eScience Intrastructure T2-T3 for High Energy Physics data analysis

Presented by:
Álvaro Fernandez Casani (Alvaro.Fernandez@ific.uv.es)
IFIC – Valencia (Spain)

Santiago González de la Hoz , Gabriel Amorós , **Álvaro Fernández**, Mohamed Kaci, Alejandro Lamas, Luis March, Elena Oliver, José Salt, Javier Sánchez, Miguel Villaplana, Roger Vives

IFIC
INSTITUT DE FÍSICA CORPUSCULAR

Cracow '08 Grid Workshop

Outline

- Introduction to the LHC and ATLAS Computing Model
  - LHC and ATLAS experiment
  - The Event Data Model
  - A hierarchical Computing Model
- ATLAS Spanish Tier-2
  - Distributed Tier-2 Resources
  - MC production
  - Data Movement and Network
  - Storage Element
  - Analysis Use cases
  - User Support
- 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
- Spanish eScience Initiative
- Conclusions

Large Hadron Collider (LHC)

The LHC is p-p collider:

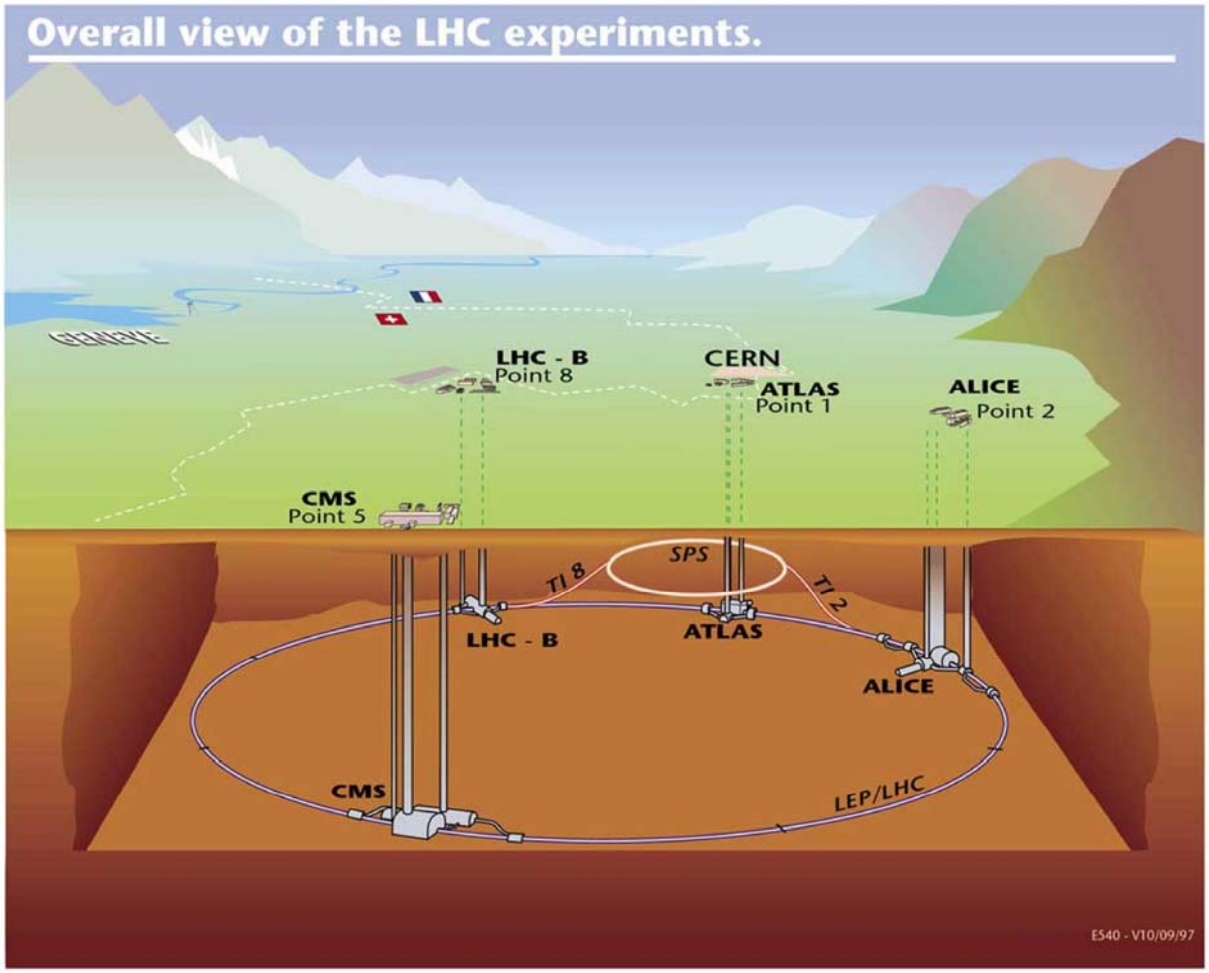
$\sqrt{s}$ = 14 TeV and

$\int \mathcal{L}$ = 10^{34} cm^{-2} s^{-1}

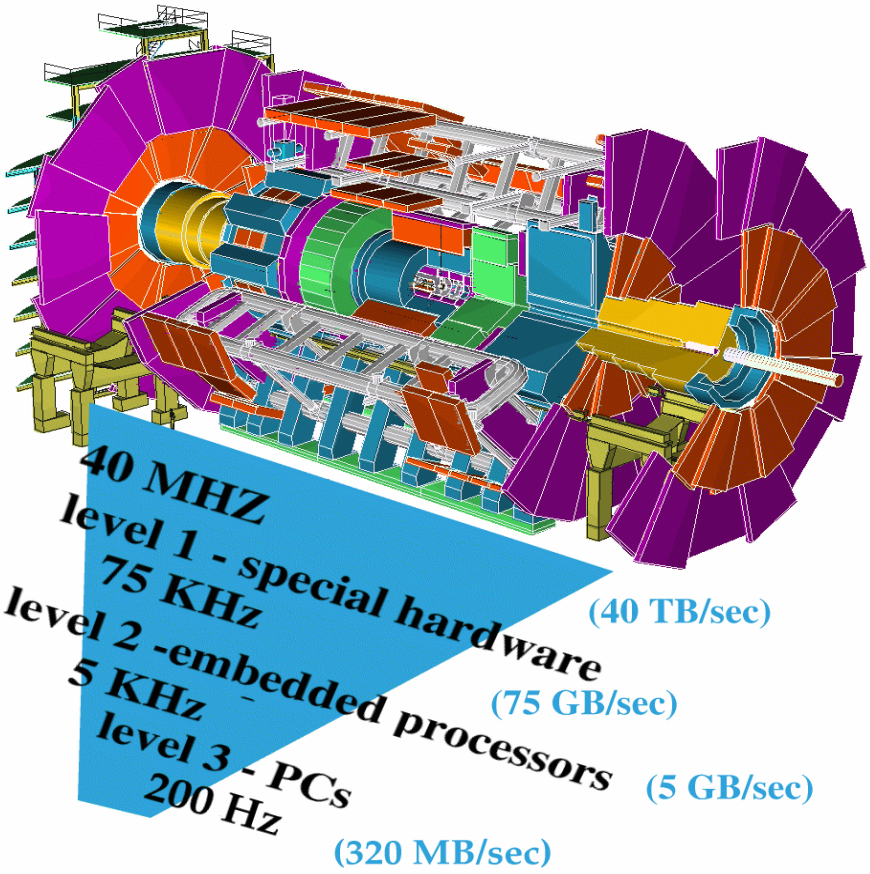
(10^{35} in high lumi phase)

There are 4 detectors:

- 2 for general purposes: ATLAS and CMS
- 1 for B physics: LHCb
- 1 for heavy ions: ALICE



ATLAS Computing



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**

↓

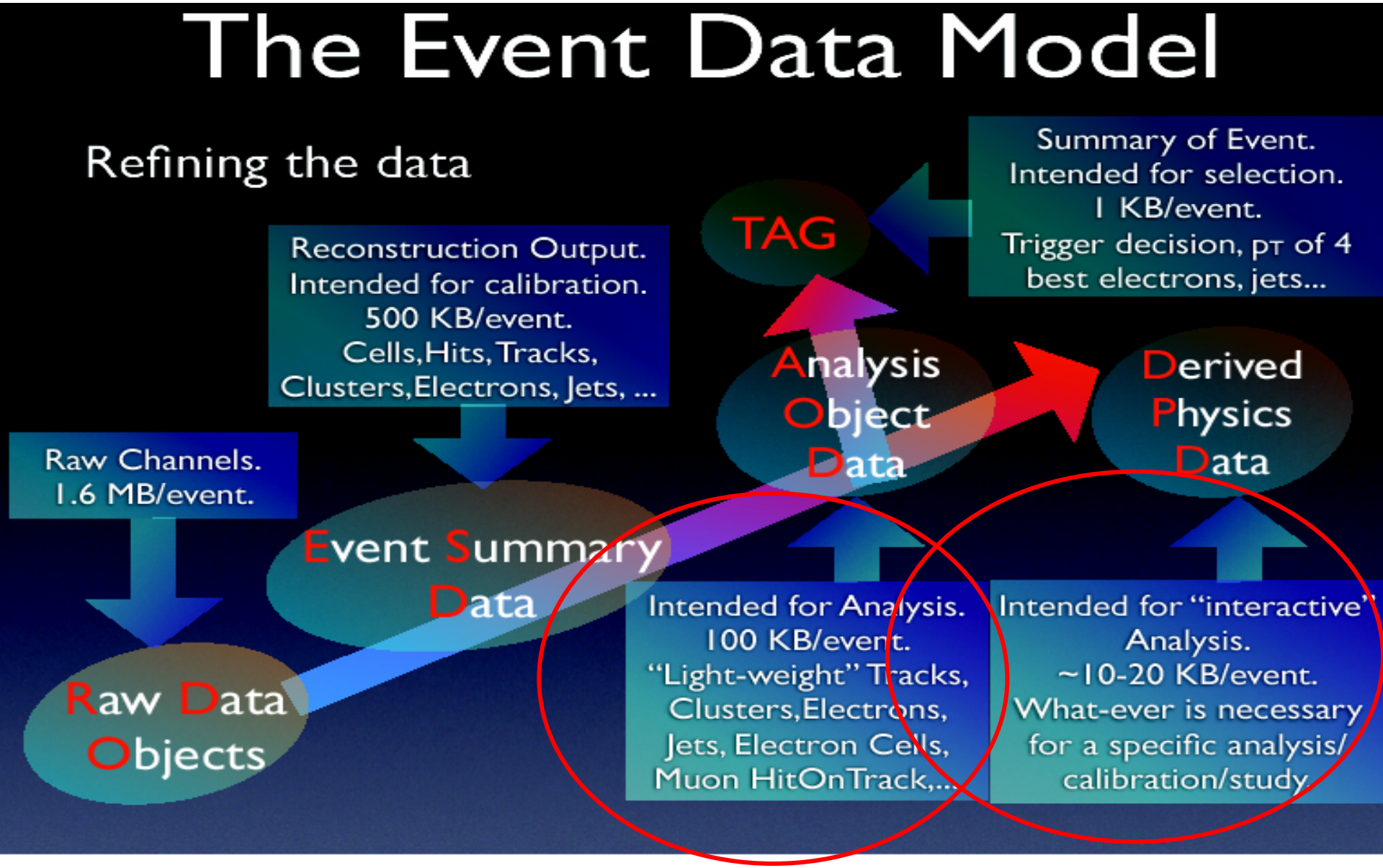
Worldwide LHC Computing Grid (WLCG)

