XXVII Workshop on Recent Developments in High Energy Physics and Cosmology

Demokritos, Athens, Greece

Spanish ATLAS computing cloud: Facing data taking

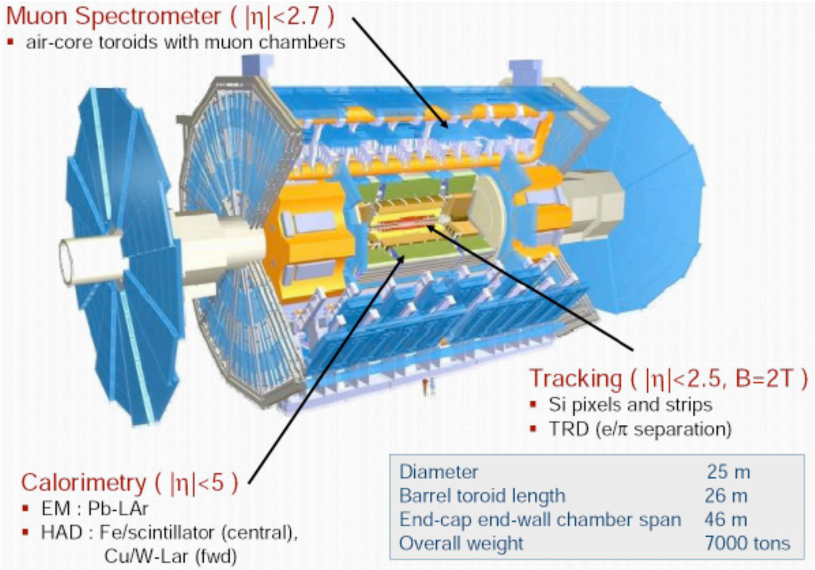
Santiago González de la Hoz,
Instituto de Física Corpuscular
IFIC-Valencia, Spain
(Centro Mixto Universitat de València-CSIC)

On behalf of the Iberian/Spanish ATLAS cloud

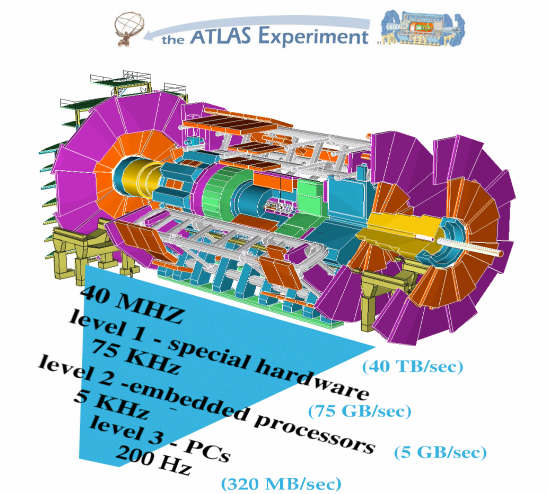
Outline

- The ATLAS Experiment
  - The event data Model
  - The hierarchical computing model
- The Spanish/Iberian cloud for ATLAS
  - Tier1 resources
  - Spanish Distributed Tier2 resources
- Simulated event production
- Data Transfer and Distributed Analysis activities
- Tier3 prototype at IFIC-Valencia
- Conclusions

The ATLAS experiment


Muon Spectrometer (|η|<2.7)
▪ air-core toroids with muon chambers
Tracking (|η|<2.5, B=2T)
▪ Si pixels and strips
▪ TRD (e/π separation)
Calorimetry (|η|<5)
▪ EM : Pb-LAr
▪ HAD : Fe/scintillator (central),
Cu/W-Lar (fwd)

Diameter	25 m
Barrel toroid length	26 m
End-cap end-wall chamber span	46 m
Overall weight	7000 tons

