Software & Computing status

- Resource usage in 2017
- Requests for 2018
- Finalizing Run-2
- Preparing Run-3
- Towards HL-LHC

Institutional Committments

- Inputs
  - **Class 3**: Management, Database, Distributed Computing , Software Project, Squad support
  - **Class 4**:Tier-1 & Tier-2 sites operations tasks
- In ATLAS

  Clear increase in required S&C Class 3 manpower
  - 2016: 147.25 FTE, 2017: 157.19 FTE
  - Manpower allocated
    - 2016 vs 2017: -5.9 FTE (4%)
  - Lack of allocated vs required
    - 2016: -9.1 FTE, 2017: -23.4 FTE

French S&C ICs

Class 4	Institute	#FTE	Fraction of ATLAS members
	LPSC	0.85	S.Crépé (0.1)
	CPPM	0.60	E.Knoops (0.4)
	LPC	0.75	
	LAPP	1.25	S.Jézéquel (0.1), F.Chollet (0.1)
	CC-IN2P3	3.30	E. Vamvakopoulos (0.8)
	LAL	0.30	
	LPNHE	0.90	F.Derue (0.1)
	IRFU	1.15	J-P.Meyer (0.15)

Class 3	Institute	#FTE	Activity	Persons
	LAL	1.2	Event Index	J. Hrivnac, J. Yuan
	LAL	1.0	Condition DB	G. Rybkin
	IRFU	0.65	Condition DB	A. Formica
In Data Preparation	LPSC	2.7	AMI	J. Odier, F. Lambert, J. Fulaquier

ATLAS pledges 2017

- 2017 (50% more data, 20% more CPU & Disk)

All ATLAS	Increase (2016->October requests)	Net increase for France	Fraction FR/all-ATLAS
T1 CPU KH	77%	36%	9.5%
T1 Disk pB	45%	13%	10.4%
T1 tape pB	62%	44%	9.5%
T2 CPU kH	99%	43%	7.4%
T2 Disk pB	15%	7%	9.1%

- Outcome (all ATLAS)
  - Shortage: CPU -14% Disk -2% Tape -5%. Thx to FAs!
  - #MC evts reduction, w/ 1kHz HLT & processing all
- In France (all LCG)
  - Extra 200k€ from IN2P3 & 100k€ from IRFU
- Ongoing optimization
  - Train production from tape,
  - AOD size reduction (~30%)
  - Workflow improvements

ATLAS pledges 2018

CERN-RRB-2017-057

Resource	Site	2017 Pledge	2018 ATLAS	Growth	2018 CRSG	Growth
CPU (kHS06)	T0+CAF	404	411	2%	411	2%
	T1	808	949	17%	949	17%
	T2	982	1160	18%	1160	18%
Disk (PB)	T0+CAF	25	26	4%	26	4%
	T1	69	72	4%	72	4%
	T2	78	88	13%	88	13%
Tape (PB)	T0+CAF	77	94	22%	94	22%
	T1	174	195	12%	195	12%

- ATLAS requests are within the expected flat budget increase and below the average 2013-2017 increase.
- CRSG recommends the requests.
- Beyond pledge resources is about 30% of the pledges, ATLAS expect to continue to receive a sizeable amount of over pledge CPU, which remain a risk for the experiment.

- Increase wrt 2017@T1 & T2s: in range 12-18%
- Except Disk@T1, T0: 4%

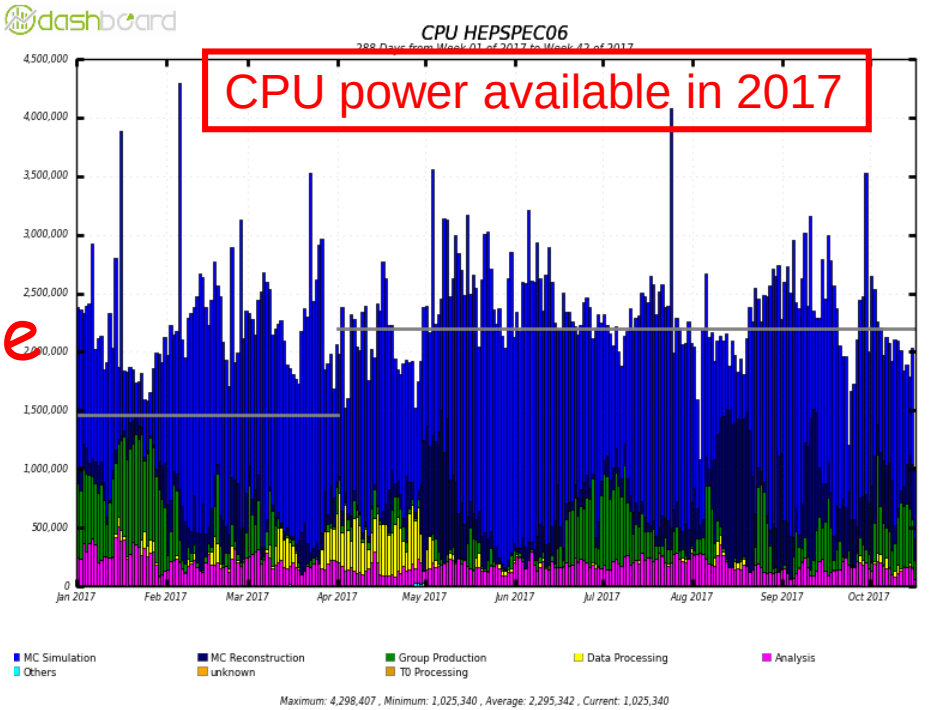
ATLAS pledges 2018 for France

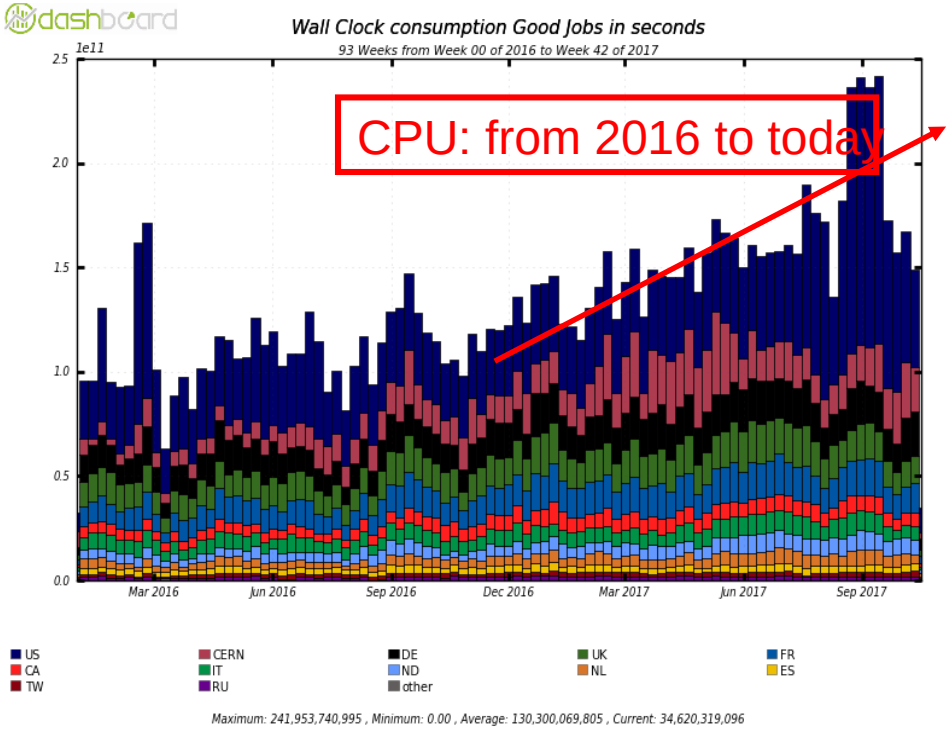
- Inputs
  - Same budget/cost basis as in 2017
  - Tape renewal has a cost: 1pB for 4VOs
  - Potential change in technology affects only the drives. No change foreseen in 2018
- Proposal for T1 (total)
  - 105 kHS06 (11.1% total ATLAS requests)
  - 8.1 pB disk (11.3% total ATLAS requests)
  - 22 pB tape (11.3% total ATLAS requests)
- Tentative for T2s (total)
  - 96 kHS06 (9% total ATLAS requests)
  - 8.3 pB disk (9% total ATLAS requests)