↓

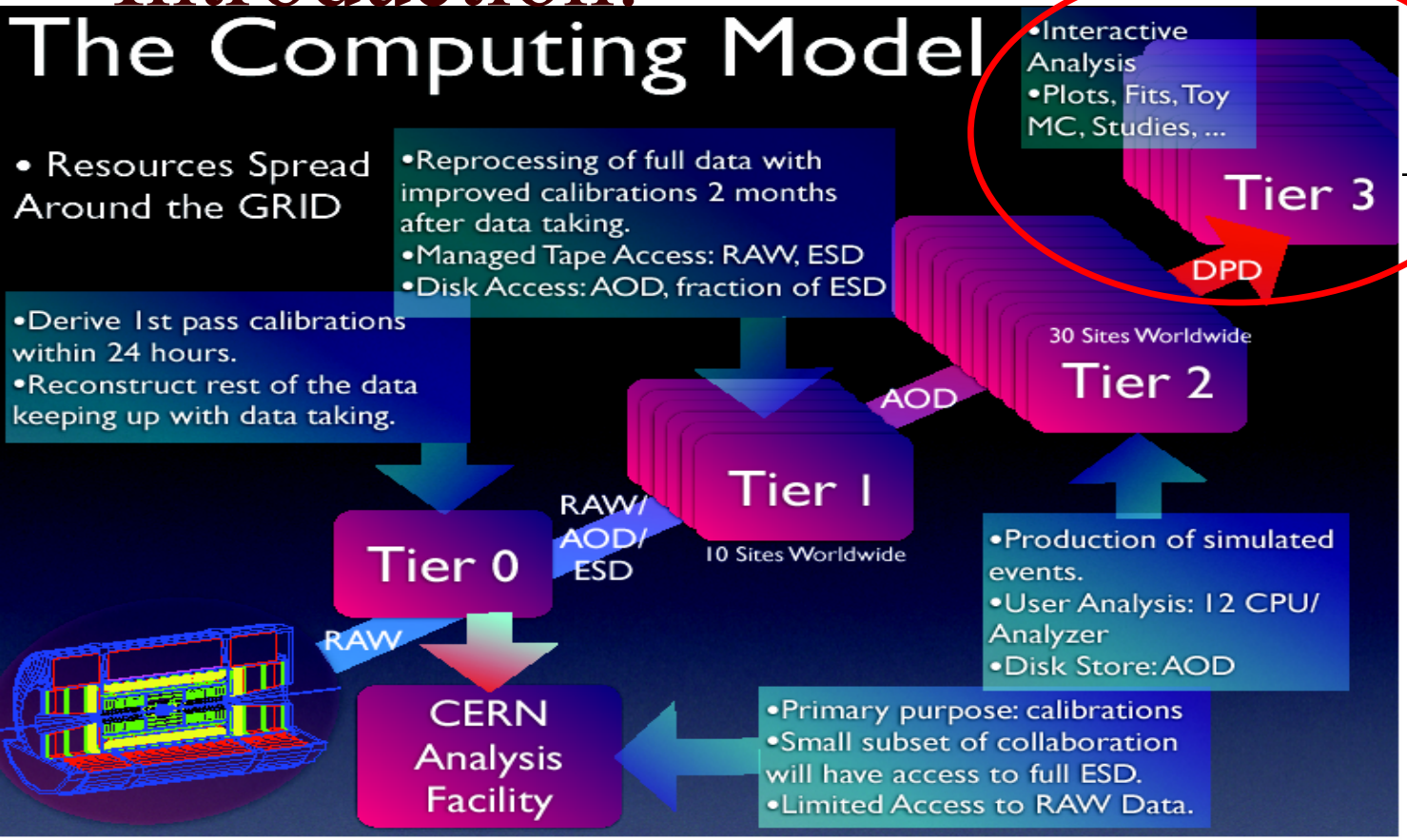
ATLAS Data Challenge (DC)

ATLAS Production System (ProdSys)

Introduction:

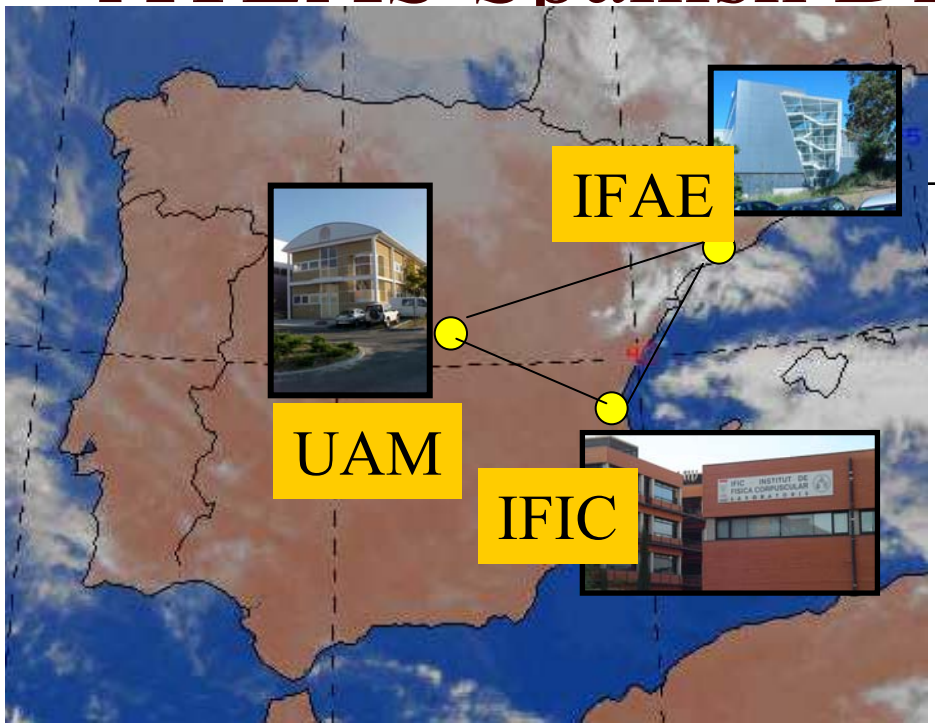


Introduction:



- 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

ATLAS Spanish Distributed Tier2



- **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**

T1/T2 Relationship

- **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)**

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 (October'08)