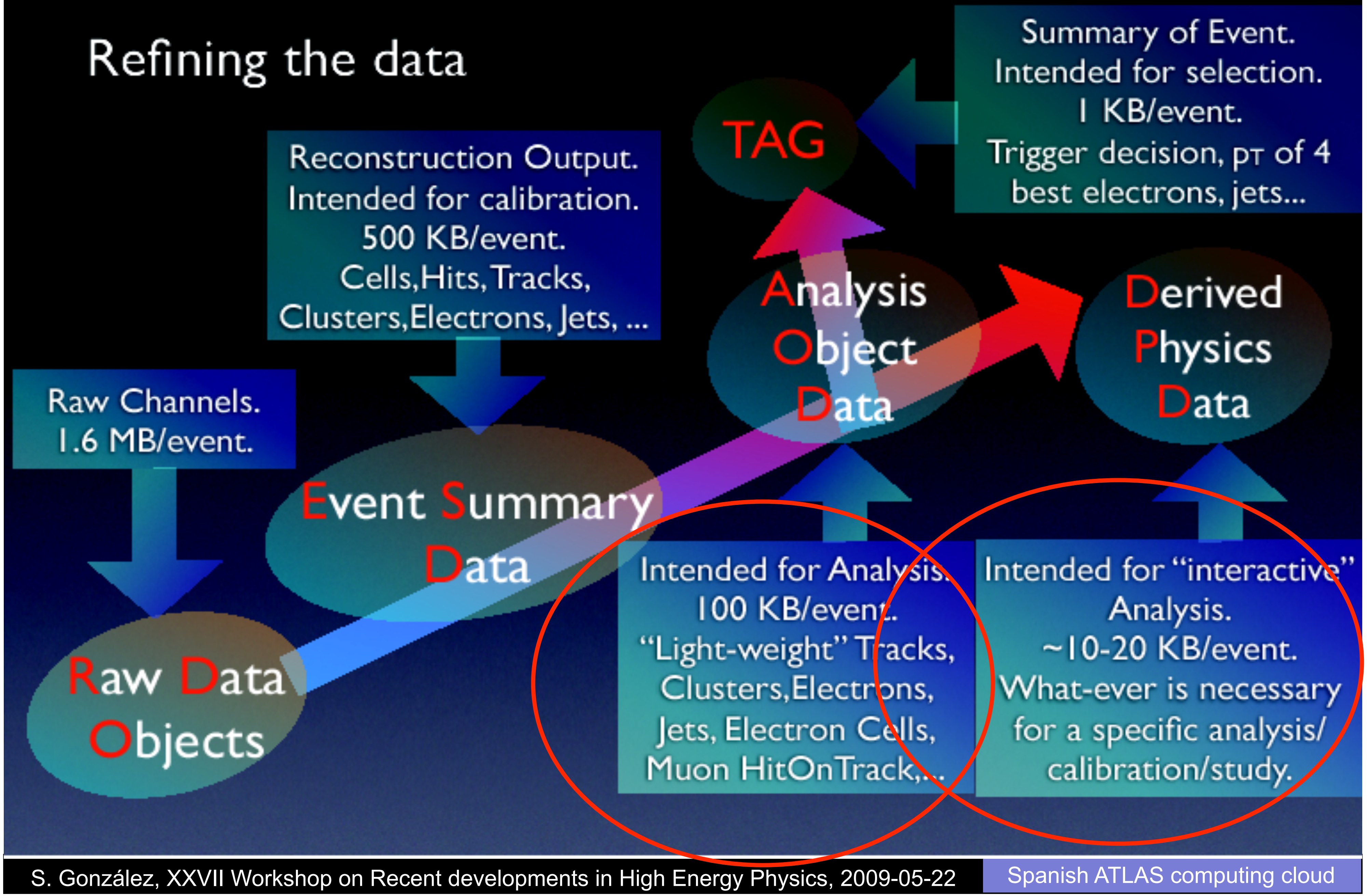


- **The offline computing**:
  - **Output event rate: 200 Hz ~ 10^9 events/year**
  - **Average event size (raw data): 1.6 MB/event**
- **Processing**:
  - **40,000 of today's fastest PCs**
- **Storage**:
  - **raw data recording rate 320 MB/sec**
  - **Accumulating at 5-8 PB/year**

- **A Solution: Grid Technologies**
  - **GRID is used to solve problems of data simulation, storage, reprocessing and analysis.**
  - **Data per year: ≈ Petabytes**
    - **event generation**
    - **simulation of what happens in the detector**
    - **reconstruction of an event from what happened in the detector**

The Event Data Model

Refining the data



TAG

Summary of Event.
Intended for selection.
1 KB/event.
Trigger decision, p_T of 4 best electrons, jets...

Reconstruction Output.
Intended for calibration.
500 KB/event.
Cells, Hits, Tracks, Clusters, Electrons, Jets, ...

Raw Channels.
1.6 MB/event.