Resource usage in 2017 (1)

CPU availability in 2017

- Above pledges
- Dominated by Simulation
- MC limited stat is an issue (eg VH H to bb)





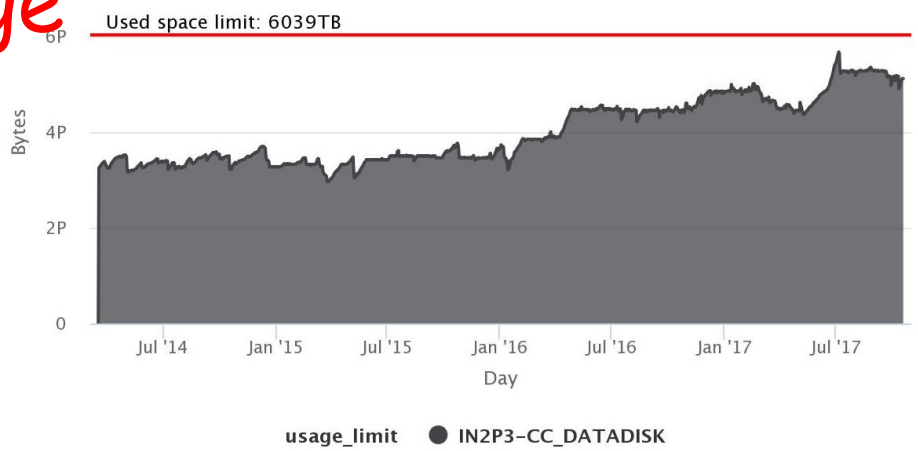
- 5-15 billion evts / day
  - few million/day simu.
- 2-5 PB of input/day
- >1 million jobs/day

Resource usage in 2017 (2)

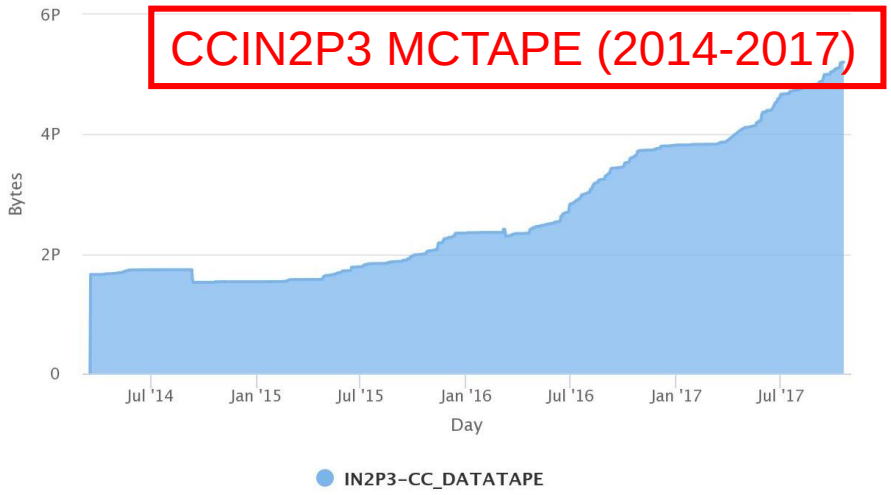
Storage increase & usage

- T1s disks full at 85%
  - +5-10% for tape staging
- T2s disks full at 90%
  - Old problem solved

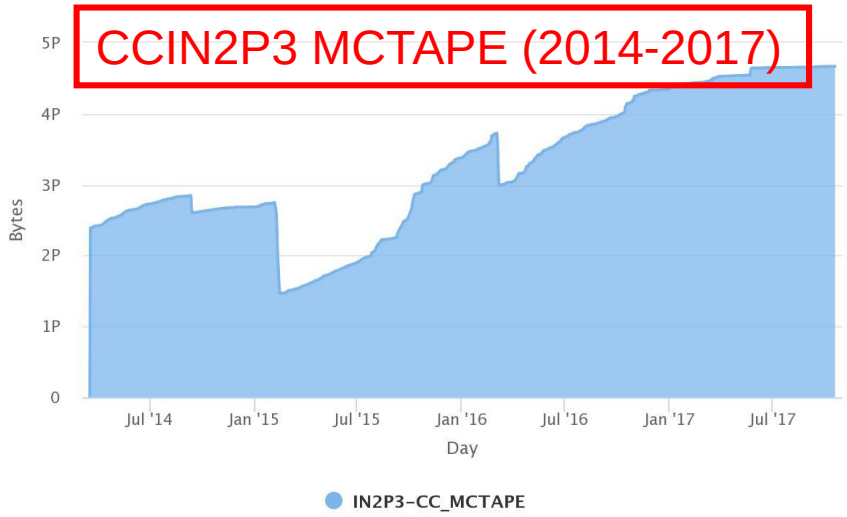
CCIN2P3 DATADISK (2014-2017)



CCIN2P3 MCTAPE (2014-2017)



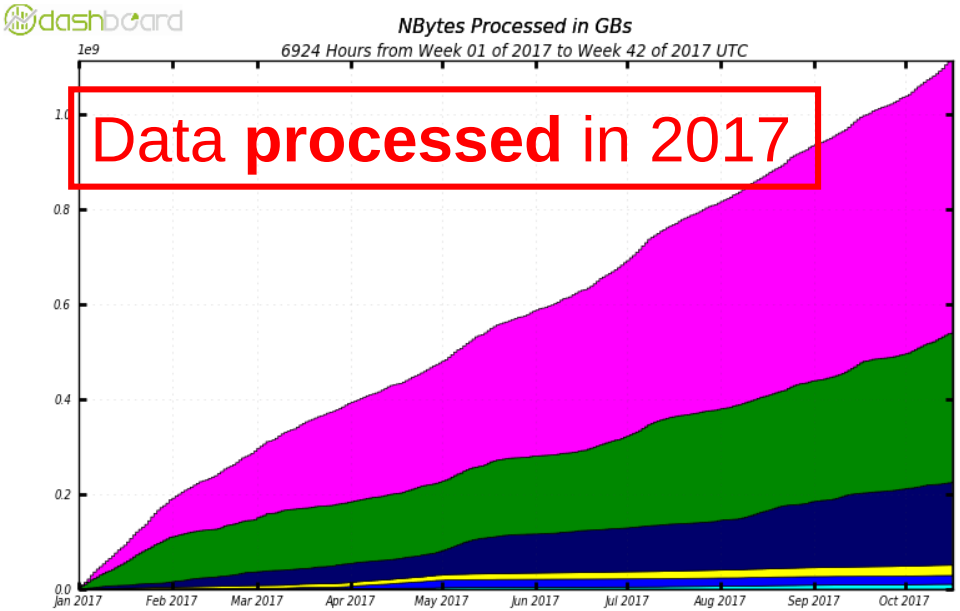
CCIN2P3 MCTAPE (2014-2017)



