



Erevan - 29 September 2011

Data handling and processing for the ATLAS experiment

Dario Barberis

(Genoa University/INFN)

On behalf of the ATLAS Collaboration



Overview

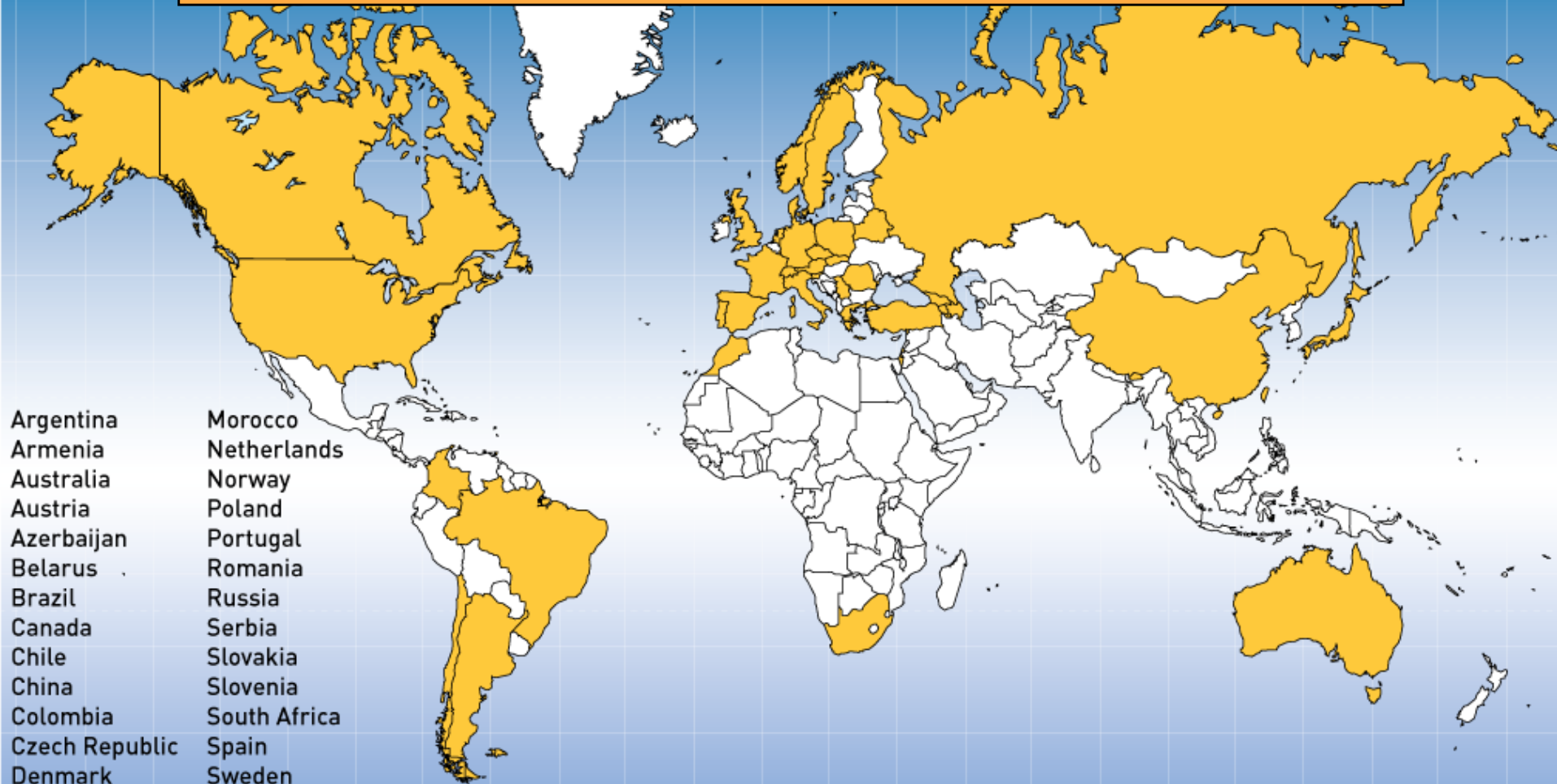
- Data collection
- Tier-0 processing
 - Fast calibration loop
 - Event reconstruction
 - Data export
- Some of the key distributed computing (Grid) technologies:
 - Data management and distribution with DDM/DQ2
 - Workload management with Panda
 - Re-processing campaigns and simulation production
 - Distributed analysis
 - Conditions Databases with Frontier
- Evolution of the computing model
- R&D projects
- Outlook



3000 active scientists:

- ~ 1800 with a PhD → contribute to M&O share
- ~ 1200 students

174 Institutions, 38 Countries, 6 Continents



- | | |
|----------------|--------------|
| Argentina | Morocco |
| Armenia | Netherlands |
| Australia | Norway |
| Austria | Poland |
| Azerbaijan | Portugal |
| Belarus | Romania |
| Brazil | Russia |
| Canada | Serbia |
| Chile | Slovakia |
| China | Slovenia |
| Colombia | South Africa |
| Czech Republic | Spain |
| Denmark | Sweden |
| France | Switzerland |
| Georgia | Taiwan |
| Germany | Turkey |
| Greece | UK |
| Israel | USA |
| Italy | CERN |
| Japan | JINR |

