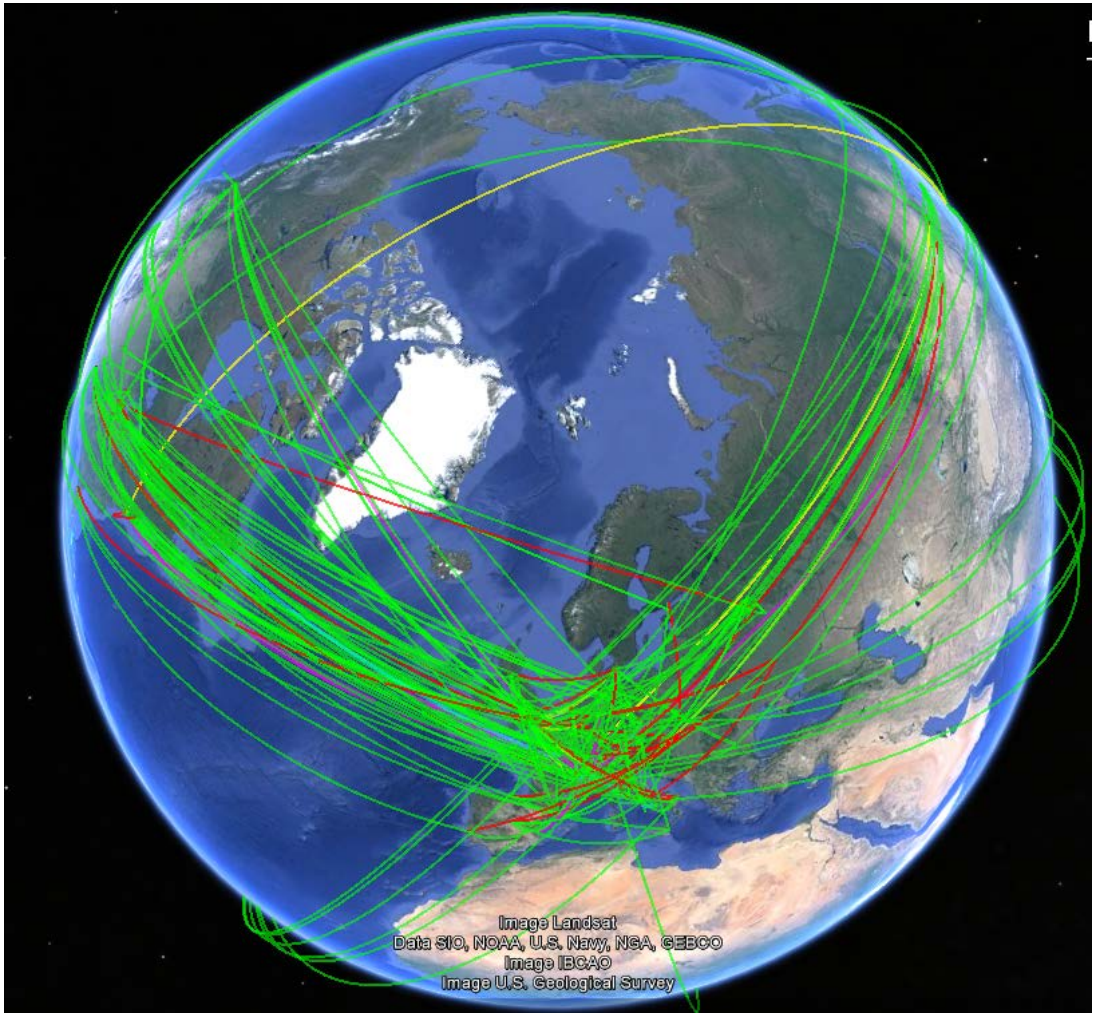
ATLAS Software & Distributed Computing Readiness for LHC Run2

