Distributed ATLAS computing activities in Iberia

Helmut Wolters (LIP Coimbra)

X. Espinal[1], H. Wolters[6], E. Acción[1], G. Amorós[3], G. Bernabeu[1], G. Borges[5], A. Bria[1], C. Borrego[2], M. Campos[2], J. Carvalho[6], M. David[5], M. Delfino[1], N. Dias[5], F. Fassi[3], A. Fernández[3], P. Fernández[4], J. Gomes[5], S. González de la Hoz[3], M. Kaci[3], A. Lamas[3], L. March[3], F. Martinez[1], J.P. Martins[5], G. Merino[1], M. Montecelo[5], L. Muñoz[4], J. Nadal[2], M. Oliveira[6] , A. Pacheco[2], J.J. Pardo[4], J. del Peso[4], J. Salt[3], J. Sánchez[3]

[1]PIC – UAB Barcelona, Spain
[2]IFAE – UAB Barcelona, Spain
[3]IFIC – CSIC Valencia, Spain
[4]UAM Madrid, Spain
[5]LIP, Lisboa, Portugal
[6]LIP, Coimbra, Portugal

Outline

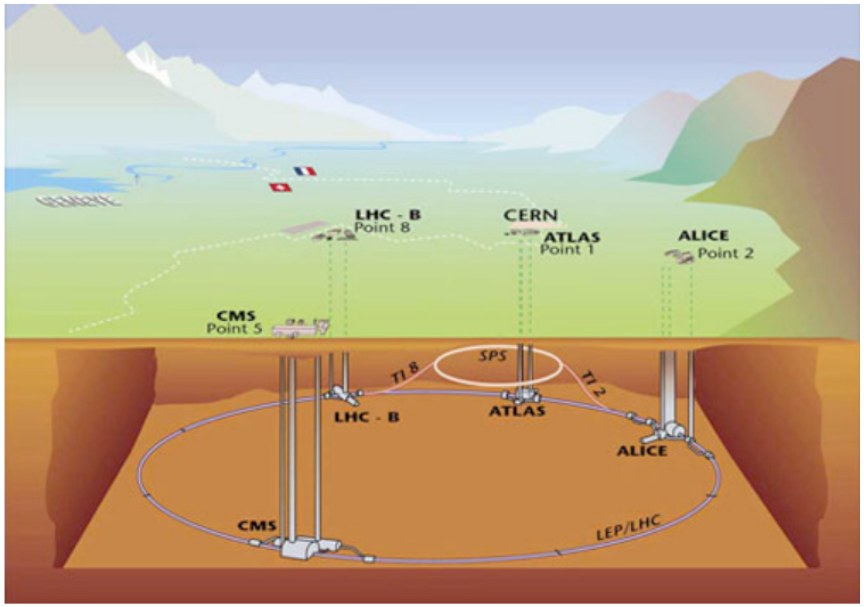
- The ATLAS experiment
- The Iberian cloud
- The ATLAS Production System
- Test results
- Conclusion

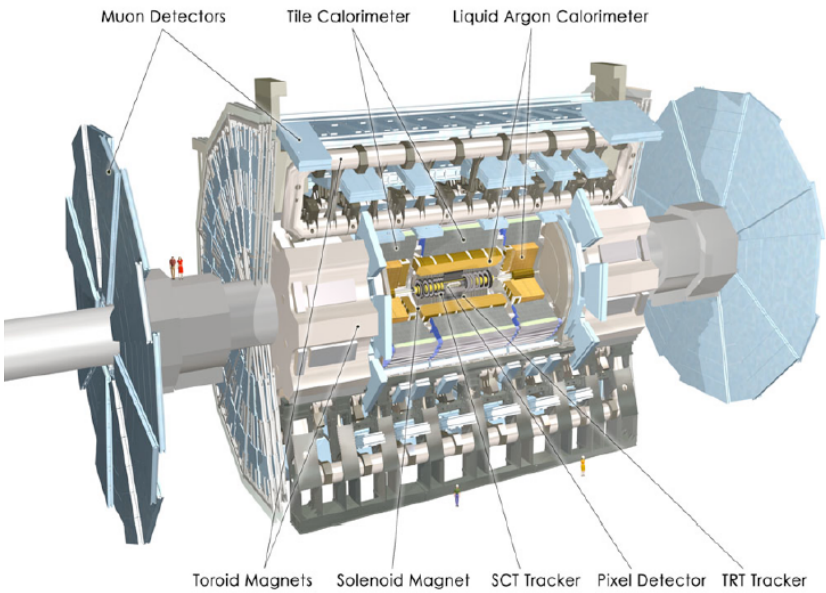
The Atlas Experiment

- The Large Hadron Collider (LHC) will be worlds largest and most powerful particle accelerator.

- Installed in an underground tunnel of 27 km in circumference astride the border between Switzerland and France.

- Will produce 800 million proton-proton collisions per second, with 14TeV center of mass energy.





- The ATLAS (A Toroidal LHC Apparatus) detector measures:
  - Diameter: 25 m
  - Barrel toroid length: 26 m
  - Endcap end-wall chamber span: 46 m
  - Overall weight: 7000 Tons

- ATLAS is one of the four LHC detectors, devoted to the study of high-energy proton-proton collisions and heavy ions.

