Analysis Facility Infrastructure: ATLAS in Valencia

S. González de la Hoz

IFIC – *Institut de Física Corpuscular de València*

DESY Computing Seminar
Hamburg, 30th June 2008

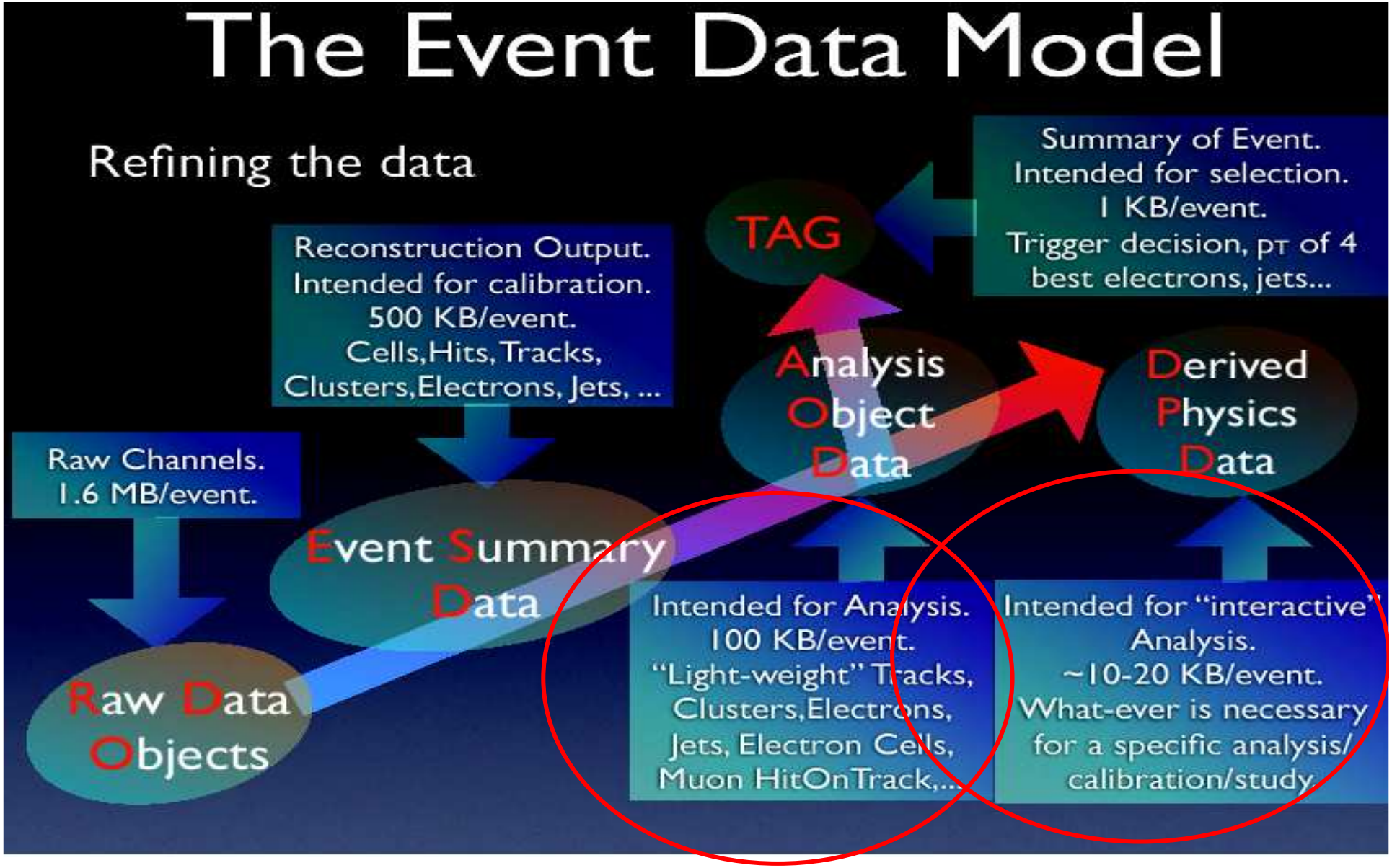
Outline

- Introduction to the ATLAS Computing Model
  - A Hierarchical Model
  - The Event Data Model
  - The Computing Model
  - Distributed Analysis in ATLAS
- ATLAS Spanish Tier-2
  - Distributed Tier-2
  - Resources
  - MC production
  - Data Movement and Network
  - Storage Element
  - Analysis Use cases
- What is an ATLAS Tier-3?
  - Minimal requirements
  - Grid and non-Grid resources
- Tier-3 prototype at IFIC-Valencia
  - Extension of our Tier-2
  - A PC farm outside grid for interactive analysis: Proof
  - Typical use of a Tier-3
- Conclusions

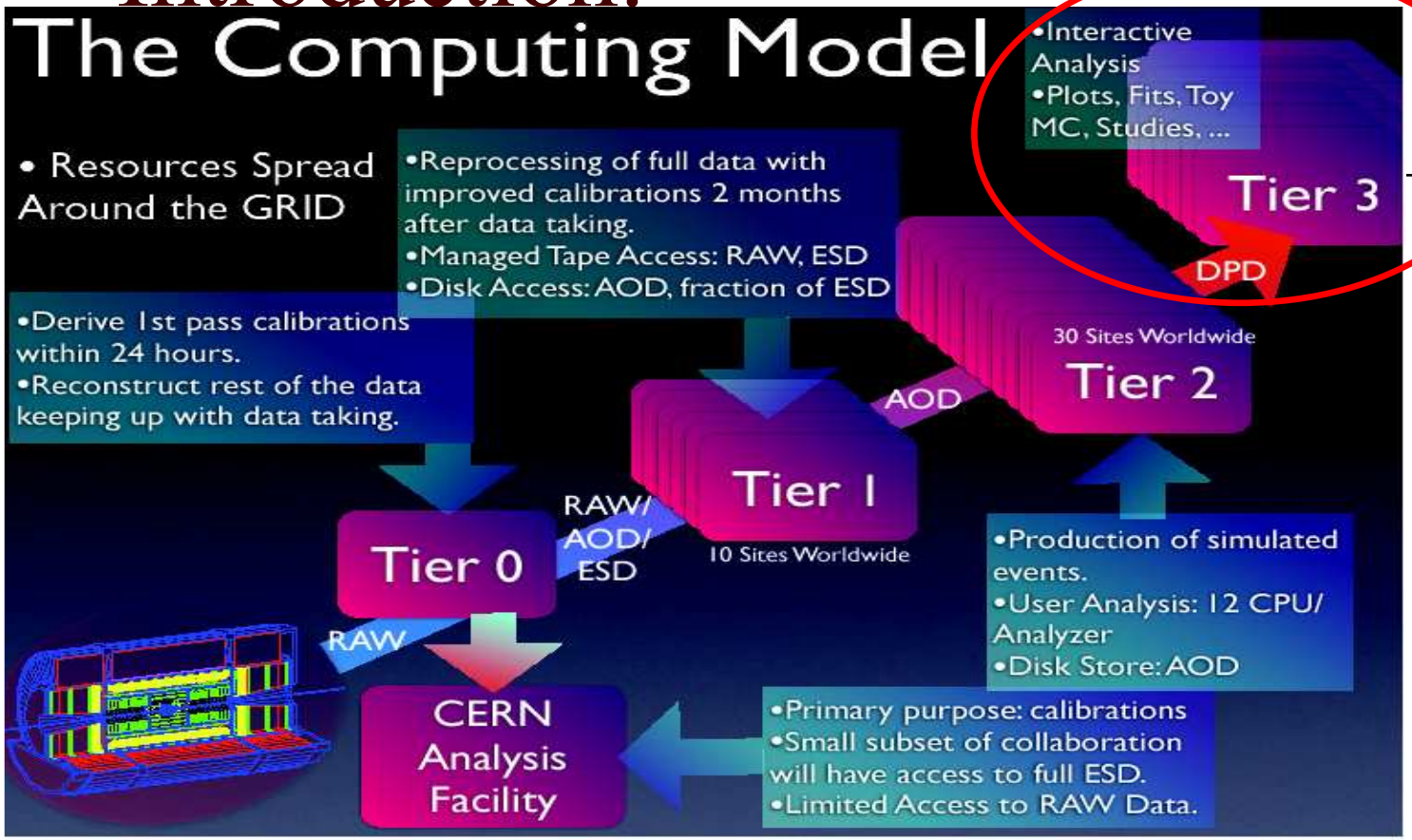
Introduction: A hierarchical model

- 1 Tier-0 centre at CERN
  - Copy Raw data to CERN Castor tape for archival
  - Copy Raw data to Tier-1s for storage and subsequent reprocessing
  - Run first-pass calibration/alignment (within 24 hrs)
  - Run first-pass reconstruction (within 48 hrs)
  - Distributed reconstruction output (ESDs, AODs & TAGs) to Tier-1s
- 10 Tier-1s centres: FZK (Karlsruhe, DE), PIC (Barcelona, ES), etc..
  - Store and take care of a fraction of Raw data **(1/10 + 1/10 from another Tier1)**
  - Run slow calibration/alignment procedures
  - Rerun reconstruction with better calib/align and/or algorithms
  - Distribute reconstruction output to Tier-2s
  - Keep current versions of ESDs and AODs on disk for analysis
  - Run large-scale event selection and analysis jobs
- ~35 Tier-2s centres: IFIC (Valencia, ES), DESY-CMS (Hamburg, DE), Munich-Federation-ATLAS..
  - Run simulation (and calibration/alignment when/where appropriate)
  - Keep current versions of AODS **(1/3 – 1/4)** and samples of other data types on disk for analysis
  - Run analysis jobs
- Tier-3s centres in all participating institutions
  - Provide access to Grid resources and local storage for end-user data
  - Contribute CPU cycles for simulation and analysis if/when possible
- CAF (CERN Analysis Farm)
  - Designed for all analysis work related to data calibration and commissioning

Introduction:



Introduction:



- Analysis Data Format
  - Derived Physics Dataset **(DPD)** after many discussions last year 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

