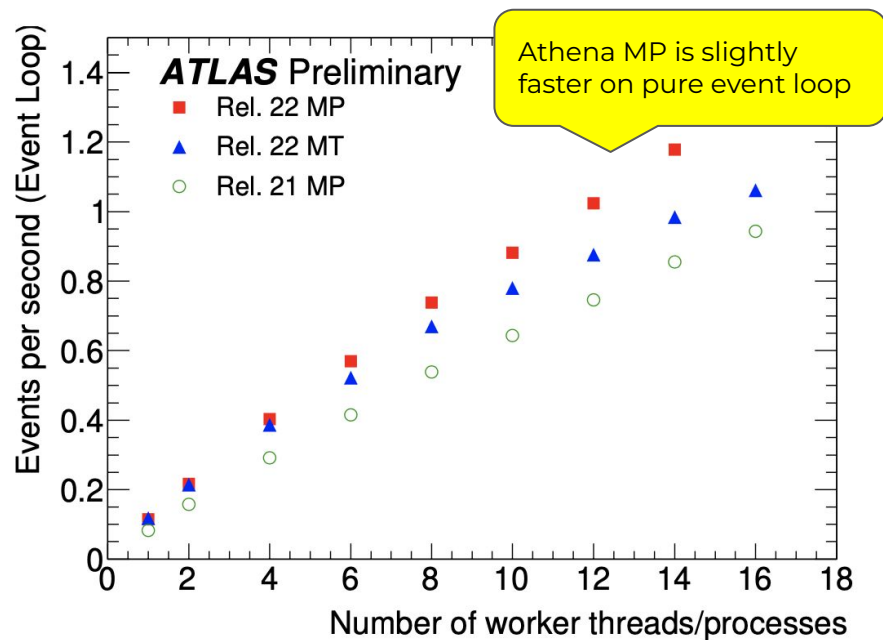
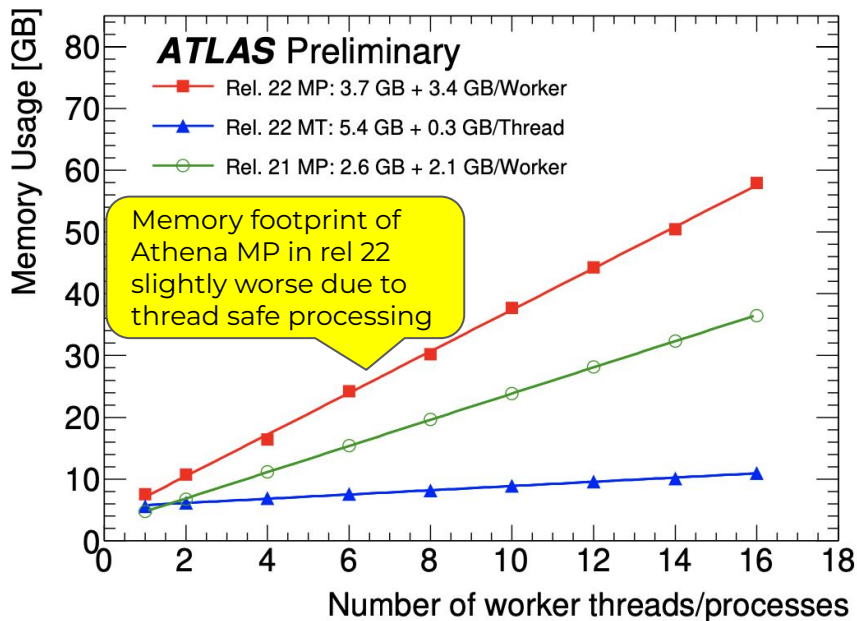
- ❑ multiple threads cannot write to the same memory at the same time;
- ❑ threads must not attempt to read memory that is actively being written
- ❑ algorithms must be scheduled such that all input is fully available before they run.