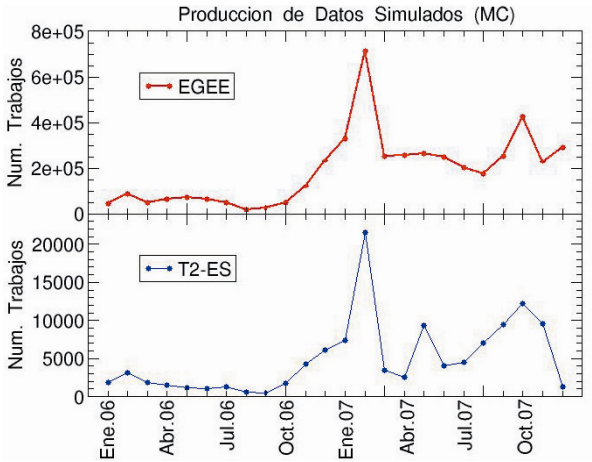
	IFAE	UAM	IFIC	**TOTAL**
CPU (ksi2k)	**201**	**275**	**132**	**608**
Disk (TB)	**104**	**100**	**36**	**240**

New acquisitions in progress to get the pledged resources

Accounting values are normalized according to WLCG recommendations

Spanish Distributed Tier2: Monte Carlo production

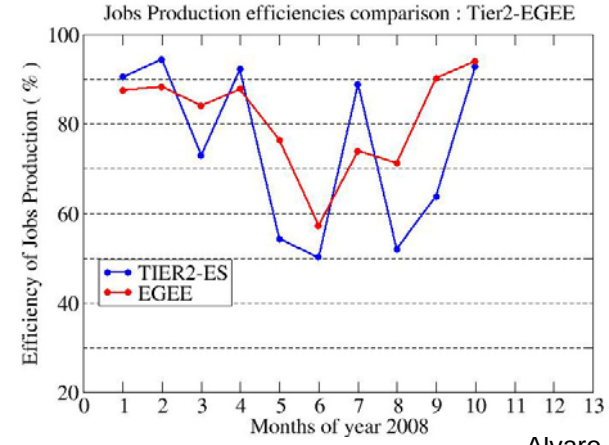
2006 & 2007



-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. Taking into account that 250 centers/institutes are participating, 10 of them are Tier-1

2008



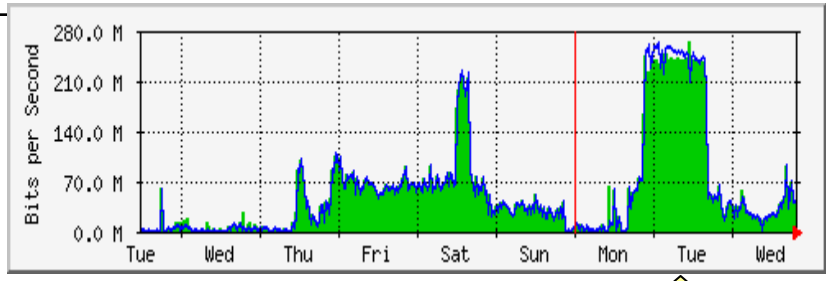
- Since January-2008, ATLAS has migrated to PANDA executor

-The production efficiency has been positively affected; the average efficiency was 50% and now is **75%-80%** @ T2-ES

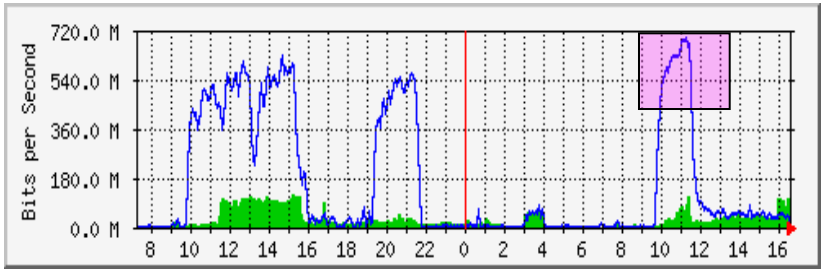
-T2-ES contribution to jobs production is being **1.6%** up to now

Spanish Distributed Tier2: Network and Data Transfer

- **It is provided by the Spanish NREN RedIRIS**
  - **(Aug'08) Connection at 10 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

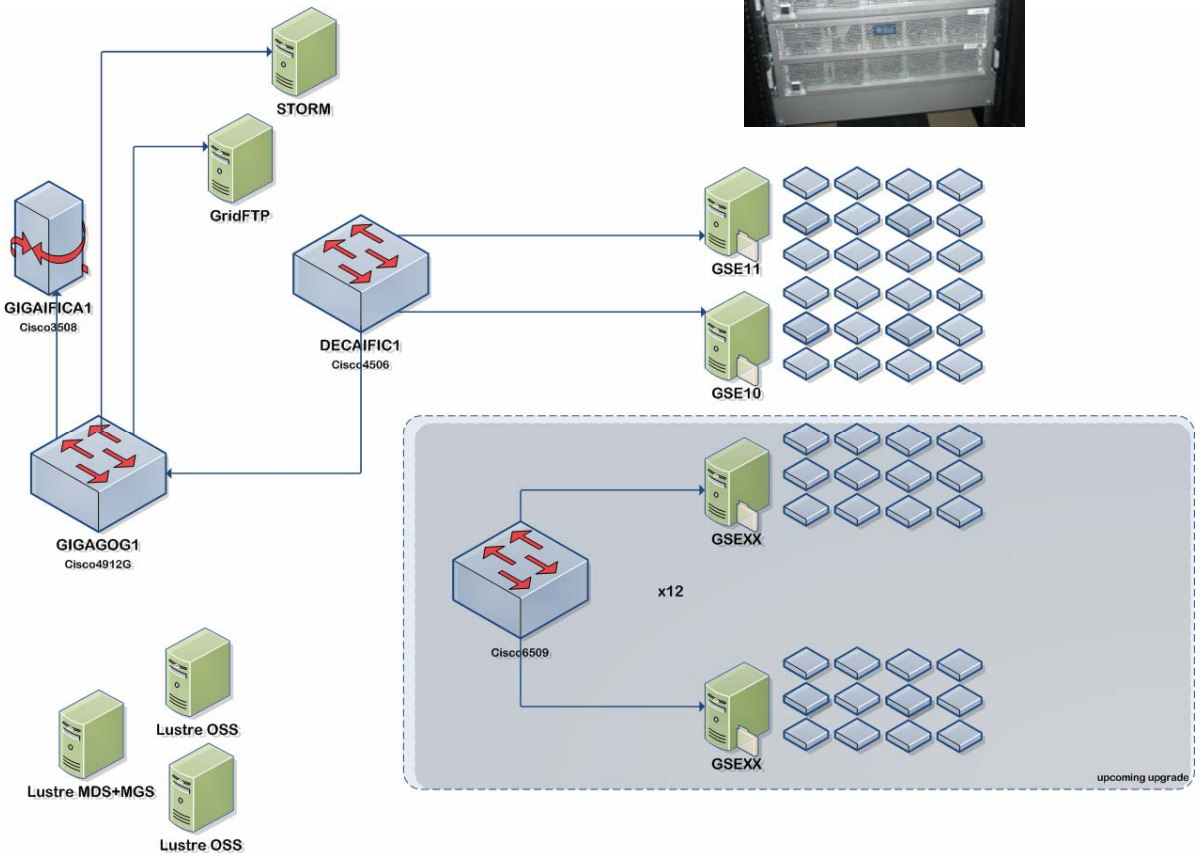
- Inside our Tier2 two SE options are used. In case that Storm/Lustre won't work as expected we will switch to dCache

