Summary of the ATLAS T3 Activities

Wolfgang Ehrenfeld, DESY

ATLAS T3 Working Groups

> **ATLAS T3 effort chaired by**

- Massimo Lamanna
- Doug Benjamin
- Rik Yoshida

> **6 Working groups:**

- Distributed Storage (xrootd, Lustre/GPFS)
- DDM
- Software and Conditions DB
- PROOF
- Support
- Virtualisation

> **What is a Tier3 for ATLAS?**
> https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasTier3

> **ATLAS T3 Jamboree: 25./ 26. January 2010**
> http://indico.cern.ch/conferenceDisplay.py?ovw=True&confId=77057

> **1st ATLAS T3 Meeting: 21. April 2010**
> http://indico.cern.ch/conferenceDisplay.py?confId=89039

> **2nd ATLAS T3 Meeting: 15. July 2010**
> http://indico.cern.ch/conferenceDisplay.py?confId=76895

Tier 3 Types

- Tier 3's are non pledged resources
  - Does not imply that they should be chaotic or troublesome resources though
- Atlas examples include:
  - Tier 3's collocated with Tier 2's
  - Tier 3's with same functionality as a Tier 2 site
  - National Analysis facilities
  - Non-grid Tier 3 (most common for new sites in the US and likely through Atlas)
    - Very challenging due to limited support personnel
- Tier 3 effort now part of the ADC
  - Massimo Lamanna (leads effort)

Non-grid site Tier 3 design/Philosophy

- Design a system to be flexible and simple to setup (1 person < 1 week)
- Simple to operate   - < 0.25 FTE to maintain
- Scalable with Data volumes
- Fast - Process 1 TB of data over night
- Relatively inexpensive
  - Run only the needed services/process
  - Devote most resources to CPU's and Disk
- Using common tools will make it easier for all of us
  - Easier to develop a self supporting community.

New US Tier 3 sites

- Most US Atlas institutions received funds from the funding agencies to support Tier 3 computing at their home institutions.
- The funds are set to flow shortly.
- Expect to see ~30 new (or greatly enhanced) Tier 3 sites in the US over next 6-12 weeks.
- All most all of these sites will be a non-grid Tier 3

ATLAS T3 Working Group

- **Storage**

- **Xrootd:**
  https://twiki.cern.ch/twiki/bin/view/Atlas/XrootdTier3

- **Lustre/GPFS:**
  https://twiki.cern.ch/twiki/bin/viewauth/Atlas/LustreTier3

Xrootd WG activites Summary

- ✓ Developed installation methods & instructions:
  - ✓ non-priv install, config & run
  - ✓ single configuration for a cluster
  - ✓ ACL method for a default T3
  - ✓ multi-homed redirector/data server
  - ✓ Integration with AtlasLocalRootBase
  - ✓ gridFTP door on top xrootd system