Raw Data Objects

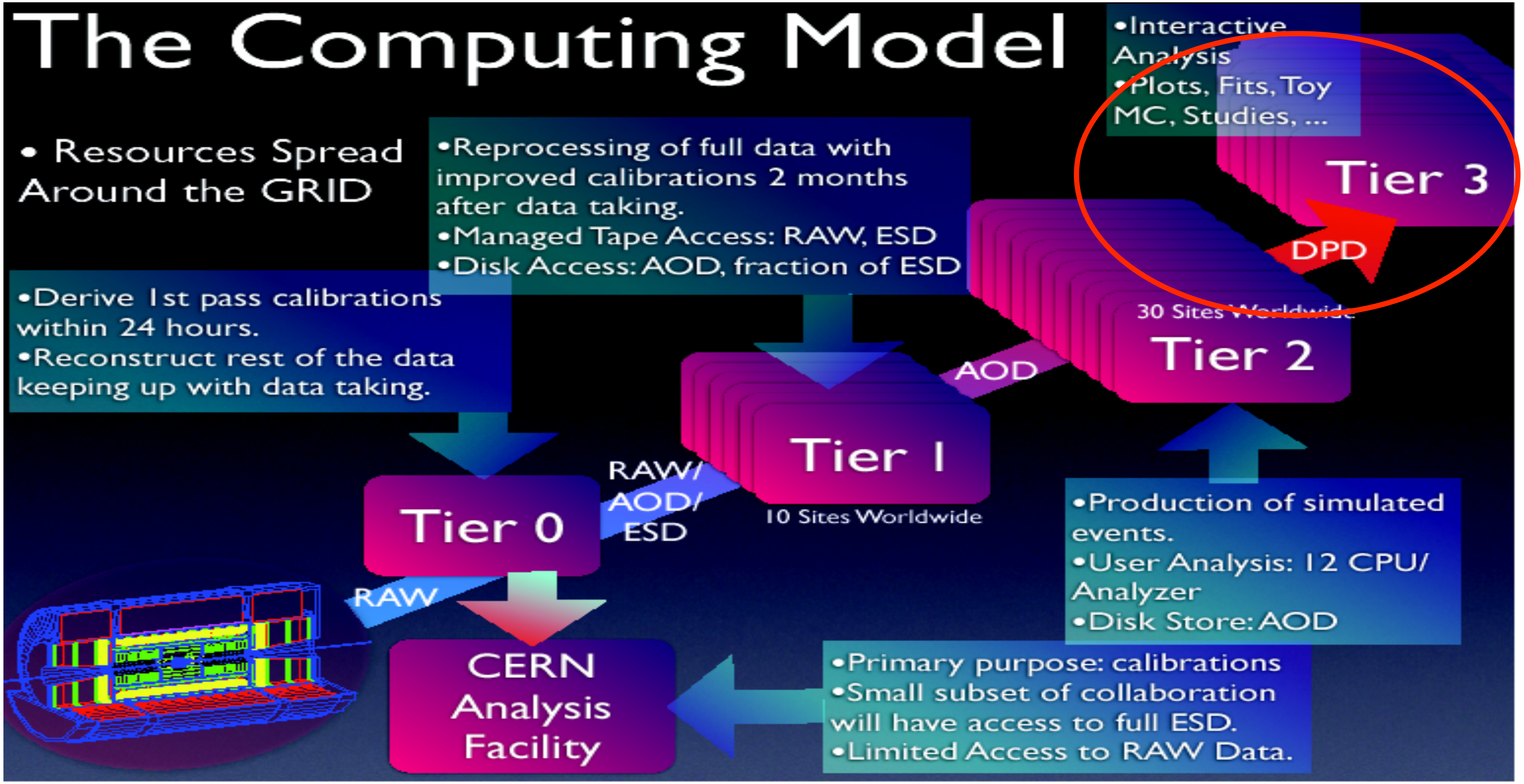
Event Summary Data

Analysis Object Data

Derived Physics Data

Intended for Analysis.
100 KB/event.
"Light-weight" Tracks, Clusters, Electrons, Jets, Electron Cells, Muon HitOnTrack,...

Intended for "interactive" Analysis.
~10-20 KB/event.
What-ever is necessary for a specific analysis/calibration/study.



The Computing Model

- Resources Spread Around the GRID

•Interactive Analysis
•Plots, Fits, Toy MC, Studies, ...

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD
•Disk Access: AOD, fraction of ESD

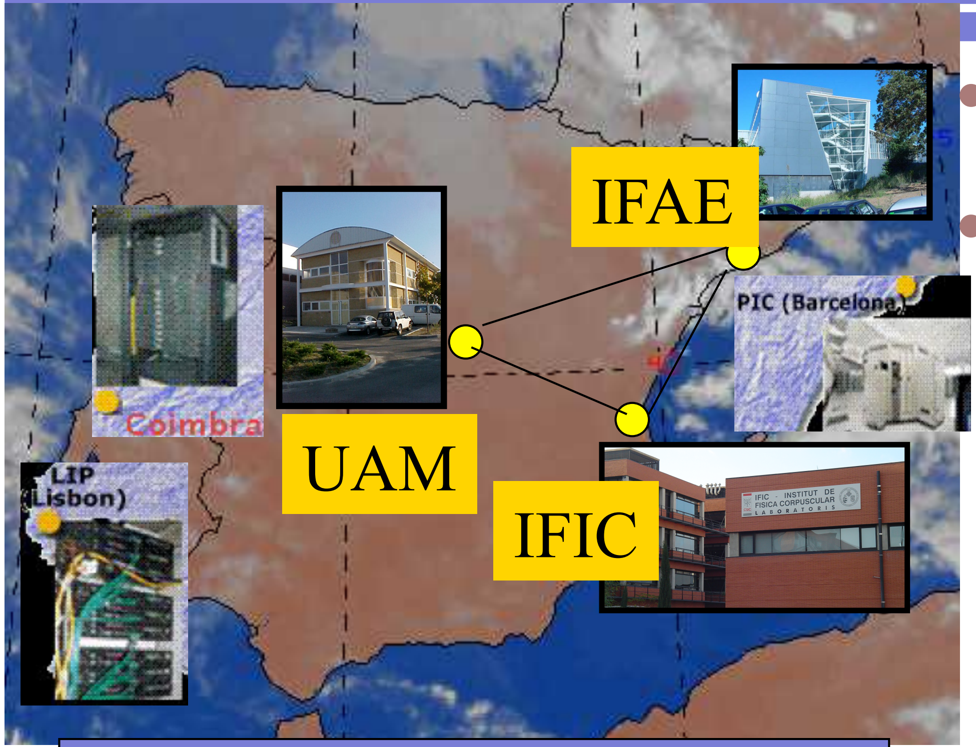
•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

•Production of simulated events.
•User Analysis: 12 CPU/Analyzer
•Disk Store: AOD

•Primary purpose: calibrations
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

- **Analysis Data Format**
  - Derived Physics Dataset **(DPD)** after many discussions last years in the context of the Analysis Forum will consist (for most analysis) of **skimmed/slimmed/thinned AODs plus relevant blocks of computed quantities** (such as invariant masses).
    - **Produced at Tier-1s and Tier-2s**
    - **Stored in the same format as ESD and AOD at Tier-3s**
    - **Therefore readable both from Athena and from ROOT**