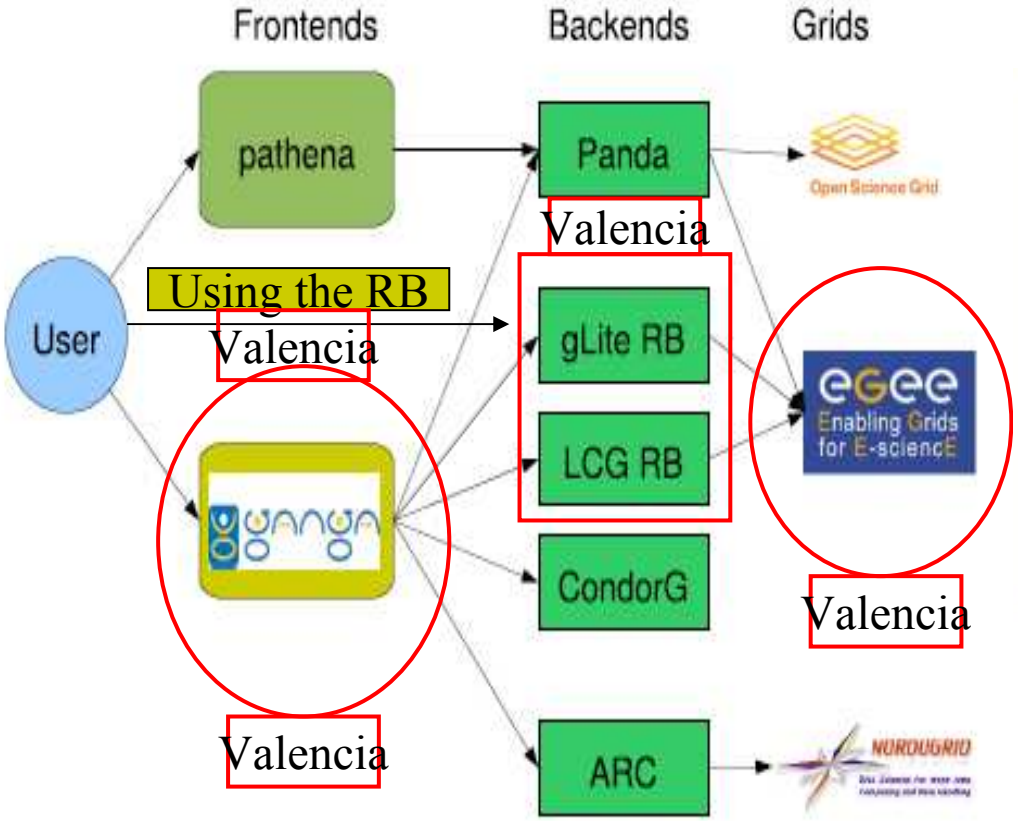
Introduction: Atlas Distributed Analysis using the Grid

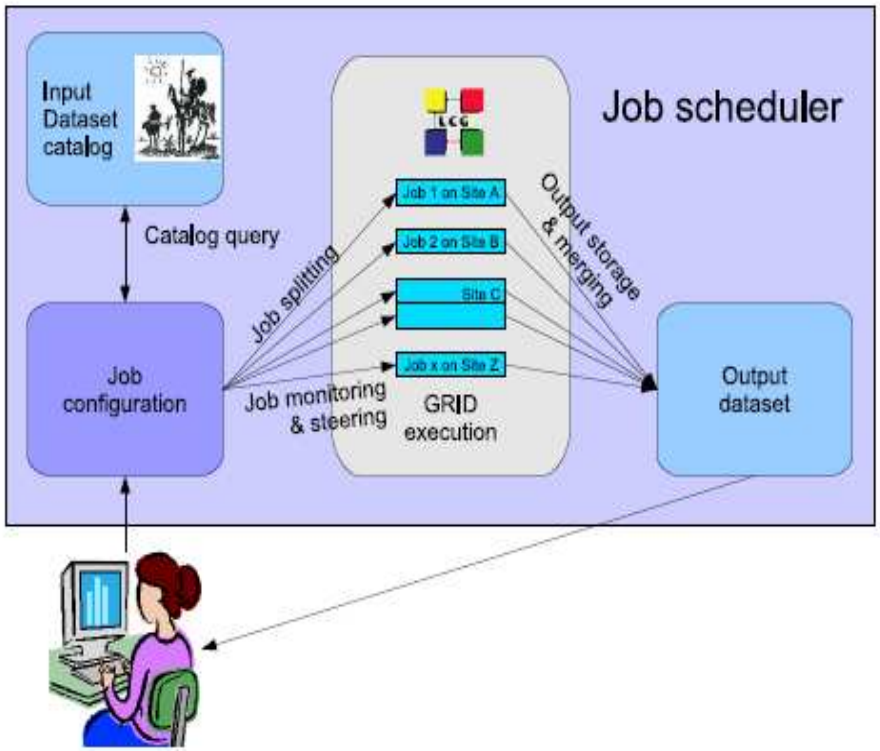
- Heterogeneous grid environment based on 3 grid infrastructures: OSG, EGEE, Nordugrid

- Grids have different middleware, replica catalogs and tools to submit jobs.
- Naive assumption: Grid ~large batch system
  - Provide complicated job configuration jdl file (Job Description Language)
  - Find suitable ATLAS (Athena) software, installed as distribution kits in the Grid
  - Locate the data on different storage elements
  - Job splitting, monitoring and book-keeping
  - Etc..
    - → NEED FOR AUTOMATION AND INTEGRATION OF VARIOUS DIFFERENT COMPONENTS
    - We have for that Two Frontends: Panda & Ganga

Introduction: Atlas Distributed Analysis using the Grid – Current Situation

- How to combine all these: Job scheduler/manager: GANGA

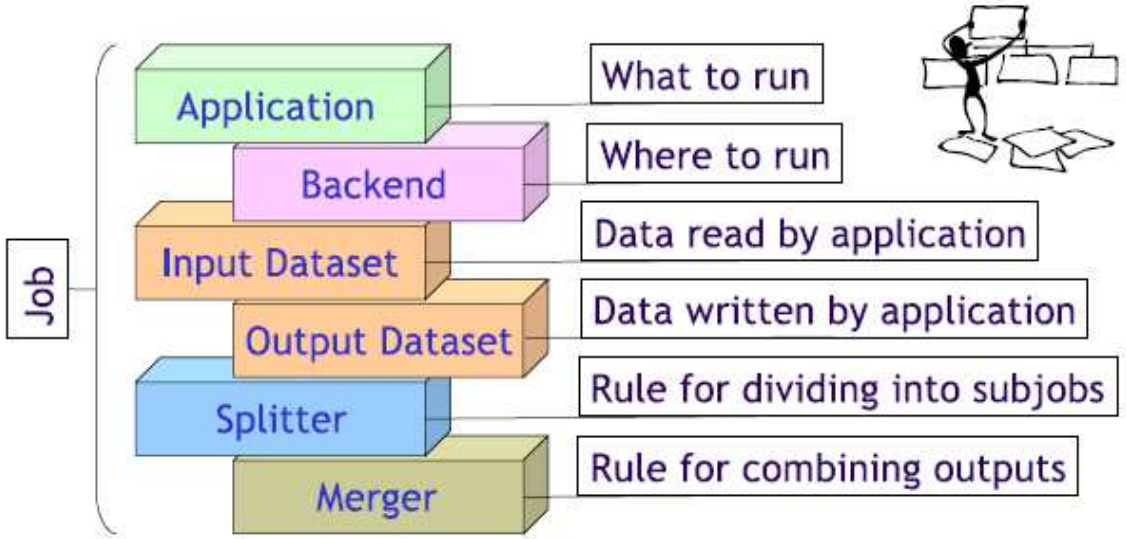




Ganga

https://twiki.cern.ch/twiki/bin/view/Atlas/DistributedAnalysisUsingGanga

- A user-friendly job definition and management tool
- Allows simple switching between testing on a local batch system and large-scale data processing on distributed resources (Grid)
- Developed in the context of ATLAS and LHCb
- Python framework
- Support for development work from UK (PPARCG/GridPP), Germany (D-Grid) and EU (EGEE/ARDA)

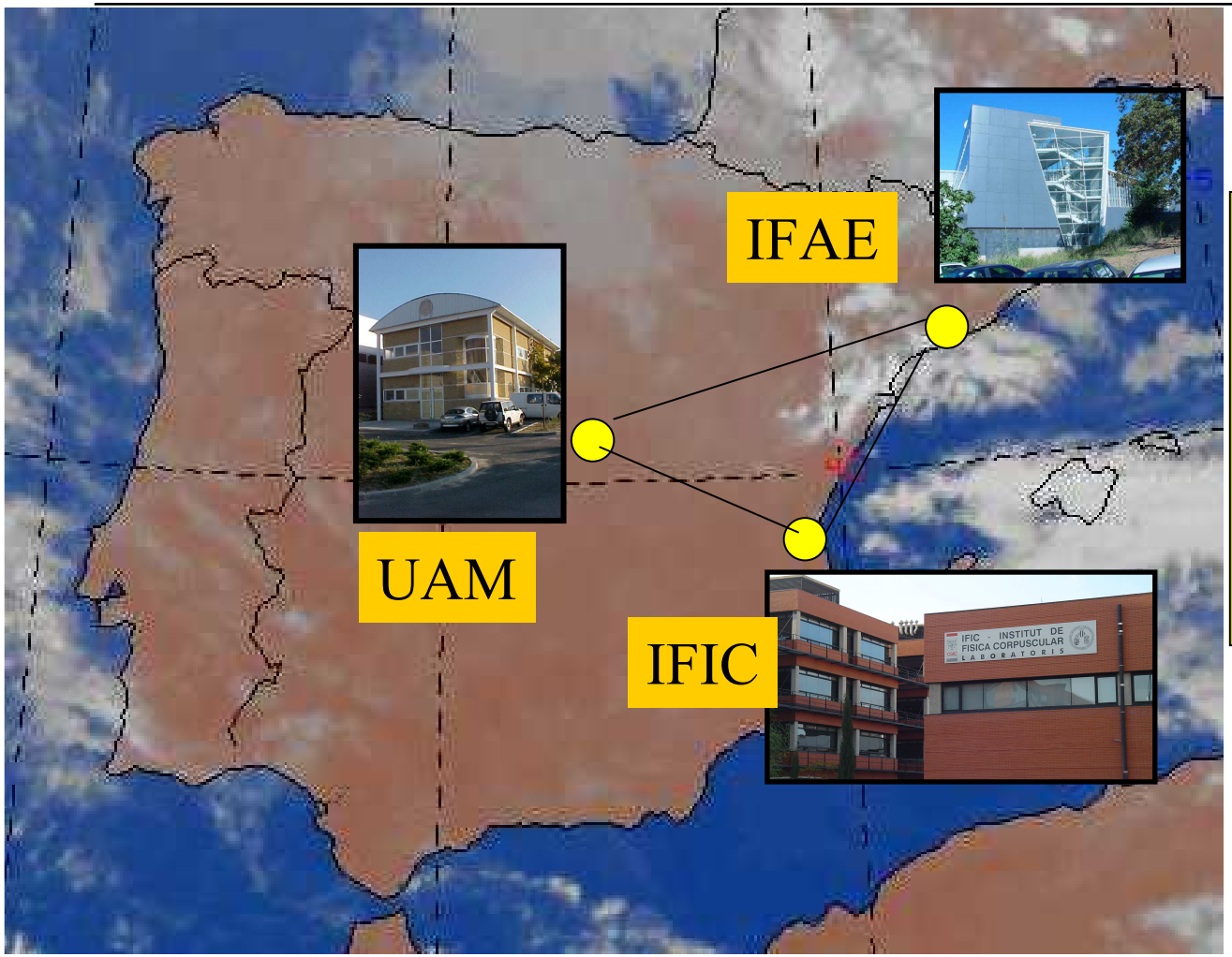


- Ganga is based on a simple, but flexible, job abstraction
- A job is constructed from a set of building blocks, not all required for every job
- Ganga offers three ways of user interaction:
  - Shell command line
  - Interactive IPython shell
  - Graphical User Interface

