USATLAS Tier-2 Status and Plans

Shawn McKee / University of Michigan

USATLAS Facilities Meeting

March 13, 2019

Overview

- Our USATLAS Tier-2s have a number of upgrades, configuration changes and capabilities to provide
  - CentOS7 (SL7) migration, IPv6, Ucore migration, perfSONAR updates, Edge-services nodes, SCRATCHDISK reconfig
  - Planning for upgrades/purchase for our batch systems, storage, networking and compute (including GPU)
  - Testing, prototyping and preparation for both Run-3 and HL-LHC in conjunction with IRIS-HEP, HSF and WLCG
- We want to track our progress and understand site plans to ensure we stay ahead of ATLAS needs and benefit from our shared experience.

Planning and Status

- Because of the recent request for our sites to provide details for ATLAS ADC, we have useful information in the Google Doc that Xin setup.
- In the next few slides, I will try to summarize the various components for our facility

Batch Systems

- All of our sites use the HTCondor-CE
  - BNL, AGLT2 and MWT2 use HTCondor for batch
  - NET2 uses SGE
  - OU uses Slurm while UTA uses Torque
- No changes planned over the next two years except that NET2 will be incorporating PaaS (“hardware cloud”) access to MGHPCC for dynamic, demand driven growth

Networking

- Three parts to networking: Capacity, Connectivity, IPv6
- Capacity
  - BNL: 2x100G, upgrade to 3x100G this year
  - AGLT2: 2x40G UM to Chicago and MSU, MSU-Chicago 10G with planned upgrade to 100G if NSF proposal is funded.
  - MWT2: 2x40G(UC), IU/UIUC?, no plans for upgrades
  - NET2: 100G with move to shared 2x100G (NESE migration)
  - SWT2: 2x10G OU/OSCER, 40G UTA, OSCER -> 100G this year
- IPv6
  - BNL, AGLT2 dual-stacked for services/storage. WN in 2019
  - MWT2: UC this month?, IU/UIUC by April
  - NET2: Not done, IPv6 when NESE migration complete
  - SWT2: OU only done internal, UTA not done, OU external IPv6 by summer
- LHCONE connection: **Done** at BNL, AGLT2, NET2, MWT2, SWT2 (and SLAC!)

Storage

- dCache deployed as primary site storage: BNL(45 PB) , AGLT2(7 PB), MWT2
  - BNL has tape system, AGLT2 and MWT2 have Ceph
- GPFS at NET2 (6 PB)
  - Will be migrating to Ceph (NESE) over time
- Xrootd based storage at SWT2 (6 PB)
  - OU has 700 TB, UTA has 5.3 PB

Storage caching planned for USATLAS sites using XCache (via SLATE edge server nodes)

SCRATCHDISK sized at BNL, AGLT2, NET2, SWT2

- **Need info from MWT2**

Services and OS

- UCORE/Harvester setup at NET2 and SWT2-OU
  - BNL testing as of Feb 8, AGLT2 enabling this week
- CentOS7 (SL7) migration
  - **BNL** has SL7 hosts but runs SLC6 via container. Going native April
  - **AGLT2** has all worker nodes with SL7 and singularity. Servers will be upgraded during 2019
  - **MWT2** has most hosts SL7 with a few servers running SL6
  - **NET2** is planning to upgrade; need to migrate LSM->rucio-mover
  - **SWT2** is split. OU has CentOS7+singularity. UTA upgrade in May

Containerization, Edge Node Status

- All sites have singularity installed
- AGLT2 and MWT2 have purchased and deployed at least one edge node
  - The other sites have received an example quote (Dell R740)
  - NET2, SWT2 indicated they are planning to purchase
  - BNL didn't indicate plans to purchase