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



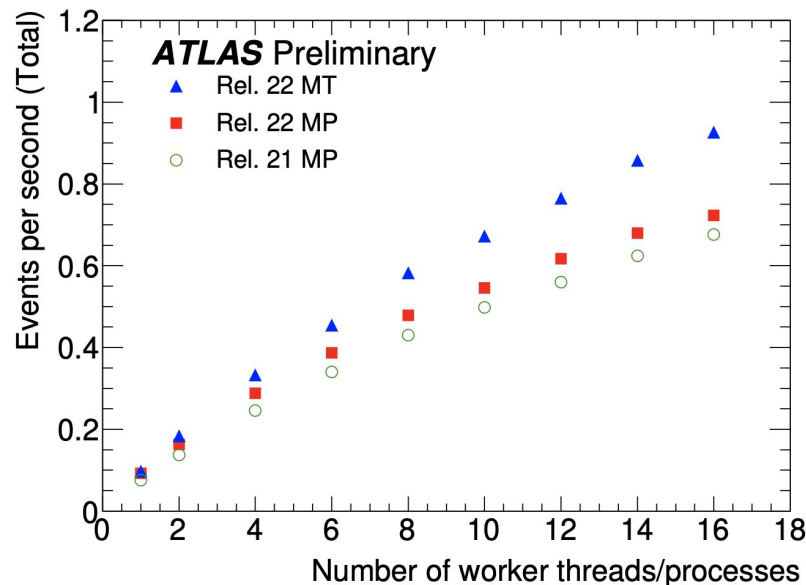
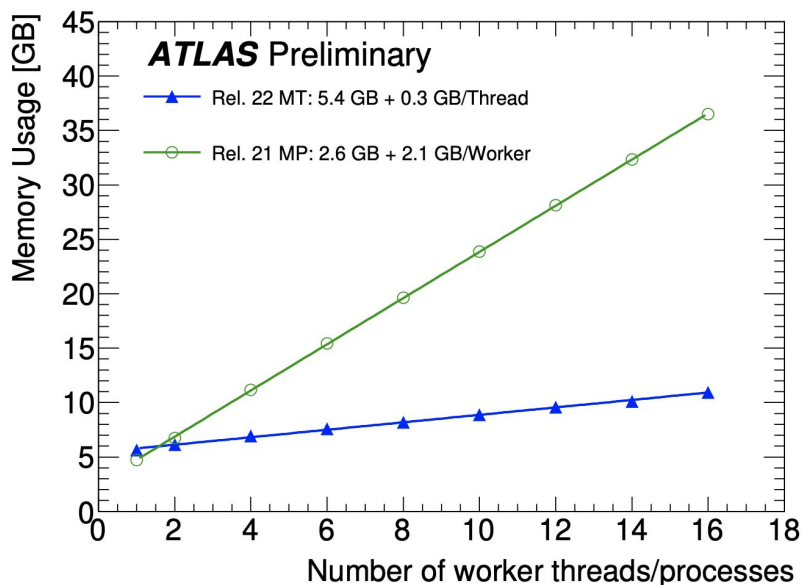
- ❑ Multi-threaded reconstruction software allows a better exploitation of opportunistic resources (eg HPC)

- ❑ 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.



AthenaMT (MultiThreading) : uno degli achievement fondamentali degli sviluppi software degli ultimi anni

1. L'ultima release (rel 22) del software di ricostruzione di ATLAS (Athena) che verra' utilizzato di default per reprocessing run2 e per il run3 e' basata su multi-thread
2. Considerevole risparmio di memoria (all threads share the same allocated memory)



Highlights (II): simulation optimisation

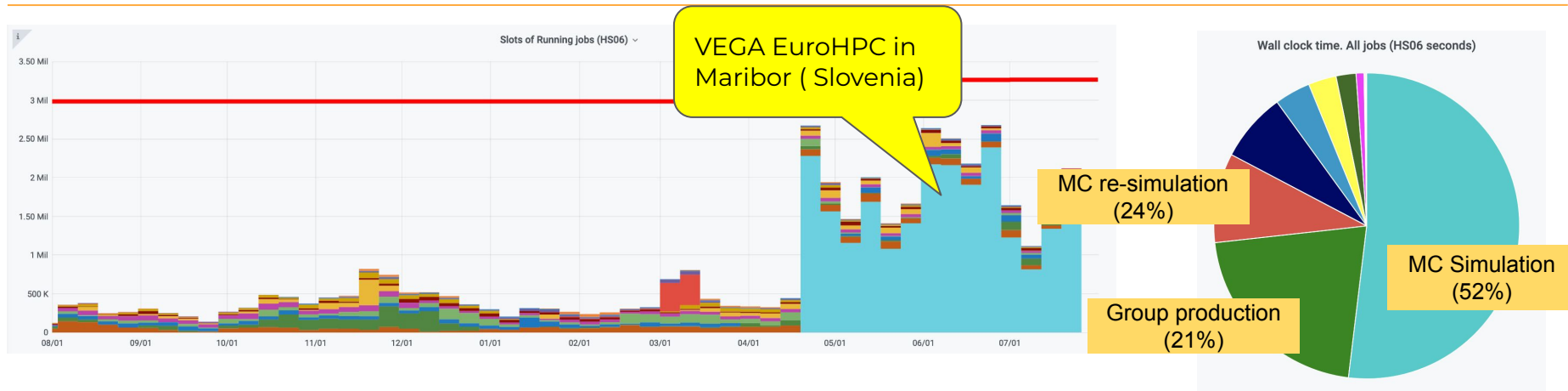
1. A Geant4 (G4) Optimisation Task Force (TF) has been setup for optimising the performance of the ATLAS G4 simulation software:
 - 1.1. Taking advantage of **intrinsic performance optimizations** coming with newer Geant4 versions
 - 1.2. Optimization with **tuning of G4 parameters** (physics models, physics lists per regions)
 - 1.3. **Neutron** and **Photon Russian Roulette + EM range cuts** (ongoing physics validation)
 - 1.4. **Geometry** optimisations (new EMEC variants + R&D on ML guided steppers in geometry)
 - 1.5. **Magnetic field** tailored switch-off
 - 1.6. **Geant4 linking** as static library (a.k.a. Big library)
2. 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. This workflow was applied on 16B events.

