AGLT2 Site Report

Benjeman Meekhof

University of Michigan

HEPiX Fall 2013

ATLAS Great Lakes Tier 2

University of Michigan AGLT2 Michigan State University

Outline

- Site Summary and Status
- Transition to SL6 and Rocks 6
- Provisioning with Cobbler
- Complete transition to CFEngine 3
- AFS on Linux ZFS
- dCache Status
- Virtualization Status
- Networking Upgrade
- Conclusion

Site Summary

- The ATLAS Great Lake Tier-2 (AGLT2) is a distributed LHC Tier-2 for ATLAS spanning between UM/Ann Arbor and MSU/East Lansing. Roughly 50% of storage and compute at each site
  - 4616 single core job slots
  - 10 eight-core (multi-core) job slots
  - 350 Tier-3 job slots usable by Tier-2
  - Average 9.03 HS06/slot
  - 3.5 Petabytes of storage
- Most Tier-2 services virtualized in VMware
- 10Gb redundant ring for inter-site connectivity, lots of 10Gb internal ports and 12 x 40Gb ports
- High capacity storage systems have 2 x 10Gb bonded links
- 40Gb link between Tier-2 and Tier-3 physical locations

Rocks 6 and SL6

- Installed using Rocks 6.1 jumbo roll
- Uses SL6.4 as OS roll
  - Created from SL6 rpm repo using "rocks create mirror“
- DB of appliances, WN, etc created using ASCII command dumps from previous Rocks5 headnode
  - Utilized various "rocks dump" and "rocks list" commands
  - Separate head nodes for WN at MSU and UM sites
- Initial rollout to a subset of WN, fed by an SL5 gatekeeper
  - Re-tested with SL6 gatekeeper
- Full rollout in May, 2013

Rocks 6 and SL6

- Rocks software distribution via rolls (T. Rockwell original development)
  - Eg, agl-osg3, agl-condor, agl-cvmfs....
  - Keeps various functionality logically separated
  - Understanding of rpm significance is greatly simplified
- Rolls updated to SL6, built, and distribution tested
  - Based upon existing SL5 xml maintained in AGLT2 svn repository
  - SL5 functionality maintained with additional SL6 requirements
- Additional post-build configuration applied via CFEngine 3

Provisioning with Cobbler

- Before Cobbler: simple PHP script to generate kickstart files using information inserted into Rocks database (for lack of better place)
- Now we have Cobbler server configuration managed by CFEngine and duplicated at both sites for building service nodes (excepting site-specific network/host info)
  - Created flexible default kickstart template with Cobbler's template language (Cheetah) to install a variety of "profiles" as selected when adding system to Cobbler (server, cluster-compute, desktop, etc).
  - Simple PXE based installation from network
  - Cobbler handles (with included post-install scripts) creating bonded NIC configurations – used to deal with those manually
  - Cobbler manages mirroring of OS and extra repositories
- Kickstart setup is kept minimal and most configuration done by CFEngine on first boot
- Dell machines get BIOS and Firmware updates in post-install using utils/packages from Dell yum repositories
- Long term we plan to replace Rocks with Cobbler for compute nodes

Complete transition to CFEngine 3

- Our first configuration management system was CFEngine 2
  - Implemented at UM, not MSU
  - Not used in compute cluster, only service nodes
- About 1+ years ago introduced CFEngine 3 for basic management of compute cluster – other systems still version 2.
- Finally this year we've migrated all necessary CFEngine 2 logic to version 3
  - UM and MSU sites use identical policy code for compute and service nodes
  - Policy changes versioned in SVN
- Developed workflow which easily allows any person to work on their own SVN branch and export it for testing from policy server
- Performed extensive evaluation of Puppet/Foreman
  - Combination of Foreman/Puppet/Hiera/PuppetDB has interesting features
  - We have already extensive CFEngine expertise and working policy
  - Prefer the simplicity and flexibility of CFEngine
  - Cobbler isn't as pretty as Foreman but good CLI, ISO export, simpler setup.

