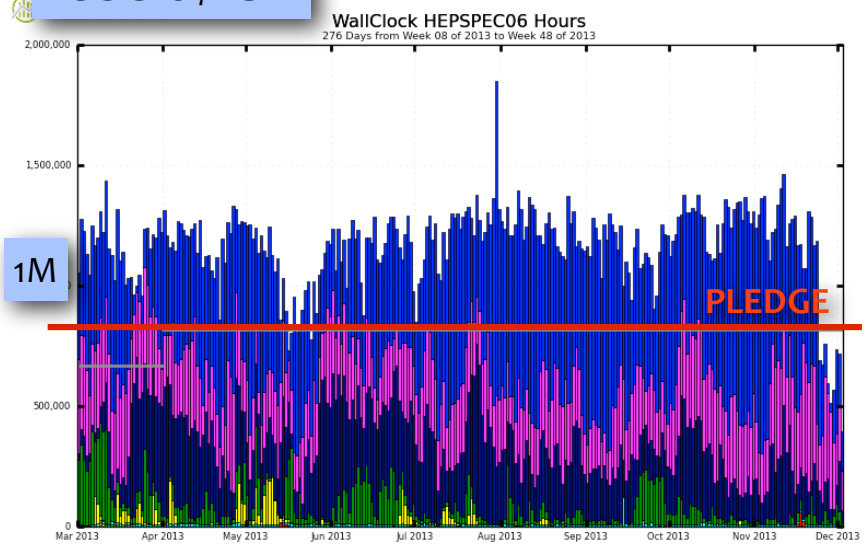
ATLAS Computing Status

Borut Kersevan
Richard Mount

LHCC Referees December 2013

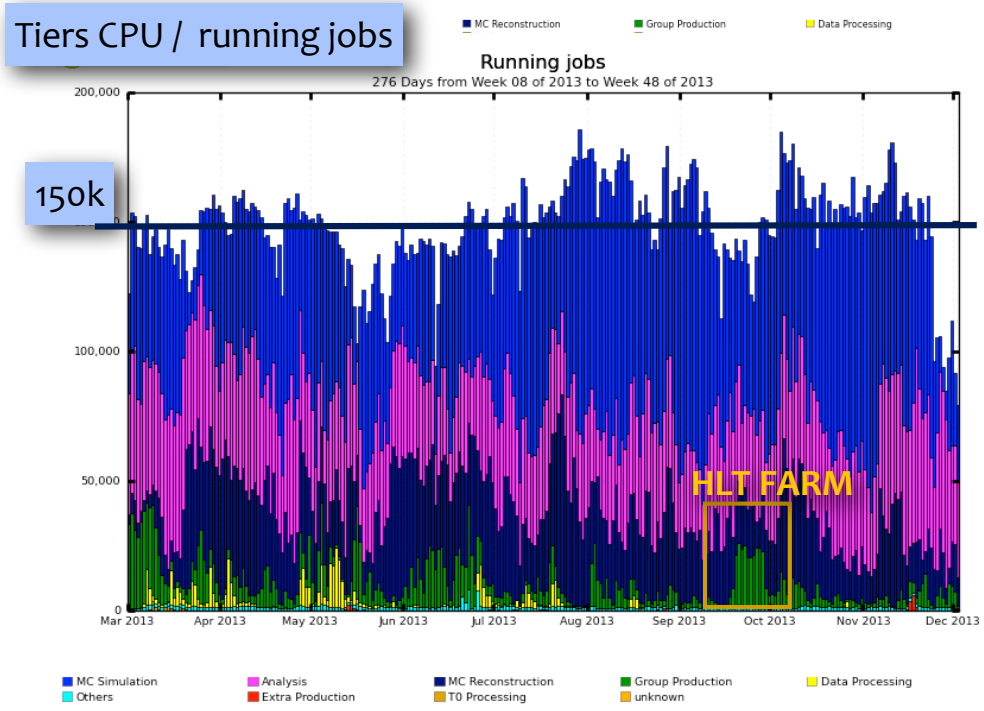
ATLAS Resource Utilization in 2013

Tiers CPU / HS06

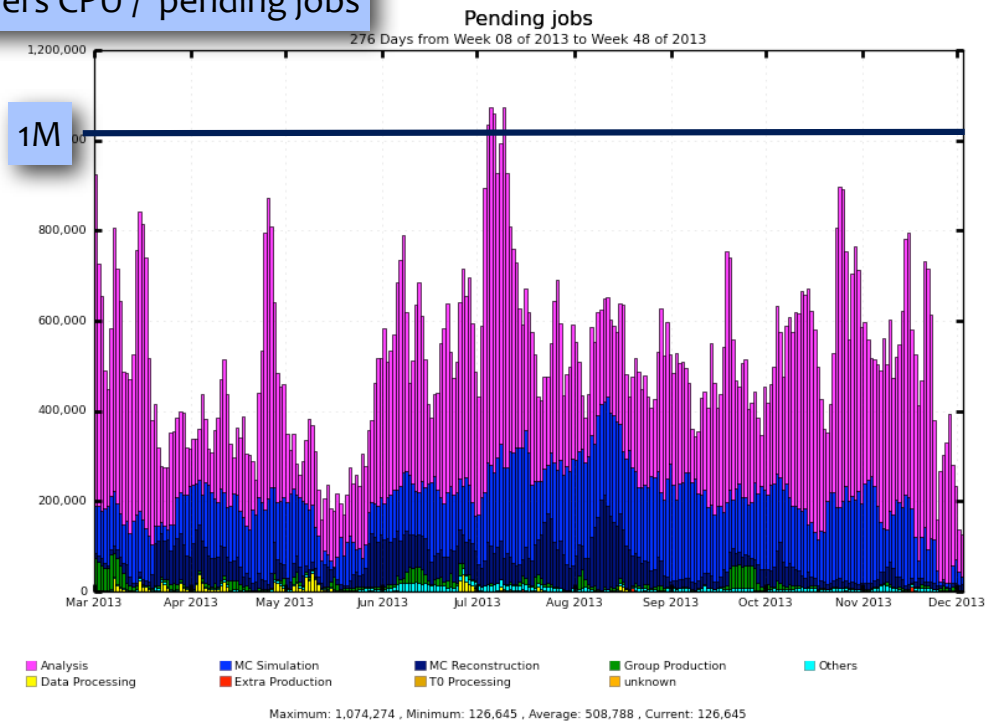


- ATLAS has utilized the computing resources in its Tiers well in the last year: *many thanks to sites for resources and excellent operating!*
  - **We manage to provide a timely throughput of analyses to meet the physics requirements.**