IFIC- Storage Network

Sun X4500 disk servers:

- 2 servers with 48x500Gb disks ~= 36 Tb
- (installing) Tier-2 infrastructure: 9 servers with 48x500Gb disks ~= 216 Tb
- (installing) Grid-CSIC infrastructure: 5 servers with 48x1Tb disks ~= 200 Tb
- Configuration:
  - RAID 5 (5x8 + 1x6). Usage ration 80%
  - 1 raid per disk controller
  - Performance (Bonie++ tests)
    - Write: 444,585 Kb/s
    - Read: 1,777,488 Kb/s
- Network performance:

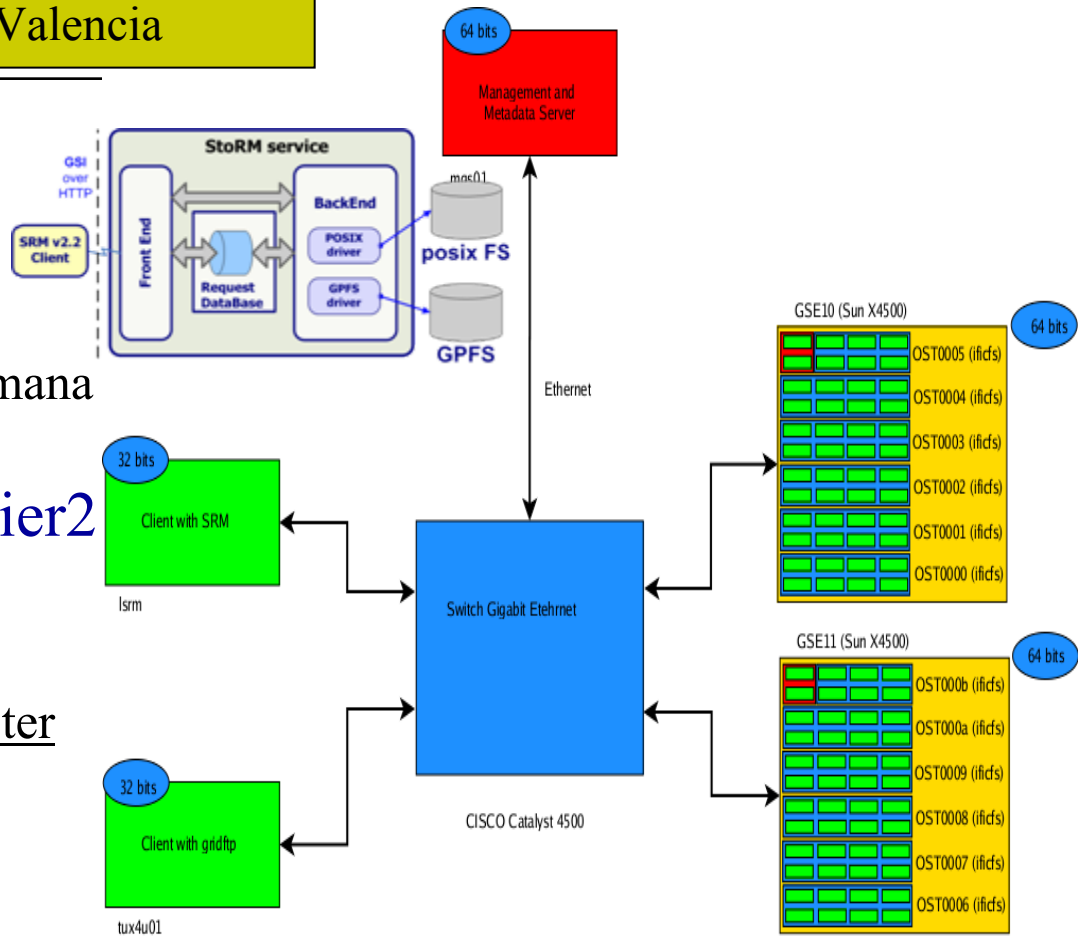
  4 Gb ports. OSS tested with channel bonding configuration ocupies link over 4 clients



