ATLAS Tier-3s

Dario Barberis

Genoa University/INFN & CERN

with substantial contributions by

Doug Benjamin

Duke University

ATLAS Tier-3s

What is a Tier-3?

- Working definition
  - "Non pledged resources" (no WLCG Memorandum of Understanding)
  - "Analysis facilities" at your University/Institute/...
- Tier-3 level
  - The name suggests that it is another layer continuing the hierarchy after Tier-0, Tier-1s, Tier-2s...
    - Probably truly misleading...
    - Qualitative difference here:
      - **Final analysis vs simulation and reconstruction**
      - **Local control vs ATLAS central control**
      - **Operation load more on local resources (i.e. people) than on the central team (i.e. other people)**

What is a Tier-3?

- Comments:
  - No concept of size (small Tier-3 vs big Tier-2...)
  - Tier-3s can serve (and be controlled by) a subset of the ATLAS collaboration (local or regional users)
  - They can be just a share in an institute or university computing cluster
- *Non-pledged* resources does not mean *uncontrolled or incoherent*
  - Need to provide a **coherent** model (across ATLAS)
    - Small set of templates to be followed while setting up a Tier-3 for ATLAS users.
  - Coherent because:
    - Guarantee no negative repercussions on the ATLAS Grid (service overload, additional complex manual support load) by the proliferation of these sites

Purpose of a Tier-3

- Tier-3s (and experiment computing in general) are tools to aid the physicists in their work
  - Work – analyzing the data to make measurements and scientific discoveries
  - The computing is a tool and a means to an end
- Tier-3 productivity
  - The success of the Tier-3s will be measured by
    - The amount of scientific output
      - Papers written
      - Talks in conferences
      - Students trained (and theses written)
      - Not in CPU hours or events processed for ATLAS
        - But funding agencies may have a different view

Tier-3 types

- ATLAS examples include:
  - Tier-3s co-located with Tier-2s
  - 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

Tier-3: interesting features

- Key characteristics (issues, interesting problems)
  - Operations
    - Must be simple (for the local team)
    - Must not affect the rest of the system (hence central operations)
  - Data management
    - Again simplicity
    - Different access pattern (analysis)
      - I/O bound, iterative/interactive
      - More ROOT-based analysis (PROOF)
      - Truly local usage
    - "Performances"
      - Reliability (successful jobs / total)
      - Efficiency (CPU/elapsed) → events read per second