- An ongoing effort in software development to optimise the resource utilization by reducing the CPU consumption, event sizes - for Run-2...

Tiers CPU / running jobs



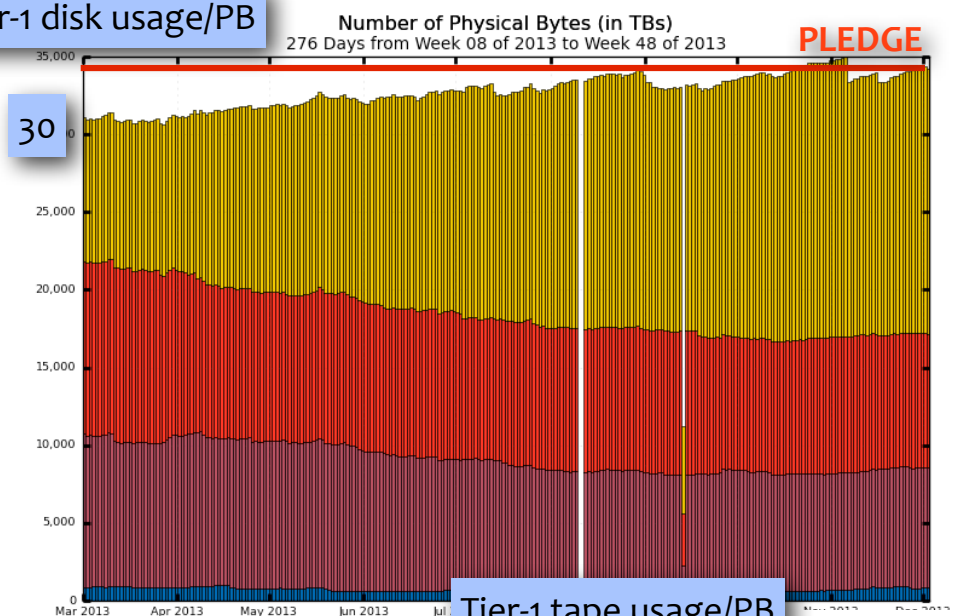
Tiers CPU / pending jobs



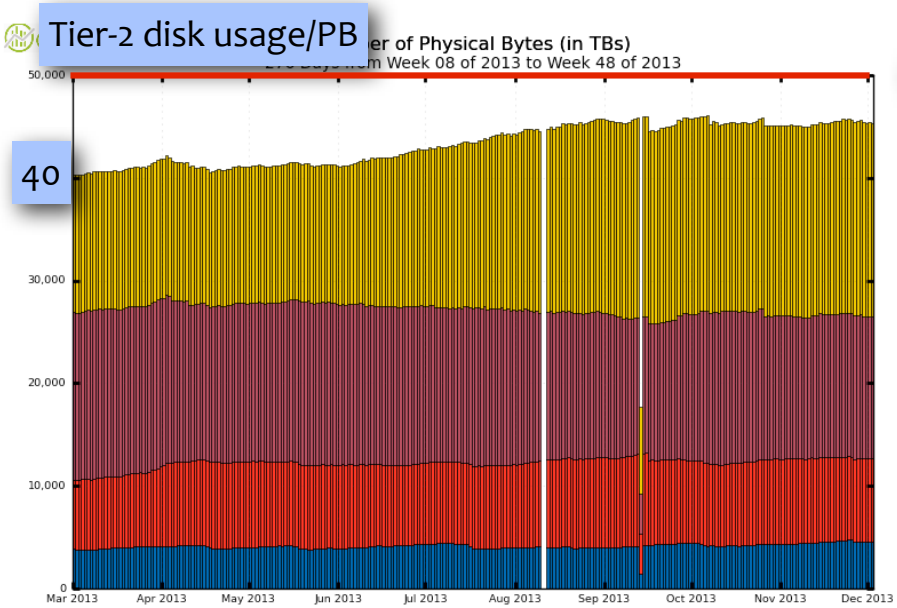
ATLAS Disk and Tape Usage in 2013

- A constant 'struggle' with full T1 disk - 'lifetime' of 2011 and 2012 analyses under-estimated in our resource models.
- Complex (= work intensive) handling of cleanup actions is being automated.