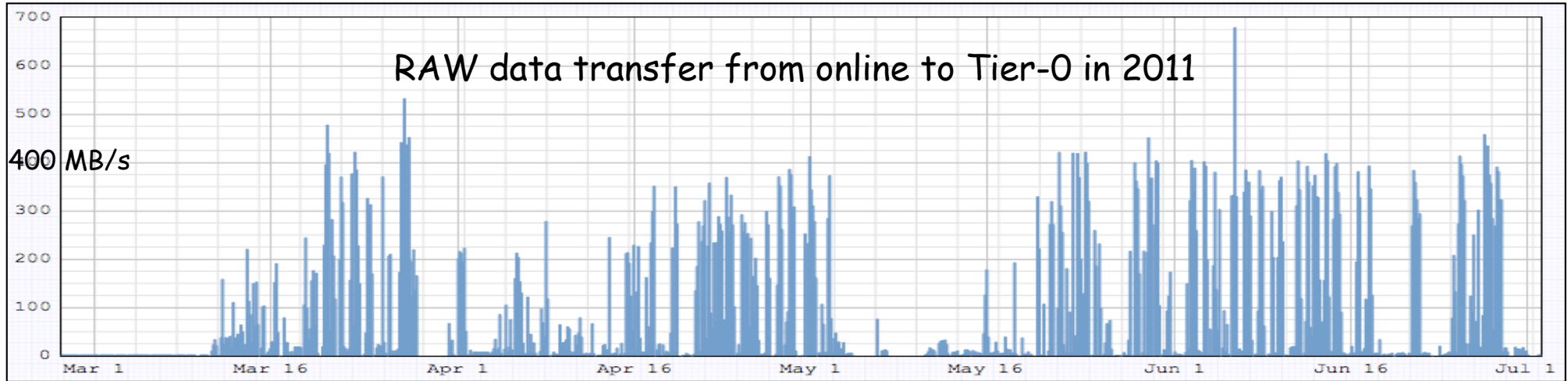
ATLAS
Collaboration



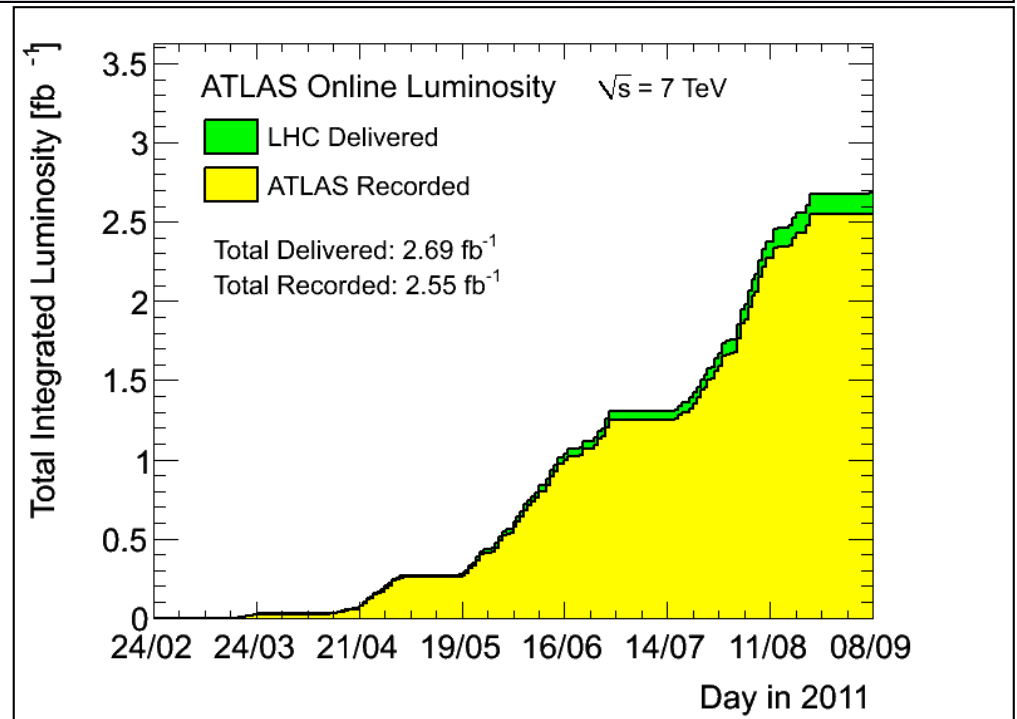


Erevan - 29 September 2011

Data taking in 2011



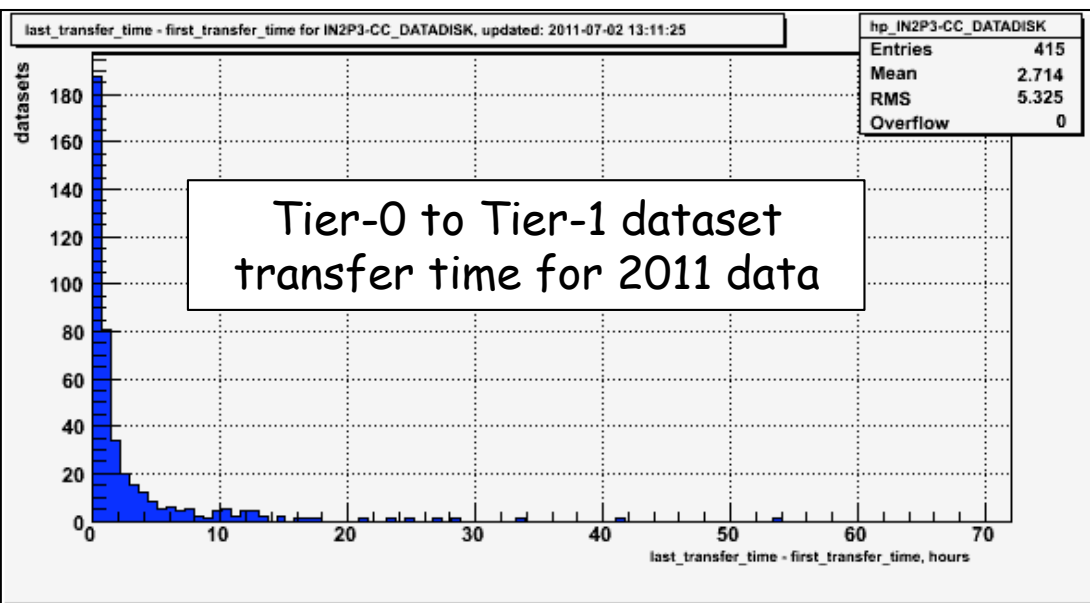
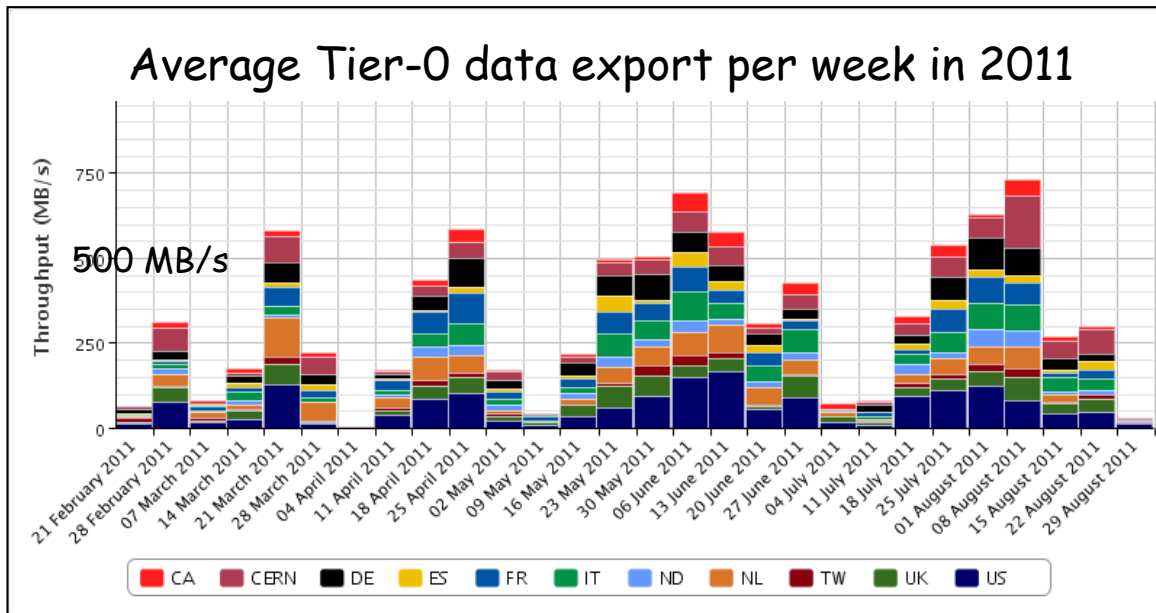
- We took until end August 2011
2.5 PB of RAW data. All data were:
 - Calibrated in real time (within 36 hours)
 - Reconstructed at Tier-0
 - Distributed on the Grid to 10 Tier-1 and many Tier-2 sites
- In total we produced >5 PB of distributed data





Data distribution on the Grid

- Data export from Tier-0 to Tier-1s:
 - RAW: 1 primary copy (on disk) + 1 custodial copy (on tape)
 - ESD: 1 primary + 1 secondary copy (both on disk at different sites)
 - DESD: 2 primary copies
 - AOD: 2 primary + 1 secondary copy
- Also several secondary copies to Tier-2s

