AGLT2 Site Report

Shawn McKee/University of Michigan

Bob Ball, Chip Brock, Philippe Laurens, Mike Nila, Forrest Phillips

HEPiX Spring 2018 / UW Madison

ATLAS Great Lakes Tier 2
University of Michigan AGLT2 Michigan State University

AGLT2 Numbers

- 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
  - 10680 logical cores
  - M(ulti)CORE job-slots 1188 (dynamic) + 10 (static)
  - Additional 936 Tier-3 job-slots usable by Tier-2
  - Average 10.69 HS06/job-slot
  - **6.9 Petabytes** of storage
  - Total of **124.2 kHS06**
  - Tier-2 services virtualized in VMware 5.5 (by end of May upgrading to 6.5)
- 2x40 Gb inter-site connectivity, UM has 100G to WAN, MSU has 10G to WAN, lots of 10Gb internal ports and 20 x 40Gb ports, 32x100G/40G or 64x50G/25G ports
- High capacity storage systems have 2 x 50Gb or 2 x 100Gb bonded links
- 40Gb link between Tier-2 and Tier-3 physical locations

Personnel Updates

- Over the last year we have lost two tier-2 cluster administrators, one at UM and one at MSU
- As of September 1, 2017 we have added a new physics grad-student at MSU working on AGLT2 at 50% time: **Forrest Phillips**
- As noted in the fall meeting, **Bob Ball** our Tier-2 Manager since we started in 2006, plans to retire this year ☹
  - Bob has started a phased retirement as of May 1 and will be fully retired on November 1. He will be missed!
    - Hopefully Bob can make it to the Fall 2018 HEPiX!
  - Wenjing Wu has been hired to begin in July and overlap with Bob for several months.

