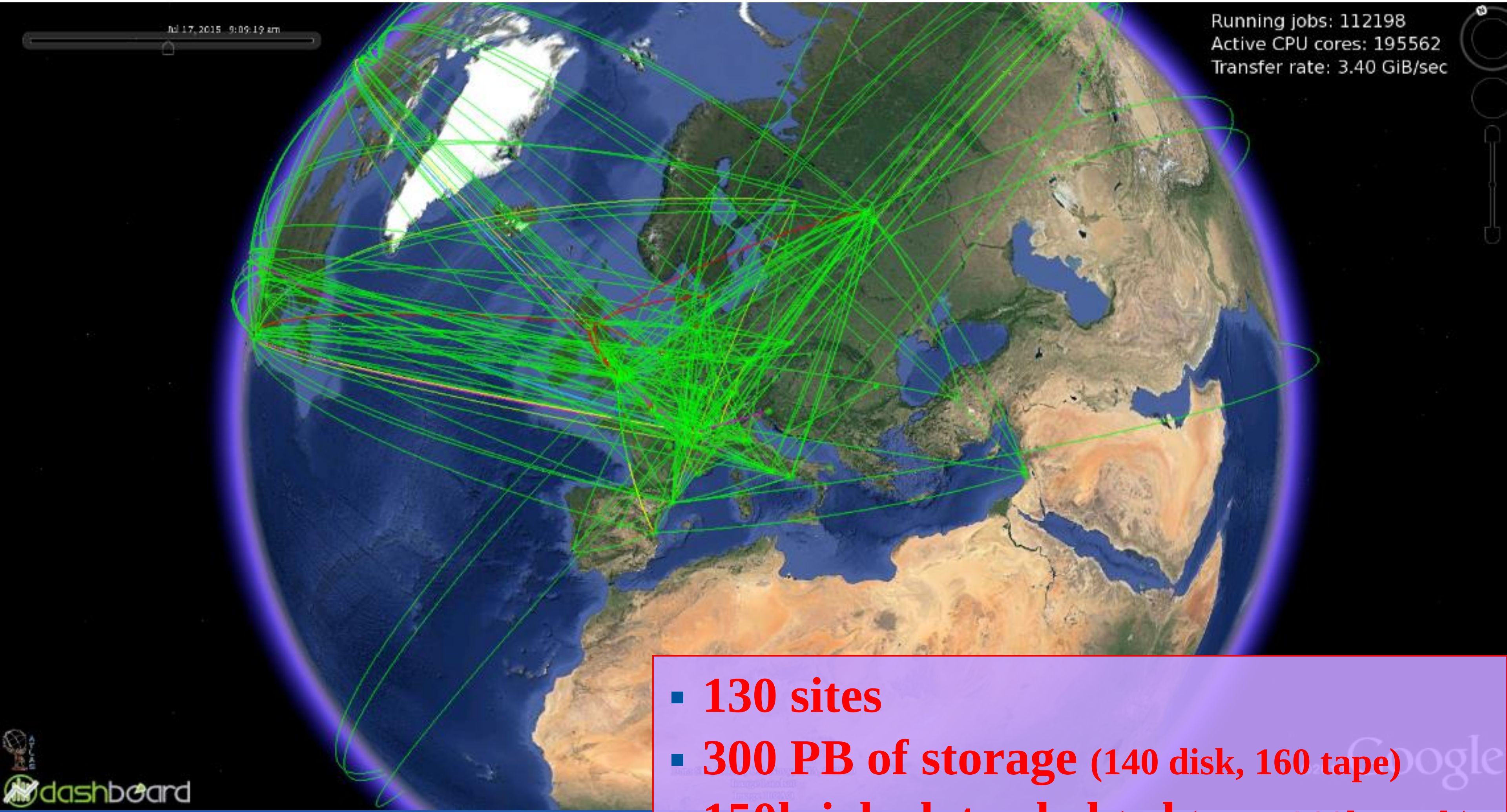
The evolution of ATLAS Distributed Computing

Alessandro Di Girolamo *CERN-IT*

On behalf of ATLAS Distributed Computing

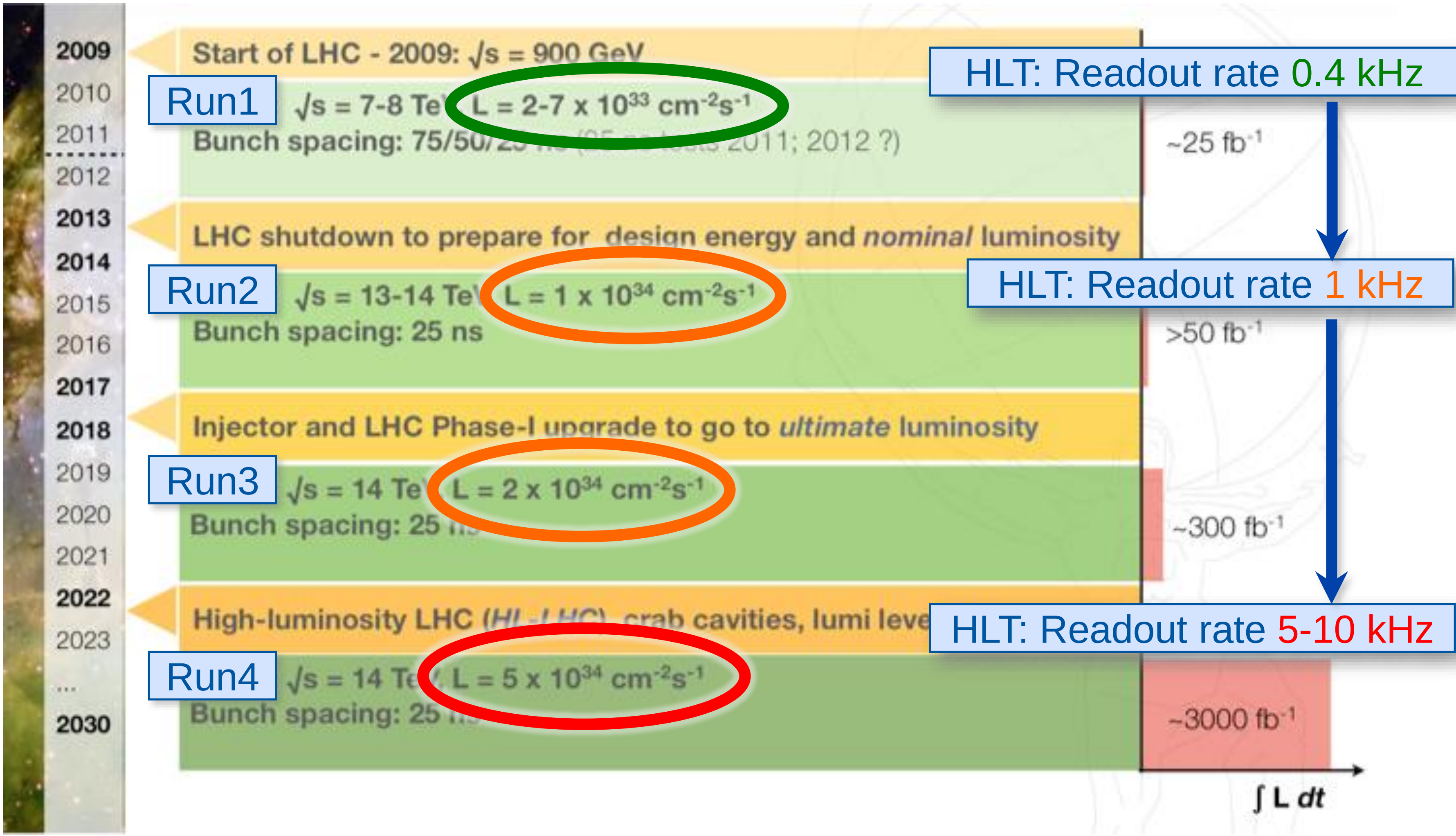

ATLASDistributedComputing
Sites
Security
Trust
Kibana
StorageConsistency
WALLCLOCK
Panda
HTCondor
Batch
Rucio
dCache
WLCG
DERIVATIONFRAMEWORK
MADGRAPH
HPC
ADCoS
EVNT
ARCCE
Deft
Efficiency
WAN
FTS
CVMFS
Condor
EventIndex
RAW
Dataset
DAST
DPM
ARC
AMI
SWRelease
Xrootd
HADOOP
HTTP
Users
Federation
Pilot
JEDI
FAX
Rules
AOD
LifetimeModel
OBJECTSTORE
EventService
LostHeartbeat
Yoda
HIST
Analytics
Reprocessing
CLOUDSQUAD
EGI
BigData
APF
RSG
PSS
CloudProvider
HAMMERCLOUD
PQ
DAOD
ANALYSIS
RSS
Memory
DashBoard
MonteCarlo

ATLAS Distributed Computing



- **130 sites**
- **300 PB of storage (140 disk, 160 tape)**
- **150k job slots pledged (up to 300k used!)**
- **3000 users**

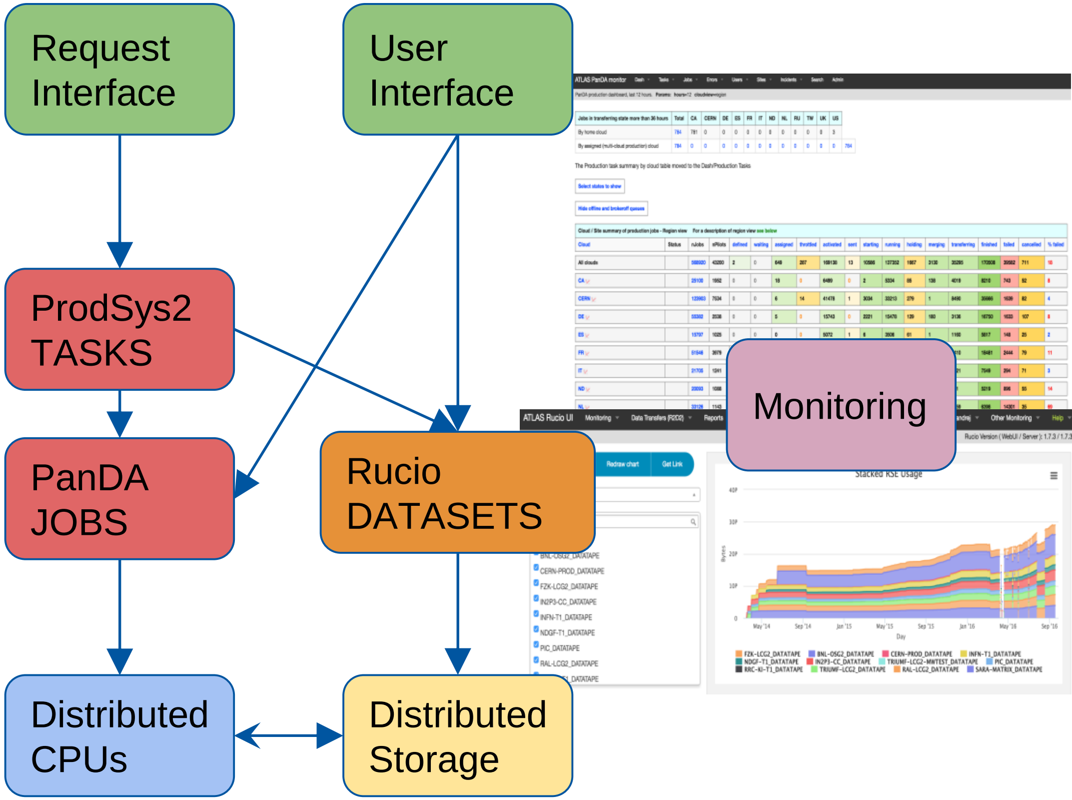
LHC: from Run1 to Run4



Preparing Run2: new Production and Data Management

Workflow Management System

- PanDA/JEDI
  - Dynamic resources, jobs
  - Analysis and production use the same infrastructure
- ProdSys2
  - Workflow organization relies on input transformation
  - Any kind of workflow is quickly implemented