L. Poggioli, LAL, Orsay

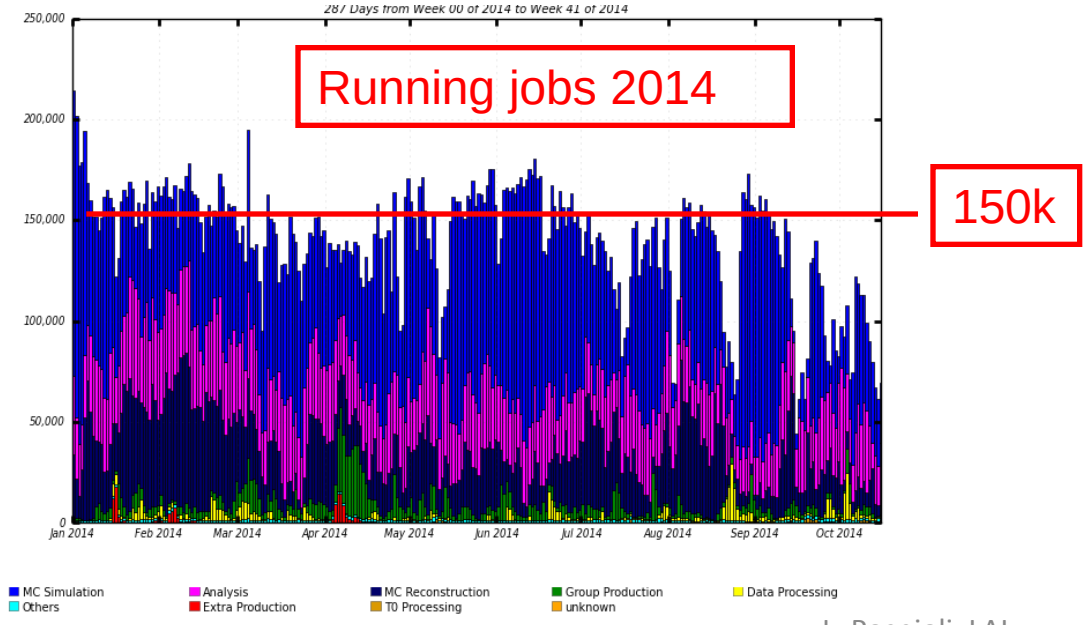


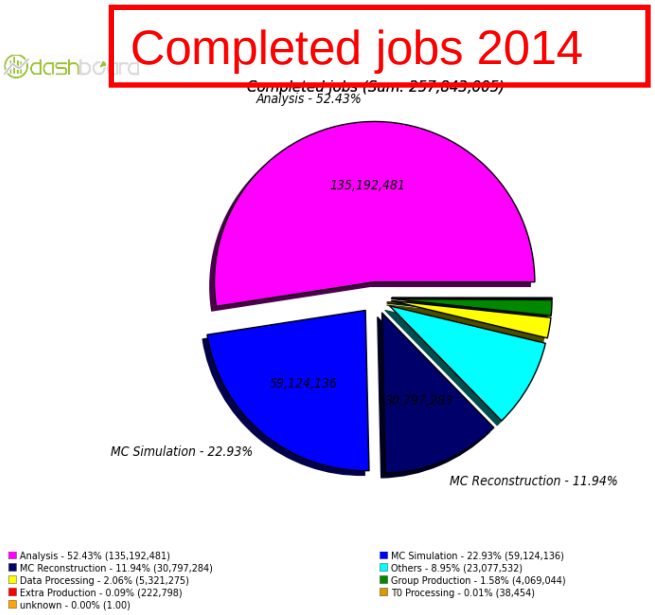
http://dashb-earth.cern.ch/

Run1 Grid Activity

- 350M jobs in 2013
  - Analysis jobs > 50%
- 80 pB disk, 70 pB tape
  - 1.2EB data read-in
    - > 82% by analysis

Analysis: Main driver of storage & I/O capacity





Run2 Challenges

- Flat budget constraints
  - h/w increase from Moore's law gain
    - Estimated factors of 1.2/year for CPU & 1.15/yr for disk & tape
- Data from Run-1
  - Proper data preservation
- LHC operation
  - HLT rate x 2.5
  - Pile-up > 30
    - -> Reco time x 2-2.5
  - 25ns bunch spacing
  - c.m. energy x 2
- 'New' detector
  - To be integrated in simul & reco

With shrinking manpower (-10% in 2013)

Limitations of current model & tools

- Partitioning of resources
  - Analysis vs Central Production - T1s versus T2s
- Data distribution management & Production systems limits to scale to new conditions
- Memory increase of MC pile-up digi & reco
- Multitude of data format for analysis

-> Gain needed in Simulation (CPU), Reconstruction (CPU, memory), Analysis (Data format, disk space, CPU)

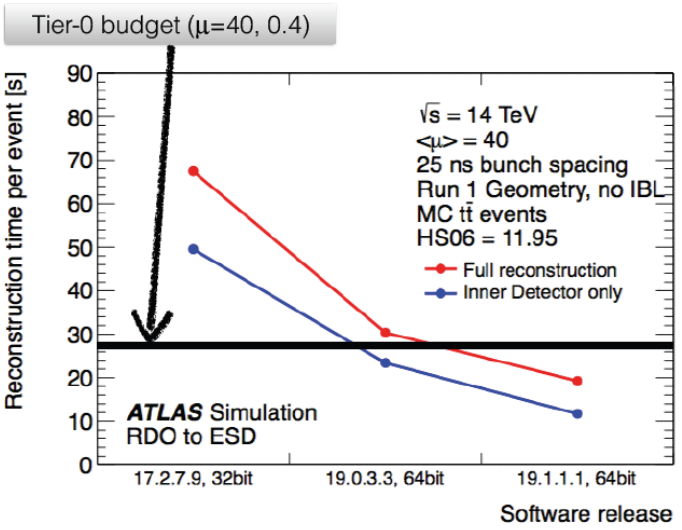
Run2: Extrapolation & extension of end of Run1 framework

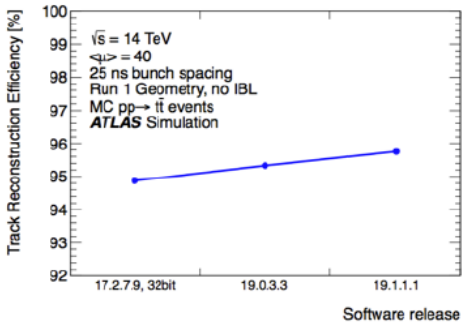
ATLAS Challenges

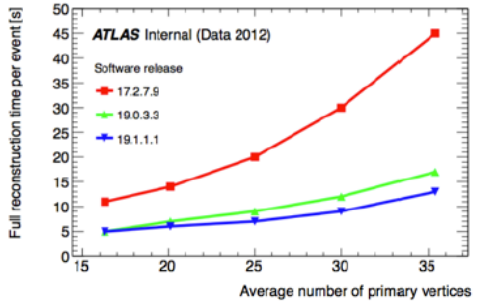
- Software
  - Get Integrated Simulation Framework (ISF) in prod
  - Speedup reconstruction by a factor 2-3
  - Migration to new data format xAOD readable both from Athena and ROOT and new Analysis model
  - Changes/Upgrade of the infrastructure : ROOT6, CMake, Tag Collector, JIRA…
- Computing
  - New data management system : Rucio
  - New production system : ProdSys2
  - New data management strategy : each dataset has a lifetime (disk & tape)

Software

- Improvements
  - Technical
    - 32->64 bits, faster vectorized maths library, algebra package CLHEP->Eiger
  - Track seeding
    - From strips + IBL use
  - Backtracking w/ RoI seeds
    - In TRT only w/EM calo deposits
- Gain
  - Reconstruction time
  - Pile-up & Tracking efficiency