Stima : si riesce a ottenere un miglioramento della velocità della fullsimulation ~ 20-25%

Highlights (III) : data carousel

- ❑ More dynamic use of the TAPE capabilities: “by ‘data carousel’, we mean an orchestration between workflow management (WFMS), data management (DDM/Rucio) and tape services whereby a bulk production campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X% ($X \sim 5\%-10\%$) of inputs onto buffer disk, such that only X% of inputs are pinned on disk at any one time”
 - ❑ Retrieve a fraction of RAW data, process them, delete from disk, start again
- ❑ ATLAS started the Data Carousel R&D in June, 2018, to study the feasibility to get inputs from tape directly, for various ATLAS workflows, such as derivation production and RAW data re-processing.
 - ❑ CNAF is actively (and very efficiently!) participating in the Data Carousel activities
- ❑ Complete Run2 RAW ($\sim 18.5\text{PB}$) reprocessing for long lived particles searches (DRAW_RPVLL) , staged & reprocessed, January~April, 2020
 - ❑ run at real scale: finished on time, no complains from data preparation group.
 - ❑ Much less disk space occupied ($\sim 2\text{PB}$)
- ❑ Run in data carousel mode on AOD : they take 1/3 of ATLAS DISK space, move them to tape !?
 - ❑ Current goal : run derivation production in Data Carousel mode for Y2016-Y2018 datasets, produce DAOD_PHYS and DAOD_PHYSLITE for physics analysis
- ❑ Tape is becoming more and more important in the ATLAS reconstruction workflows

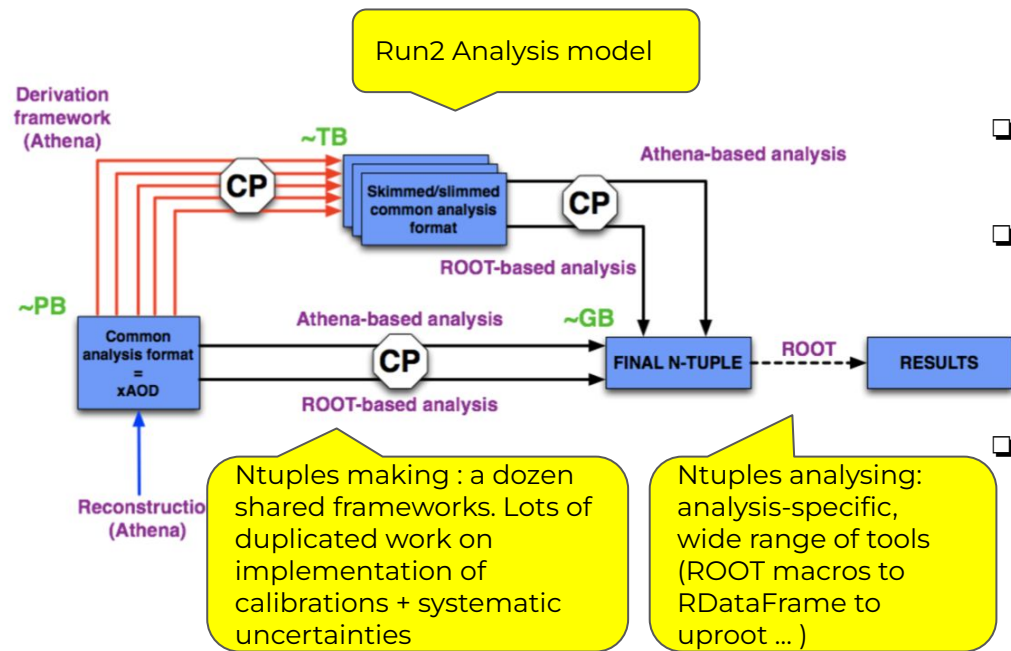
Highlights (IV) : HPC resources



1. Risorse HPC : grande attenzione di ATLAS verso la possibilita' di utilizzare queste risorse (lo sforzo su AthenaMT va anche in questa direzione)
 - 1.1. il successo dipende fortemente dalla configurazione delle macchine (open/closed), dalla possibilita' di intervenire nel processo di definizione del progetto e dell'architettura e dalla disponibilita' di persone esperte dedicate
 - 1.2. su risorse 'open' (accesso dei wn alla rete esterna) si possono girare molti workflows di ATLAS
 - 1.3. Grant su VEGA (EuroHPC in Slovenia) : copre circa 50% del pledge ATLAS, quasi tutti i workflows (anche se principalmente MC)
2. In Italia ATLAS ha partecipato positivamente al grant CINECA 2020
 - 2.1. non chiaro se riusciremo a partecipare al nuovo grant Marconi 100 (non esiste un build del software ATLAS per Power9 serve manpower per costruirlo)

