(Programa de Colaboración Interuniversitaria e Investigación científica)

MINISTERIO DE ASUNTOS EXTERIORES Y DE COOPERACIÓN

aecid

## II PCI 2009 Workshop

# ATLAS Tier3s

Santiago González de la Hoz

IFIC-Valencia



# Tier3 Coordination

- ATLAS Tier3 coordinator: Andrej Filipcic
- Technical issues coordinator: Douglas Benjamin
- Mandate:
  - Coordinate T3 approval and status with ICB
  - Coordinate T3 setup and operation with ADC

# ICB coordination

- Approval of T3, category
- Association to clouds
- T3 signed agreement: approved by ICB (template to be prepared)
- Ex. Rabat in Morocco
  - On behalf of the site and physics groups, the national physicist representative in Atlas collaboration should address an official application to the International Computing Board (with Eric Lancon as the president). The letter should present the site resources dedicated to Atlas (as opposed to other projects), site support for Atlas and describe the usage of the site in terms of Atlas activities (analysis, production,...). The application should be supported by all the Moroccan Atlas physics groups which are going to use the site. The application shall propose the association to an Atlas cloud for technical and support reasons.
  - Before the site is officially included, a memorandum of understanding should be signed between the Moroccan representative and the ICB

# Tier3 Category

- Tier-3s co-located with Tier-2s (not signed agreement)
- Tier-3s with same functionality as a Tier-2 site
- National Analysis Facilities
- Grid-enabled Tier-3s as part of a multi-user Grid site (Tier3gs)
  - Useful setup when a Grid site is already active and ATLAS can use it
- Non-grid Tier-3 (Tier3g):
  - Most common for new sites in the US and likely through ATLAS
  - Very challenging due to limited support personnel

# Tier3 Category

- Depending on the resources and site scope, defined in the agreement.
  - Minimal resources per category to be defined (under discussion).
- T3 definition needed on central services (panda, DDM,..) to distinguish from pledged resources.
- Not (grid) managed: no SE, no CE (limited data transfer, 1TB/week)
- disk only: SE, no CE (local batch only or interactive)
- local grid analysis: SE and CE, brokeroff
- Grid analysis: T2-like setup, analysis panda queues only
- low I/O production: T2-like setup, analysis + production queues for MC simulation
- high I/O production: full T2-like setup

# Tier3 Category

- Rabat example:
  - Doug and Andrej shall discuss within the ADC (Atlas Distributed Computing) the technical terms of the site inclusion in the central production system, with the purpose of testing the site for data transfers, software installation and test jobs. The result of the testing shall tell what is the most appropriate way for the site to be included as an Atlas Tier-3, and this will be recommended by ADC.
- ADC policies:
  - T3s must not affect the T0/1/2 operations
  - write-only storage endpoints
  - transfer throttling (limited)
  - black-listing if central services affected
  - obsolete data can be removed by ADC
  - job/data distribution according to the T3 category