Data volume in 2017

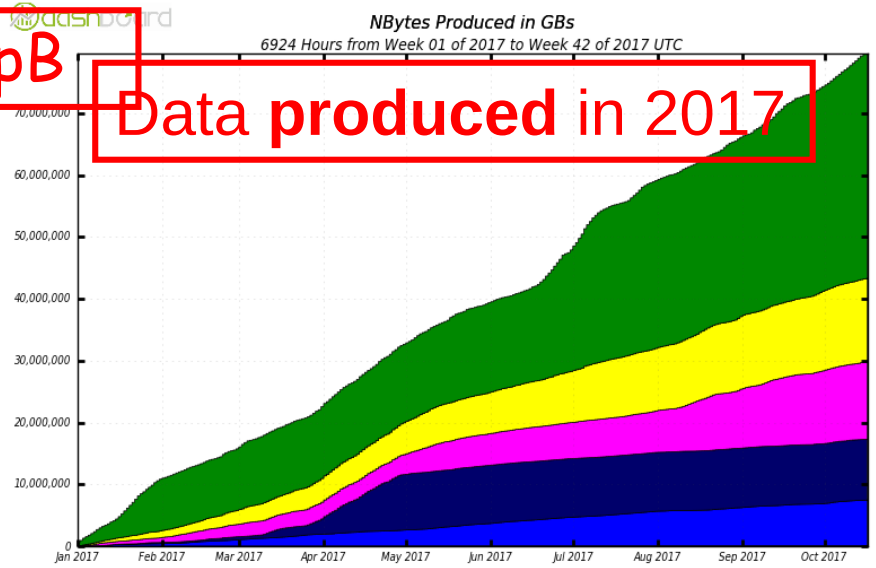
1.2 EB!!

The Exabyte Area/Era
Dominated by Analysis



Data **processed** in 2017

70pB



Data **produced** in 2017

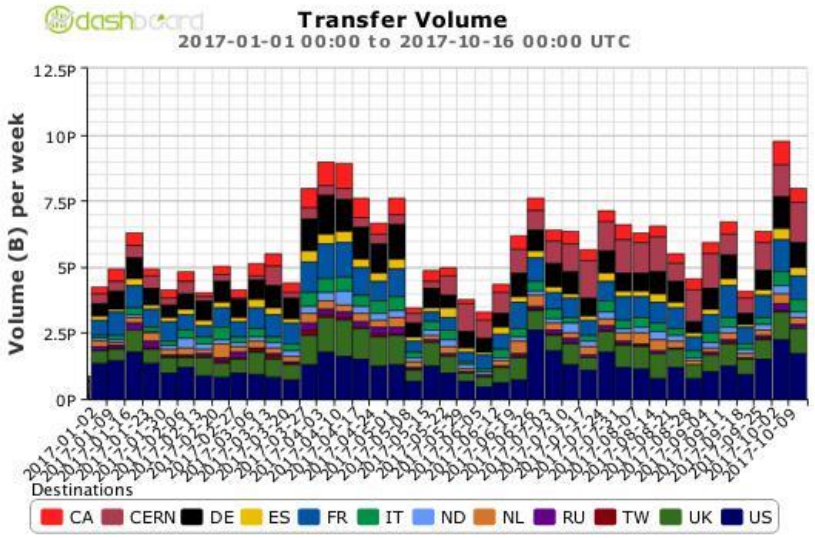
942,141)
9,254)
))

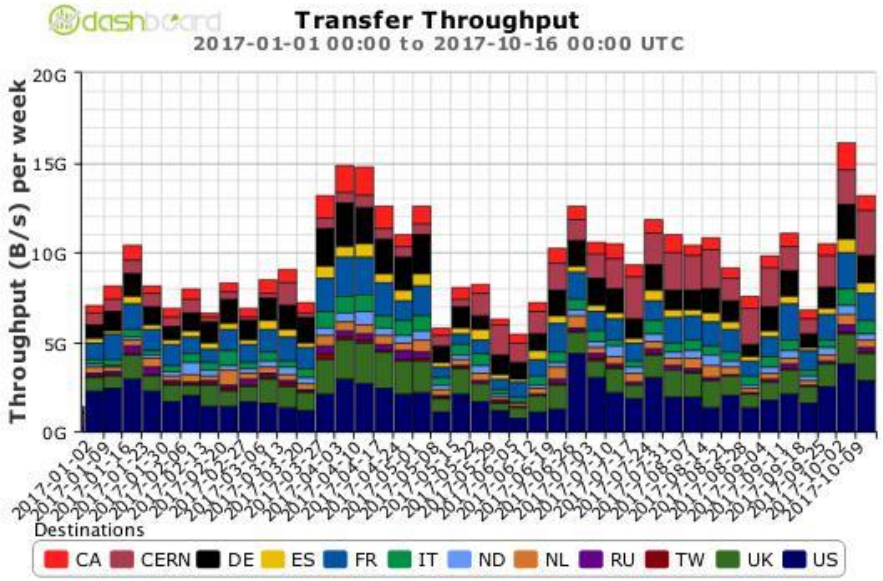
Group Production (314,917,469)
Data Processing (18,547,935)
T0 Processing (0.00)
MC Reconstruction (173,950,675)
MC Simulation (10,493,575)

Total: 1,114,591,053 , Average Rate: 44.71 /s

Group Production (36,407,411)
MC Simulation (7,383,844)
Analysis (13,496,897)
Others (70,050)
MC Reconstruction (12,499,589)
unknown (0.00)
Data Processing (9,844,247)
T0 Processing (0.00)

Data Transfers (Rucio)





Per week:

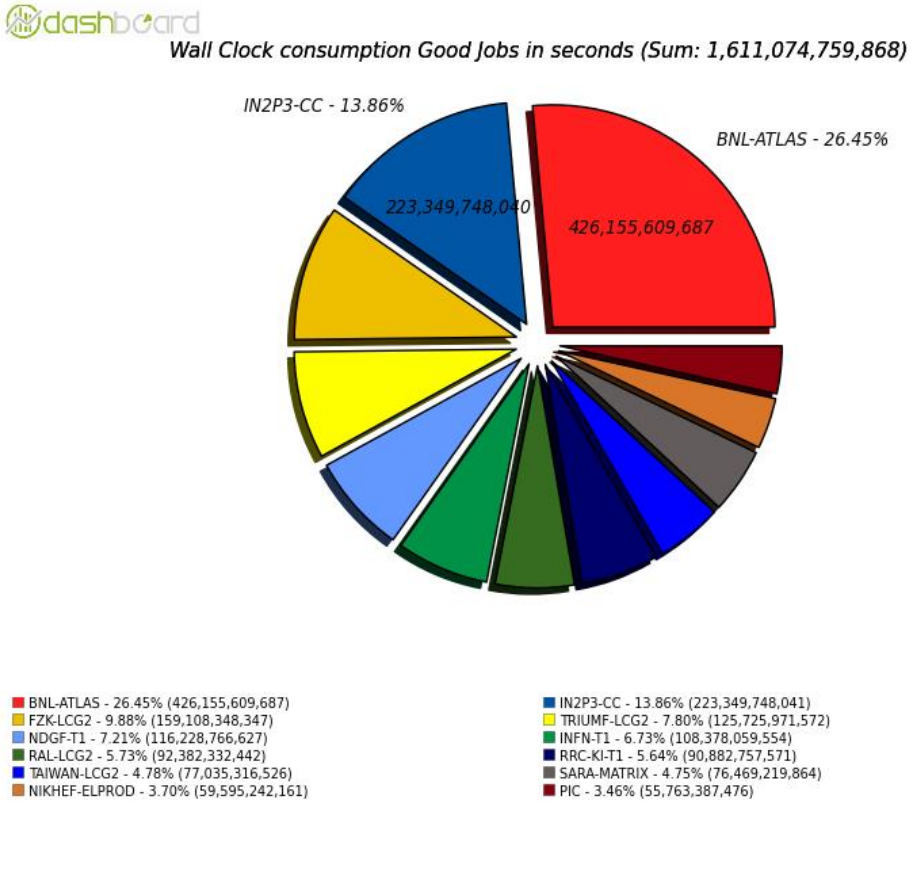
- 7pB data transferred
- i.e. 15M files
- ~ 100Gb/s bandwidth
- 10pB deleted!

Automation helped:

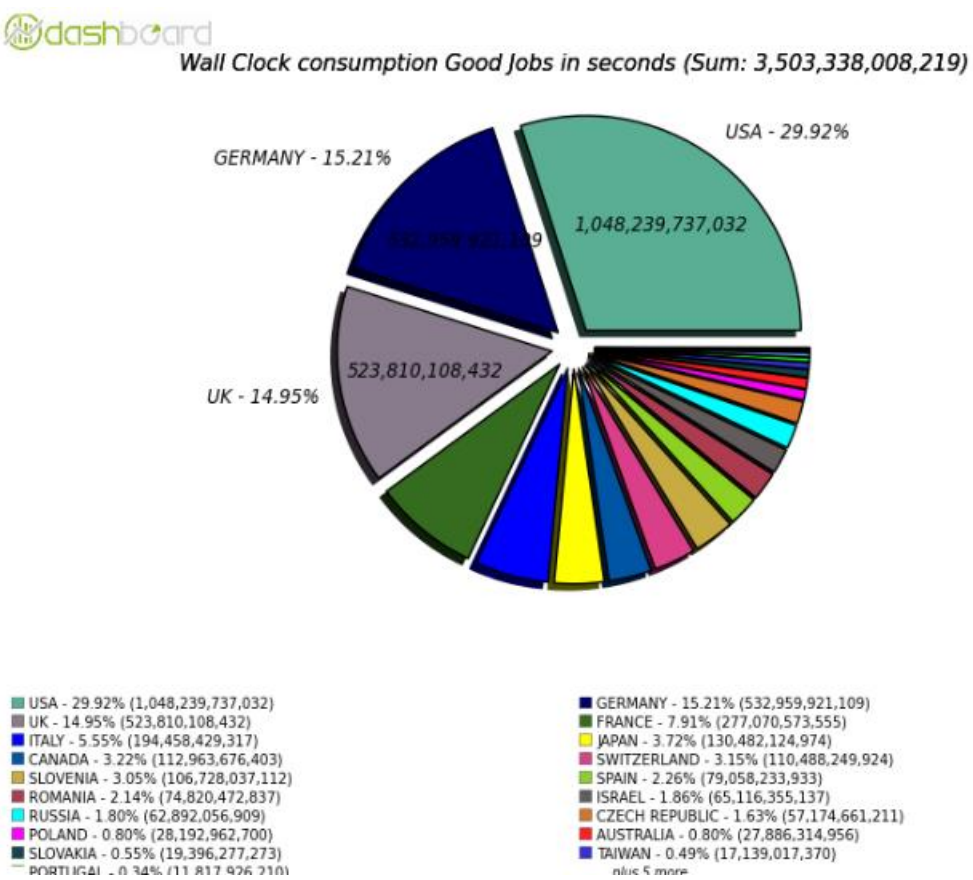
- Data pre-placement ✓
- Data replication ✓
- Data rebalancing ✓

France in 2017: CPU

- **T1s**



- **T2s**

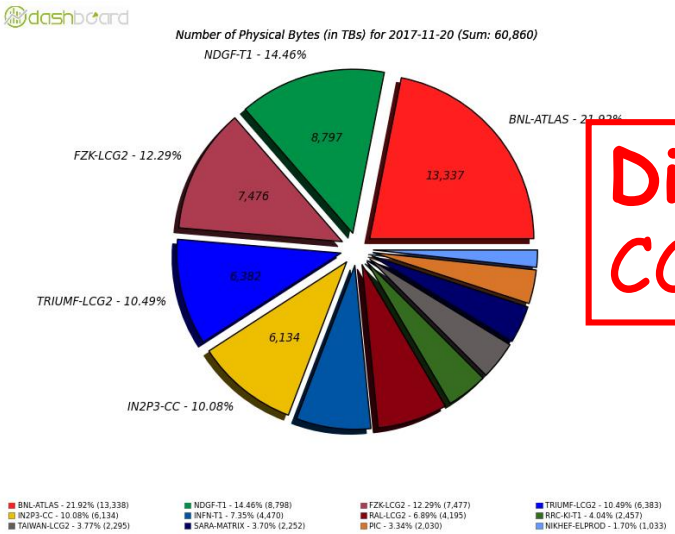


CCIN2P3: 14% (2nd)

France: 8% (4th)

