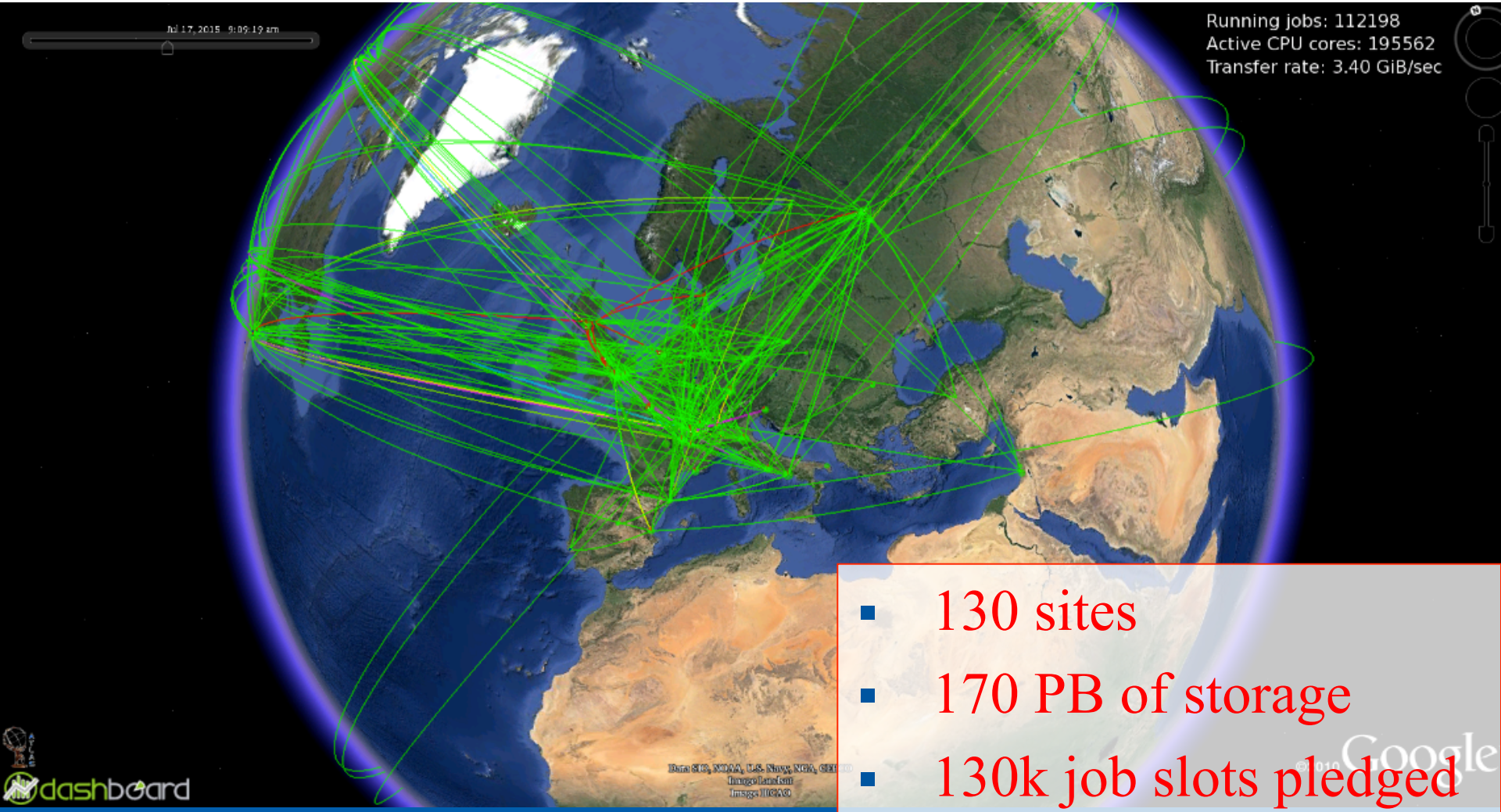
ATLAS Distributed Computing for LHC Run-2 and beyond

Alessandro Di Girolamo
(CERN IT - Support for Distributed Computing)
on behalf of the ATLAS collaboration

ATLAS Distributed Computing

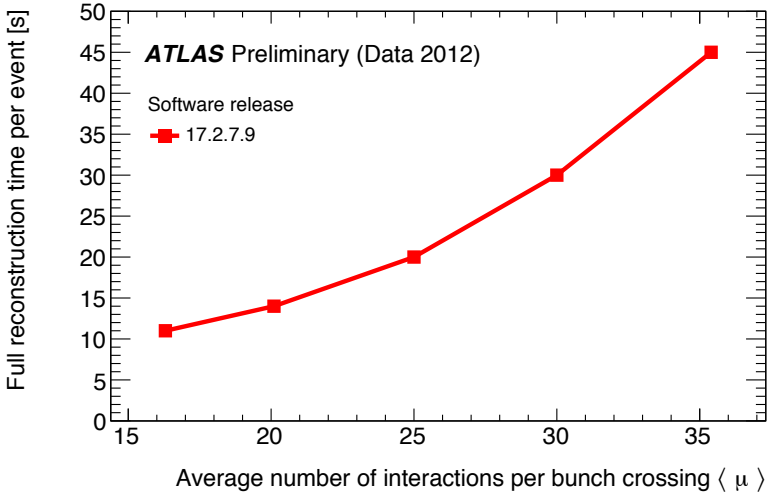


- 130 sites
- 170 PB of storage
- 130k job slots pledged
- 3000 users

LHC Run2: Computing Challenges

- Trigger rate 1kHz:
  - ~ 350Hz Run1
    - ~ 3 times more data
- Pile up μ ~ 40:
  - ~ 20 in Run1