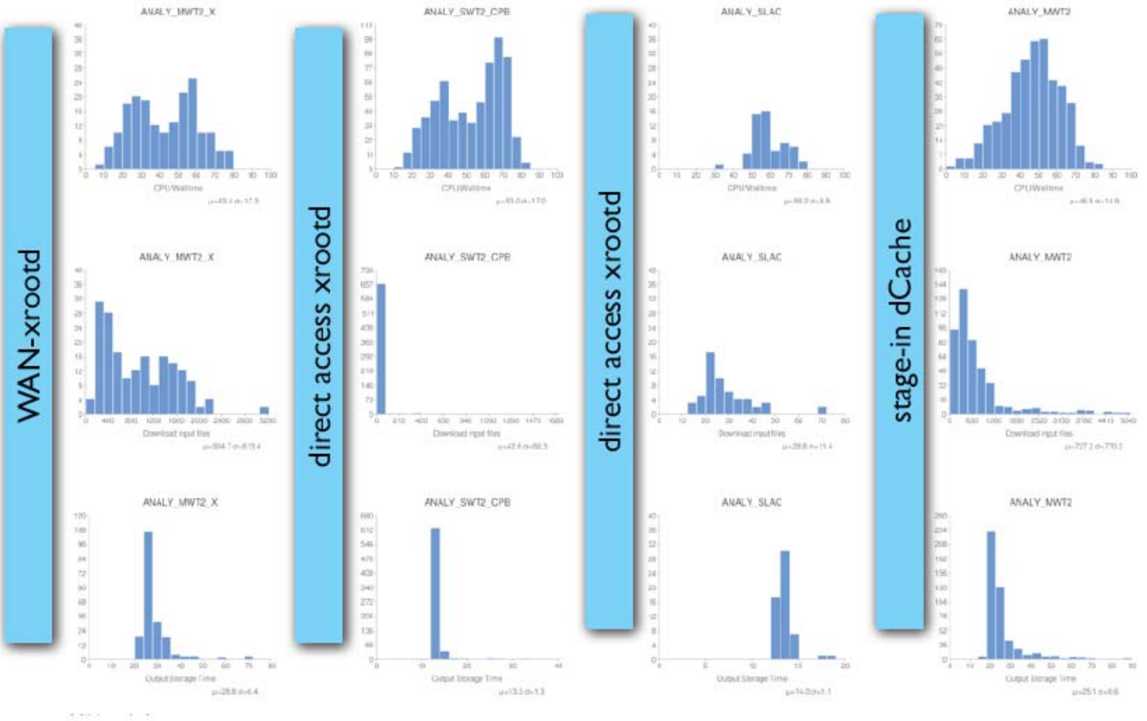
Xrootd WG activites Summary

- ✓ Testing configurations: load balancing on writing & Posix client access
- ✓ Performance studies: xrdcp and Hammercloud stress tests - comparison of access methods (direct vs stage-in)
- ✓ Next steps identified
- ✓ Final report: https://twiki.cern.ch/twiki/bin/view/Atlas/XrootdTier3Report

A Tier3 Storage Component

- Suggests a low cost high performing Xrootd data server appropriate for T3
- Optimal for 64 clients (eg. if max rate by Athena or Root is 10MB/s). But holds up okay if twice #clients, rate only reduced ~1/2
- Also suggests disk-heavy workers would handle the load okay - probably would want to do bonded gigabit or put in a 10G if affordable