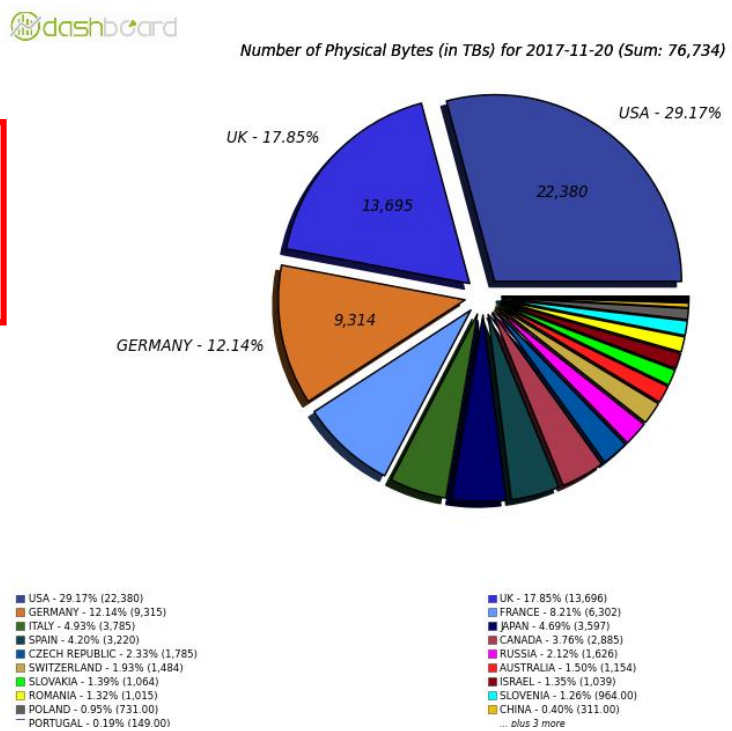
France in 2107: Storage

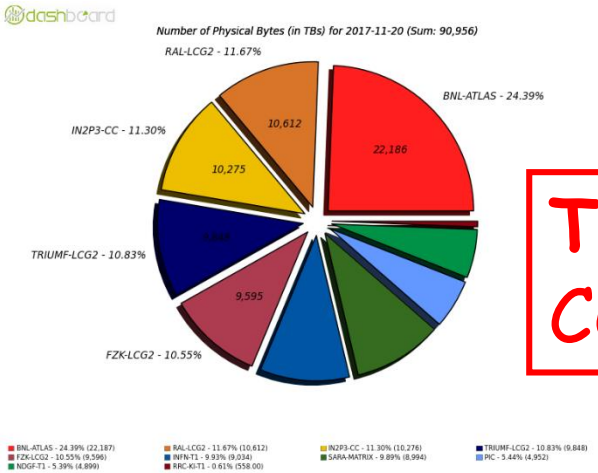
- **T1s**



Disk
CC: 10% (5th)

- **T2s**





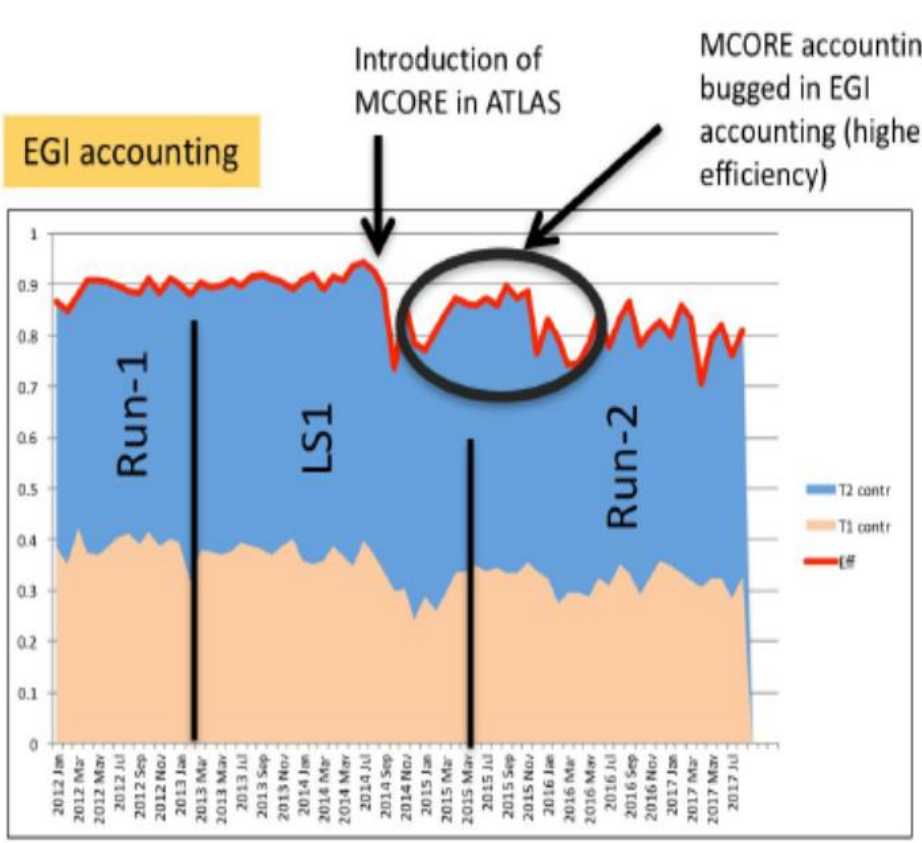
Tape
CC: 11% (3rd)

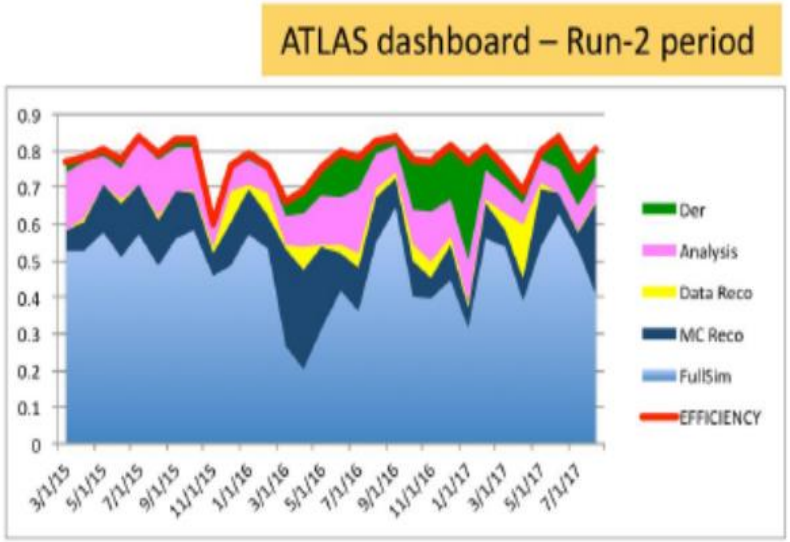
Disk
France: 8% (4th)

CPU Efficiency

Loss of CPU efficiency (CPU/WT) surveyed by LHCC

Introducing MCORE caused a loss of 10% efficiency. Some of this is real, due to serial operations in MCORE environment (e.g. I/O). Some is artificial (initialization accounted differently in MCORE and SCORE)