Reconstruction Time vs μ with old Run1 SW release



- Flat budget for computing resources!
  - Relying on HW improvements
  - Exploit to the last bit what we have

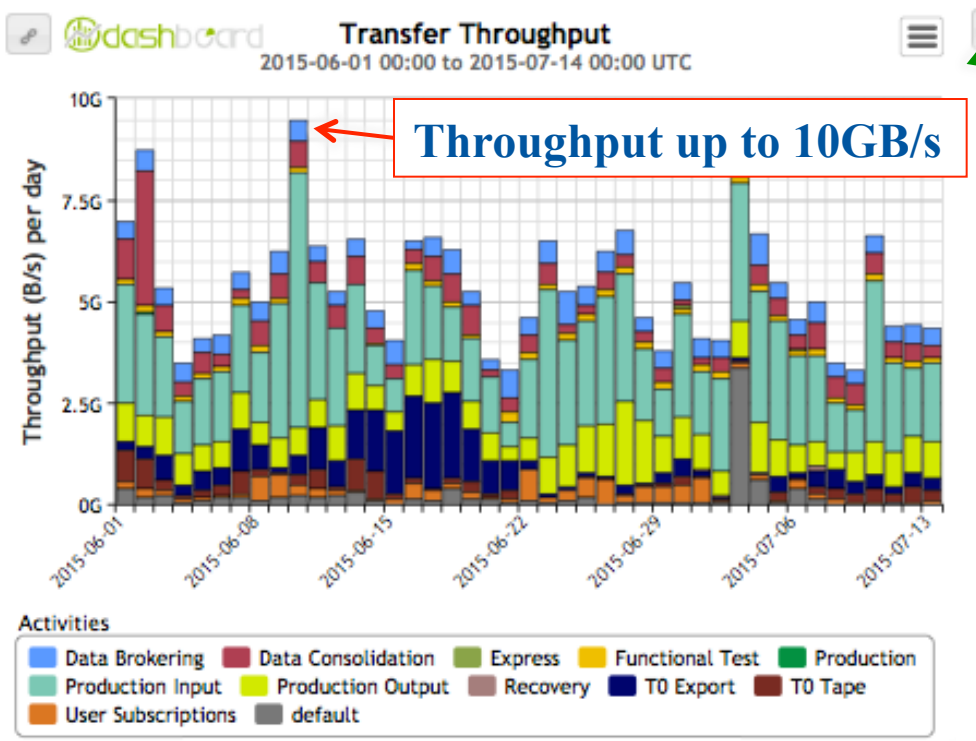
LHC Run2: Tackling the challenges

- (Re)Evolution of ATLAS Distributed Computing frameworks:
  - New Distributed Data Management (Rucio), Production System (ProdSys2) and Job Management System (JEDI)
- Completely new frameworks:
  - ATLAS EventIndex and ATLAS Event Service
- Leveraging all type of (opportunistic) resources
  - HPC, Volunteer Computing, Cloud resources
- Optimize space management: new Data Lifetime Policy
- More efficient utilization of resources
  - Understand/limit resource consumption (reconstruction speedup, memory sharing in multicore)
  - Workflows optimization: new Analysis Model and Derivation Framework

