ATLAS Computing in Run-2

L. Poggioli, LAL

- History
- Limitations of current model
- Run-2 challenges & solutions

Most inputs from:
- Borut Kersevan, current ATLAS computing coordinator
- Eric Lançon, future ATLAS computing coordinator

ATLAS resource utilization in 2013

- ATLAS has utilized computing resources in Tiers well in last year: Many thanks to sites for resources & excellent operating!!
- Manage to provide timely analyses thruput to meet physics requirements
- An ongoing effort in s/w development to optimize resource utilization by reducing CPU consumption, event sizes for Run-2,…

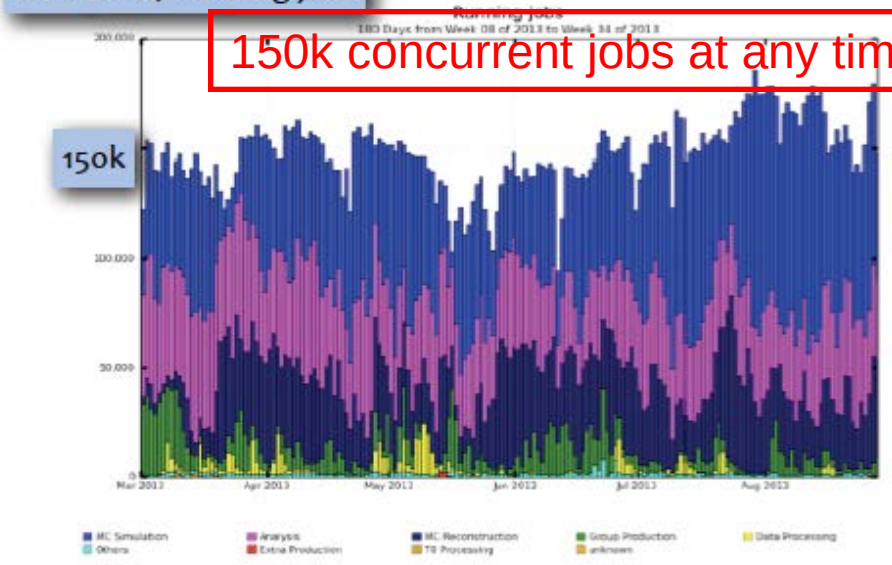
ATLAS RESOURCE USAGE IN FIRST HALF OF 2013 (RRB)

	Location	Requested	Used
CPU [kHS06]	CERN	111	111
	Tier-1	316	435
	Tier-2	360	713
Disk [PB]	CERN	9.6	8.9
	Tier-1	35 [38]	35
	Tier-2	51 [52]	48
Tape [PB]	CERN	25	29 (incl. 9 PB of ESD)
	Tier-1	42	33

CPU delivered much above pledges

Tiers CPU / running jobs

150k concurrent jobs at any time



The Challenges of Run-2

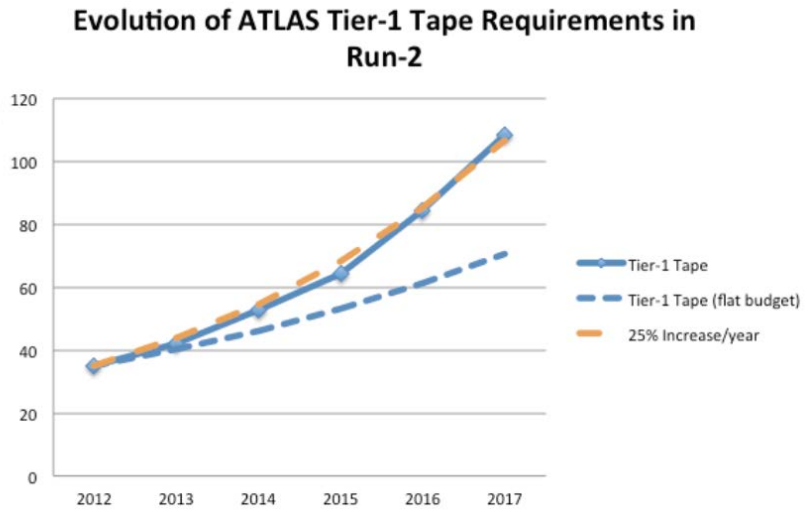
- Flat budget constraints
  - Both for h/w & operation & dev't
  - h/w increase from Moore's law gain
    - Estimated factors of 1.2/year for CPU and 1.15/year for disk
- Data from Run-1
  - Proper data preservation
- LHC operation
  - HLT rate 1 kHz
  - Pile-up > 30
  - 25ns buch spacing
  - c.m. energy x 2
- 'New' detector
  - To be integrated in simul & reco
- (New CPU architec.)
  - Less memory/core