Spanish-Iberian cloud for ATLAS

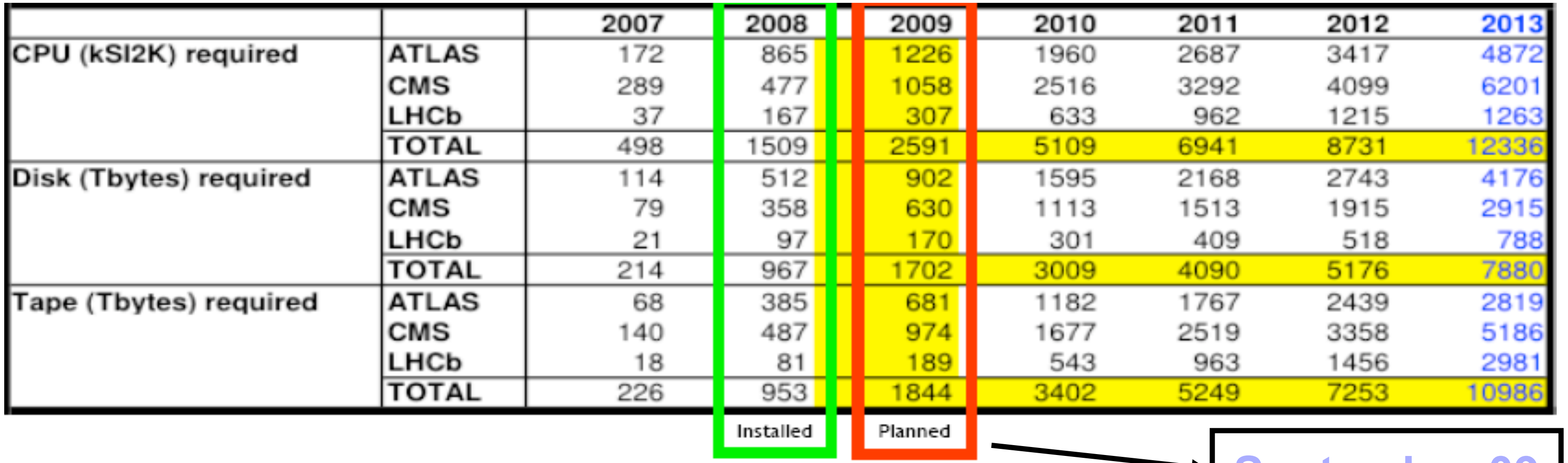


SWE Cloud:
Spain-Portugal
Tier1:
PIC-Barcelona
Tier2's:
UAM, IFAE & IFIC
LIP & Coimbra

- **Tier1 and its related Tier2s are organized in so called clouds**
- **Tier1 at PIC Barcelona**
  - Offers **storage and processing resources** for 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-1s centres in order to guarantee the conservation and integrity of the data.
  - **~10% of the raw data** from the LHC accelerator will be stored at PIC.
  - **Optical Private Network** (OPN) Tier0 (CERN) ↔ Tier1's.
  - More than **9 PetaBytes in/out** PIC in 2008.

Tier1 Resources

- It will provide the infrastructure for **data re-processing**, as the **raw data stored** will be reprocessed several times per year with new parameters, as calibration and alignment constants improve.



		2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K) required	ATLAS	172	865	1226	1960	2687	3417	4872
	CMS	289	477	1058	2516	3292	4099	6201
	LHCb	37	167	307	633	962	1215	1263
	TOTAL	498	1509	2591	5109	6941	8731	12336
Disk (Tbytes) required	ATLAS	114	512	902	1595	2168	2743	4176
	CMS	79	358	630	1113	1513	1915	2915
	LHCb	21	97	170	301	409	518	788
	TOTAL	214	967	1702	3009	4090	5176	7880
Tape (Tbytes) required	ATLAS	68	385	681	1182	1767	2439	2819
	CMS	140	487	974	1677	2519	3358	5186
	LHCb	18	81	189	543	963	1456	2981
	TOTAL	226	953	1844	3402	5249	7253	10986
			Installed	Planned				

September 09

- Data Storage:
  - Experiments do need **large, reliable and scalable storage services**.
  - To server the data at the required speed in order to maximize the efficiency of the cluster.
  - **Multi-Gigabit Ethernet network architecture**, specially designed to enhance high **speed data movement between WAN** (Tier0, Tier1s, Tier2s) and **LAN** (CPU farm).
  - **dCache storage system**.

Spanish Distributed Tier2 resources

- ATLAS Spanish Federated Tier2

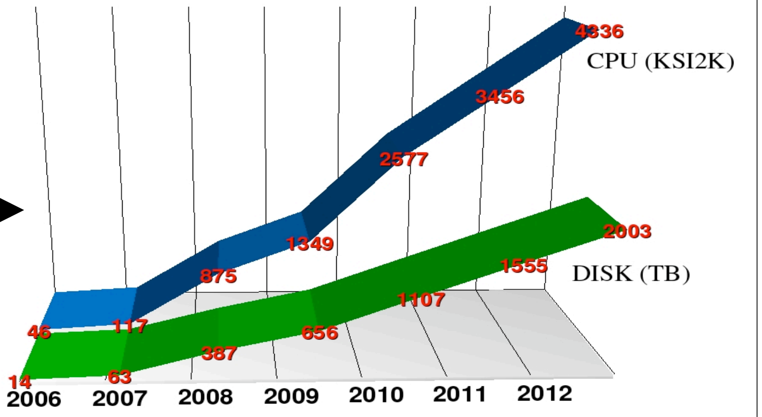
- •IFIC: Valencia (coordinator)
- •IFAE: Barcelona
- •UAM: Madrid

Ramp-up of Tier-2 Resources (after LHC rescheduling) numbers are cumulative.
Evolution of ALL ATLAS T-2 resources according to the estimations made by ATLAS CB (Oct.06)