ATLAS Spanish Tier2

- **The ATLAS distributed TIER-2 is conceived as a Computing Infrastructure for ATLAS experiment. Its main goals are:**
  - **Enable Physics Analysis by Spanish ATLAS Users**
    - **Tier-1s send AOD data to Tier-2s**
  - **Continuous production of ATLAS MC events**
    - **Tier-2s produce simulated data and send them to Tier-1s**
  - **To contribute to ATLAS + LCG Computing Common Tasks**
  - **Sustainable growth of infrastructure according to the scheduled ATLAS ramp-up and stable operation**
- **In the ideal world (perfect network communication hardware and software) would not need to define default Tier-1 – Tier-2 associations.**
- **In practice, it turns out to be convenient (robust?) to partition the Grid so that there are default data path between Tier-1s and Tier-2s.**
  - **FTS (File Transfer System) channels are installed for these data for production use**
  - **All other data transfers go through normal network routes**
- **In this model, a number of data management services are installed only at Tier-1s and act also on their "associated" Tier-2s:**
  - **VO Box, FTS channel server, Local file catalogue (part of Distributed Data Management)**

ATLAS Spanish Distributed Tier2



SWE Cloud:
Spain-Portugal
Tier-1:
PIC-Barcelona
Tier-2:
UAM, IFAE & IFIC
LIP & Coimbra

Spanish Distributed Tier2: Resources

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 (October 2006)

Año	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

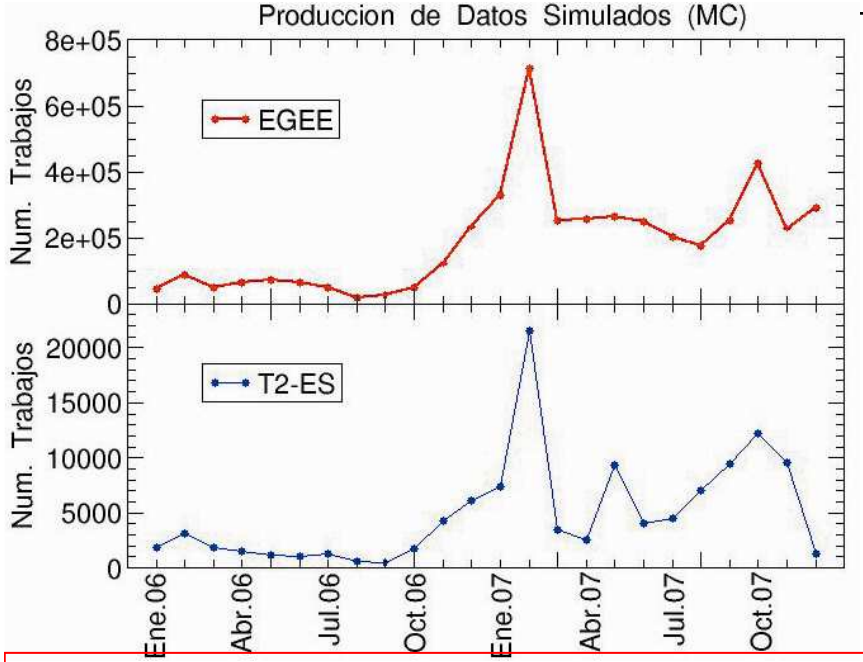
Present resources of the Spanish ATLAS T-2 (June'08)

	IFAE	UAM	IFIC	**TOTAL**
CPU (ksi2k)	**60**	**144**	**132**	**336**
Disk (TB)	**30**	**38**	**36**	**104**

New acquisitions in progress to get the pledged resources

Accounting values are normalized according to WLCG recommendations

Spanish Distributed Tier2: Monte Carlo production



MC Production in 2006 and 2007:

-The production in T2-ES follows the same trend as LCG/EGEE (good performance of the ES-ATLAS-T2)

-The ES-ATLAS-T2 average contribution to the total Data Production in LCG/EGEE was 2.8% in 2006-2007 (take into account that 250 centers/institutes are Participating, 10 of them are T-1)

MC Production in 2008:

- Since January-2008, ATLAS has migrated to PANDA executor
- The production efficiency has been positively affected; the average efficiency was 50% and now is 75%-80% @ T2-ES
- T2-ES contribution is 1.2% in the first 2008 quarterly (essentially due to 2 downtime of PIC (Tier-1))

Spanish Distributed Tier2: Data Movement Management

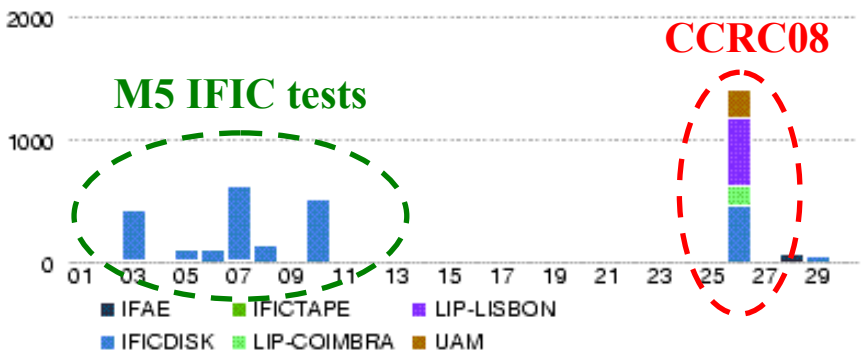
Distributed data management exercises at Tier-2

Cosmic rays data taking (Tier-1 => Tier-2)

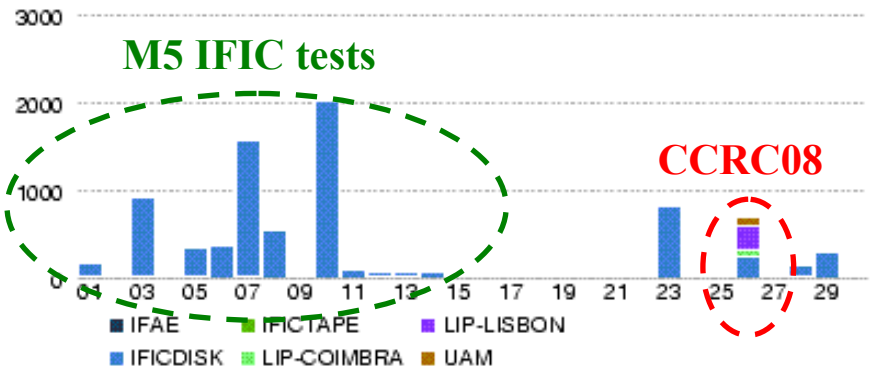
1) End Sept. 2007: M4 exercise (ESD & CBNT)
2) End Oct. – begin Nov 2007: M5 exercise (ESD & CBNT)
3) 12 Feb. 2008: FDR-1 (fdr08_run1, AOD)
4) 25 Feb. 2008: CCRC08 (ccrc08_t0, AOD)

Distributed Data Management (DDM) activity at Tier-2 in February 08

Data transfer (GB/day)



Number of completed file transfers per day



Spanish Distributed Tier2: Data Movement Management

- Data stored at the Spanish ATLAS Tier-2:

Data at IFIC

Dataset type	# sets	#files DDM	#files site	Total size (GBytes)
EVNT	154	60532	46926	1212.633
HITS	217	1167706	16414	463.433
RDO	181	1197532	32752	1805.928
ESD	256	170924	47282	9367.806
AOD	867	328118	303612	9790.303
SAN	18	3908	86	4.625
HPTV	13	3494	72	2.484
CBNT	5	3624	2164	79.533
HIST	31	5082	608	0.000
TAG	10	3540	66	0.027
TOTAL	1752	2944460	449982	22726.773

Data at UAM

Dataset type	# sets	#files DDM	#files site	Total size (GBytes)
EVNT	63	14724	1482	86.989
HITS	227	1184298	36800	927.925
RDO	185	1127594	61158	3119.308
ESD	178	158202	17952	3202.476
AOD	583	530406	175032	4005.110
SAN	18	3698	140	10.779
HPTV	27	3788	154	5.935
CBNT	4	352	350	4.909
HIST	38	5374	460	0.000
TAG	22	3768	124	0.053
TOTAL	1345	3032204	293652	11363.484

Data at IFAE

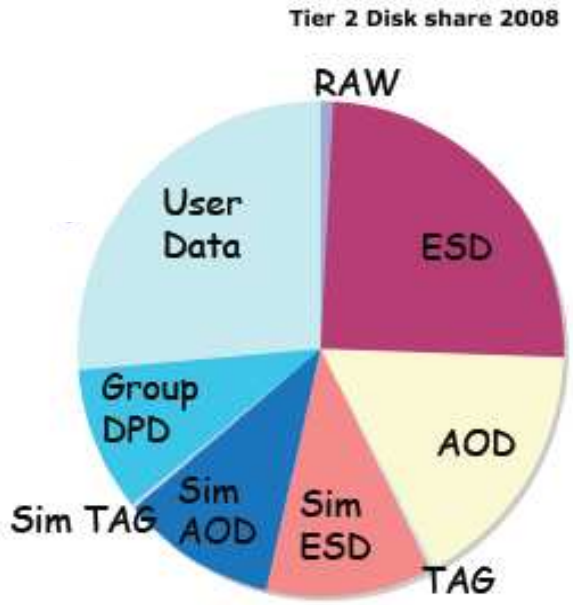
Dataset type	# sets	#files DDM	#files site	Total size (GBytes)
EVNT	10	3856	0	0.000
HITS	96	351860	12	0.059
RDO	43	349076	49464	2767.835
ESD	29	55134	0	0.000
AOD	142	235014	22862	853.392
SAN	5	2748	0	0.000
HPTV	4	2554	0	0.000
CBNT	0	0	0	0.000
HIST	6	958	0	0.000
TAG	5	2328	0	0.000
TOTAL	340	1003528	72338	3621.287

More info: http://ific.uv.es/atlas-t2-es (June 2008)

Tier-2	Total data (TB)	AODs (TB)	AOD contribution
IFAE	3.6	0.9	25 %
UAM	11.4	4.0	35 %
IFICDISK	22.7	9.8	43 %

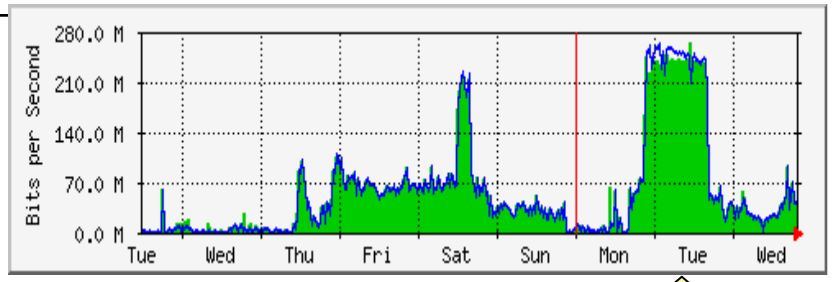
Spanish Distributed Tier2: ATLAS Tier-2 data on disk