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

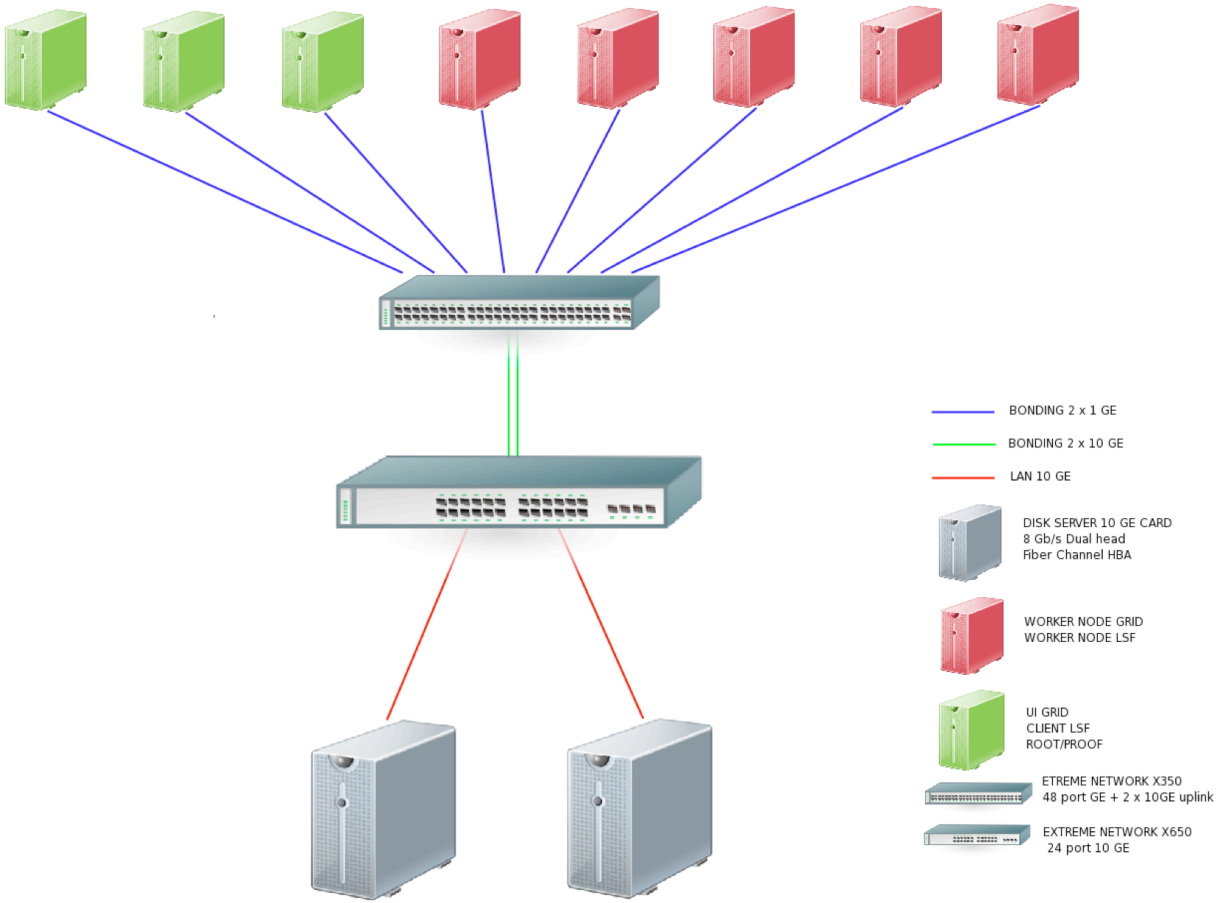
Minimal Tier3gs (gLite) requirements

- 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

A Tier3gs example

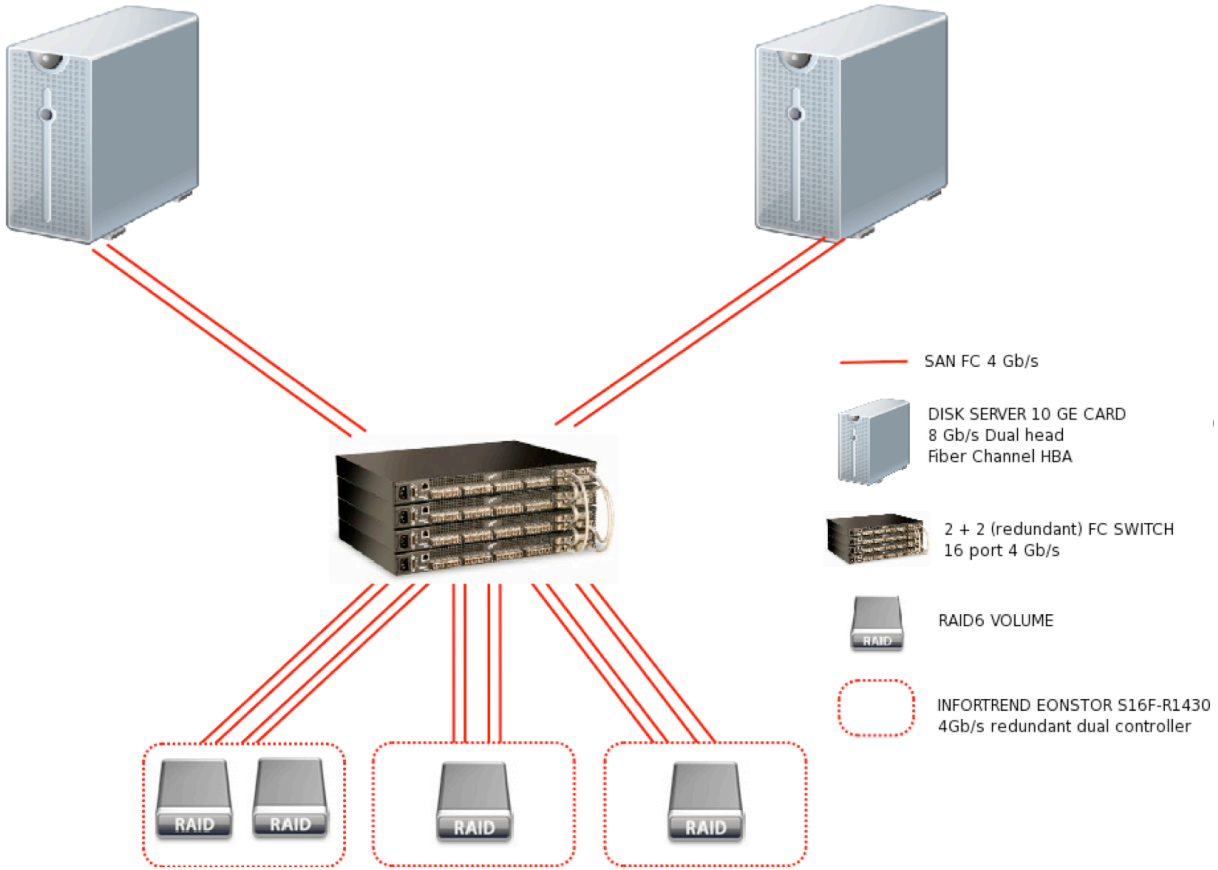
- My own institute (Genoa University and INFN, Italy)
  - ~15 ATLAS members (9 staff physicists, 2 postdocs, 4 PhD students, plus diploma students and engineers)
  - Computing power: 9 machines with 42 cores (3 different generations), total 364 HS06, 2 GB/core of RAM with LSF as batch scheduler
  - Storage capacity: 20 TB (soon 40 TB) managed through GPFS (national licence)
  - Grid Computing Element: LCG-CE
  - Grid Storage Element: StoRM+GPFS
  - Local Area Network: 10 Gb/s
  - Wide Area Network: 100 Mb/s (shared) being upgraded to 1 Gb/s (shared) to the Italian Academic and Research Network (GARR)
  - Support: after setting up, 0.1-0.2 FTE from computing service and 0.1-0.2 FTE from ATLAS group
- Depending on the number of local users, each site should size its infrastructure appropriately
  - The important part is to have an access point to the ATLAS/WLCG Grid
    - To submit jobs and retrieve the outputs
    - To copy data samples locally and run interactively on them

Example: Genoa Tier-3 (CPU)



- Interactive CPUs: 3 bi-processor quad-core Intel Xeon E5520 nodes running at 2.27 GHz with hyper-threading and 24 GB of RAM
- Batch CPUs:
  - 3 bi-processor dual-core AMD Opteron 2218 nodes running at 2.6 GHz with 8 GB of RAM
  - 2 bi-processor quad-core core Intel Xeon E5410 nodes running at 2.3 GHz with 16 GB of RAM
- Each node has a 250 GB local SATA disk and a double integrated GE network card

Example: Genoa Tier-3 (Storage)