Budget: 'Flat' budget model

- Cost inputs
  - cpu -20%/yr, disk -15%/yr tape -15%/yr
  - Under validation (ICB h/w survey)
- C-RSG followed requests
  - eg Tape
- Outcome
  - Seems doable for CPU & disk
  - May be problematic for tapes
- NEW (LHCC)
  - Drop 'flat budget' framewk
  - Physics motivated needs should be stated instead…

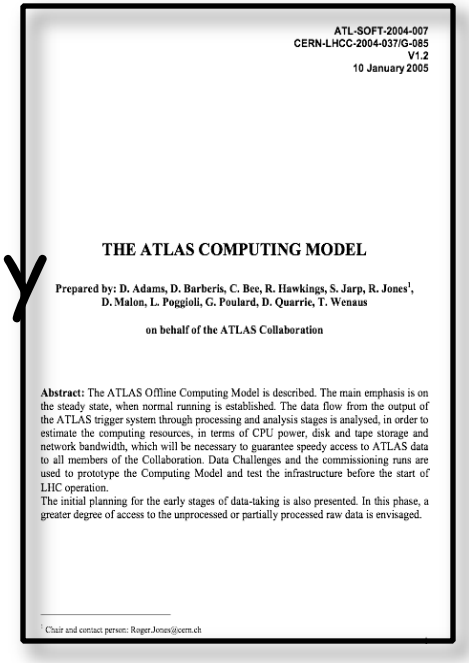
C-RSG Recommendation

Resource	Site	2014 ATLAS	2014 CRSG	2015 ATLAS	2015 CRSG
CPU (kHS06)	T0+CAF	111(111)	111	205(240)	205
	T1	365(385)	355	462(478)	450
	T2	425(412)	390	530(522)	520
Disk (PB)	T0+CAF	12(12)	11	14(15)	14
	T1	35(35)	35	39(47)	37
	T2	52(56)	49	55(65)	52
Tape (PB)	T0+CAF	29(29)	27	33(38)	33
	T1	53(55)	53	65(74)	65



Initial Computing Model (2005)

- Derived from MONARC (1999)
- CERN-Tier0 the center
- 10 T1s connected by dedicated 10Gb/s links (LHCOPN)
- O(100) T2s each attached to a T1
- The data flows along the hierarchy
- Insufficient networking assumed
- Hierarchy of functionality and capability



ATL-SOFT-2004-007
CERN-LHCC-2004-037/G-085
V1.2
10 January 2005

THE ATLAS COMPUTING MODEL

Prepared by: D. Adams, D. Barberis, C. Bee, R. Hawkings, S. Jarp, R. Jones[1], D. Malon, L. Poggioli, G. Poulard, D. Quarrie, T. Wenaus

on behalf of the ATLAS Collaboration

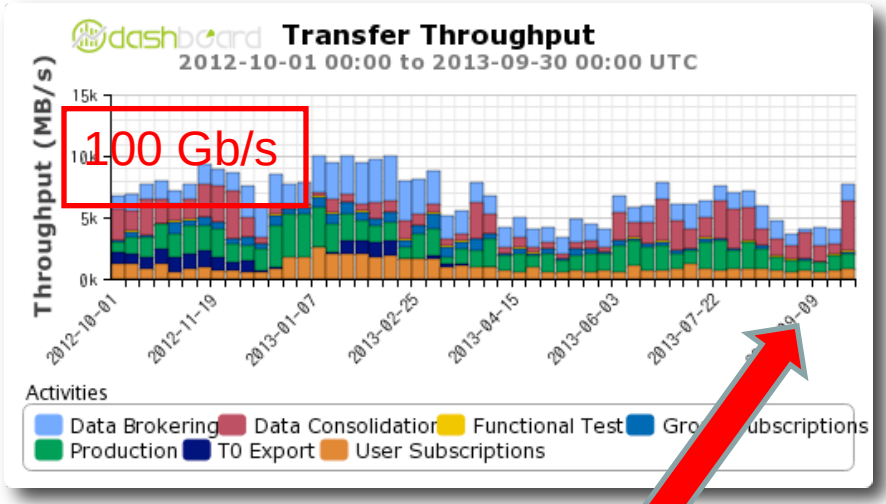
Abstract: The ATLAS Offline Computing Model is described. The main emphasis is on the steady state, when normal running is established. The data flow from the output of the ATLAS trigger system through processing and analysis stages is analysed, in order to estimate the computing resources, in terms of CPU power, disk and tape storage and network bandwidth, which will be necessary to guarantee speedy access to ATLAS data to all members of the Collaboration. Data Challenges and the commissioning runs are used to prototype the Computing Model and test the infrastructure before the start of LHC operation.
The initial planning for the early stages of data-taking is also presented. In this phase, a greater degree of access to the unprocessed or partially processed raw data is envisaged.