- ~35 Tier-2 sites of very, very different size contain:
  - Some fraction of ESD and RAW
    - In 2008: 30% of RAW and 100% of ESD in Tier-2 cloud
    - In 2009: 10% of RAW and 30% of ESD in Tier-2 cloud
  - 10 copies of full AOD on disk
  - A full set of official group DPD
  - Lots of small group DPD
  - User data

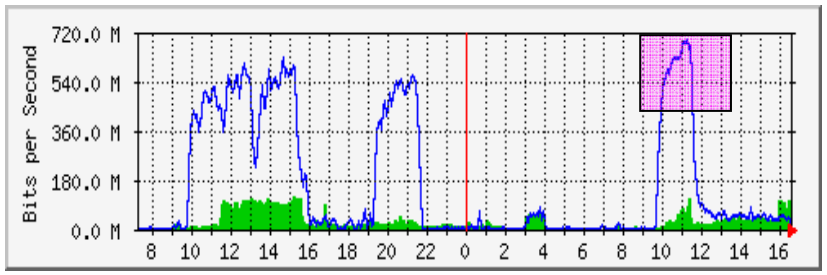


Spanish Distributed Tier2: Network and Data Transfer

- **It is provided by the Spanish NREN RedIRIS**
  - **Connection at 1 Gbps to University backbone**
  - **10 Gbps among RedIRIS POP in Valencia, Madrid and Catalunya**
- **Atlas collaboration:**
  - **More than 9 PetaBytes (> 10 million of files) transferred in the last 6 months among Tiers**
- **The ATLAS link requirement between Tier-1 and Tier-2s has to be 50 MBytes/s (400 Mbps) in a real data taken scenario.**



Data transfer between Spanish Tier1 and Tier2. We reached 250 Mbps using gridftp between T1 -> T2 (CCRC'08)



Data transfer from CASTOR (IFIC) for a TICAL private production. We reached 720 Mbps (plateau) In about 20 hours (4th March 08) High rate is possible

Spanish distributed Tier2: Storage Element System

- Distributed Tier2: UAM(25%), IFAE(25%) and IFIC(50%)

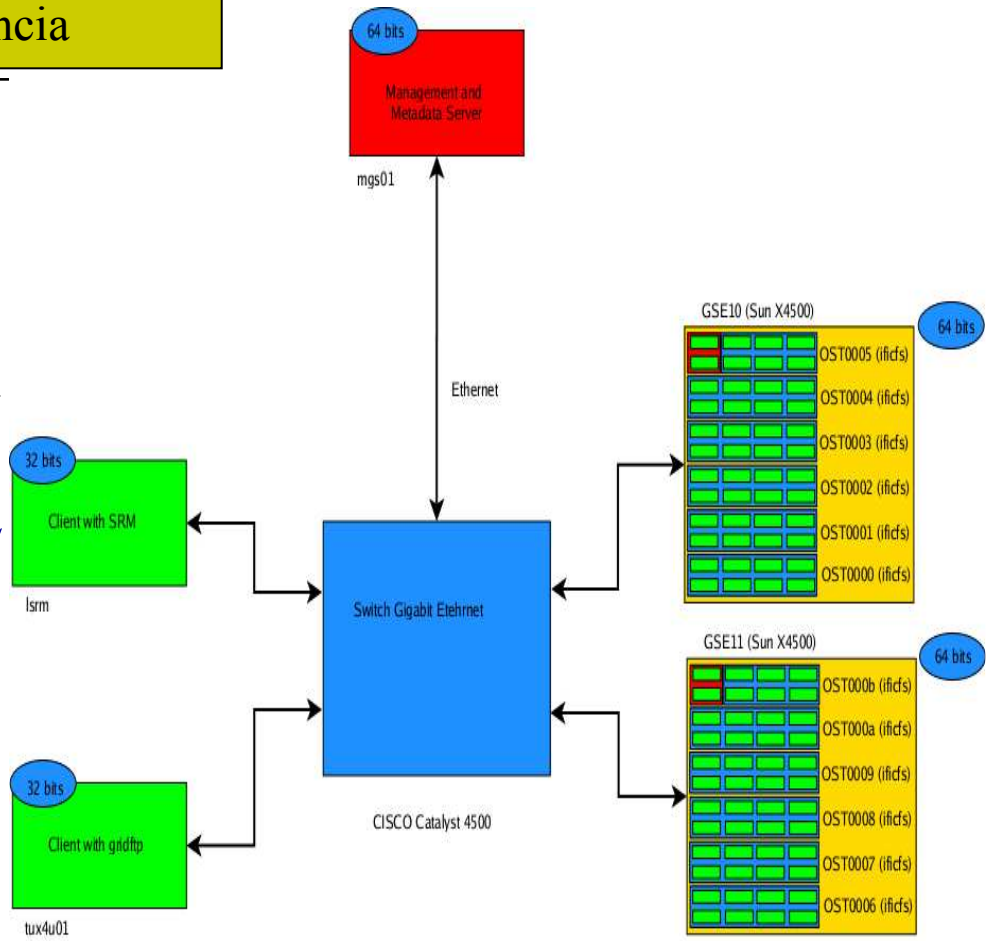
	SE (Disk Storage)
IFIC	Lustre+StoRM
IFAE	dCache/disk+SRM posix
UAM	dCache

- Inside our Tier2 two SE options are used. In case that Lustre won't work as expected we will switch to dCache

StoRM + Lustre

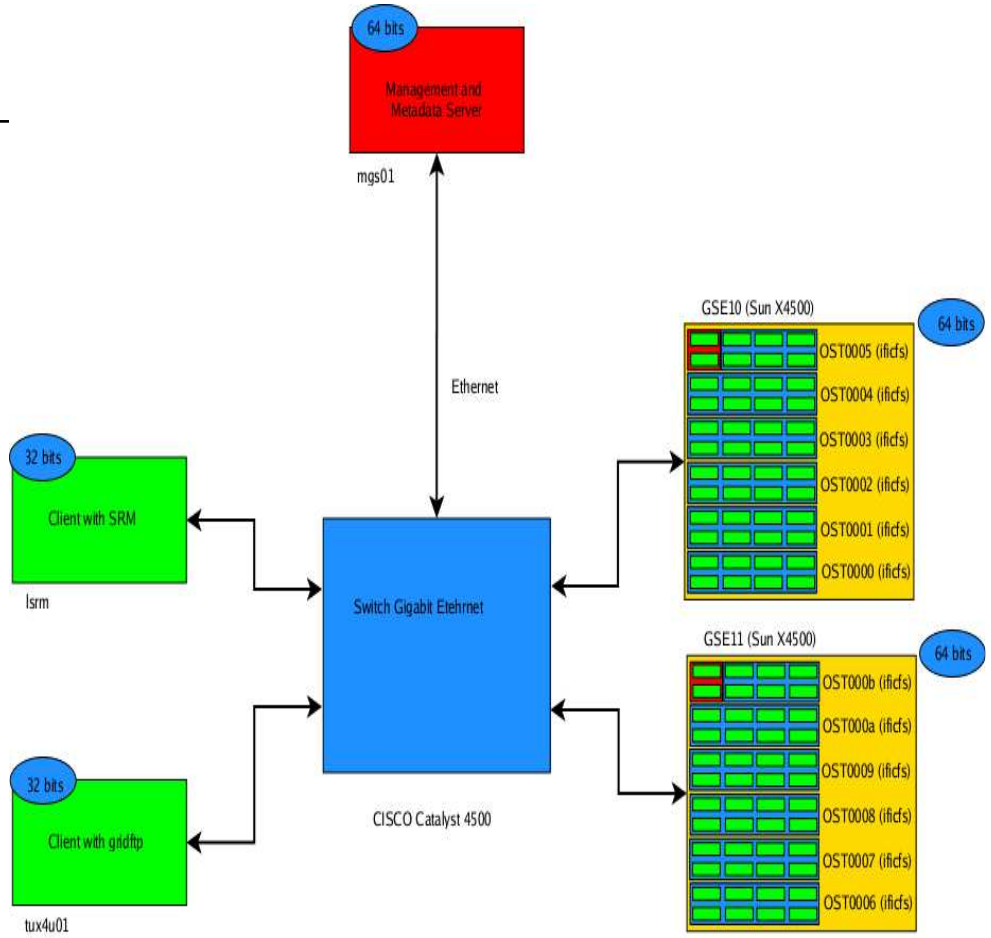
Storage Element system at IFIC-Valencia

- StoRM
  - Posix SRM v2 (server on Lustre)
  - Being used in our IFIC-Tier2.
  - Endpoint: srm://srmv2.ific.uv.es:8443:/srm/managerv2
- Lustre in production in our Tier2
  - High performance file system
  - Standard file system, easy to use
  - Higher IO capacity due to the cluster file system
  - Used in supercomputer centers
  - Free version available
  - Direct access from WN
  - www.lustre.org



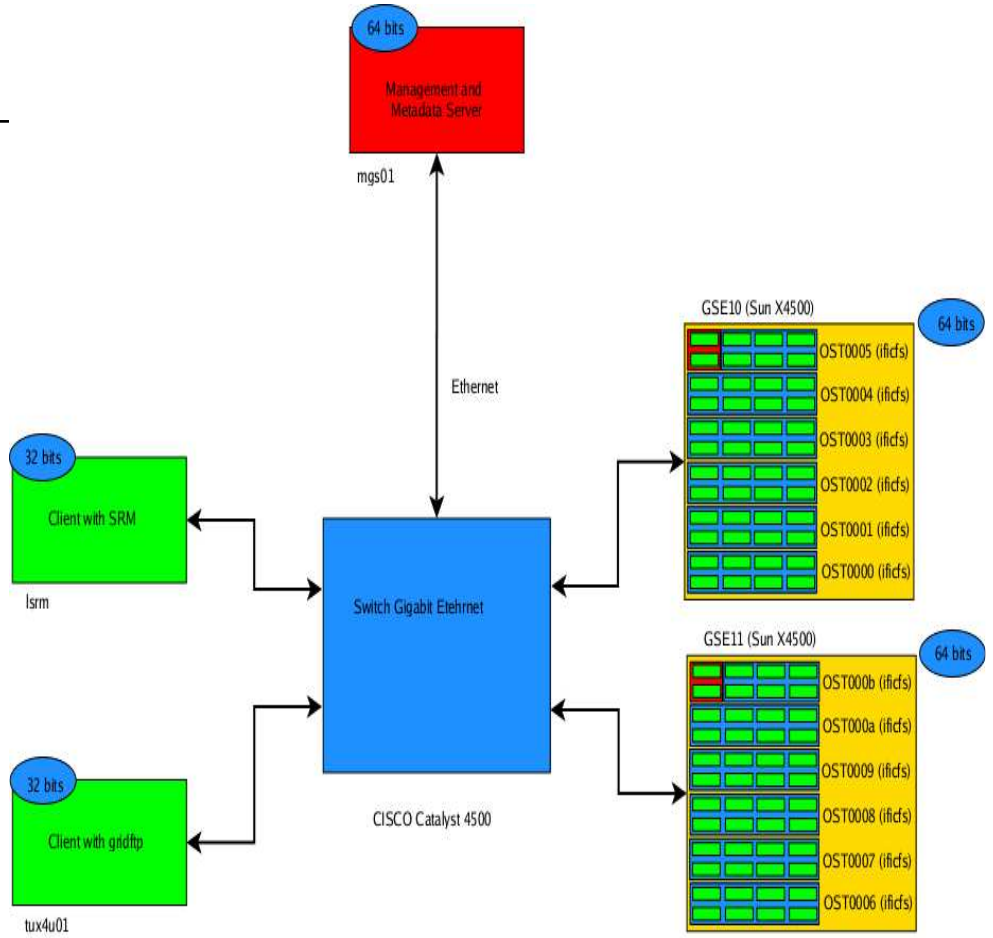
StoRM + Lustre

- Hardware
  - Disk servers:
    - 2 SUN X4500 (two more in place to be installed in the near future, used for testing)
    - 36 TB net capacity
  - Connectivity:
    - Switch Gigabit CISCO Catalyst 4500
  - Grid Access:
    - 1 SRM server (P4 2.7 GHz, GbE)
    - 1 GridFTP server (P4 2.7 GHz, GbE)
  - Lustre Server:
    - 1 MDS (Pentium D 3.2 GHZ, R1 disk)