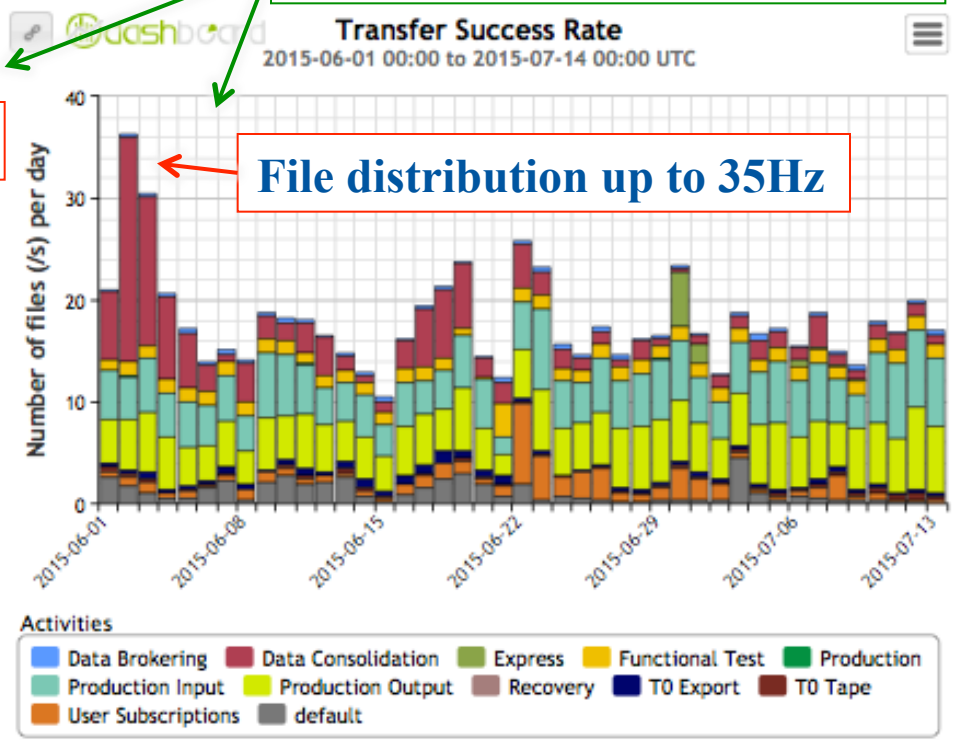
New Data Management: Rucio

- ATLAS DDM in charge to manage all ATLAS data:
  - 170 PB, >650M files
- Dataset and File catalog consolidated
- Protocols plugins: SRM, Webdav, gridftp, xrootd, S3…
- Smart automated data placement tools

No limitations faced up to now



Throughput up to 10GB/s

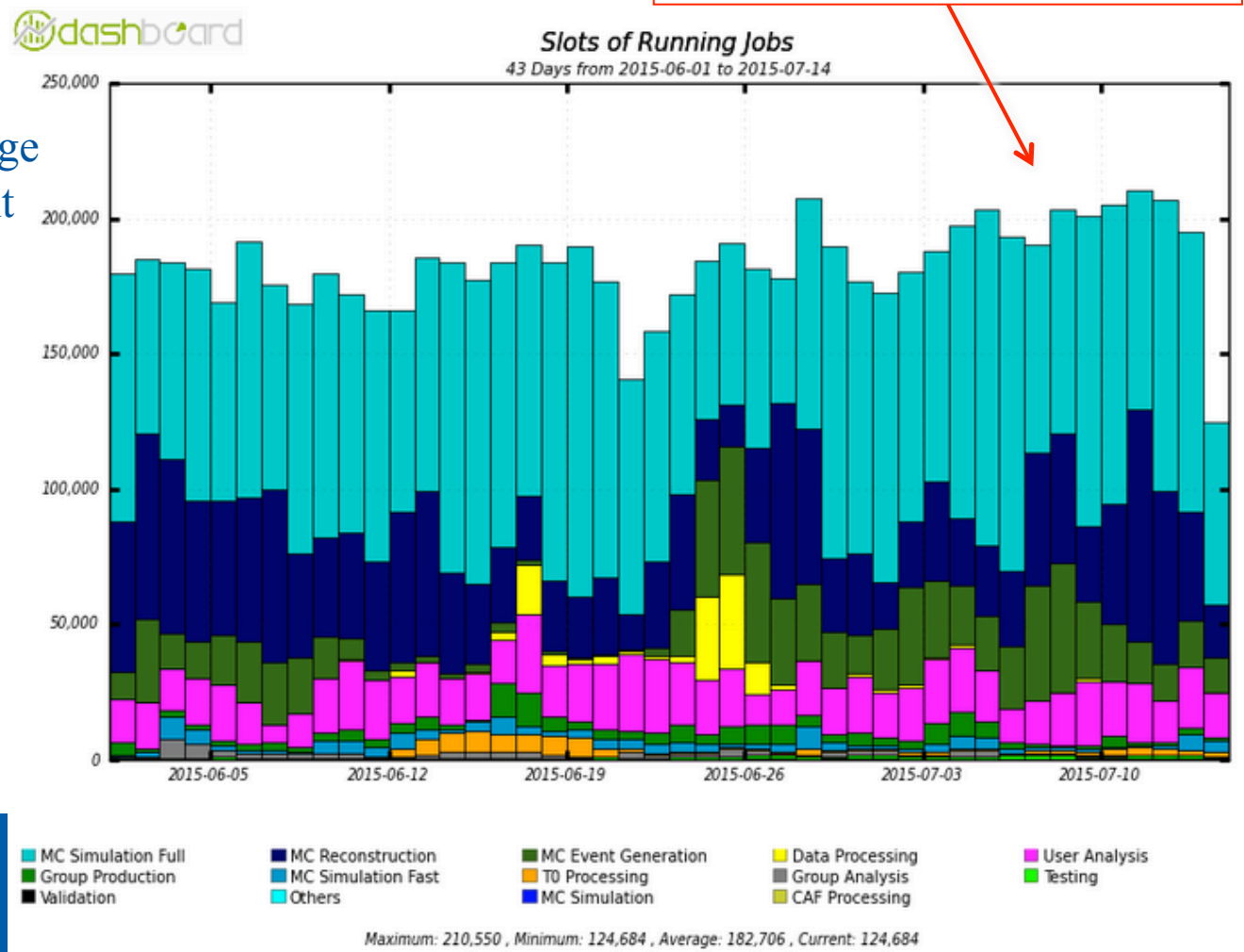


File distribution up to 35Hz

Workload Management (eco)System

- Production ANd Distributed Analysis system (Panda) is the core
- New components for Run2 (and beyond):
  - JEDI (Job Execution and Definition Interface): dynamic job definition, manage workload up to event level, automatic merging
  - DEFT (Database Engine for Task): requests and tasks management
- Integrated Network awareness to optimize brokering

