VM Technologies (and others) for Site and Service Resiliency

Shawn McKee/University of Michigan

OSG AHM - Lincoln, Nebraska

March 19th, 2012

Service and Site Resiliency Motivations

- I'm sure every site providing any service is concerned with resiliency: they want their services to remain useable
- For AGLT2 we have a large distributed LHC Tier-2 for ATLAS that spans two locations: UM/Ann Arbor and MSU/East Lansing with roughly 50% of the storage and compute at each site (Total ~4500 job slots and 2.2 PB of dCache)
- However almost all of the critical site services are at the UM location! **What happens if the UM site is down?**
- We are very interested in overall resiliency of the Tier-2 and would like to continue operations even if one of our sites is down/offline for an extended period of time. (Either one of our two sites represents a significant amount of resources)
- **First GOAL:** Allow AGLT2 to continue to run with one site up

Virtualization of Service Nodes

- Our current USATLAS grid infrastructure requires a **number** of services to operate:
  - Grid gatekeepers and authentication/authorization services
  - Job scheduler
  - Grid storage and distributed file-systems
  - Various meta-data services
- These services need to be robust and highly-available
- Can Virtualization technologies be used to support some of these services?
- Depending upon the virtualization system this can help:
  - Backing up critical services
  - Increasing availability, reliability via enterprise features
  - Easing management

Virtualization Technology

- In the last two years, we have tested and deployed **virtualization technologies** for our services
- AGLT2 uses VMware Enterprise V5.0 Plus which runs:
  - LFC, 4 Squid servers, USATLAS Gatekeeper, **OSG Gatekeeper**, Condor headnode, ROCKS headnodes (dev/prod), Kerb/AFS nodes, **central syslog-ng host**, muon calibration splitter, Oracle DB, all our AFS file servers (storage in iSCSI), dCache headnode
- "HA" mode can ensure services run even if a physical server fails. Backup is easy as well.
- Can "live-migrate" VMs between 3 servers or migrate VM storage to alternate back-end iSCSI storage servers.
- Downside is initial/on-going costs for VMware.