[1] Chair and contact person: Roger.Jones@cern.ch

http://monarc.web.cern.ch/MONARC/

2010-2013: Many changes

- Hide grid complexity from users, **simplif'ions**, less middleware dependence
- **Caching** vs **centralized** DB
  - Conditions data access from any site, not only at T1s (**squid**, **frontier**)
  - No more need to pre-install s/w releases at sites (**cvmfs**)
- **Dynamic** data placement & deletion based on **popularity**
  - Better usage of disk space
  - Reduced job waiting times
- T2→N-T1s & T2↔T2 exchanges (**T2D**)



Network performing over expectations

LHCONE (2011)
Dedicated netwk between (some) WLCG sites

2010 Evolution 2013

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

~20 AOD copies distributed worldwide

2010 Evolution 2013

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

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

~20 AOD copies distributed worldwide

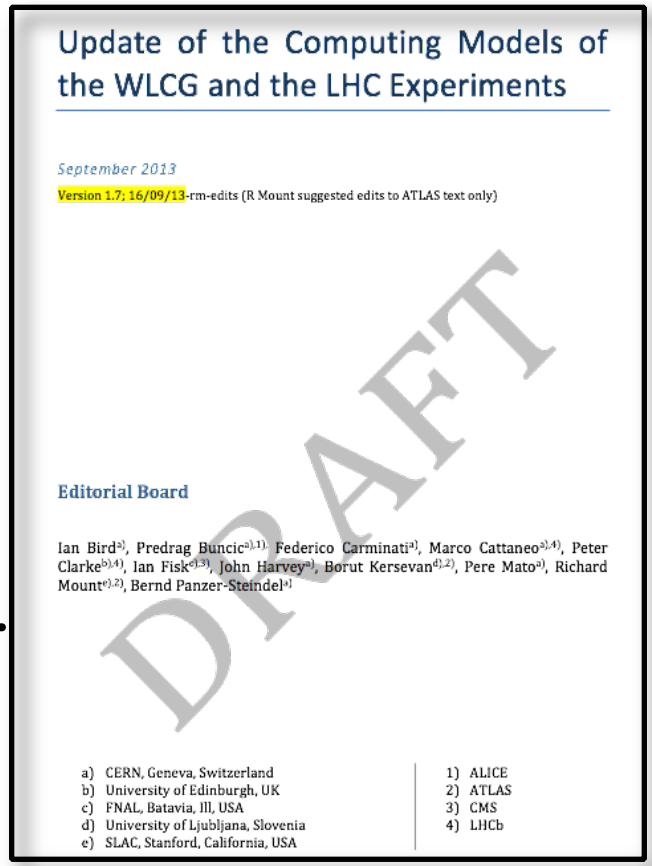
4 AOD copies distributed worldwide

Some limitations of current model & tools

- Partitioning of resources
  - Analysis vs Central Production
  - T1s versus T2s
- Difficulties of current Data Distribution Management & production systems to accommodate new use cases & technologies
- Memory increase of MC pile-up digitization & reconstruction
- Multitude of data format for analysis
- Full reprocessing once a year

Run-2 Computing Model

- Common document from all experiments submitted to LHCC
- This is not a completely new model but an **extrapolation** and **extension** of end of Run-1 framework
- New tools under development for higher scalability to meet Run-2 challenges



Update of the Computing Models of the WLCG and the LHC Experiments

September 2013

Version 1.7; 16/09/13-rm-edits (R Mount suggested edits to ATLAS text only)

