MWT2 Site Report

Fred Luehring

US ATLAS Tier2 Meeting @ IU

June 22, 2007

MWT2 Overview

- Production resources partitioned at two sites, roughly equally
  - MWT2_IU
  - MWT2_UC

- Leveraged resources, remnants of iVDGL and other projects
  - IU_ATLAS_TIER2 (Retiring soon)
    - Replacemnt =
      - 112 IBM HS21 blades in two racks
      - each blade has two quad-core Intel 5335 Xeons
      - 2.0GHz, 8GB RAM per blade
      - 2 x gigE
      - 36GB SAS drive per blade
  - UC_ATLAS_MWT2
    - Plan to upgrade to SLC4 ~next few months
  - UC_Teraport
- Operating model
  - Using weekly shift model
  - Admins work across sites, on all resources above

MWT2 Hardware Profile

- Phase I (operational)
  - Processors
    - 31 Dual-CPU, dual-core AMD Opteron 285 (2.6 GHz): 216k SI2K
    - 124 batch slots
  - Storage
    - 80 GB local scratch
    - 5 x 500GB Hardware RAID5 / node (~1.9TB/node)
    - ~68 TB dCache-based
  - Edge servers for dCache, DQ2, GUMS, NFS (OSG, /home), GridFTP, mgt services
  - Gigabit switching Cisco 6509/UC, Force10/IU; 10G NICs (for four hosts, 2 at each site)
  - Cluster management
    - Cyclades terminal servers for serial console access with logging
    - Network accessible power distribution units for remote power management
- Phase II (operational)
  - Additional 44 nodes (307k SI2K), compute only, 176 additional batch slots
  - Additional scratch disk for all worker nodes (500 GB); retrofit Phase I
  - Delivered mid-January

Planned Acquisitions

- Phase III (underway)
  - Fill Phase II nodes with dCache disk pools
  - 6 x 750 GB drives (~3.4 TB RAID 5)
  - Adding ~150TB
  - After Phase III, installed capacity:
    - CPU: 523K SI2K
    - Storage: ~216 TB
- Phase IV (late summer)
  - Based on operational experience with a ~200 TB scale dCache system we will evaluate technology options
  - If we continue with the same architecture
    - Increase CPU and storage capacity with a ~$135K purchase
    - Roughly 140k SI2K, 50 TB
  - After Phase IV, installed capacity:
    - CPU: ~660K SI2K
    - Storage: ~266 TB

Summary: MWT2 Capacity Analysis

- Assumptions
  - SI2K cost based on Opteron 285 FY06 purchase, CPU doubling every 24 months, and server retirement after 3 years
  - Distributed dCache disk model (disk on compute node), 24 month doubling (proposal assumed 18 months and did not account for server retirement)

523K SI2K, 68 TB

MWT2 Capacity Evolution Study

CPU Proposed Tier2 Facility	2005	2006	2007	2008	2009	2010	2011
CPU (Project) (SI2K)	97538	244178	464636	698537	1050185		
CPU (DL, UC) (SI2K)	212992	212992	212992	212992	0		
CPU (DL, IU) (SI2K)	54400	54400	217600	217600	217600		
CPU Total Dedicated (Proposed)	364930	511570	895228	1129129	1267785		
CPU Updated Tier2 Facility							
CPU (Project) (SI2K)	0	460416	601706	959098	1004109	1577602	2231066
CPU (DL, UC) (SI2K)	212992	212992	212992	212992	0	0	0
CPU (DL, IU) (SI2K)	54400	54400	217600	217600	217600	0	0
CPU Total Dedicated (Updated)	267392	727808	1032298	1389690	1221709	1577602	2231066
Disk Proposed Tier2 Facility							
Disk (Project) (TB)	51	132	261	465	790		
Disk (DL, UC)	15	15	15	15	15		
Disk (DL, IU)	10	10	50	50	50		
Disk Total, Dedicated (Proposed) (TB)	76	157	326	530	855		
Disk Updated Tier2 Facility							
Disk (Project) (TB)	0	61	222	350	470	565	800
Disk (DL, UC)	15	15	15	0	0	0	0
Disk (DL, IU)	10	10	50	50	50	0	0
Disk Total, Dedicated (Updated) (TB)	25	86	287	400	520	565	800

660K SI2K, 268 TB late summer

Software Profile

- Platform: SLC4
  - Linux 2.6.9-42.0.3.EL.cernsmp #1 SMP i686 athlon i386 GNU/Linux
  - RAID partitions formatted with ext3
- Torque/Maui
  - Simple: one queue with a 120 hour wall-time limit
- Cluster management tools from ACT
  - Image “cloner” and “beo_exec” command script
- dCache 1.7.0 full bundle (server, client, postgres, dcap)
- OSG 0.6.0
- GUMS
  - Configured to authorize only usatlas1, usatlas2, usatlas3, usatlas4, mis, ivdgl, osg, sam, samgrid
- ATLAS
  - Releases: 11.0.3, 11.0.42, 11.0.5, 12.0.3, 12.0.31, 12.0.4, 12.0.5, 12.0.6, 12.3.0, 13.0.10 and the corresponding versions of Kit Validation
  - DQ2 0.2.12 site services stopped; DQ2 0.3 update underway