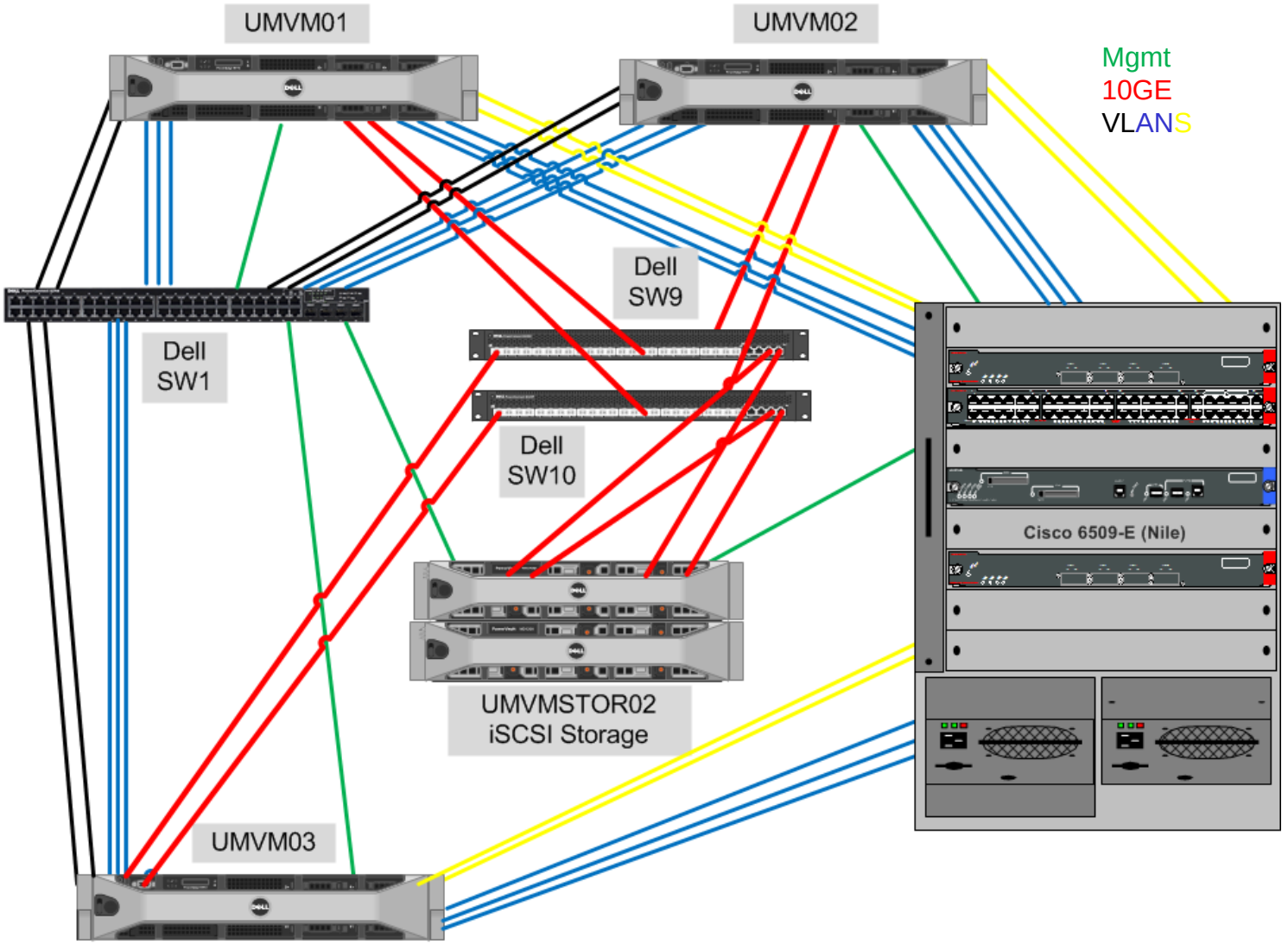
Virtualization Installation at MSU

- MSU interim virtualization purchase (final delivery **today**):
  - Two Dell R710s, 192GB, 2 X5675 processors, 2 X520 dual 10GE
  - MD3220 (8 x 6Gbps ports, SSD 150G, 23 x 300G 15K SAS; 7TB)
  - MD1220 (24 x 900G 10K SAS disks; 21.6 TB raw)
- VMware Ent Plus licenses with Production support via Merit
- Sized to hold all needed services from UM site (has 33% more memory, disks selected to provide IOPS and space)
- Limitation: SAS-connected disks are not visible from UM site
- MSU VMware nodes can copy from UM. Plan to setup replication process for needed VMs from UM to MSU

Virtualization Considerations

- Before deploying any VM technology you should plan out the underlying hardware layer to ensure a robust basis for whatever gets installed
- Multiple VM "servers" (to run VM images) are important for redundancy and load balancing
- Multiple, shared back-end storage is important to provide VM storage options to support advanced features
  - iSCSI or clustered filesystems recommended
- Scale hardware to planned deployment
  - Sufficient CPU/memory to allow failover (N-1 svrs)
  - Sufficient storage for range of services (IOPS+space)
  - Sufficient physical network interfaces for # of VMs
- Design for no-single point-of-failure to the extent possible

