USATLAS Tier 2 Status

Frederick Luehring
luehring@indiana.edu
Indiana University
Virtual USATLAS Facility Meeting
Monday, March 16, 2020

Outline

- This talk will be my impressions after 6 weeks or so of being the Level 2 Tier 2 manager.
- Outline:
  - Work areas / issues
  - Status of USATLAS Tier 2 Sites
  - Recent incidents
  - Plans

Work Areas

- Middleware / software / OS requests upgrade requests
  - dCache 5.2
  - OSG 3.5, Condor 8.8
  - CentOS 7 / SL7
  - ipv6
- New equipment
  - Can't just buy stuff
    - Careful study required to spend budget efficiently
  - Takes more effort than people realize
    - Prototype
    - Choose
    - Order
    - Install
    - Test

Issues

- Stability Issues
  - Lots of variation from month to month in our HS06 hours
  - Slow response to reported issues
    - Too much hardware and too few admins? Remote management tools may help with this if we can agree to use them
  - Also we are slow to respond to external requests and struggle to get to stable running after acting on request.
- Monitoring difficulties
  - Problems are being missed
  - The number of monitoring tools is approaching infinity
- The sites vary significantly making managing the project complicated and it harder to share expertise.
  - Each physical site is in its own bubble

Site Status: OS, OSG, Batch

Site	Linux Dist	OSG Vers	Batch Sys	Batch Vers
AGLT2	SL7 / CentOS7	3.4? ATLAS 3.5 OSG, Test	HTCondor	8.6.13 head 8.8.7 GKs
MWT2	SL 7.7	3.5 / 3.4 on a few SL6 sys.	HTCondor HTCondor-CE	8.8.7 4.0.1
NET2	CentOS 7.5	3.4	SGE	
SWT2 - UTA	CentOS 7.6	3.4	SLURM	17.11.13
SWT2 - OU	CentOS 7.3	3.4	SLURM	17.11.5

AGLT2 is in the process of upgrading to OSG 3.5 / HTCondor 8.8.7.

Site Status: CPU, Network

Site	# Cores	HS06	HS06/ core	Uplink
AGLT2	11,768	134,536	11.02	100G UM up, 10G MSU up. 100G between sites
MWT2	27,896	308,968	11.08	80G up, 100G pipe UC 80G up, 100G pipe IU 200G up UIUC / NCSA
NET2	13,028	135,157	10.37	100G GPFS & NESE
SWT2 – UTA CPB	10,296	109,164	10.60	40G up, 100G pipe
SWT2 – OSCER	2,744	26,558	9.68	10G up, 100G pipe
SWT2 – Total	16,244	164,801	10.15	

AGLT2 has 208 cores set aside for BOINC.

MWT2: There may be some retirements at NCSA where servers are retired based on the calendar rather than running them into the ground which is our normal operational mode.

Site Status: Storage

Site	PB	Backend	Vers	IPV6 Storage	IPV6 Other
AGLT2		dCache	5.2.16	Yes	
MWT2	11.3	dCache	5.2.15	Yes	PerfSONAR UC & UIUC
NET2	5.1 1.6	GPFS CEPH		Planned	Could be delayed
SWT2 – UTA CPB	5.5	XRootD	4.10.0	No	Engin. hired
SWT2 – OSCER	.7	XRootD	4.11.1	Close	Need server
SWT2 – Total	6.2			No	No

MWT2 was planning to retire 1.4 PB in 2019 but did not.

NET2 could see further delays setting up IPV6 because network team is busy dealing with setting up for online classes. The CEPH storage is buy-in and 1.6 PB of 5.9 PB is used.

SWT2 numbers don't reflect an ongoing process of retiring ~0.7 PB and installing ~1.5 PB of new storage. The SWT2 numbers also don't include include .3 PB of old storage at UTA_SWT2.

CPB Low Efficiency

- On about October 23, 2019 something happened affecting the CPB site at SWT2.
  - The number of wall hours stayed about the same ~5.5 million hours per month (± ~0.5 million hours/month).
  - The CPU efficiency dropped below 40%.
  - The CPU hours dropped from 4-5 million/month to under 2 million hours/month.
- On March 4, 2020 Patrick McGuigan fixed an AGIS setting causing RUCIO mover to be tried before LSM.
  - This caused a long waits for RUCIO transfers to timeout.
  - Reversing the AGIS setting solved the issue.
- Patrick believes that someone at CERN changed the order when there were other ongoing problems.

MWT2 Job Failures

- After the recent dCache upgrade, MWT2 had a very high Panda job failure rate for productions jobs in MWT2_UCORE – up to 80% over 12 hours.
  - A lot of investigation led to asking the ADC DPA group for help and Ivan Glushkov of that team eventually narrowed the problem down settings for queue and asked Tadashi Maneo for assistance.
  - Tadashi quite quickly identified that a MWT2 test queue for HPC work was set to a different default RSE in AGIS. Since this queue was defined as a production queue this caused panda to get confused and the output files to be lost (not great for debugging).
  - Correcting the setting fixed the failures immediately.
- The likely cause is on the next slide.

MWT2 Job Failures

- Later Tadashi explained the likely underlying cause of the job failures
  - Tadashi explained that while the HPC test queue was configured incorrectly during summer 2019, that order that that Panda looks at the AGIS info is essentially random and that presumably the misconfiguration was not reached.
  - However during the MWT2 dCache upgrade something must have been changed in AGIS and the new order that the queue info was randomly different. Presumably the misconfiguration was seen and caused the trouble.

Tier 2 Plans

- These are my initial impressions and of course we should discuss them:
  - Reduce the variations between the Tier 2 sites.
    - Clearly there are limits here because we want to take advantage of whatever special, local features can help us.
  - Look into centralized management of anything that can be.
    - The variations from site to site coupled with a low number of sysadmins leads to risk of failures.
    - If we could do a more central job of managing with less diversity from site to site, the risk could be reduced.
  - Improve the monitoring and check it systematically.
    - There are times when the monitoring shows issues and is ignored.
    - There are so many monitors that we end up confused about which are working, which have problems, and how to compare the systems.
    - Some more coherent approach with the right plots to look at could lead to effective monitoring and problem detection.

Bonus: Communication

- Our communication particularly with the ADC/EU seems weak.
  - We tend not to respond the the US Cloud mailing list.
  - We tend to process GGUS tickets slowly.
  - This leads to unhappiness on their side.
    - Some of this seems justified.
  - It also seems to lead to them jumping into fruitful discussions and trying to get other open problems solved.
    - In at least one case I was involved in, the problem ultimately was to be at CERN but the ADC team was sure it was at our sites. Since we don't always respond it seemed like they missed some rather obvious hints that problem was not a site problem.
- So I would really like to take a step back and see if we can improve the cordiality of our communication.