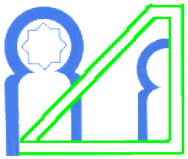


PCI2009 Workshop at Rabat

Tier3 status and plans in Spain

S. González de la Hoz, M. Villaplana, J. Salt, E. Oliver, J. Sánchez. A. Lamas
Instituto de Física Corpuscular (IFIC)
Valencia, Spain
(Centro Mixto Universitat de València-CSIC)

Outline

- ATLAS Tier3@CERN
- Common Tier3 parts in Spain:
  - Batch analysis (resources coupled to Tier2)
  - Interactive analysis
- Tier3 facilities in Spain
  - UAM-Madrid
  - IFAE-Barcelona
  - IFIC-Valencia
- General remarks and conclusions

ATLAS Tier3

- Tier-3s are non-ATLAS funded or controlled centers
- It is up to the different institutions to propose possible Tier-3 configurations
- ATLAS Tier-3 Taskforce: https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasTier3
  - Try to converge the various existing Tier-3 prototypes on a small number of models
- Document the current usage in Atlas Tier-2 and Tier-3 sites
- Determine and make available best practices guidelines
- Develop suggestion for deployment at all Tier-3 sites
- Propose test metrics for the considered design and tabulate the results.

Main Goal: Provide a document + some twiki pages with installation recommendations in a Tier-3

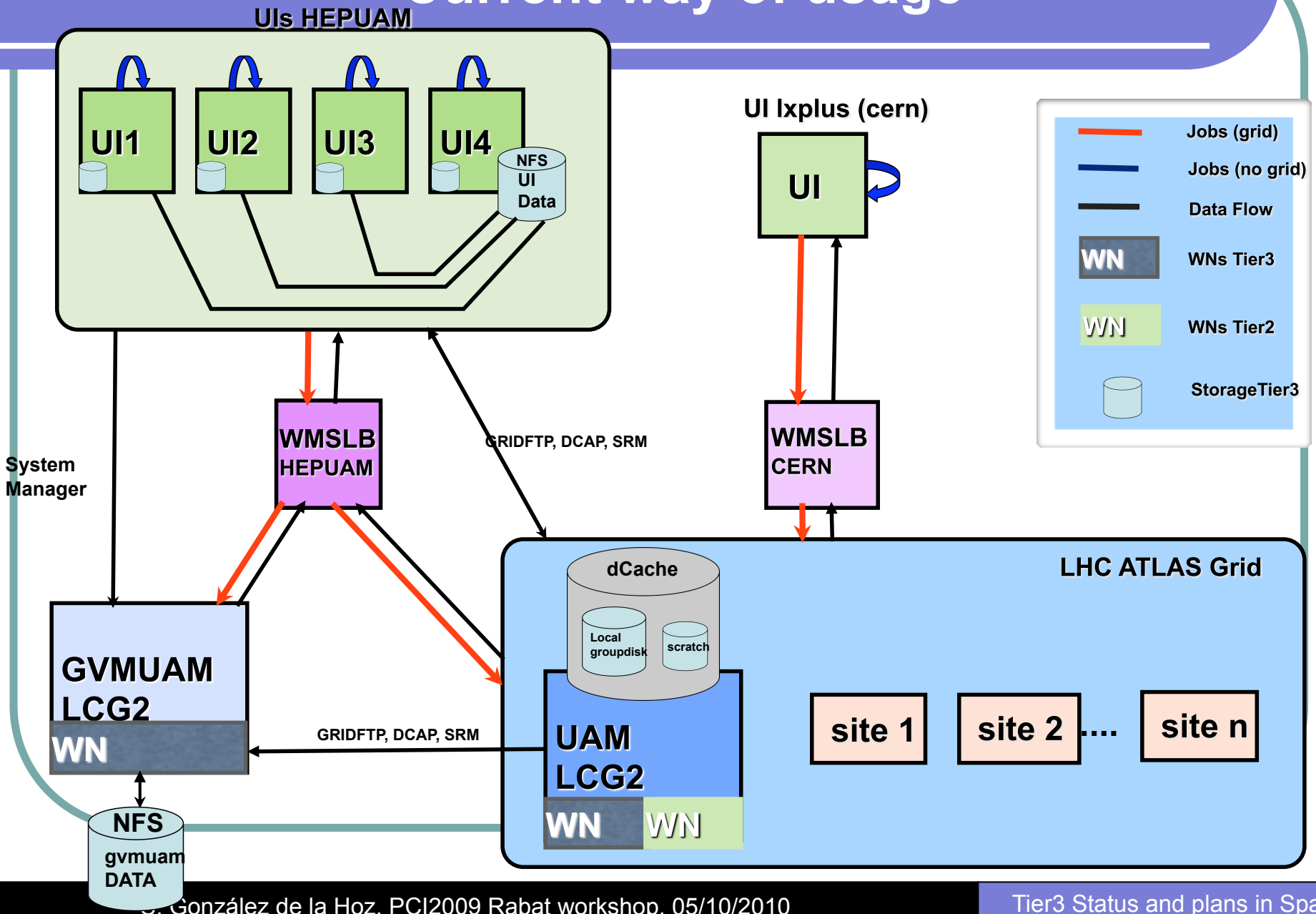
Batch Analysis ("a la Grid") in Spain

- Storage
  - Tier3 has access to the Tier2' SE (AOD, DPD,..) in native mode
  - Grid access to the Tier3 storage
- CPU (dedicated WN's)
  - Using different/shared analysis queues
    - WMS jobs need a local public queue
      - Ganga can manage to filter queues using my proxy certificate, that means neither ATLAS Central Production jobs nor user outside our cloud/site
    - New Analysis queue in Panda working on that