Year	2006	2007	2008	2009	2010	2011	2012
CPU(KSI2k)	925	2336.11	17494.51	26972.76	51544.64	69128.42	86712.2
Disk (TB)	289	1259.04	7744.37	13112.04	22132.3	31091.45	40050.92

Spanish ATLAS T-2 assuming a contribution of a 5% to the whole effort

Year	2006	2007	2008	2009	2010	2011	2012
CPU(KSI2k)	46	117	875	1349	2577	3456	4336
Disk (TB)	14	63	387	656	1107	1555	2003



Strong increase of resources

Present resources of the Spanish ATLAS T-2 (April'09)

	IFAE	UAM	IFIC	**TOTAL**
CPU (ksi2k)	201	338	96	435
Disk (TB)	94	165	34	293

New acquisitions in progress to get the pledged resources

Accounting values are normalized according to WLCG recommendations

Spanish Resources

- Storage Element System

	SE (Disk Storage)
IFIC	Lustre+StoRM
IFAE	dCache/disk+SRM posix
UAM	dCache

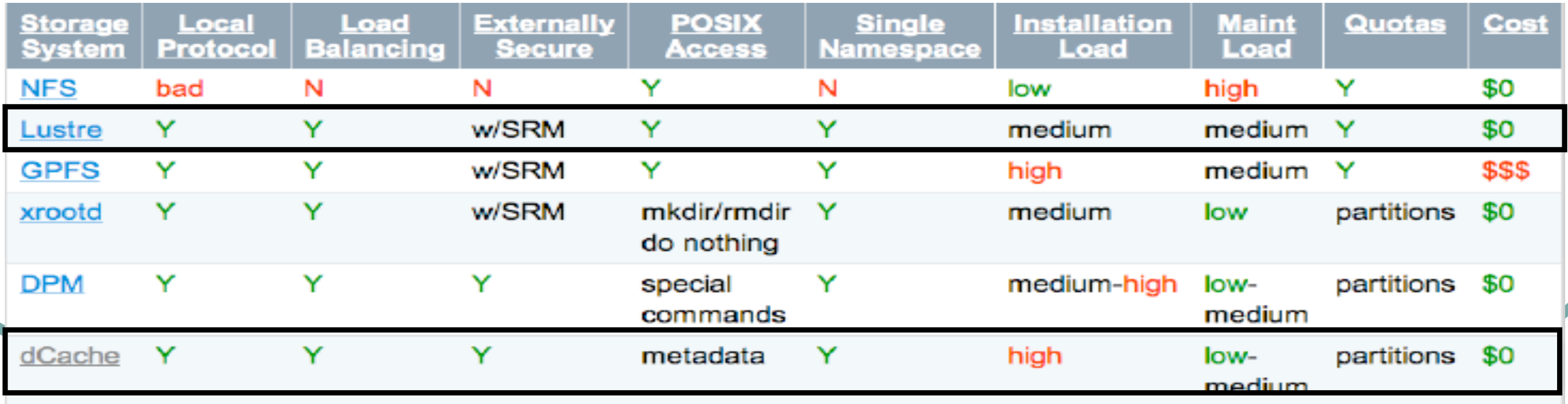
•StoRM: Posix SRM v2 (as the SRM interface)

•Lustre: High performance standard file system

Shares: 50% IFIC, 25% IFAE and 25% UAM. Data (AOD) distribution and DDM FT continuously running from Tier1 to Tier2

•**A Tier needs a reliable and scalable storage system that can hold the users data, and serve it in an efficient way to users.**

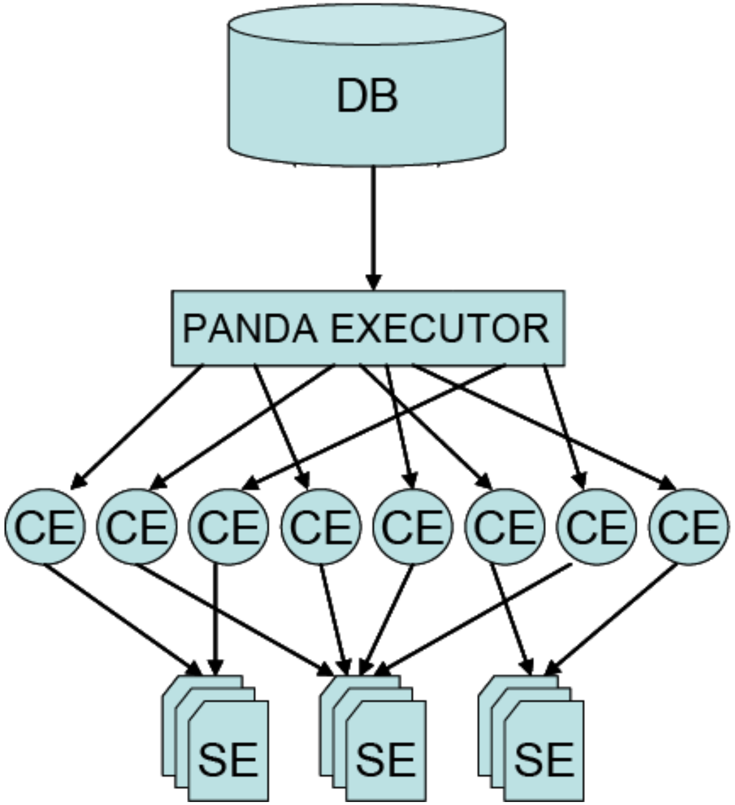
•**A first sketch of a Storage system matrix (evaluation of different systems on going at CERN):**