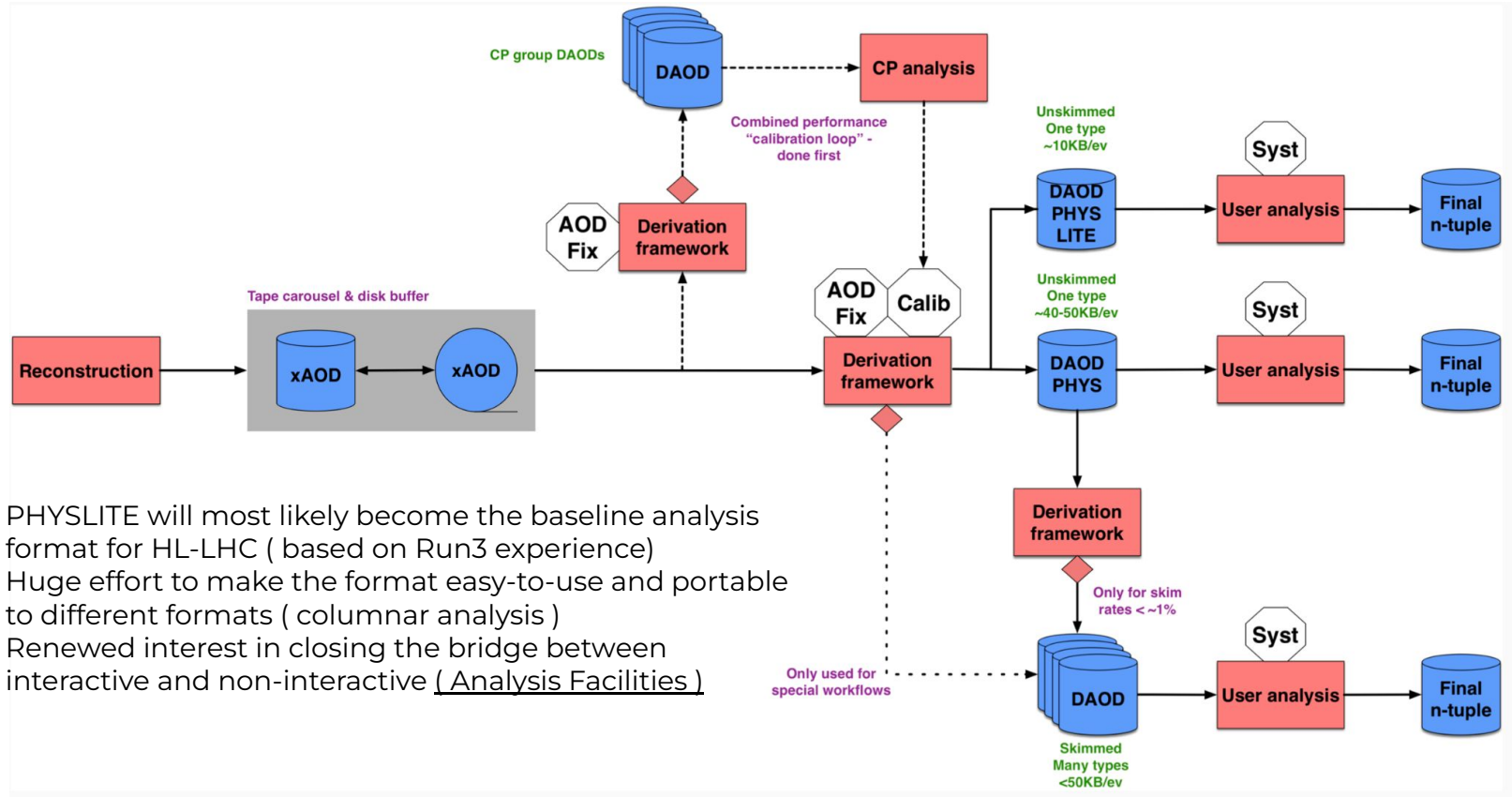
A new analysis model

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.