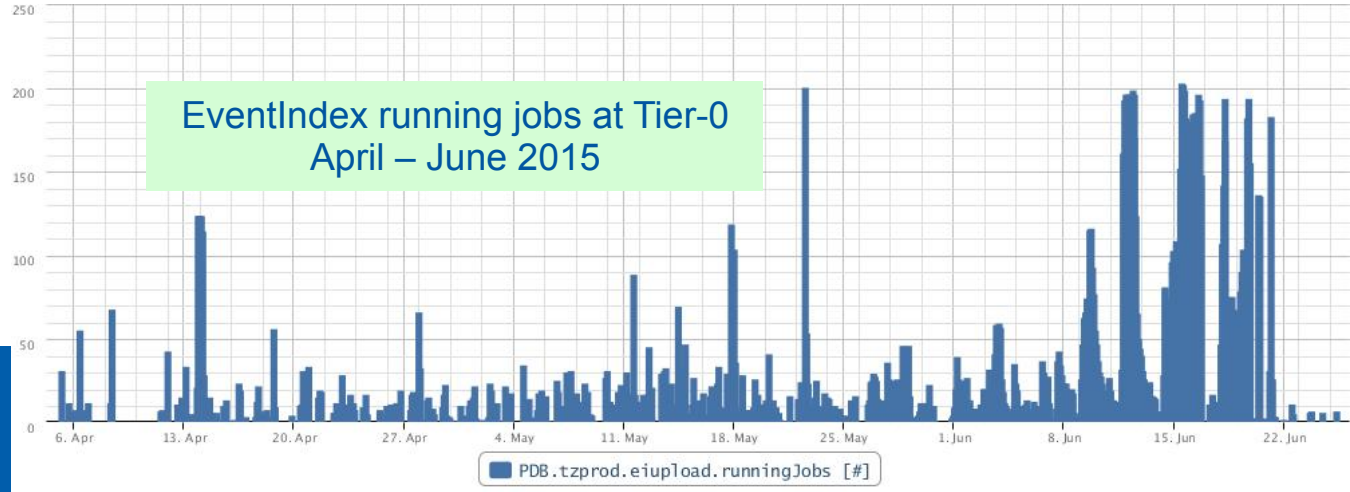
180-200k slots steadily used!



CERN ATLAS EXPERIMENT

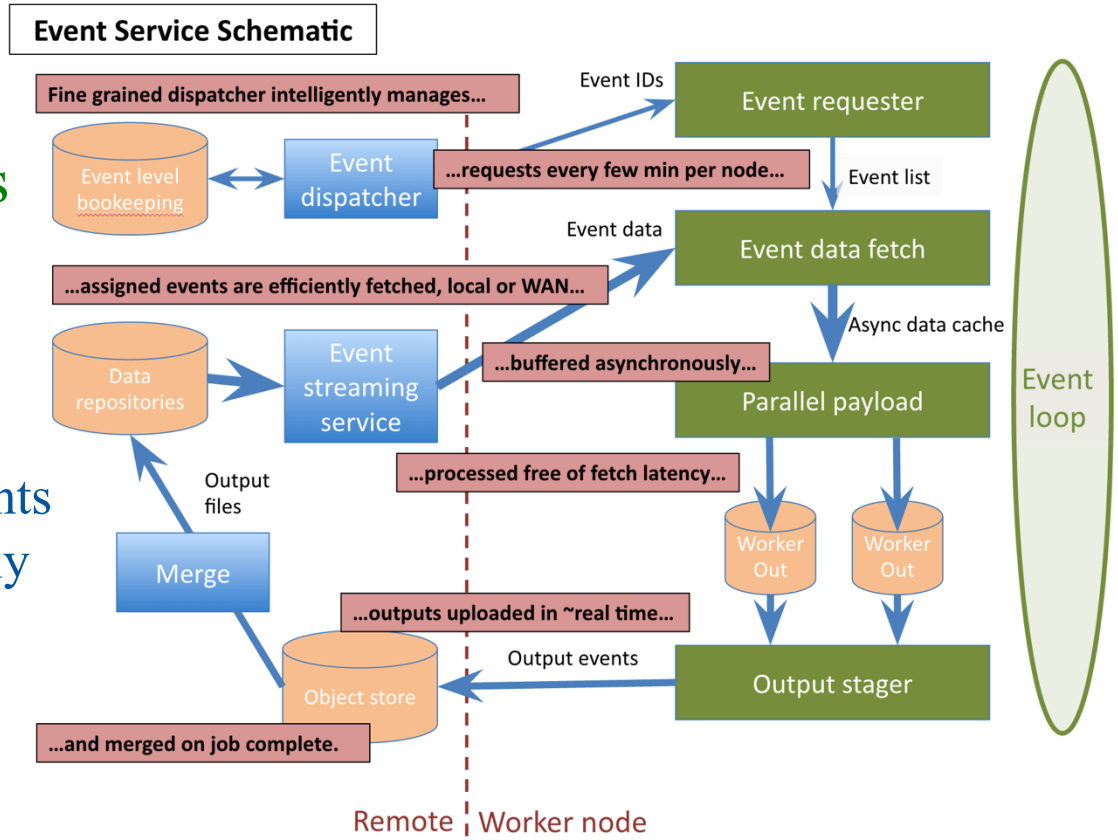
ATLAS EventIndex:
find the events you really want