DRAFT

Editorial Board

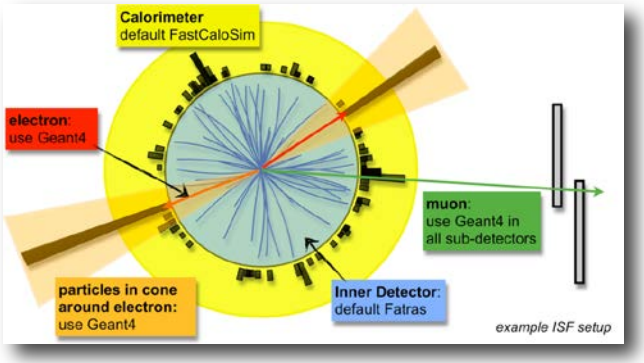
Ian Bird[a], Predrag Buncic[a),1)], Federico Carminati[a], Marco Cattaneo[a),4)], Peter Clarke[b),4)], Ian Fisk[c),3)], John Harvey[a], Borut Kersevan[d),2)], Pere Mato[a], Richard Mount[e),2)], Bernd Panzer-Steindel[a]

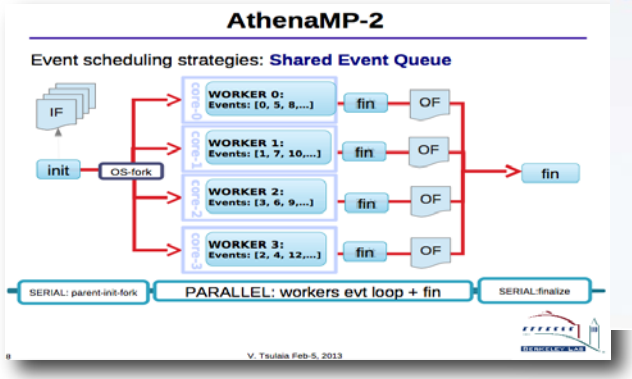
a) CERN, Geneva, Switzerland
b) University of Edinburgh, UK
c) FNAL, Batavia, Ill, USA
d) University of Ljubljana, Slovenia
e) SLAC, Stanford, California, USA

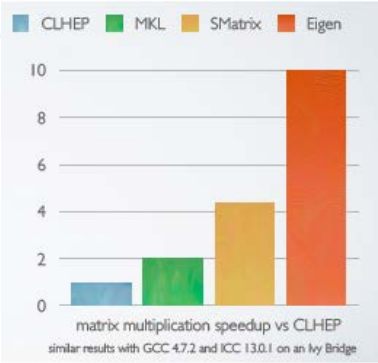
1) ALICE
2) ATLAS
3) CMS
4) LHCb

Working towards solutions

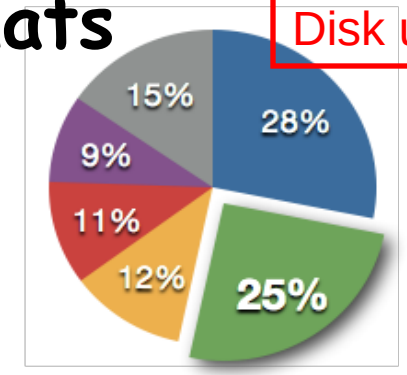
- Simulation: **CPU**
  - Integrated Simulation Framework
- Reconstruction: **Memory & CPU**
  - Parallelism, code speedup
  - MP solution to reduce memory footprint
- Analysis Model : **Multiplication of data formats**
  - Common analysis data format, xAOD
  - Streamlining analysis flow