Interactive Analysis (Outside Grid)

- UI's (n User Interfaces)
  - Same software as a WN.
  - Local checks, to develop analysis code before submitting larger jobs to the Tier2's via Grid.
- Proof: Parallel ROOT Facility
  - Exploiting intrinsic parallelism to analyze data in reasonable times.
  - Install PROOF in a PC farm for interactive analysis on DPD.

Tier3 prototype at UAM-MADRID
Current way of usage



Tier3 prototype at UAM-MADRID
A real example: LAr Drift Time

Jobs (grid)
Jobs (no grid)
Data Flow
WN WNs Tier3
StorageTier3

dCache
scratch
Local groupdisk

Input:
Cosmic RAW data (size ≈ 150 GB)
Cond DB

STEP 1

Framework: *Athena*
RAM < 2GB
Input access: dCache (dcap)

- On UIs (28 cores)
- Time lasted≈ 2 weeks
- Comments:
  - ✓ Good dCache perfomance.
  - Uncomfortable for users (input access and submiting jobs to all UIs)

UIs

Step 1 is done only once

NFS
gvmuam DATA

Input:
ROOT ntuple file (size ≈ 4 GB)

STEP 2

Framework: *ROOT batch mode*
RAM < 1GB
Input access: NFS - posix

- On GVUAM-LCG2 ≈ 200 cores
- wms grid job type
- Time lasted: ≈ 10 hours
- Comments:
  - Rather efficient but need faster input file access.

GVMUAM LCG2
WN

Steps 2 and 3 have to be repeated many times for different studies of the analysis.

NFS
UI DATA

Input:
≈700 ROOT files (size ≈ 20 MB each)

STEP 3

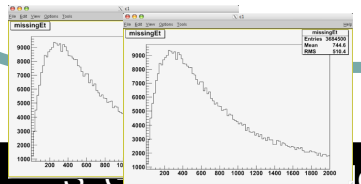
Framework: *Interactive ROOT*
RAM < 1GB
Input access: NFS - posix

- On UI (one core only)
- Time lasted: ≈ 3 hours
- Comments:
  - Too much time.