- Global catalogue of ATLAS events
  - One record per event in each processing stage
    - Event identification (run and event number, luminosity block, data stream and type), trigger info, location pointers (unique identifiers of the file(s) plus internal pointer)
- Useful to e.g.:
  - Run sample selections on rare trigger combinations and retrieve those events
  - Retrieve one or a few events for event display or detailed comparisons of calibration or software versions
  - Minimize events overlaps in the various output of the DerivationsFramework
- Run2 data indexed automatically during Tier-0 reconstructions
- Run1 data filled in

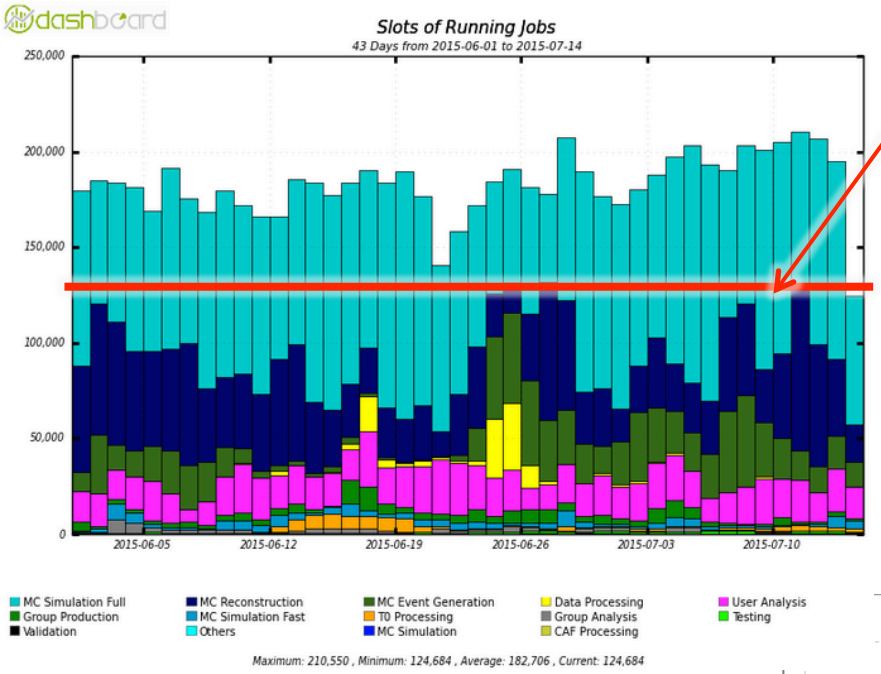


ATLAS Event Service:
exploiting the resources to the last drop

- Agile and efficient to exploit diverse, distributed, potentially short-lived resources:
  - HPCs (including preemptive mode), Spot market clouds, Volunteer computing (ATLAS@Home)
- Quasi-continuous event streaming to worker nodes
- Stream out output away quickly
  - Minimize the loss if WN vanish, minimal requirements on storage, outputs promptly available
- Commissioning and integration ongoing now



Leveraging all type of resources

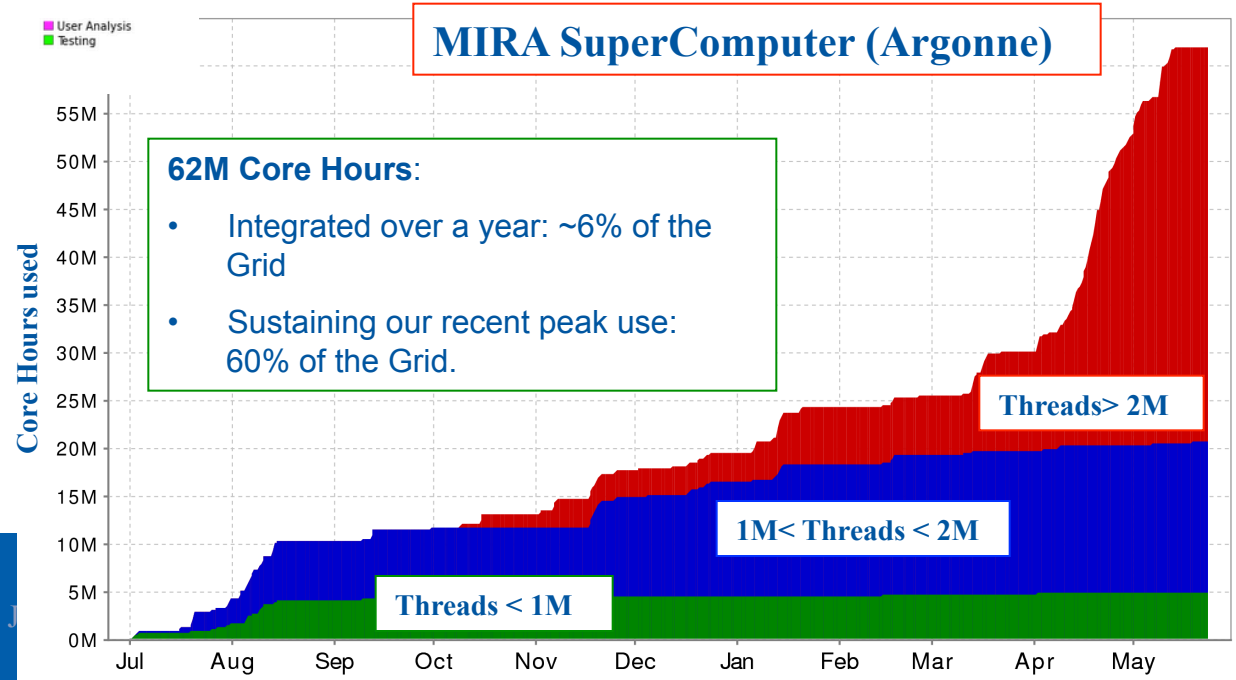


Pledges: 1.2M HS06 ≈ 130 K slots

- Over-pledge site resources
- Volunteer computing, Cloud resources, HPC
  - Require often dedicated effort
  - Potentially a huge amount of resources