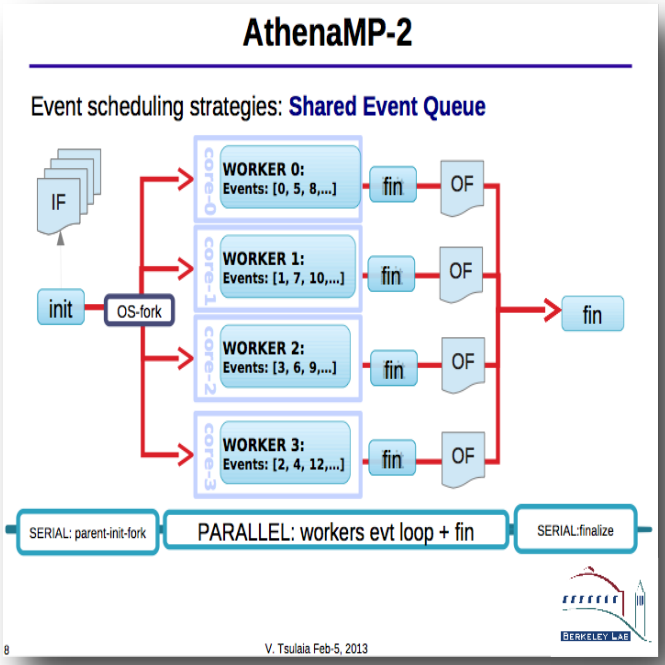


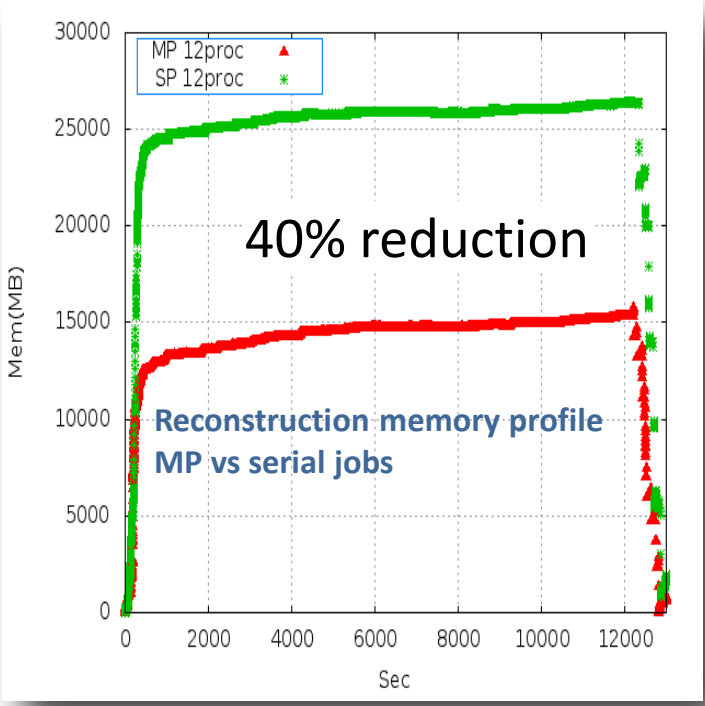




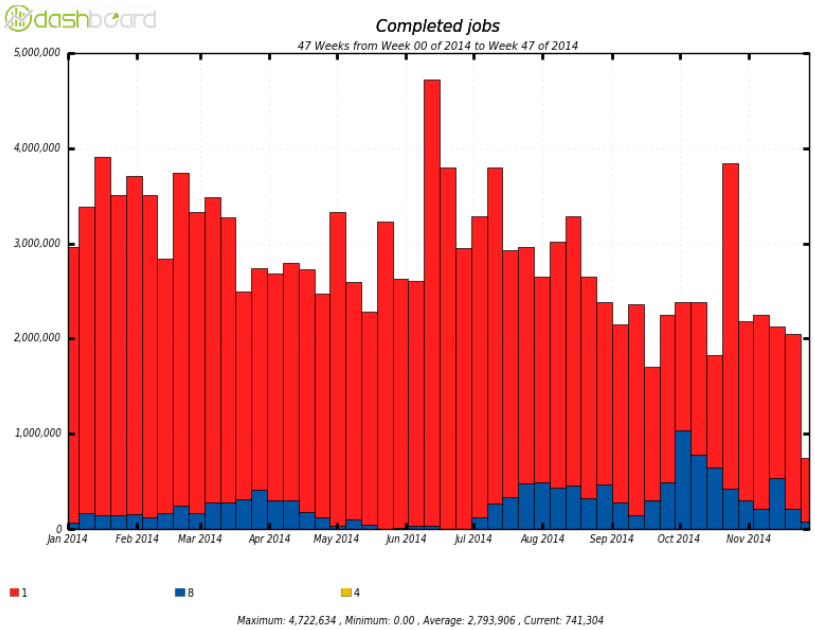
Memory saving with MultiCore jobs (1)

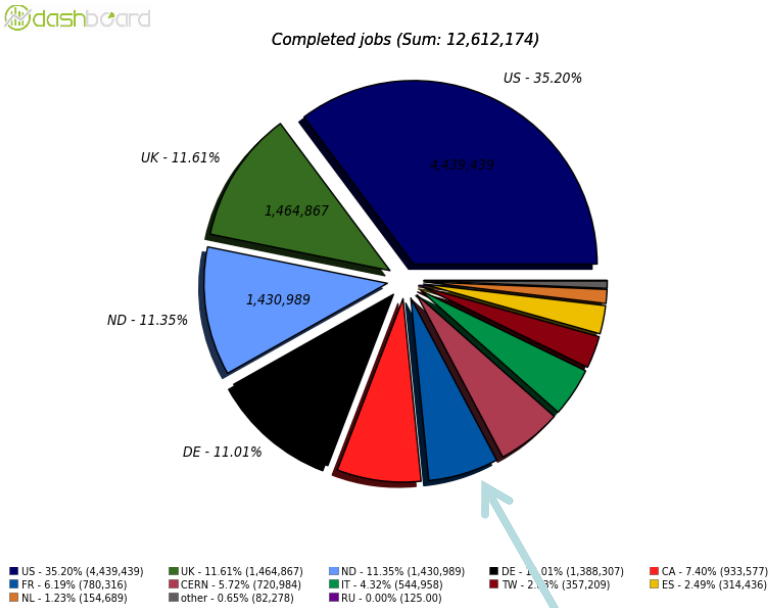
- Multi-Process Athena (AthenaMP)
  - Event parallelism to reduce memory footprint
  - For Simulation & Reconstruction jobs (not Analysis)





Memory saving with MultiCore jobs (2)