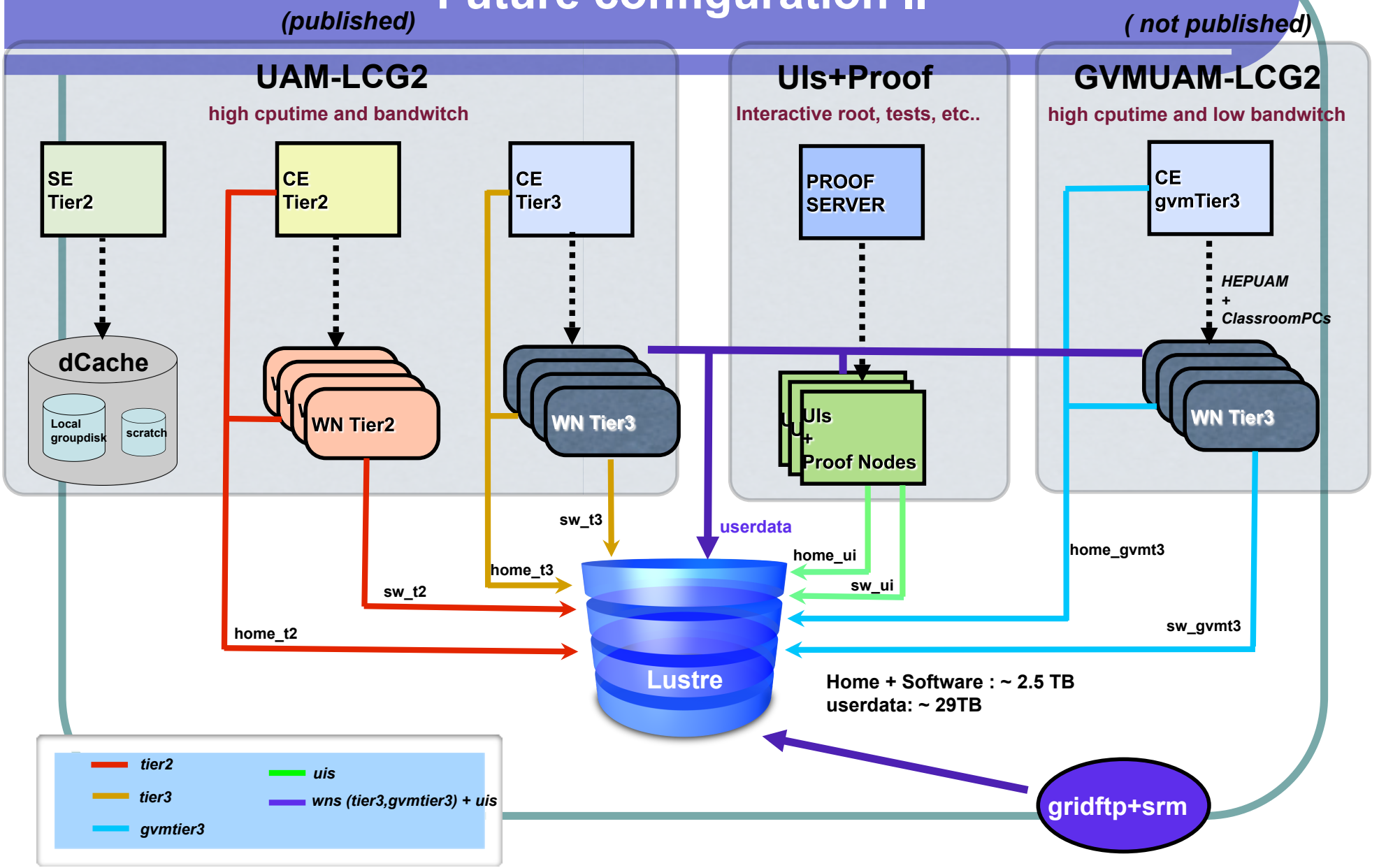
UIs

- PROOF trials - enables interactive analysis with *ROOT - (28 cores)*
- *All UIs as Nodes PROOF*
- Time lasted:≈ 10 min
- Comments:
  - Job was adapted.

UIs



Tier3 prototype at UAM-MADRID
Future configuration II



Tier3 prototype at IFAE-Barcelona

IFAE tier3 (what we have)

- We just started building the IFAE tier3 infrastructure. Until now physicists in the Institute using few User Interfaces for interactive analysis and resources from the tier2 (sending jobs via Ganga)
- The tier3 is hosted at PIC together with the PIC tier1 and the IFAE tier2.
- 4 user interfaces (two slots each):
  - For grid jobs and root/athena interactive analysis.
  - In addition to glite-ui, Atlas software needed for analysis installed: root, ganga/panda clients, dq2...
- 1 CE (not dedicated, but from IFAE tier2). ATLAS production releases are shared with tier2.
- ~30 TB of NFS-mounted disks & 2 Disk severs. The I/O load on the servers has been low in the past and nfs performed well. Studying the possibility of implementing other file systems .

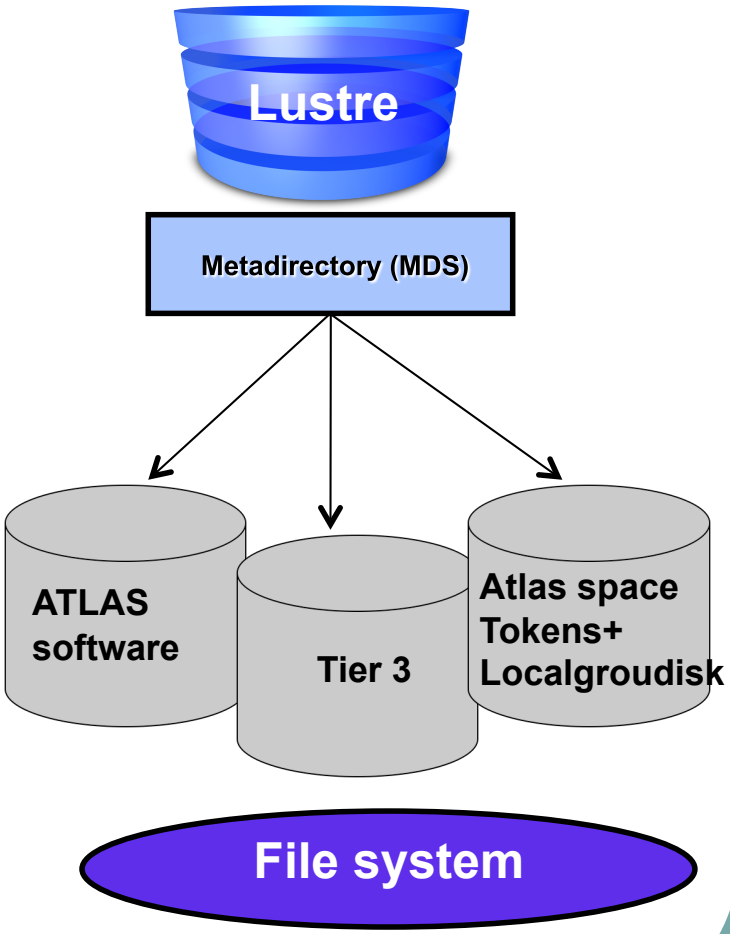
Tier3 prototype at IFAE-Barcelona

IFAE tier3 (plans for 2010)