# Grid-enabled Tier-3s (Tier3gs)

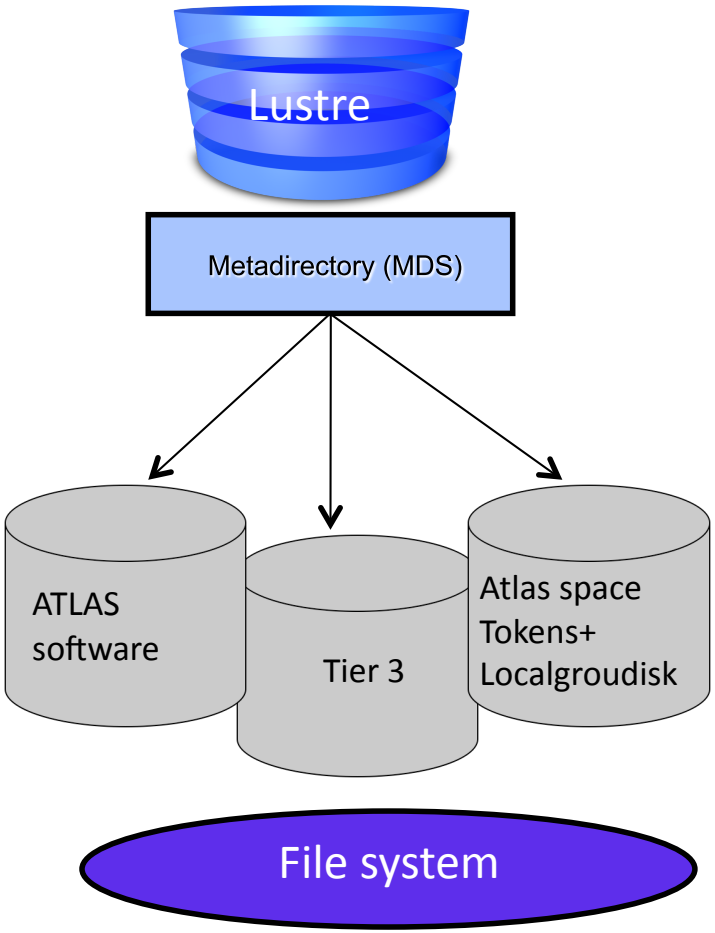
- Grid-enabled Tier-3s for ATLAS are usually part of a larger Grid site that serves several Communities
- They receive automatic software installations and are fully enabled to run Athena jobs, interactively, in local batch mode and as Grid jobs
- They can be used to develop software and to test tasks on a small scale before submitting them to the Grid
- Data throughput is of the highest importance at Tier-3s, as they run mostly analysis jobs
- Several Grid-enabled Tier-3s have adopted GPFS+StoRM or Lustre+StoRM as storage solution
  - In this way it is possible to share the file system between Grid and direct access as all servers are mounted on all batch and interactive nodes
  - Files can be imported to a given SRM storage area and analysed directly with an interactive job
  - There is no need of separate storage servers for grid and non-Grid users
    - Easier system management
    - Easier life for users
  - Direct access to the Grid file system is much more intuitive for non-expert users
  - "HammerCloud" tests show excellent performance