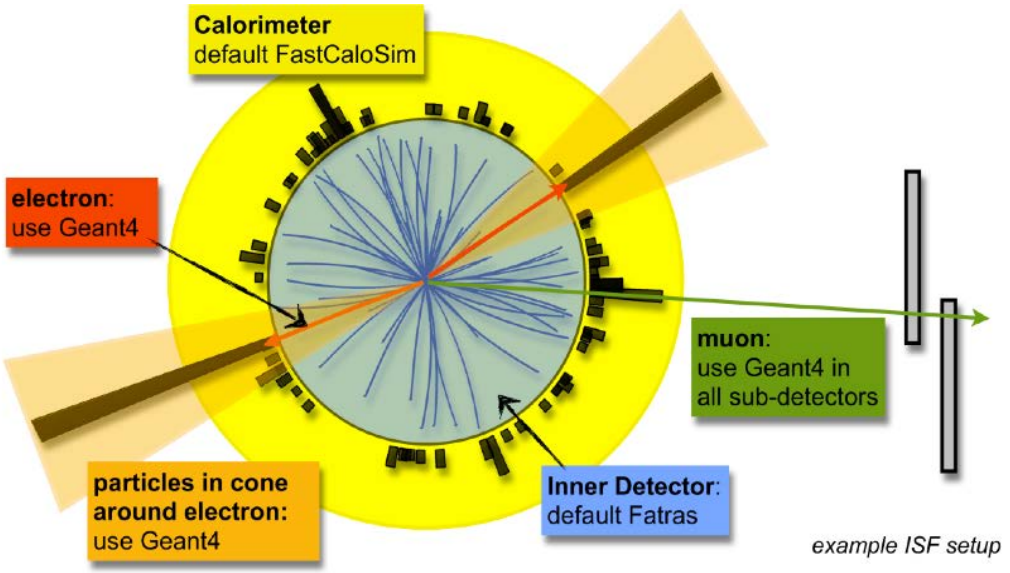


FR: 6.2%

- In FR-cloud
  - Started late but OK now
  - Mostly static (CCIN2P3, Tokyo, RO-07)
  - Dynamic (LAPP scheme: LAPP, CPPM, LPSC, LPC)
  - IRFU: Also test ARC-CE + HTCondor

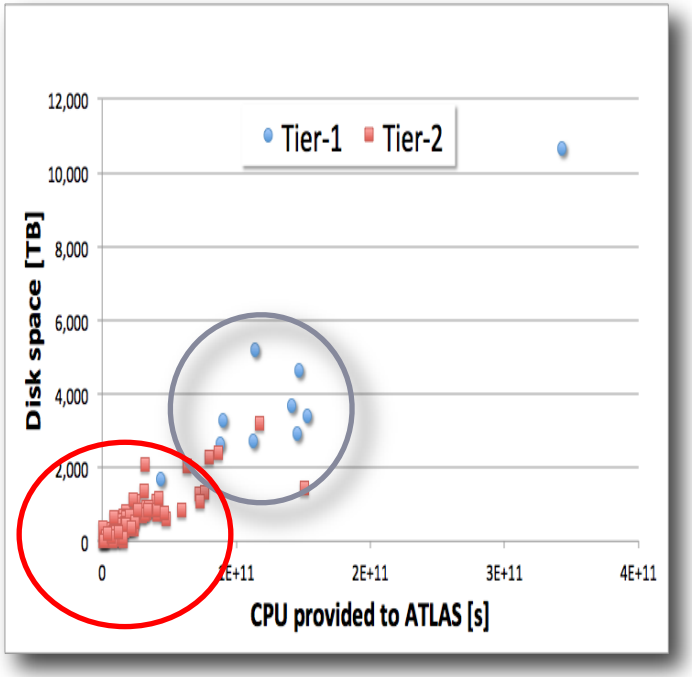
Fast Simulation

- Take advantage of fast sim when appropriate
- Tradeoff accuracy for speed: Smearing, Frozen Showers, Parametric techniques
- Integrated Simulation Framework (ISF): clever mixing of fast and full simulation within same evnt
  - Keep high precision for some particles and regions
  - Use fast simulation in areas that are not so important
  - x100 speed ups possible with much better results than normal fast simulation



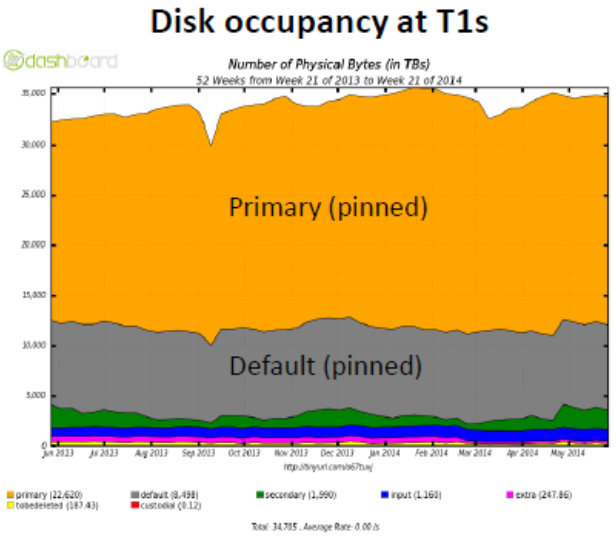
Data Processing

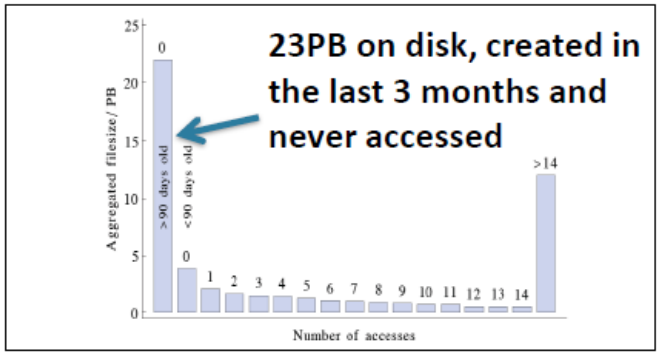
- Flexibility to be introduced
  - Some T2s are equivalent to T1s in disk & CPU
- Operational load of many 'small' sites
  - Less & larger sites be better
- Data processing
  - Optional extension of 1st pass processing from T0 to T1s in case of resource shortage at T0
  - T1s & **some** T2s used for most demanding wkflows high memory and I/O intensive tasks
    - Data reprocessing & MC reco also done at **some** T2s