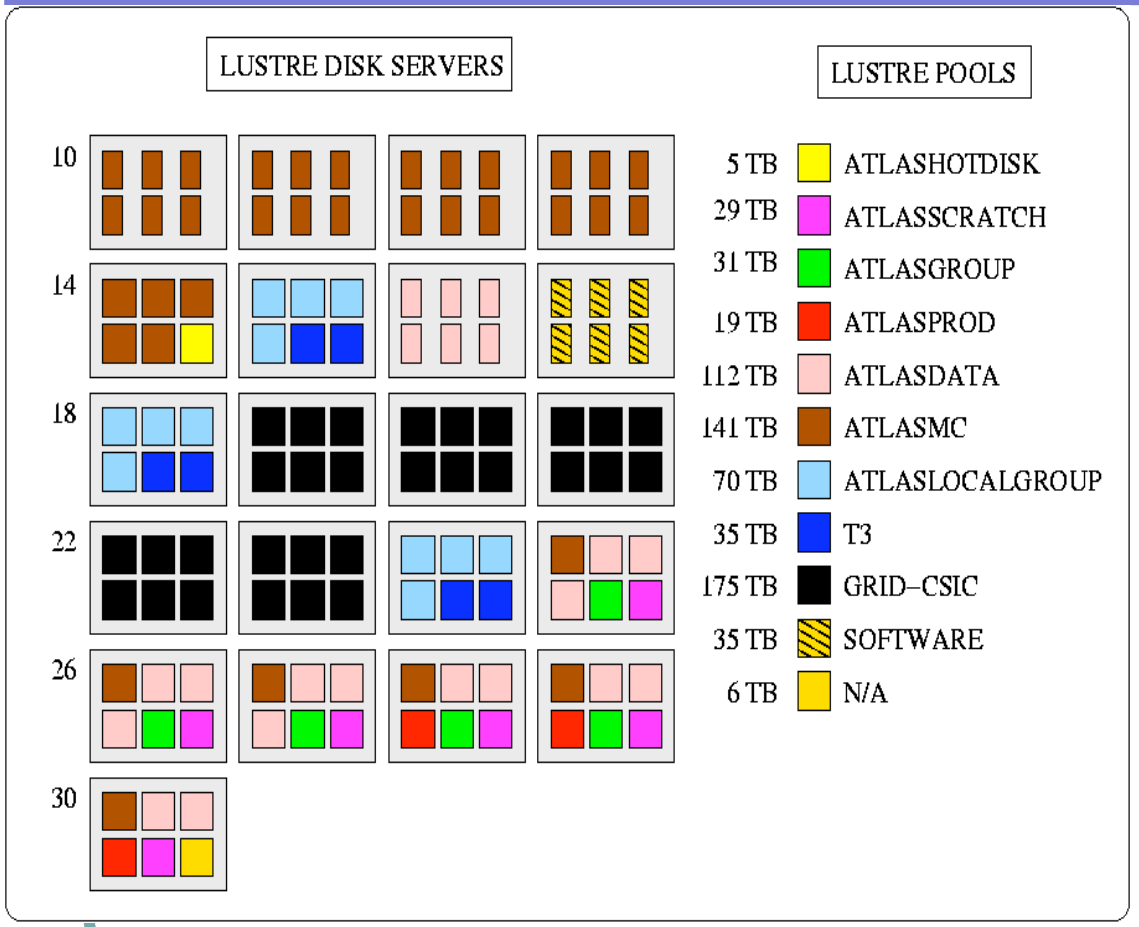
- 1 dedicated person just recruited for the three sites: Carlos Osuna
  - **Increase the storage up to ~80 TB of disk for the data.**
  - **~100 batch cores as WNs**
  - **10-15 user interfaces for interactive analysis**
  - **A dedicated CE**
- Setup PROOF for parallel processing of root/AthenaRootAccess in the farm of User Interfaces.
- This together with an increase in the storage will allow users to reduce the need of data distributed analysis in favour of analysis in local cluster (currently users expend 95% of the total CPU time of the analysis in the tier2's).
- With the increase of disk space we plan to study other storage file systems (rather than nfs, for instance xrootd) already installed in some tier3's.

Tier3 prototype at IFIC-Valencia

- Use a common technology for both Tier2 and Tier3
  - Lustre for data storage (+ Storm SRM)
    - Local access from all the machines (Posix)
  - UI to access Grid and run local test/ program
  - WNs to provide CPU for both Tier2 and Tier3
    - Using share/dedicated CEs/queues
  - Take profit from ATLAS software installation
  - Proof on dedicated nodes accessing lustre
    - Evaluating the possibility to share Tier3's WNs



Tier3 prototype at IFIC-Valencia



- The Tier3 is using three disk servers: gse15,18 and 24
- No overlap with Tier2 disk servers
- The only shared resource is the Lustre metadirectory (MDS).