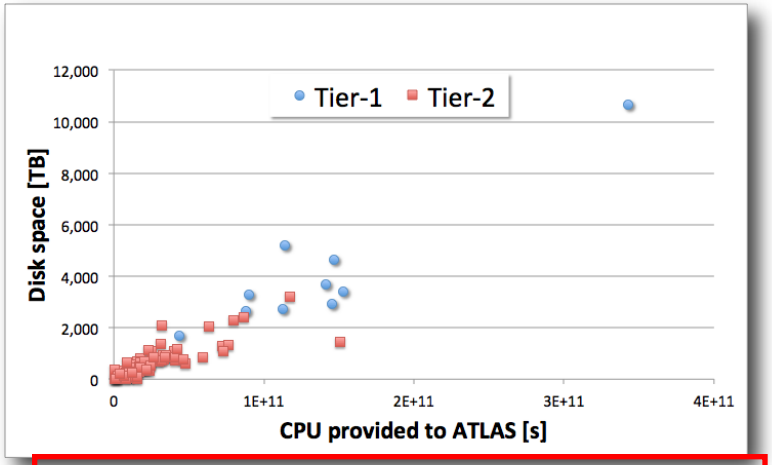


Disk usage @ T1s & T2s



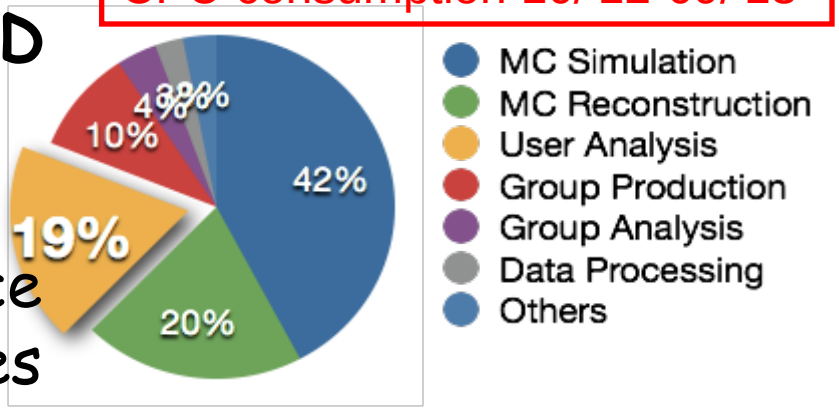
Data Processing

- Optional extension of 1st pass processing from **T0 to T1s** if resource shortage at T0
- T1s and some T2s used for most demanding workflows: **high memory & I/O intensive tasks**
- Data reprocessing & MC reco. also performed at **some T2s**
- Still one full reprocessing from RAW /yr, but multiple **AOD2AOD repro/yr**
- **Derivation Framework** (train model) to centrally produce TB size data samples for analyses



Some T2s equivalent to T1s wrt disk storage & CPU power

CPU consumption 10/'12-09/'13



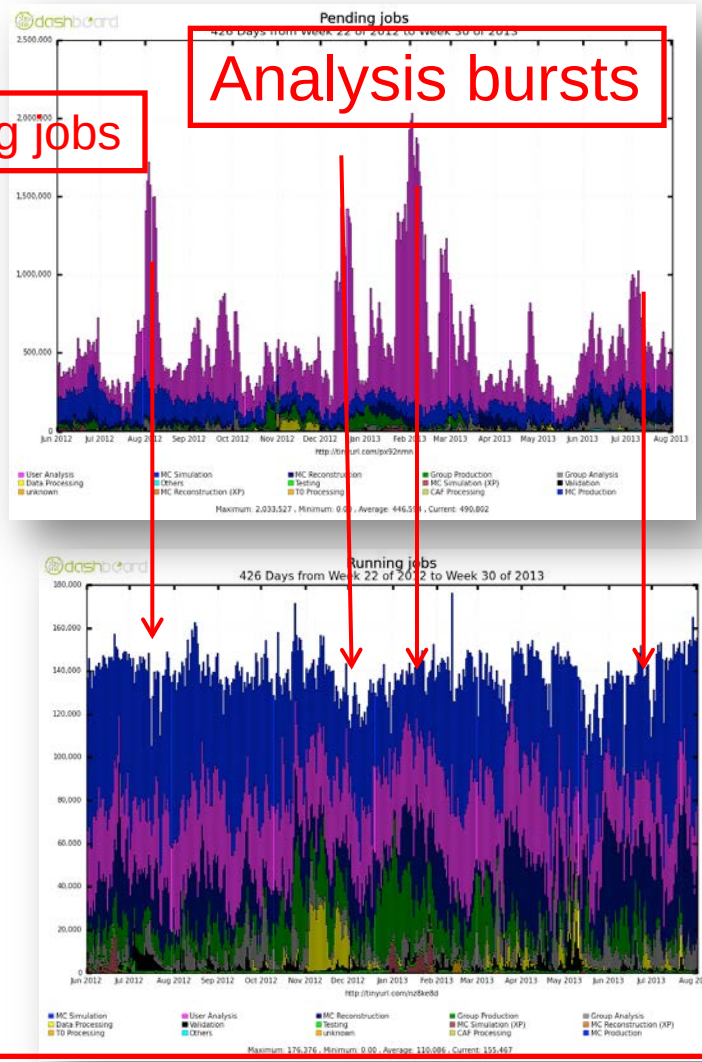
Data Placement in Run-2

- Initially **2 copies** of analysis data formats (xAOD: 1 at T1s, 1 at T2s)
  - Already being implemented to gain disk space
- **Non-popular** data -> archived **to tape** at T1s
- In addition
  - In October recovered 9% disk space from data not accessed over the last 9 months
  - Minimal number of copies on disk not guaranteed
  - User access to data on tape granted through centralized tools