Rucio

- Optimized and scalable data management
- Transfer latencies are minimized

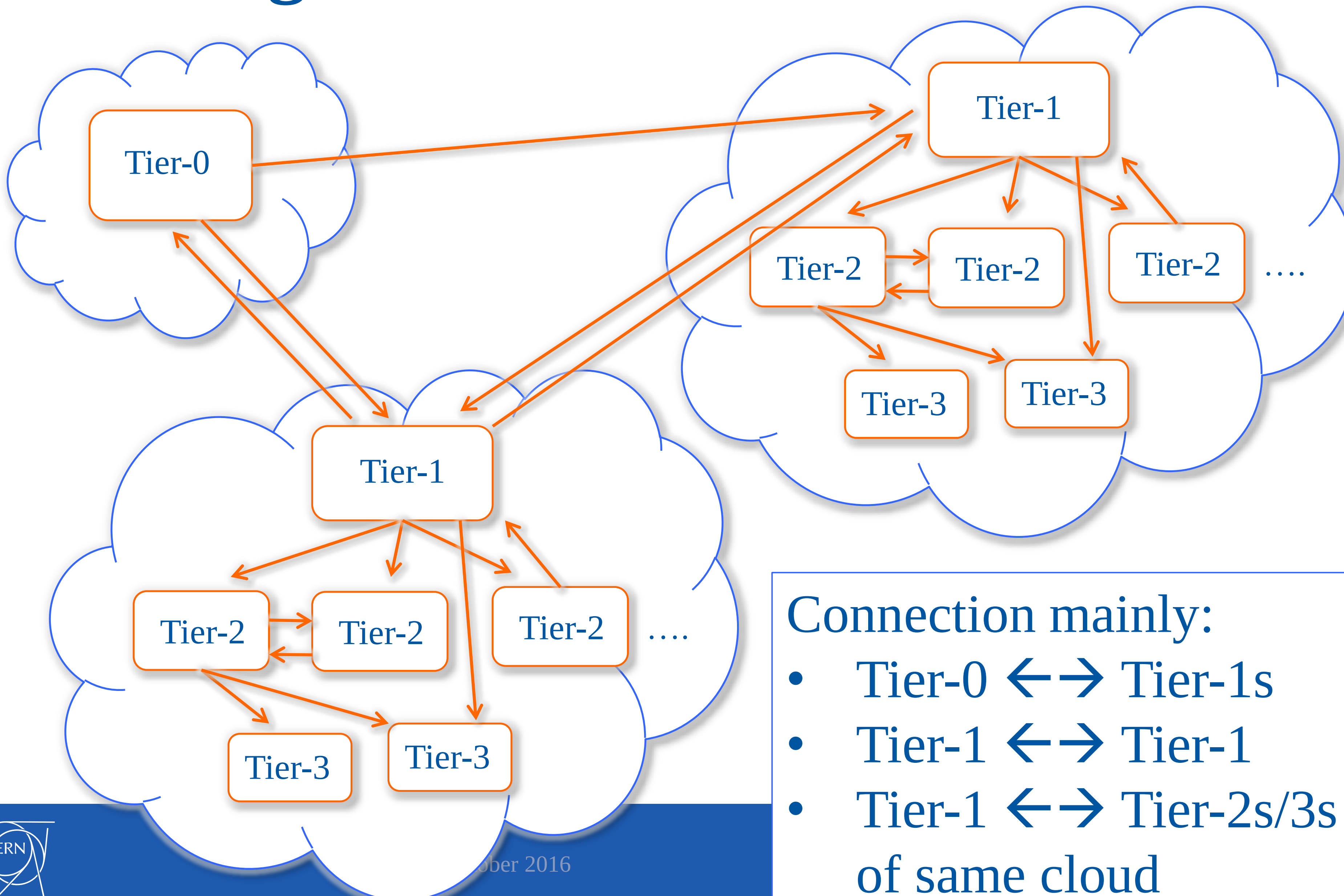
... More changes for Run2 (and during it)

- Many changes/renovation/rethinking/build-from-scratch. Just few examples here:
- Auto-tuning of jobs:
  - Jobs memory and walltime measured for first 10 (scout) jobs of a task and set for the rest
  - Retries of failed jobs have increased memory or walltime if that was the reason for failure
- Task completion
  - Requests and tasks are monitored for progress: almost completed tasks or tasks with a close deadline are auto boosted to complete the remaining jobs
- From *Clouds* to WORLD: MONARC model is gone!
  - Every reliable site can store single replica (primary) data → Nucleus
  - Every site well connected to nucleus can process data: → Satellite
  - Associations are dynamic at the task and job brokering level

ATLAS *Clouds*

- ATLAS *Clouds*
  - ≠ Cloud resources (AWS, Google Compute, Rackspace)
  - Logical grouping of sites:
    - one Tier1 plus several Tier2s and Tier3s
    - Mostly belonging to the same country/funding agency
  - Support provided by *Cloud* Squads
    - close to each site, often same language
- Historical concept
  - <u>Useful in the past</u>: networking limitations
  - <u>Still useful </u>especially for the support model

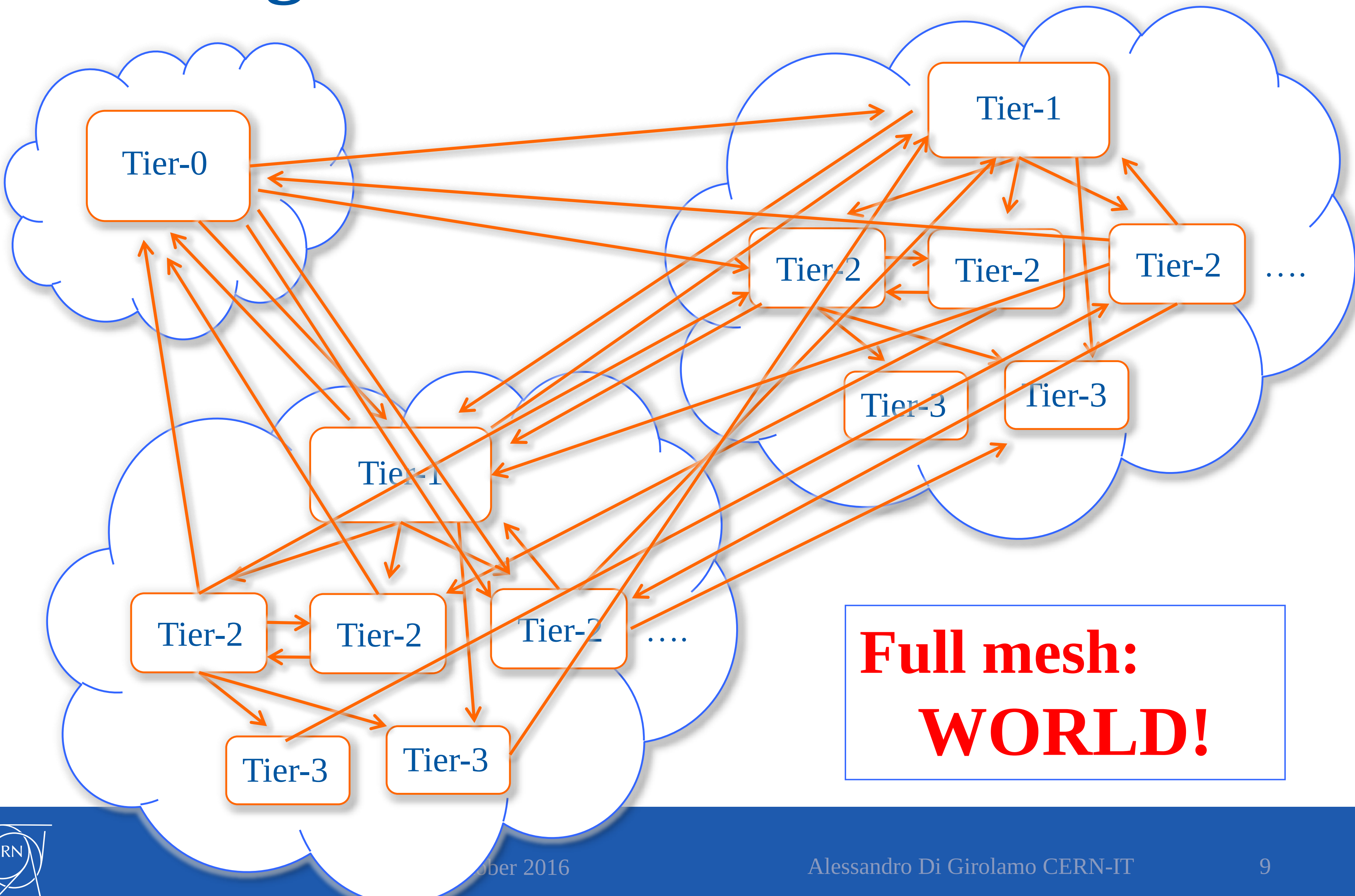
Breaking the Clouds boundaries: before



Connection mainly:

- Tier-0 ←→ Tier-1s
- Tier-1 ←→ Tier-1
- Tier-1 ←→ Tier-2s/3s of same cloud

Breaking the Clouds boundaries: now!



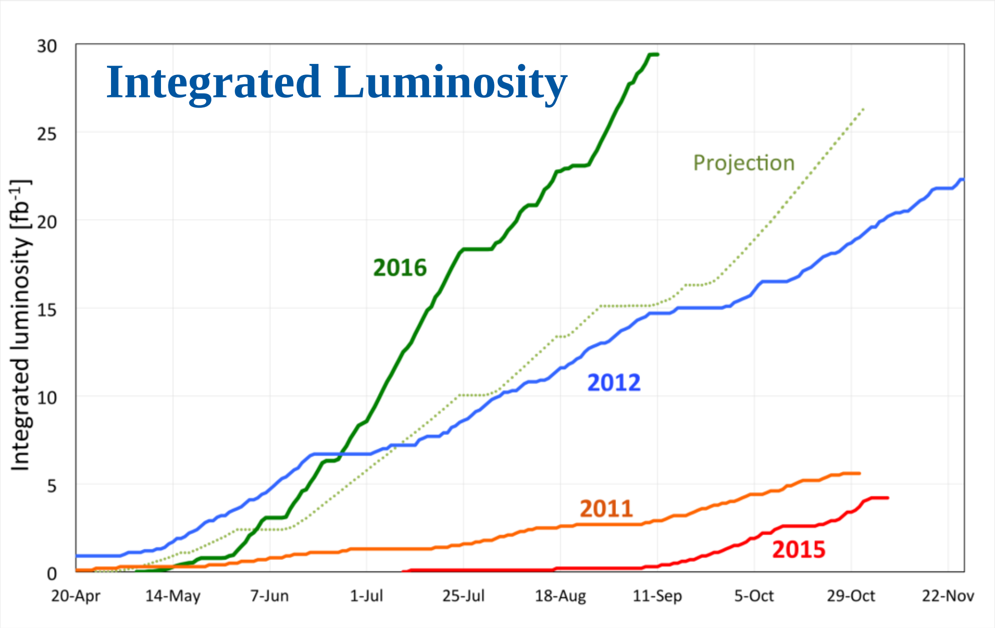
Full mesh: WORLD!