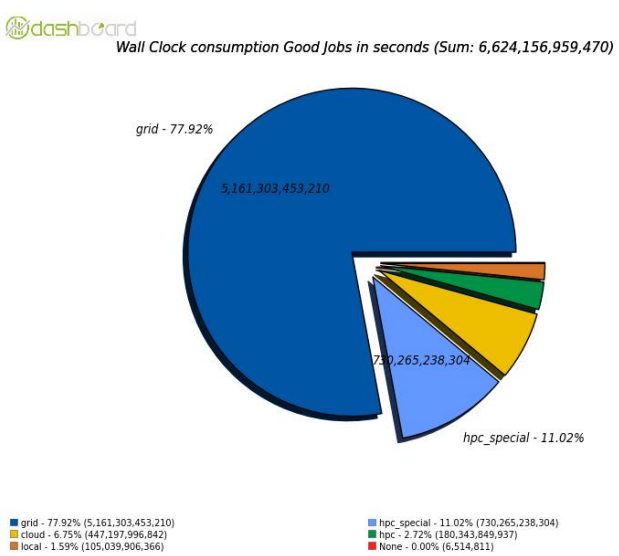


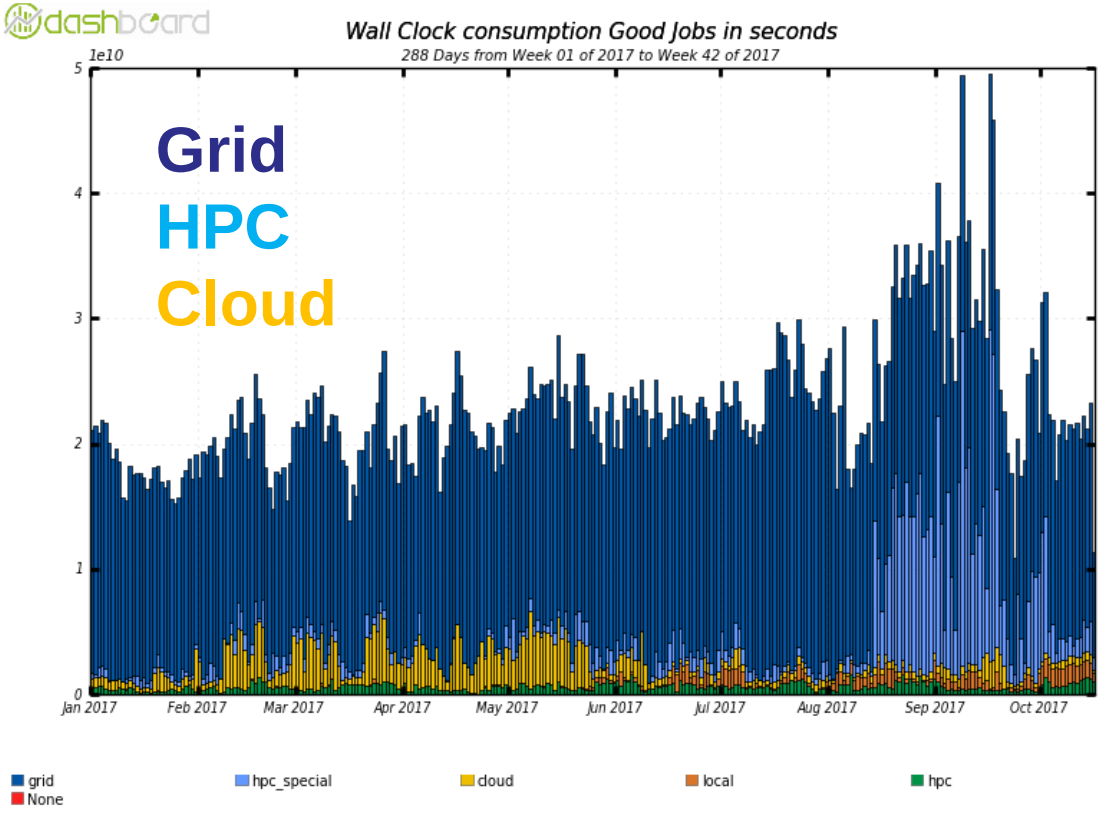
Conclusion-I: no obvious drop of ATLAS efficiency in Run 2

Conclusion-II: that does not mean we should accept 80% is good. We continue our work understanding performance and improving it

e.g. using checkpointed images to reduce serial init

Extra resource (1)



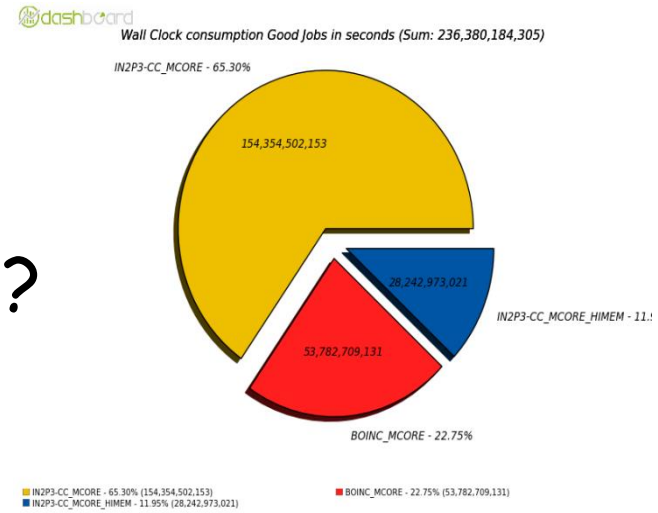


Over 2017

- Grid 78%
- Cloud (academic, commercial, BOINC) 6.7%
- HPC_Special (NERSC_CORI, ORNL_Titan) 11%
- HPC (Local) 2.7%

Extra Resource (2)

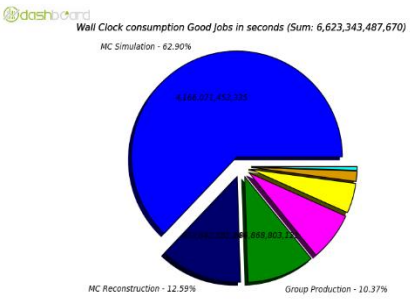
- Cloud
  - Stable but no real increase
  - Cost gain wrt grid? Manpower?
- ATLAS@Home: Increasing
- HPC
  - Complex to set up
  - Test by CC@IDRIS. OK but 2.5k/10k slots max
- General
  - Harvester under development (common interface for ALL type of resource)
  - Event Service: Work at event level (simulation)

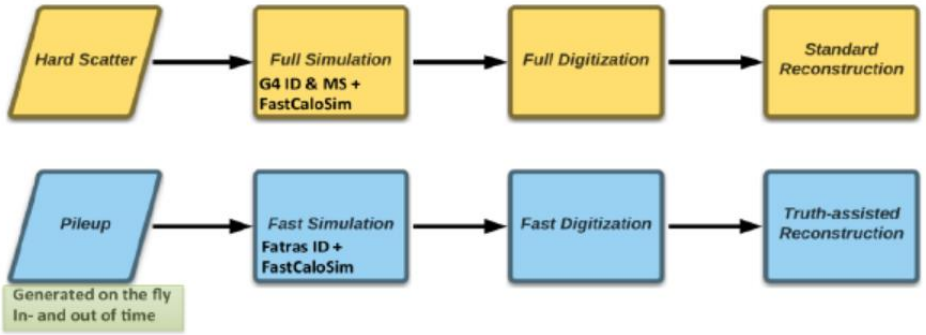


2017: WT BOINC (11k slots,22%) vs CCIN2P3

Progress on simulation

- Dominates CPU needs and adds systematic to some physics studies (eg VHbb)
- FastSimulation
  - Validated in Sep'2017
  - Uses FastCaloSim
  - Test of FastChain
- Pile-up treatment
  - Standard vs Overlay





- Combining multiple HITS level events during pile-up digitization is not efficient in terms of CPU or I/O.
  - Event Digitization time w/pile-up ~ Event Simulation time w/ATLFASTII
  - Current approach relies on high I/O to avoid high memory usage.
  - The situation will get worse as <mu> increases.
- Data Overlay (combining a digitized hard-scatter event with a Zero-bias data event) has been used successfully for heavy ion campaigns.
  - Lower CPU, memory and I/O requirements.
  - Places a huge load on conditions database infrastructure, due to jobs requesting different conditions IoVs.