Extra-load on tape system

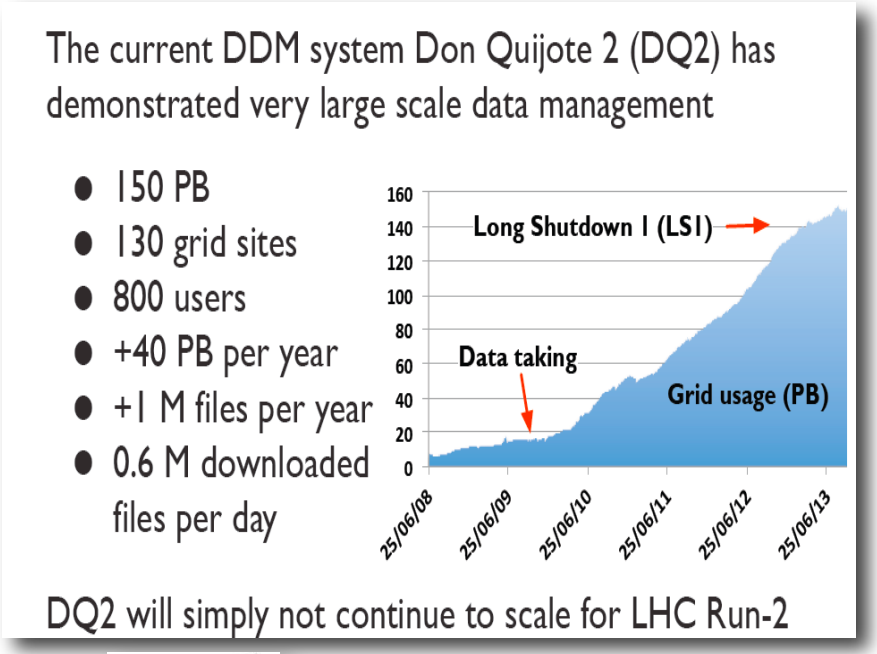
New production system: PRODSys2

- Same engine for analysis & production
  - PanDA+JEDI+DEfT
  - Current analysis vs prod. shares managed by sites not by ATLAS
  - Better reactivity to analysis loads
- Data traffic minimized
- Optimized job to resource matching



Data Distribution Management & Databases

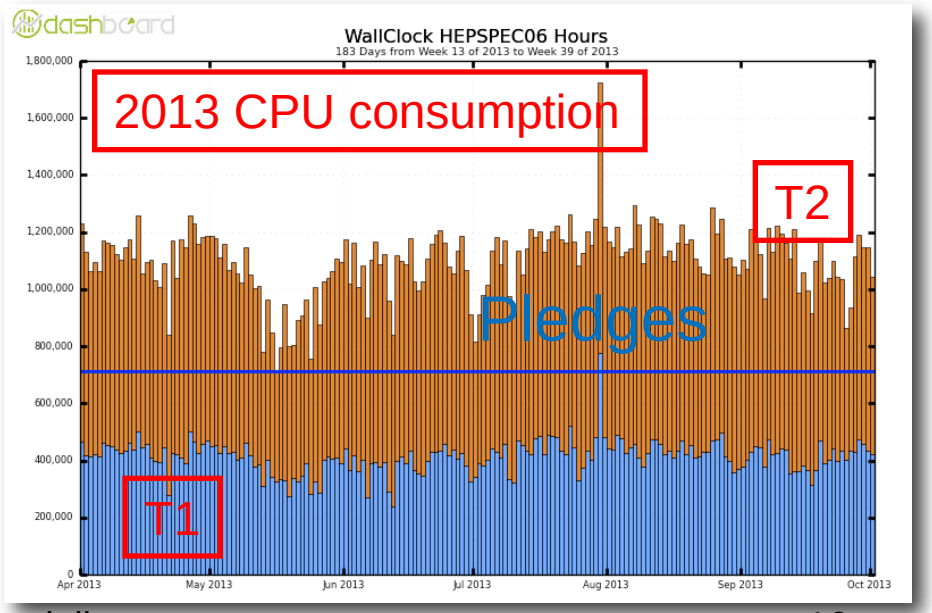
- New DDM system: RUCIO will replace DQ2
  - New scalable architecture
  - File level vs dataset functionality
  - Built-in data replication policy for space & network optimization
  - Multi-protocol (http,…)
- Database infrastructure: simplification and streamlining