HC test comparisons


WAN-xrootd

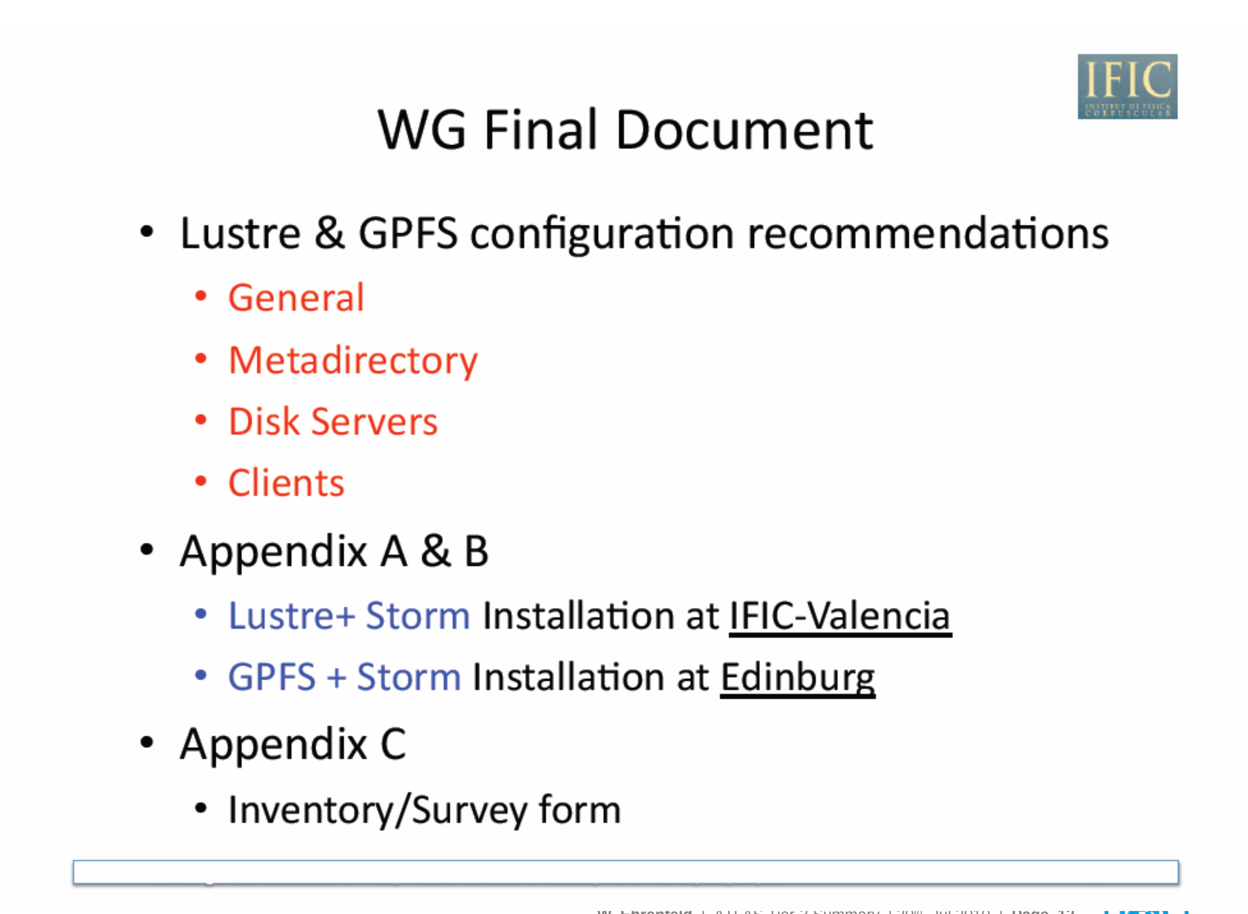
direct access xrootd

direct access xrootd

stage-in dCache

WG Final Document

- http://ific.uv.es/~sgonzale/WP-Lustre-GPFS-conclusions.pdf
- Introduction
  - Lustre & GPFS components
    - Minimal setup
    - Base version
    - Existing installation and contacts



WG Final Document

- Lustre & GPFS configuration recommendations
  - General
  - Metadirectory
  - Disk Servers
  - Clients
- Appendix A & B
  - Lustre+ Storm Installation at IFIC-Valencia
  - GPFS + Storm Installation at Edinburg
- Appendix C
  - Inventory/Survey form

ATLAS T3 Working Group

- DDM

Types of ATLAS T3s

- The SERVICE-based T3: T3gs
  - Looks like a T2 but with unpledged resources
    - Runs a Storage and a Computing element
    - Is known by the WLCG information system
    - Is known by the ATLAS DDM system and served by DDM Site Services
- The CLIENT-based T3: T3g
  - Needs a minimal installation of the Grid middleware
  - Needs a minimal installation of DDM clients