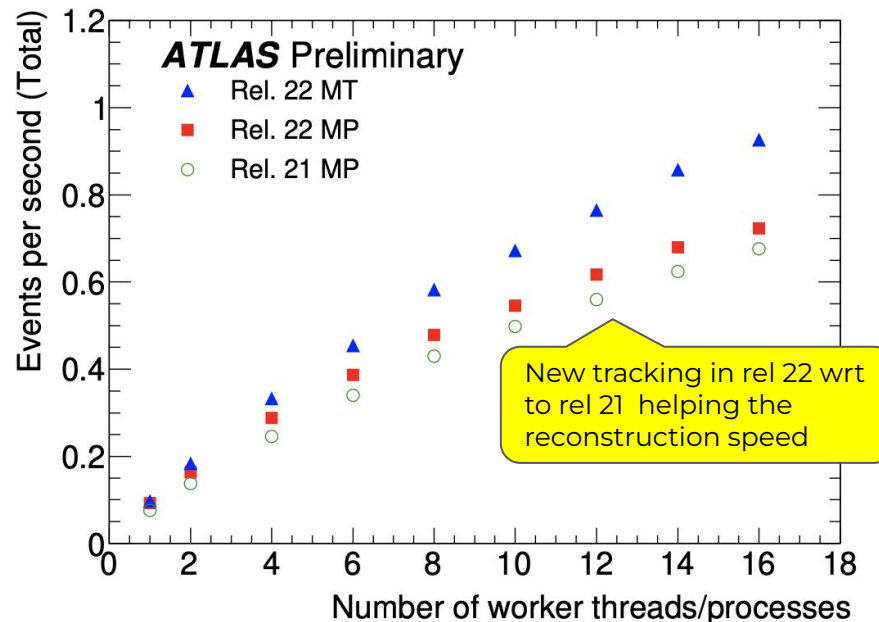
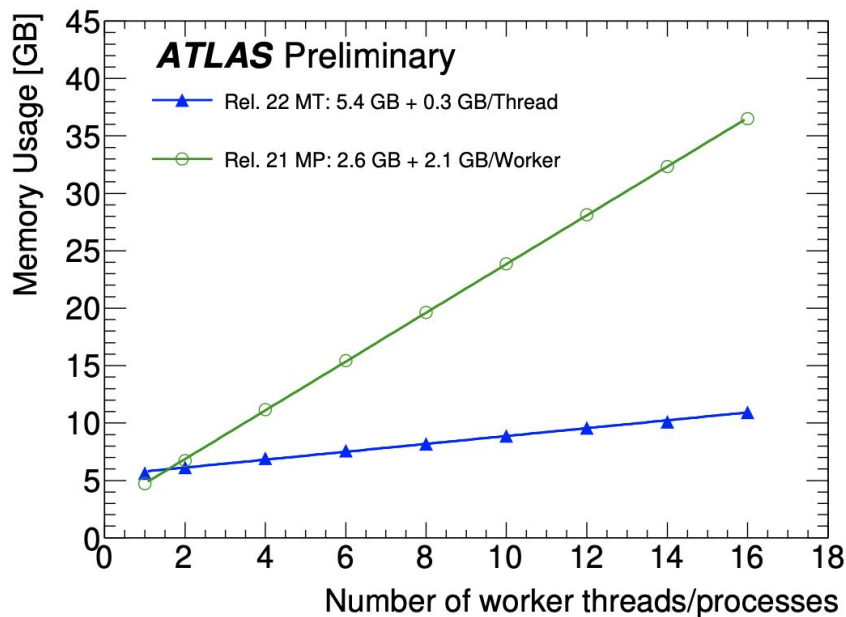


- ❑ The output of the data and MC reconstruction is stored in Analysis Object Data (AOD) files and grouped in datasets on the various Grid sites.
- ❑ These datasets are processed in the derivation framework which produces about 80 different derived AOD (DAOD) formats that contain a subset of events and 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



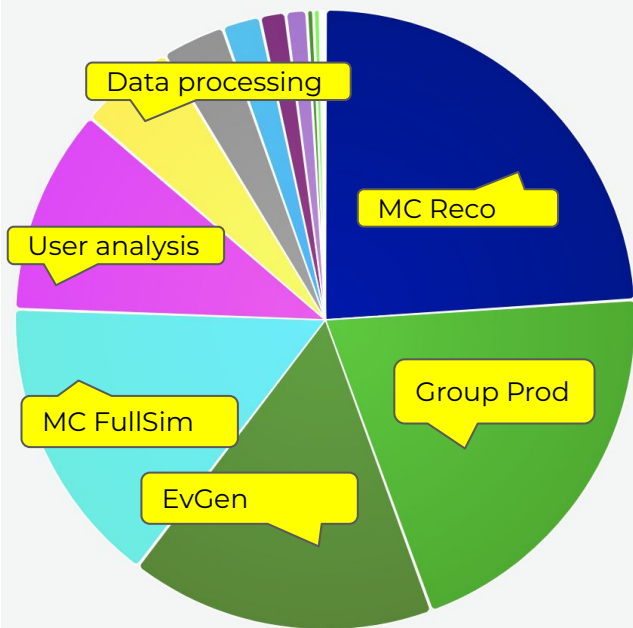
- ❑ The benchmark jobs use real data from run 357750 taken during 2018, with 250 events per worker process or thread.
- ❑ The data events have an average number of interactions per bunch crossing (μ) = 50, which is approximately that expected for the luminosity-leveling period during Run 3.
- ❑ Tests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



ATLAS Computing per type of resources

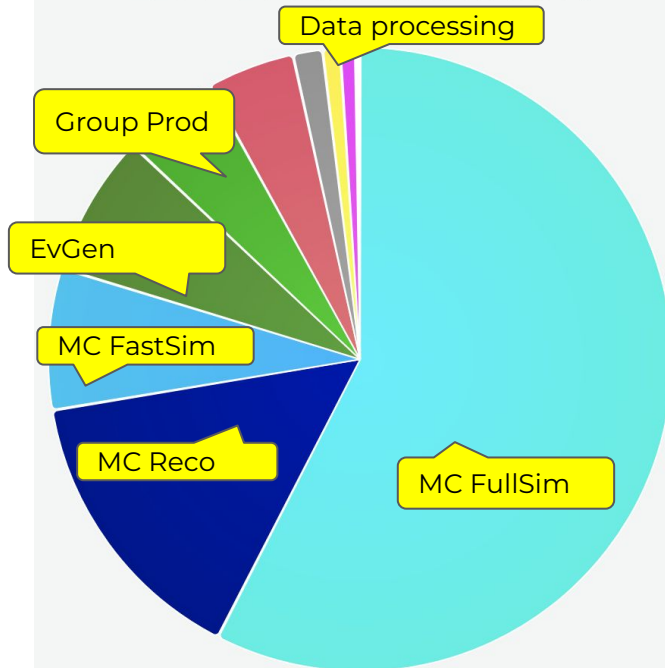
GRID RESOURCES

Wall clock time. Successful jobs (HS06 seconds) ▾



VEGA(HPC)

Wall clock time. Successful jobs (HS06 seconds) ▾



- ❑ 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.