Storage System	Local Protocol	Load Balancing	Externally Secure	POSIX Access	Single Namespace	Installation Load	Maint Load	Quotas	Cost
NFS	bad	N	N	Y	N	low	high	Y	$0
Lustre	Y	Y	w/SRM	Y	Y	medium	medium	Y	$0
GPFS	Y	Y	w/SRM	Y	Y	high	medium	Y	$$$
xrootd	Y	Y	w/SRM	mkdir/rmdir do nothing	Y	medium	low	partitions	$0
DPM	Y	Y	Y	special commands	Y	medium-high	low-medium	partitions	$0
dCache	Y	Y	Y	metadata	Y	high	low-medium	partitions	$0

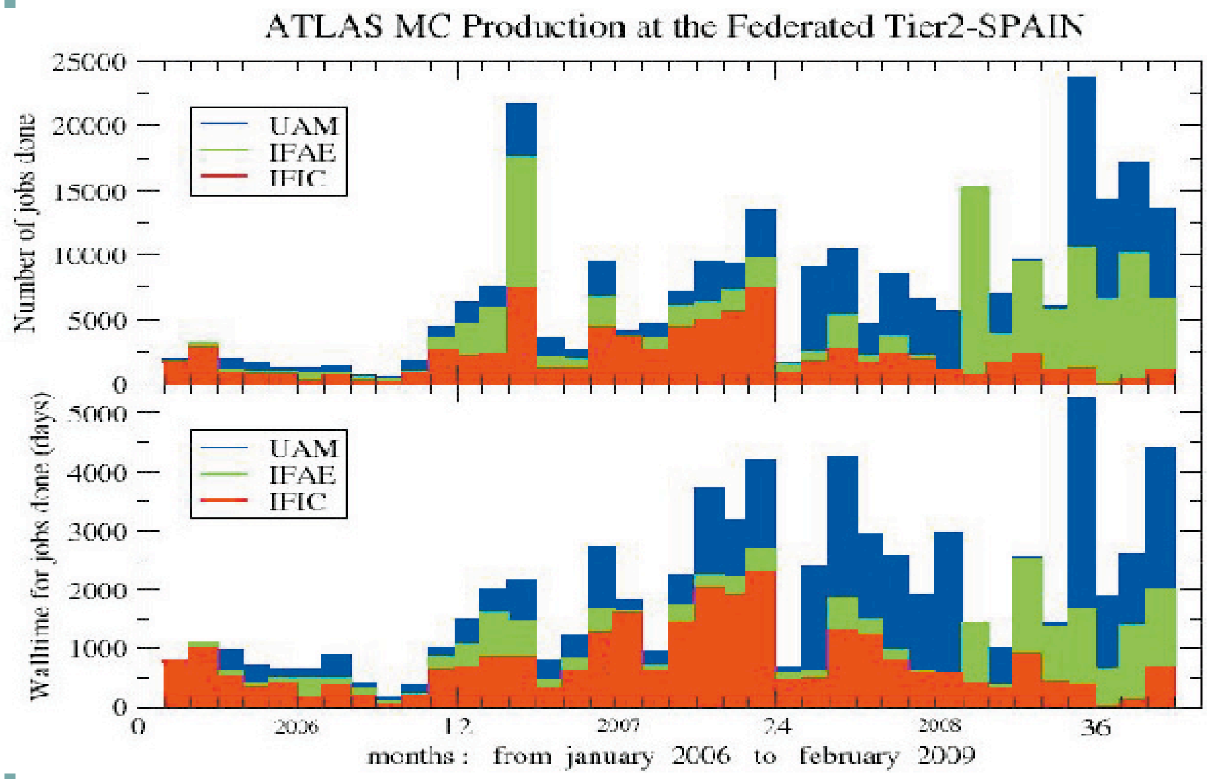
Simulated event production

Production System for Simulated Data (MC) :



- **The ATLAS production system:**
  - **A database (DB) where jobs to be run are defined as well as their run-time status.**
  - **An executor (PANDA) which takes the jobs from the DB and manages sending them to the ATLAS computing resources, using pilots jobs.**
    - **Check the correct environment for running the jobs**
    - **Have the ability to report free resources on the cluster they are running.**
    - **Together with DDM transfer the needed data to the site before running the job**
  - **A distributed data management (DDM) system which stores the produced data on the adequate storages resources and register them into the defined catalogues.**

Simulated event production



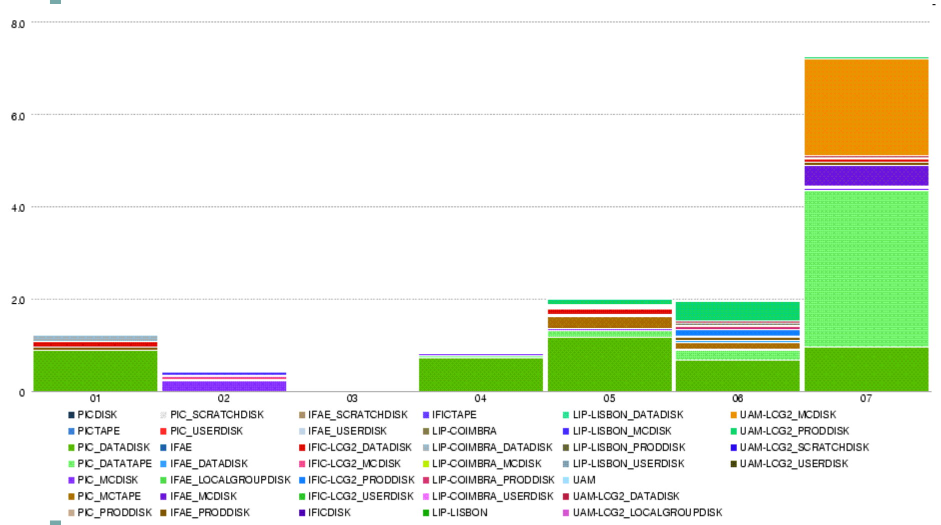
- Pilot job schema was deployed in January 2007 at the Spanish cloud.
- Walltime Efficiency for the ES cloud has been pretty stable at around 95% during last months.
- The Spanish Tier2 contribution to the massive production is around 2.5%