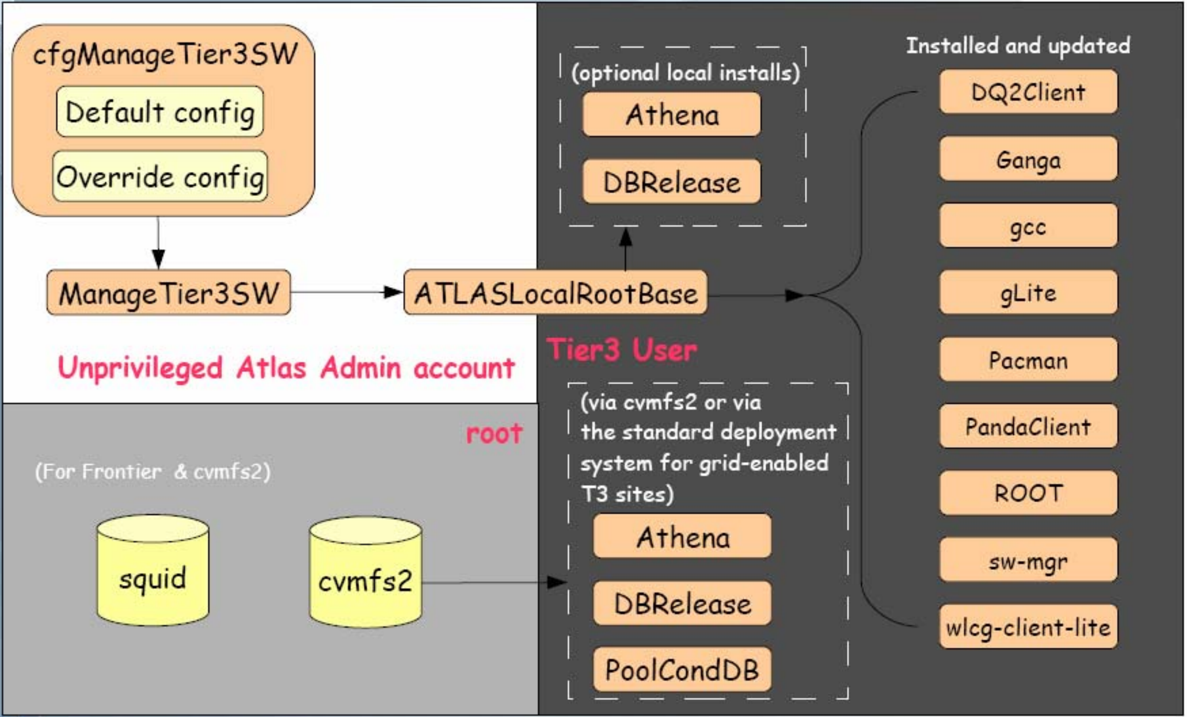
How to use dq2-get + FTS

- Bad News: you need to setup a storage element
  - A simple gridFTP server on top of a filesystem would do
  - If you plan to have a storage larger than i.e. 10TB, you should be considering a real pool manager + SRM
    - For example dCache or BestMan
- You need to inform your Cloud contact about your storage
  - He needs to setup a FTS channel and grant permissions
  - So, yes, you have to belong to a cloud
- You need to be careful about which source you use
- Prototype. Get in touch with Hiro for more infos
  - we will provide a quick twiki once we have some experience)

ATLAS T3 Working Group

- **Software and Conditions DB:** https://twiki.cern.ch/twiki/bin/view/Atlas/Tier3SoftwareWorkingGroup

Tier3 Software: Interactions



Tier3 User ...

1 Define in ~/.bashrc

```
export   ATLAS_LOCAL_ROOT_BASE=<path>/ATLASLocalRootBase
alias  setupATLAS='source ${ATLAS_LOCAL_ROOT_BASE}/user/atlasLocalSetup.sh'
```

2 Setup

```
setupATLAS
```

```
> setupATLAS
...Type localSetupDQ2Client to use DQ2 Client
...Type localSetupGanga to use Ganga
...Type localSetupGcc to use alternate gcc
...Type localSetupGLite to use GLite
...Type localSetupPacman to use Pacman
...Type localSetupPandaClient to use Panda Client
...Type localSetupROOT to setup (standalone) ROOT
...Type localSetupWlcgClientLite to use wlcg-client-lite
...Type saveSnapshot [--help] to save your settings
...Type showVersions to show versions of installed software
...Type createRequirements [--help] to create requirements/setup files
...Type setupDBRelease to use an alternate DBRelease
...Type diagnostics for diagnostic tools

> diagnostics
...Type runKV [--help] to test the kit or your desktop
...Type checkOS to check the system OS of your desktop
...Type supportInfo to dump information to send to user support
...Type db-fnget to run fnget test for Frontier-squid access
...Type db-readReal to run readReal test for Frontier-squid access
...Type rootsysTest to test for a ROOTSYS length issue
```