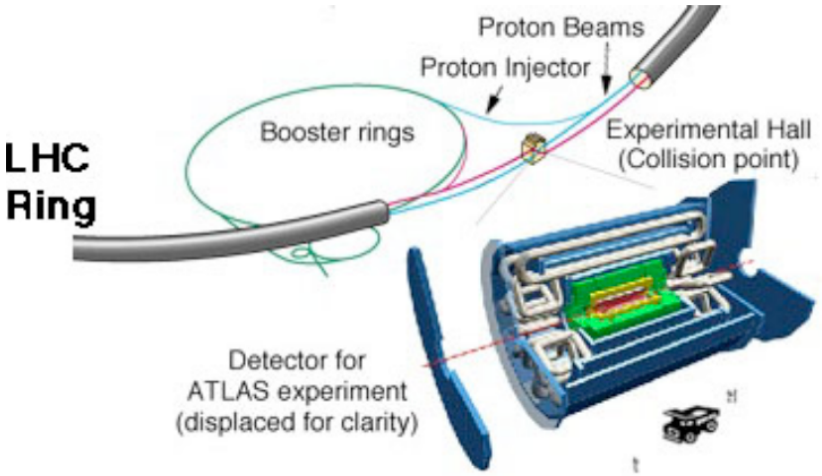
(X. Espinal — 3rd IEEE International Conference on e-Science and Grid Computing - Bangalore 11-13 Dec. 2007)

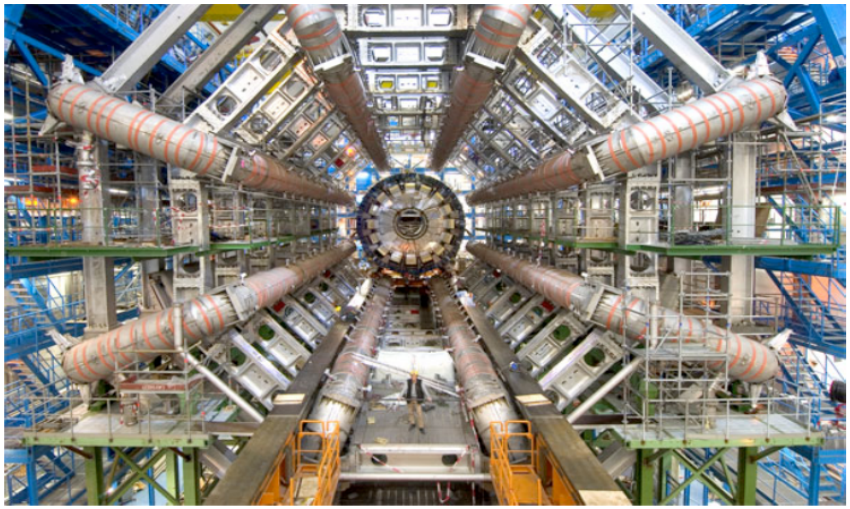
The Atlas Experiment

- The offline computing will have to deal with an output event rate of 200 Hz i.e $2x10^9$ events per year with an average event size of 1.6 Mbyte (320 MB/s).

- ATLAS will produce an amount of data of about 10 Pb per year, to be analyzed by ~2000 physicists from all over the world.

- The design and construction of an experiment like ATLAS requires a large amount of simulated data in order to optimize the detector design, estimate physics performance, and test the software and computing infrastructure.





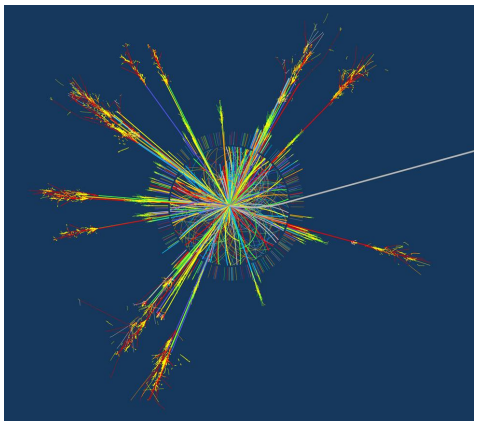
- The funds, electrical power, and human resources necessary for a single, all-purpose computing site would be too great for one laboratory.

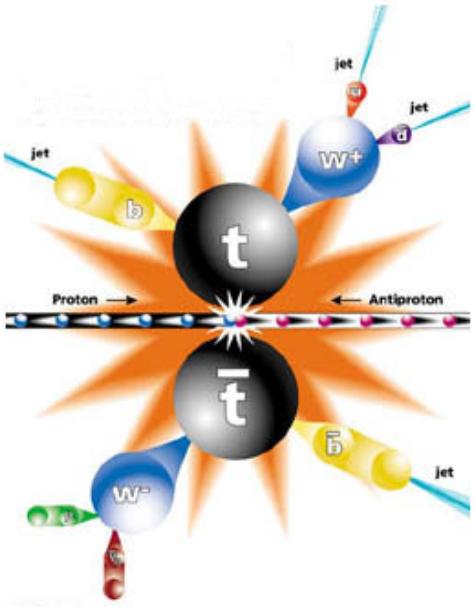
- Physicists and computer scientists create a grid-computing system for the experiments, in which more than 100 small and large computing centers share the responsibility for storing, generating, and processing the data.

- Monte Carlo simulated production is performed all over the world both at large computer centers, called Tier-1s, and at smaller sites, called Tier-2s, as well as in institute or university sites, called Tier-3s (ATLAS tiered structure).

(X. Espinal — 3rd IEEE International Conference on e-Science and Grid Computing - Bangalore 11-13 Dec. 2007)