Storage management

- Saturation of T1s DATADISK
  - Lot of primary space occupied by unused data
  - 'Less ' critical for T2s



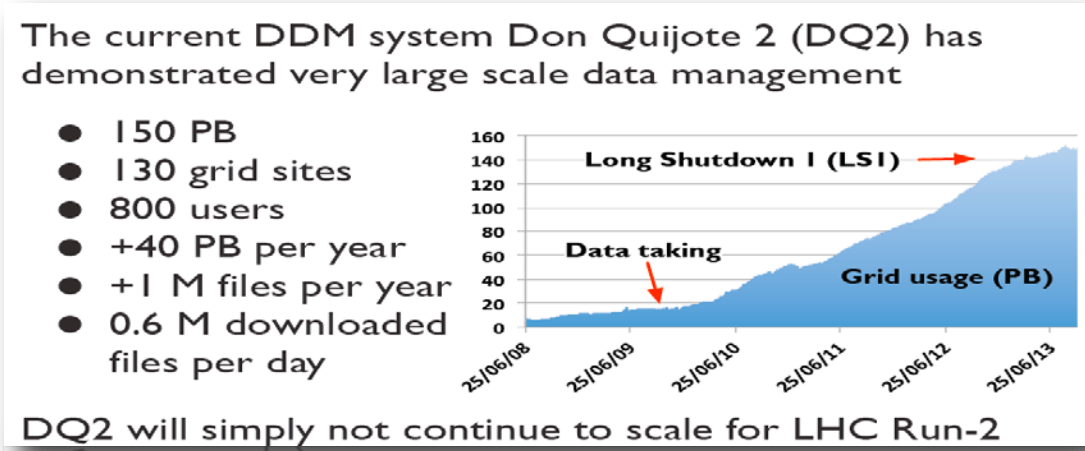


8 PB of data on disk never been touched

- Very agressive dataset policy being set up
  - Each DS -> lifetime (disk&tape) & retention policy
  - Tape more used for Run2

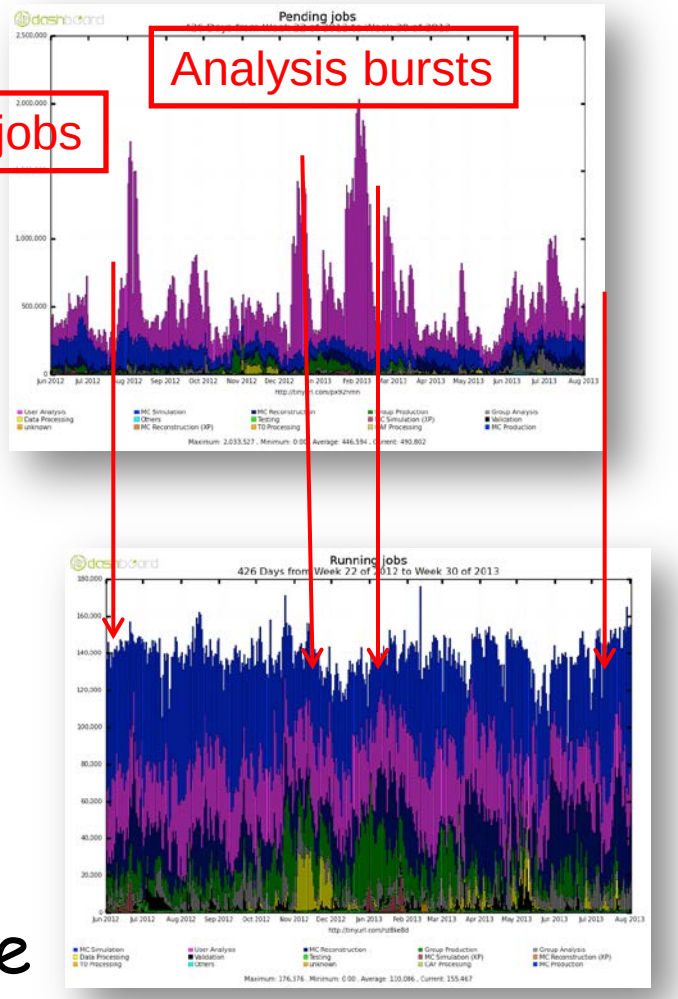
Data Management System

- New system: Rucio to replace DQ2
  - New scalable architecture
  - File level functionality instead of dataset
  - Built-in data replication policy for space and network optimisation
  - Multi-protocol
- Status
  - WebDAV deployed for data renaming
  - No more LFC
  - Final migration DQ2-> Rucio NOW!



New production system: ProdSys2

- Same engine for analysis & production
  - PanDA+JEDI+DEfT
  - Current analysis vs prod. shares managed by sites
  - Better reactivity to Ana. loads
  - Job <-> resource optimized
- Status
  - Moved to JEDI for Analysis in August
  - Tested against most workflows
  - New BigPandaMonitoring in place
  - Final migration NOW