Analysis: Issues & Progress

- Run2 Model (xAOD, Train & Derivations) OK
- Assumed
  - Derivations run fast enough -> able to run new full derivations every few weeks if necessary
  - Derivations output small -> grid analysis jobs able to process all data/MC in O(1 day)
- In real
  - ~6 weeks for repro all data into O(80) DAOD
    - -> Run only major productions
  - Analysis jobs tails >>1 day
    - Write larger outputs to reduce #grid iterations
- Solutions
  - DAOD production: More efficient merging, reduce size
  - DAOD processing: Improvement in Distributed Analysis

HEP Software Foundation (HSF)

- Goal: facilitates coordination & common efforts in HEP software and computing internationally
- ATLAS part of it & benefit from HSF
  - In WFM and DDM we have strong software we should advance as possible community standards
    - Rucio, PanDA, Harvester, the package of all of them
- Output: CWP (Community White Paper)
  - Scope: HL-LHC
  - Items addressed: Flexible management of facilities, use of heterogeneous resource, Computing Models, Facilities, Distributed Computing
  - LAL involved (M. Jouvin et al.)
  - http://hepsoftwarefoundation.org/index.html

Lines of effort

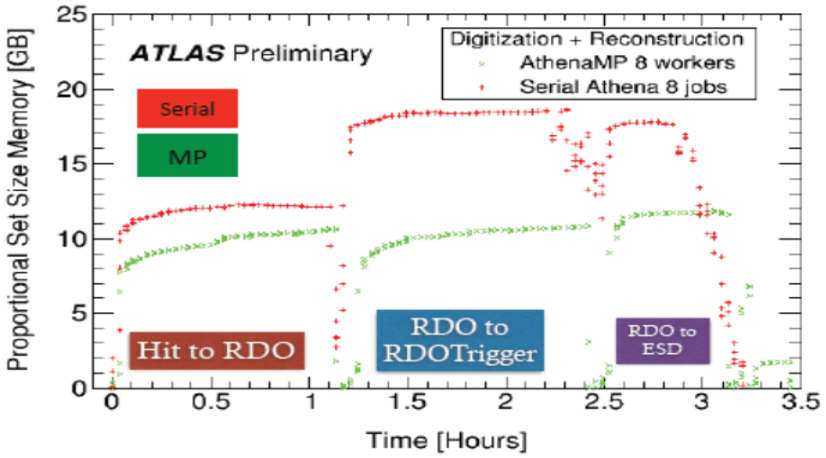
- T1s continue to exercise and improve perf. of DAOD production from tape inputs
- Promote support for software development
  - Supporting software activity where the effort is crucial and currently insufficient
- Support the roll-out of containers
  - Powerful technology for improving uniformity, ease of ops & security across resources. Singularity (see Eric's talk) deployment model agreed by WLCG and ATLAS. Proactive deploymen on grid sites is proceeding
- T3s policy: Will to limit 'bad' T3s / small sites

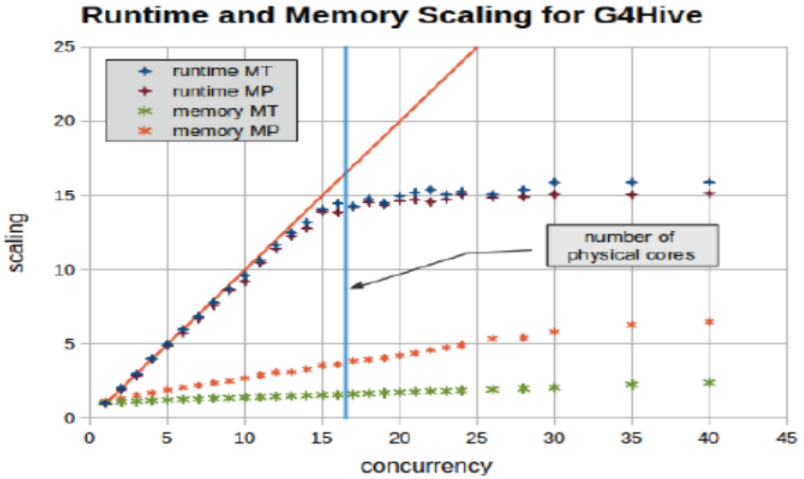
Towards Run-3

- Going slowly away from hierarchical T0/T1/T2 mode: Nucleus/satellites model
  - Storage, transfers & resource optimized
    - eg Now T2s disk are full
  - I/O further optimized
- New Database scheme
  - To replace COOL & handle Overlay (IRFU, LAL)
- Key words
  - Harvester, Event Service, Overlay
  - Containers: Virtualization for batch execution (Singularity: see Eric's talk)

Reconstruction for Run-3

- Going from multi-event parallelism (MP) to inter-event parallelism @algo level (MT)
- Compulsory for high-pileup tracking (Run-3)
  - More memory needed for Reco (pile-up)
  - Memory/core available not increasing
- Ready for Run-3 start
  - Algorithmic migration will take some time





Towards HL-LHC: Inputs

T. Wenaus

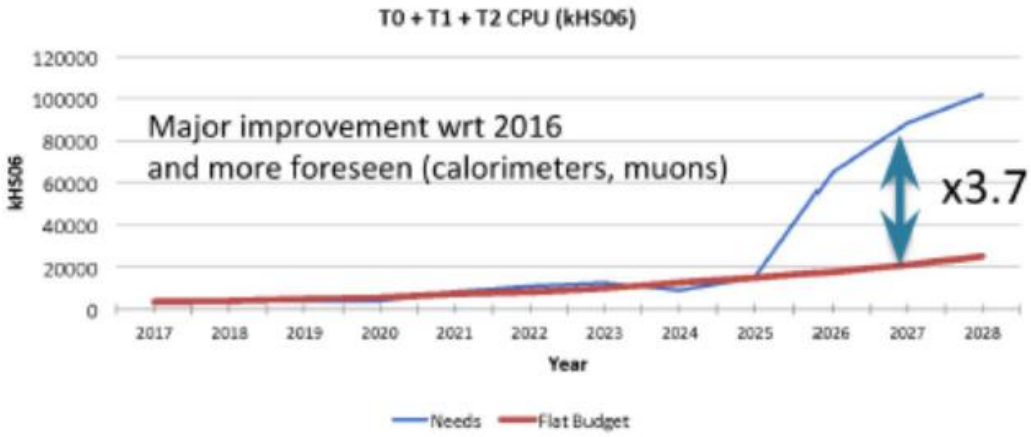
Input Parameters at HL-LHC,
updated after the conclusion of the Layout Task Force

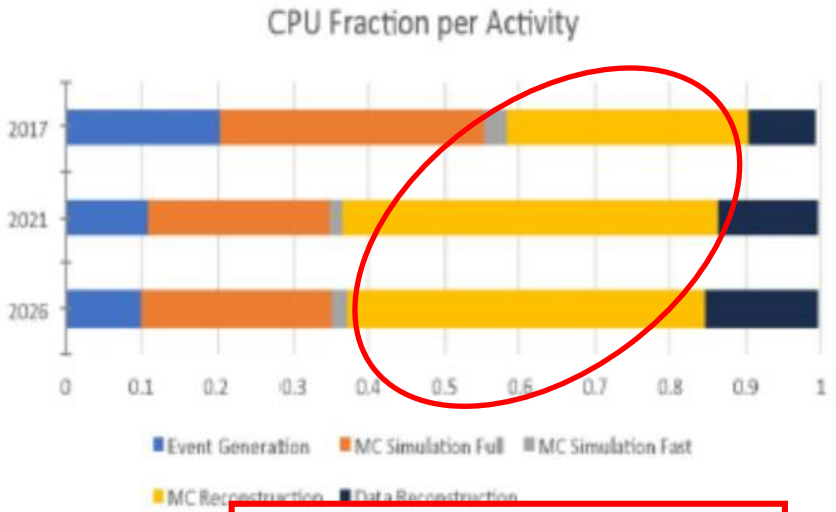
Output HLT rate: 10kHz
Reco time: 130s/event at mu=200, Simul Time: 454 s/event
Nr Events MC / Nr Events Data = 1.5
N events with Fast Simulation: 50% of Full Simulation
LHC live seconds /year: 7.3 M