Tier-1 disk usage/PB

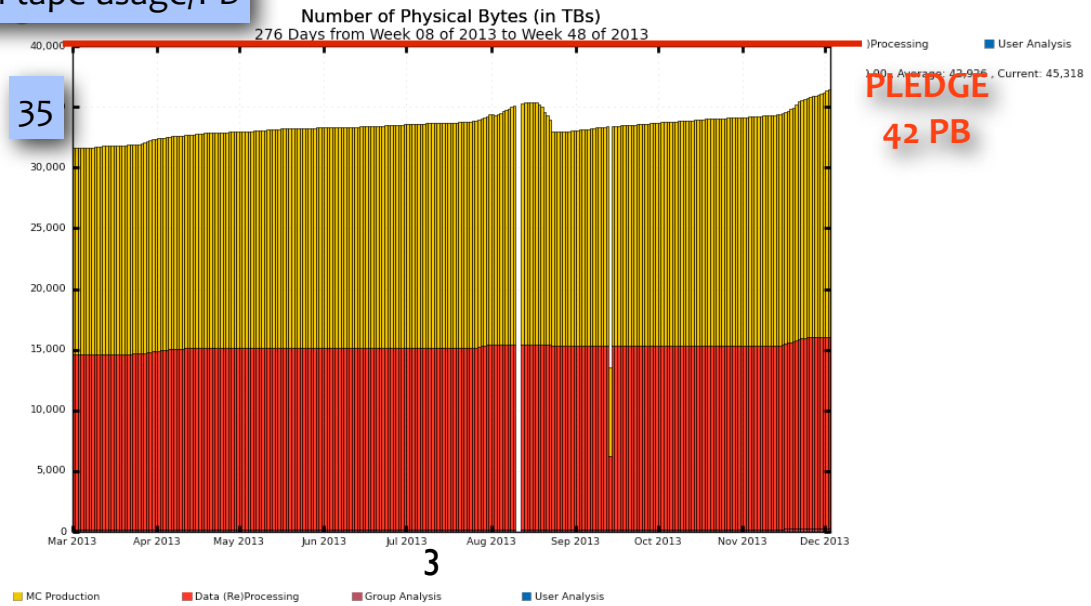


Tier-2 disk usage/PB



PLEDGE
51 PB

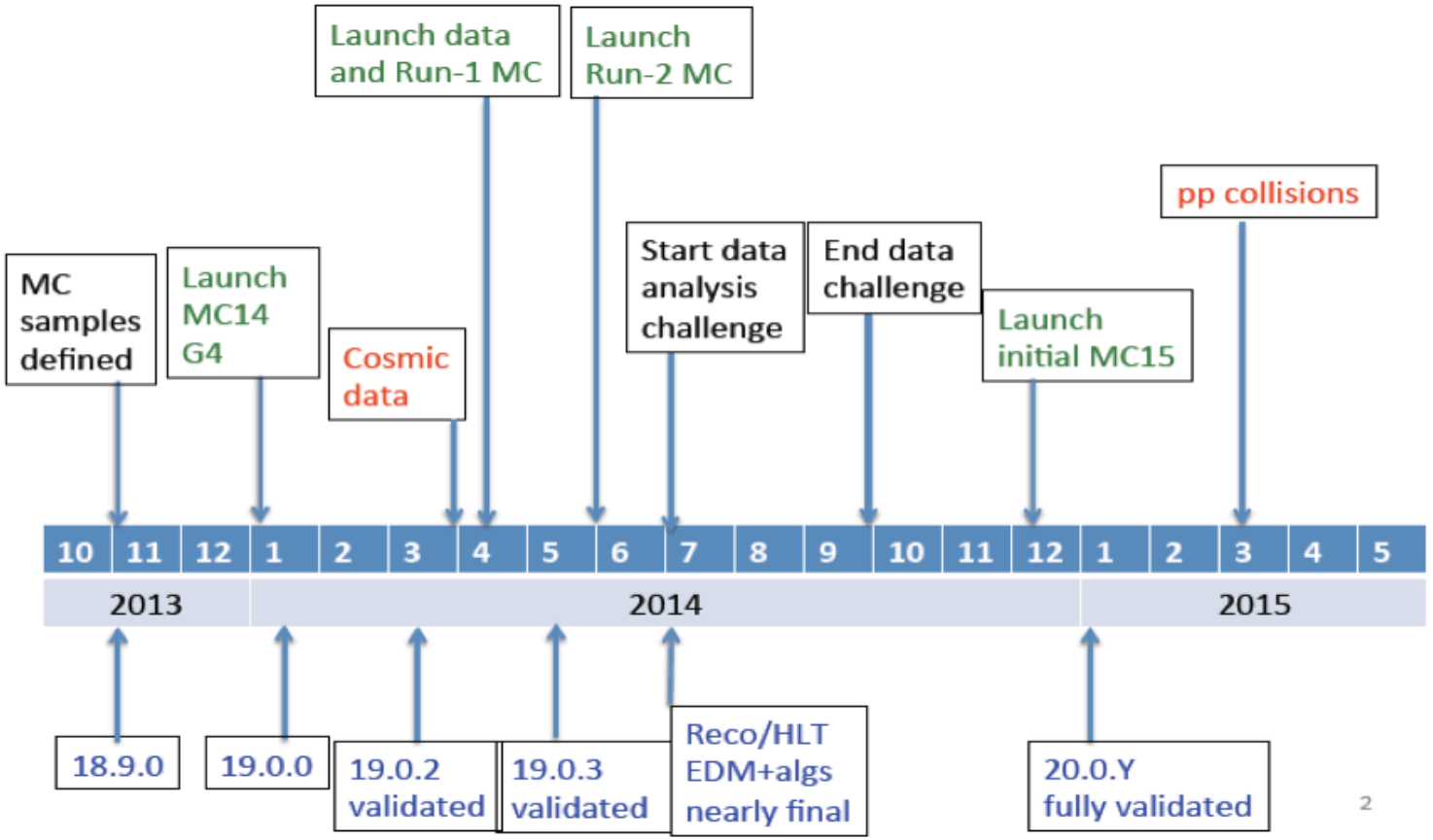
Tier-1 tape usage/PB



PLEDGE
42 PB

Data Challenge(s) 2014

- A detailed set of Data Challenge milestones in 2014/2015 defined by the ATLAS Offline Coordination Board to exercise commissioning and readiness of all the Run-2 improvements for data analysis.



Software Improvements for Run-2

ATLAS has set strict benchmark targets for the CPU/event consumption and AOD event sizes of data reconstruction, Monte Carlo production chain and analysis, and is currently investing a lot of manpower to achieve this.

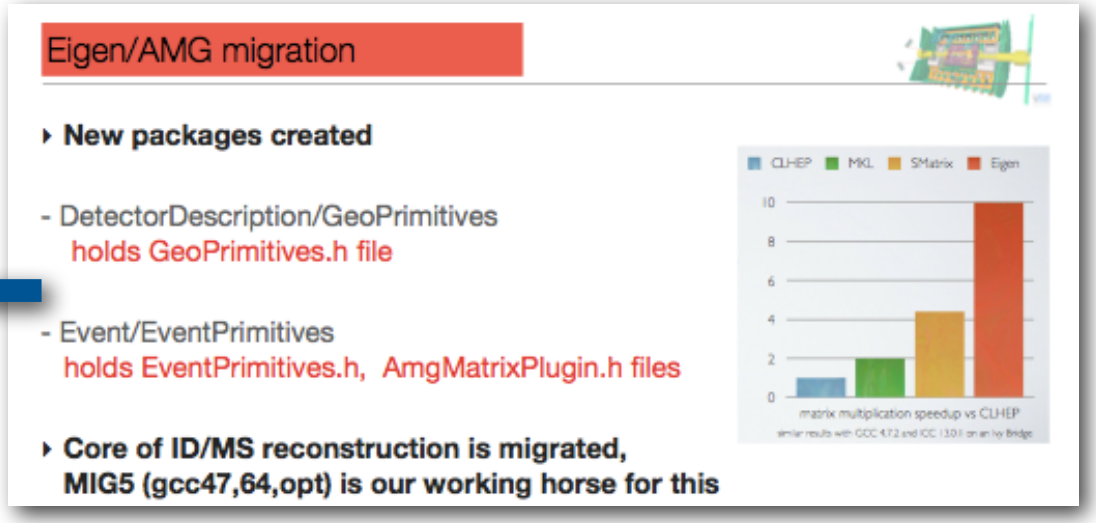
A second Software Coordinator appointed to manage the load.

- **We rely heavily on software improvements to keep within a reasonable resource budget.**
  - The processing times (CPU/event) will in reconstruction need to improve by at least a factor of 2 compared to direct use of current Run 1 software in Run 2 conditions (**1 kHz data recording rate, 14 TeV, 25 ns bunch spacing, average pile-up of 25 to 40**).
  - The event sizes will need to be optimized to be kept (approximately) at Run-1 values.
  - Work is needed on memory consumption of the reconstruction chain (AthenaMP).
  - The development of fast simulation flavors within Integrated Simulation Framework as well as fast digitization and fast reconstruction is of high importance.
    - We will have to rely more on fast simulation than in Run 1 to achieve reasonable Monte Carlo statistics for analysis.

Work ongoing (manpower struggle), deliverables will be ready for DC14. At present not possible to do a relevant resource gain evaluation.