Dell Portal and Standardized Configs

- We have been working to revive the Dell portal and made good progress at the ANL facility meeting, agreeing that locked down configs could work for all of us.
  - SFP+ (10G) NICs
  - iDRAC9 express on compute, iDRAC9 enterprise on servers
  - Spinning disk (4x1.2TB 10k SAS) on compute instead of SSD because of cost
  - Smaller core-count, larger memory / HT-core config (56 vs 80; 3.4 G/HT-core vs 2.4G/HT-core). $/HS06 more but systems should perform better; RAM headroom
  - Form factor either R440 or C6420 for compute
  - Servers (storage headnode or virtualization node) R740
  - Storage MD1400 or new dense Dell systems (SAS or iSCSI option)
  - Networking (Top-of-Rack(N or S series) and Aggregation (Z9100/S4248-ON)
- Next few slides show configs we will send to Dell if everyone agrees

Dell R440 Compute Node

- Trusted Platform Module 2.0
- 2.5" Chassis with up to 8 Hot Plug Hard Drives
- Intel® Xeon® Gold 6132 2.6G,14C/28T,10.4GT/s, 19M Cache,Turbo,HT (140W) DDR4-2666 **Qty 2**
- Riser Config 1, 1 x 16 FH
- 16GB RDIMM, 2666MT/s, Dual Rank **(Qty. 12**)
- PERC H330 RAID Controller, LP
- 800GB SSD SATA Mix Use 6Gbps 512n 2.5in Hot-plug Drive, Hawk-M4E,3 DWPD,4380 TBW (**Qty. 2**)
- **or**
- 1.2 TB 10K SAS 12 Gbps disks (**Qty. 4** or **Qty 2**?!)
- iDRAC9 Express
- Intel X710 Dual Port 10Gb Direct Attach, SFP+, Converged Network Adapter
- No Internal Optical Drive
- Single, Hot-plug Power Supply (1+0),550W
- No Bezel for x4 and x8 chassis
- ReadyRails Sliding Rails Without Cable Management Arm

Add in Quick Sync 2 (At-the-box mgmt) [350-BBKQ] / 5104112 ? What about power cords?

Dell C6420 Compute Node

- 210-ALBP PowerEdge C6420
- 461-AADZ No Trusted Platform Module
- 321-BCPD PE C6420 Motherboard
- 340-BLEY PowerEdge C6420/C6400 Shipping
- 338-BLMN Intel Xeon Gold 6148 2.4G, 20C/40T, 10.4GT/s , 27M Cache, Turbo, HT (150W) DDR4-2666 **Qty 2**
- 370-ADNU 2666MT/s RDIMMs
- 370-AAIP Performance Optimized
- 405-AAND PERC H730P Controller Card
- 540-BBWM PERC Bridge Card for C6420
- 575-BBNY MiniPerc Bracket for C6420
- 780-BCEK C14A, PERC H730P Controller, C6420 1U Direct BP, NO RAID, Supports up to 6x2.5in Hard Drives
- 470-ACJQ MiniPerc Cable for C6420
- 370-ADRU M.2 Blank Riser for C6420
- 385-BBKX iDRAC9,Enterprise
- 379-BCQV iDRAC Group Manager, Enabled
- 540-BBWN PCIe Riser for C6420
- 800-BBDM UEFI BIOS Boot Mode with GPT Partition
- 813-8553 Dell Hardware Limited Warranty Plus On Site Service
- 813-8556 Basic Hardware Services: Business Hours (5X10) Next Business Day On Site Hardware Warranty Repair 5 Year
- 370-ADND 16GB RDIMM, 2666MT/s, Dual Rank **Qty 12**
- 400-ASHI 1.2TB 10K RPM SAS 12Gbps 512n 2.5in Hot-plug Hard Drive (**Qty 2 or Qty 4!?)**

  OR
- 800GB SSD SATA Mix Use 6Gbps 512n 2.5in Hot-plug Drive, Hawk-M4E,3 DWPD,4380 TBW (**Qty. 2**)
- 540-BBWC Intel X710 Dual Port 10Gb, SFP+, OCP Mezzanine card
- 818-BBGR OCP Mezzanine Bracket for C6420