Consistent look
Menu driven
Encapsulated setups

SetupATLAS … some examples

1 Conditions Pool File (cvmfs2 or local) , Frontier automatically setup

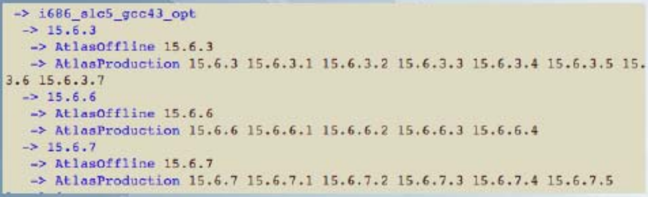
```
> env | grep -e POOL -e FRONTIER
ATLAS_POOLCOND_PATH=/opt/atlas/conditions/poolcond/catalogue
FRONTIER_SERVER=(serverurl=http://frontier.triumf.ca:3128/ATLAS_frontier)(proxyurl=...
```

2 createRequirements: works for cvmfs or local Athena kits; eg.

```
source ~/cmthome/setup.sh -tag=15.6.7,use_cvmfs
```

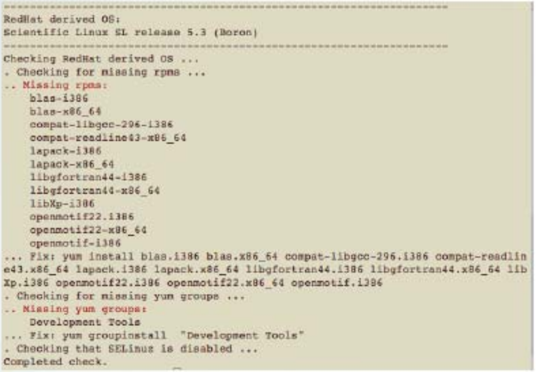
3 showVersions (Athena also shows cvmfs); eg.

```
showVersions -show=athena --cvmfs
```



```
 -> i686_slc5_gcc43_opt
  -> 15.6.3
   -> AtlasOffline 15.6.3
   -> AtlasProduction 15.6.3 15.6.3.1 15.6.3.2 15.6.3.3 15.6.3.4 15.6.3.5 15.
3.6 15.6.3.7
  -> 15.6.6
   -> AtlasOffline 15.6.6
   -> AtlasProduction 15.6.6 15.6.6.1 15.6.6.2 15.6.6.3 15.6.6.4
  -> 15.6.7
   -> AtlasOffline 15.6.7
   -> AtlasProduction 15.6.7 15.6.7.1 15.6.7.2 15.6.7.3 15.6.7.4 15.6.7.5
```

4 CheckOS for missing rpms:



```
--------------------------------------------------------------------
RedHat derived OS:
Scientific Linux SL release 5.3 (Boron)
--------------------------------------------------------------------
Checking RedHat derived OS ...
. Checking for missing rpms ...
.. Missing rpms:
    blas-i386
    blas-x86_64
    compat-libgcc-296-i386
    compat-readline43-x86_64
    lapack-i386
    lapack-x86_64
    libgfortran44-i386
    libgfortran44-x86_64
    libXp-i386
    openmotif22.i386
    openmotif22-x86_64
    openmotif-i386
... Fix: yum install blas.i386 blas.x86_64 compat-libgcc-296.i386 compat-readlin
e43.x86_64 lapack.i386 lapack.x86_64 libgfortran44.i386 libgfortran44.x86_64 lib
Xp.i386 openmotif22.i386 openmotif22.x86_64 openmotif.i386
. Checking for missing yum groups ...
.. Missing yum groups:
    Development Tools
... Fix: yum groupinstall  "Development Tools"
. Checking that SELinux is disabled ...
Completed check.
```