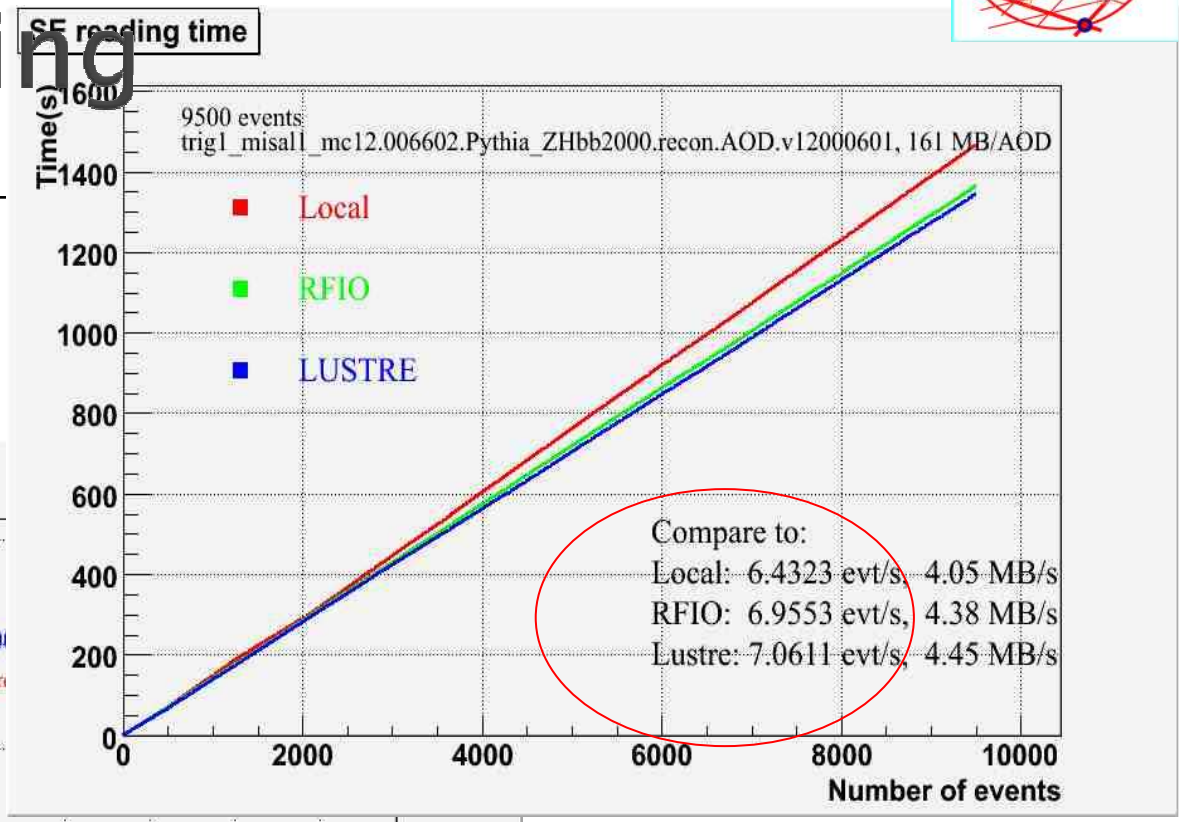
StoRM + Lustre

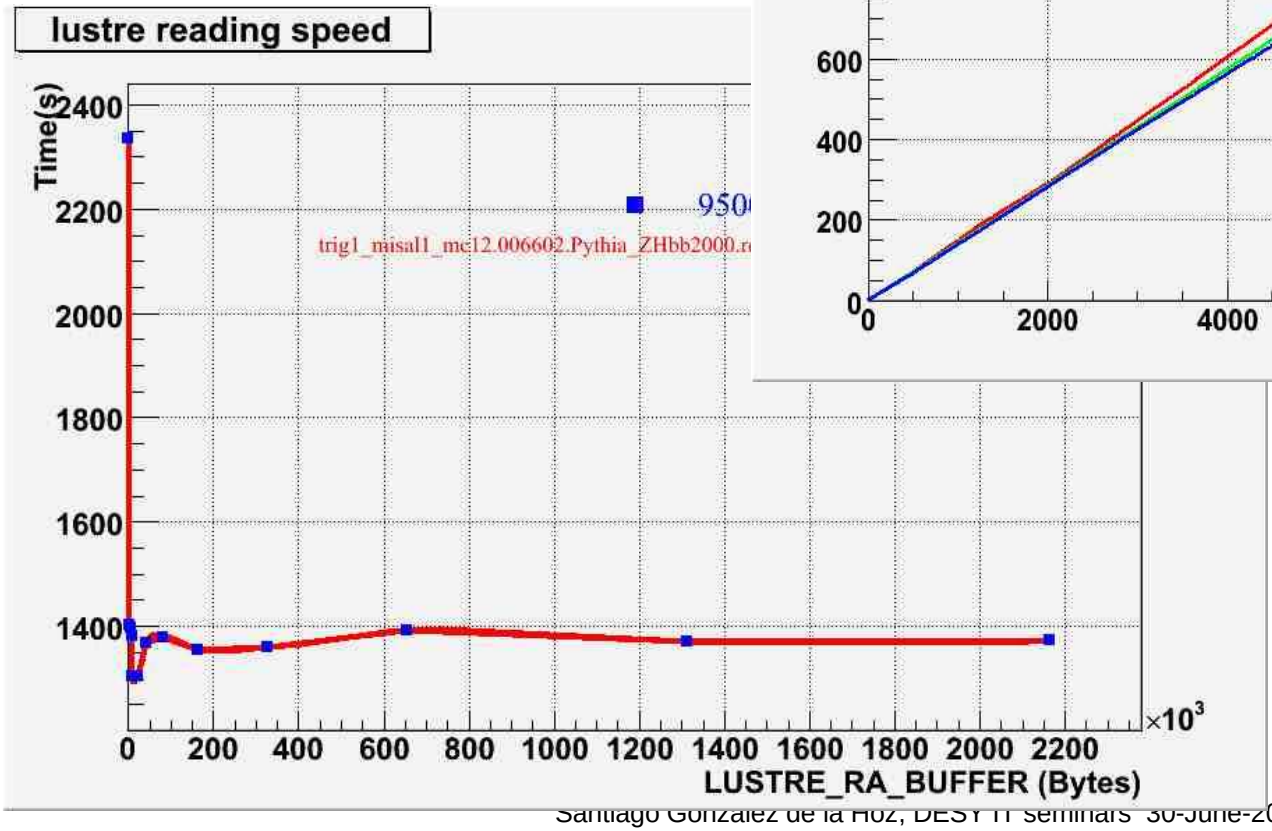
- Plans
  - Add more gridftp servers as demand increases
  - Move the Lustre server to a High Availability hardware
  - Add more disk to cope with ATLAS requirements and use
  - Performance tuning



Lustre Tuning

Tests:
RFIO without Castor

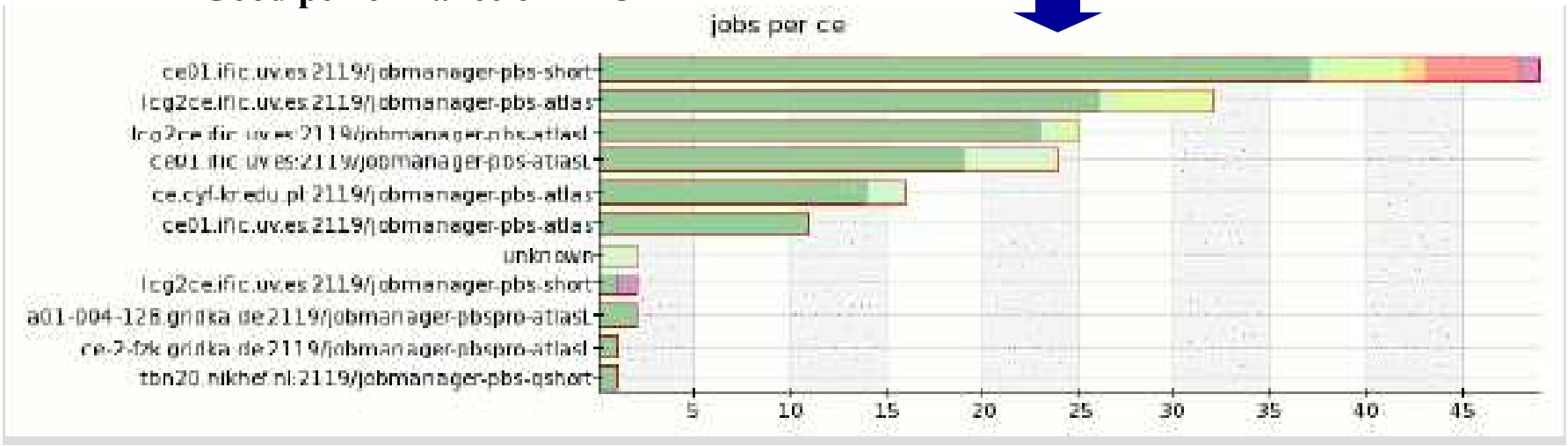




Athena analysis 12.0.6
AOD's
4 MB/s CPU limited and
Athena

Spanish Distributed Tier2: Analysis Use cases (1): MC production

- **AANT ntuple (DPD) generation from a given dataset to analyze events with semileptonic channels to study Top Polarization**
  - **Challenge: to process 595.550 events**
  - **Distributed Analysis is needed**
  - **GANGA and TopView (Analysis software) are used**
  - **Several sites in the game: IFIC (mainly), Lyon (France), FZK (Germany), CYF (Poland)**
  - **Good performance of IFIC**



Spanish Distributed Tier2: Analysis Use cases (2): AOD Analysis

- **Analysis of AOD for the process Z_H -> t tbar (Little Higgs)**
  - **4 ROOT files x 5000 of event generation in local**
  - **80 files AOD x 250 reconstruction events in GRID (20000 evts)**
  - **Whole process takes 5 days to get an AOD file in local**
  - **Strategy;**
    - **To test with a small number of events**
    - **Once problems are fixed, jobs are sent to Tier-2 of 12 countries**
  - **We save time : 400 days running in local / 14 days using GRID**

Proceso	Lugar	Ficheros	Eventos	Tiempo (horas)	Nº veces enviar x fallos	Tiempo por fallos
Generación	Local	4	5000	5	0	0
Simulación+Digitalización	Grid	400	50	96	2	144
Reconstrucción	Grid	80	250	18	2	72
			Total:	119		216
			Todo tiempo:	335	Todo tiempo (días):	13.96

What is an ATLAS Tier3?