Example: AGLT2_UM VMware Hardware



iSCSI for VM Storage

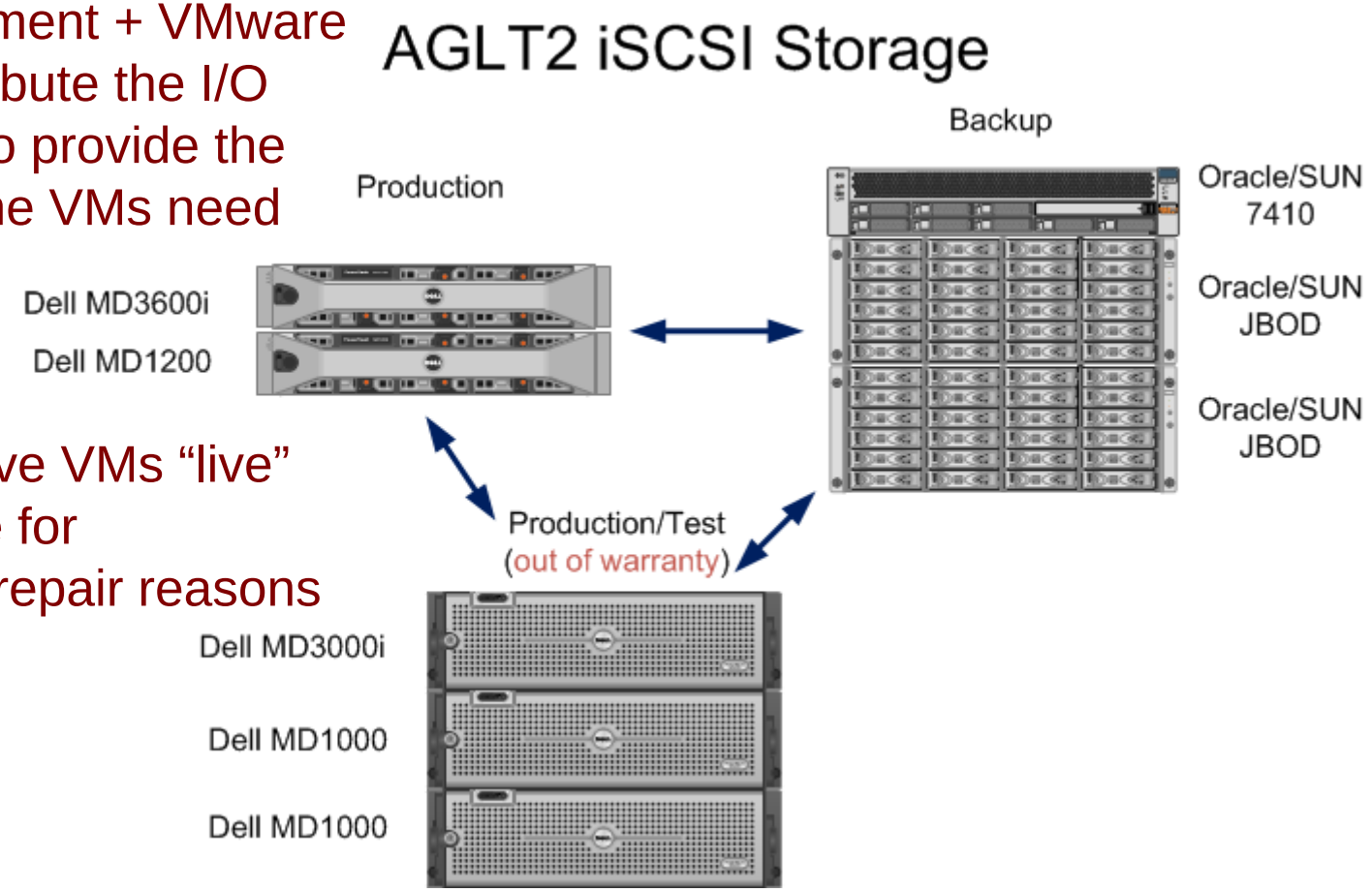
- **iSCSI** has been around for a while. Lots of nice features
  - Relative to fiber channel it can be inexpensive
  - Allows **multi-host access** to the same storage
  - Typically supports features like snap-shots and cloning for LUNs. Easy to migrate via LAN/WAN
  - With 10GE and hardware iSCSI offloading, performance can exceed fiber channel
- Roll-your-own with Open-iSCSI or Openfiler
- Allows advanced features (like “Storage vMotion” in VMware)

iSCSI Hardware Options

- Can buy iSCSI appliances. ..lots of choices
  - Dell offers MD32xxi/ D36xxi. Relatively inexpensive via LHC matrix. Can configure 12 disks up to 3TB each or 24 disks up to 1TB each. Can add MD12xx shelves (up to 4).
  - Oracle(Sun) 74xx (Amber Road) systems provide higher-end capabilities and nice interface and multiple access options.
  - Very nice systems with higher-end features from other vendors (**$$$**) (Isilon, HP, EMC VNX5300)
- Can build your own direct attached storage and enable iSCSI (e.g., OpenFiler)
- Must balance features vs cost as usual

iSCSI for “backend” Live Storage Migration

- This set of equipment + VMware allows us to distribute the I/O load as needed to provide the level of service the VMs need
- We can also move VMs “live” between storage for maintenance or repair reasons