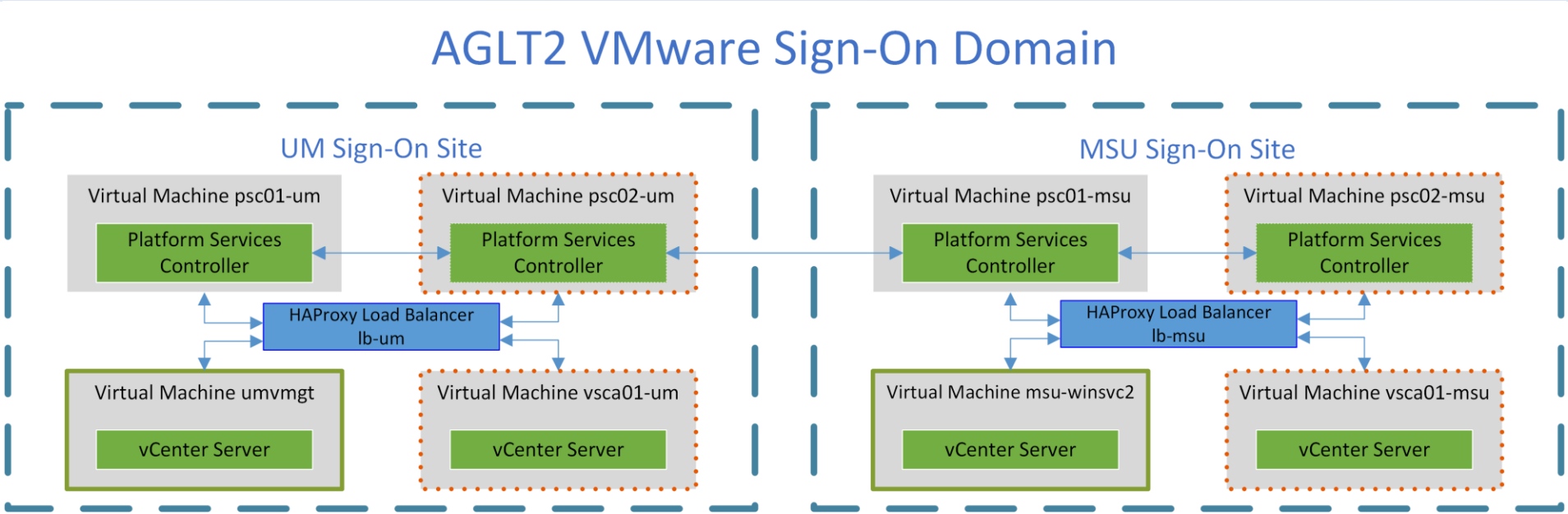
Hardware Additions

- We completed our FY17 hardware purchasing and also used some FY18 funds
  - Purchased Dell C6420 platform: 5 at UM, 3 at MSU
  - Config: Each sled 2xGold 6132 (14C/28T), 192G, 2x1.2TB 10K SAS, dual 10G
- Bought some VMware storage at MSU
  - Need storage under warranty: went with SAS attached MD3460, 20x1.2TB 10K SAS disks plus 40x6TB 7.2K NL-SAS disks
  - MSU added a third VMware ESXi host (repurposed R630 WN)
- Both MSU and UM sites purchased more network ports
  - S4048-ON at MSU and N2048 at UM
- New "edge server" (Dell R740) ordered in early May (http://slateci.io)
- Additional purchased some equipment to optical split incoming/outgoing WAN wavelengths (see later slide)

Lustre at AGLT2

- Lustre at AGLT2 is used for “local” users (Tier-3)
  - About **1.5 PB** of space running on SL7/Lustre 2.10.1/ZFS 0.7.7-1
  - ZFS 0.7.7 on the OSS, ldiskfs on MGS/MGT
  - Upgrading “soon” to 2.10.4 to support SL7.5 kernels
- 20Gb/s bonded Ethernet on OSS
- Lustre Re-Exported to MSU WN via NFS from Lustre SL7.4 client
  - Same OS and kernel as on the OSS
- Data read rate shows significant increase (~x2; up to 2GB/s aggregate seen) over older SL6/2.7 combination using older SL6/2.7 clients

VMware Update



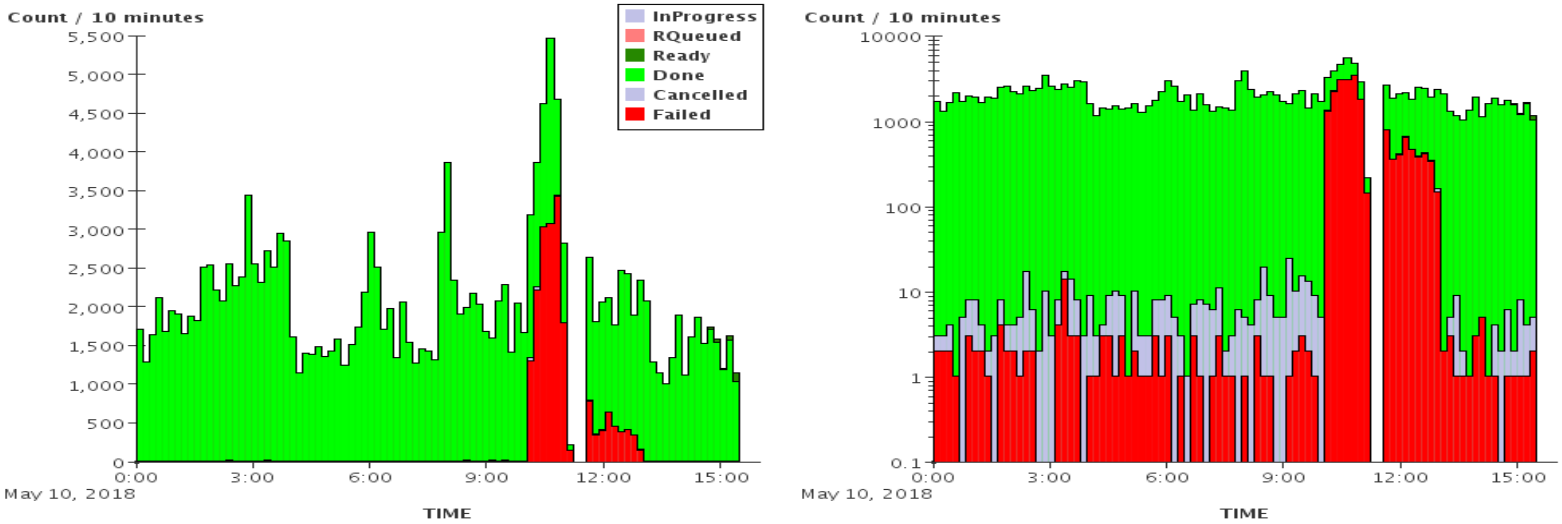
We are planning to upgrade our VMware vSPhere deployment from 5.5 t 6.5 by the end of the month