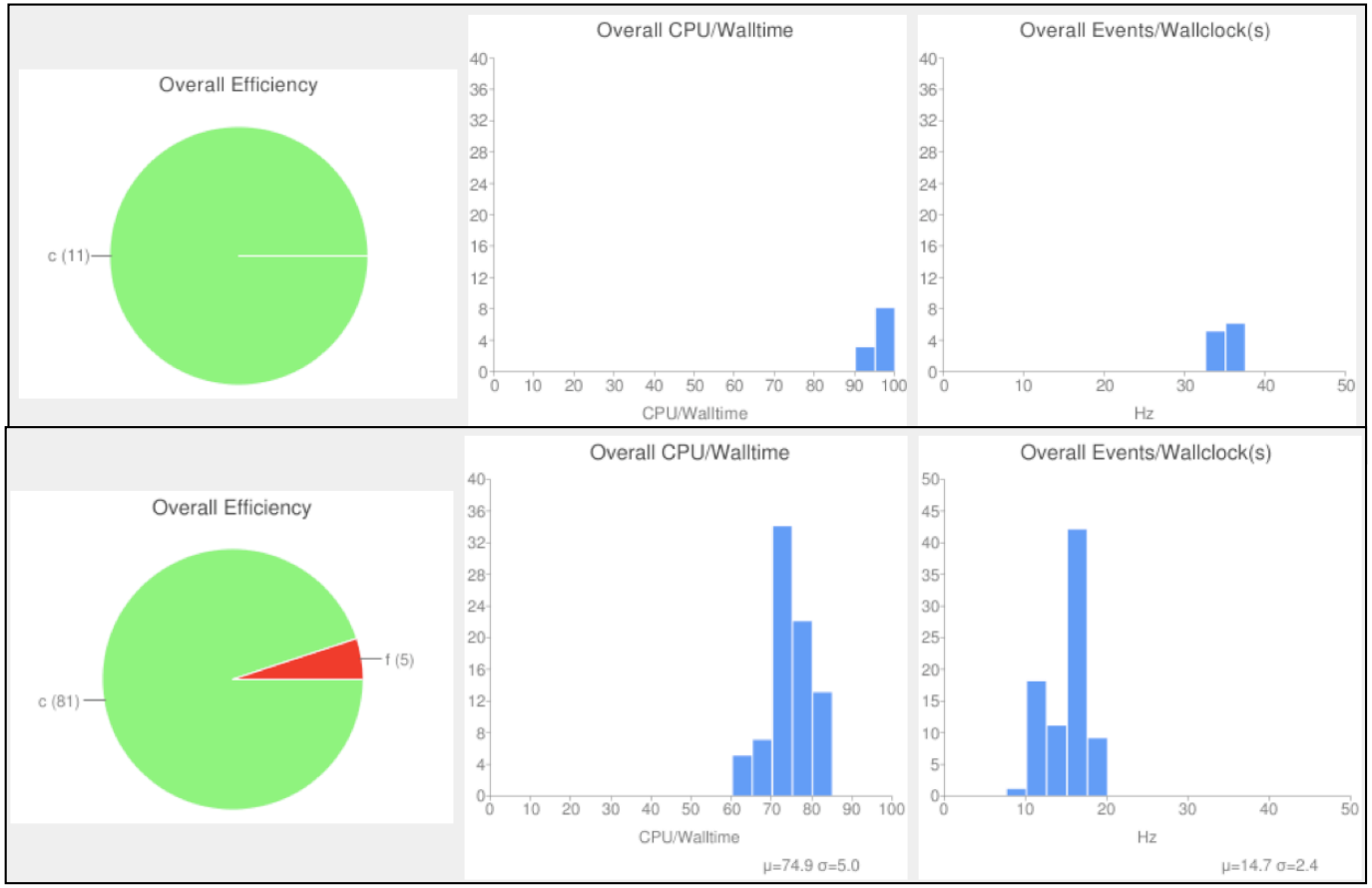
- The storage is based on GPFS with StoRM as SRM interface:
  - 4 RAID sets configured as RAID6 and exported by 3 Infortrend Eonstor S16F-R1430 controllers, each with double 2-channel FC controller at 4 Gbps.
  - The SAN is made of 2 SanBox 5600 FC switches in redundant configuration

Exported via GPFS by 2 bi-processor Intel Xeon L5520 disk servers running at 2.2 GHz, each with 24 GB of RAM.

- The servers are connected through a copper 10 GE interface directly to the backbone switch.
- The SAN connection is through a dual-head 8 Gbps FC card (used at 4 Gbps because of the SAN structure), configured with fail-over and load balancing
- Guaranteed theoretical throughput is 800 MB/s per server

Job performance at Tier-3s

- 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



Example 1: HammerCloud functional test in Genova last Summer run on small simulated AODs:
Very good results!
CPU/WCT = 96%
35 events/sec

Example 2: HammerCloud functional test in Genova last Summer run on larger real AODs:
Still good results, but
CPU/WCT = 75%
15 events/sec

A Tier-3 is not necessarily a small Tier-2

- Of course the recipe *Tier 3 = (small) Tier2* could make sense in several cases
- But in several other cases:
  - Too heavy and a significant investment in the resources to provide a minimum working setup for small new sites
    - "Human cost"
    - The new model is appealing for small Tier2-like centre as well
- In all cases:
  - We got data!
  - The focus is more and more on doing the analysis than supporting computing facilities ;-)