Interim purchase as MSU will only allow MSU nodes to replicate VM images from UM site. This is OK for our primary goal as long as we maintain a timely replica

Service Multi-Site Resiliency Options

- What about site level resiliency? Various options by service:
  - DB can use "internal" replication methods: Postgresql – streaming/hot-standby, Oracle – Streams/Clustering, MySQL – Replication
    - Already using Oracle Streams to CERN and Postgresql 9.0 hot-standby for dCache DBs at AGLT2_UM. Must extend to include MSU once up.
  - Services without state: AFS, Kerberos, LFC, ROCKs head nodes, web servers, etc. use virtualization features like HA which makes sure an instance is always running
    - AFS is a special case: need both DB/auth and file-server components
    - Plan to add AFS VM nodes at MSU and leverage AFS volume replication
  - Grid services with state: Condor, Gatekeepers: **still looking for a good solution**. Currently both virtualized at AGLT2. Plan: replication of VM image to keep "cold" copy at remote site. Startup requires network changes to instantiate at secondary site. Host-cert issues…

Multi-Site Virtualization Challenges

- VMware (or other virtualization technologies) can provide high availability and ease management but **require that the same networks are present at two sites for transparent operations**
- If the networks available at each site differ, we need some additional work:
  - Readdress VMs before bringing up at the remote site
  - "Move" the network from the down site to the up site
  - Use network aliases and DNS to allow different back-end addresses to serve the same DNS name (LVS is one example)
  - Share a "service" subnet between two sites (allows hot-migration)
  - VMware has "Site Recovery Manager" which automates some of this and we intend to test it out

Grid Storage Considerations

- In addition to services, we need to worry about storage.
- AGLT2 uses dCache to host its 2.2 Petabytes of storage.
- Storage is split between UM and MSU. We cannot afford to replicate this volume of data. How can the site continue to operate with 50% of its storage offline?
  - New files can continue to be written to the remaining online nodes
  - Files already written may be either offline or online
  - AGLT2 uses a site-caching configuration which ends up replicating "hot" (in use) files at each site (see next slide)
  - Federated Xrootd configuration might be able to augment this by transparently getting any missing files from other sites in the federation (might even use dCache "cache" to store them)