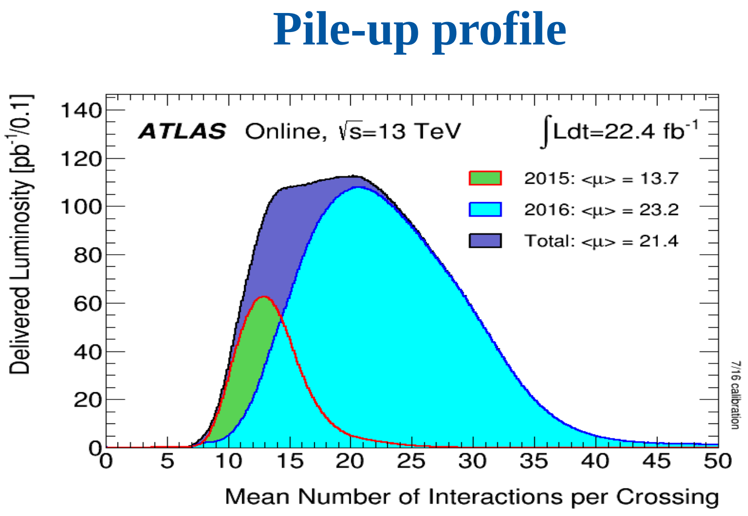
How are things going?

Many changes ...

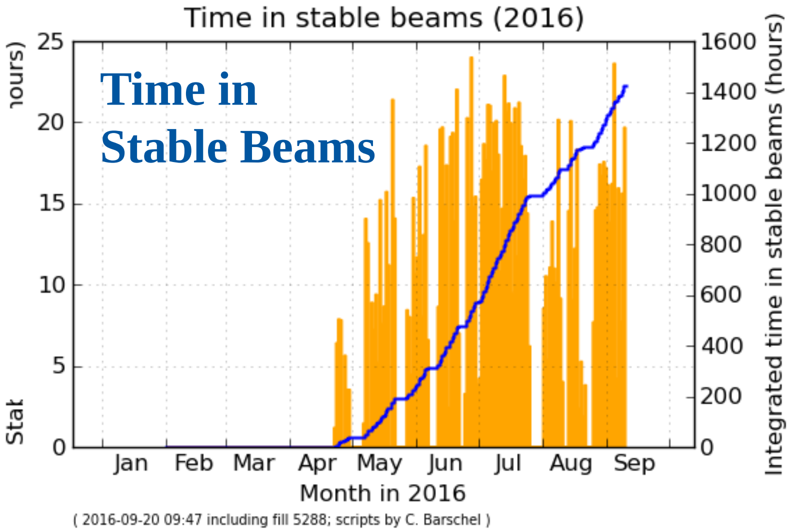
focus now on Data Taking,
Distributed Processing and
Distributed Data Management

LHC Run2 experience





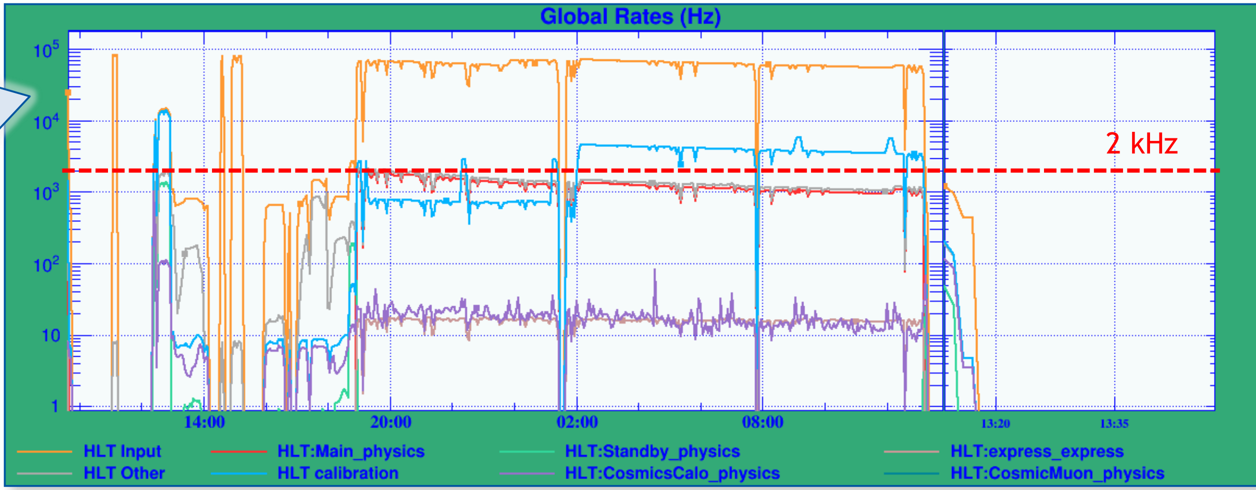
- Integrated luminosity: ~50% more than expected!
- Up to 80% duty cycle!

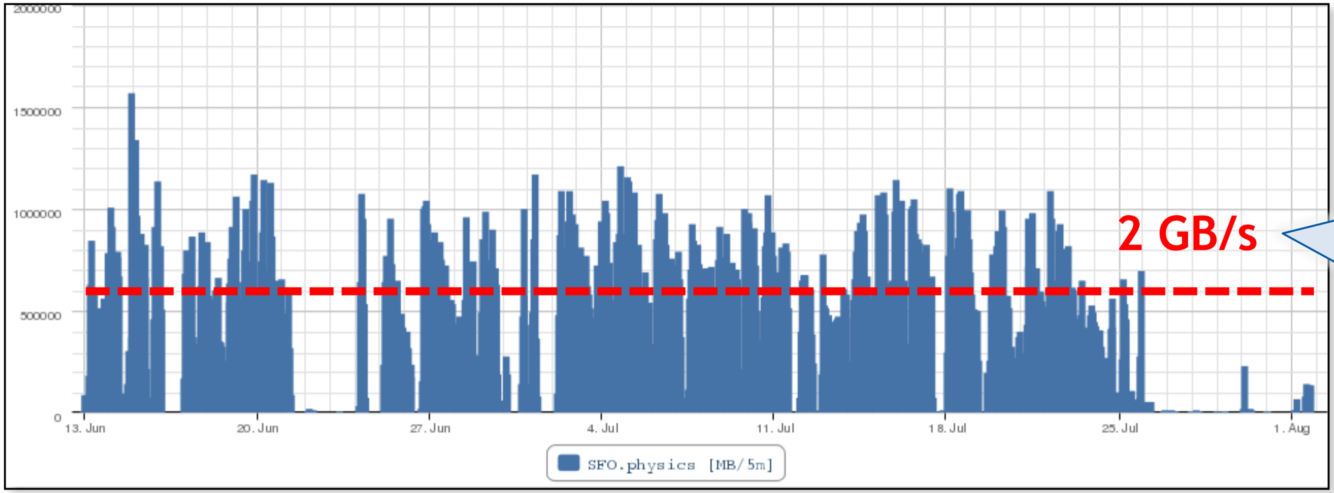


- Computing resources stretched to the max to cope with the impressive LHC performance
  - Thanks to the sites and to the framework renovation we did during LS1

From ATLAS detector to the Tier-0

HLT global rates
(run 302054, Jun 15):
typical profile of first weeks
in June, later adjusted

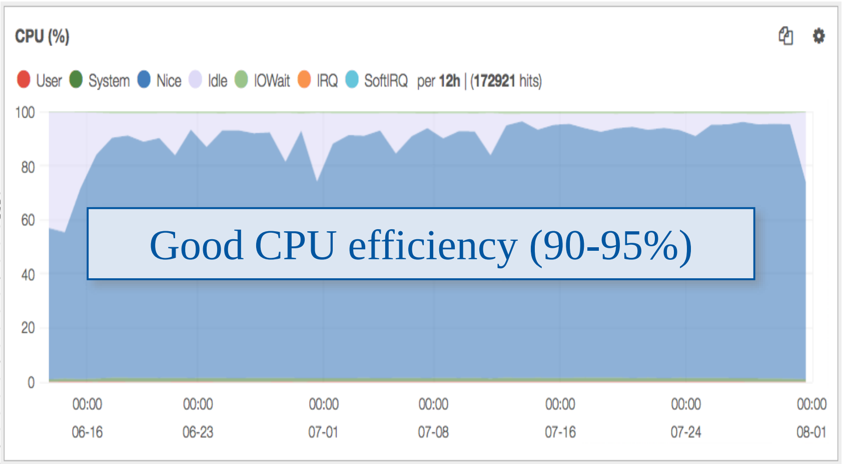


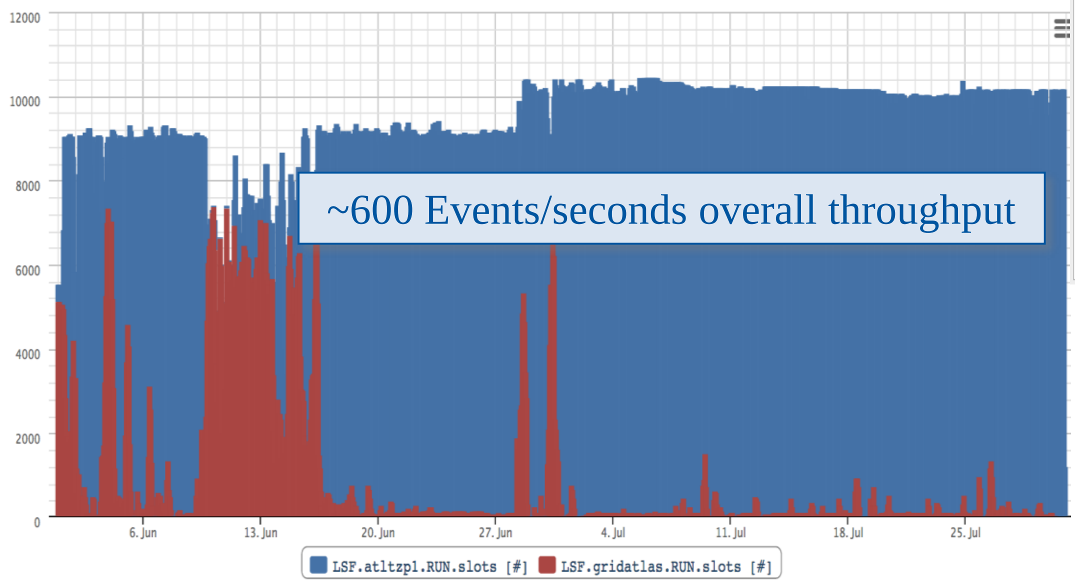


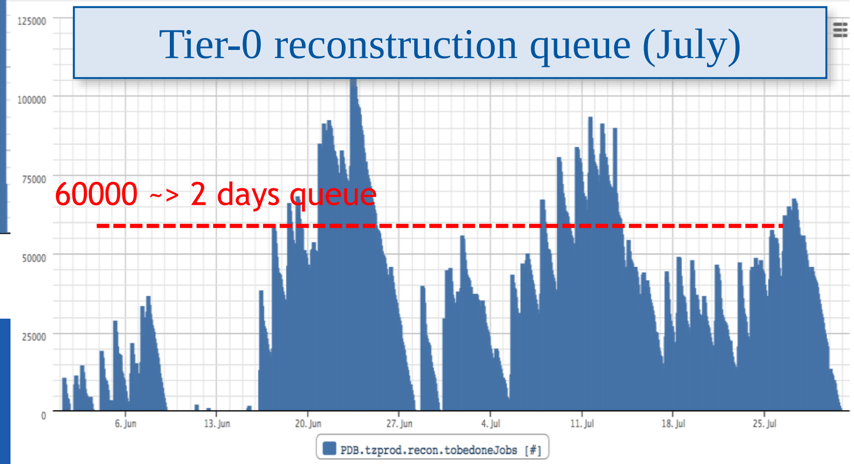
ATLAS SFO → Tier-0
transfers
(Jun 13 − Aug 2)