Highlights (V): new analysis model

1. Attualmente l'analysis model prevede ~80 diversi streams (DAOD) basati su skimming-menu diversi e con diverso contenuto a seconda dell'analisi
2. Analysis Model completamente diverso per run3 (e preparazione run 4)
 - 2.1. Solo un tipo di output (DAOD_PHYS) unskimmed con tutte le informazioni principali per il ricalcolo delle correzioni e sistematiche
 - 2.2. DAOD_PHYSLITE, pre-calibrati, possono essere riprodotti a partire da DAOD_PHYS
 - 2.3. open the discussion for analysis facilities

	MC size/event (KB/event)	Data size/event (KB/event)	Nominal MC events/year in run 3 (billion) from CDR*	Nominal data events/year in run 3 (billion) from CDR*	Total size/year/ version (PB)
PHYS	40	30	25	10	1.3
PHYSLITE	15	10			0.475

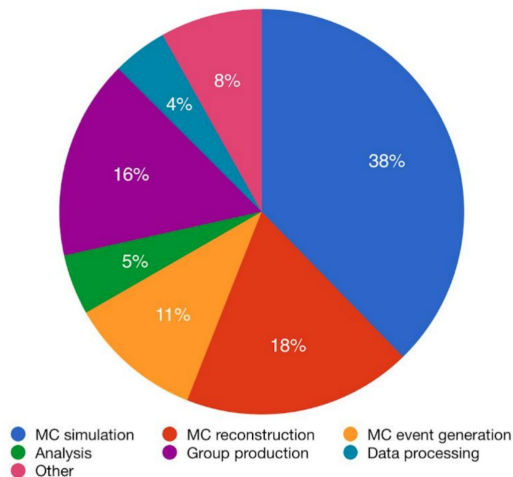
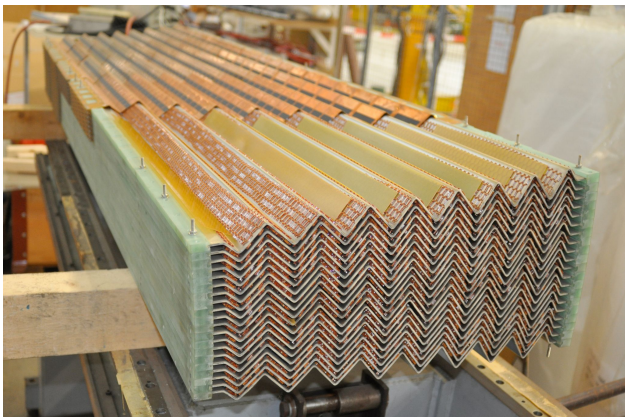
Default in Run4

Stima : assumendo 4 copie complete di DAOD_PHYS si passa da 60 PB (modello attuale) a ~ 32 PB (nuovo analysis model)

3. Produzione di DAOD_PHYS/PHYSLITE attraverso tape carousel !

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)

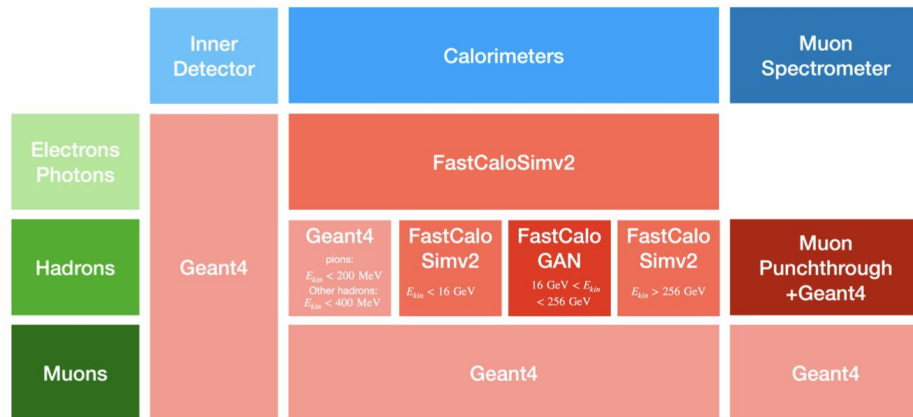
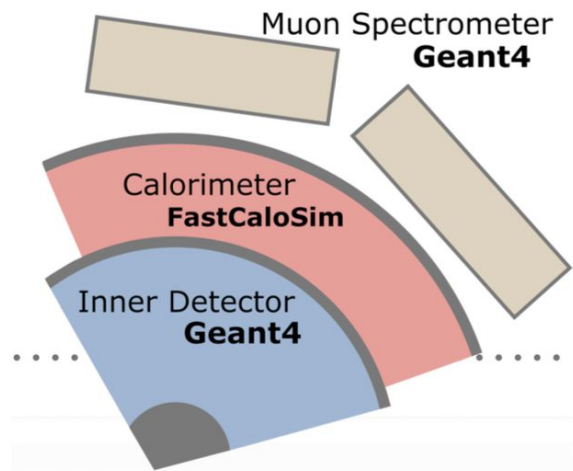