Dell R240 perfSONAR node

Trusted Platform Module (TPM)
Trusted Platform Module 2.0
Chassis Configuration
3.5" Chassis with up to 4 Hot Plug Hard Drives
Processor
Intel® Xeon® E-2146G 3.5GHz, 12M cache, 6C/12T, turbo (80W)
Memory DIMM Type and Speed
2666MT/s UDIMMs
Memory Capacity
(2) 16GB 2666MT/s DDR4 ECC UDIMM
RAID/Internal Storage Controllers
PERC H330 RAID Controller, Adapter, Full Height
Hard Drives
1.2TB 10K RPM SAS 12Gbps 512n 2.5in Hot-plug Hard Drive, 3.5in
Additional Network Cards
On-Board Broadcom 5720 Dual Port 1Gb LOM
Embedded Systems Management
iDrac9, Express

Internal Optical Drive
DVD +/-RW, SATA, Internal for Hot Plug Chassis
Rack Rails
1U/2U 2/4-Post Static Rails
Bezel
No Bezel
Power Cords
C13 to C14, PDU Style, 12 AMP, 2 Feet (.6m) Power Cord, North America
Power Supply
Single, Cabled Power Supply, 250W
Password
iDRAC,Factory Generated Password
PCIe Riser
PCIe Riser with Fan with up to 1 LP, x8 PCIe + 1 FH/HL, x16 PCIe Slots
Hardware Support Services
3 Years, Basic Hardware Warranty Repair: 5x10 HW-Only, 5x10 Next Business Day Onsite
Deployment Services
No Installation

Web price $2451. Waiting for Dell quote
This system missing 10G+ NIC options
3-year warranty probably OK. (update ~3yrs)

Dell E740 Edge-Server

- Trusted Platform Module 2.0
- Chassis with Up to 12 x 3.5 Hard Drives for 2CPU Configuration 321-BCPU
- 2xIntel Xeon Silver 4110 2.1G, 8C/16T, 9.6GT/s , 11M Cache, Turbo, HT
- PERC H730P RAID Controller, 2GB NV Cache, Adapter, Low Profile 405-AAOE
- BOSS controller card + with 2 M.2 Sticks 240G (RAID 1),FH 403-BBPT
- iDRAC9,Enterprise 385-BBKT
- iDRAC,Factory Generated Password 379-BCSF
- Riser Config 3, 2 x8, 3 x16 slots 330-BBHE
- Intel X520 DP 10Gb DA/SFP+, + I350 DP 1Gb Ethernet, Network
- Dual, Hot-plug, Redundant Power Supply (1+1), 750W 450-ADWS
- No Power Cord 450-AAGG
- No Bezel 350-BBBW
- UEFI BIOS Boot Mode with GPT Partition 800-BBDM
- ReadyRails Sliding Rails Without Cable Management Arm 770-BBBQ
- Dell Hardware Limited Warranty Plus On-Site ServiceBasic Hardware Services: Bus Hours (5x10) NBD
- On-Site Hardware Warranty Repair, 5 Years
- 12x16GB RDIMM, 2666MT/s, Dual Rank
- 12x12TB 7.2K RPM NLSAS 12Gbps 512e 3.5in Hot-plug Hard Drive 400-AWIP

Cost $13,696

Discussion

- Need to select either 80 (HT)core vs 56 (HT)core option
  - 80 (HT)Core drove MWT2 to select 2x960GB SSD vs 2x1.2T 10K SAS
  - AGLT2 uses C6420 with 2x1.2T 10K SAS
    - Wenjing tested today 100 runs of IOSTAT spaced by 3 secs
      - Nodes fully loaded already with both grid and BOINC jobs
      - C6420(2x1.2TB SAS disks) **IOwait=0.01** CPU=99.89 R=20.2MB/s W=23.9
      - BL(2x150G SAS) **IOwait=0.90** CPU=95.44 R=2.4MB/s W=3.1 MB/s
      - All other WN **IOwait=0.02** CPU=98.7 R=8.0MB/s W=8.9 MB/s
  - If we get 80 HTcore config we need 2xSSD or 4xSAS versus 56 HTcore config which is working with 2xSAS (cheaper)
- Need to converge ASAP to get suggested configs to Dell