- **Summary of ATLAS Tier3 workshop in January 2008 (https://twiki.cern.ch/twiki/bin/view/Atlas/Tier3TaskForce)**
- **These have many forms: No single answer**
  - **Size**
    - **Very small ones (one professor and two students)**
    - **Very large ones (regional/national centers)**
  - **Connectivity and Access to data**
    - **A Tier3 next to a Tier1/2 looks different than a Tier3 in the middle of nowhere**
- **Basically represent resources not for general ATLAS usage**
  - **Some fraction of T1/T2 resources**
  - **Local University clusters**
  - **Desktop/laptop machines**
  - **Tier-3 task force provided recommended solutions (plural)**
    - http://indico.cern.ch/getFile.py/access?contribId=30&sessionId=14&resId=0&materialId=slides&confId=22132

What is an ATLAS Tier3?

- **LHC Tiers (T0,T1 & T2) are governed by MoU with experiments.**
- **Possible definition of a Tier3: "A facility for LHC analysis without MoU with the experiment".**
  - **An ATLAS Tier3 TaskForce could only provide advice**
- **Tier3 has Grid and local/interactive resources.**
- **The AODs are stored at Tier-2s**
- **The DPDs (10-20 times smaller than AOD) reside at Tier-3s, and analyze with ROOT + AthenaROOTAccess**
  - **20 TB = 100% of 10^9 events in 20KB/event DPD format (1 LHC year)**
- **Rough size estimate:**
  - **25 cores and 25 TB per analysis**
    - **To produce in one month all the DPDs (20KB/event) and Plotting in one day**

What is an ATLAS Tier3?

- **Operating System**
  - **ATLAS SW best supported for SL(C)(4)**
- **Installation depends on the size of your setup**
  - **Very few machines: Installation by hand**
  - **Large cluster: You need installation and maintenance tools (i.e Quattor)**
- **PROOF: Parallel ROOt Facility to analyze DPDs**
  - **Exploiting intrinsic parallelism to analyze data in reasonable times**
  - **Show good scalability: Will scale with the 5-10 average user/analysis number of an institute**

What is an ATLAS Tier3?

Storage System

- **A Tier3 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:**

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

- **For your choice, consider what is available and check support!!**
- **Evaluation of different system on going at CERN, follow their work**

What is an ATLAS Tier3?

- **Minimal Tier-3 requirements (from D. Barberis)**
  - **A Computing Element known to the Grid, in order to benefit from the automatic distribution of ATLAS software releases**
  - **A SRM-based Storage Element, in order to be able to transfer data automatically from the Grid to the local storage, and vice versa**
  - **The local cluster should have the installation of:**
    - **A Grid User Interface suite, to allow job submission to the Grid**
    - **ATLAS Distributed Data Management (DDM) client tools, to permit access to the DDM catalogues and data transfer utilities**
    - **The Ganga/pAthena client, to allow the submission of analysis jobs to all ATLAS computing resources**

prototype at IFIC

We started in September 2007

<table>
<tr><td rowspan="2">Desktop
or
Laptop
(I)</td><td>Atlas Collaboration Tier2 resources (Spanish T2)
(II)</td></tr>
<tr><td>Extra Tier2: ATLAS Tier3 resources (Institute)
(II)</td></tr>
<tr><td colspan="2">PC farm to perform interactive analysis (Institute)
(III)</td></tr>
</table>

Tier3 IFIC prototype: Step 1 (each user desktop):

- Desktop (similar to lxplus at CERN):

a) Access to ATLAS software via AFS-IFIC

/afs/ific.uv.es/project/atlas/software/releases

(Athena+Root+Atlantis+ …)

- **Local checks, to develop analysis code before submitting larger jobs to the Tier-1s-2s via Grid**

b) User Interface (UI) (Glite; middleware Grid)

- **To find data and copy them**
- **To send jobs to the Grid**

Tier3 IFIC prototype: Step 1 (each user desktop)

- Every PC can be an **User Interface (UI)** in AFS:
  - source /afs/ific.uv.es/sw/LCG-share/sl4/etc/profile.d/grid_env.sh
  - source /afs/ific.uv.es/sw/LCG-share/sl4/etc/profile.d/grid_env.csh

  - source /afs/ific.uv.es/sw/LCG-share/sl3/etc/profile.d/grid_env.sh
  - source /afs/ific.uv.es/sw/LCG-share/sl3/etc/profile.d/grid_env.csh

- To use the **dq2(DDM)** client in AFS:
  - source /afs/ific.uv.es/project/atlas/software/ddm/current/dq2.csh
  - source /afs/ific.uv.es/project/atlas/software/ddm/dq2_user_client/setup.sh.IFIC
  - dq2-list-dataset-site IFIC

- And the **Ganga**:
  - source /afs/ific.uv.es/project/atlas/software/ganga/install/etc/setup-atlas.sh
  - Source /afs/ific.uv.es/project/atlas/software/ganga/install/etc/setup-atlas.csh

- And of course to use the Tier2 UI (ui02 & ui03.ific.uv.es)

Tier3 IFIC prototype: Step 1 (each user desktop)

- Another possibility discussed in ATLAS Tier3 task force:

(https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasComputing?topic=Tier3TaskForce)

a) To install some User Interfaces and at least one CE dedicated to the Tier3:

- **To have the ATLAS software (production releases & DDM tools) installed automatically**
- **The user has to login in the UI's and they can send jobs to the Grid**
- **It is possible to ask for development releases installation**
- **In our case, every UI can see "Lustre" (/lustre/ific.uv.es/grid) as a local file system (Useful to read files).**

b) The Ganga client is installed in AFS

Tier3 IFIC prototype: Step 2 (Resources coupled to Tier2)

- Nominal:

a) ATLAS Collaboration Resources: TB (SE) y CPU (WN)

- Extra resources(Tier3):

a) **WNs & SEs used by IFIC users and the collaboration**

b) **This assures free resources to "private" send jobs:**

1) Private AOD and DPD productions

2) AOD analysis on the Grid

3) To store "private" data or interesting data for analysis

c) **Using different/share queues**

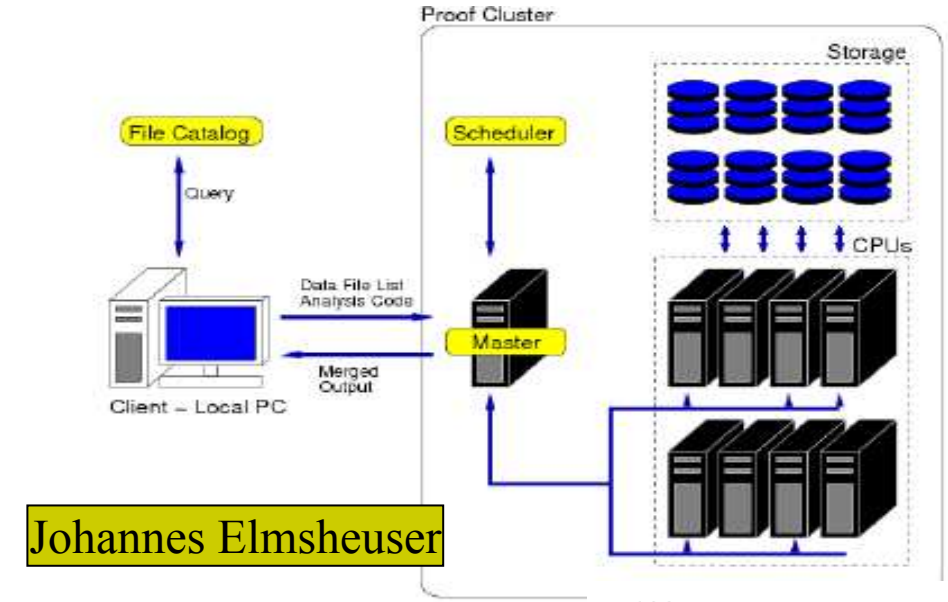
Tier3 IFIC prototype: Step 3 (PC farm outside Grid)
(It is being deployed)

- Interactive analysis on DPD using ROOT-PROOF
- Install **PROOF** in a PC farm:
  a) Parallel ROOT facility. System for interactive analysis of very large sets of ROOT (DPD) data files.
  b) Outside the Grid
  c) ~20-30 nodes
  d) Fast Access to the data: Lustre and Storm
     a) To use same technology as in our Tier2

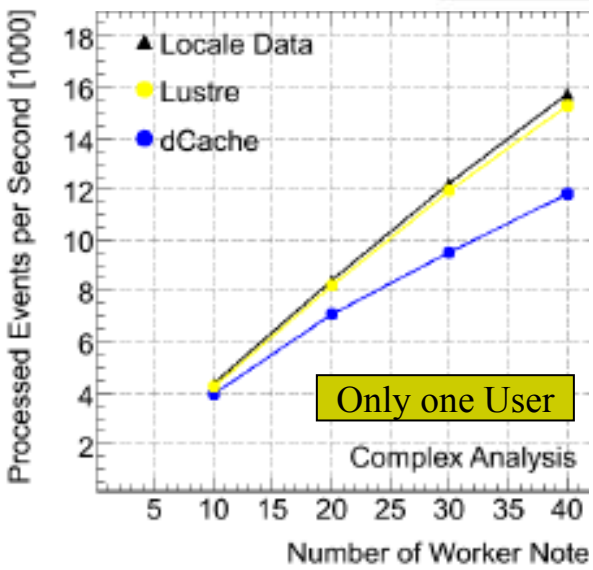
Tier3 IFIC prototype: Step 3 (PC farm outside Grid)