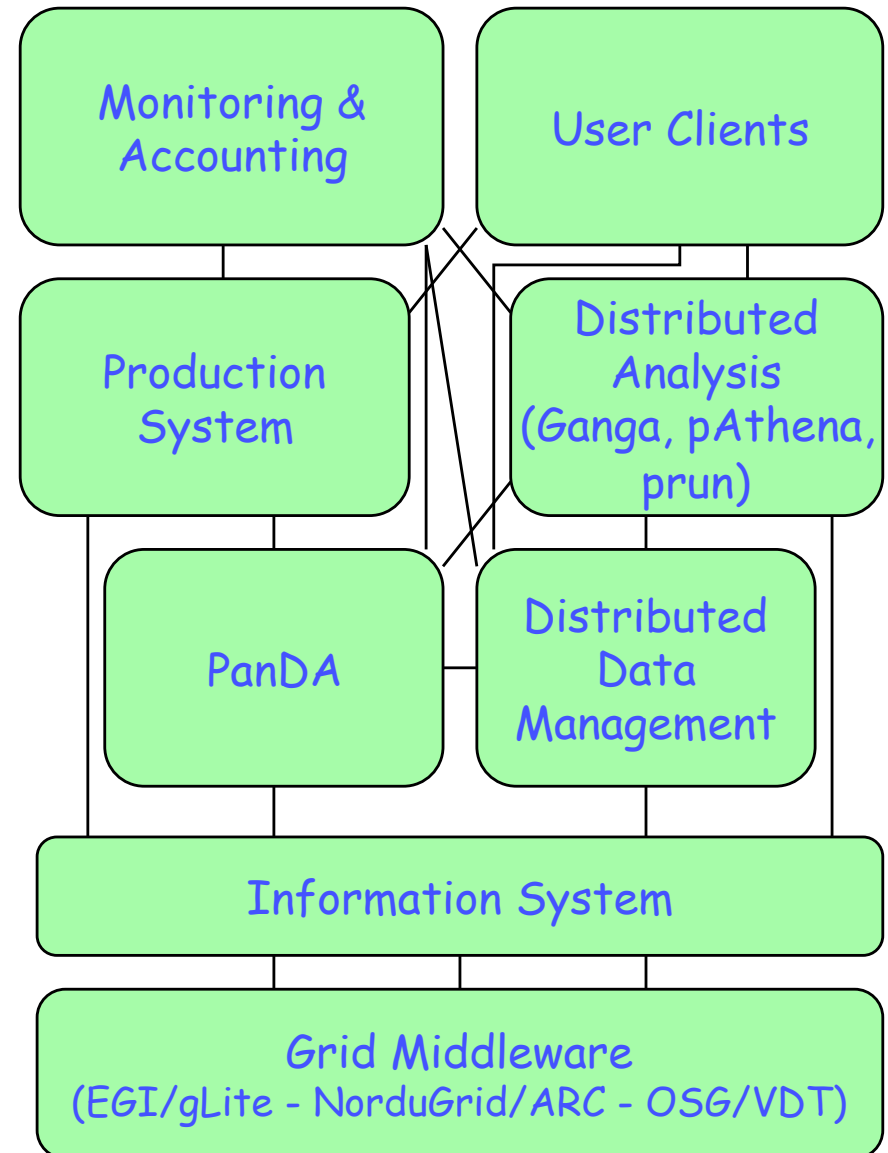


- Data are available for analysis in "almost-real" time. Example:
 - data11_7TeV AOD distribution (to one specific Tier-1 but they are all similar):
 - on average 2.7 hours to complete the dataset
 - Exact time depends on dataset size and how much other traffic is on the network



ATLAS Grid Architecture

- ATLAS runs on 3 middleware suites:
 - gLite in most of Europe and several other countries (including all A-P countries)
 - ARC in Scandinavia and a few other small European countries
 - VDT in the USA
- ATLAS Grid tools interface with the middleware and shield the users from it
 - They also add a lot of functionality that is ATLAS specific
- The ATLAS Grid architecture is based on few main components:
 - Information system
 - Distributed data management (DDM)
 - Distributed production and analysis job management system (PanDA)
 - Distributed production (ProdSys) and analysis (Ganga/pAthena/prun) interfaces
 - Monitoring and Accounting tools
- DDM is the central link between all components
 - As data access is needed for any processing and analysis step!





Distributed data management: DDM/DQ2

- The Distributed Data Management (DDM) architecture is implemented in the DQ2 tools and additional services
- The unit of storage and transfer is the dataset:
 - A dataset contains all files with statistically equivalent events
- DDM and associated tools take care of:
 - Distributing data produced by Tier-0 to Tier-1s and Tier-2s
 - Distributing simulated and reprocessed data produced by Tier-1/2s
 - Distributing user and group datasets as requested
 - Managing data movement generated by production activities
 - Cataloguing datasets (files, sizes, locations etc.)
 - Providing usage information for each dataset replica
 - Deleting obsolete or unnecessary replicas of datasets from disk when disks are full
 - Providing end-users with client tools to operate on datasets (import/export/move etc)

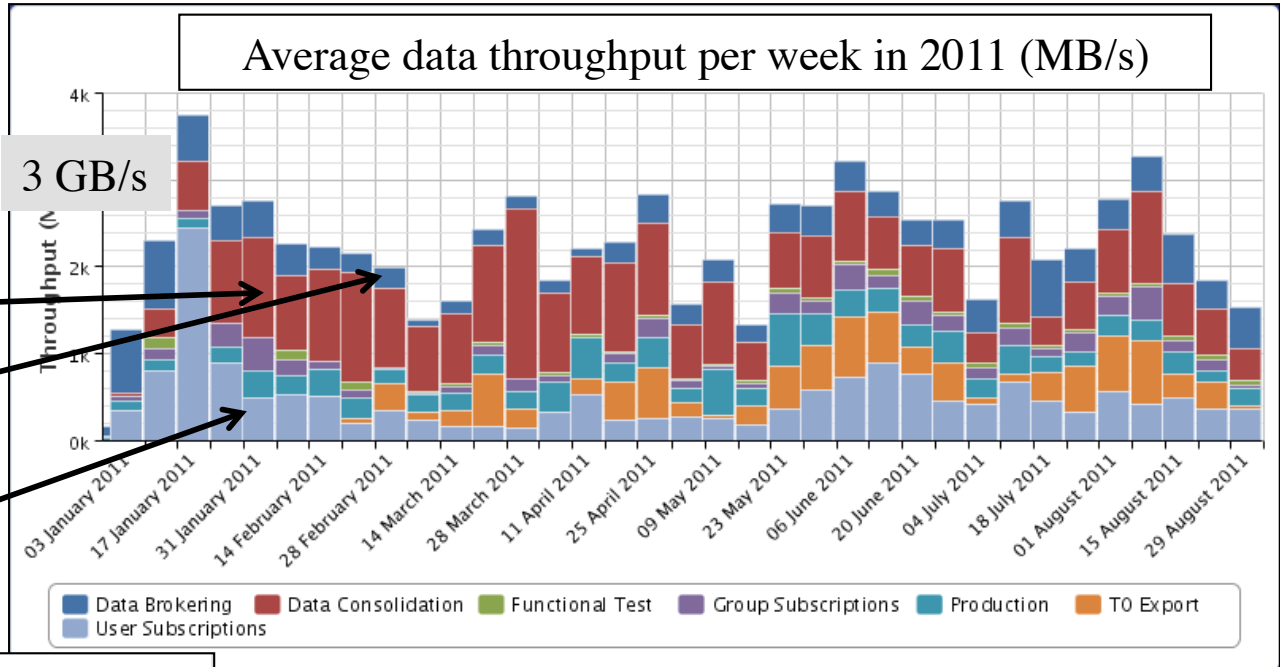


Distributed data management: DDM/DQ2

- Data are transferred around the world steadily at high rates (the Grid never sleeps!)