Tier3 prototype at IFIC-Valencia

- Storage in our Tier3 (Lustre)
  - LOCALGROUPDISK
    - 60% (around 60 TB)
    - Under DDM, not quotas
  - T3
    - 40% (around 40 TB)
    - 1-2 TB per user
    - With quotas
    - Write enabled from UIs (Seen as local disk)

Tier3 prototype at IFIC-Valencia

- Several User Interfaces and two CE no-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).**
- Access to ATLAS software and DDM tools via Lustre as a local file system in our UI

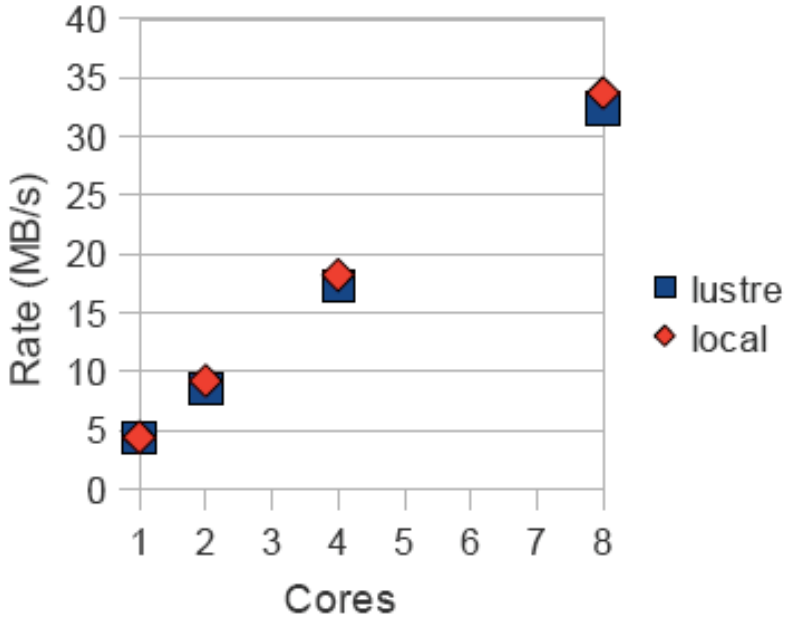
  /lustre/ific.uv.es/sw/atlas/releases
  - **local checks, to develop analysis code before submitting larger jobs to the Tier-1s-2s via Grid**
- The Ganga client is installed locally (AFS)

  source /afs/ific.uv.es/project/atlas/software/ganga/install/etc/setup-atlas.sh

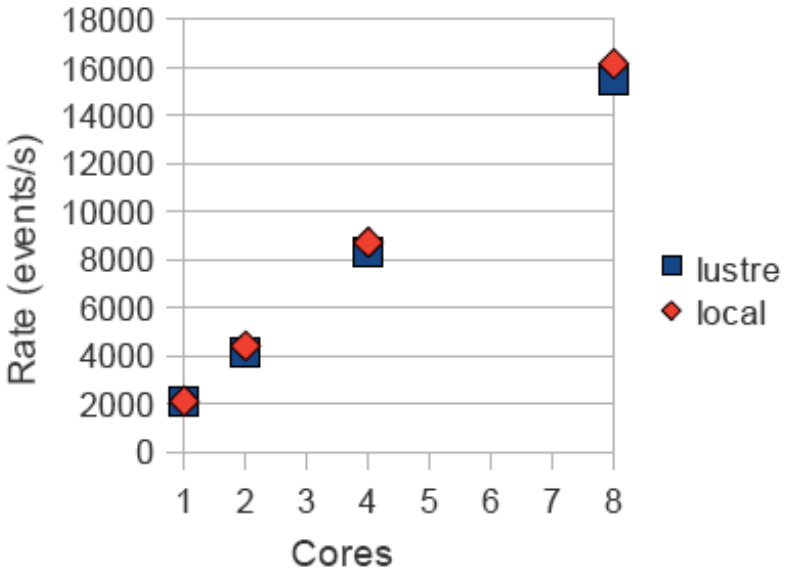
PROOF test at IFIC-Valencia

- Interactive analysis on DPD/ntuples using PROOF
- Test using one UI with 8 cores (PROOF-Lite)
  - Dataset with 3684500 events (7675.24 MB), 372 files, 22MB per file
  - The data was stored locally and on Lustre file system
- Test on a cluster of machines
  - 128 cores (16 nodes)
    - 16 x HP BL460c, 8 cores, 2 x Intel Xeon E5420@2.5 GHz
    - 16 GB RAM
    - 2 HD SAS 146 GB (15000 rpm).
  - Access to the data: Lustre
    - To use the same technology as in our Tier2
  - Xrootd used to start proof servers.