Example: Tracking 'migration'/improvements

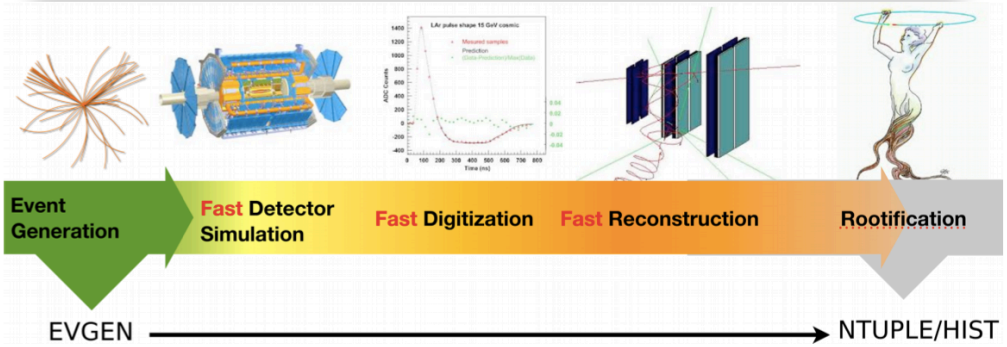
	stage	migration status	ID CPU time	comments
	space point formation	done	-1%	fully validated
	space point seeded track finding	done	-20%	slightly increased track finding efficiency
	ambiguity solving & track fitting	done	N/A	biggest relative CPU saving potential, debugging phase
	vertex reconstruction	ongoing	N/A	bound together with **xAOD** via xAODParticle design, prototype in preparation
	particle creation	ongoing	M/A	

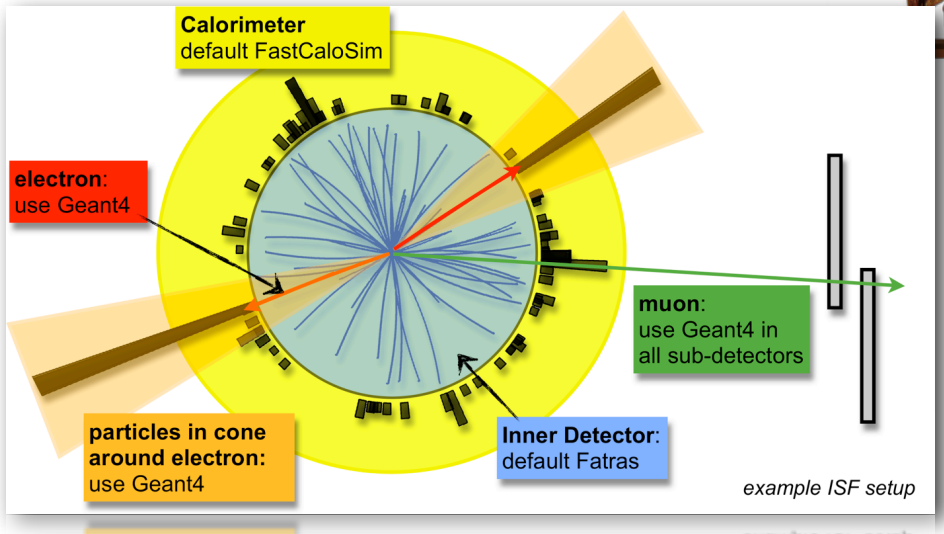


A special Task Force dealing with Eigen/ xAOD software migration set up.

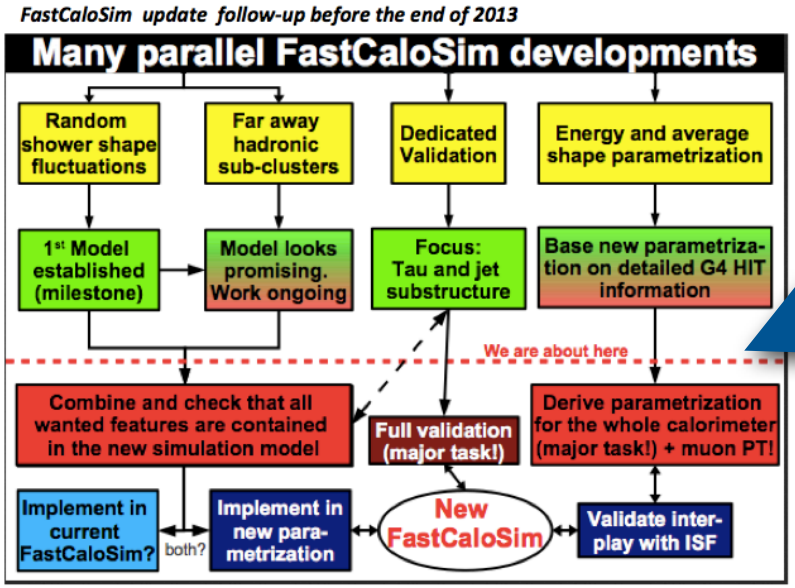
ATLAS Simulation - ISF

ATLAS is developing several flavours of fast simulation within the new Integrated Simulation Framework (ISF), as well as very-fast digitization and reconstruction truth-seeded procedures for Monte-Carlo data. The aim is to achieve an optimal balance between required precision for analyses and resource requirements.





On track to be deployed between start of 2014 to 2015 depending on the ISF flavor.