To reduce the impact of the preparation of MC events :

- ❑ Optimise G4 full simulation : expect up to 30% reduction
- ❑ 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

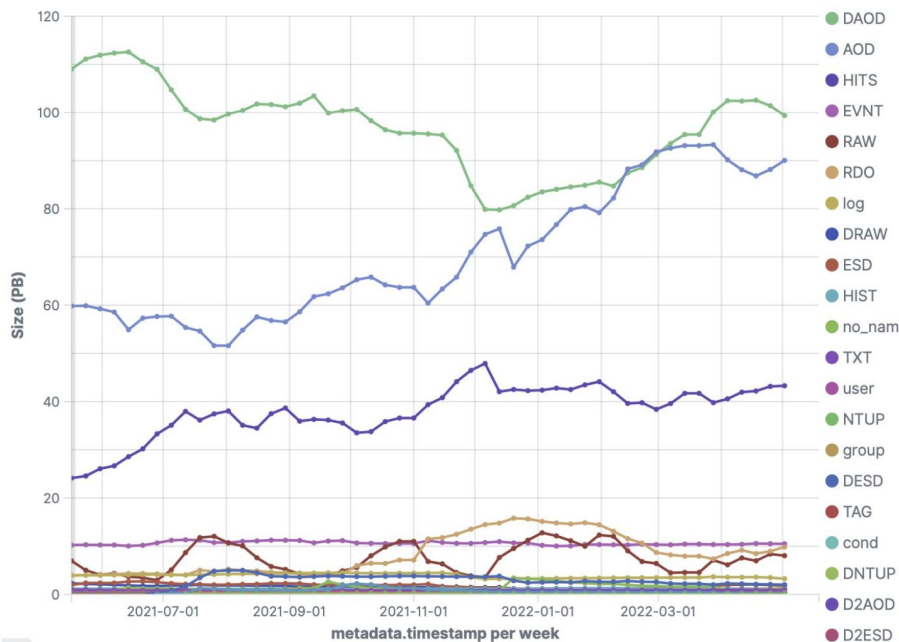


- ❑ Atlfast-III (AF3) is the successor of the Atlfast-II (AF2) simulator
- ❑ Full simulation for tracking (ID + muons) and parameterized simulation of the calorimeter
- ❑ AF3 implements two distinct approaches of shower generation:
 - ❑ FastCaloSimV2: parameterized modelling (separate parameterisations 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

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

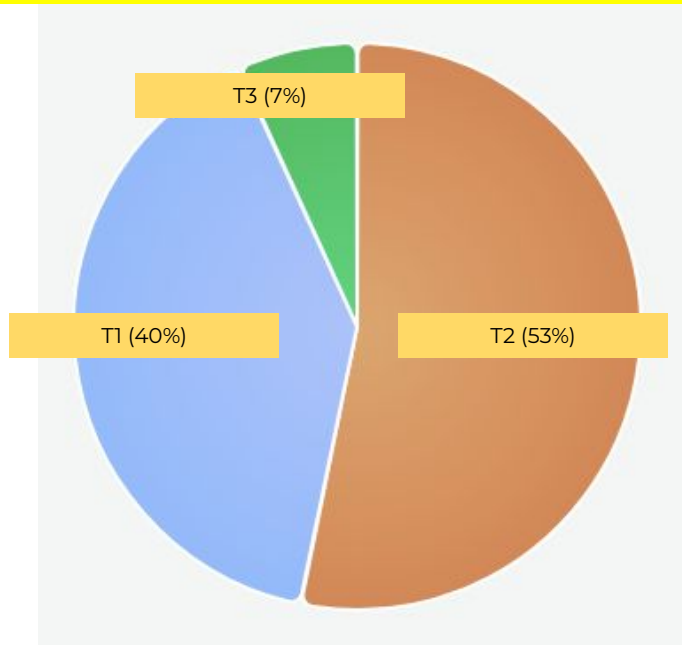


- ❑ The size/event for the AOD is about 600 kB
- ❑ Around 80 DAOD formats, size in the range of 40-450 kB depending on the type of the physics selection and the information retained.
- ❑ only 1-2 replicas of each dataset and campaign can be kept on disk.
- ❑ AODs and DAODs which are the two formats taking 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