- Delicate balance between

- Pre-placement
- Dynamic data placement
 - With automatic caching and cleaning
- User requests



	TOTAL-	CA+	CERN+	DE+	ES+	FR+	IT+	ND+	NL+	TW+	UK+	US+
TOTAL-	93 % 2 GB/s	94 % 109 MB/s	91 % 414 MB/s	93 % 331 MB/s	91 % 94 MB/s	93 % 259 MB/s	89 % 119 MB/s	93 % 103 MB/s	89 % 178 MB/s	89 % 51 MB/s	92 % 217 MB/s	96 % 430 MB/s
CA+	96 % 116 MB/s	96 % 58 MB/s	95 % 17 MB/s	98 % 7 MB/s	96 % 2 MB/s	98 % 4 MB/s	98 % 3 MB/s	97 % 2 MB/s	97 % 5 MB/s	92 % 1 MB/s	95 % 4 MB/s	99 % 13 MB/s
CERN+	90 % 204 MB/s	91 % 6 MB/s	87 % 78 MB/s	89 % 22 MB/s	81 % 5 MB/s	95 % 16 MB/s	84 % 8 MB/s	91 % 10 MB/s	88 % 10 MB/s	92 % 6 MB/s	84 % 15 MB/s	95 % 28 MB/s
DE+	92 % 380 MB/s	90 % 9 MB/s	94 % 45 MB/s	92 % 201 MB/s	87 % 9 MB/s	95 % 9 MB/s	92 % 14 MB/s	93 % 18 MB/s	91 % 18 MB/s	86 % 5 MB/s	91 % 15 MB/s	93 % 37 MB/s
ES+	93 % 111 MB/s	91 % 3 MB/s	91 % 16 MB/s	95 % 7 MB/s	93 % 53 MB/s	92 % 6 MB/s	86 % 3 MB/s	96 % 3 MB/s	86 % 4 MB/s	86 % 2 MB/s	90 % 5 MB/s	94 % 9 MB/s
FR+	93 % 308 MB/s	87 % 6 MB/s	96 % 48 MB/s	94 % 19 MB/s	94 % 6 MB/s	92 % 148 MB/s	93 % 9 MB/s	98 % 9 MB/s	91 % 14 MB/s	90 % 5 MB/s	90 % 15 MB/s	96 % 30 MB/s
IT+	90 % 138 MB/s	87 % 4 MB/s	92 % 35 MB/s	94 % 10 MB/s	91 % 3 MB/s	91 % 6 MB/s	89 % 54 MB/s	95 % 4 MB/s	91 % 5 MB/s	78 % 3 MB/s	84 % 4 MB/s	94 % 11 MB/s
ND+	93 % 87 MB/s	97 % 2 MB/s	94 % 20 MB/s	97 % 5 MB/s	95 % 1 MB/s	95 % 3 MB/s	85 % 2 MB/s	92 % 38 MB/s	92 % 3 MB/s	95 % 689 kB/s	93 % 3 MB/s	96 % 9 MB/s
NL+	89 % 185 MB/s	85 % 4 MB/s	96 % 35 MB/s	92 % 14 MB/s	72 % 3 MB/s	91 % 8 MB/s	84 % 4 MB/s	91 % 5 MB/s	88 % 91 MB/s	89 % 2 MB/s	93 % 6 MB/s	89 % 14 MB/s
TW+	92 % 54 MB/s	94 % 1 MB/s	91 % 13 MB/s	94 % 4 MB/s	91 % 829 kB/s	97 % 4 MB/s	89 % 1 MB/s	91 % 2 MB/s	94 % 1 MB/s	90 % 20 MB/s	92 % 2 MB/s	93 % 4 MB/s
UK+	93 % 226 MB/s	85 % 4 MB/s	91 % 34 MB/s	92 % 10 MB/s	73 % 4 MB/s	92 % 11 MB/s	88 % 6 MB/s	92 % 5 MB/s	90 % 8 MB/s	83 % 3 MB/s	94 % 126 MB/s	91 % 17 MB/s
US+	95 % 497 MB/s	92 % 13 MB/s	91 % 74 MB/s	94 % 31 MB/s	92 % 9 MB/s	95 % 36 MB/s	90 % 15 MB/s	95 % 17 MB/s	92 % 19 MB/s	85 % 5 MB/s	92 % 22 MB/s	96 % 257 MB/s