StoRM + Lustre

Storage Element system at IFIC-Valencia

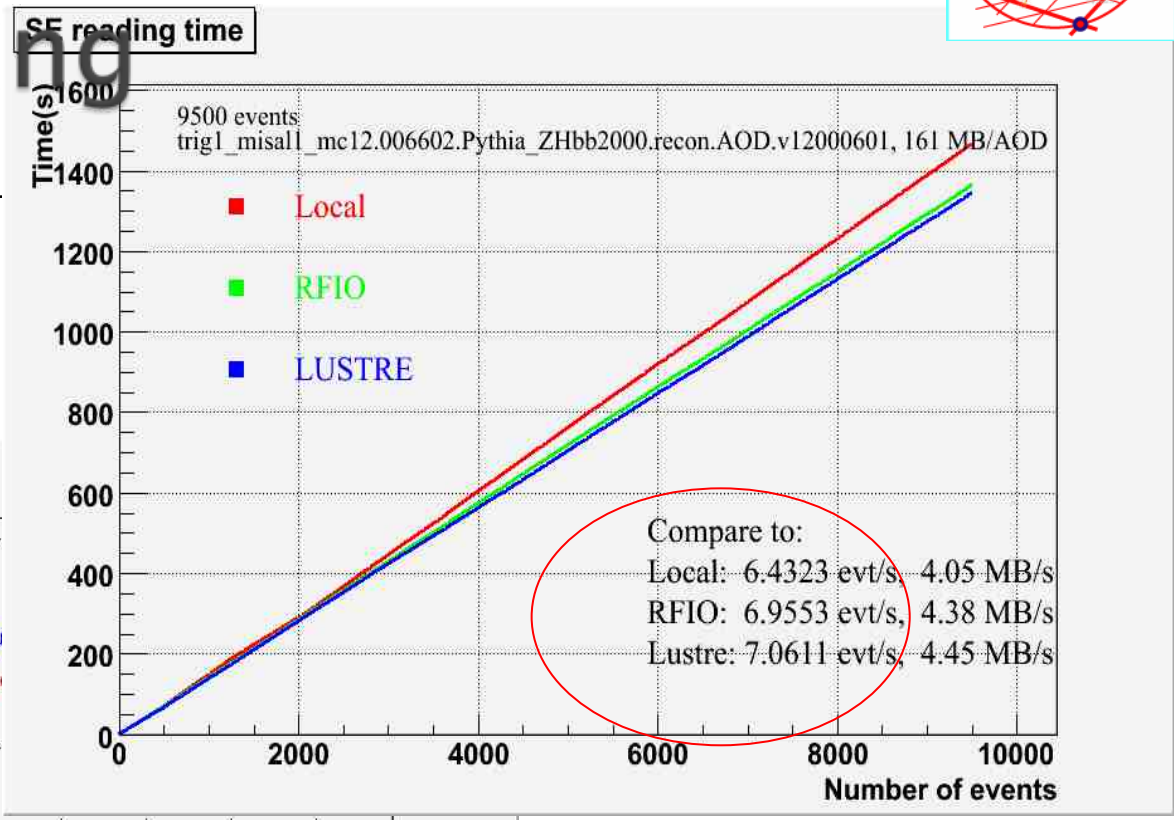
- StoRM
  - Posix SRM v2 (server on Lustre)
  - Being used in our IFIC-Tier2.
  - Endpoint: srm://srmv2.ific.uv.es:8443:/srm/managerv2
- Lustre in production in our Tier2
  - High performance file system
  - Standard file system, easy to use
  - Higher IO capacity due to the cluster file system
  - Used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