The Atlas Experiment



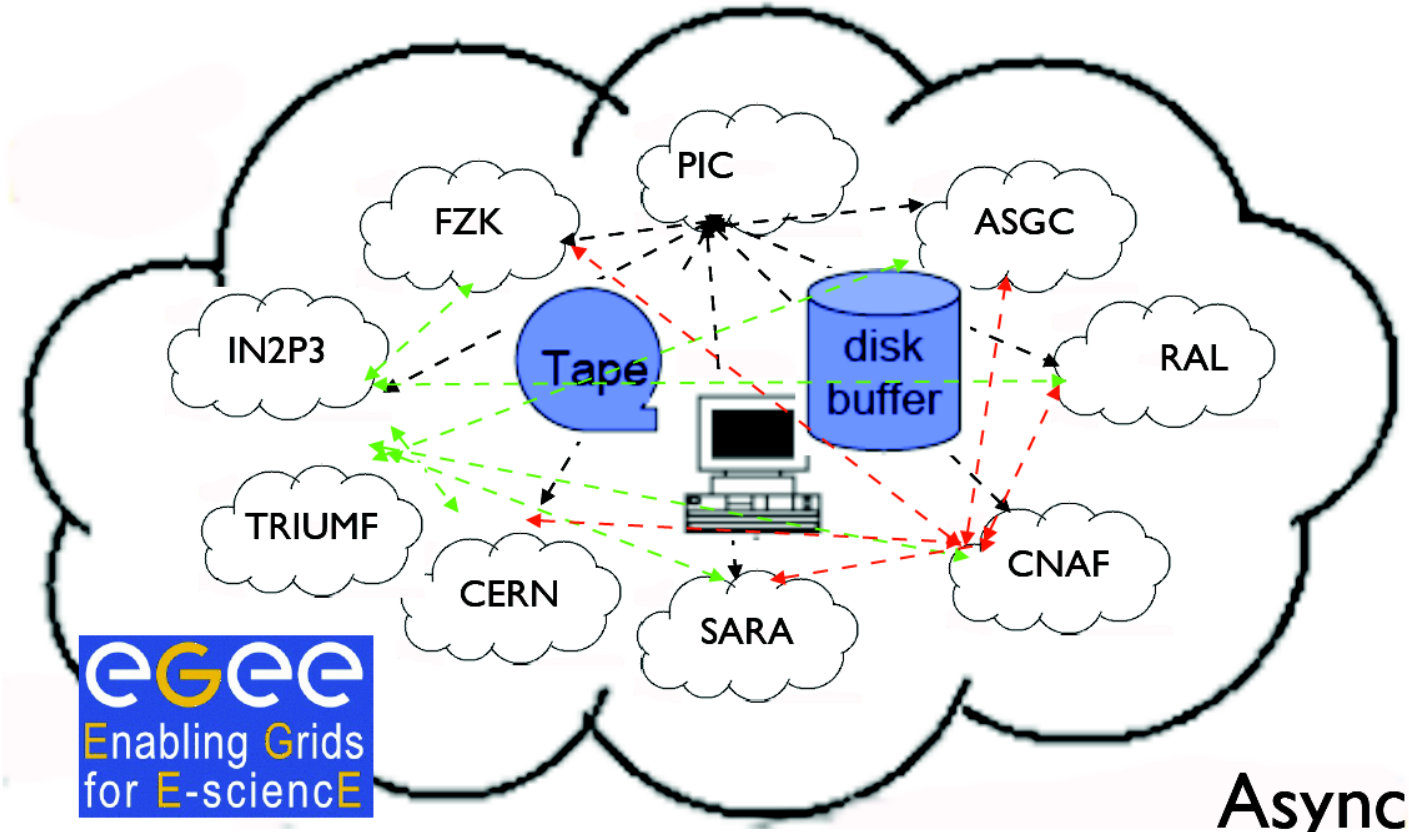


GRID computing

GRID is used to solve problems of data simulation, storage and analysis
Data per year: $\approx$ 10 PetaBytes

- event generation
- simulation of what happens in the detector
- reconstruction of an event from the signals in the detector