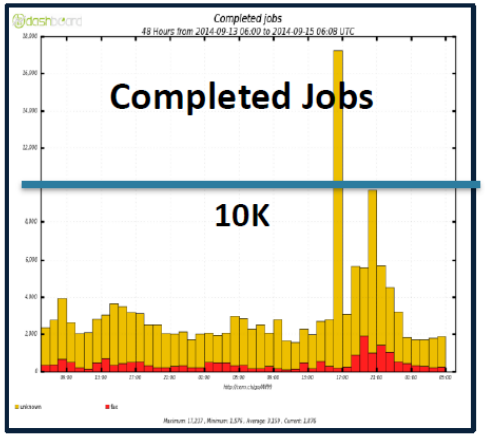


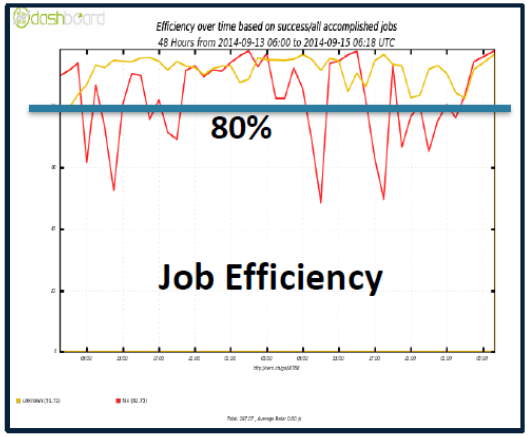
Remote Data Access

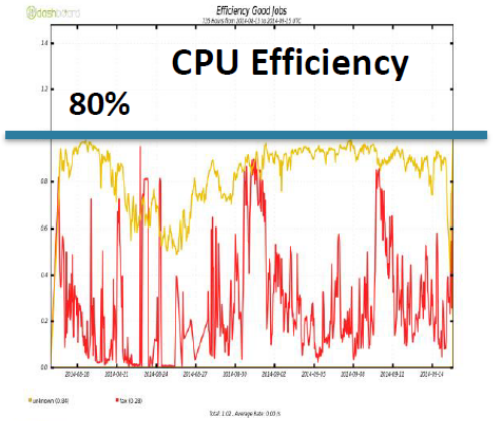
- Networking keeps on progressing fast
  - x10 every 4.25 yrs / Already 100Gb/s among US
- -> Jobs can access data remotely via network
  - Allows better usage of storage resources
  - Breaks the 'jobs go to data' Grid paradigm!!
  - Better suited to Analysis jobs
- Protocols
  - http: Allows direct download files from Grid to local: Not in quality production today
  - Xrootd: Allows direct data acces in ROOT & analysis s/w

XrootD/FAX

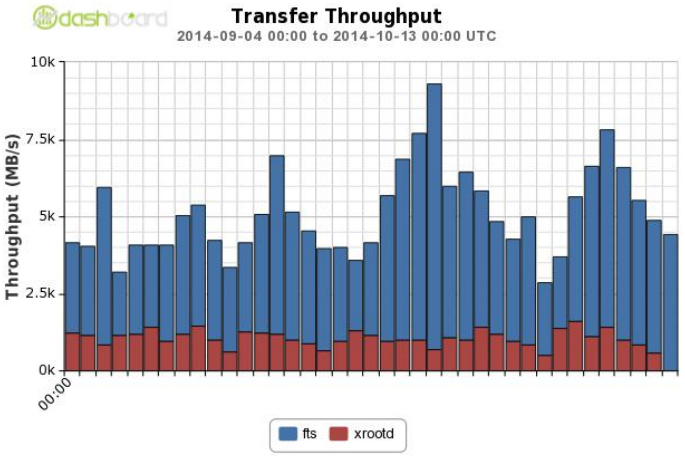
- FAX
  - >96% data coverage. All French sites OK (+Tokyo)
  - Overflow tested in US (CPU eff. to be improved)





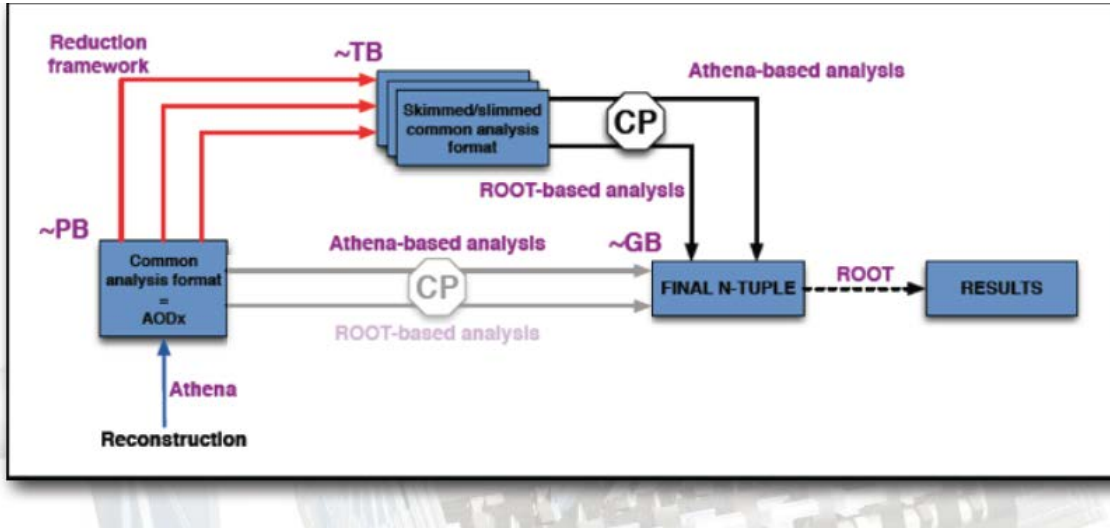


  - Monitoring: WLCG dashboard
    - Allows to choose FTS/XrootD & Xrootd remote/local



- Goal: 10% max access via FAX to protect sites

New Analysis Model



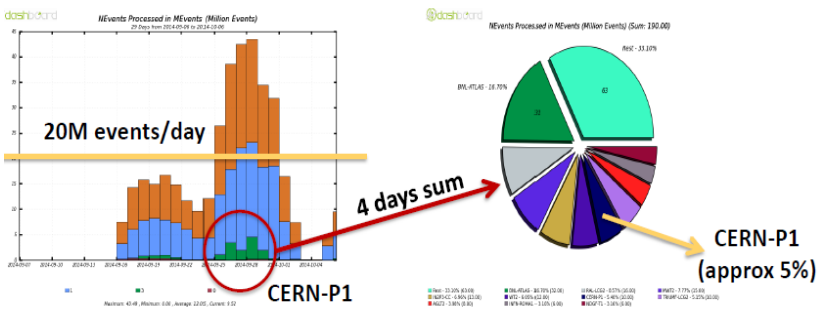
- Common analysis data format: xAOD
  - replacement of AOD & group ntuple of any kind
  - Readable both by Athena & ROOT
- Data reduction framework
  - AthenaMP to produce group data sample
    - Centrally via Prodsys
  - Based on train model
    - one input, N outputs
    - from PB to TB