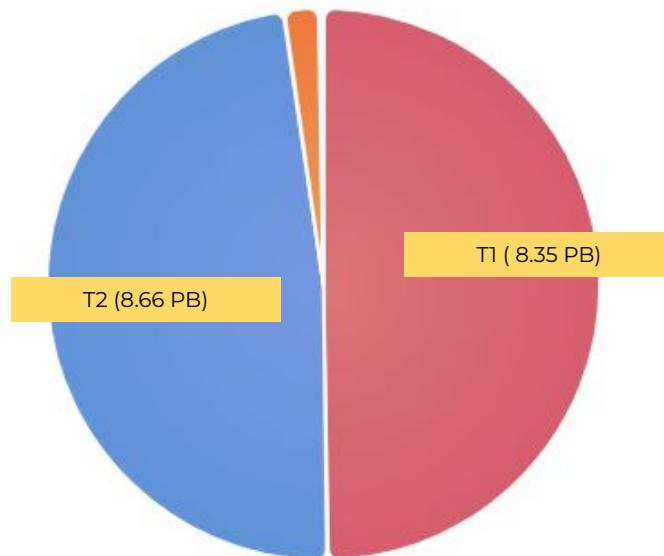
Distribuzione carico lavoro ATLAS in Italia

Lavoro distribuito sul T1 (CNAF), i quattro T2 (Frascati, Milano, Napoli e Roma1) e quattro T3 (Cosenza, Genova, Lecce, Roma3)

CPU

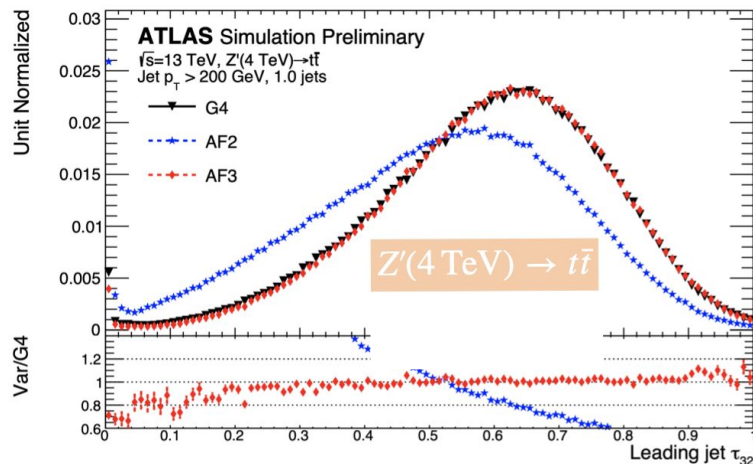
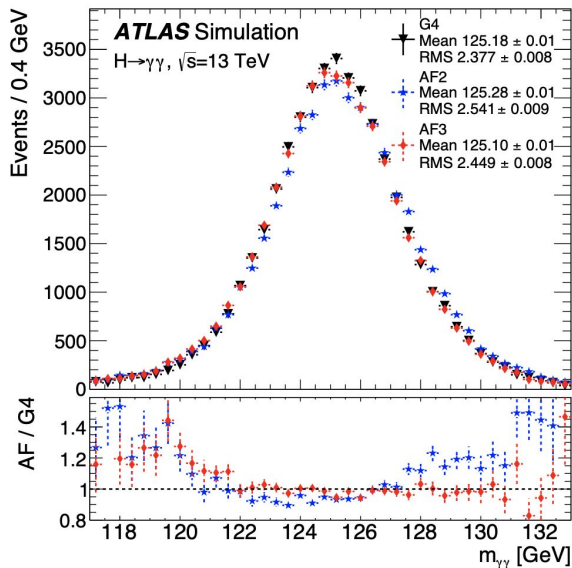


STORAGE



Fast simulation : Atfast-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

Richieste dettagliate T2 : CPU

CPU	Frascati	Milano	Napoli	Roma1	Totale	Pledged
Finanziato 2022 (HS06)	36812	35781	34097	35204	141894	142920
Aggiornato 2022 (HS06) (*)	35312	36286	38886	30511	140995	
Obsolescenza 2023 (**)	16866 (8433/8433)	8004 (0/8003)	0	8004 (0/8003)	32874	
Obsolescenza 2022 non finanziata	0	422	0	6517		
Delta 2023	3000	3000	3000	3000	12000	
Totale richieste 2023 (HS06)	19866	11426	3000	17521	51813	
Totale CPU nel 2023 (HS06)	38312	39708	41886	40028	159934	157230

(*) numeri aggiornati dopo inserimento nuove risorse da gare in corso

(**) in parentesi (obsolescenza primo semestre 2023 / obsolescenza secondo semestre 2023)

Richieste dettagliate T2 : disco

Nella configurazione di quest'anno stiamo un po' rompendo la simmetria dei siti

Disco	Frascati	Milano	Napoli	Roma1	Totale	Pledged
Finanziato 2022 (TB)	2839	2432	2355	2432	10058	9940
Aggiornato 2022 (TB) (*)	3328	2489	4327	2624	12768	
Totale disco 2022 aggiornato (TB) senza LOCALGROUPDISK (**)	3148	2309	4147	2444	12048	
Obsolescenza 2023	0	0	0	0	0	
Delta 2023 (TB)	0	0	0	0	0	
Totale richiesta 2023 (TB)	0	0	0	0	0	
Totale disco 2023 (TB)	3148	2309	4147	2444	12048	11760

- ❑ (*) numeri aggiornati dopo inserimento nuove risorse da gare in corso
- ❑ (**) come al solito 180 TB a sito per LOCALGROUPDISK
- ❑ Napoli ha anticipato 2PB di disco IBISCO per crisi disco ATLAS