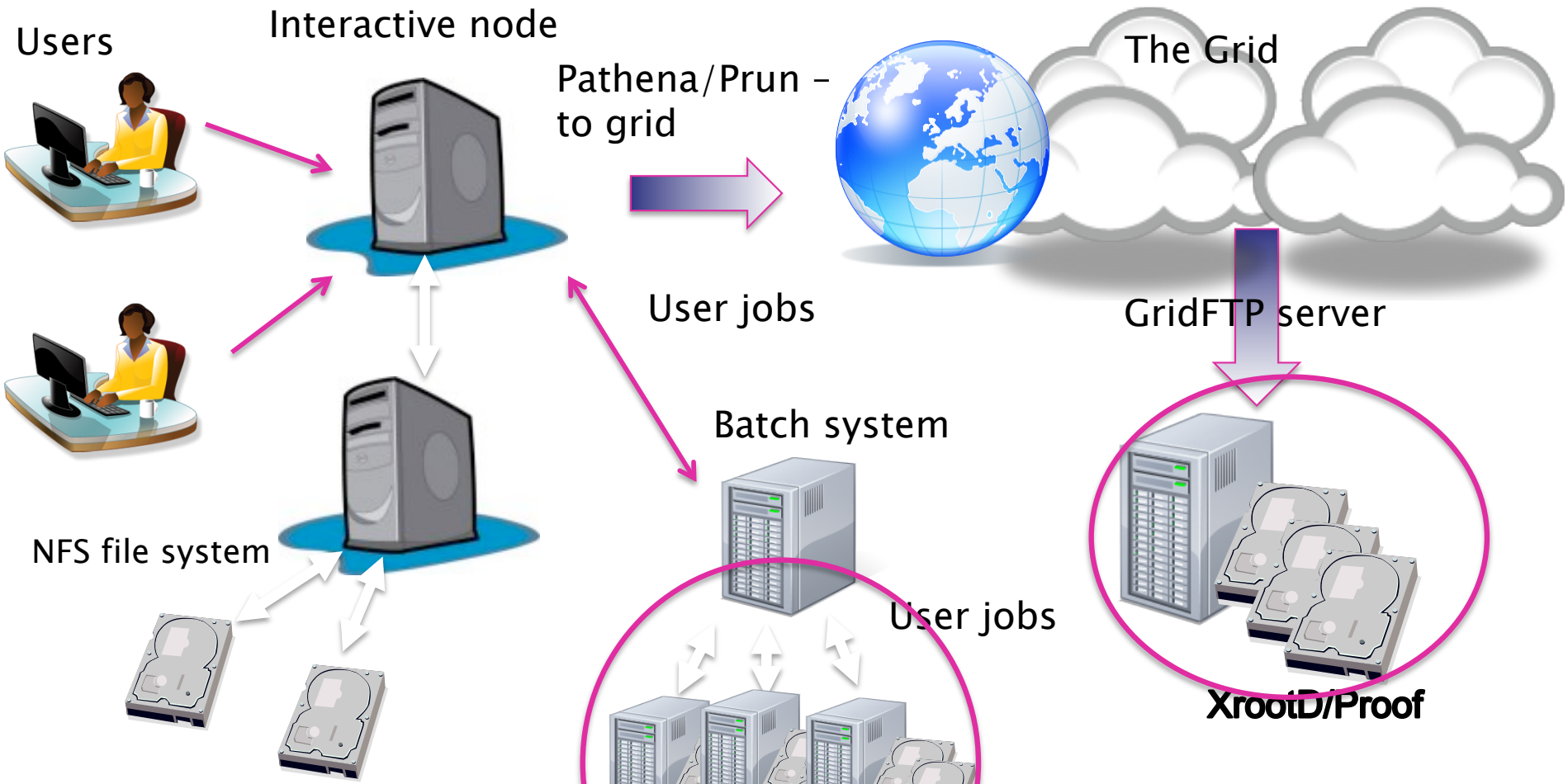
Tier3g design/philosophy

- 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.