# Minimal Tier3gs requirement

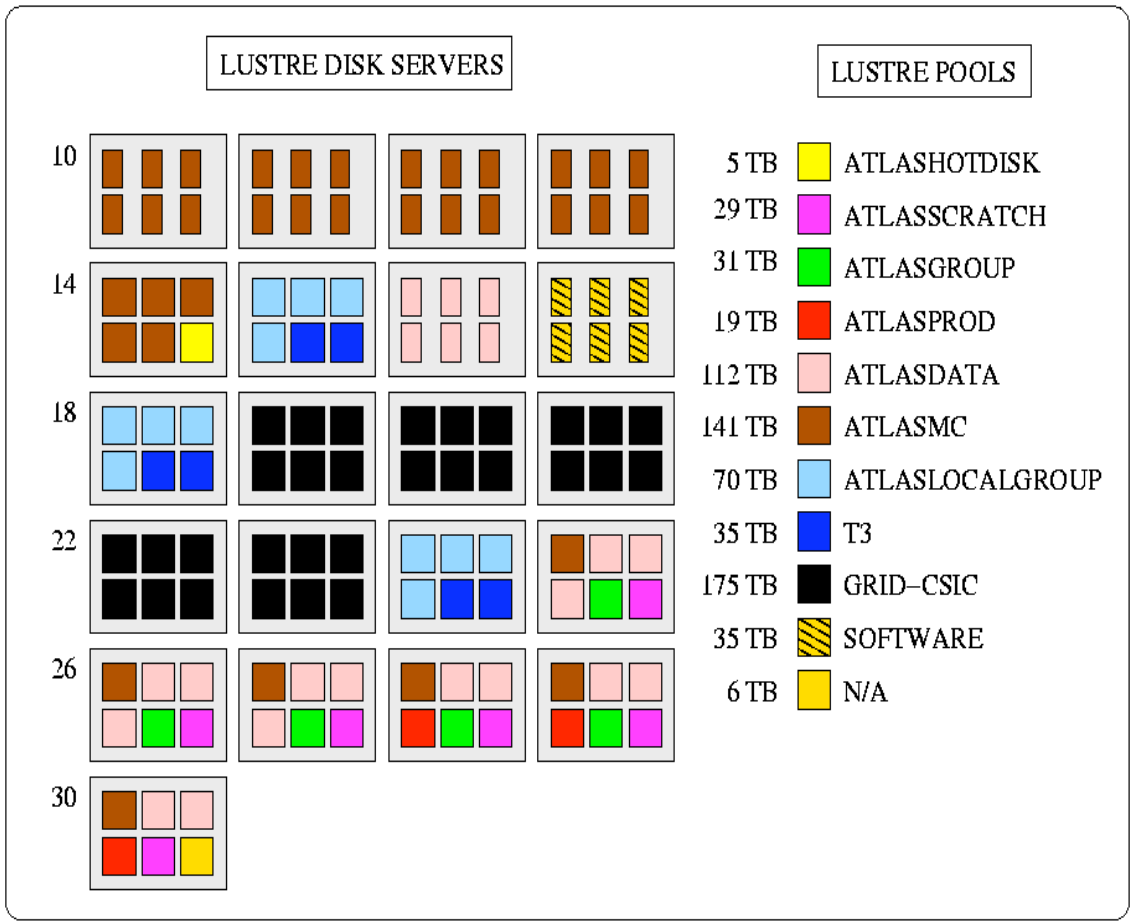
- The minimal requirement is on local installations, which should be configured with a Tier-3 functionality:
  - A Computing Element known to the Grid, in order to benefit from the automatic distribution of ATLAS software releases
    - Needs >250 GB of NFS disk space mounted on all WNs for ATLAS software
    - Minimum number of cores to be worth the effort is under discussion (~40?)
  - A SRM-based Storage Element, in order to be able to transfer data automatically from the Grid to the local storage, and vice versa
    - Minimum storage dedicated to ATLAS depends on local user community (20-40 TB?)
    - Space tokens need to be installed:
      - LOCALGROUPDISK (>2-3 TB), SCRATCHDISK (>2-3 TB), HOTDISK (2 TB)
    - Additional non-Grid storage needs to be provided for local tasks (ROOT/PROOF)
- 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.