- Walltime from January 2006 to March 2009 in our Tier2
- Demonstrate the stability of the pilot job schema, which is capable to fill all available resources at the sites.

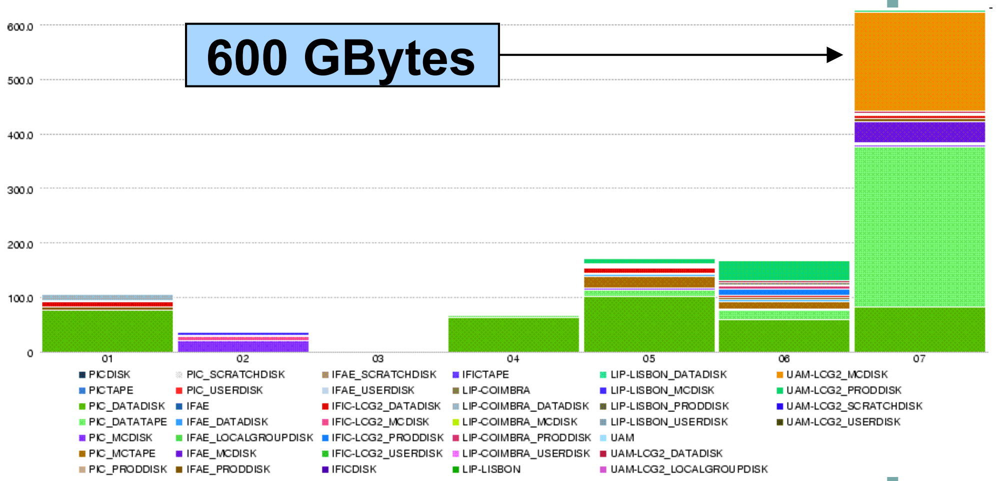
Data Transfer among Spanish Cloud

Data distribution (1-8 May 2009)

Throughput (MB/s)



Data transfer (GBytes)



600 GBytes

- There are millions of file transfer throughout the day coming from simulation production, AOD distribution, Functional tests, etc.
- Dataset are broadly distributed from Tier0 to Tiers1s and then among the Tier2s inside the cloud.
- All sites within the Spanish cloud are involved in this data transfer tests, and results have been pretty stable since the lasts months.

Distributed Analysis (DA) activities

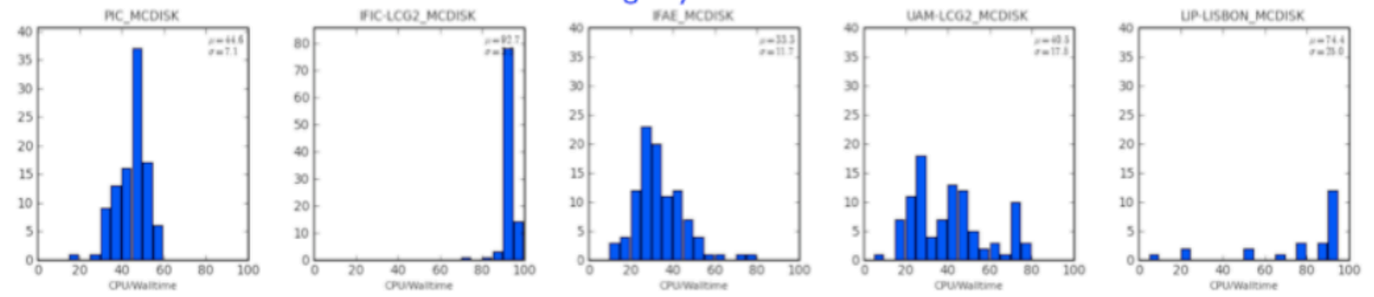
- Final target of grid computing for the LHC is to provide a solid framework to perform analysis over the real data.
- DA test are being executed on a regular basis, in order to spot potential problems at the sites.
- DA test jobs are defined centrally and rely on a real user analysis case.
- Jobs are sent in bulk to the cloud and dispersed among the active sites.
  - Job Brokering is done through Ganga (Analysis toolkit for ATLAS and LHCb) with direct submission to the Computing Element (CE)
  - Ganga uses the native data access protocols, which depend on the site architecture. These protocols in our cloud are: *dcap* for dCache, *rfio* for Castor/DPM and *file* for Storm (Posix I/O through Lustre).
  - A parallel way to get input data is being tested: File Stager
    - Pre-copies the next input file with lcg-cp (or anything else) in a background thread.