Tier3g

- Interactive nodes
- Can submit grid jobs.
- 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 work model



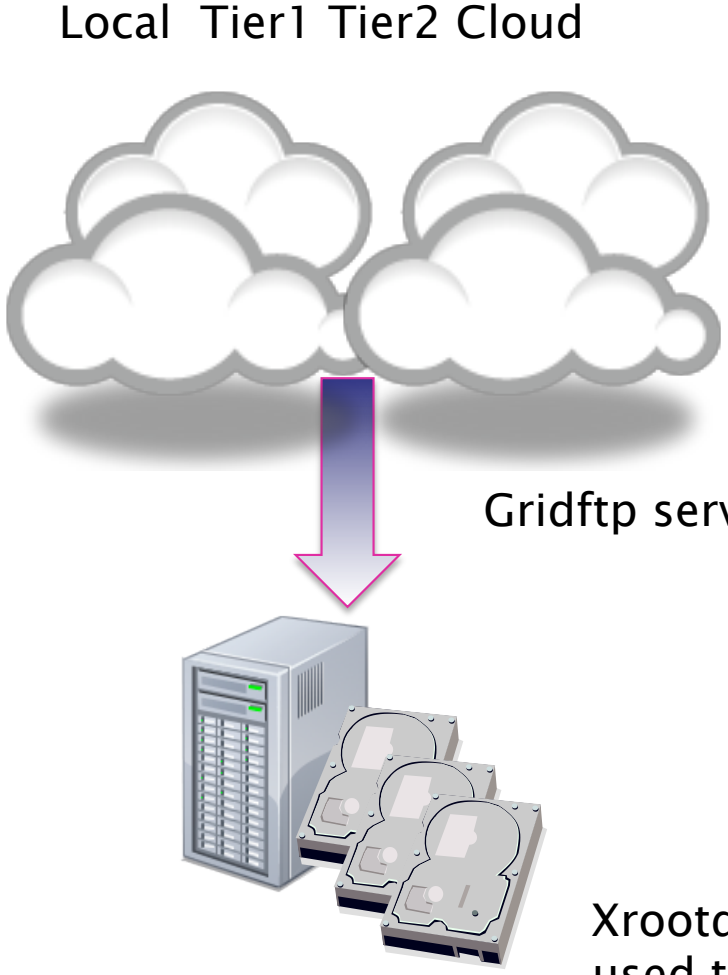
Example: how data come to OSG Tier3g's

Two methods
• Enhanced dq2-get (uses fts channel)

•Data subscription
  •SRM/gridftp server part of DDM TiersOfATLAS

Bestman Storage Resource Manager (SRM) (fileserver)

•Sites in DDM ToA will tested frequently
•Troublesome sites will be blacklisted (no data) extra support load