The final topology will follow the above diagram with resilient services at each site fronted by load balancers and all using a single sign-on domain

IPv6 Updates

- While AGLT2 has had a /48 in IPv6 for a couple years we had only used it for perfSONAR
- About a month ago we added IPv6 addresses to all our dCache systems
  - Simple IPv6 addressing scheme <IPv6_Prefix>:aaa:bbb:ccc:ddd where IPv4 aaa.bbb.ccc.ddd
  - Example umfs01.aglt2.org **192.41.230.21** IPv4 and **2001:48a8:68f7:1:192:41:230:21** IPv6
- Last week we enabled AAAA record…no issues
- Next day we "fixed" missing IPv6 default gateway…issues!

IPv6 Updates(2)



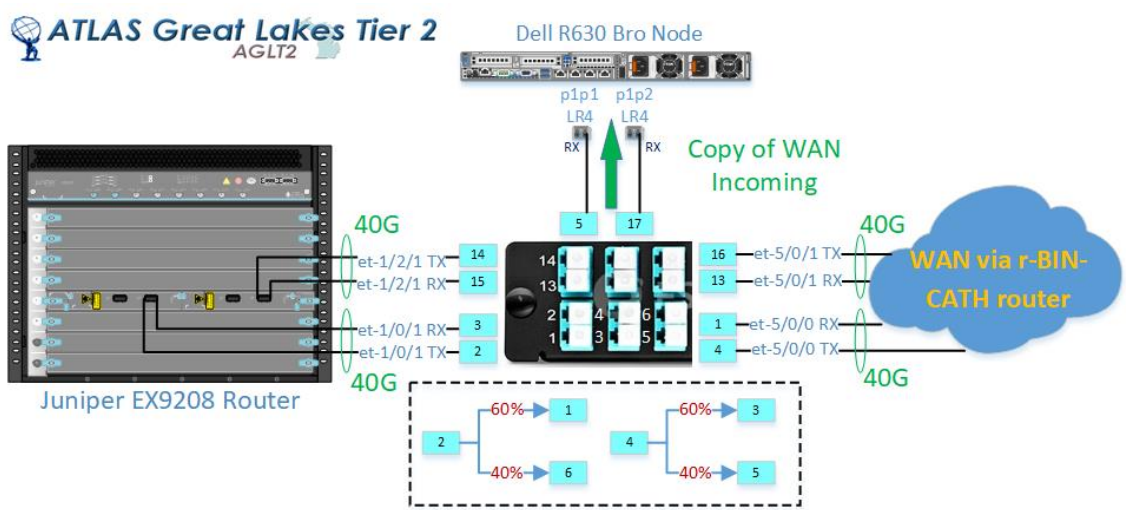
Once the IPv6 gateway was in place we really started using IPv6.

- The problem was we missed defining IPv6 networks for dCache; AGLT2 uses networks to identify the UM vs MSU locations to determine read/write behavior
- Reconfiguring fixed most of the problems…restarting some problem services on specific nodes fixed the rest
- Has been running great since then

dCache and Open vSwitch

- AGLT2 implemented Open vSwitch (OVS) on our dCache nodes last Fall
  - **Goal**: provide a mechanism to test Software Defined Networking for LHC physics tasks
  - OVS gives us a means to get to the "source & sink" of data (visibility and control)
  - https://www.aglt2.org/wiki/bin/view/Main/Open_vSwitch/InstallOVSonSL73
- Now all the AGLT2 dCache storage servers are running SL7.4 and have their public IPs (IPv4 and IPv6) on OVS
  - Installing OVS and cut-over the IP to the virtual switch was done "hot" on production servers. Approximately 40 second downtime not a problem for transfers
- **We are trying to encourage other ATLAS sites to join us so that we can start experimenting to determine the impact of SDN / NFV**
  - **MWT2 and KIT are interested**
  - Work will be managed in the context of the HEPiX NFV working group
  - See report tomorrow afternoon in the network session