Tier-0 processing

- Data taking is pushing the infrastructure to the limits
- powerful WNs: 10k cores, SSD w 4GB/core
- Grid jobs overspilling when Tier-0 not running



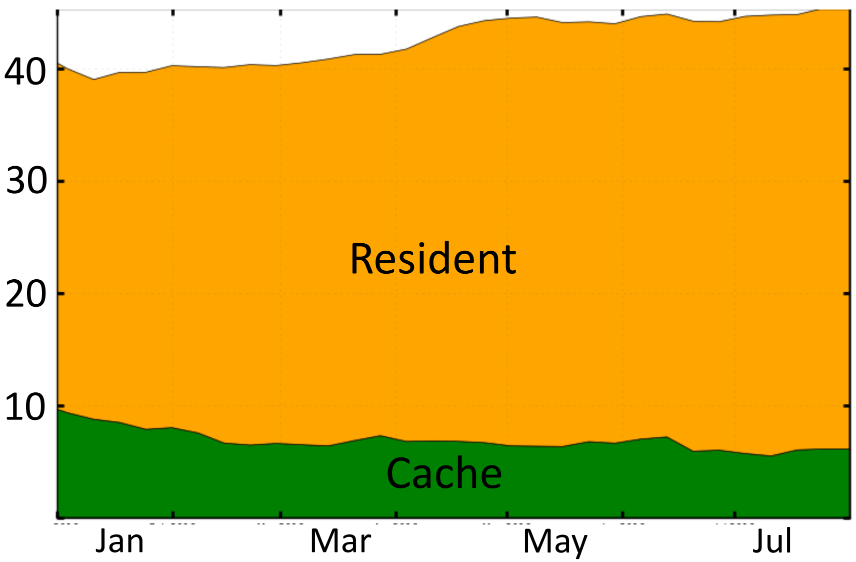




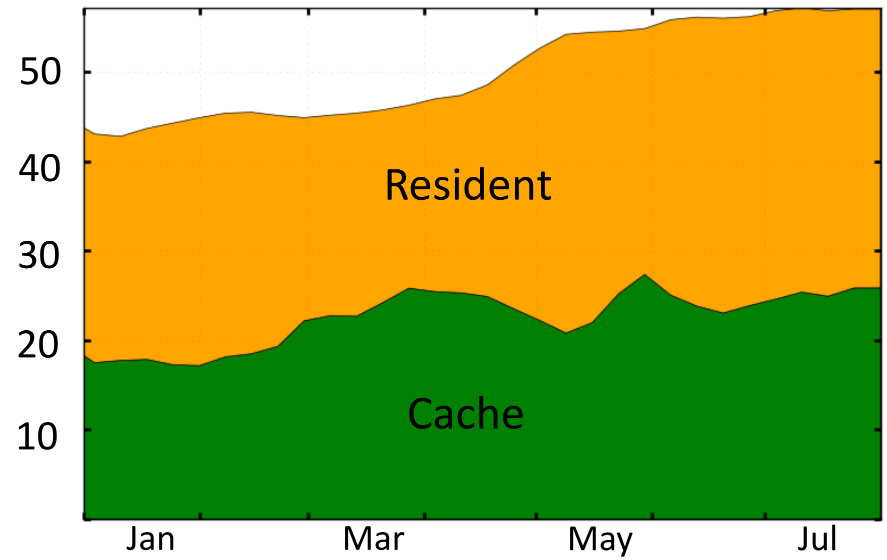
Distributed Data Management

- 300PB between disk and tape
  - 1B files, 100k datasets
- **Primary (resident) data** is partially **replicated** (cache)

Resident vs Cache data at T1s (PB)

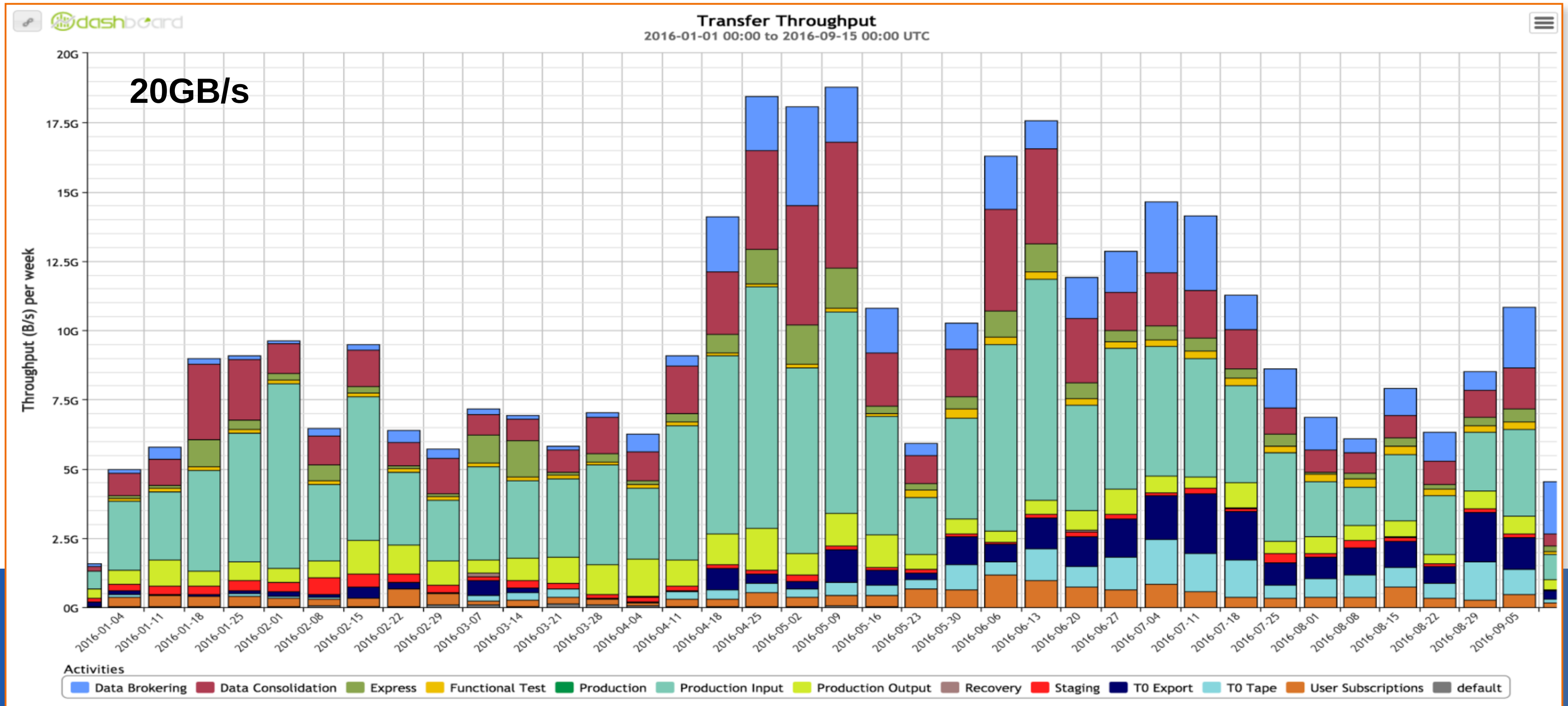


Resident vs Cache data at T2s (PB)

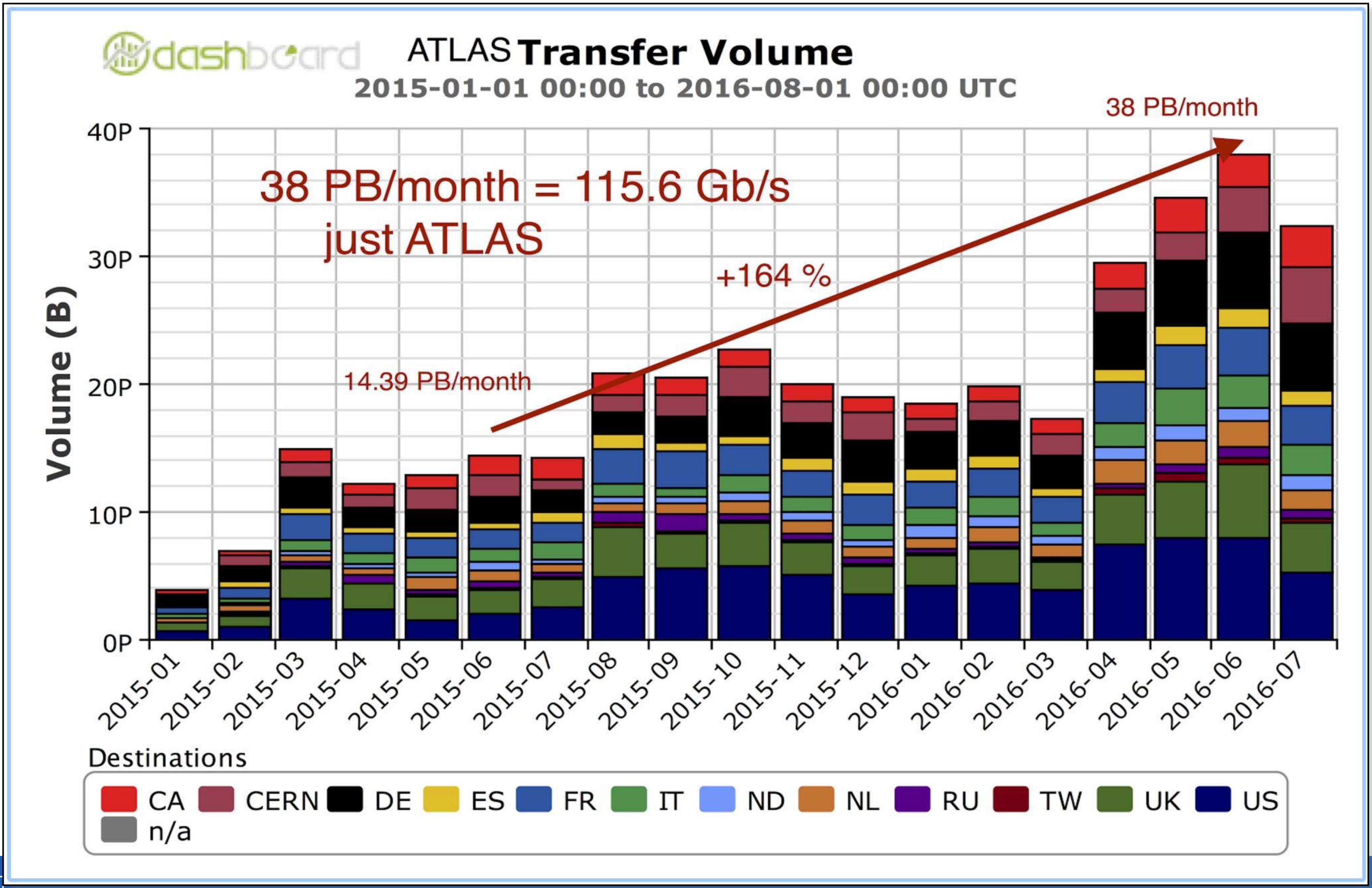


Distributed Data Management

- Data transfers peaks at **20 GB/s** weekly
  - with days at 40+ GB/s
- More than 50 files/s
- Largest activity - **input transfers**