- Excellent data transfer efficiency achieved overall

- 93% average success rate in 2011

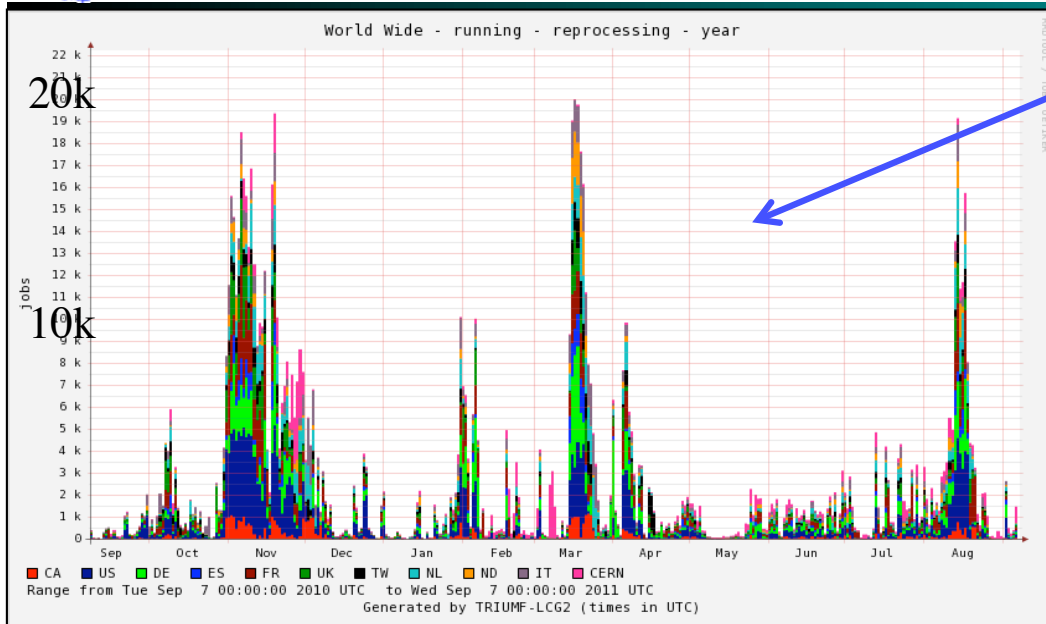
➤ First retry always succeeds

- Users can direct the outputs of their analysis jobs to their "home" on the Grid

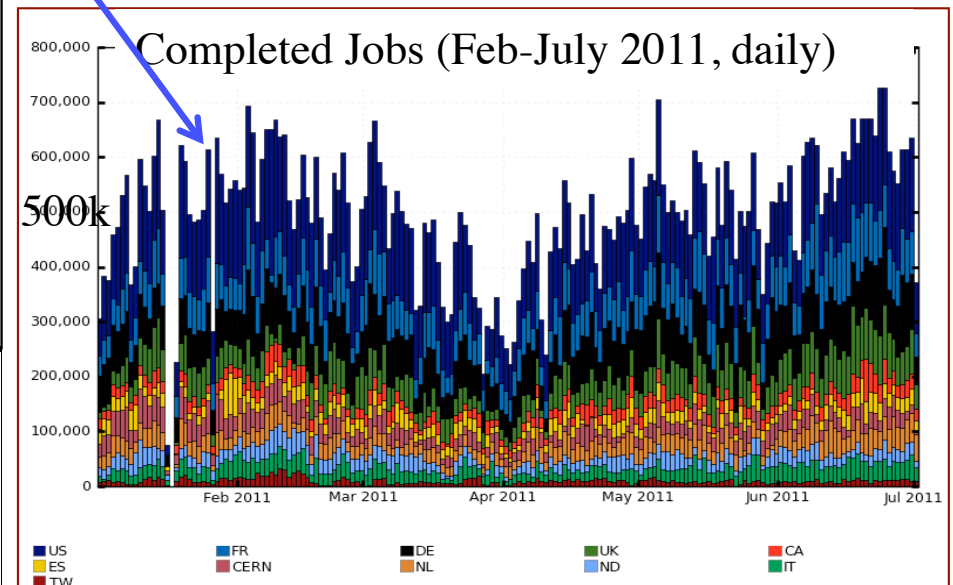
- Asynchronous transfer (plus retry) guarantees success in shortest time



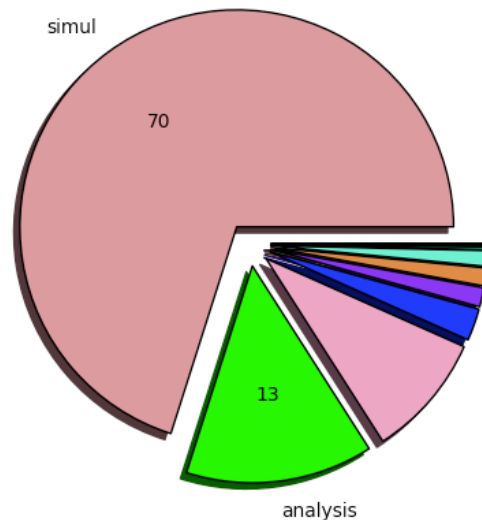
Reprocessing and simulation production



- One year of reprocessing campaigns
- Over 500,000 simulation production and data reprocessing jobs/day on the Grid



CPU share by activity
(April-August 2011):
83% simulation
14% analysis
1% validation
2% reprocessing