Flat budget
20% more CPU/year, 15% more storage/year
Evolution from a 2017 baseline
Data from previous years taken into account
Tier-0 contribution added to the total

Towards HL-LHC: CPU

T. Wenaus





Dominated by MC Reco

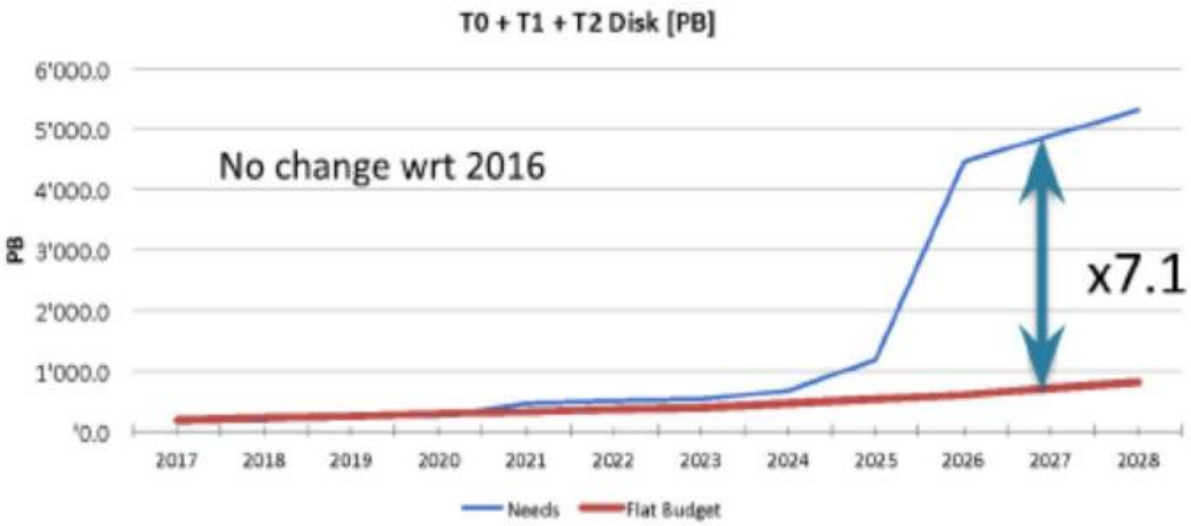
CPU: the gap has shrunk significantly with e.g. substantial improvements in reco time and improvements of the model

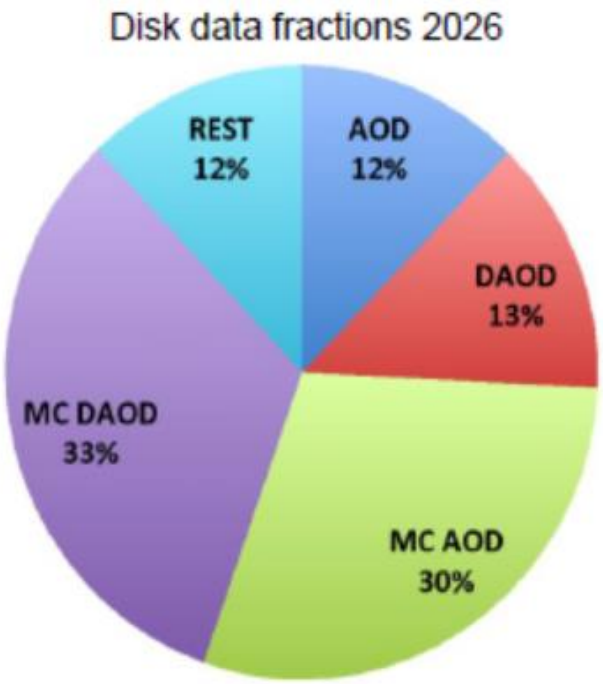
As other subsystems follow ITk in making reco improvements we can expect the gap to shrink further

Also relies on a much larger role for fast simulation

Towards HL-LHC: Disk

T. Wenaus





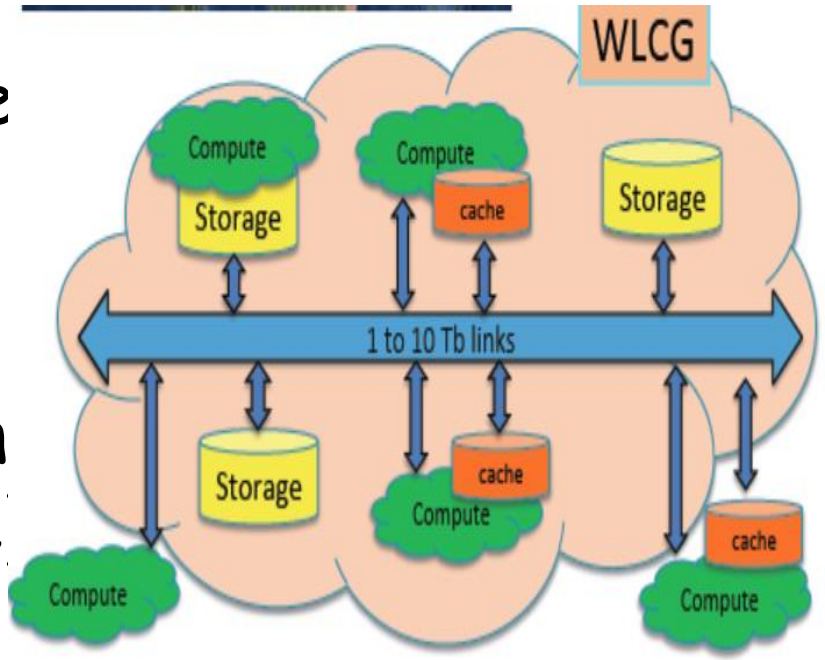
Storage: the gap won't shrink until we develop and quantify the strategies to bring it down

This we are starting to do, e.g. with a program of testing and improving train workflows using tape as input

Possible gains

- Improvement on CPU
  - Detector layout (TDR)
  - Machine learning technique
  - Fast simulation/Fast chain
- Improvement in Storage
  - No AOD on disk (Run Train analysis from AOD on tape
- Not enough. Gain to come
  - From re-thinking of distributed storage an& data access
  - A network driven data model allows to reduce the amount of storage, eg disk

Network driven 'data lake'



Summary

- ATLAS computing in very good shape!
- Now able to focus on refinements, performance, and look to future with R&D
- ATLAS should be front and center in common R&D (inside HSF community)
- Run-3 a priori OK within flat budget. Key issue is software: AthenaMT
- HL-LHC
  - Trend lines are good in CPU (constant progress)
  - Plans in storage to be quantified (today critical)
  - R&D, 'Data lake' model, non-flat budget?