HPC main challenges:

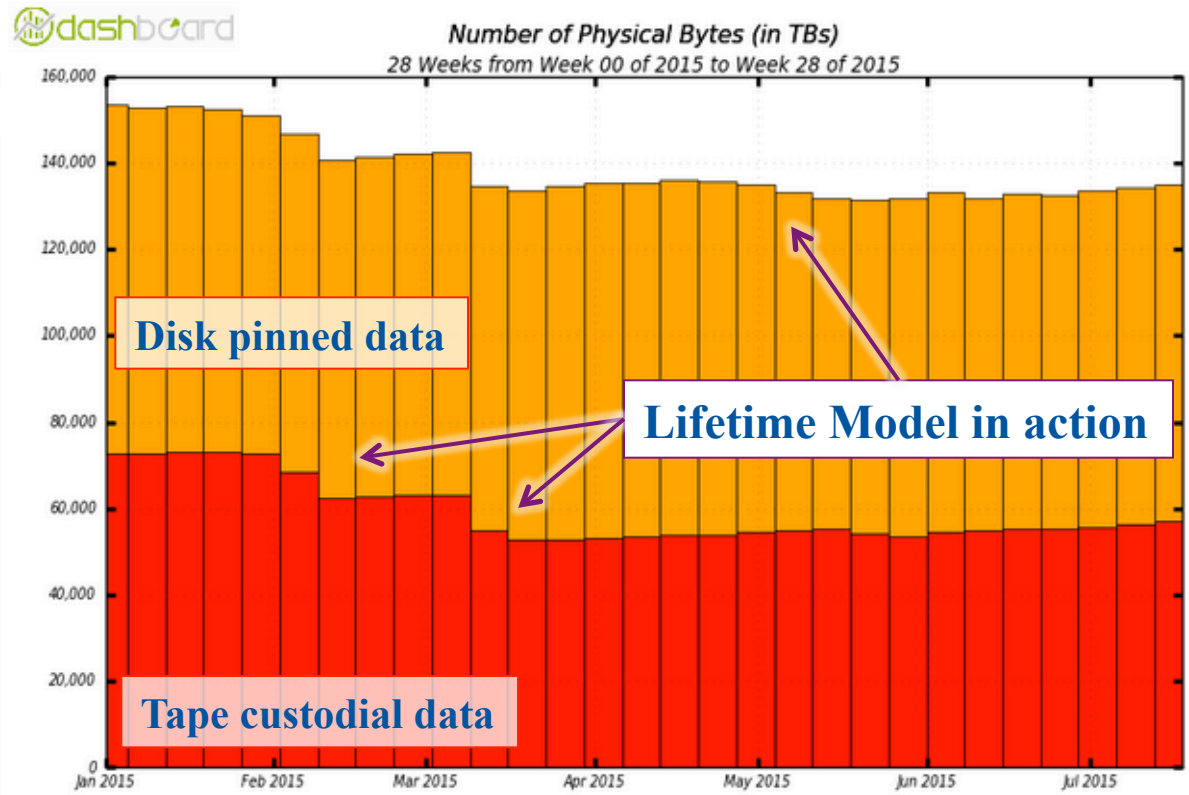
- Porting SW to scale to millions of threads requires substantial effort!
- each HPC is different: ad-hoc solutions and integration



Automatic space management: Data Lifetime policy

"easy" to get over-pledge CPU, much more difficult to get over-pledge storage

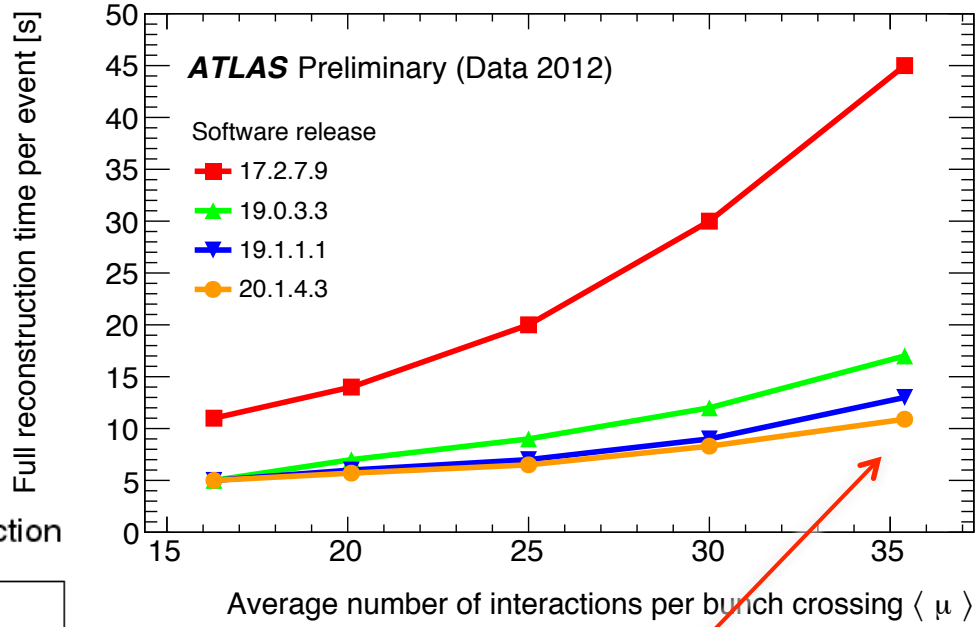
- Assign a Lifetime to all new and existing ATLAS data
  - Infinite for RAW
- Lifetime is extended if data are accessed
- At the end of the Lifetime datasets are deleted from tape and from disk (if space is needed)