# Tier3gs example: 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



# Tier3gs example: 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).

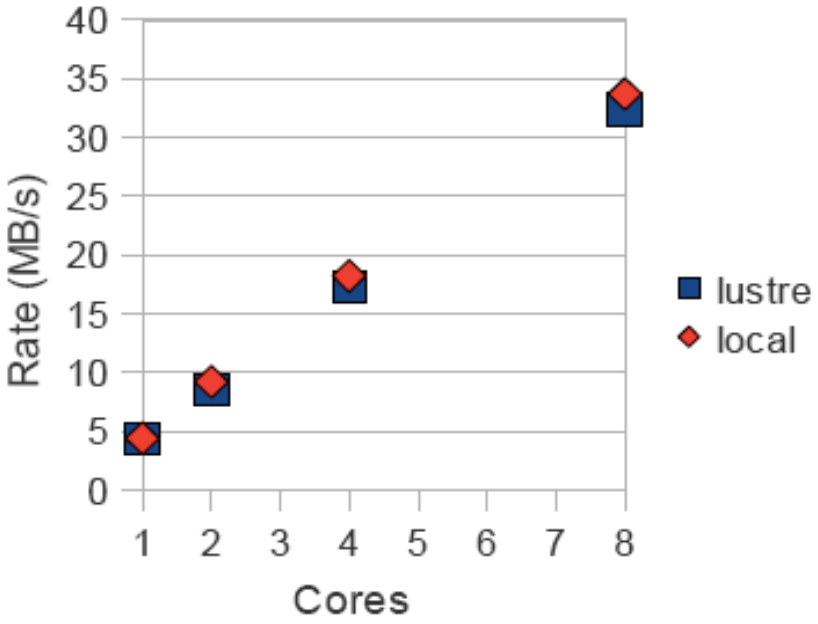
# Tier3gs example: IFIC-Valencia

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

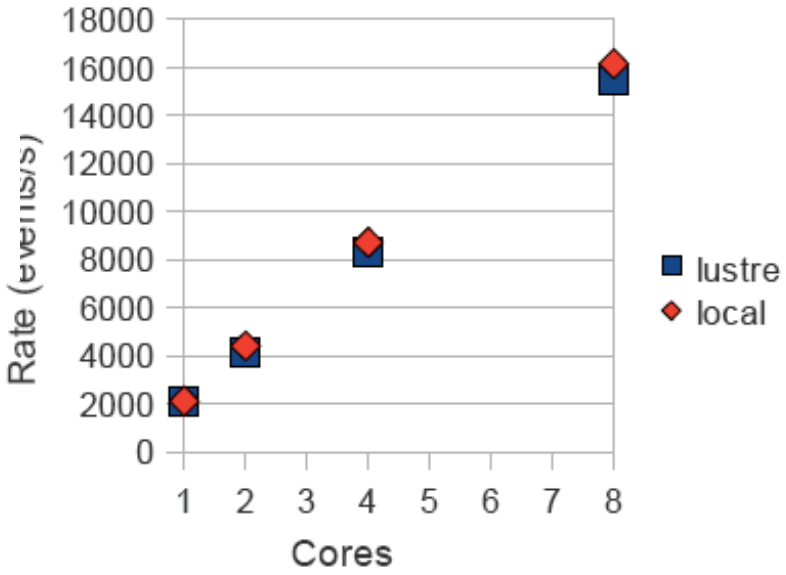
# Tier3gs example: IFIC-Valencia

- 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.

# Tier3gs example: IFIC-Valencia

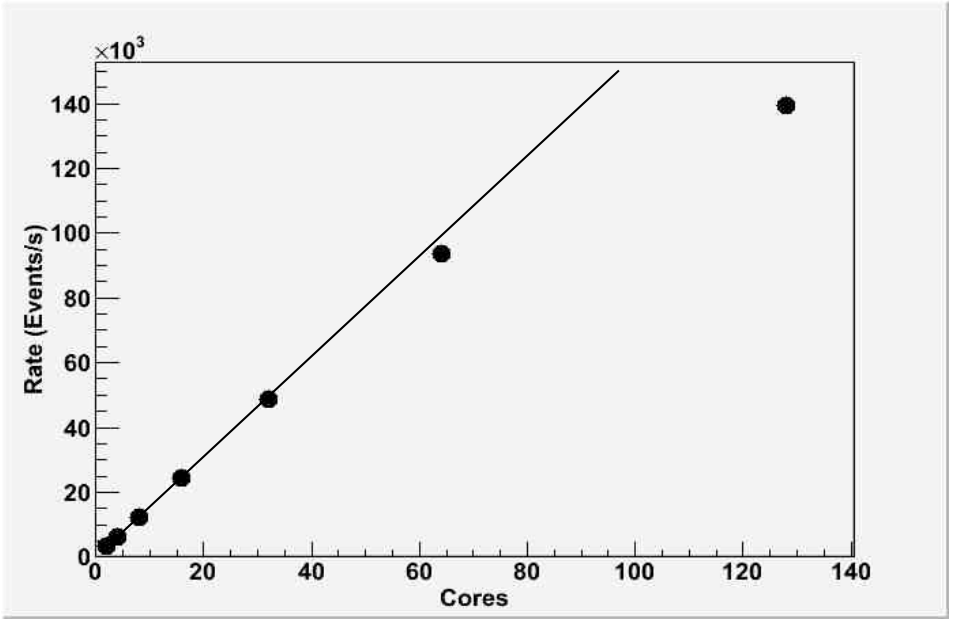



- PROOF-Lite with 8 cores (in our UI).
- The lustre file system shows a nearly equivalent behaviour as the local storage.