PROOF test at IFIC-Valencia

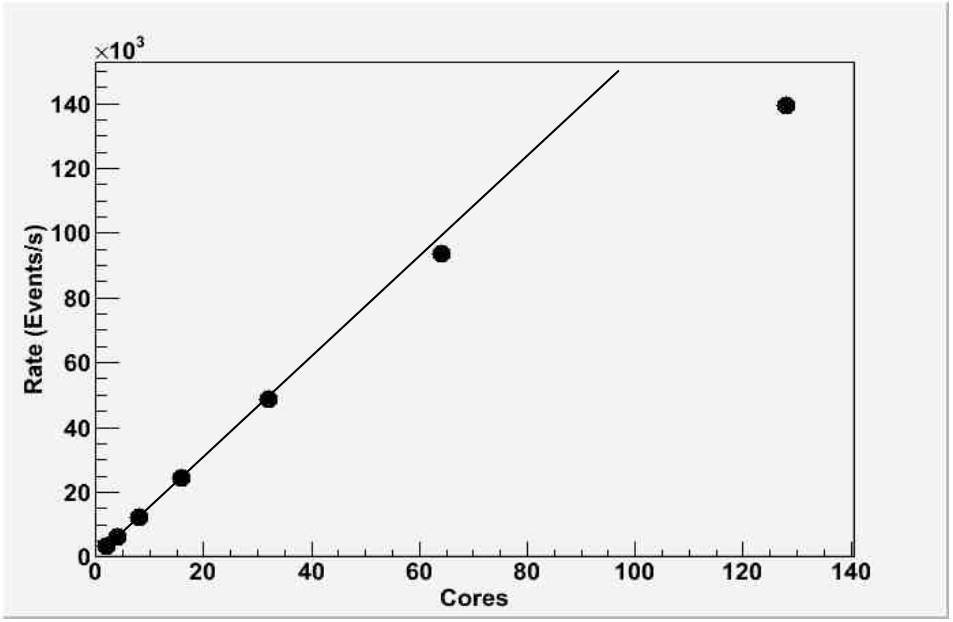


- PROOF-Lite with 8 cores.
- The lustre file system shows a nearly equivalent behaviour as the local storage.



PROOF test at IFIC-Valencia

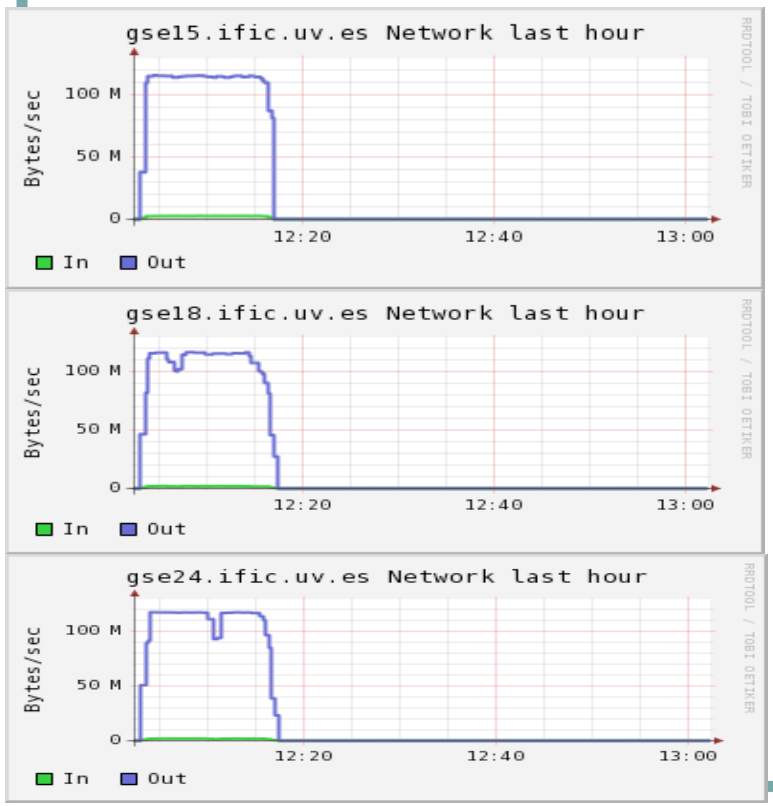
- Test using 128 cores
  - 16 nodes x 8 cores
  - ~ 1440 files
  - ~ 32 GB
  - Data was stored on Lustre file system

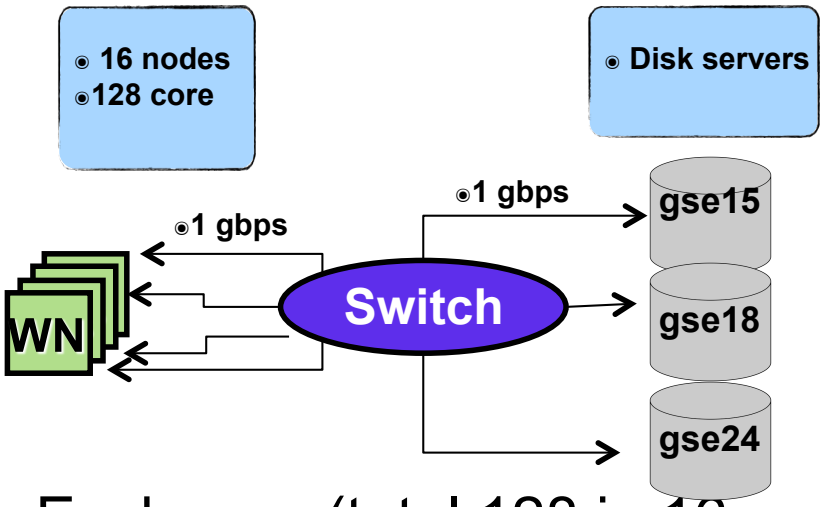


- With 128 cores we are loosing linearity because we are limited by our disk server interface

PROOF test at IFIC-Valencia

- Sequential read test





- Each core (total 128 in 16 nodes) reads 100 random files with dd (bs=32k)
- 10995 files (225 GB)
- Test result BW = 357 MB/s
- Disk servers interfaces were saturated.