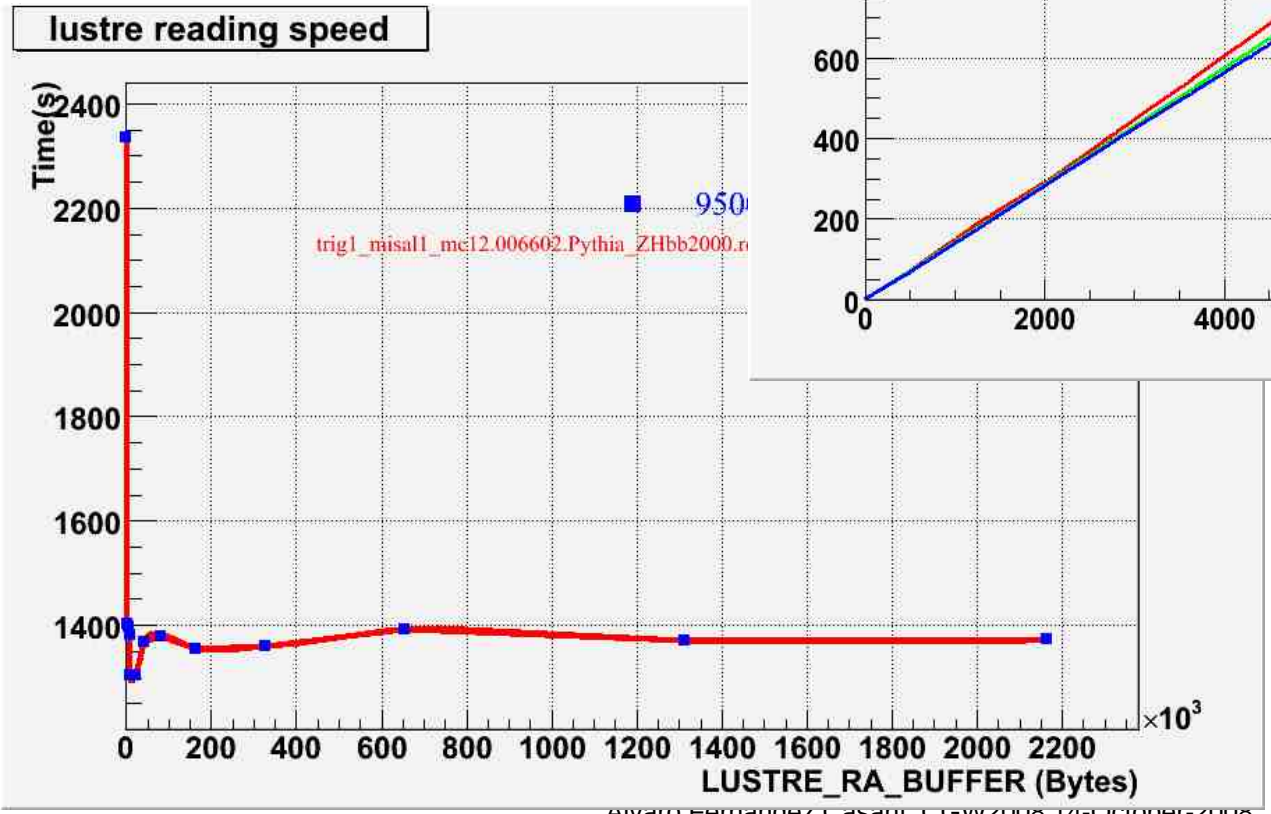


https://twiki.ific.uv.es/twiki/bin/view/Atlas/LustreStoRM

Lustre Tuning

Tests:
RFIO without Castor





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

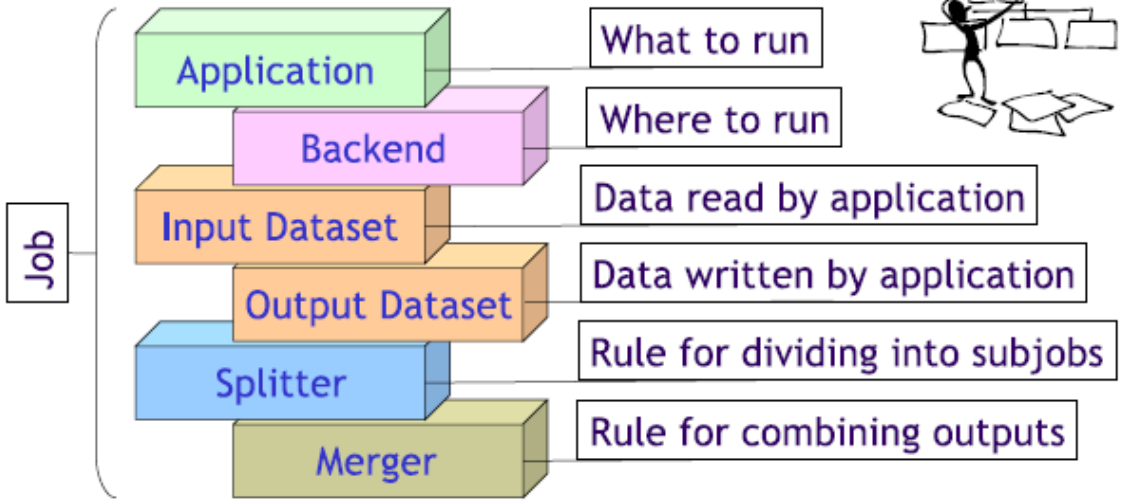
Aplication: Atlas Distributed Analysis using the Grid

- 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

Aplication: Atlas Distributed Analysis using the Grid – Current Situation

- 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)



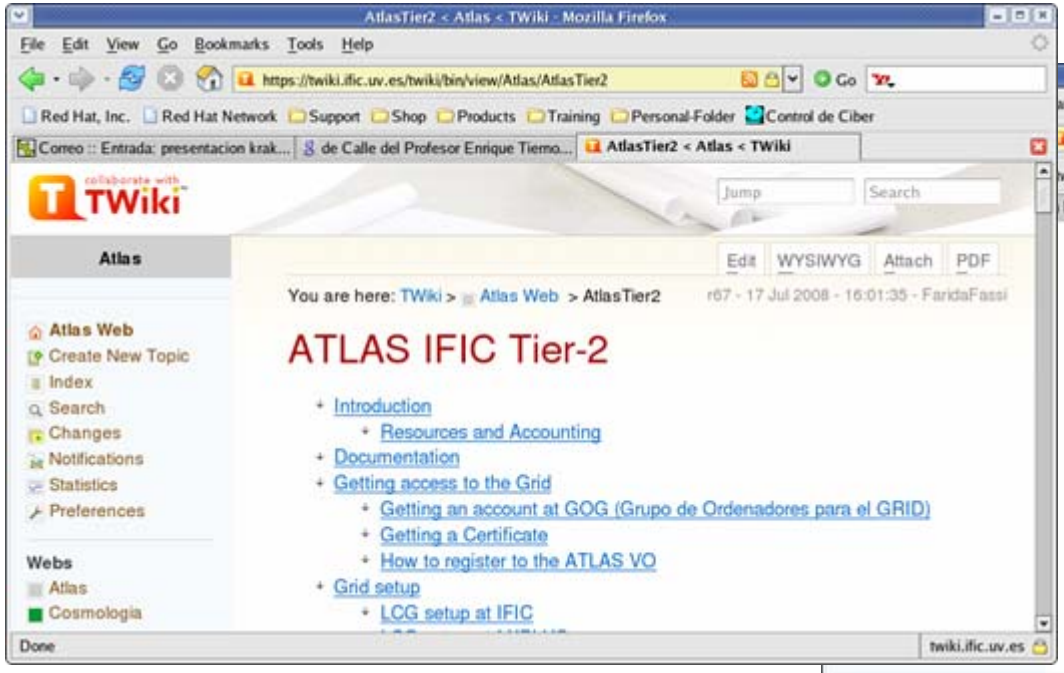
- Ganga is based on a simple, but flexible, job abstraction
- A job is constructed from a set of building blocks, not all required for every job
- Ganga offers three ways of user interaction:
  - Shell command line
  - Interactive IPython shell
  - Graphical User Interface