Frontier setup

- Instructions at https://twiki.cern.ch/twiki/bin/view/Atlas/Tier3SoftwareWorkingGroup
- Suggested best practices for Tier3 sites
  - Install a local squid or use a close T2 facility for Frontier access if
    - The site has > 100 cores and no CVMFS setup
    - The site has more than 2 cores and uses CVMFS to access the ATLAS software
  - The squid used for Frontier can also be used to cache the CVMFS files (normal proxies)

Access to CD files

- Possible options
  - Rsync from CERN AFS
    - Only ~10% of the data volume will be actively used
    - Needs to rebuild the PFC to use the local paths
  - Distribution via CVMFS
    - Some tuning still needed to allow fast updates in the central server and availability on the client side
    - Will probably need to use a different catalog/server for the CD files, separated from the one used for the releases
    - For the moment this is the most performant solution
  - Http file access
    - Limited scalability
    - https://twiki.cern.ch/twiki/bin/view/Atlas/AthenaDBAccess#HTTP_CD_file_access

ATLAS T3 Working Group

- **PROOF:**
  https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasProofWG

MAIN ISSUES

- For system administrators
  - Instruction for installing PROOF on top of existing file systems.
  - A good solution for local file registration with PQ2 tools.
  - Some real benchmarks with D3PD files for different sites to understand the performance of their cluster.
- For normal PROOF users
  - Instruction of using PROOF from scratch.
  - How to analysis with Sframe.
  - How to run analysis from using PROOF-lite.
  - Analysis examples for D3PD ntuples.

STATUS OF PROOF

- Version 5.26b_PROOF is still under testing.
- What's new in 5.26b_PROOF?
  - New pq2-ls and pq2-ls-files, much faster and less load on PROOF master.
  - pq2-redistribute
    - redistribute data on a given (sub-)set of servers making the number of files or the total file size even
    - file movement currently based on xrootd tools
  - Improvement of using datasets in PROOF code.
  - Some instabilities have been solved.
- All the installation and benchmark instructions will be based on it.

- Example analysis code for Egamma D3PD files. (They are not using CollectionTree so there is some modification to normal PROOF analysis code.)
- We will collect more example analysis code for different D3PDs.

STATUS OF SFRAME

- SFramealready provides many convenience functionality for running analyses on (D3PD) ntuple files.
- The main page is: http://sourceforge.net/projects/sframe/
- A package under https://svnweb.cern.ch/trac/atlasgrp/browser/Misc/sframe/D3PDVariables/trunk
- This package will help reading D3PD files. The idea is that for each python D3PDObject that is available in the D3PDMaker code, we would put one such class in this package. So if someone saved the muon information with detail level 2 in his/her D3PD, (s)he could read all these variables with something like this code:

  - In the cycle header: D3PD::MuonD3PDObject m_muons;

  - In BeginInputData: m_muons.ConnectVariables("d3pdTreeName", 2);
- SFrame with PQ2 is still under developing. It still work with file list now.
- SFrame with PROOF-batch is still under developing.

ATLAS T3 Working Group

- **Support:** https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasTier3SupportWG

Where do the users start?

- Usage of PanDA/Ganga/Pathena for Tier 3 Analysis means that most of the workflows that worked on Tier 2 analysis should also work on Tier 3.

- User Documentation for the Tier 3 will be integrated to the existing entry-points for Computing and Analysis:
  - ATLAS Computing Workbook is the starting point for users:
    - https://twiki.cern.ch/twiki/bin/view/Atlas/ComputingWorkbook
  - Links to Ganga/Pathena documentation
    - https://twiki.cern.ch/twiki/bin/view/Atlas/FullGangaAtlasTutorial
    - https://twiki.cern.ch/twiki/bin/view/Atlas/DAonPanda