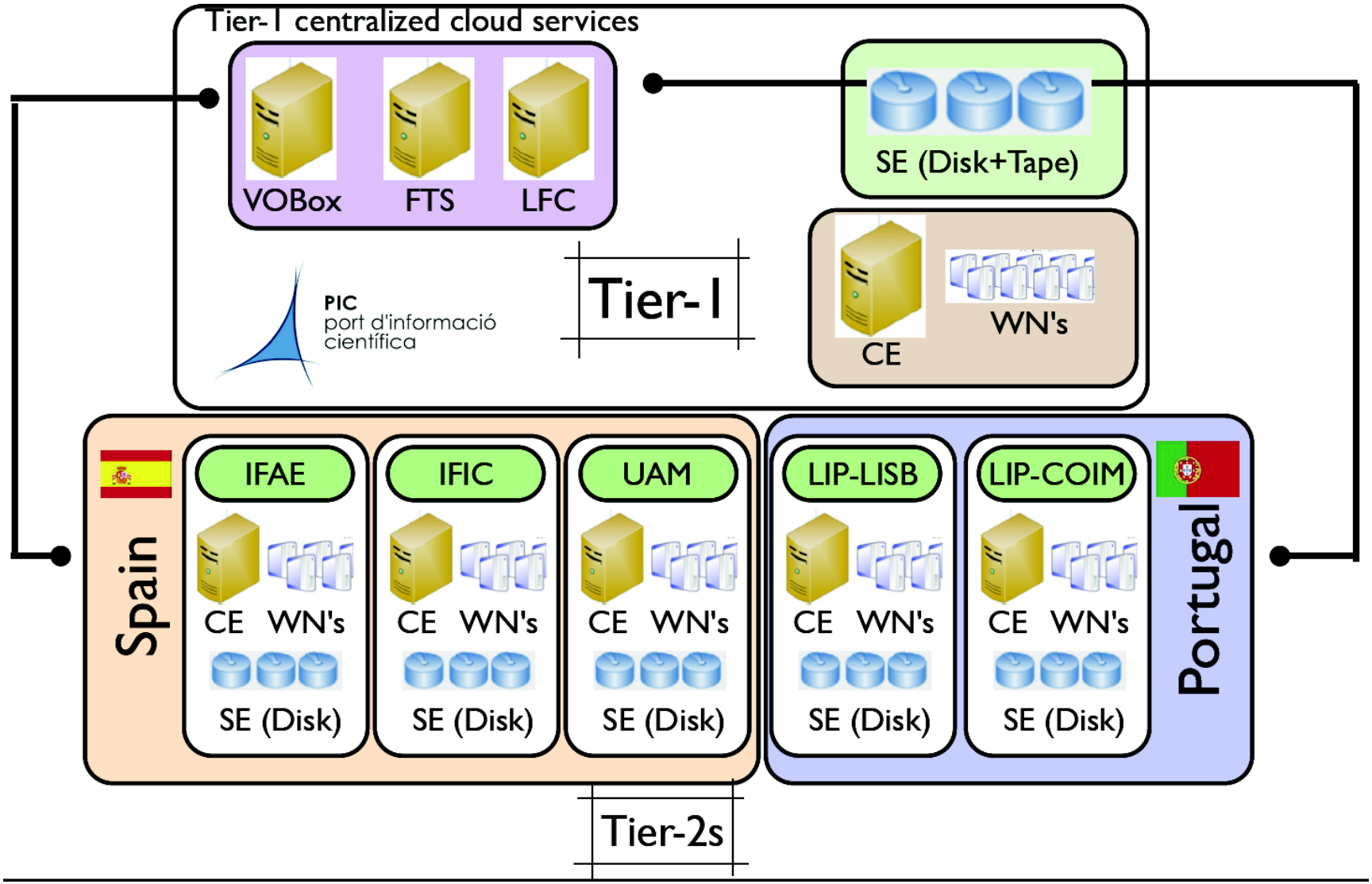
The Atlas Cloud Model



Asynchronous
Data
Movement

(X. Espinal — 3rd IEEE International Conference on e-Science and Grid Computing - Bangalore 11-13 Dec. 2007)

The Iberian Cloud around the Tier1 at PIC



(X. Espinal — 3rd IEEE International Conference on e-Science and Grid Computing - Bangalore 11-13 Dec. 2007)

Tier1 at PIC Barcelona

Data Flow:

- offers storage and processing resources to three LHC experiments: ATLAS, CMS and LHCb
- LHC experiments will store a copy of the collected data from the accelerator at CERN and dispatch a secondary copy to the Tier-1 centers, in order to guarantee the conservation and integrity of the data.
- $\approx$ 5% of the raw data from the detectors will be stored at PIC.
- Optical Private Network (OPN) Tier0 (CERN) $\leftrightarrow$ Tier1's
- $>$ 3 PetaBytes in/out PIC in 2007

Tier1 at PIC Barcelona

Data Processing:

- PIC will provide the infrastructure for data processing, as the raw data stored will be reprocessed several times per year with new parameters, as calibration and alignment constants improve.
- PIC has a cluster which will deliver 1500 kspecint2000 of CPU power (June-2008), which can be seen as 350 double core CPUs
- These reprocessing tasks will access the new calibration and alignment constants through the LHC Distributed Data Base Services, the LCG-3D (located at CERN and at the Tier-1s).
- PIC is running as of today more than 600 jobs simultaneously and finishing more than 1000/day

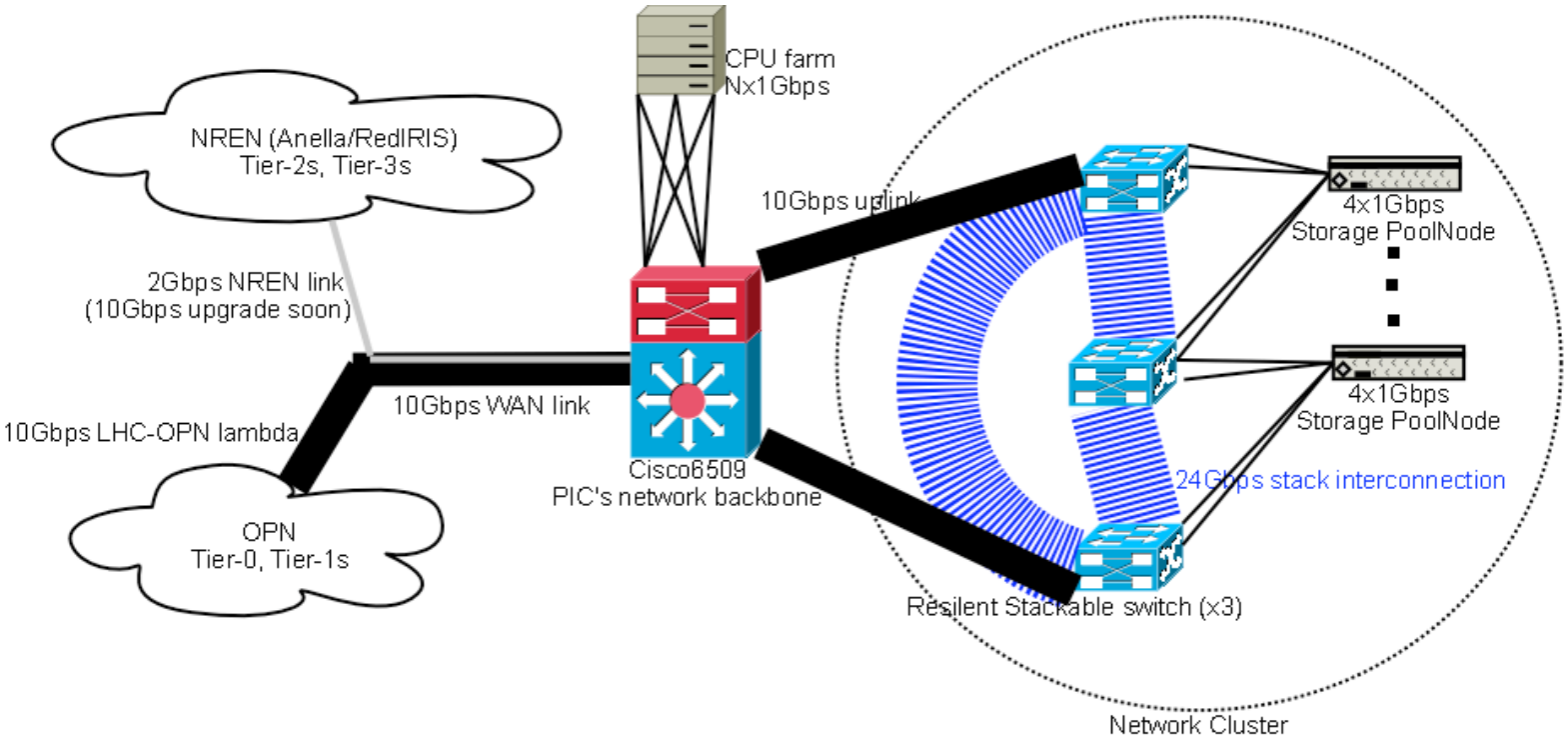
Tier1 at PIC Barcelona

Data Storage:

- experiments do need large, reliable and scalable storage services.
- to serve the data at the required speed in order to maximize the efficiency of the cluster,
- to have a scalable system which can easily grow in parallel with the LHC demands without penalizing the performance.
- Multi-Gigabit Ethernet network architecture, specially designed to enhance high speed data movement between WAN (Tier-0, Tier-1s, Tier-2s) and LAN (CPU farm)
- dCache storage system

Tier1 at PIC Barcelona

Network Architecture



Each LAN cluster has up to 24Gbps available bandwith for pool2pool data replication and 20Gbps (upgradable to 30Gbps) for data movement with the outside (WAN and other LAN clusters)

Tier2s in PIC cloud

Federated Spanish Tier2

- **IFIC: Valencia (coordinator)**
  - 126 Worker Nodes
  - 5 disk servers with a capacity of 36 TB
  - tape robot with a potential capacity of 140 TB + 5 TB disk
- **IFAE: Barcelona**
  - 6 Worker Nodes
  - 80TB of disk capacity
- **UAM: Madrid**
  - 133 Worker Nodes
  - 63 TB (update to 120 TB soon)

Tier2s in PIC cloud

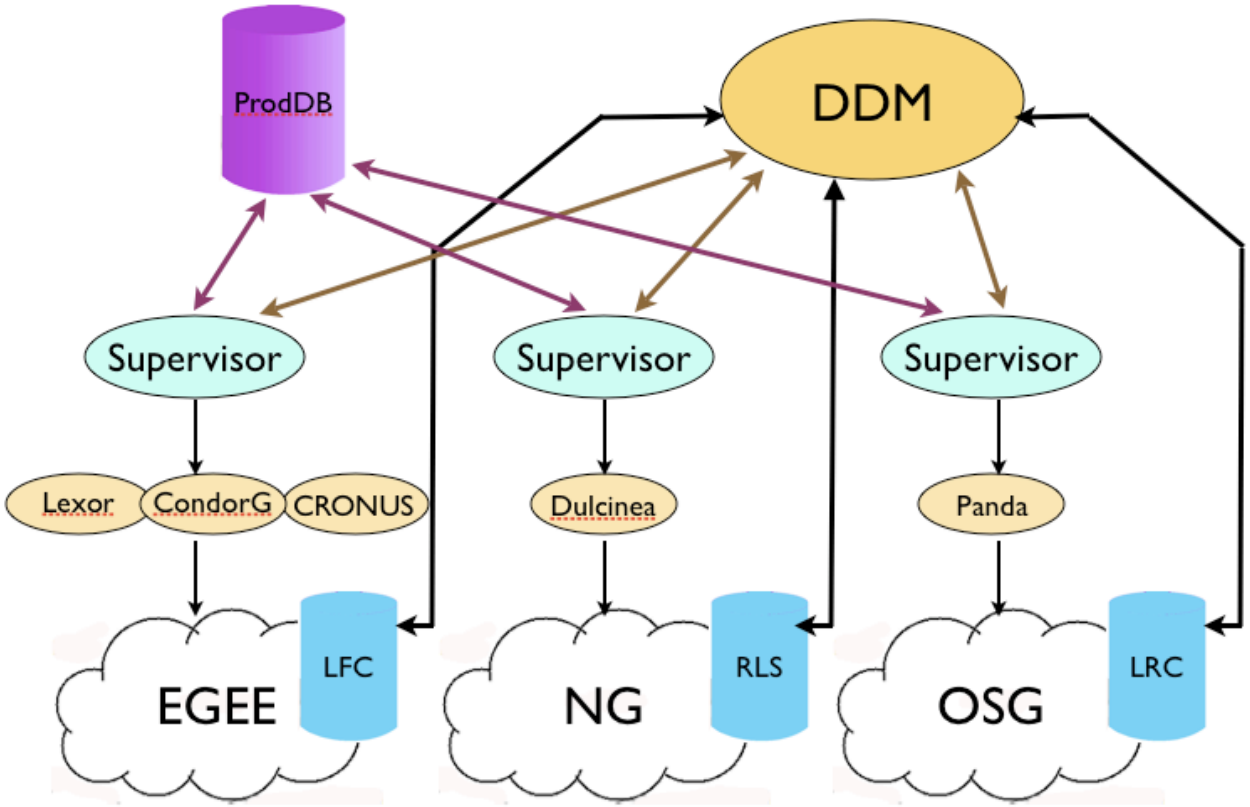
Federated Portuguese Tier2

- **LIP Lisbon:**
  - 80 CPU cores
  - 21 TB storage, 4 TB for ATLAS
  - upgrade to 100 TB soon
- **LIP Coimbra:**
  - 85 Worker nodes (fairshare for ATLAS 40%) (upgrade to 130 soon)
  - 2.8 TB storage for ATLAS (upgrade do 65TB soon)
- **The whole Tier2 will be further upgraded by additional**
  - 400 CPUs
  - 400 TB storage
  - $\approx$ 2/3 Lisbon, 1/3 Coimbra, considerable part already end of May

The ATLAS production system

- Simulated production jobs are fairly different from the user specific jobs, not planned or validated.
- It is formed from several individual elements providing the required functionality for: job submission, tracking, recovery and validation.
- The individual elements of the production system are
  - Common database for the production jobs (ProdDB),
  - Data management system (DDM),
  - Common Supervisor (Eowyn),
  - Executors.

Former Atlas Production System Schema



- several types of executor and file catalogues for each one of the Grids
- since beginning of 2008 moving to a common framework with **one single executor**, based in the pilot jobs schema, and **a single file catalogue** (LFC) **for all Grids**.