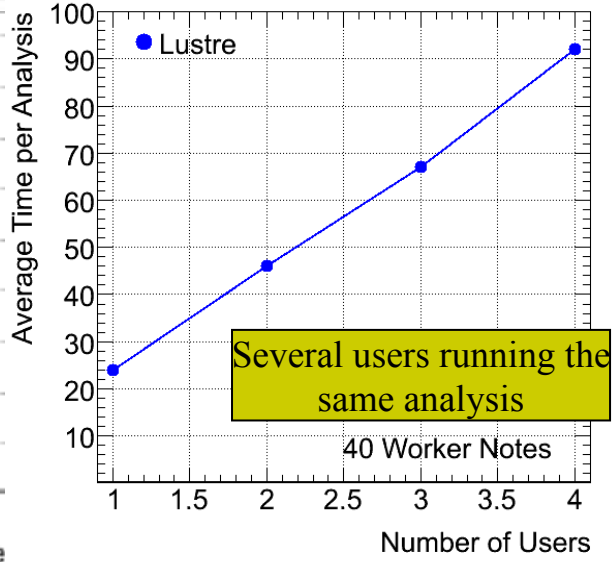
- We have just bought 4 machines to install PROOF and make some tests with Lustre:
  - PE1950 III 2 Quad-Core Xeon E5430 2.66GHz/2x6MB 1333FSb
  - 16 GB RAM (2GB per core; 8x2GB)
  - 2 HD 146 GB (15000 rpm)
- Under disk and CPU performance tests:
  - RAID0 (*Data Stripping*): distributes data across several disks in a way which gives improved speed and full capacity, but all data on all disks will be lost if any one disk fails.
    - Hardware y Software

PROOF and Lustre in Munich



Johannes Elmsheuser





- DPD production from AOD and then analyse with PROOF
- PROOF take care of the parallel processing at a local cluster
- 10 nodes used:
  - 2 dual core processor
  - 2.7 GHz & 8 GB RAM
- Dataset ($Z \rightarrow e^+e^-$) with 1.6 millions of events
- 4 KB per event, in total 6GB
- The data was stored locally at each node, on a Lustre and a dCache-file system, respectively.
- Number of processed events depending of the number of slaves processes is showed.

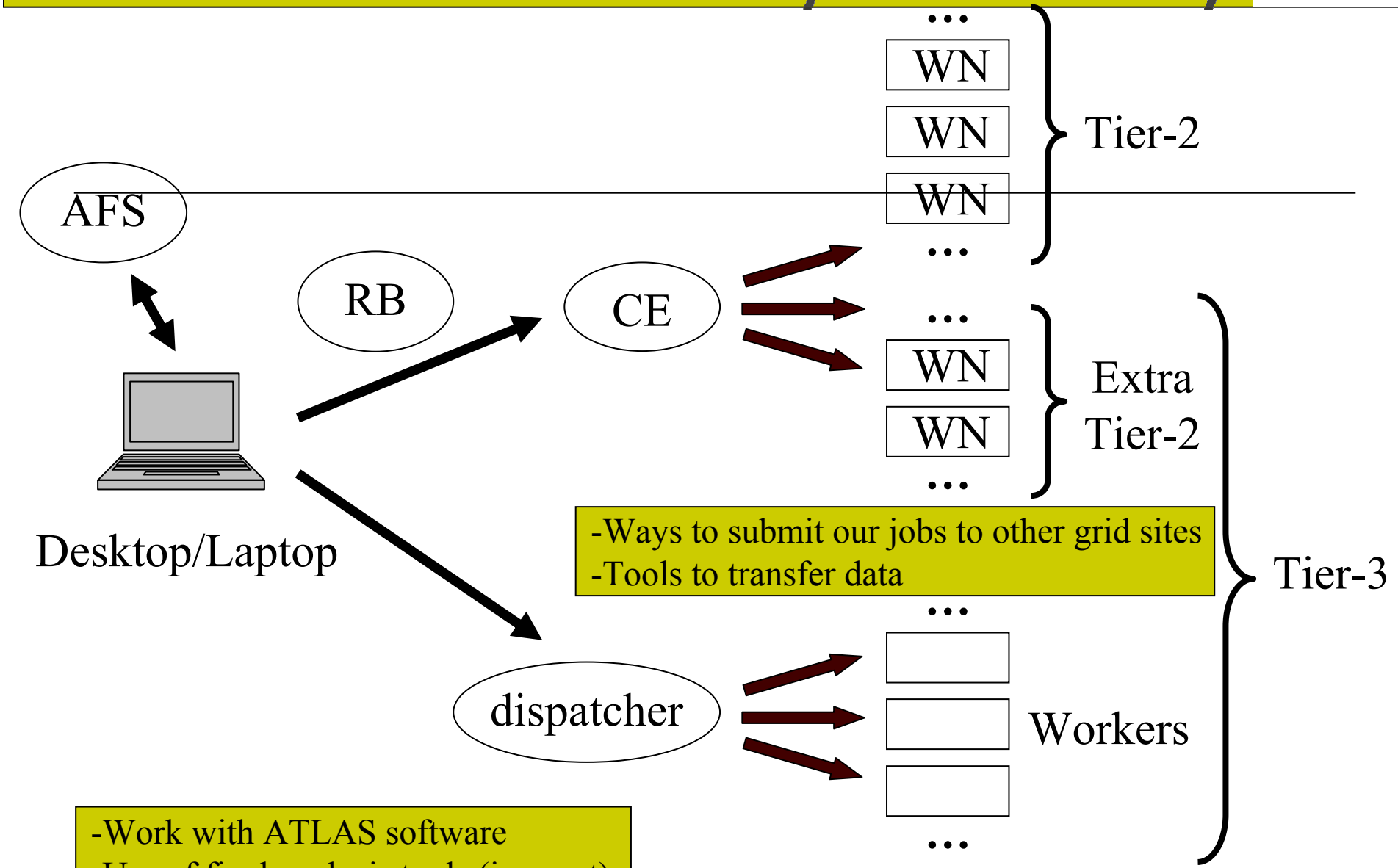
The Lustre filesystem shows a nearly equivalent behaviour as the local storage. dCache performed in this test not as good as the others.

- dCache performance was not optimised, since many files were stored in the same pool node

We could observe anearly linear speed up to 40 nodes.

As expected, the average total time for one analysis is proportional to the number of users.

IFIC Valencia Analysis Facility



-Work with ATLAS software
-Use of final analysis tools (i.e. root)
-User disk space

Tier3 IFIC prototype: Typical use of a Tier-3

- Interactive analysis on n-tuples.
  - It does require access to the data when these n-tuples are generated.
- Development of analysis code.
  - This would motivate a local copy of a small number of data (AOD or even ESD).
- Running small local test jobs before submitting larger jobs to the Tier-1s or Tier-2s via the Grid.
  - This would motivate similar copies of the data as above.
- Running skimming jobs of the Tier-1 and Tier-2s via the Grid, and copying the skimmed AOD back to the Tier-3 for further analysis.
  - The output of this skim must be a very small subset of the AOD.
- Analyzing the above skimmed data using the ATLAS framework (Athena)
- Production of Montecarlo samples of special interest for the local institution

Conclusions

- At IFIC in Valencia we are proposing a possible Tier-3 configuration and software setup that matches the requirements according to the DPD analysis needs as formulated by the ATLAS analysis model group.
  - Some local resources, beyond Tier-1s and Tier-2s, are required to do physics analysis in ATLAS.
  - These resources could consist of workstations on each physicist's desk or computer farms, and could be operated as a shared facility provided by the institution own resources.
- Support from the Tier-1s and Tier-2s to such Tier-3 centres in term of expertise (installation, configuration, tuning, troubleshooting of ATLAS releases and the Grid stack) and services (data storage, data serving, etc.) is very important.

Backup

Stream de datos

Lumi	egamma	jetTauEtMiss	muon	minBias	bphysics
10^{31}	✓	✓	✓	✓	✗
10^{32}	✓	✓	✓	✓	✓

Streams	egamma	muon	jetTauEtmiss	cósmicos	bphysics
Análisis	SUSY,Top, exóticos y taus	Alineamiento, SUSY,Top, exóticos, taus, Higgs MSSM, Tilecal calibración	SUSY, Top, exóticos, Higgs MSSM y taus	Alineamiento, Commisioning, calibración y optimal filtering	Alineamiento y calibración (lumi 10^{32})
ESD	Pequeñas muestras (SUSY)	Pequeñas muestras (SUSY) y grandes muestras (Top)	Pequeñas muestras para SUSY y para Top	Muestras para el alineamiento y commisioning (pequeña cantidad en disco)	Muestras para el alineamiento (pequeña cantidad en discos)
AOD	SUSY, Top y exóticos	SUSY, Top, exóticos y Higgs MSSM	SUSY, Top y exóticos	X	X
DPD	SusyView, TopView y EventView	SusyView, TopView, EventView y AtautauView	SusyView, TopView y EventView	X	X

Requisitos grupos de física de Valencia

- Análisis
  - No necesitan express stream data
  - Varias versiones de athena instaladas
  - Máquinas para análisis de AOD y DPD (y quizás algunos ESD)
  - Espacio en disco
  - Producciones de Montecarlo (AOD y DPD)
- Detector
  - Si necesitan los express stream data utilizarían la CAF
  - Comissioning (esto seria de inmediato, antes de septiembre):
    - Versiones estables del software para reconstrucción
      - Runes de calorímetros e Inner detector
      - Reconstrucción del Tier0 y Tier1 ESD y DPD
      - Raw Data (Run M6 40 TB)
  - Alineamiento
    - Alamcenamiento en disco
  - Calibración y optimal filetering del Tilecal
    - Básicamente en la CAF del CERN (necesitan 12 CPUs)
    - Si necesitan los express y otros data stream, también lo harían en la CAF

Preliminary list of signatures for the Express Stream

Interés de algunos grupos como Tilecal de Valencia

Streams interesting for Physics/Detector performance preferences of the group

	Decay or signature	Motivation
	$Z \to l^+l^-$	calibration and data quality
	minimum bias	data quality
	lepton pair with high mass	alert on rare events
	$B \to \mu^+\mu^-$	alert on rare events
	≥ 3 high p_{T} leptons	alert on rare events
✓	lepton + jets + ETmiss	calibration
	$W \to l\nu$	calibration and data quality
✓	large missing E_{T}	alert on rare events
	lepton with large p_{T}	alert on rare events
✓	large ΣE_{T}	alert on rare events
	large M_{eff}	alert on rare events
	high multiplicity of trigger objects	alert on rare events

¿Qué son Express Stream?

http://indico.cern.ch/getFile.py/access?contribId=2&resId=0&materialId=0&confId=a06527

- Motivations for the Express Stream
  - Calibration
    - "Physics calibration data streams for rapid processing
  - Check of general data quality
    - Only sample that can be used for as rapid and complete check of the data quality
  - Rapid alert on interesting physics events
    - Monitoring about rare avents while the data are being taken.
- Pero tienen una vida de 48 horas, después existe el Stream

- **¿El Tier-2 o Tier-3 tiene que tener datos de este tipo?**
  - **¿Se van a hacer análisis por parte de nuestros usuarios sobre este tipo de datos? EN PRINCIPIO NO, si es SI parece que seria en el CERN**
  - **¿ Y los Stream of data? SI, ver tabla de principio presentación**