AFS on Linux ZFS

- At AGLT2 we run our own AFS cell **atlas.umich.edu**
- Originally this was run on dedicated hardware nodes and had 3 DB servers and 3 file servers (1TB /vicepX partitions)
- We added an additional two file servers at CERN in 2008 to provide specific storage for ATLAS muon chamber data
- Once we invested in VMware we migrated our AFS services there
  - All three DB servers were virtualized
  - The three file servers at UM were virtualized and the /vicepX partitions were moved on iSCSI locations (still using ext3)
- As ZFS on Linux was available we migrated to it for our AFS /vicepX storage: compression, snapshots and data integrity were primary motivations
  - Compression providing factor of **1.26-1.6** increase in space depending on partition
  - 'zfs get "all" zfs/vicepg | grep compressratio'
- One issue: snapshots "steal" filesystem space! Have had "full" partitions a few times requiring us to manually fix. Just need to tune snapshots and balance space
  - Check space used via 'zfs list –t snapshot'
- **Has worked well for us over the last few months.**

dCache Status

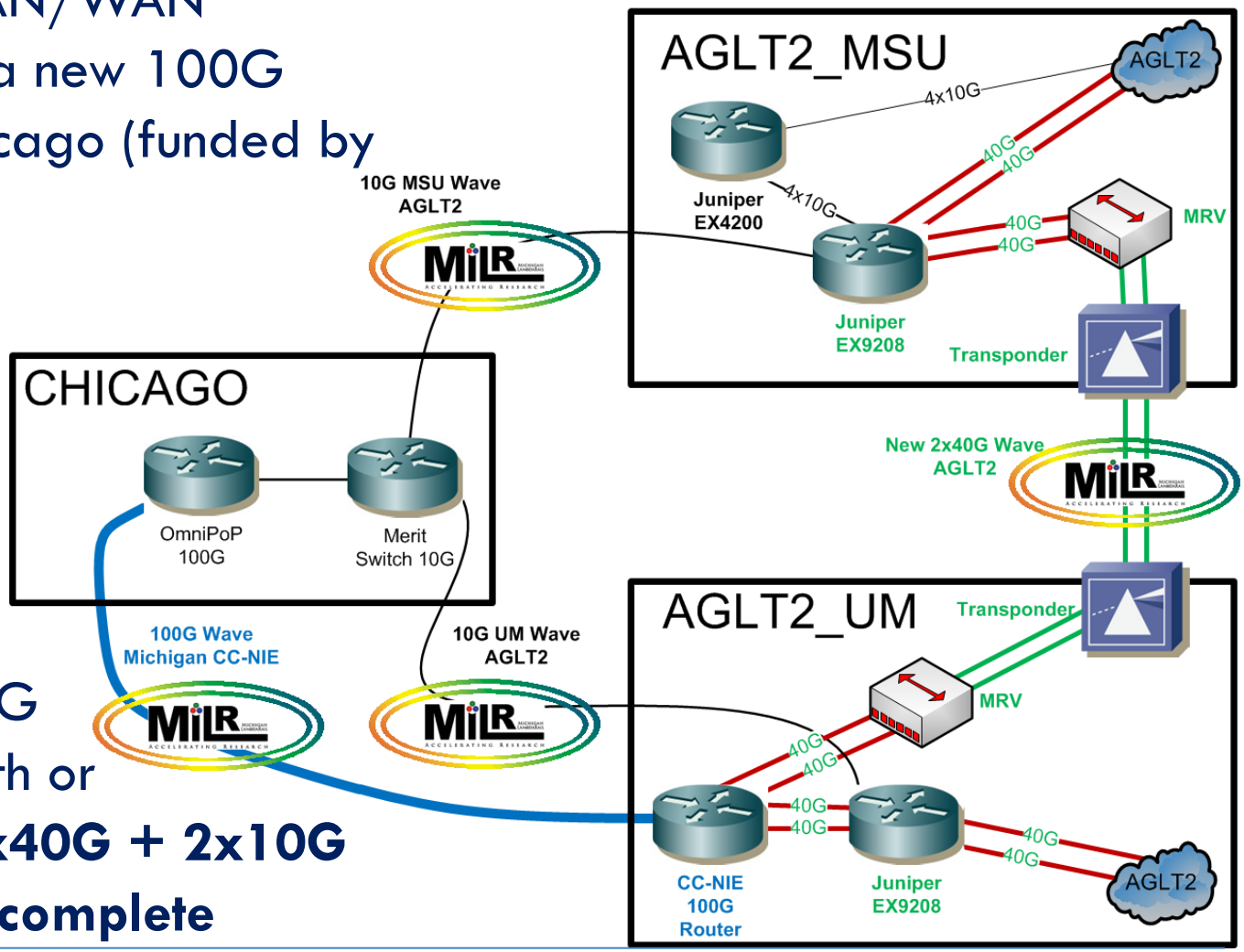
- AGLT2 running dCache 2.2.17 now
- We hit unusual problem with billing DB related to rate
  - Seems billing was unable to keep up with load
  - Cause out-of-memory in other cells as backlog piled up.
  - Lots of support emails and debugging resulted in a revised/tuned set of postgresql triggers. Problem solved!
- Problem: FAX xrootd backport doesn't work
  - Can use 2.2.15 for xrootd backport but no SHA-2
  - **Temp-fix: run 2.2.15 on xrootd door node only**
- dCache 2.6.x supports BOTH xrootd backport and SHA-2 BUT requires gPlazma2 only…no gPlazma1 support
  - AGLT2 had some issues converting their gPlazma1 config into gPlazma2.
  - Tigran helped get the conversion working…now ready! Plan to upgrade soon