The ATLAS production system

- It is independent of the implementation of the underlying Grid infrastructure that is actually used.
- The ATLAS production system provides a common framework where any Grid flavour may be integrated.
- It is formed from several individual elements which provide the required functionality for the submission, tracking, recovery and validation of the jobs.

The ATLAS production system

- *ProdDB*: single logical production database for EGEE, OSG and NG
  - holds the entries of the jobs requested by the physics groups
  - keeps the information for the operative workflow:
    - *Job transformation*: describes a particular combination of executables (Athena) and the ATLAS software release.
    - *Job definitions*: points to its associated job transformation. Keeps track of the current attempt of executing this job (lastAttempt), supervisor, priority, etc.
    - *Job execution*: zero, one, or more records corresponding to each attempt of executing the job, start- and end-time, resources consumed, where outputs were stored, etc.

The ATLAS production system

- *Supervisor*: takes free jobs from the production database and hands them on to one of the executors to which it is connected, that finally send the jobs over the Grid.
- *Executors*: retrieves the jobs, creates the wrapper files and submits them to the Grid, taking into account the free slots in the sites, installed software, etc.
- *Software installation*: is based on the Lightweight Job Submission Framework for Installation (LJSFi). This system is able to automatically discover, check, install, test and tag the full set of resources made available in the LCG/EGEE sites to the ATLAS Virtual Organization in a few hours.

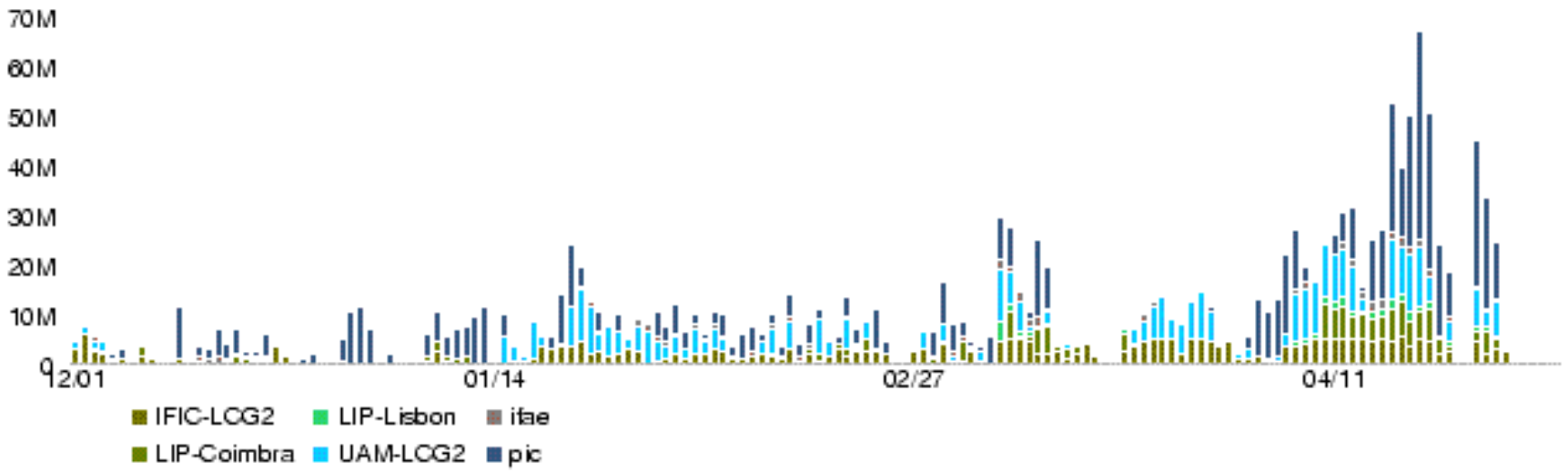
The ATLAS production system

The pilot job schema

- *Pilot Jobs* check the correct environment for running the jobs, and once verified it transfers a job from the Job database, where the jobs are defined and waiting for green light from pilots, to a Worker Node.
  - deployed in December at the Iberian cloud
  - showing a stable and very promising resources occupancy and walltime efficiencies.
  - There are basically three types of jobs defined in the ATLAS production system:
    - event generation,
    - simulation
    - reconstruction
  - Each one of these jobs spends different amounts of time to finish.

Walltime in Cloud

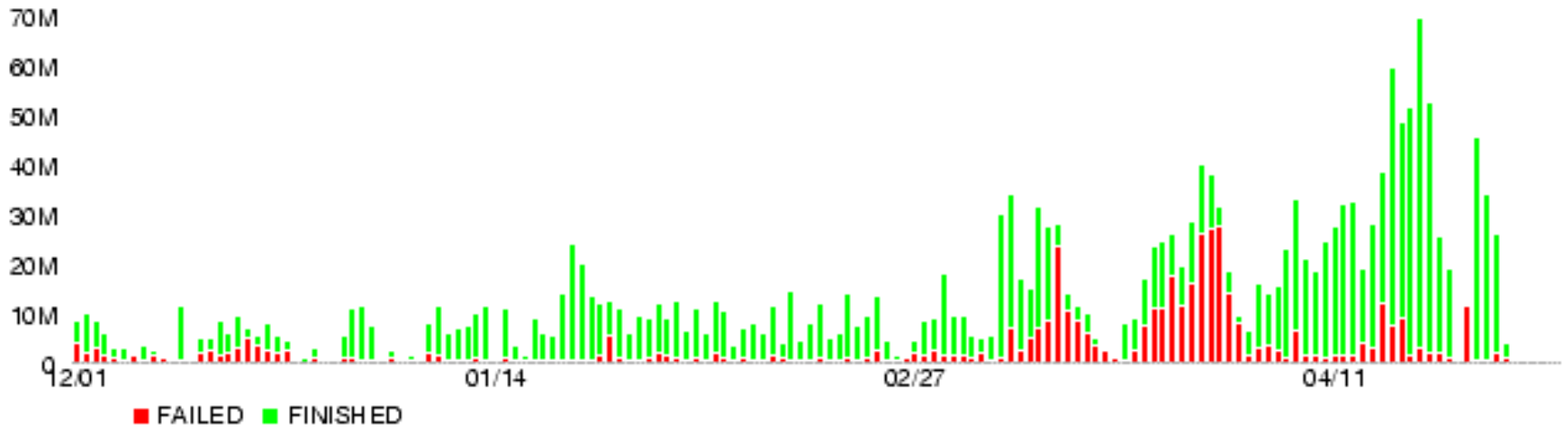
Successfully finished jobs in the Iberian cloud in 2008
[Walltime days]



Demonstrates the stability of the pilot job schema,
which is capable to fill all available resources at the sites.