PROOF test at IFIC-Valencia

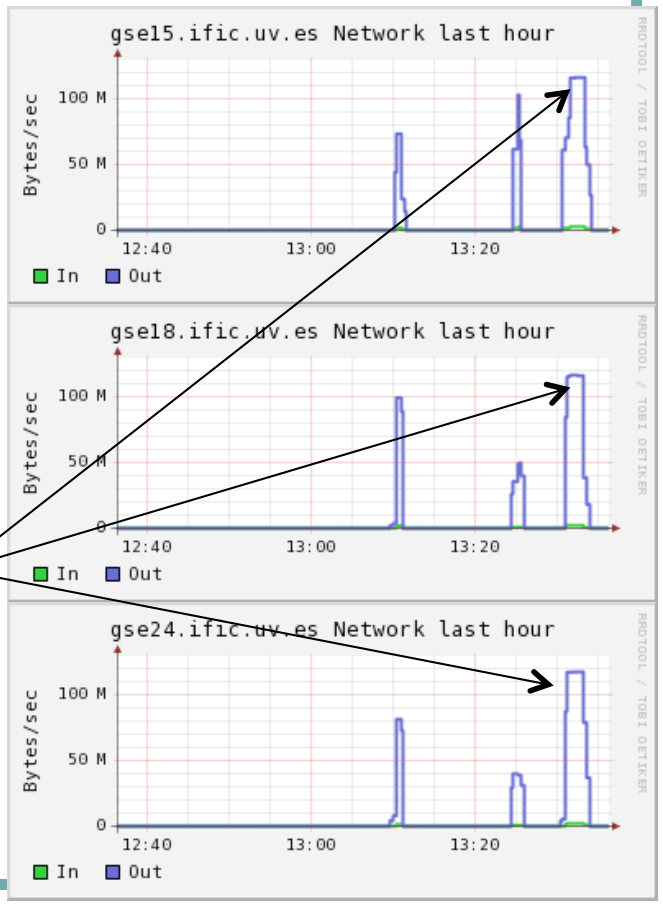
- Test using 4 proof simultaneous sessions with 128 cores each one over 3684500 events (372 files, 7 GB)

N	INI (s)	Time (s)	Rate (evt/s)	BW (MB/s)
128	2.5	36	101634.4	228.3

1 Proof session

N	INI(s)	Time (m:s)	Rate (evt/s)	BW (MB/s)
128	6.0	2:38	23234.3	53.8
128	8.1	2:39	23133.0	53.8
128	8.1	2:36	23530.9	54.4
128	7.3	2:37	23362.0	54.7
Total			93260.2	216.7

4 Proof session



PROOF test at IFIC-Valencia

- Good PROOF behaviour. Scalability is correct with this kind of user analysis. Concurrent use is possible without added degradation.
- Lustre performance is adequate and no sensible degradation was observed while concurrent access is made.
- Lustre performance is limited by disk server ethernet interface. Room is still open to improvement aggregating a second interface (channel bonding). Tests were already done (not presented here).
- The Tier-3 at IFIC-Valencia is no longer a prototype but a real working facility with around 20 users
- The design might change in the future acording to users needs

General remarks and conclusions

- Transfers to the LOCALGROUPDISK space token via DDM/dq2 are not using the user proxy certificate. Quotas are not used and therefore a user could fill in the whole disk space.
- We need a dedicated Panda/Ganga analysis queues for our Tier3 and only "for our users". How these queues are going to be managed?
- How the Tier3 Grid resources are going to be considered by the ATLAS global accounting?
- Tier2 and Tier3 interaction
  - Batch analysis
    - Actually it is to add more resources to our Tier2 (like to have a bigger Tier2)
    - If the Tier2 is well structured, this should not be a problem.
      - To add more WNs and see that our CE is working fine
  - Interactive analysis
    - You are accessing to the data from the UI using the native protocol (Lustre, dcache. Etc.) and this could interfere with the Tier2.
    - But we are talking about 1 UI against all the Tier2's WNs, and therefore the effect should be negligible.
- All Spanish sites have a batch and an interactive analysis facility.
  - Batch: Extra Tier2 resources but in the same infrastructure and using the same technology (Lustre, dcache, etc..)
  - Interactive: UIs + PROOF+(Lustre, nfs,…)

Backup

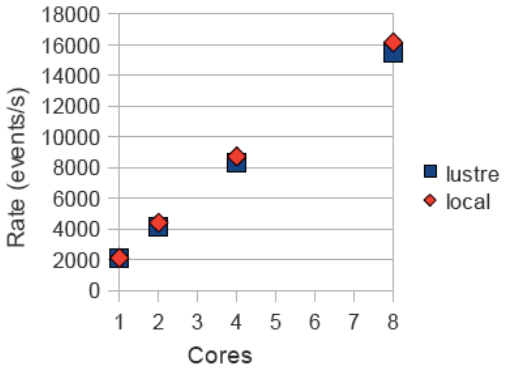
General remarks and conclusions

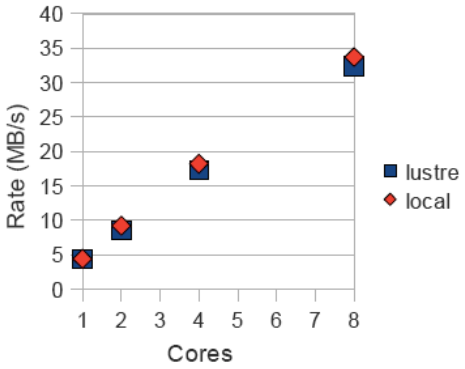
- UAM, IFAE and IFIC Conclusions
  - These Tier3 facilities are starting just now. So far, there has not been a lot of analysis jobs. Therefore the Tier3 facilities has not been used in real cases (a lot of users running a lot of jobs) and they could change.
  - Using Lustre because is a posix distributed file system.
    - Add disc in an easy way
    - Posix file system: User can read Tier2 data from the UIs
  - Lustre giving similar performance to local storage
  - We need a dedicated Panda analysis queue for our Tier3 and only “for our users”