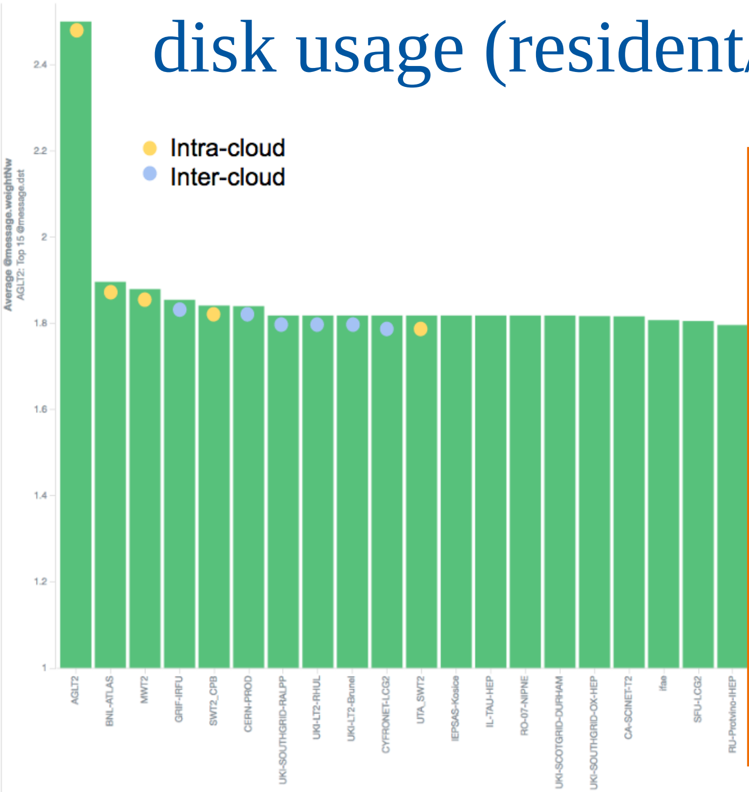


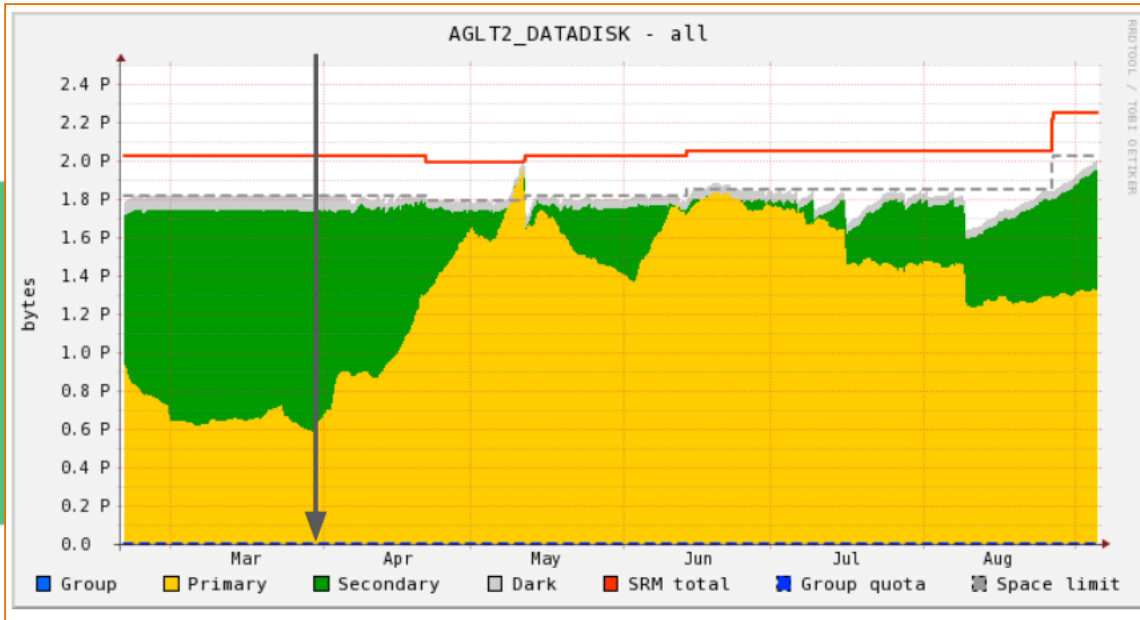
Throughput evolution



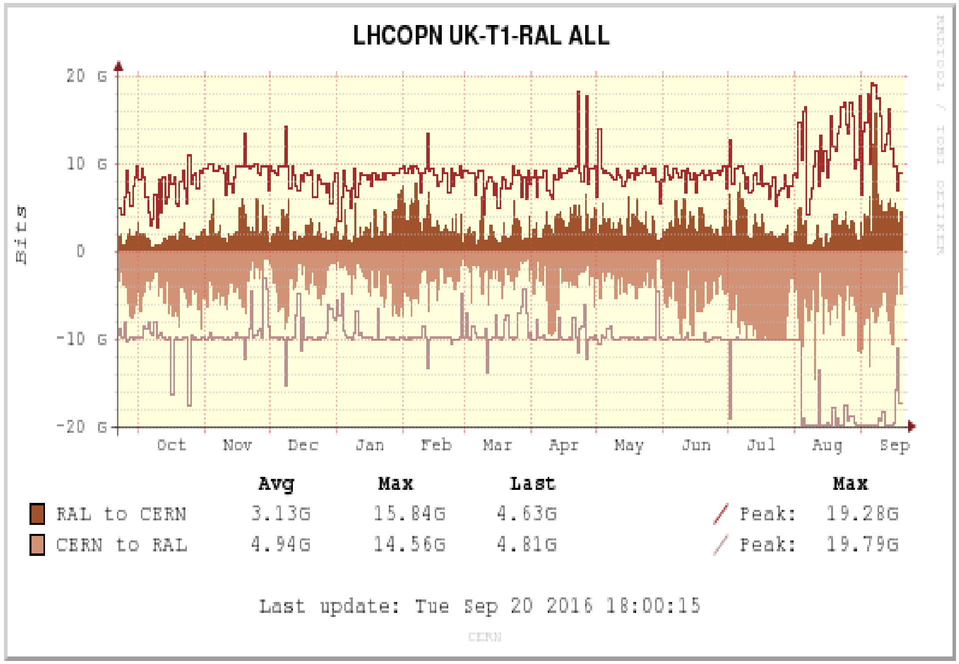
Network-aware brokering

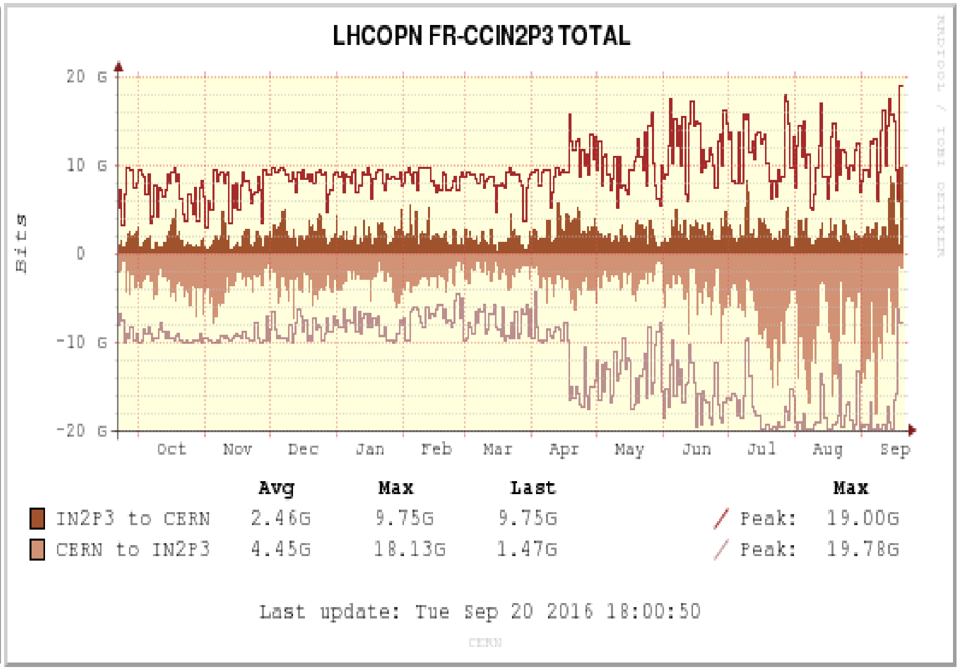
- WORLD was fully activated end March 2016
- Nuclei being added progressively
  - Currently T1s and ~20 (out of 80) T2s
- Task output to Nuclei T2s: positive impact on the overall disk usage (resident/cache ratio)





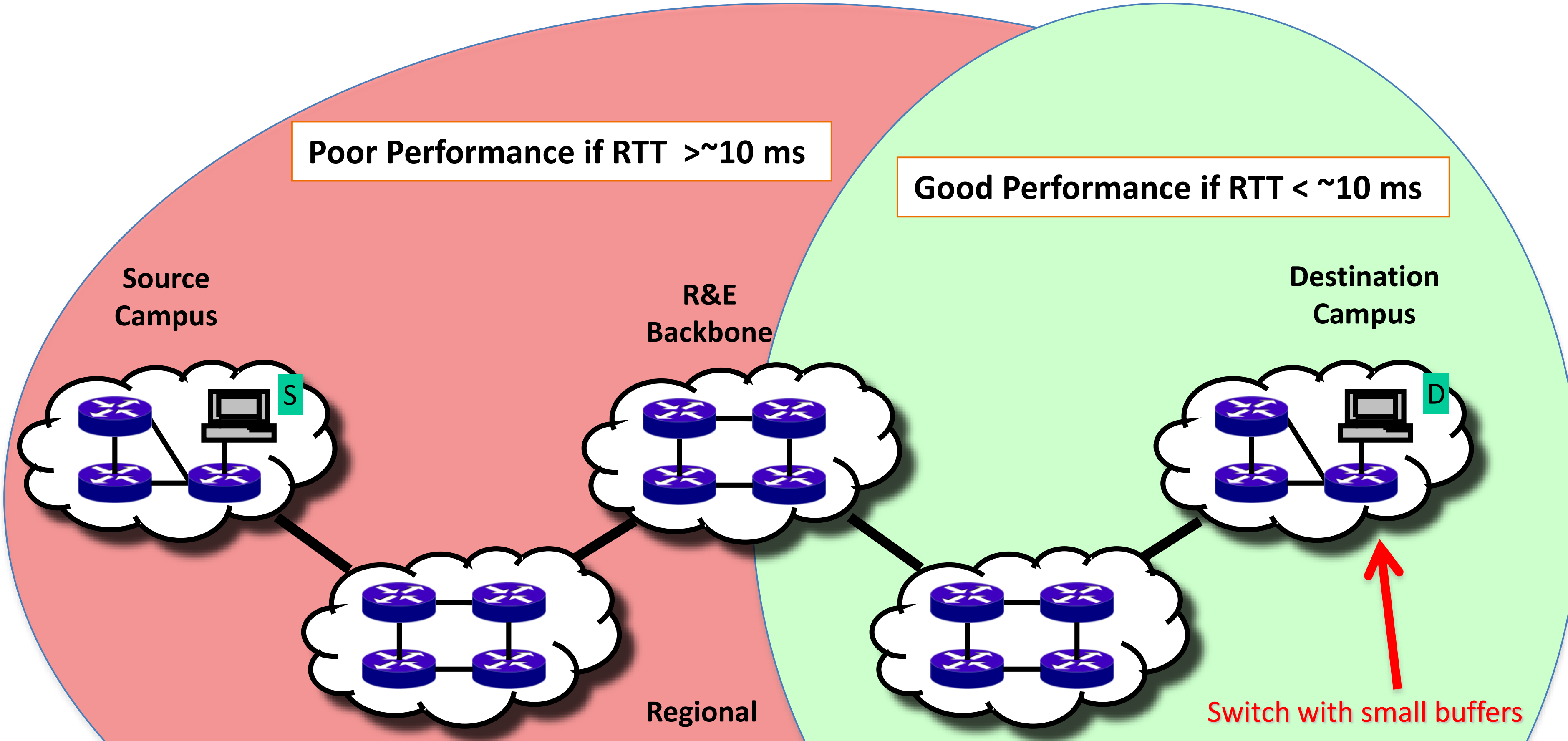
…but network is not infinite





- Just an example, Tier-0 to 2 Tier-1s:
  - Secondary links, usually used for resiliency, are fully exploited