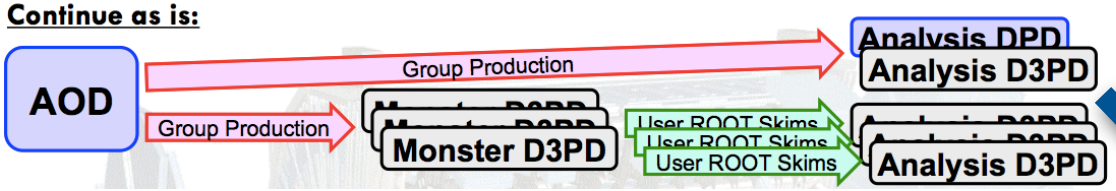
ISF simulation setup	Speedup	Accuracy
Full Geant4	1	best possible
Geant4 with FastCaloSim	~25	approximated calorimeter
Fatras with FastCaloSim	~750	all subdetectors approximated
Fatras with FastCaloSim simulate only particles in cones around photons	~3000	all subdetectors approximated event simulated only partially

gg→H→γγ no pileup

Analysis Changes

ATLAS has reworked its analysis model in order to optimize and reduce the resource requirements. The model involves innovations impacting resource requirements as follows:

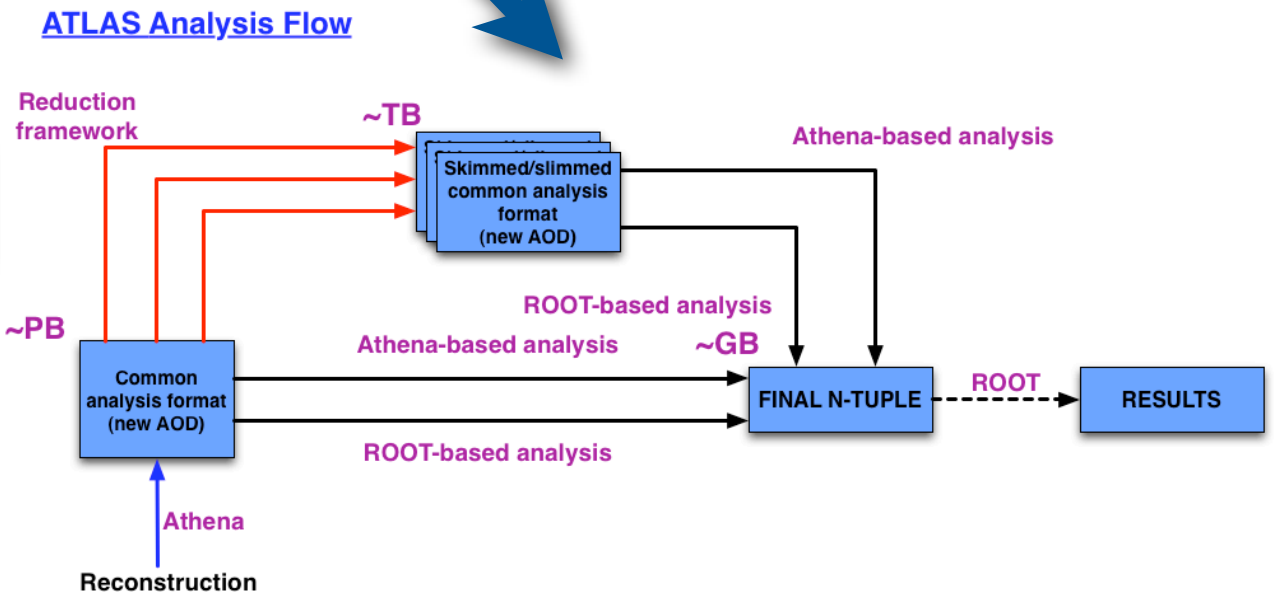
- an update to the AOD format, to make the analysis objects ROOT-readable, thus eliminating the need for 'intermediate' big ROOT-readable group analysis ROOT N-tuples used in Run-1.
- Introducing AOD2AOD re-processings to reduce the CPU consumption of group analysis, which currently introduces reconstruction improvements at the group analysis level.
- Introduction of a centralized 'Reduction Framework' to derive the group analysis specialized N-tuples with reduced effort and processing, thus reducing the CPU consumption.



New Analysis Model

Disk saving of 'group' formats by a factor of 3-4
CPU savings of 'group' analysis by a factor 2-3