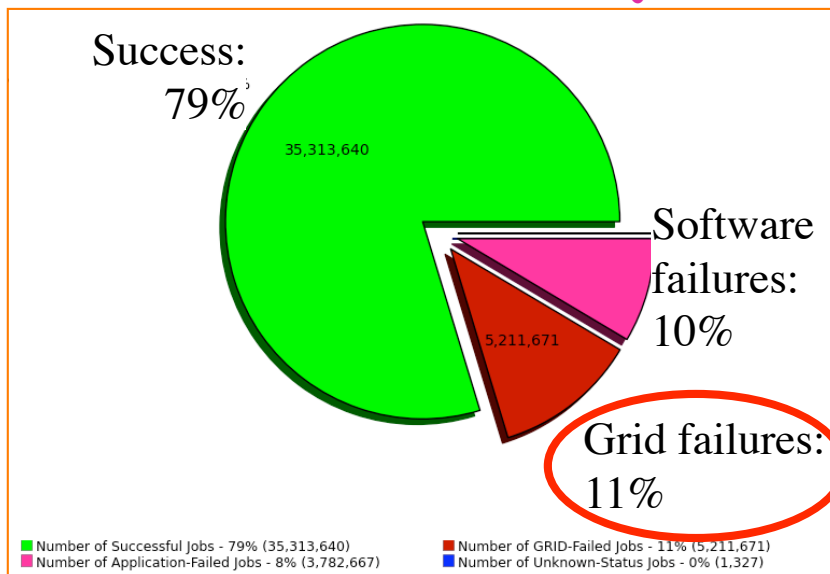
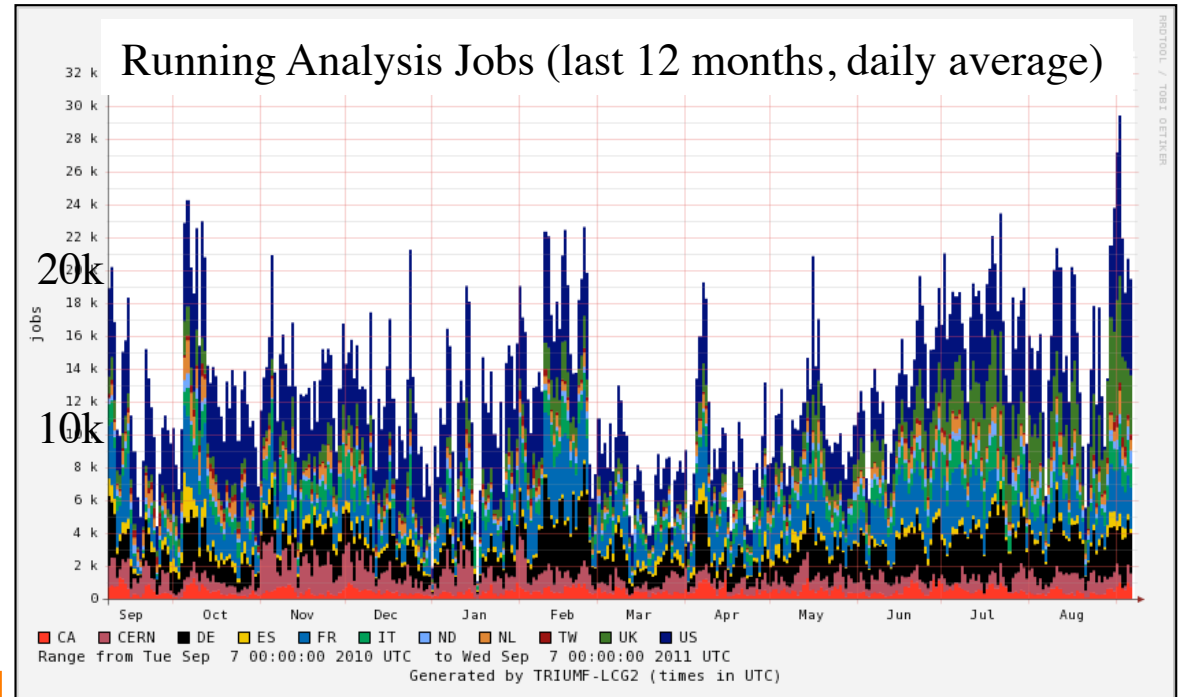
simul (70.09)	analysis (13.97)	pile (9.28)	reprocessing (2.38)
validation (1.43)	reco (1.32)	evgen (1.16)	merge (0.23)

- ~80k jobs running simultaneously
- Analysis tasks are 50% of the jobs but use 14% of total available CPU time
 - Re-run frequently to produce newer n-tuples



Distributed analysis on the Grid

- Analysis jobs run world-wide
 - Jobs go to the data as much as possible
- Grid reliability issues...
 - automatic exclusion (and re-inclusion) of analysis queues that do not perform well, measured via automatic HammerCloud test jobs