- Here are some Tier3-specific docs:
  - https://atlaswww.hep.anl.gov/twiki/bin/view/UsAtlasTier3/Tier3gUsersGuide
  - https://twiki.cern.ch/twiki/bin/view/Atlas/PandaTier3

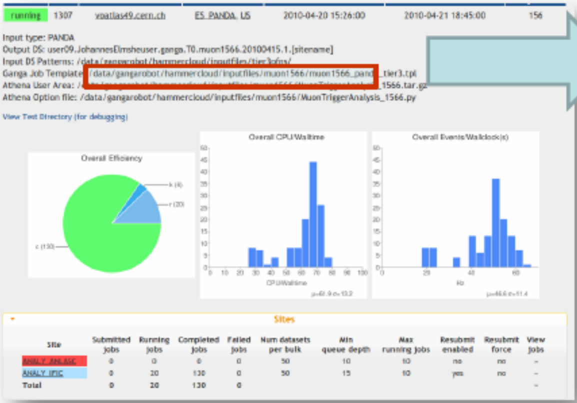
Contacting your site in case of a problem

- In the case that a central support shifter detects a problem with a Tier 3 site, the shifter will:
  - Submit a GGUS or RT ticket to the site, if possible, and
  - Send a mail to the relevant atlas-support-cloud eGroup:
  - Relevant cloud support persons should be connected to this list so that issues can be routed to the Tier 3
  - (For the US the list is atlas-support-cloud-us@cern.ch, and at the moment it has one member: racf-wlcg-announce-l@lists.bnl.gov)
- Why this model?
  - We polled all ATLAS clouds about local plans for support and found:
    - 8/10 clouds have Tier 3's associated with a Tier 2, so the current procedures for ticketing a site via GGUS and using the atlas-support-cloud-xy eGroup will apply to Tier3's.
    - 2 clouds have a dedicated FTE to support Tier 3s.
  - Some Tier 3's may not be connected with a cloud. In this case, the site should provide a contact address.

Getting Help from Other Sites

- Tier 3 system managers should join the global eGroups list atlas-adc-tier3-managers@cern.ch, which will be the primary communication channel *to* and *between* Tier 3 sites:
  - This list has two main goals:
    - To act as the primary contact point for all the ATLAS Tier3 system managers
    - To facilitate discussions within the Tier 3 community regarding software, configurations, or other deployment and maintenance issues
  - "System managers" can mean local system administrators and/or local ATLAS contact persons; the list is not intended for users or user support.

HammerCloud



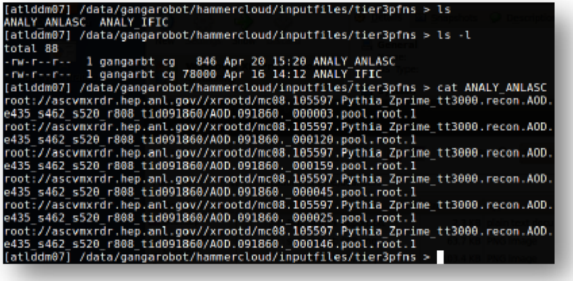
running 1307 voatlas49.cern.ch ES_PANDA_US 2010-04-20 15:26:00 2010-04-21 18:45:00 156

Input type: PANDA
Output DS: user09.JohannesElmsheuser.ganga.T0.muon1566.20100415.1.[sitename]
Input DS Patterns: /data/gangarobot/hammercloud/inputfiles/tier3pfns/
Ganga Job Template: /data/gangarobot/hammercloud/inputfiles/muon1566/muon1566_panda_tier3.tpl
Athena User Area: /data/gangarobot/hammercloud/inputfiles/muon1566/MuonTriggerAnalysis_1566.tar.gz
Athena Option file: /data/gangarobot/hammercloud/inputfiles/muon1566/MuonTriggerAnalysis_1566.py

View Test Directory (for debugging)