Work organized in three Task Forces, deliverables will be ready for DC14

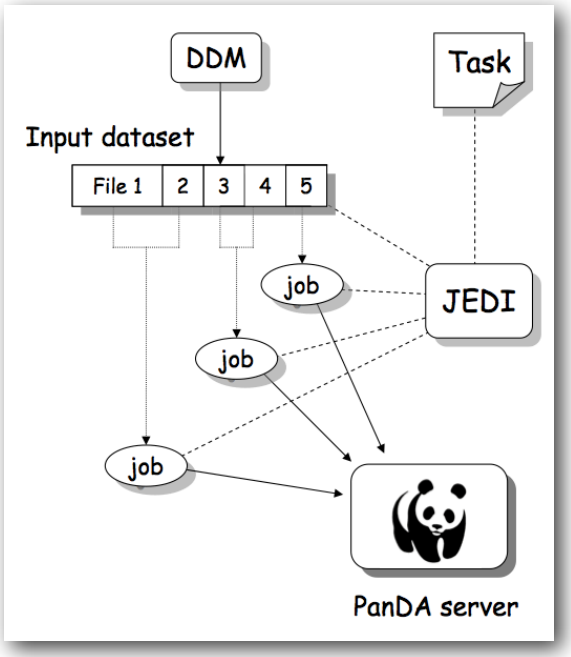


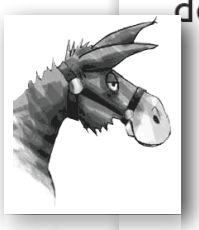
Rucio and Prodsys-2

- **Distributed computing Software Improvements:**
  - **The new data management system (Rucio)** will introduce more granular handling of ATLAS data (file replicas rather than dataset replicas) together with a more intelligent implementation of data ownership for users and groups. This will allow optimization of the space usage while offering more functionalities in terms of local and global quotas, thus e.g. **relaxing the need for dedicated group storage (SRM space tokens) at sites.**
  - **The new Prodsys-2 (PanDA+Jedi+DEfT)** will simplify handling of different workflows (e.g. production vs analysis jobs and pilots).
  - The combination of Rucio and Prodsys-2, will also provide the functionality for the **use of federated resources and remote access**, i.e. xRootD, HTTP federations...
  - **"Event Service"**, combining functionalities of Panda/Prodsys2, the xROOTd federation and AthenaMP, is under development for Run 2.

Rucio and Prodsys-2 will be ready for DC14 (early next year)

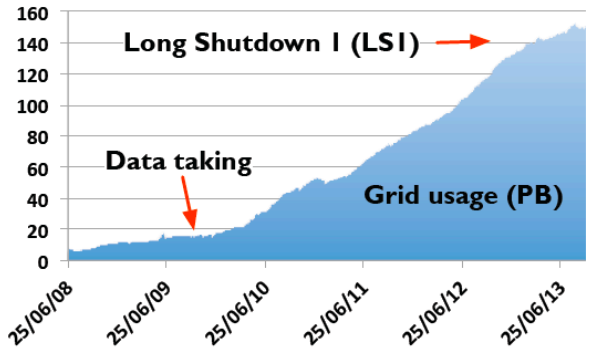
Federations (XRootD) and Event service (pilot use case) Will be ready in 2015





The current DDM system Don Quijote 2 (DQ2) has demonstrated very large scale data management

- 150 PB
- 130 grid sites
- 800 users
- +40 PB per year
- +1 M files per year
- 0.6 M downloaded files per day



DQ2 will simply not continue to scale for LHC Run-2

http://rucio.cern.ch

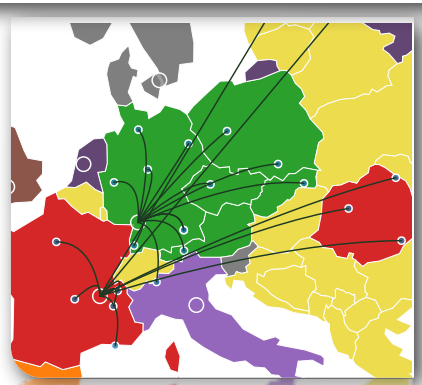
Data Placement Policies

ATLAS has reduced the pre-placement policy of AOD replicas to **one replica on Tier-1 disks and one replica on Tier-2 disks** (At the end of Run-1 ATLAS used two replicas on Tier-1 disks and two replicas on Tier-2 disks). **ATLAS will consequently rely even more on the existing popularity-based dynamic data replication (PD2P) and deletion (Victor).**

Implementing now, used in in 2014 and onwards.

2010

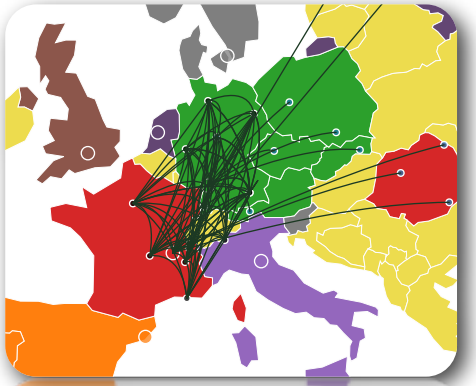
Planned data distribution
Jobs go to data
Multi-hop data flows
Poor T2 networking across regions



~20 AOD copies distributed worldwide

2013

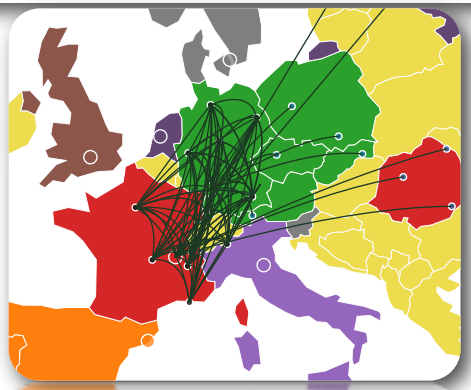
Planned *& dynamic* distribution data
Jobs go to data *& data to free sites*
Direct data flows *for most of T2s*
Many T2s connected to 10Gb/s link