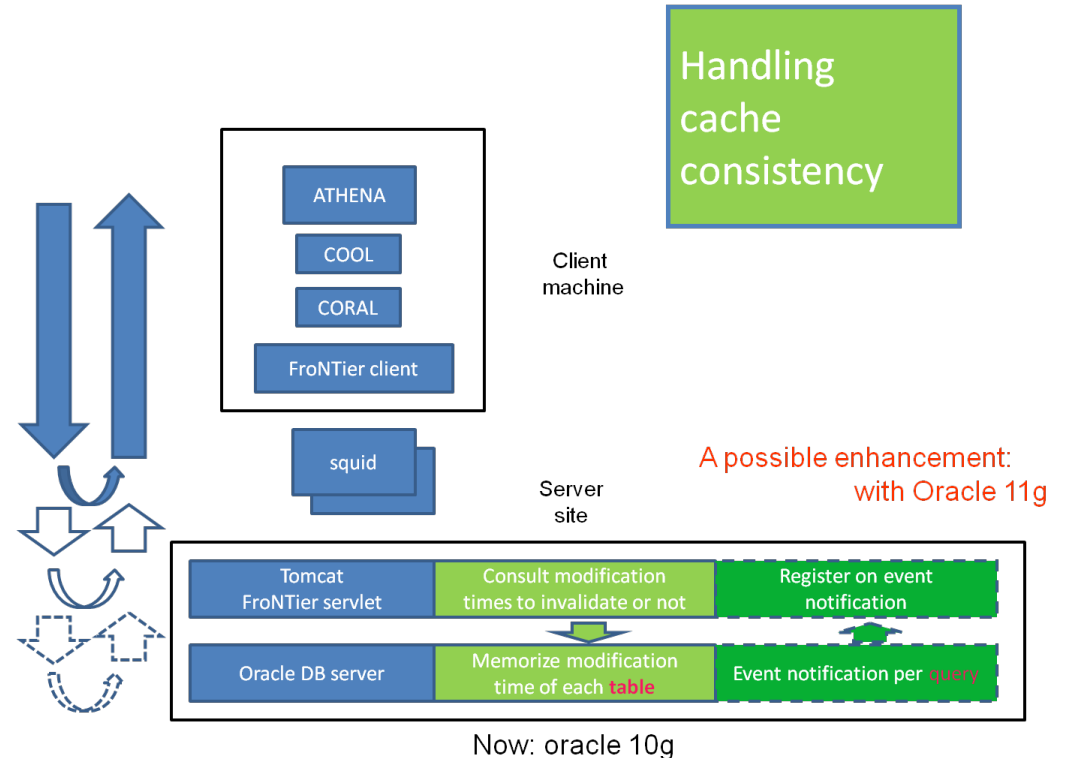


- Work in progress to improve task efficiency (and user happiness)
 - Merging of output files
 - Automatic retrieval of jobs that fail for well-defined Grid-related reasons
 - Improved analysis tasks book-keeping, to better keep track of the whole workflow



Conditions Databases

- Frontier deployed in 2009 to enable distributed access to the conditions DB
- Flow of database data:
 - Oracle: CERN online -> CERN offline -> 3D (BNL, TRIUMF, RAL, KIT, IN2P3-CC)
 - Frontier server at each of the above sites connects to local Oracle database
 - Local Squid contacts nearest Frontier server
 - With failover to next-to-nearest



Map of installed Squids



- Frontier reduces considerably the access time to DB data from remote sites
- It is particularly important for sites with low bandwidth and high latency towards Oracle servers
 - but in reality good for all sites!
- Now under test also for Tier-0 and CAF at CERN



Web access to software and databases

- Placing web servers and a sequence of cascading caches in front of major services is a very cost-effective way to provide robust and distributed read access to popular information
- **Frontier** shields Oracle servers from overloads due to repeated identical queries from jobs accessing the same conditions data (same runs or time intervals)
 - Data are cached in the Frontier launchpad (server) and the site Squid before getting to the worker node
 - Access times for conditions data decreased from several minutes to a few seconds for Tier-2s with large latency times to the nearest Tier-1
- **CVMFS** is a web-based file system with Squid caches at each site
 - No need to pre-install all software releases on each site; software is pulled and cached locally when used
 - Conditions data files are also available through CVMFS, saving space on the SE
 - Deployment for ATLAS in rapid progress (1/4 of sites now, aim for completion by the end of 2011)



Evolution of the Computing Model in 2011

- ✓ Break the cloud* boundaries
 - Introduce flexibility in data distribution and job assignment
- ✓ Allow inter-cloud direct Tier-1 \leftrightarrow Tier-2 and Tier-2 \leftrightarrow Tier-2 transfers according to network connectivity
 - For data placement, user subscriptions and job I/O
- ✓ Allow job distribution from Tier-1s to Tier-2s in other clouds
 - Output files are then collected back to the original Tier-1 (of course)
- ✓ Reduce the number of data replicas to have more data on disk
- ✓ Introduce dynamic data replication and deletion based on dataset popularity
- ✓ Reduce the multiplicity of Oracle database servers and equip all remaining ones with Frontier web servers
- Integrate all 11 LFCs into a single catalogue at CERN (work in progress)
 - No longer one catalogue for each cloud
- Move towards using CVMFS (web-based file system) for software release and conditions data files distribution (deployment in progress)
- ★ (An ATLAS Grid cloud includes a Tier-1 and all associated Tier-2/3s)



ADC R&D projects

- All software projects are continuously evolving
 - Following the technological evolution in the IT domain
- ATLAS software also evolves
 - Adiabatic evolution takes place all the time
 - From time to time some major re-thinking is needed
- A few R&D projects were started within the ATLAS Distributed Computing (ADC) community earlier this year to take advantage of the experience of the first years of LHC data-taking:
 - DDM data organisation, catalogues, services
 - Data placement and file, event, sub-event caching
 - Use of NoSQL databases for some of the ADC tools
 - Interfacing to cloud computing technology in addition to Grid middleware
 - Using many-core processors



Erevan - 29 September 2011

Summary and Outlook

- The ATLAS Distributed Computing infrastructure is working thanks to many efforts in preparation and many people working in operations
- We are able to
 - Process, distribute, and reprocess the data
 - Analyse the data
 - Provide support to our large community
- As we get experience with *reality* we are looking at the evolution of the model and our implementations, e.g.
 - Less-strict cloud model
 - Better data distribution for analysis
 - Improved support for analysis
- A number of R&D projects for future distributed computing tools have been started this year in order to improve further the performance and be able to scale up to higher data and job rates