Requisitos mínimos para un Tier3

- The ATLAS software environment, as well as the ATLAS and grid middleware tools, allow us to build a work model for collaborators who are located at sites with low network bandwidth to Europe or North America.
- The minimal requirement is on local installations, which should be configured with a Tier-3 functionality:
  - A Computing Element known to the Grid, in order to benefit from the automatic distribution of ATLAS software releases
  - A SRM-based Storage Element, in order to be able to transfer data automatically from the Grid to the local storage, and vice versa
- The local cluster should have the installation of:
  - A Grid User Interface suite, to allow job submission to the Grid
  - ATLAS DDM client tools, to permit access to the DDM data catalogues and data transfer utilities
  - The Ganga/pAthena client, to allow the submission of analysis jobs to all ATLAS computing resources

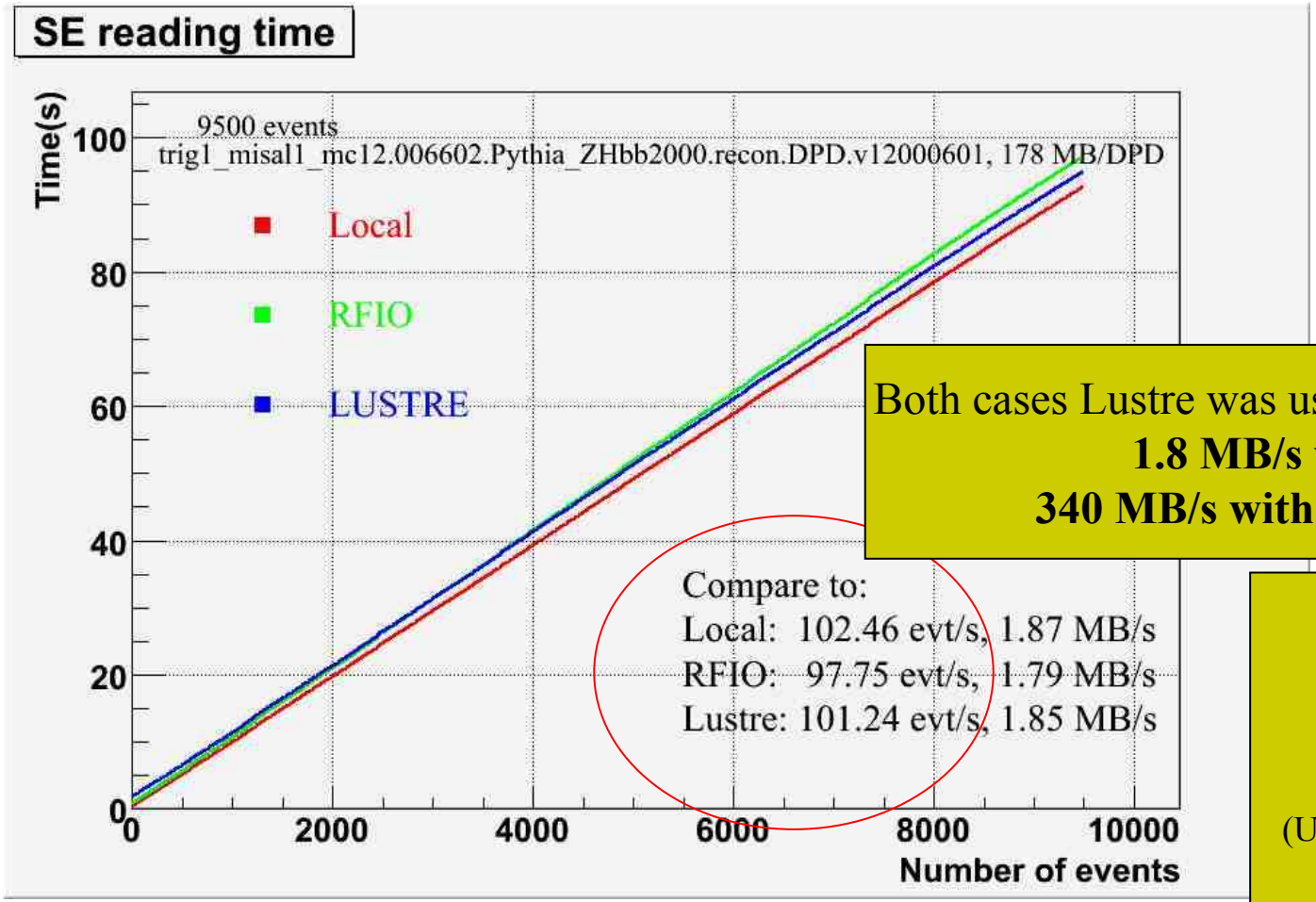
Nos lo da nuestro Tier2

Tenemos del Tier2, tendría que haber para el Tier3

Instalado en el IFIC

Dario Barberis: The ATLAS Computing Model

The same test with DPD



Both cases Lustre was used and data in the cache
1.8 MB/s with Root
340 MB/s with a simple "cat"

2 x Intel Xeon 3.06GHz
4 GBytes RAM
1 NIC Gigabit Ethernet
HD: ST3200822AS
(Using a 3Ware card: 8006-2LP)

The Express Stream of ATLAS Data

Algunos grupos han expresado su interés (ej Tilecal Valencia)

- Need to define what calibration and express streams the group would like to have replicated in Spain. Need to develop common strategy with this other Spanish groups and the Tier1
  - As TileCal experts the groups needs calibration streams.
    - TileCal community is in the process of defining the composition and specifics of calibration streams
      - Preliminarily: laser triggers or the charge injection triggers generated in the empty bunches. Use muon streams
  - At the same time, due to physics interests and for the sake of competitiveness, the group also needs access to express streams (see next slide)

Summary

- From her/his desktop/laptop individual physicist can get access to:
  - IFIC Tier2-Tier3 resources.
  - ATLAS software (Athena, Atlantis, etc..), DDM/dq2 and Ganga tools.
- IFIC Tier3 resources will be split in two parts:
  - Some resources coupled to IFIC Tier2 (ATLAS Spanish T2) in a Grid environment
    - AOD analysis on millions of events geographically distributed.
  - A PC farm to perform interactive analysis outside Grid
    - To check and validate major analysis task before submitting them to large computer farms.
    - A PROOF farm will be installed to do interactive analysis.

¿Qué son Express Stream?

http://indico.cern.ch/getFile.py/access?contribId=2&resId=0&materialId=0&confId=a06527

- The Express Stream, like any other stream of data, will be defined by the trigger selection criteria.
- The computing model assumes that the first-pass event reconstruction will be completed in 48 hours since the data was taken (in the Tier-0 operations).
  - The bulk of the processing will begin after 24 hours.
- The data of the Express Stream, and of the calibration estreams, will be reconstruct in less than 8 hours.
- It will be possible to achieve feedback from the Express Stream in a few hours for the following reasons:
  - The data volume will be small (15% of the total)
  - Existing calibration constants can be used
  - A regular and rapid offline processing of the Express Stream will be made a part of the operation,
  - Useful feedback can be obtained already online, by monitoring the trigger rates that contribute to the Express Stream

Table 2-2 The assumed event data sizes for various formats, the corresponding processing times and related operational parameters.

Item	Unit	Value
Raw Data Size	MB	1.6
ESD Size	MB	0.5
AOD Size	kB	100
TAG Size	kB	1
Simulated Data Size	MB	2.0
Simulated ESD Size	MB	0.5
Time for Reconstruction (1 ev)	kSI2k-sec	15
Time for Simulation (1 ev)	kSI2k-sec	100
Time for Analysis (1 ev)	kSI2k-sec	0.5
Event rate after EF	Hz	200
Operation time	seconds/day	50000
Operation time	days/year	200
Operation time (2007)	days/year	50
Event statistics	events/day	10^7
Event statistics (from 2008 onwards)	events/year	$2 \cdot 10^9$

- AOD Event/year:
  - **200TB**
- DPD Event/year
  - 10KB/event → **20TB**

- El Tier2 español en 2008 y 2009 dispondrá de 387 y 656 TB
  - ¿Suficiente para nuestro análisis?

- **Para el modelo de análisis:**
  - **¿Tenemos que almacenar todos los AODs? Tier 2 solo 1/3**
  - **¿Cuantas versiones de ellos?**
  - **¿Y cuántos ESD? Tier1 1/10**
  - **¿Y los datos de Monte Carlo?**
  - **¿Y los DPD? En el Tier3**
- **Petición formal española para el 1/10 de ESD (Tier1) y 1/3 AOD (Tier2) a ATLAS**
- **Necesidad de más espacio (Tier3)**

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

Ejemplo de uso de un Tier3
(basado en la experiencia de Luis, Elena y Miguel)

- Análisis interactivo de Ntuplas.
  - No es necesario acceso a los datos desde donde estas Ntuplas se han generado
- Desarrollo de código de análisis.
  - Se necesita un copia local de una pequeña cantidad de sucesos ESD, AOD o quizás RAW
- Correr pequeñas pruebas locales antes de en enviar una gran cantidad de trabajos a los Tier1s o Tier2s utilizando el Grid.
  - También necesito una pequeña cantidad de datos copiados localmente, igual que el caso anterior.
  - Incluso igual necesito tener acceso a los TAG data
- Correr trabajos vía Grid en los Tier1s o Tier2s pero copiando los AOD (o quizás raramente los ESD) en el Tier3 para un posterior análisis.
- Analizar los anteriores AOD usando Athena (Con Ganga).
- Producción de muestras privadas de Monte Carlo de interés especial para los análisis que se lleven a cabo en el instituto.

¿Recursos en los Centros o en el CERN? Discutido en diversas reuniones T1-T2

- En el CERN siempre se pueden poner recursos. En principio sólo es cuestión de poner dinero, ellos gestionan esos recursos
  - Entonces, ¿por qué este modelo de *computing* jerárquico?
    - ¿Por qué no están todos los Tier-1s y todos los Tier-2s en el CERN?
- Si algún día se decide crear y mantener una infraestructura en el centro, cuanto antes se empiece a ganar experiencia mucho mejor
  - Yo gestiono estos recursos, en el CERN no.
- ¿Qué pasa con los Físicos de nuestro centro en el CERN?
  - ¿De cuánta gente estamos hablando?
    - Que utilicen los recursos que les da el CERN
    - Que se conecten a los recursos de su centro
- ¿Qué pasa con los Físicos que se quedan en su Centro?
  - Que utilicen los recursos de su centro
  - Que se conecten a los recursos del CERN

Requisitos de los físicos del IFIC

- Reunión de 9 de Mayo 2008 en Valencia
- Actividades del grupo del IFIC:
  - Análisis de Física
    - Física del Top y leptones
    - SUSY
    - Higgs en el MSSM
    - Física de Exóticos (Little y Twin Higgs)
  - Detector:
    - Commissioning (General, pero mirando calorímetros e Inner Detector)
    - Alineamiento en el SCT
    - Calibración y Optimal filtering en el TileCal