Optical Splitter / Bro / MISP



AGLT2 has been working with the WLCG Security Operations Center (SOC) effort (uses Bro/MISP)

Have enabled cost effective ($1.1K) optical tap configuration to track 80 Gbps WAN traffic to AGLT2 at UM

Future Plans

- Participating in **SC18**
  - Will demo use of **OSiRIS** storage service: sign up on the floor and use the storage during SC18 and possible use of object store for ATLAS
  - Possible **AGLT2** / **ATLAS** demos showing use of OVS and NFV
- Experimenting with SDN/OVS in our Tier-2 and as part of LHCONE point-to-point testbed and in coordination with the HEPiX NFV working group.
- Update to **VMware** soon: new version (5.5->6.5), new configuration for HA, new ESXi host at MSU
- Working on IPv6 dual-stack for **all nodes** in our Tier-2 with aglt2.org interfaces
  - Have IPv6 address block for AGLT2 (spans UM/MSU)
  - **Dual-stacking our dCache system is now in effect.**
- Moving all worker nodes to **SL7** in the next months
  - All of MSU WN, 1/3 of UM WN now updated. Complete by end of summer?
- **Continue work with Security Operations Center effort**
  - **Exploring cost-effective means to capture and analyze traffic to/from AGLT2**

Summary

- Tier-2 services and tools are evolving. Site continues to expand and operations are smooth.
- FUTURE: IPv6 for other than storage, SL7 WN, SDN testing, SC18

Questions ?

Additional Slides

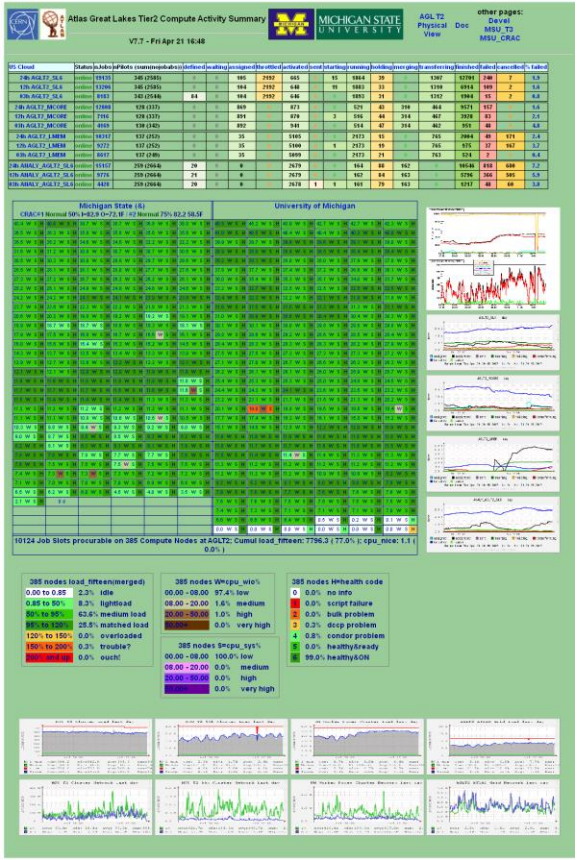
Backup slides follow

AGLT2 Monitoring

AGLT2 has a number of monitoring components in use
As shown before we have:

- Customized “summary” page ------->
- **OMD** (**O**pen **M**onitoring **D**istribution) at both UM/MSU
- Ganglia
- Central syslog’ing via ELK: Elasticsearch, Logstash, Kibana
- SRMwatch to track dCache SRM status
- GLPI to track tickets (with FusionInventory)

We find this set of tools indispensable for proactively finding problems and diagnosing issues.



Software Updates Since Last Mtg

- Tier-2 VMs rebuilt to use SL7 (some were old SL5!)
- dCache updated to 4.0.5 (still Postgresql 9.5)
- HTCondor running version 8.6.10
- OSG CE updated to 3.4.11
- Various switch firmware updates applied
- Monitoring updates: OMD/check_mk to 1.4.0p25
- Two “major” updates: Lustre and dCache/OVS