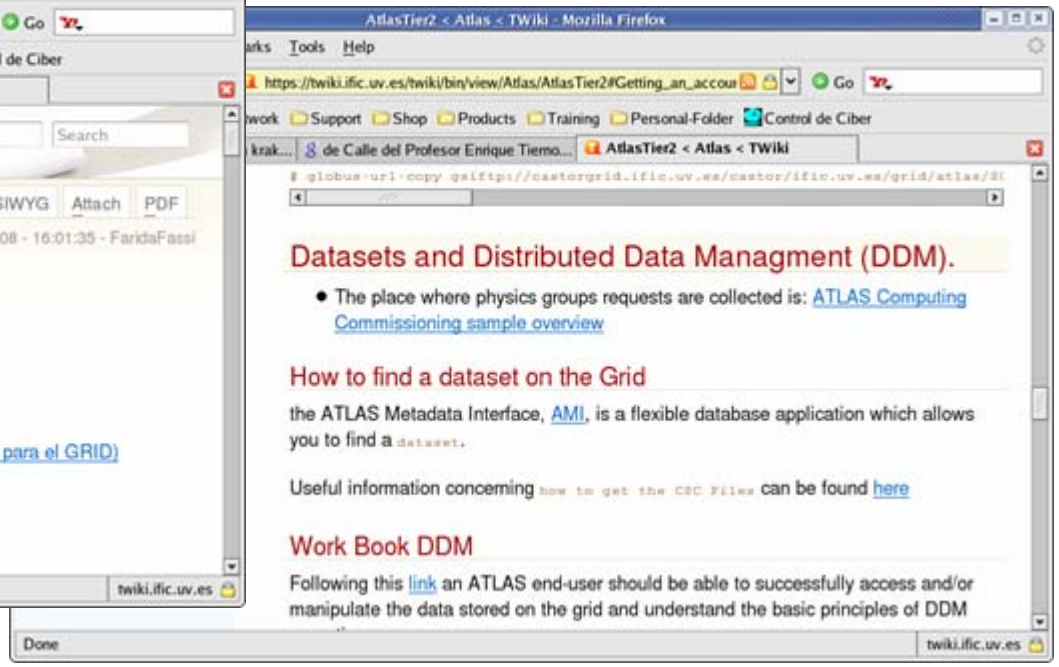
Spanish Distributed Tier2: User Support

A twiki web page has been provided in order to guide new users on how to do for using the grid and local ressources for the atlas data analysis

https://twiki.ific.uv.es/twiki/bin/view/Atlas/AtlasTier2

It is envisaging the introduction of a Ticketing-System-like





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

Tier3 IFIC prototype: user access

- Discussed in ATLAS Tier3 task force and currently taken:

(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:

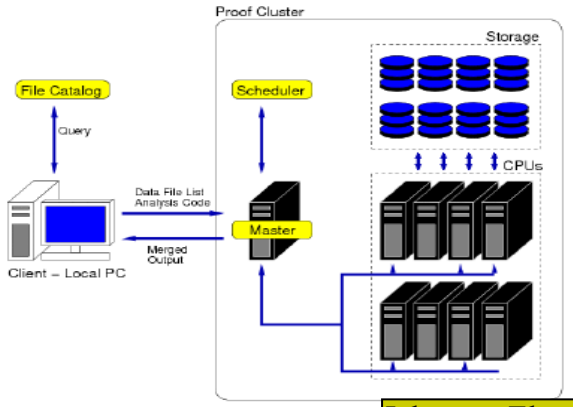
- **To have the ATLAS software (production releases & DDM tools) installed automatically**
- **The user has to login in the UI's and they can send jobs to the Grid**
- **It is possible to ask for development releases installation**
- **In our case, every UI can see "Lustre" (/lustre/ific.uv.es/grid)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

b) SEVIEW developed for https access

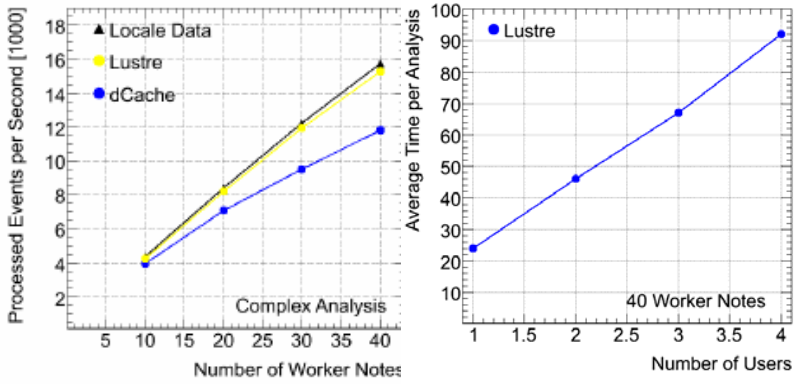
c) The Ganga client is installed in AFS

Tier3 IFIC prototype: (PC farm outside Grid)

- Interactive analysis on DPD using ROOT-PROOF
  a) Outside the Grid
  b) ~20-30 nodes
- 4 machines to install PROOF and make some tests with Lustre:
  - DELL PE1950 III 2 Quad-Core Xeon E5430 2.66GHz/2x6MB 1333FSb
  - 16 GB RAM (2GB per core; 8x2GB)
  - 2 HD 146 GB (15000 rpm)
- TESTS (Johannes Elmsheuser – munich):
  - 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.



Johannes Elmsheuser



IFIC Valencia Analysis Facility

prototype at IFIC

AFS / Lustre

RB/W MS

CE

WN

WN

WN

...

Tier-2

...

WN

WN

...

Extra Tier-2

Resources coupled to Tier2)

Desktop/Laptop

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

dispatcher

...

Workers

Tier-3

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

Spanish eScience Initiative

Red Española de e-Ciencia

- See Talk by J. Marco on this conference
- NGI initiative to be part of EGI
- IFIC provides its T2-T3 infrastructure for HEP. Additionally more resources from GRID-CSIC to the NGI.
- IFIC Members in the Infrastructure and Middleware panels
- Study and support of aplications to be ported and supported on this eInfrastructure (i.e- Medical Physics .HadronTherapy)

Conclusions

- eScience Infrastructure for Tier2 and Tier3 was presented
- Distributed Tier-2 among 3 sites in Spain. October '08:
  - 608 KSI2000
  - 240 TB
- Tier-3 at IFIC 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.

  .
- IFIC is active member of the Spanish eScience NGI. Commitee members at middleware and infrastructure panels. Lessons learned in high energy physics to be applied to new applications