http://rucio.cern.ch

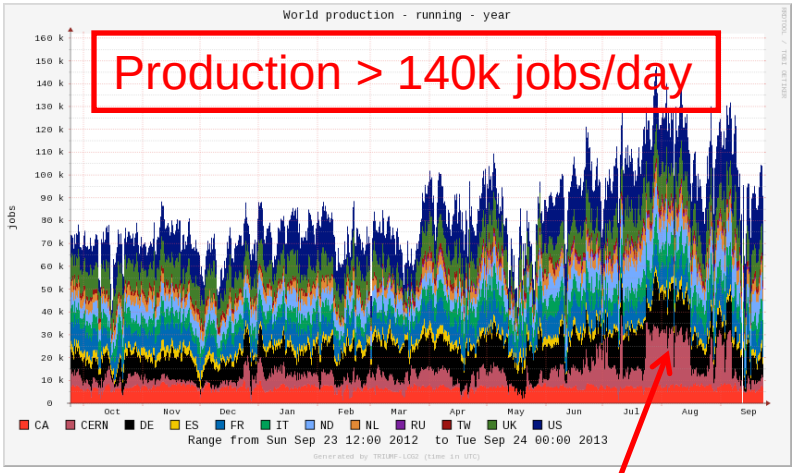
Opportunistic resources

- CPU consumption above pledges at T1s & T2s
  - Needs larger than official requests
- Sites and Funding Agencies provide more than pledged resources (thank you!)
- Additional solutions
  - HLT farm at P1
  - Cloud computing
  - Large HPC centers
  - Volunteer computing: ATLAS@home



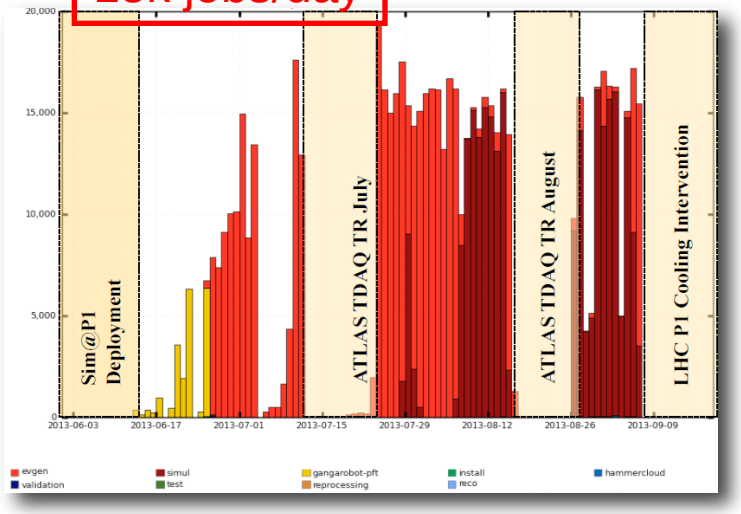
HLT Farm @ P1

- HLT farm cloudified mid-2013
  - Reached >15k concurrent simulation jobs
  - Switch between trigger & simulation mode tested
- Availability in Run-2
  - For MC production during shutdowns or LHC technical stops
  - ~30% over a Run2 year?



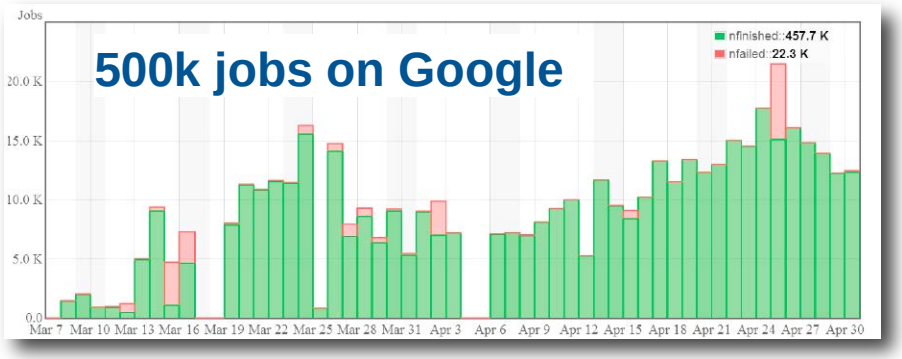
HLT Farm

15k jobs/day



Equivalent of T1 or big T2

Cloud Computing



- Ongoing R&D on academic clouds and Amazon or Google (AUS,CA, US,…)
- Issues with long jobs and I/O
- Plan to use academic clouds & cheap commercial (opportunistic) is possible
- Some providers -> cost-competitive offers (with some limitations)