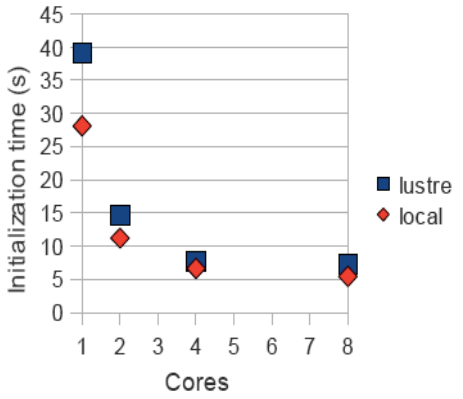
Tier3 prototype at IFIC-Valencia

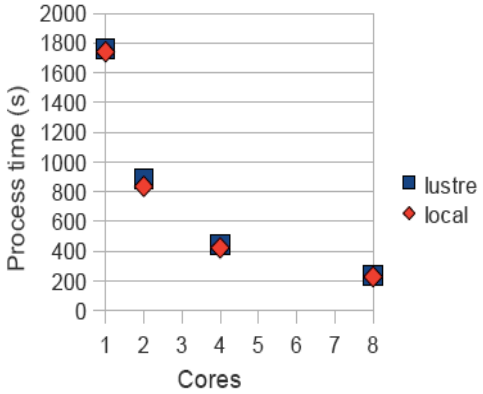
MUESTRA: 3684500 evts (7675.24 MB) 372 files 22MB po

	cores	ini	proc	rate (evt/s)	rate (MB/s)
lustre	1	39,1	1762	2090,2	4,4
	2	14,6	888	4145,8	8,6
	3				
	4	7,7	443	8307,1	17,3
	5				
	6				
	7				
	8	7,2	237	15542,3	32,4
	cores	ini	proc	rate (evt/s)	rate (MB/s)
local	1	28,1	1741	2115,1	4,4
	2	11,2	836	4402,8	9,2
	3				
	4	6,6	422	8718,7	18,2
	5				
	6				
	7				
	8	5,4	228	16158	33,7









Tier3 prototype at IFIC-Valencia

- Same Hammer Cloud test 993. Submitting jobs with Ganga to LOCAL (our UI) and PANDA backend (ANALY_IFIC) using LUSTRE storage

	Gangajob	Dataset number	# Events	CPU	Elapsed time	Rate
Local	16	121050	4	9%	354	0.01
	17	121182	1701	24%	184	9.24
	18	121226	47	9%	354	0.13
	19	121244	1645	19%	462	3.56
	22	121457	504	9%	435	1.16
	25	121679	99	11%	297	0.33
	26	121683	1718	14%	319	5.39
	23	121847	13185	34%	479	27.53
	24	121874	2755	36%	692	3.98
	Gangajob	Dataset number	# Events	CPU	Wallclock	Rate
Grid	0.1	121050	4	15%	317	0.01
	0.0	121182	1701	17%	380	4.48
	1.0	121226	47	16%	318	0.15
	1.2	121244	1645	14%	684	2.4
	2.2	121457	504	13%	381	1.32
	4.0	121679	99	16%	318	0.31
	4.1	121683	1718	20%	319	5.39
	12.0	121847	13185	28%	624	21.13
	12.1	121874	2755	21%	1172	2.35

Tier3 prototype at IFIC-Valencia

- Same HC test 993. With athena running on our UI (interactive analysis, time athena joboption.py)

Datasets (15613 events)
121874 92 files 1.7 GB
121847 11 files 7.8 GB
total 103 files 9.5GB

LOCAL STORAGE

real	user	sys
12m10s	5m59s	13.689s
10m4s	5m53s	12.201s
9m47s	5m57s	12.479s
12m29s	5m54s	13.587s **without cleaning run directory

LUSTRE STORAGE

16m43s	6m26s	19.815s
11m29s	6m24s	19.608s

Tier3 prototype at UAM-MADRID
Future configuration I

Storage

- Shared home directory by UIs
- Shared storage area (posix) by UIs and WNs Tier3 (UAM-LCG2 and GVMUAM-LCG2)
- Lustre File System
  - high performance
  - scalability
  - posix access
- Lustre allows to share storage areas with posix access (home, software, etc...) for Tier2 and Tier3.
- The space token LOCALGROUPDISK will be kept on dCache Tier2 file system.

Computing

- The publication of Tier3 resources is useful to be able to run ATLAS tools (e.g. private MC productions).
- PROOF: Parallel processing for interactive root analysis.
- UIs prepared to be used as PROOF nodes as well.