Latency and packet loss matters



0.0046% loss (1 out of 22k packets) on 10G link

- with 1ms RTT: 7.3 Gbps
- with 51ms RTT: 122Mbps
- with 88ms RTT: 60 Mbps (factor 80)

perfSONAR deployment status

- Network monitoring is critical: perfSONAR
  - http://grid-monitoring.cern.ch/perfsonar_report.txt for stats

249 Active perfSONAR instances
199 Running latest version (3.5)

95 sonars in latency mesh
- 8930 links measured at 10Hz
- packet-loss, one-way latency, jitter, ttl, packet-reordering

115 sonars in traceroutes mesh
- 13110 links
- hourly traceroutes, path-mtu

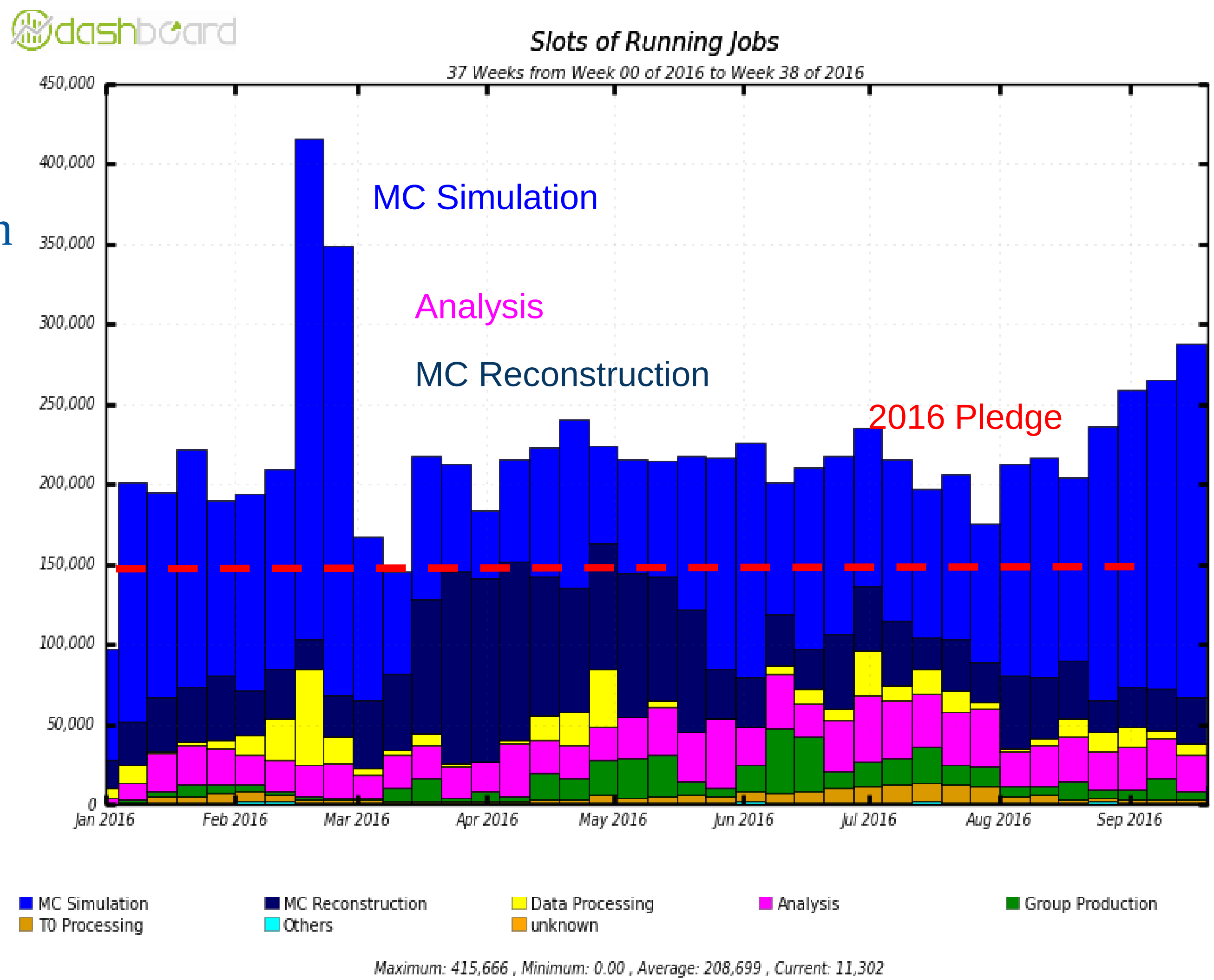
102 sonars in bandwidth mesh
- 10920 links (iperf3)

https://www.google.com/fusiontables/DataSource?docid=1QT4r17HEufkvnqhJu24nIptZ66XauYEIBWWh5Kpa#map:id=3

- Initial deployment coordinated by WLCG perfSONAR TF
- Commissioning followed by WLCG Network and Transfer Metrics WG

Workload management: CPU usage

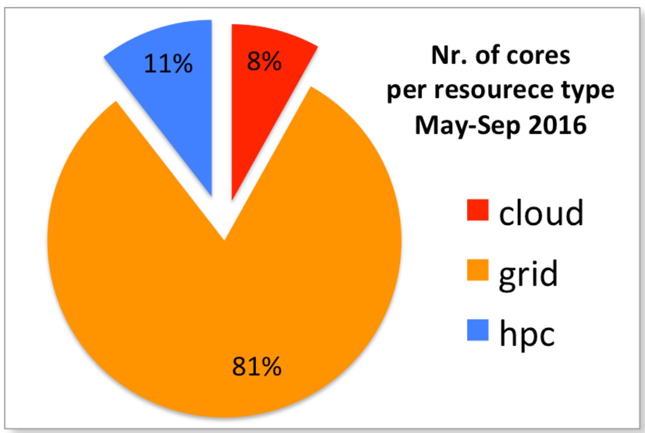
- Using much more CPU than pledged
- Significant I/O stress:
  - Higher pile up - MC reconstruction
  - Longer I/O intensive campaigns

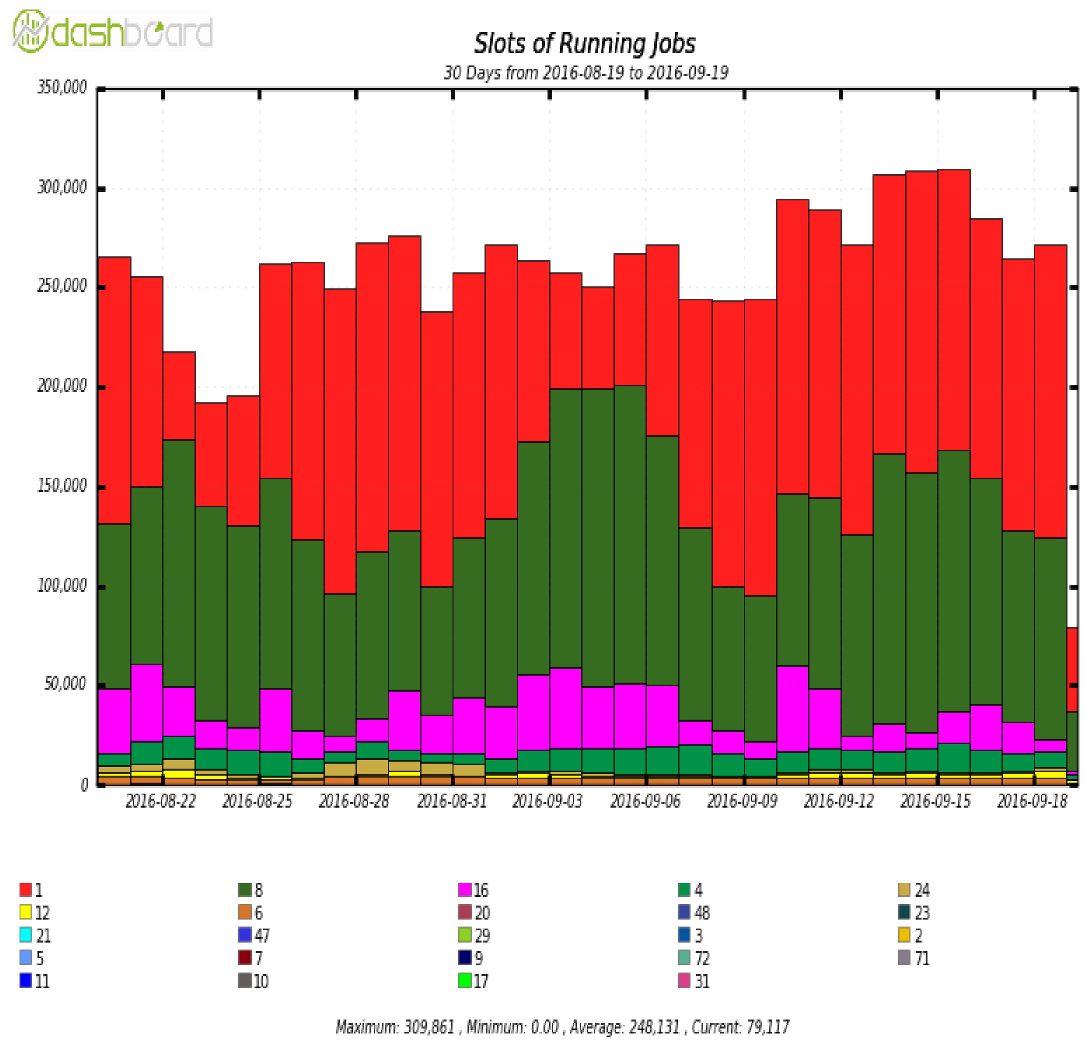


CERN

Workload management: flexible

- Single and MultiCore at all the sites
  - Also (on quite many sites) High Memory slots