# Tier3gs example: 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

# Tier3gs example: IFIC-Valencia

- Proof tests (Today)
  - Setup:
    - 1master (2xquad core + 300 GB HD SAS)
      - Redirector de xrootd
    - 5 workers (2xquad core + 300 GB HD SAS)
      - Access to Lustre in native
      - Xrootd file serving in each worker

Tests:

- Direct access to Lustre
- Copy files to the Workers local discs

# Tier3gs example: IFIC-Valencia

## Results:

- Same performance between files stored in lustre as in local (xrootd).
  - But now the data are shared in different WNs and the job is running in a different WN where the data is.
  - Optimize Proof and xrootd to run the job in the worker where the data.

## In the near future:

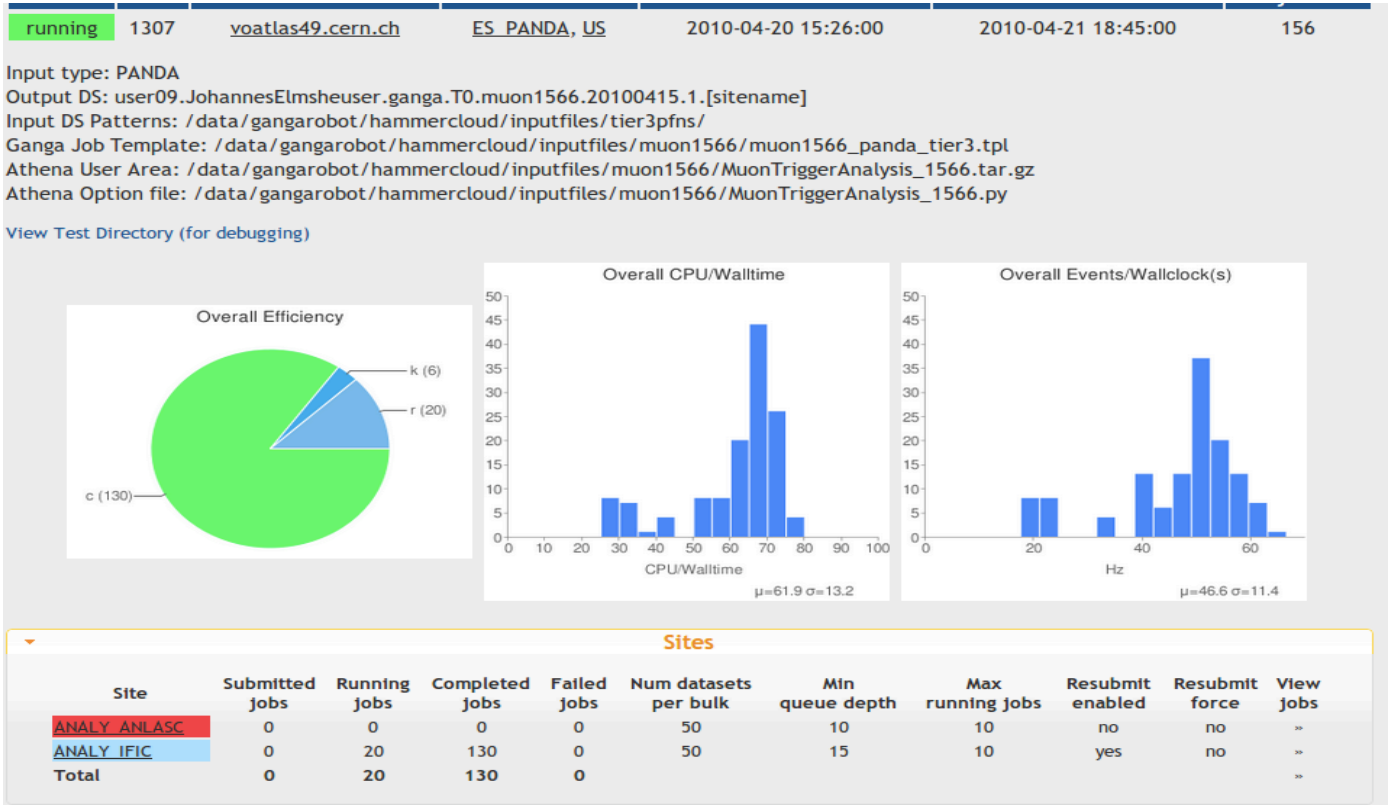
- To use Lustre as a MSS system and to cache the files to the workers in a transparent way using garbage collection automatically

