Atlas Tier 3
Overview
Doug Benjamin
Duke University
Purpose of a Tier 3

- Tier 3’s (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 3’s 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
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
  - Doug Benjamin (technical coordinator)
Atlas Tier 3 Workshop

- Jan 25-26 2010
  - [http://indico.cern.ch/conferenceDisplay.py?ovw=True&confId=77057](http://indico.cern.ch/conferenceDisplay.py?ovw=True&confId=77057)
  - Organizers  Massimo Lamanna, Rik Yoshida, DB
  - Follow on to activities in the US the year before
  - Showed the variety of Tier 3’s in Atlas
  - Good attendance from all across Atlas
  - 6 working groups formed to address various issues
    1. Distributed storage(Lustre/GPFS and xrootd subgroups)
    2. DDM – Tier3 link
    3. Tier 3 Support
    4. Proof
    5. Software and Conditions DB
    6. Virtualization
Tier 3 G
(most common Tier 3 in US)

- Interactive nodes
- Can submit grid jobs.
- Batch system w/ worker nodes
- Atlas Code available
- 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)
Tier 3g 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.
Tier 3g configuration

Users

Interactive node

Pathena/Prun - to grid

User jobs

Batch (Condor)

Local Tier 2 Cloud

The Grid

User jobs

Gridftp server

XrootD/Proof

User's home area

Atlas/common software

NFSV4 mounts

+ web server for Panda logs
Tier 3g – Interactive computing

Users

Interactive node

Pathena/Prun – to grid

The Grid

File server

NFSv4 mounts

User’s home area

Atlas/common software

Common User environment (next slide)
Atlas software installed (two methods)
manageTier3SW
Web file system CVMFS
Atlas Code installation

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

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

Well tested straightforward to use
CVMFS 0.2 (v2)
NFS V4 vs CVMFS
Comparison
Athena Compilations

Rik Yoshida (ANL)
Dell R710: 8 cores (16 hyperthreaded)

<table>
<thead>
<tr>
<th>No. Simultaneous Condor jobs:</th>
<th>1</th>
<th>4</th>
<th>8</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFS4</td>
<td>7 min</td>
<td>15 min</td>
<td>60 min</td>
<td></td>
</tr>
<tr>
<td>CVMFS2</td>
<td>7 min</td>
<td></td>
<td>8 min</td>
<td>11 min</td>
</tr>
</tbody>
</table>


How data comes to Tier 3g’s

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

• Data subscription
  • SRM/gridftp server part of DDM Tiers of Atlas

Bestman Storage Resource Manager (SRM) (fileserver)

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

US Tier 2 Cloud

Data will come from any Tier 2 site

Gridftp server

Xrootd/ Proof (pq2 tools) used to manage this
Implications of ToA

- Tier 3 site will be both a data sink and data source
- Atlas DDM tests required to run at T3 site on a fixed cycle (appropriate for T3’s)
- File remove implies remove in database before remove at Site or Dark data
- Site can/will be black listed if you fail too many DDM tests. -> No data
- Must provide good quality of service
Tier 3g Storage issues

- Monitoring of Data storage
  - Tier 3’s will have finite amount of storage. Storage will be like a cache – Need to monitor data usage patterns to determine Data longevity at site. (clean up old data)
  - Storage system performance monitoring
  - Xrootd and Proof master might help here
  - Need help from XrootD team and OSG for better packaging and installation scripts
  - Alice uses **Mona Lisa** to monitor its XROOTD systems

- Efficient Data Access
  - Tier 3g sites typically have 1 Gbe between worker nodes and storage. Want to minimize network traffic with cluster
  - Implies moving jobs to data if possible
  - Xrootd - perhaps HADOOP as a solution
Tier 3g Storage issues (cont)

- Data security (data safety)
  - In a Tier 3 want to maximize the amount of space available at minimum copy
  - Implies commodity hardware
  - Some data will be able to be fetched again from BNL cloud (US Tier 1 and Tier 2 sites)
    - This will put an added tax on wide area network between Tier 3 and Tier 1 & 2 sites... Why not think transfer among Tier 3 centers (Data cloud)
  - Some sites will store save multiple copies of data on different platforms (Raid NFS fileserver – Xrootd system)
  - Why not use a file system that can provide automatic replication (HADOOP)

- Networking
  - We often take networking for granted – yet it needs to optimized for efficient transfers. - implies interaction with Campus Network admin.
  - Internet 2 has agreed to help us (Thanks!)
IllinoisHEP T3gs Storage

• Atlas T3 with Grid Services

• Panda site
  – Production queue (IllinoisHEP-condor)
  – Analysis queue (ANALY_IllinoisHEP-condor)

• Software
  – Scientific Linux 5.5 (64 bit)
  – dCache 1.9.5-21 (Chimera) installed via VDT 3.0.3

• Hardware
  – 8 nodes (3 doors, 1 head, pNFS, and 3 pool nodes)
  – Dell R710 (E5540, 24GB) and Intel (E5345, 16GB)
  – H800/MD1200/2TB SAS disks (144TB raw)
  – 12 Drives (1 Tray) per Raid 5 set with 512KB strip size, XFS file system
  – 10Ge network (HP5400, Intel Dual CX4)
IllinoisHEP T3gs Storage

• Good performance
  – Pool nodes are over 1GB/s read, 800MB/s write via dd, 600MB/s Bonnie++
  – FTS transfers over 700MB/s

• Issues
  – dCache 1.9.5-19 and -21 fixed many problems
  – Update using VDT package is very easy

• Network tuning very important
  – 10Ge tuning different than 1Gb
  – Cards need to be in 8x PCIe slots (R710 has both 8x and 4x)
  – Much larger memory needs

• Problems seen with bad tuning
  – Broken network connections
  – Files transferred with errors (bad adler32 checksums)
IllinoisHEP T3gs dCache Tweaks

• Some tweaking of dCache parameters recommended by T2 sites
  – Use 64bit java with memory increase to 2048/4096M
  – gsiftpMaxLogin=1024
  – bufferSize=8388608
  – tcpBufferSize=8388608
  – srmCopyReqThreadPoolSize=2000
  – remoteGsiftpMaxTransfers=2000

• Use Berkley Database for meta data on pool nodes

  metaDataRepository=org.dcache.pool.repository.meta.db.BerkeleyDBMetaDataRepository
Conclusions

- Tier 3 sites come in different varieties
- The staffing levels at Tier 3 sites is smaller than that of Tier 2 sites
  - Should keep things as simple as possible
  - Need to be as efficient as possible
- The success of the Tier 3 program will be measured by **physics productivity**
  - Papers written, Conference talks given and Student theses produced
- In the US Tier 3’s are being fund with ARRA funds - we can expect enhanced oversight...
- With collaboration w/ CMS and OSG we can make it a success.