- Status (tested on DC14)
  - 57 defined (ongoing), data will go to DATADISK
  - Total space for all derived datasets ~ 2x primary xAOD size
  - Dieting campaign on xAOD size ongoing

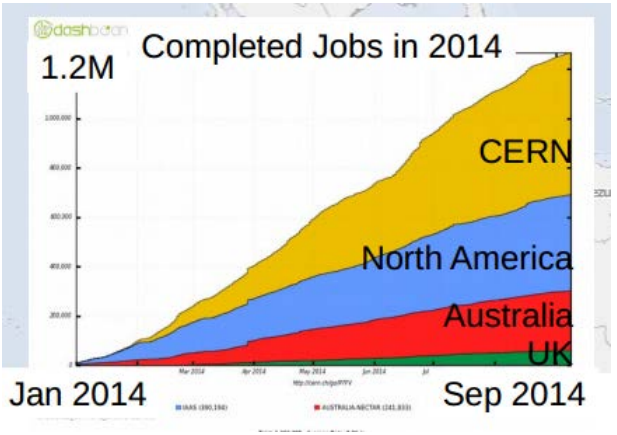
Opportunistic resources: Virtualization

Ask for resources thru interface and get access & control of (virtual) machine, i.e. job slot on Grid

- High Level Trigger (HLT)
  - Use resource between fills
  - CPU power ~T1 or big T2



- Clouds usage
  - Academic (focus on OpenStack @ CERN)
  - Commercial (Amazon AWS, Google GCE)
  - Use image distribution service (addition to OpenStack, GLINT)
  - 1.2M jobs (Jan-Sep 2014)
    - 1core & Mcore

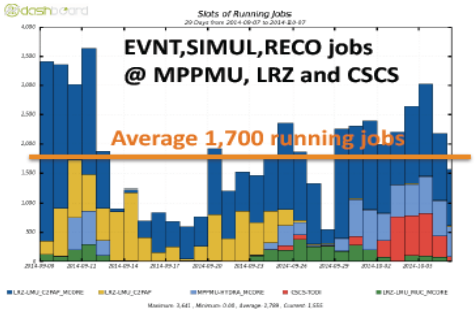


Opportunistic resources: HPC

- From Peta to ExaFLOPS
  - Large # CPU cycles used parasitically
    - eg MC simulation (10% Grid production, 10-20k cores)
  - Issues: I/O & outbound connectivity
- Natural interface to HPC: ARC-CE
  - Handles I/O, integrated into ATLAS prod via aCT

TITAN: 20pFlops, 700k Cores

Oak Ridge Titan System	
Architecture:	Cray XK7
Cabinets:	200
Total cores:	299,008 Opteron Cores
Memory/core:	2GB
Speed:	20+ PF
Square Footage	4,352 sq feet



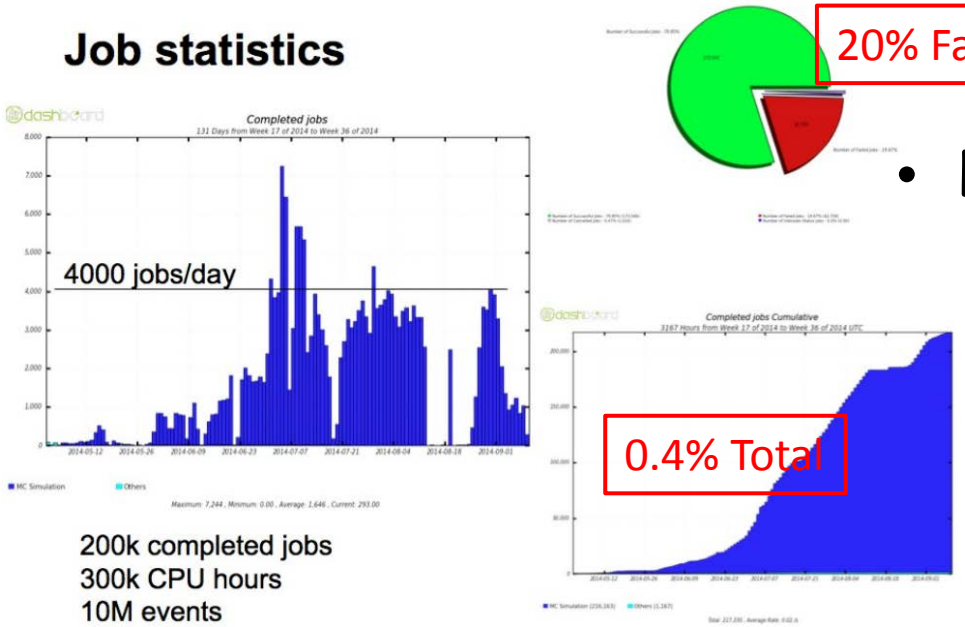
24h test at Oak Ridge Titan system (#2 world HPC machine, 299,008 cores). ATLAS event generation: 200,000 CPU hours on 90K parallel cores (equivalent of 70% of our Grid resources)

Mira@ARGONNE: Sherpa Generation using 12244 nodes with 8 threads per node, so 97,952 parallel Sherpa processes.

- In France : IDRIS (100k cores, 1.2pFlops)
  - See also F. Derue' talk

Opportunistic resources: ATLAS@Home

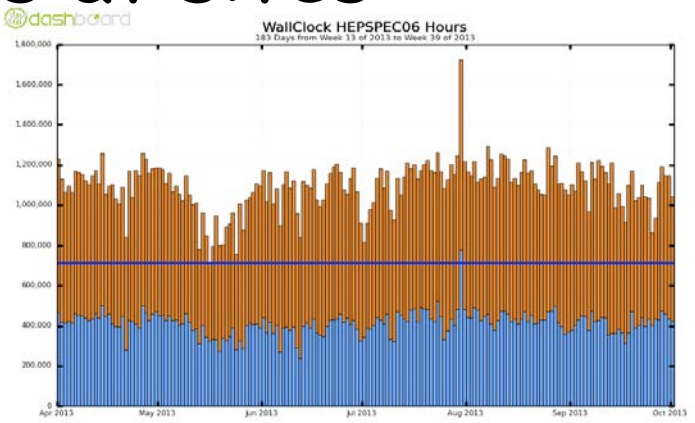
- Volunteer Computing using BOINC + ArC CE