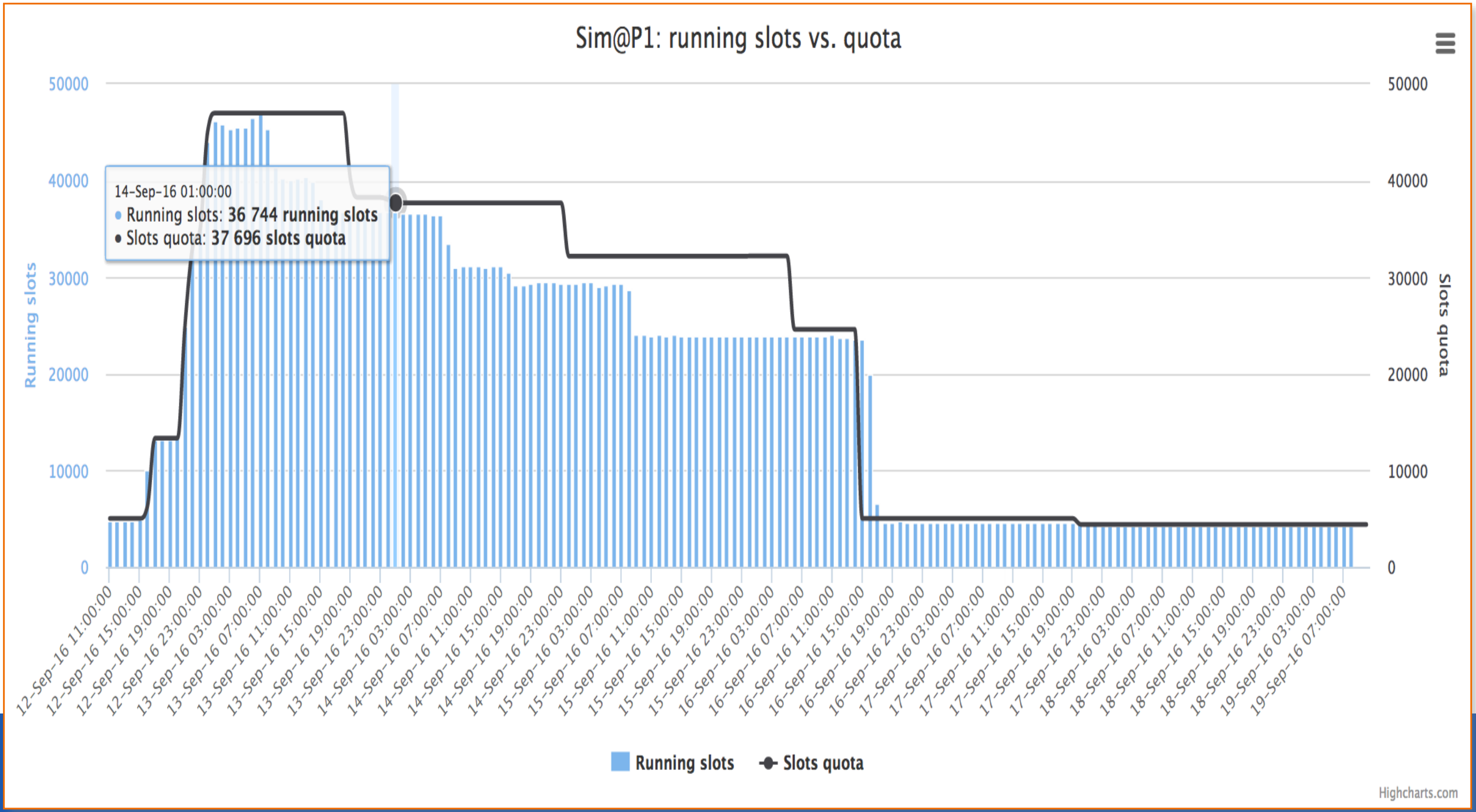


- Running opportunistic (also) on:
  - HPC
  - Clouds

Big investment –> big return!

Exploiting “opportunistic”: 1 example

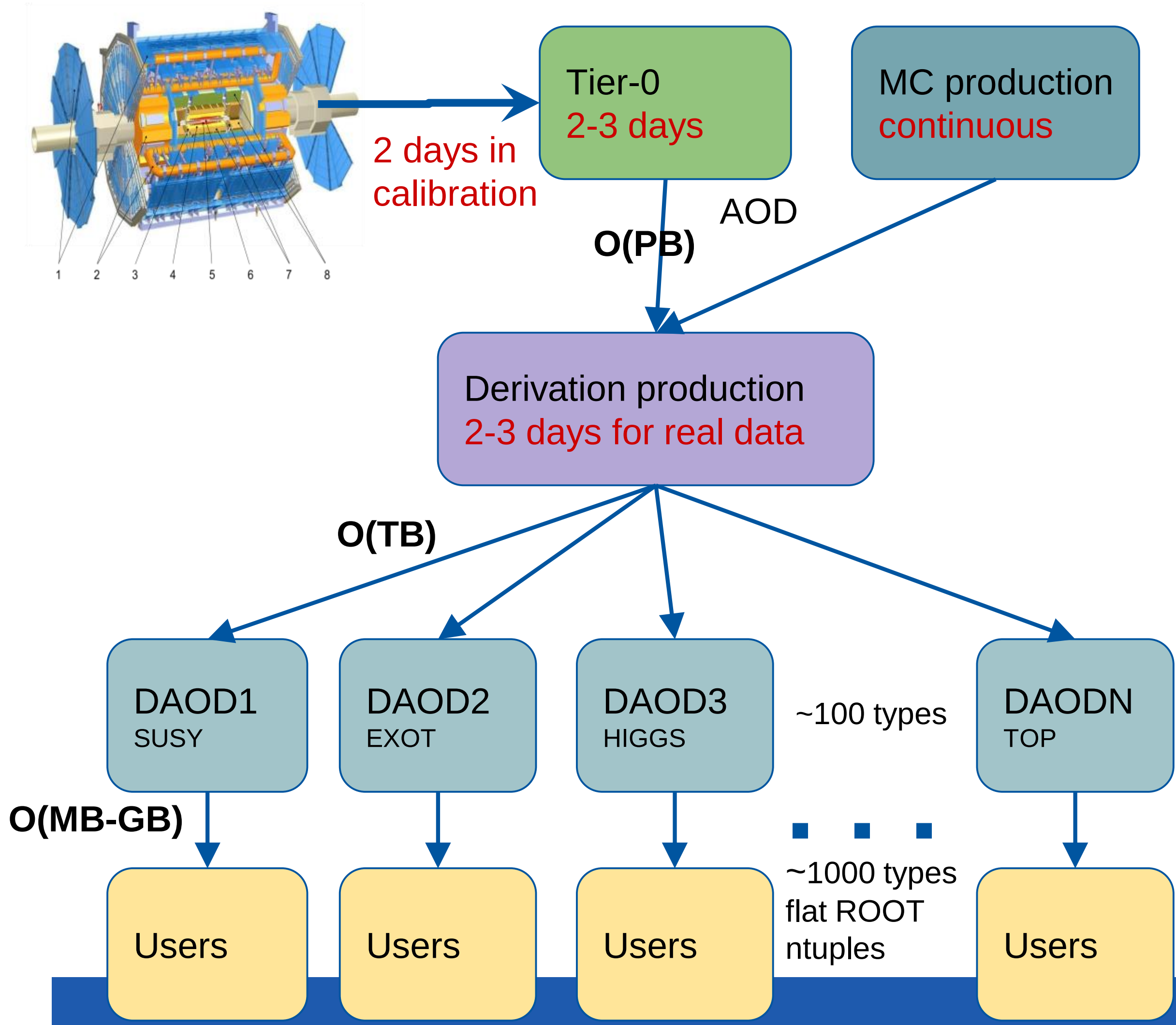
- Grid Simulation on the ATLAS HighLevelTrigger farm when not used for online



CERN

From Detector to the Physicists: Derivations

- Centrally managed production of analysis specific DAOD datasets (reduced data format from main AOD format)
- Real data:
  - Available ~1 week after data taking
- Several campaigns with improved sw on data and MC
  - 2-3 weeks to process



Data Persistency

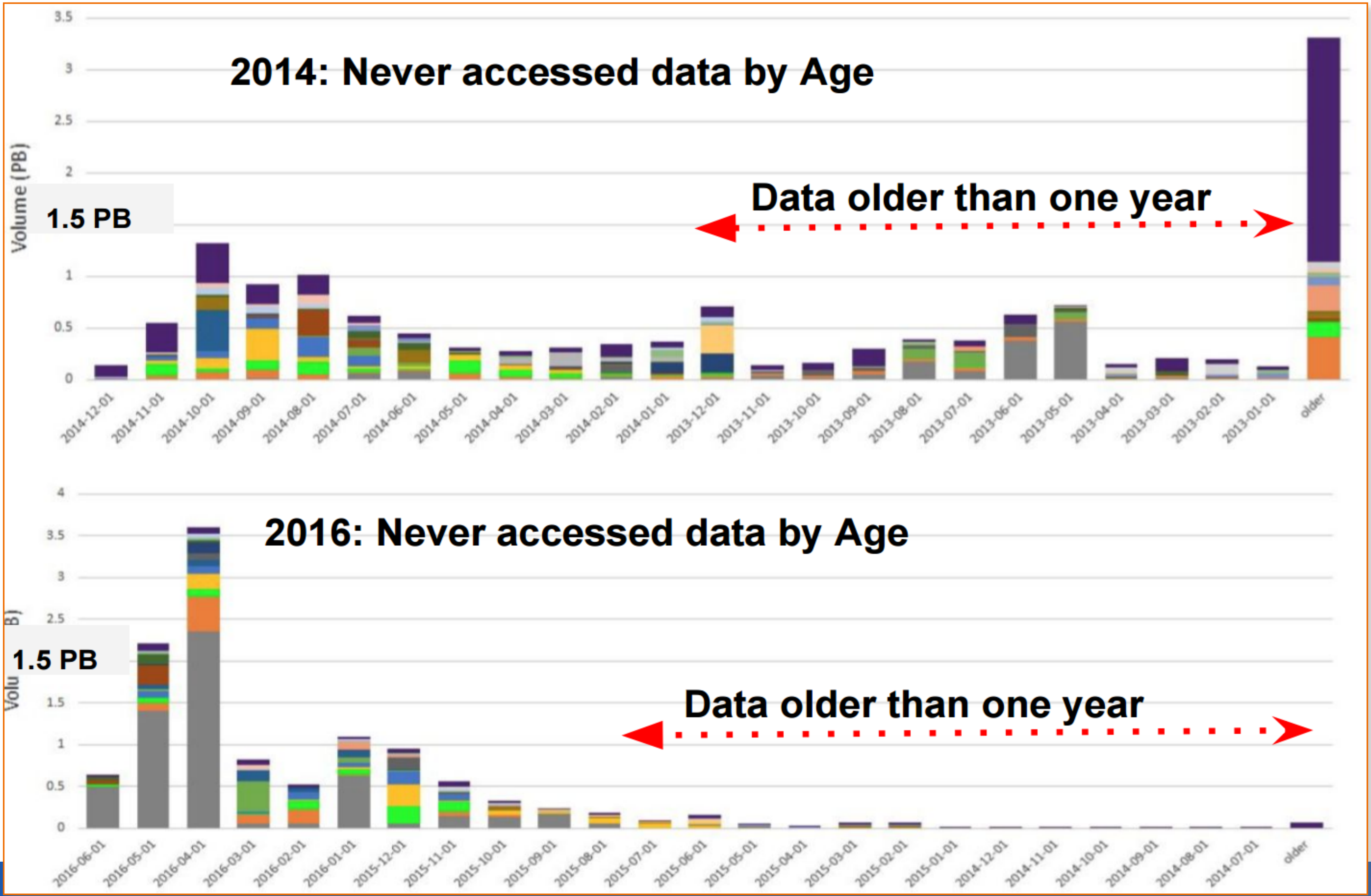
… what do we do with all these real and simlated data that we reconstruct and skim/slim/thin?