Virtualization Status

- Most Tier-2 services on VMware (vSphere 5.1)
- UM uses iSCSI storage backends
  - Dell MD3600i, MD3000i and SUN NAS 7410
  - vSphere manages virtual disk allocation between units and RAID volumes based on various volume performance capabilities and VM demand
- MSU runs on DAS – Dell MD3200
- Working on site resiliency details
  - vSphere and storage operational at MSU
  - SSO operational between sites
  - Still exploring vSphere Replication and other capabilities
  - Need to determine details of overall site resilient configuration, how to move services and inform upstream (or do transparently)
- Goal is to have MSU capable of bringing up Tier-2 service VMs within 1 day of loss of UM site
  - MSU is already operating many site-specific Tier-2 VMs (dcache doors, xrootd, cobbler) on vSphere
  - Need to finish testing and specifics of disaster recovery for UM Tier-2 services

Networking Upgrade

- AGLT2 is in the process of upgrading our LAN/WAN
- U Michigan has a new 100G connection to Chicago (funded by NSF CC-NIE)
- Tier-2 used project funds to purchase Juniper EX9208 routers at UM/MSU
- Deploying 2x40G over the next month or two. **Will have 2x40G + 2x10G to the WAN once complete**



perfSONAR-PS at AGLT2

- Shawn will report tomorrow on WLCG and perfSONAR-PS but I wanted to update you on how we use perfSONAR-PS at AGLT2 in this site report
- The UM site has 3 sets of instances on physical machines
  - Original KOI boxes (5 years old) are now in Tier-3 area
  - Test-instance Dell R410 using KVM to provide bandwidth/latency
  - New Dell R310/R610 "production" versions at Tier-2
- In addition UM has a virtualized latency instances "pinned" to each VMWare host system to verify vSwitch connectivity
- MSU also has 3 sets of instances within their Tier-2
  - Original KOI boxes at the network distribution layer
  - Production instances (Dell R310s) at access layer (near storage)
  - Additional two boxes at the AGLT2 network border
- This config allows debugging of our LAN as well as WAN

Conclusion

- The Tier-2 is running well, site resiliency work is nearing a "final" solution, and Scientific Linux 5 is close to being no more than a memory
  - Some minor service nodes remain SL5 (webservers, svn server, some others)
- The move to Scientific Linux 6 was a major transition but it pushed us to move more quickly on transitions to new provisioning system and CFEngine 3. Our site is more manageable and organized today than ever before
  - Cobbler is working out well, should be able to phase out Rocks for compute cluster
  - Still see occasional SL6 issues not specific to our site
- Our new Juniper switch is installed and will become the "core" switch very soon
- Plans to upgrade to 40G network are underway
- Despite the problems a couple months ago, dCache is running stably for us now