Data will come from **any** Tier 2 site

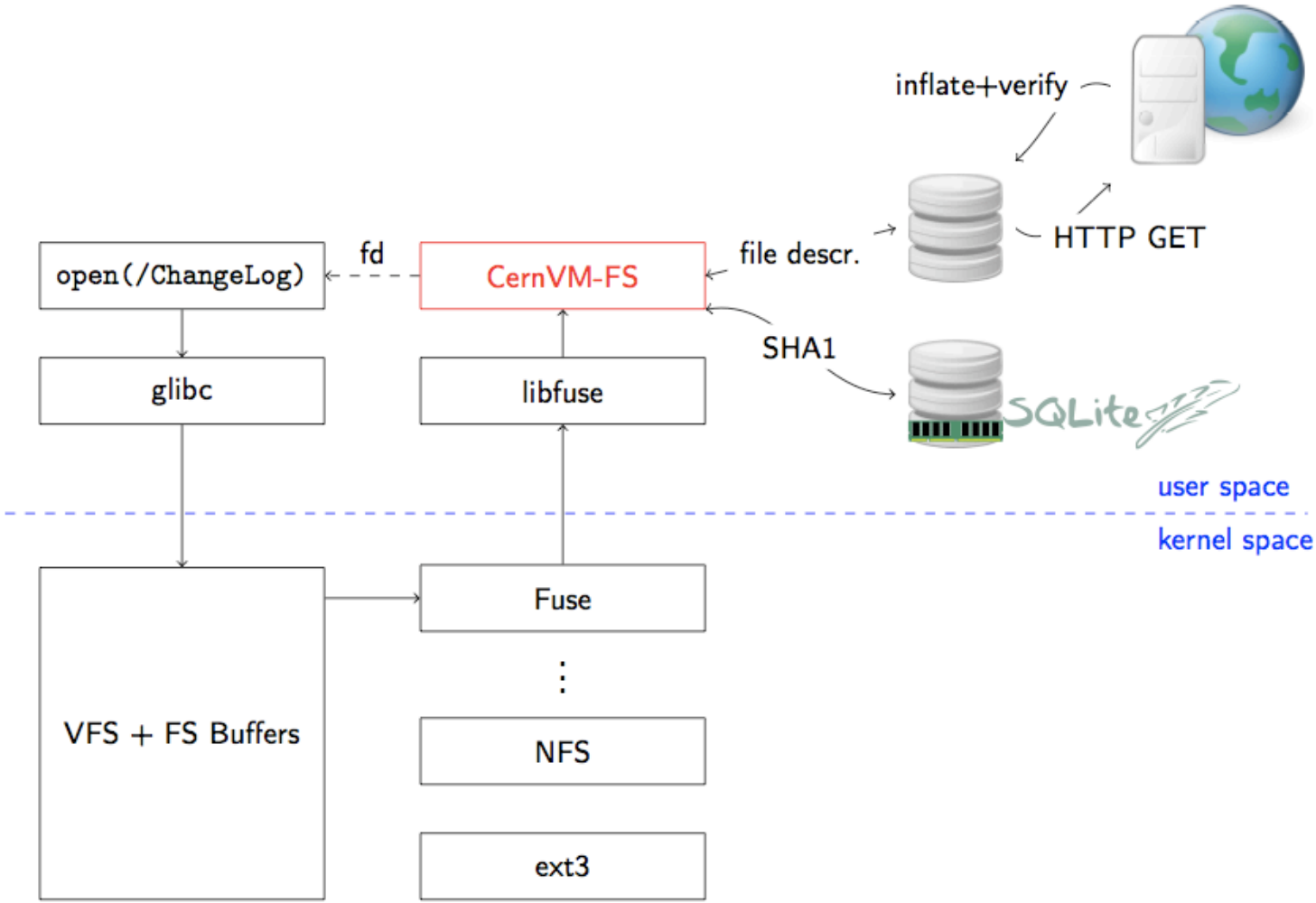
Xrootd/ Proof (pq2 tools) used to manage this

Transformative technologies (1)

- By their operational requirements non-grid Tier 3 sites will require transformative 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

CernVM
SoFTware Appliance

CVMFS (v2)



Transformative technologies (2)

- 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 code installation

- NFS file server
  - ManageTier3 SW package (Asoka DeSilva Triumf)

https://twiki.atlas-canada.ca/bin/view/AtlasCanada/ManageTier3SW

Well tested and straightforward to use

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

Discussion

Comparison of Tier-3 site set-ups

	gLite Tier3gs	NorduGrid/ARC	Tier3g
Minimum CPU	>40 cores or not worth the effort	A few boxes	A few boxes
Minimum disk	>20 TB or not worth the effort Needs space tokens	A few TB (may, or not, have an SRM storage element)	A few TB
Ease of install	Difficult	Easy	EASY
Local access to data on Grid storage	Yes if Posix file system (Lustre/GPFS/Hadoop etc)	Yes, with xrootd or gsidcap (dcache)	YES if xrootd with ROOT/PROOF (default)
Full analysis capability	Yes if complemented by local storage	Yes if local storage and remote transfers are allowed as in ND cloud	YES
Production capability	Yes if site is large and stable enough	Yes if site is stable enough	No (this is not a full Grid site)
Grid usage accounting	Yes	Yes	
Shared site with other users	Yes	Yes	
Data in ATLAS catalogues	Yes	No (only ARC cache) Work in progress on ACIX catalogue export	
Support effort needed after installation	>=0.3 FTE	0.25 FTE	<=0.1 FTE