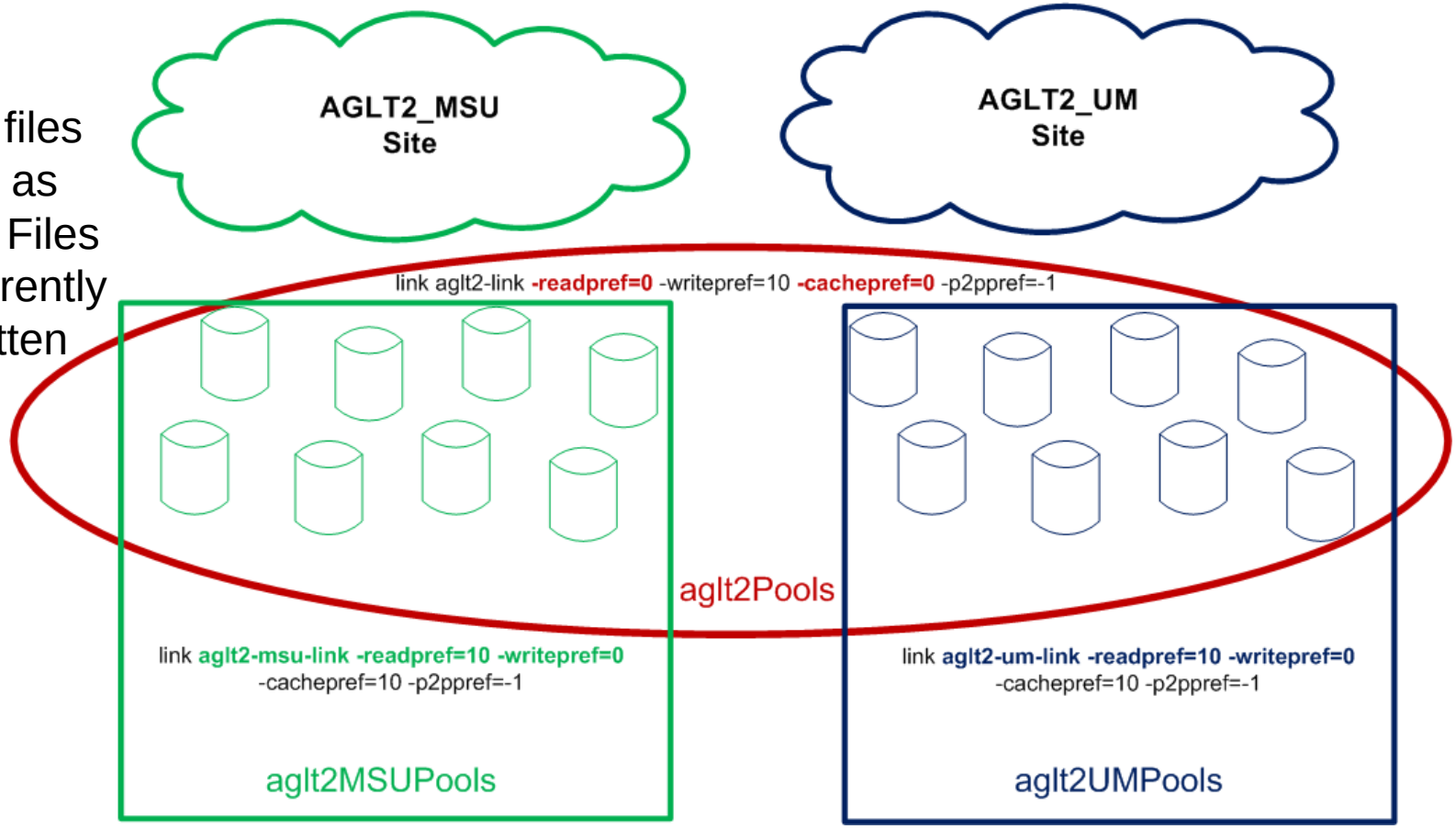
dCache Inter-site Caching

All reads **MUST** come from a local pool

AGLT2 dCache New Configuration
Site Locality Aware

Missing files staged in from remote site

Staged files marked as replica. Files transparently overwritten (LRU)



A little more on Federated Xrootd Use

- The previous slide shows how dCache can be made to transparently cache popular files by forcing a site to only "read" from local dCache pools.
- Missing files are retrieved from the remote site via pool-to-pool copying (remote site is treated just like a tape system)
- When the remote-site pools are offline we would like to fall-back to retrieve missing files via our Xrootd Federation.
- The plan is to patch dCache's pool-to-pool copy process to allow it to use xrdcp from the federation to get missing files when dCache pools containing the file are offline.
- Xrootd files end up cached at the site just as if they came from the other sites dCache pool during normal operation

Future Considerations

- Our primary goal is to allow AGLT2 to continue to function if one of the two physical sites is down for an extended period.
- Would like to have the ability to hot-migrate services between sites: requires robust replication and network changes.
- Some services will need to be replicated to be running at both sites: DB replication, DNS aliases with LVS, etc.
  - Will need well-documented procedures to promote secondary to primary for each such setup
- Software solutions may not be sufficiently robust or may have too large of a latency to meet our desired service level but we plan to try them out
- We think iSCSI **hardware** may be able to provide needed replication and storage services for site resiliency. Cost is an issue.

Summary

- Most of our focus has been on “Service Virtualization”
  - Reliability (multiple hosts and storage systems support live migration)
  - Ease of management
    - Easy to backup/clone/update
    - Maintenance do-able without downtime for VMs (most cases)
  - Works well within one site
- Goal is “site resiliency” allowing AGLT2 to function with either UM or MSU site being down
  - Currently recovering from loss of UM site would take weeks; our goal is to reduce that to hours.
- dCache and Federated Xrootd may allow storage to support single site operations
- Hardware and plans in place to meet this goal soon

Questions / Discussion?