Walltime efficiency in MC production

Successful/failed jobs in the Iberian cloud in 2008
[Walltime days]



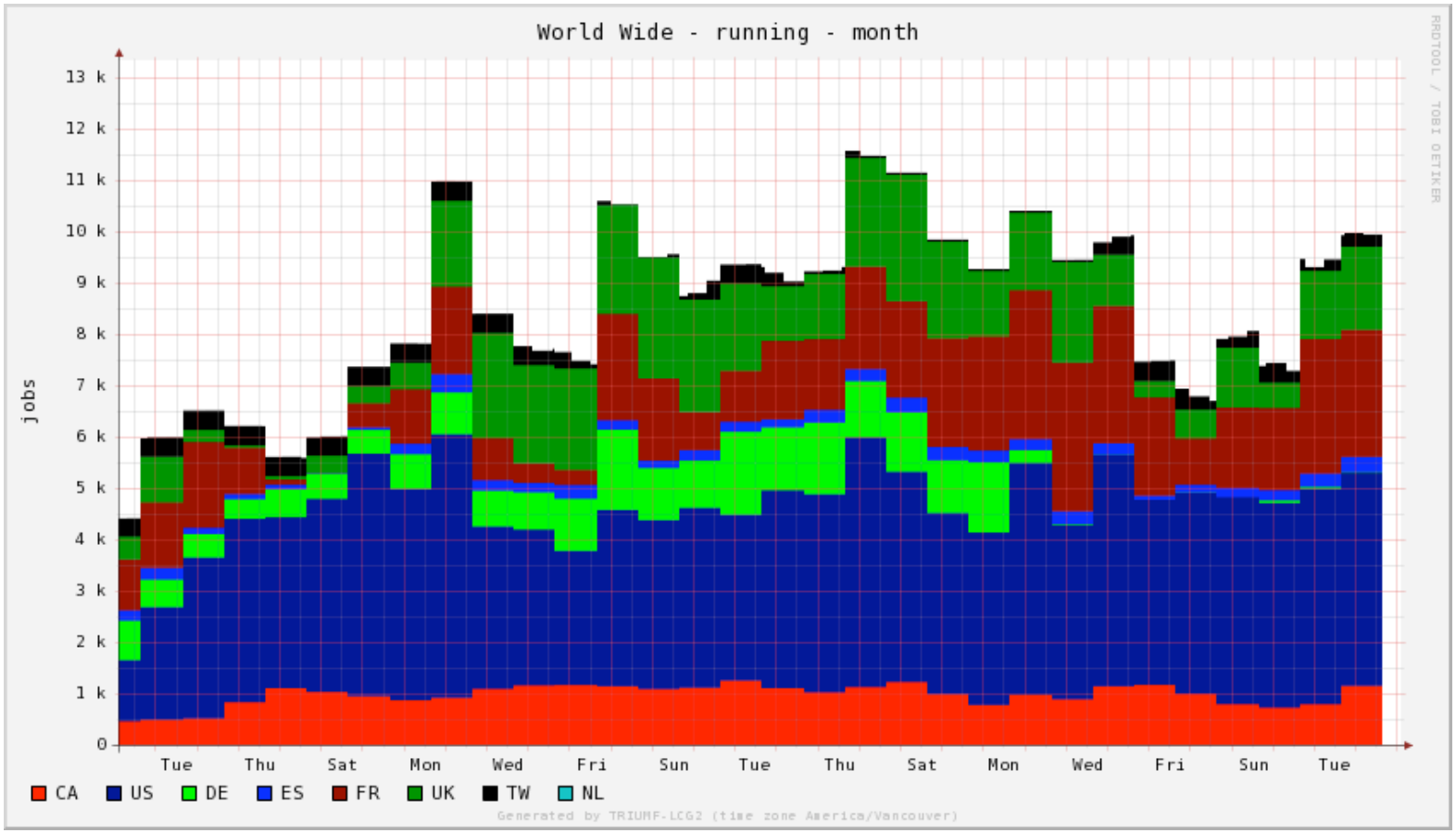
walltime efficiency for the Monte Carlo production during the current year at the Iberian cloud, which is close to 90%

Cloud Statistics

Site	Jobs	Efficiency	Walltime days	Efficiency
PIC	23760	70%	3731	86%
UAM	12244	74%	4569	95%
IFIC	4680	89%	1987	93%
IFAE	3440	82%	768	92%
LIP-COIMBRA	3202	86%	1373	94%
LIP-LISBON	717	30%	310	80%

Statistics for the cloud, starting since the 1st of January 2008 (LIP-LISBON only accounting from 1st of February).