Distributed Analysis (DA) activities

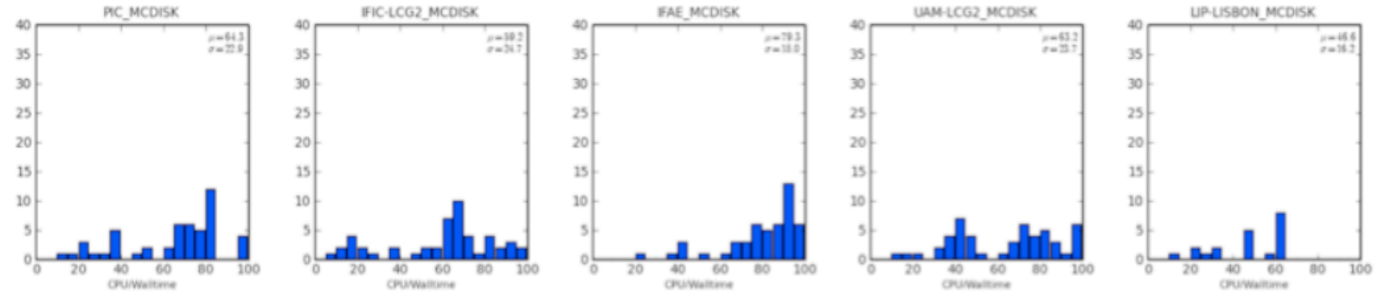
Site	Job eff.	CPU/Wall	Events (Hz)	SW setup(s)	Input(Output) (s)
PIC	100%	44%	8	119	19(32)
IFAE	100%	33%	9	30	79 (27)
IFIC	94%	92%	10	82	8 (34)
UAM	95%	41%	8	340	16 (11)
LIP-Lisbon	80%	74%	17	92	6 (26)

Site CPU/Walltime

Natural Storage System Protocol



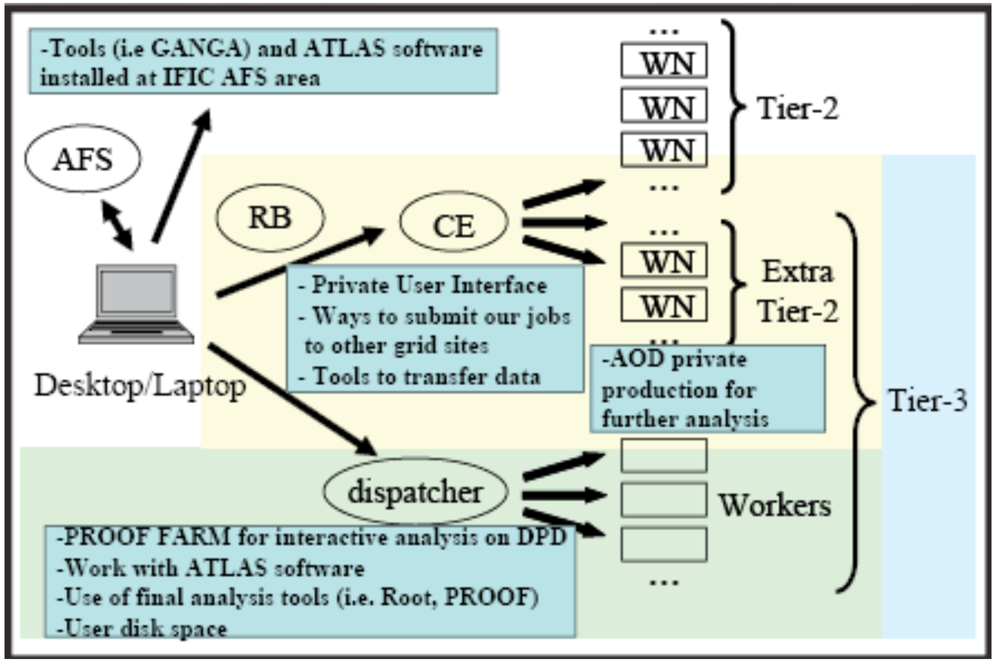
Using File Stager



- **The overall efficiency was 96%.**
- **CPU/Walltime and input/output perform much better in sites with the filesystem mounted on the worker nodes.**
- **dCache shows slightly better job efficiency**

- **File stager showed to improve the CPU/Wall efficiency for dCache sites but is not a good option for Lustre**

IFIC Analysis Facility Tier3


-Tools (i.e GANGA) and ATLAS software installed at IFIC AFS area

- Private User Interface
- Ways to submit our jobs to other grid sites
- Tools to transfer data

-AOD private production for further analysis

-PROOF FARM for interactive analysis on DPD
-Work with ATLAS software
-Use of final analysis tools (i.e. Root, PROOF)
-User disk space

- A Tier3 site-located computing infrastructure is needed as Analysis Facility for the Spanish ATLAS end-user physicists.
- It could be used by users for running jobs with few events or storing private datasets and DPDs.
- The use of the Tier3 analysis facility could be faster than the grid-use for some kind of jobs.
- However, Tier2-Tier3 interaction is necessary in order to access AODs and DPD on DDM

Conclusions

- Almost all ATLAS distributed computing areas have been tested.
  - Computing resources are increasing according to ATLAS schedule.
  - Reliability of Spanish Tier2 greater than 90% over several months.
  - Continuous Production of ATLAS Simulated Event Data.
  - High Rate Data transfer between Tier2 and its associated Tier1.
  - Enable Physics analysis by Spanish ATLAS users.
  - Efficiencies on Data Transfers, MC Production and Analysis Jobs similar to other Tier2s.
- **One can be confident for the starting of data taking, based on the results obtained during the last year in the Iberian/ Spanish sites.**
- Next steps are:
  - perform last interventions to ensure a good reliability of the sites during the expected long data taking period.
  - continue exercising the system by participating in all ATLAS activities.