4 AOD copies distributed worldwide

2015

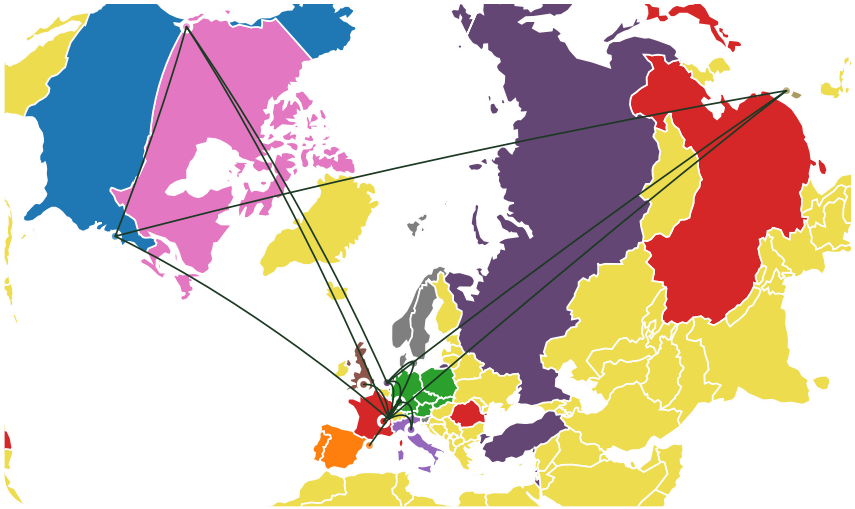
Planned *& dynamic & federated* distribution data
Jobs go to data *& data to free sites & are read remotely*
Direct data flows *for most of T2s.*



< 2 AOD copies distributed worldwide

Roles of Tiers

- **In Run 1 the border between Tier-1s and Tier-2s in terms of performing different job types became 'soft':**
  - This will continue in Run 2, i.e. workflows will be optimally shared between Tier-1s and Tier-2s to optimize disk and CPU availability, network traffic…
- **With the technological progress of WAN ATLAS has already in Run 1 gradually moved away from the hierarchical association of Tier-2s to one Tier-1 (MONARC model) and is now able to associate workflows between a well-connected Tier-2 and several Tier-1s, based on monitored connectivity metrics.**
  - This functionality will be further pursued in Run 2 and is being incorporated into the production system and distributed data management developments.
- **Tier-1s continue to keep the essential role as archival facilities and are being relied upon for the most difficult workflows (I/O,CPU,memory..)**
  - A new feature in Run 2 is the envisaged scenario of **'overspilling' the prompt reconstruction from Tier-0 in case of CPU congestions to Tier-1s.**
    - In order to ensure the readiness of Tier-1s for the prompt data processing a small stream of data will be constantly processed at Tier-1s and validated by comparing the outputs with the Tier-0 processing outputs.



Use of Opportunistic Resources

- In Run 2, the work will continue on utilizing the **opportunistic 'Cloud'/IAAS resources or High Performance Computing (HPC) centres** and the creation of a transparent and straightforward interface between these and the ATLAS distributed computing environment.
  - to complement our existing resources to **off-load CPU intensive and low I/O workloads which under-use our I/O optimized grid resources.** A typical example would be 'Monte Carlo generation of hard-process events' (Alpgen, Sherpa Monte Carlo event generators), which took ~15% of CPU resources on the ATLAS Grid sites in 2012. The Grid resources can then be better used for I/O intensive work: (simulation, reconstruction, group production, analysis), important especially during peak demand.
  - In addition, if ATLAS Grid sites **want to move to use cloud middleware** (e.g. OpenStack) or **provide dual-use resources** ('batch'+HPC) the ATLAS versatility gained by the ongoing R&D provides this option.
    - **We use this approach at our HLT farm during LS1.**

High performance Computing (HPC):

ATLAS started evaluating opportunistic usage of free HPC resources (both x86 and non-x86 architecture).

Examples of recent usages:

- Usage of supercomputers as virtual Tier2 for production of evgen and simulation
- Generated 81M of Alpgen W/Z+jets 4-vectors
- 1M CPU hours at multiple sites to set Zprime limits.

Early results suggest real potential, ATLAS has established a working group to evaluate these opportunities.

Running Jobs | Queued Jobs | Reservations

Total Running Jobs: 10

Job Id	Project	Run Time	Walltime	Location	Queue	Nodes	Mode
653066	CACS_In	15:25:12	1d 00:00:00	ANL-R10-R47-32768	R.cacs	32768	script
653009	DirectNoise	08:58:09	12:00:00	ANL-R04-R07-4096	prod-long	4096	script
653208	Peta_CESAR_2013	06:55:37	12:00:00	ANL-R02-M1-512	prod-long	512	vn
653232	ATLASP	03:28:57	06:00:00	ANL-R00-M1-512	prod-short	512	smp
[illegible]	ATLASP	03:07:30	06:00:00	ANL-R01-M1-512	prod-short	512	smp
653234	ATLASP	03:07:04	06:00:00	ANL-R01-M0-512	prod-short	512	smp
653235	ATLASP	00:45:49	06:00:00	ANL-R02-M0-512	prod-short	512	smp
653236	ATLASP	00:25:30	06:00:00	ANL-R00-M0-512	prod-short	512	smp
653237	ATLASP	00:19:09	06:00:00	ANL-R03-M1-512	prod-short	512	smp
653238	ATLASP	00:18:42	06:00:00	ANL-R03-M0-512	prod-short	512	smp

Summary and Outlook

- A lot of experience acquired in 3 years of LHC data taking.
- Run-2 will put high pressure on hardware and human resources.
- Solutions under development and manpower is needed.
- New computing model and its components will be tested during 2014 data challenge.
- LHC & ATLAS upgrades also mean resources for software & computing (but we need to be careful how to prioritize the work).