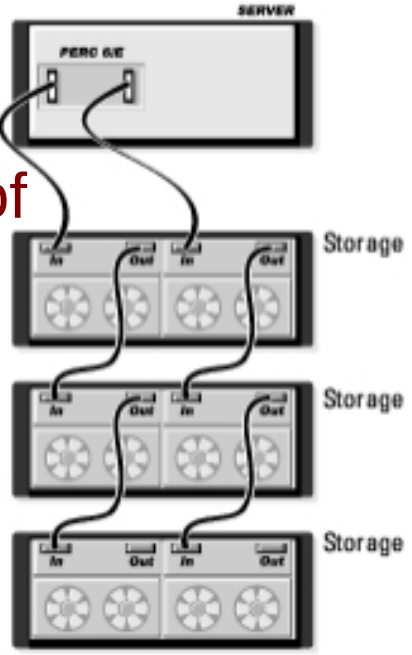
ADDITIONAL RESILIENCY CONSIDERATIONS

Storage Connectivity

- Increase robustness for storage by providing resiliency at various levels:
  - Network: Bonding (e.g. 802.3ad)
  - Raid/SCSI redundant cabling, multipathing
  - iSCSI (with redundant connections)
  - Disk choices: SATA, SAS, SSD ?
  - **Single-Host resiliency**: redundant power, mirrored memory, RAID OS disks, multipath controllers
  - **Clustered/failover** storage servers
  - Multiple copies, multiple write locations