HPC (High-Performance Computing) resources

SuperMUC a PRACE T0 center
- 155k Sandy Bridge cores
  2.8M HS06
- WLCG 2013 T0/1/2 pledges
  ~2.0M HS06

- Large investments in many countries: from Peta to Exa scales
  - http://www.eesi-project.eu/pages/menu/eesi-1/publications/investigation-of-hpc-initiatives.php
  - Latest competitive SC are familiar Linux clusters
- Large number of spare CPU cycles available at HPCs & not used by 'standard' HPC applications
  - Projects to use idle CPU cycles at HPC centers in US, China & DE
- Demonstrators working for simul & evt generation
  - Difficult to use HPC centers for I/O intensive applications
  - Outbound connectivity of HPC centers may be an issue
- Some T2s -> pledges resources on shared HPC facilities

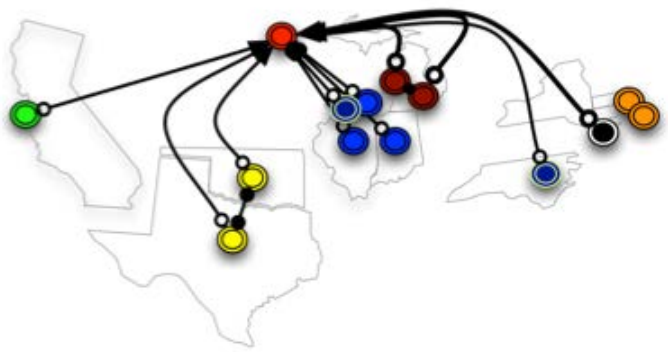
Might endanger traditional computing budget

Network potential & usage

- Networking will probably continue its progress & evolution further
  - In terms of bandwidth increase
  - In terms of new technologies (eg NaaS)
- 2 interesting ATLAS initiatives ongoing
  - **Data federation** (FAX, xrootd fed., http fed.)
    - Remote file access over WAN
  - **Event Service**: passing single evts for processing from/to storage

Adopting such solutions in full could optimize our disk space & CPU needs

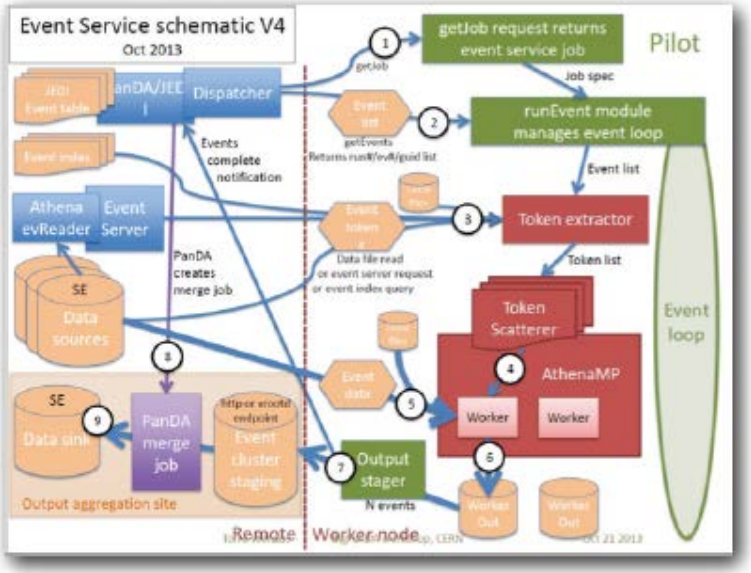
Distributed storage/Remote access



- Jobs access data on shared storage resources via WAN
- For better usage of storage resources (disk prices!)
- Bandwidth and stability needed
- FAX (Federating ATLAS data stores using Xrootd) demonstrator
  - job fail-over if access failure for 1st implementation
- http protocol also considered

Event Service

- In development: software and distributed computing effort
- Feed Virtual Machines with short jobs (simulate one single event)
- Usages
  - Backfilling of HPC centers
  - Opportunistic use of commercial clouds
  - Volunteer computing (ATLAS@home)



Conclusion & Outlook

- A lot of experience acquired by ATLAS in 3 yrs of data taking
- Run-2 will put high pressure on hardware and human resources
- New computing model and its components will be tested during **2014 data challenge**
- ATLAS **upgrades** also mean resources for software & computing
- Solutions under development & **Manpower** is crucial:
  - For development
  - For operation & support
    - In particular ATLAS support at T1s has proven to be essential