Sites

Site	Submitted jobs	Running jobs	Completed jobs	Failed jobs	Num datasets per bulk	Min queue depth	Max running jobs	Resubmit enabled	Resubmit force	View jobs
ANALY_ANLASC	0	0	0	0	50	10	10	no	no	-
ANALY_IFIC	0	20	130	0	50	15	10	yes	no	-
Total	0	20	130	0						-



```
[atlddm07] /data/gangarobot/hammercloud/inputfiles/tier3pfns > ls
ANALY_ANLASC  ANALY_IFIC
[atlddm07] /data/gangarobot/hammercloud/inputfiles/tier3pfns > ls -l
total 88
-rw-r--r--  1 gangarbt cg   846 Apr 20 15:20 ANALY_ANLASC
-rw-r--r--  1 gangarbt cg 78000 Apr 16 14:12 ANALY_IFIC
[atlddm07] /data/gangarobot/hammercloud/inputfiles/tier3pfns > cat ANALY_ANLASC
root://ascvmxrdr.hep.anl.gov//xrootd/mc08.105597.Pythia_Zprime_tt3000.recon.AOD.
e435_s462_s520_r808_tid091860/AOD.091860._000003.pool.root.1
root://ascvmxrdr.hep.anl.gov//xrootd/mc08.105597.Pythia_Zprime_tt3000.recon.AOD.
e435_s462_s520_r808_tid091860/AOD.091860._000120.pool.root.1
root://ascvmxrdr.hep.anl.gov//xrootd/mc08.105597.Pythia_Zprime_tt3000.recon.AOD.
e435_s462_s520_r808_tid091860/AOD.091860._000159.pool.root.1
root://ascvmxrdr.hep.anl.gov//xrootd/mc08.105597.Pythia_Zprime_tt3000.recon.AOD.
e435_s462_s520_r808_tid091860/AOD.091860._000045.pool.root.1
root://ascvmxrdr.hep.anl.gov//xrootd/mc08.105597.Pythia_Zprime_tt3000.recon.AOD.
e435_s462_s520_r808_tid091860/AOD.091860._000025.pool.root.1
root://ascvmxrdr.hep.anl.gov//xrootd/mc08.105597.Pythia_Zprime_tt3000.recon.AOD.
e435_s462_s520_r808_tid091860/AOD.091860._000146.pool.root.1
[atlddm07] /data/gangarobot/hammercloud/inputfiles/tier3pfns >
```

To run these tests, your site needs to be remotely accessible via:

- a Panda queue (ANALY_yoursite)
- a grid CE (this is still in development)

- HammerCloud determines the performance (e.g. Events/Second) of your site under stress
- To run a HammerCloud test you need to:
  - Download some (10's of GB's) data to your site which is useful to our test analysis (typically mc data with muons, e.g. mc09*mu*AOD*)
  - Send us a list of physical-file-names for this data
- Then we arrange a test time and we will submit the jobs.

ATLAS T3 Working Group

- **Virtualisation:** https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasTier3VM

CernVM/CernVM-FS

- CernVM 2.10 stable
- It is official:
  - All Atlas SW will be installed by releases shifters
  - You will see it on CernVM-FS as soon as you receive the email announcement
  - Same for DBReleases
- Coming Soon:
  - Provide condDB and nightlies to CernVM-FS in a separate server

Summary

- **Half year prototyping has been successfully finished**
- **(US) ATLAS has a better idea what a Tier-3 should be**
- **Tier-3 effort now an ADC subproject**
- **Move to demonstrator phase**

- **German Tier-3s mostly at Tier-2s or Grid based Tiers-3**
  - Not much to learn
- **Interesting topics for institutional clusters/desktops**

“Long term”

- Immediate tasks
  - Prepare a plan for sites to become on-line (along with the recommendations of the WG)
  - "Prepare" docs for sites and especially users (I think we are still week (i.e. very unbalanced across sites and analysis groups)
  - Attract more people (sites) for the demonstrator
    - Handful of sites
    - Interest from several places (KIT, ANL, SLAC, Valencia, Belgrade...)
    - Coordination with the CMS demonstrator