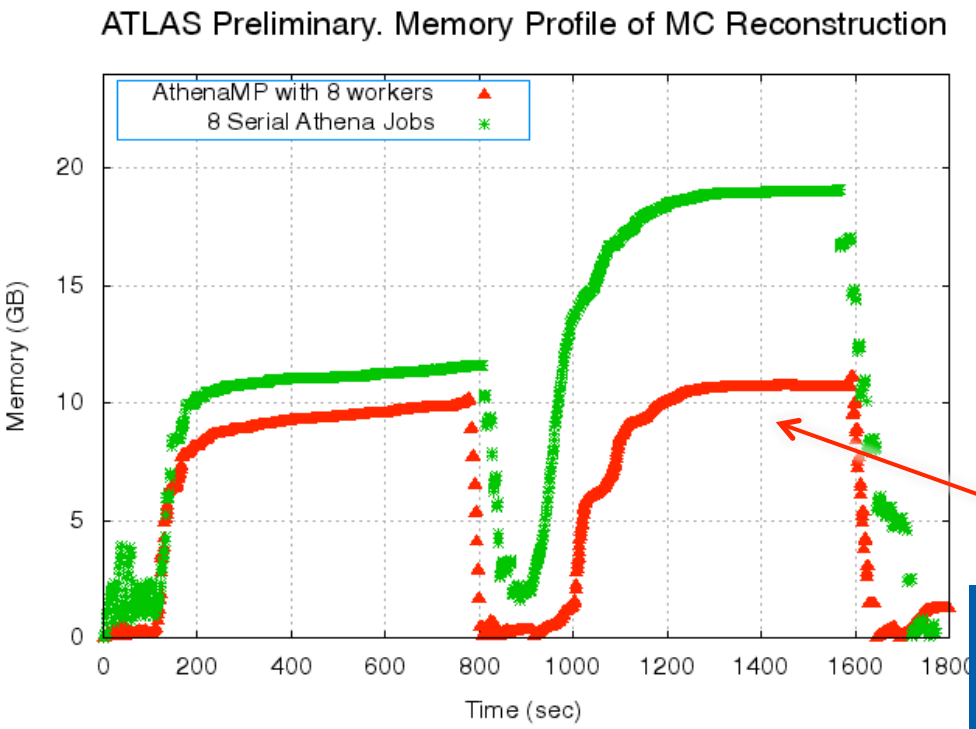
More efficient usage of resources

Software improvements during LHC shutdown, just 2 examples:

- Reconstruction time
- Memory sharing in Multicore jobs

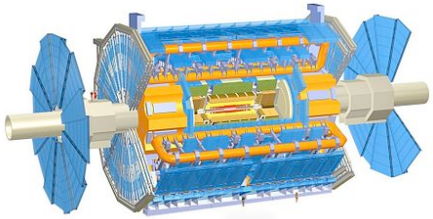


Factor 3 in reconstruction time



Memory sharing in MultiCore jobs

Data workflow: from detector to physicists



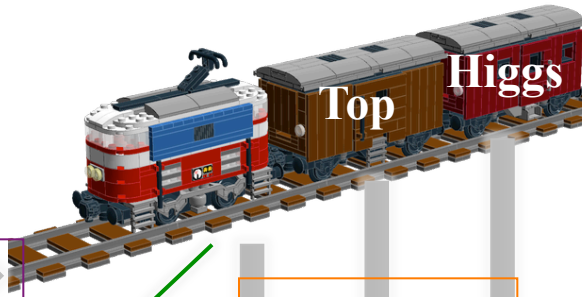
1-1.5 GB/s

Tier-0

Calibration loop and signoff: 24h

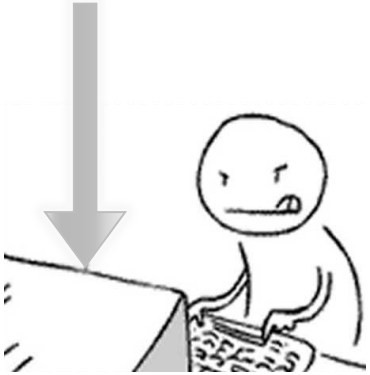
Bulk processing: ~24h

Exports and processing on the Grid

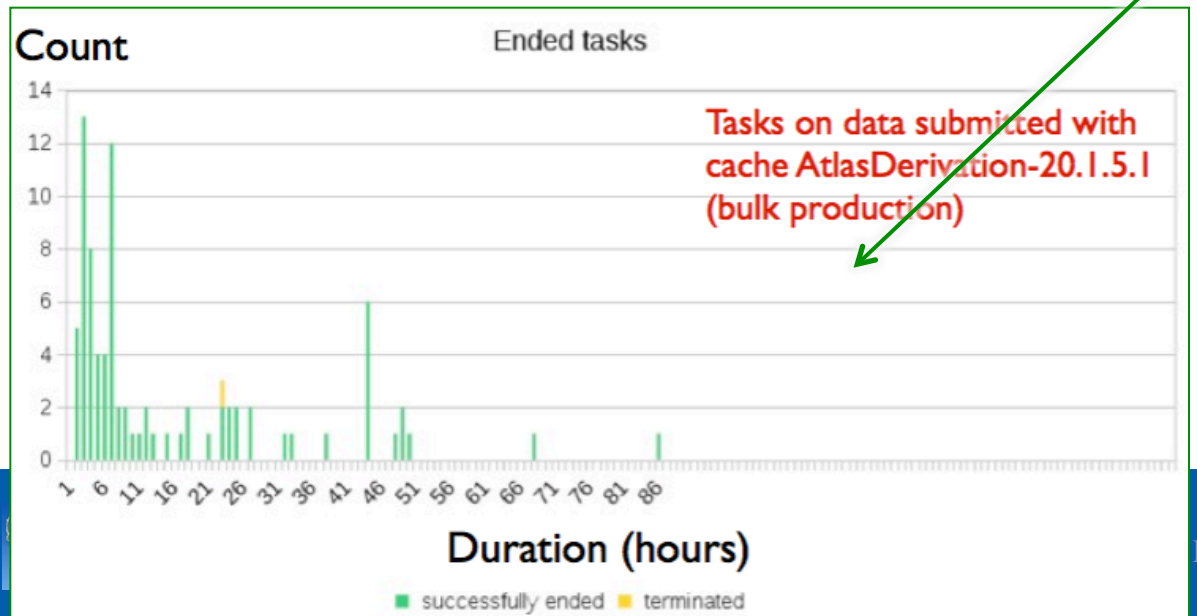


~ 24 hours

DAOD_ …



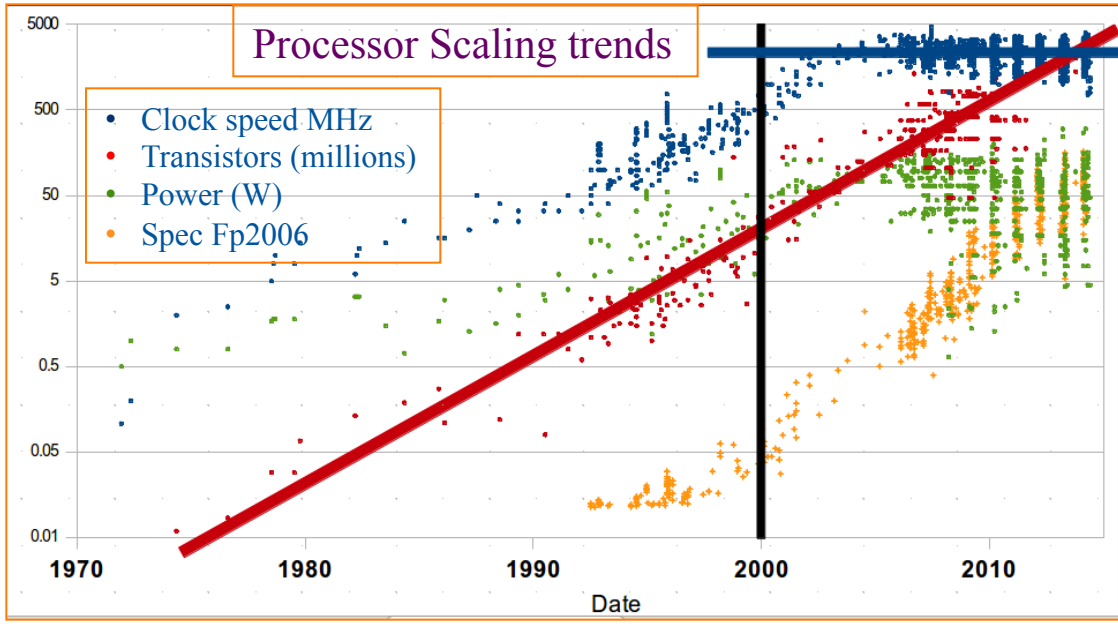
"Fresh" data delivered to analysis within ~3 days



Beyond LHC Run2

(some of) the facts:

- ! Negligible increase in clock speed
- ! O(10) growth for both storage and computing
- ! Complex new architectures (multicore and co-processor)
- ! Affordable memory per core decreasing



- New Framework: MultiThreaded based on GaudiHive and MultiThreaded G4
- Leverage (even more) all available resources
- Access data more intelligently and efficiently with less replication:
  - Use our increasingly powerful networks

Find common solutions exploiting collaborations between all the experiments!

Summary

- ATLAS Distributed Computing has considerably evolved and improved during the LHC Run1 shutdown
  - ✓ Solid frameworks with possibility for new features and improvements
  - ✓ Flexible to exploit all available resources

- ✓ Smooth first month(s) of LHC Run2 data taking
  - ➢ Quick in providing data to physicists