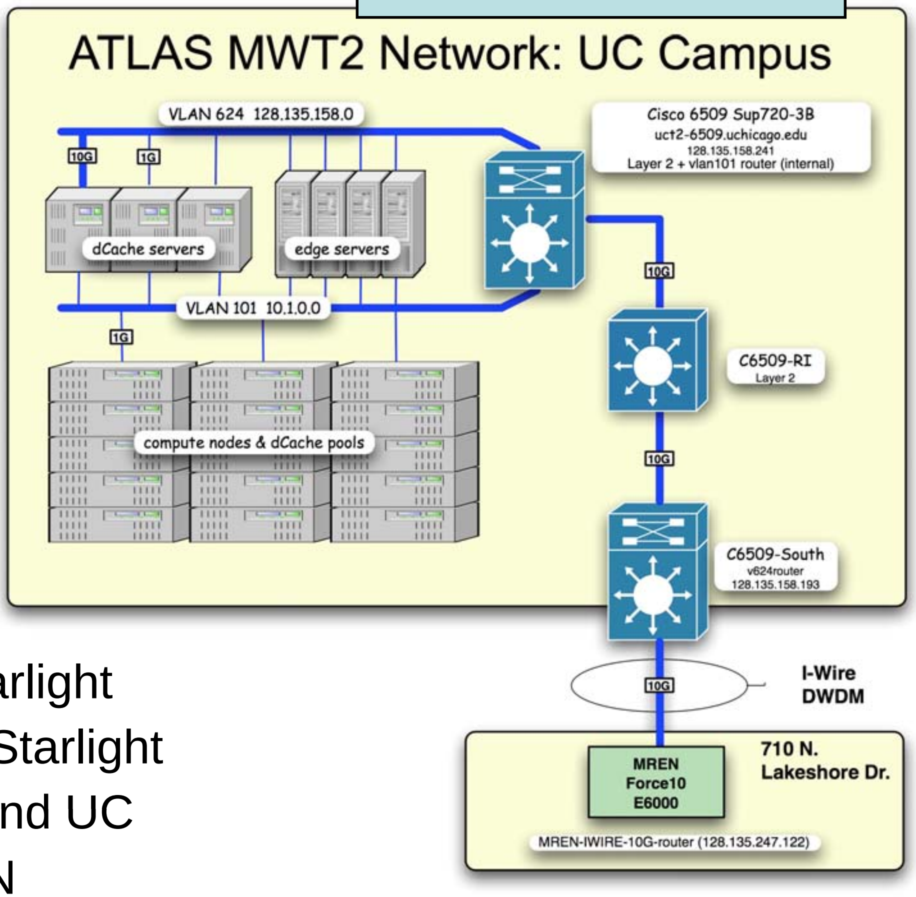
Monitoring

- Nagios
  - Host monitoring: ping, ssh, dcache gridftp door, etc.
  - Email notices and web interface
- Ganglia
  - Performance monitoring: network, cpu load, I/O, temp., etc.
  - Web interface, graphs
- Osiris
  - Security monitoring
    - Takes snapshot of client
    - Compares on a schedule for changes to binaries, open/missing ports
- Power Distribution
  - Event monitoring an history, overload warnings

Site Architecture

- Dual role for worker nodes
  - Four processing cores
  - dCache R/W pool (1.9 TB)
  - 500 GB scratch
- Edge servers
  - 3 dCache services nodes
    - dc1: gridFTP, dcap, SRM
    - dc2: pnfs server, Postgres
    - dc3: admin, gridFTP, dcap
  - DQ2, GUMS, OSG
  - Interactive logins
- Network
  - UC: Cisco, w/10G iWIRE to Starlight
  - IU: Force10, w/10G iLIGHT to Starlight
  - VLAN configured between IU and UC
  - 10G connectivity to BNL, CERN
- Other services deployed:
  - Torque/Maui, Ganglia, Nagios

IU site has nodes on public network and Force10 switch



Creation Date: 10/20/06
Contact Information: R. Gardner

Resource Allocation Policies: CPU

- Current MWT2 PBS job queue policy
  - authenticate only usatlas1, usatlas2, usatlas3, usatlas4, mis, ivdgl, osg, sam, samgrid
  - Single queue with 120 hour wall time limit
  - Priority weights: usatlas1 95%, sum of all others is 5%
- Customization for _IU:
  - Authenticate samgrid user (D0) with 18% priority weight and high water mark of 40 jobs (contributed nodes)
    - dzero has 9 additional quad-core nodes in MWT2_IU
- Planned changes (implement RAC policies):
  - weights x for usatlas3: eg. ~20%
  - y for sum of OSG VOs
  - (100-x-y) for usatlas1,2
    - with x, y given by the RAC