Global number of ATLAS MC jobs finished worldwide



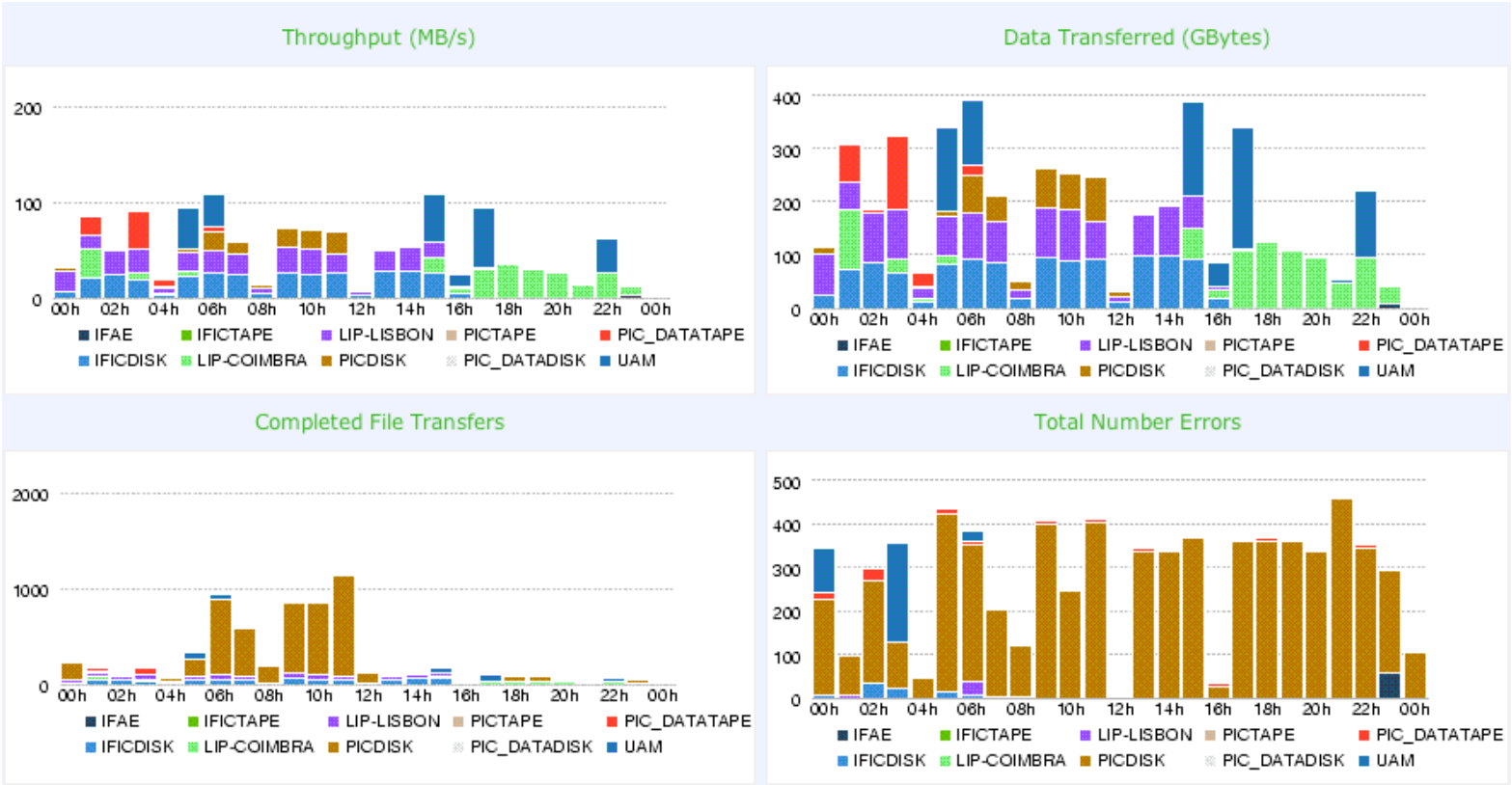
Currently all 10 ATLAS Tier-1s are using pilot job architecture, but very soon all the cloud will be integrated, once the pilot factory using glite-WMS is ready.

Runnings Tests

Common Computing Readiness Challenge - CCRC08

- Since February 2008 all LHC experiments are running tests in parallel
- CCRC-08 run 2 running since the 5th of May
- idea: test the current infrastructure at all tiers with a constant multi-VO exercise
- ATLAS simulated the data produced by the collision at the Sub-farms output level and injected these data into Castor to proceed with calibration and reconstruction.
- Unfortunately the trigger rates have been lowered to 10% of the nominal value, which is 200 Hz. So the amount of data is drastically reduced by this factor.
- After the reconstruction step, data are ready to be exported to the Tier-1 centers.

FDR — Full Dress Rehearsal — data replication



FDR data replication from PIC to all the Tier-2s

- errors are related to source when transferring form Tier-0 to PIC
- once data was consolidated at the Tier-1, it was exported to the Tier-2s without further problems

Transfer statistics

Number of files transferred, efficiencies and throughputs

Site	Files	Datasets	Efficiency	Thr. Peak
IFIC	86	43	100%	42 MB/s
IFAE	50	25	100%	40 MB/s
UAM	50	25	100%	25 MB/s
LIP-LISBON	96	48	100%	22 MB/s
LIP-COIMBRA	90	45	100%	30 MB/s

(results from the CCRC08 run 1 functional tests)

Cloud Efficiency Statistics

Site	Jobs	Efficiency	Walltime days	Efficiency
PIC	33645	70%	4091	86%
UAM	15884	74%	4439	95%
IFIC	4983	89%	1912	93%
IFAE	3922	82%	738	92%
LIP-COIMBRA	3324	86%	1099	94%
LIP-LISBON	3202	30%	256	73%
Cloud Avg.*	12351	80%	2455	92%
Atlas Avg.*	15072	76%	3536	88%

Statistics for the cloud, starting since the 1st of January 2008 to the 1st. of March. The Atlas average comprises the statistics from 122 sites around the world. * LIP-LISBON only accounting from 1st of February, excluded for average calculation due to low statistics.

Conclusions

- ATLAS activities in the Iberian cloud has successfully proved its integration within the ATLAS computing model.
- Simulated events production covered the ATLAS demands and successfully integrated the pilot job schema at all sites since the beginning of the year.
- Data distribution exercises showed good performance and reliability of our storage systems with high efficiency and few latency.
- all sites on the Iberian peninsula are in good shape and ready to fulfill the ATLAS computing demands.
- Data taking starting in November...