# Tier3gs example: IFIC-Valencia

- The performance metrics for analysis jobs is hard to quantify as the results depend strongly on the type of data and the kind of analysis people run on a given site
- ATLAS HammerCloud tests can nevertheless be used to compare the performance of different sites and hardware solutions
  - HammerCloud runs large numbers of pre-defined Athena-based analysis jobs on data that are placed at a given site and produces performance plots that can be used to better **tune the local set-up**

- Exercise HammerCloud +pfnListat a few sites to work out the bugs

–But maintaining this list will be difficult



# Tier3g

- Design a system to be flexible and simple to setup (1 person < 1 week)
- Simple to operate <= 0.5 FTE to maintain
- Relative inexpensive and fast (1 TB of data over night)
  - Devote most resources to Disks and CPU's
- Using common tools will make it easier for all of us
- Interactive nodes (ROOT & PROOF)
- Can submit grid jobs (UI).
- Batch system with worker nodes
- ATLAS code available
- DDM client tools used for fetch data (dq2-ls, dq2-get)
  - Including dq2-get + fts for better control
- Storage can be one of two types (sites can have both)
  - Located on the worker nodes
    - Lustre/GPFS (mostly in Europe)
    - XROOTD
  - Located in dedicated file servers (NFS/XROOTD/Lustre/GPFS)

# Tier3g

- By their operational requirements non-grid Tier 3 sites will require transformative technologies ideas and solutions
- ATLAS constantly producing new software releases
  - Maintaining an up to date code stack much work
  - Tier-1 and Tier-2 sites use grid jobs for code installation
- CVMFS (CERNVM web file system)
  - Minimize effort for ATLAS software releases
  - Conditions DB
- We recently officially request long term support for CVMFS for Tier-3s
  - We are starting testing cvmfs for Tier-1s and Tier-2s also

# Tier3g

- Xrootd/Lustre
  - Xrootd allows for straight forward storage aggregation
  - Some other sites using Lustre or GPFS
  - Wide area data clustering will help groups during analysis (couples xrootd cluster of desktops at CERN with home institution xrootd cluster)
- dq2-get with FTS data transfer
  - Robust client tool to fetch data for Tier-3 (no SRM required – not in ToA – a simplification)
- Medium/Longer term examples
- Proof
  - Efficient data analysis
  - Tools can be used for data management at Tier-3
- Virtualization / cloud computing

# ATLAS XROOTD demonstrator project

- Last June at WLGC Storage workshop
  - ATLAS Tier-3 proposed alternative method for delivering data to Tier-3 using confederated XROOTD clusters
- Physicists can get the data that they actually use
- Alternative and simpler than ATLAS DDM
  - In testing now
  - Plan to connect Tier-3 sites and some Tier-2 sites
- CMS working on something similar (Their focus is between Tier-1/Tier-2 – complimentary – we are collaborating )

# ATLAS XROOTD demonstrator project

- Doug Benjamin would like to include UK (GPFS-DPM), Valencia (Lustre) and KIT-GridKa (xrootd) in the project.
  - There has been much activity in the project in the US of late and I would like to include European sites also.
- Report on the progress at the upcoming Atlas Software week (nov29-dec3) meeting in the Tier 3 session.
  - On the other hand, a survey to the Tier3 sites would be sent to collect information (what they are using)
  - This information will be used to help formulate some of the Tier3 related activities within ADC (for example Tier 3 monitoring and Tier 3 support)
- In addition Graeme and Doug still have to provide a schedule for WLCG and likely some sort of progress report.

# Conclusions

- Tier-3 computing is important for data analysis in ATLAS
- A coherent ATLAS wide effort has begun in earnest
- Tier-3s must be designed according to the needs of the local research groups
- Striving for a design that requires minimal effort to setup and successfully run.
- Technologies for the Tier-3s are being chosen and evaluated based on performance and stability for data analysis.
- Ideas from Tier-3 are moving up the computing chain