→ **the Data Lifetime Model**

Data Persistency - the lifetime model

- There is too much data to keep on storage permanently
- Each data type is set a finite lifetime:
  - Analysis inputs (DAOD) - 6 months, fast turnaround
  - Monte-Carlo simulations - 2-3 years, expensive to regenerate
  - RAW data - unique and precious, infinite lifetime
- Frequently used data - lifetime extension
- Monthly cleanup procedure for expired datasets
  - Approval of exceptions
  - Permanent automated deletion of expired data

The lifetime model in action



Upcoming features: Run2 and Run3

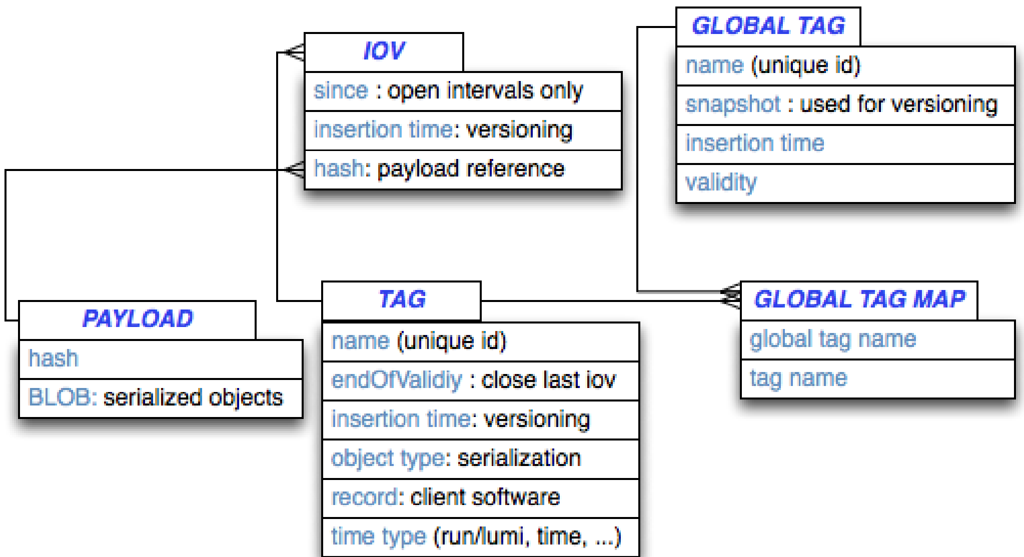
<u>Run3 will be as challenging as Run2</u>

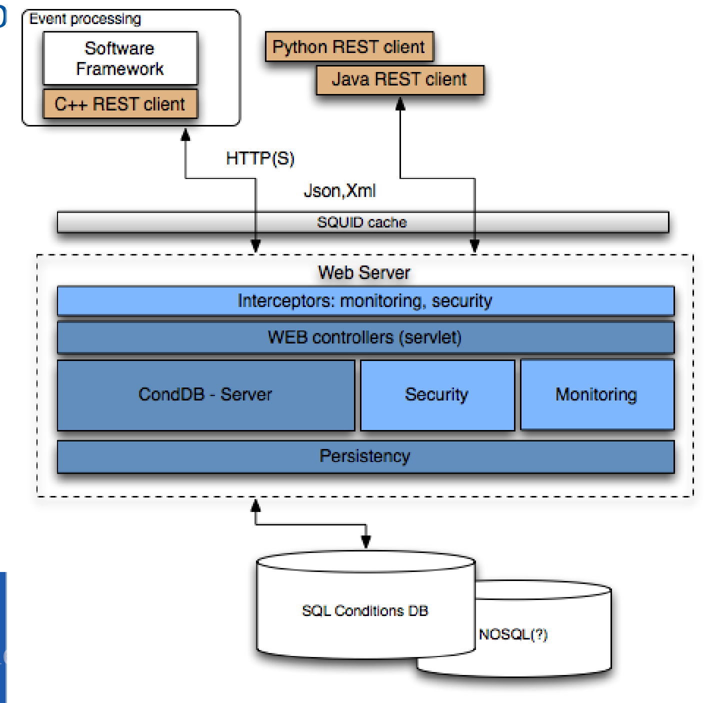
- Same Data Taking trigger rate of 1Khz (physics)
- More pile-up → more resources in particular for reconstruction
- Run4 is a completely different story – not for this talk!

- The Present Future:
  - Global fair-shares
    - Limit the cpu slots per activity, boost activity when requested
  - New Conditions Data architecture (next slides)
    - to enable new workflows today (almost) impossible
  - Machine learning studies and analytics (next slides)
    - All the monitoring records are stored in ElasticSearch for detailed analysis
  - Event service (next slides)
    - Exploit the vanishing opportunistic resources up to the last drop!
  - … and much more

ATLAS Conditions Data

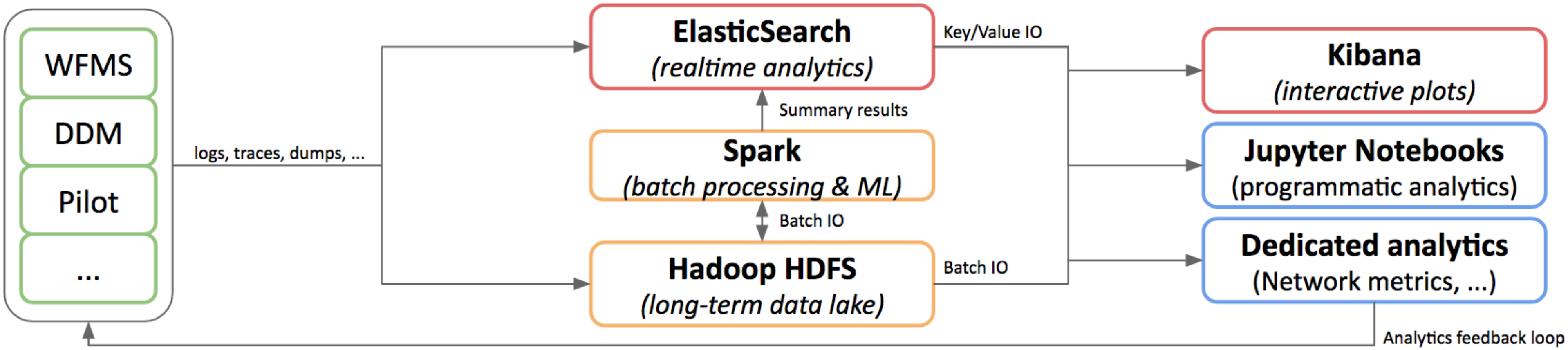
- Physics data processing relies on Conditions Data
  - Conditions: set of parameters related to the detector (alignments, calibrations, …), essential for reconstruction (and simulation) of physics data
- Simplify conditions storage (from ~10K tables to 10…)
  - Data Model : implement simple data model (few tables) by using a CMS-like approach
  - Re-enforce the multi-tier architecture (Frontier-like) providing REST management tools
  - Simplify client access (disentangle client from the b





Analytics: what and why?

- **Understand** our distributed systems and overall operational performance
- **Correlate** operational data across our systems
- Data mining or **machine learning** algorithms on raw and aggregated data
- Ability to host third party analytics services on a **scalable compute platform**
- Satisfy variety of use cases for different user roles for **ad-hoc analytics**
- Provide an **open platform** with documented collections and tools



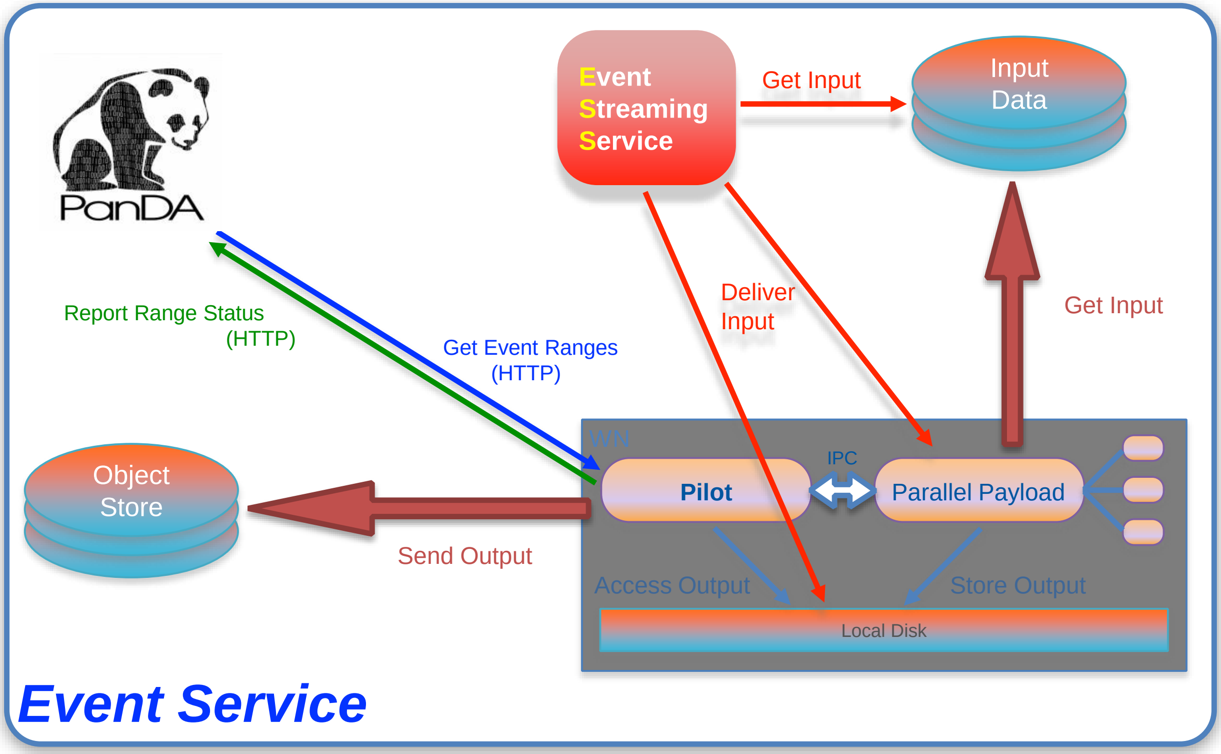
Analytics: advanced use cases

- **Ad-hoc** analytics — done by users on the open platform
- Dedicated analytics projects
  - DDM Metrics aggregation, …
  - Scrutiny group reporting, Group space accounting, ...
- Many **machine learning** projects running in parallel
  - Network performance modeling: Regression models to estimate throughput/latency
  - Time To Complete Estimation: ProdSys task duration, File Transfer duration
  - Support for computing operations: Correlate anomalies, recommend actions, automate
  - Smart data placement
    - Uses DDM metrics, network performance modeling, TimeToComplete estimation
    - Decide where to place input and output files
    - Automatic rebalancing

Event Service: the concept

- A fine-grained approach to event processing.
  - **Designed for exploiting diverse, distributed and potentially short-lived resources**
    - Quasi-continuous event streaming through worker nodes
- Exploit event processors fully and efficiently through their lifetime
  - Real-time delivery of fine-grained workloads to running application
  - Be robust against disappearance of compute node on short notice
- Decouple processing from chunkiness of files, from data locality considerations and from WAN latency
- Stream outputs away quickly
  - Negligible losses if the worker node vanishes
  - Minimal demands for the local storage

Event Service: schematics and status



- Pilot delivers fine-grained workloads to the running payload application in real time
    - Workload: **Event Ranges**
- Payload application: **process-parallel version of Athena (AthenaMP)**
    - Serial initialization in the master process
    - Then fork worker processes
    - Workers process the events

- Event Service: commissioning towards full production
    - First use case: ATLAS Geant4 simulation
    - Exploiting opportunistic resources HPC-like

Conclusions

- Big efforts to evolve and (partially) redesign the ADC systems is paying off!
  - ✓ Cope well with higher-than-expected Run-2 LHC performance
  - ✓ Presently no scaling issues! Each subsystem has demonstrated to be able to absorb ~5 more than the average load
    - Still, it might not be sufficient for high-luminosity LHC Run-4

✓ **ATLAS Distributed Computing perform extremely well**

- produce physics results on time for conferences


© Randy Glasbergen
glasbergen.com
GLASBERGEN
"It's the latest innovation in office safety.
When your computer crashes, an air bag is activated
so you won't bang your head in frustration."

მადლობა!