20% Failure vs 5% on Grid

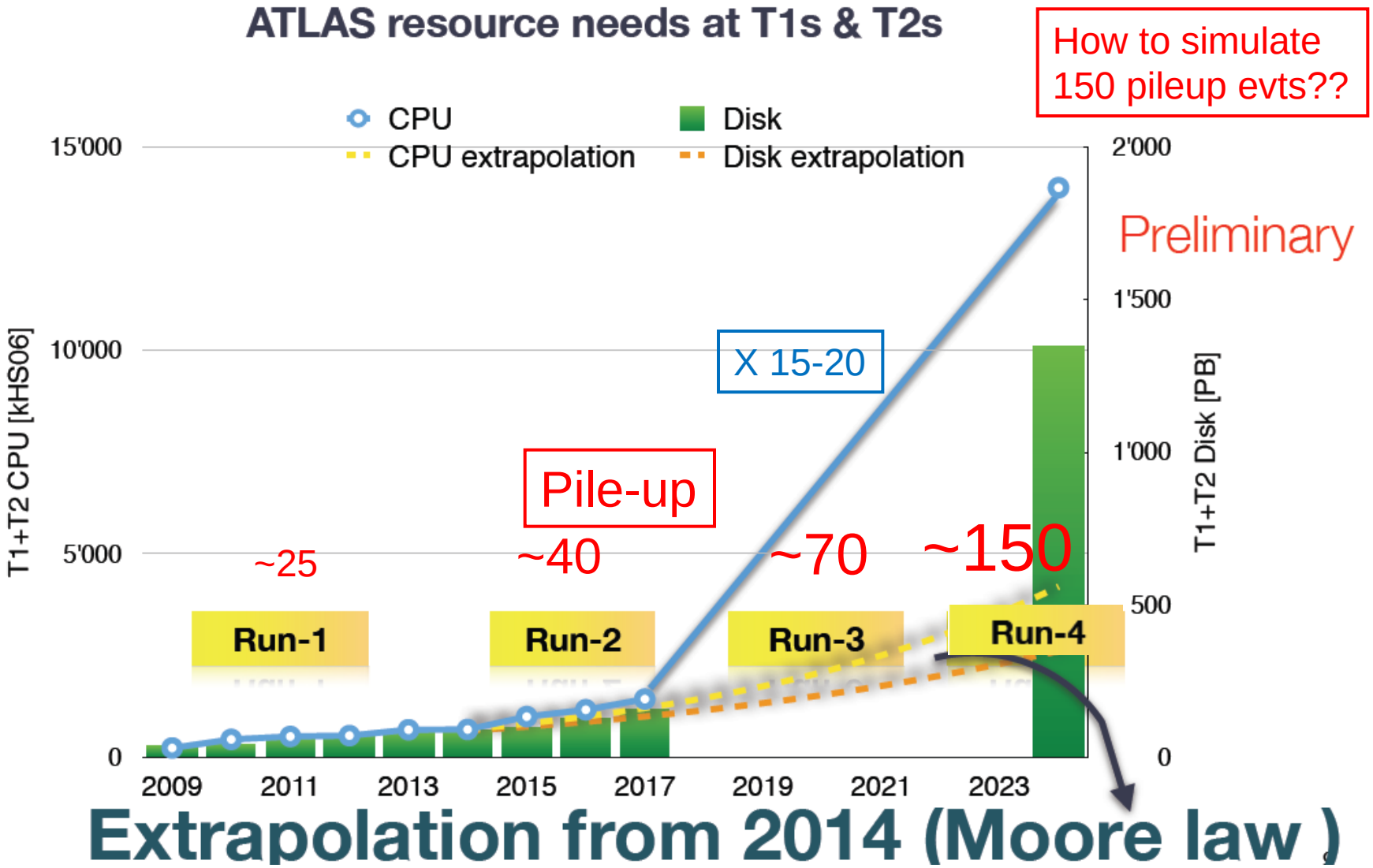
- Free!! (install Boinc + VirtualBox)
  - Solution for Institute desktop clusters
  - Can work at event level (Cf. Event Service

- Extra-unpledged resources at sites
  - eg T3s resources, opportunistic, as in Run1



CPU consumption above pledges both at T1s and T2s

Beyond Run2


ATLAS resource needs at T1s & T2s
How to simulate 150 pileup evts??
CPU
CPU extrapolation
Disk
Disk extrapolation
Preliminary
T1+T2 CPU [kHS06]
T1+T2 Disk [PB]
15'000
10'000
5'000
0
2'000
1'500
1'000
500
0
X 15-20
Pile-up
~25
~40
~70
~150
Run-1
Run-2
Run-3
Run-4
2009
2011
2013
2015
2017
2019
2021
2023
Extrapolation from 2014 (Moore law)

Preparing for Run3 – a multi-year software effort

- Re-engineer for future (or even current) computer architectures: vectorization, multi-threading,…
- Make Software Quality an integral part of software development and maintenance
- Define strategy for reconstruction and simulation of high pileup events
- Ensure that software packages on which we depend are compatibly re-engineered (Geant4, Root, Eigen, …)
- Collaboration and cooperation with external projects

Summary

- Run2 is an evolution of Run1
- Many ideas investigated for Run2
  - Cf. LHC Computing Model Update document
  - Role of Tiers, use of network, data federation, clouds, opportunistic resources
  - Big efforts by experiments to optimize & gain in resource (CPU, memory, storage)
  - All these ideas being tested at full scale
- Run3 & Run4
  - Needs & solutions already under study
- Manpower is an issue

FR-Cloud: Warmest thanks to

- Site admins
  - Interaction via LCG-TECH (Frédérique, Catherine)
- CAF members
  - Sites-ATLAS interaction, Eric
- Shifters
  - ADCoS (Edith Knoops, Claire Adam)
  - DAST (Laurent Duflot), AMOD
- French Cloud Squad support
  - Sabine Crépé-Renaudin, Emmanuel le Guirriec
- Support at T1
  - Essential role of Emmanouil Vamvakopoulos
  - His dedication & expertise has to be perpetuated