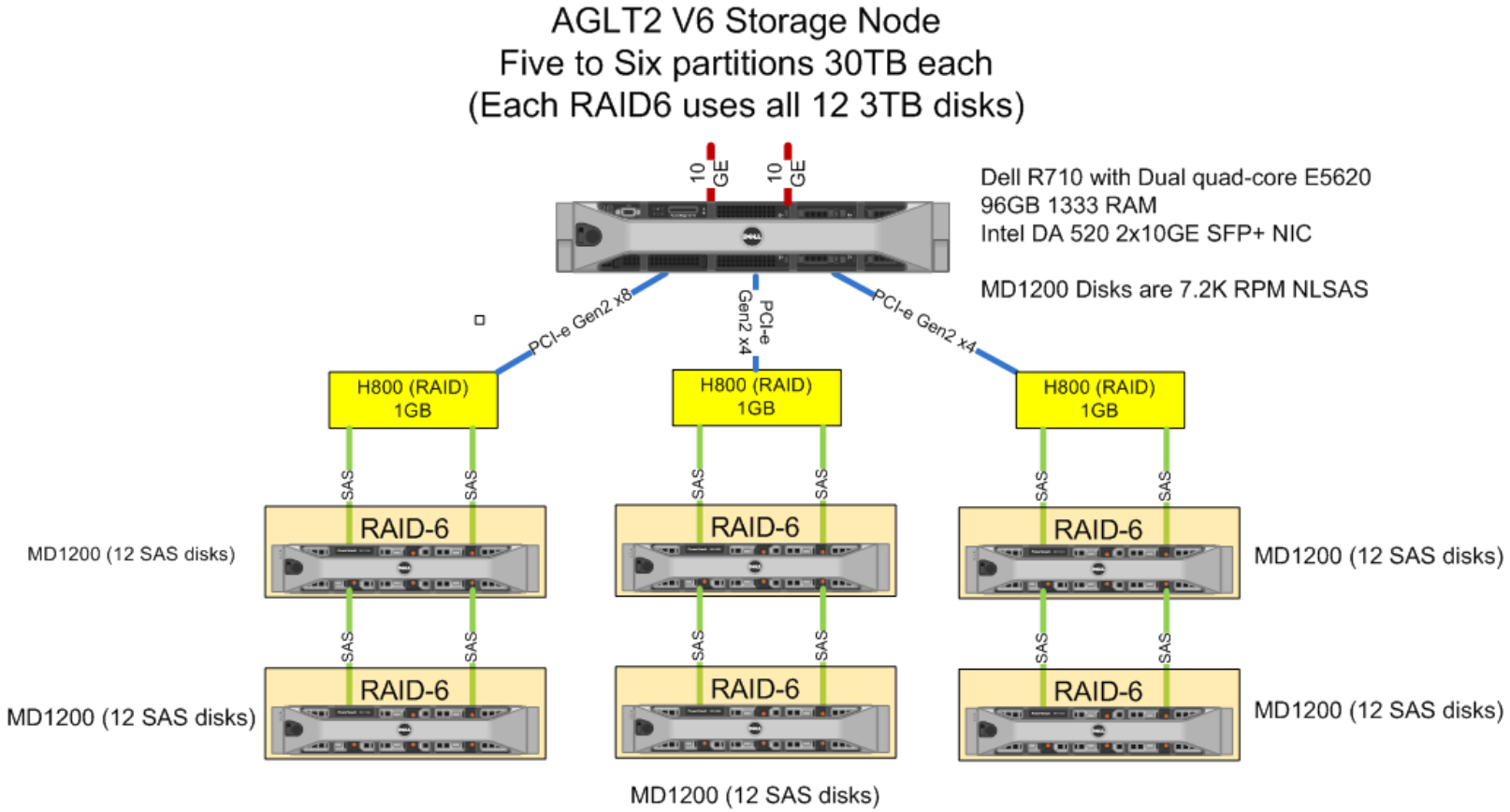
Redundant Cabling Using Dell MD1200s

- Firmware for Dell RAID controllers allows redundant cabling of MD1200s
- MD1200 can have two EMMs, each capable of accessing all disks
- An H800 has two SAS channels
- Can now cable each channel to an EMM on a shelf. Connection shows one logical link (similar to “bond” in networking): Performance and Reliability



Storage Example: Inexpensive, Robust and Powerful

Dell has special LHC pricing available. US ATLAS Tier-2s have found the following configuration both powerful and inexpensive Individual storage nodes have exceeded 750MB/sec on the WAN



Site Disk Choices

- Currently a number of disks choices:
  - SATA – Inexpensive, varying RPMs, good BW
  - SAS - Higher quality, faster interface, more $
  - SSDs – Expensive, great IOPs, interface varies
- Reliability can be very good for most choices. NL-SAS (SATA/SAS hybrid) very robust. Range of throughputs and IOPS.
- SSDs have IOPS in the 10K+ region versus fast SAS disks at ~200. SSD I/O bandwidths usually x2-3 rotating disk

SSDs for Targeted Apps

- SSDs make good sense if **IOPS** are critical.
  - DB applications are a prime example
  - Server hot-spots (NFS locations, other)
  - Example: Intel SSDs at AGLT2 decreased some operation times from 4 hours to 45 minutes
- New SSDs very **robust** compared to previous generations…capable of 8+ Petabytes of writes; 5 year lifetimes (check *endurance* figures)
- Lower power; bandwidths to **550 MB/sec**
- Choice of interface: SATA, SAS, bus-attached. Seagate/Hitachi/Toshiba/OCZ have 6 Gbps/SAS
- See https://hep.pa.msu.edu/twiki/bin/view/AGLT2/TestingSSD

Power Issues

- Power issues can frequently be the cause of service loss in our infrastructure
- Redundant power-supplies connected to independent circuits can minimize loss due to circuit or supply failure (Verify one circuit can support the required load!!)
- UPS systems can bridge brown-outs or short-duration loses and protect equipment from power fluctuations
- Generators provide longer-term bridging

General Considerations (1/2)

- Lots of things can impact both reliability and performance
- At the hardware level:
  - Check for driver updates
  - Examine firmware/bios versions (newer isn't always better BTW)
- Software versions…fixes for problems?
- **Test changes** – Do they do what you thought? What else did they break? ☺

General Considerations (2/2)

- Sometimes the additional complexity to add “resiliency” actually **decreases** availability compared to doing nothing!
- Having **test equipment** to experiment with is critical for trying new options
- Often you need to trade-off cost vs performance vs reliability (pick 2 ☺)
- **Documentation, issue tracking and version control systems** are your friends!

“Robustness” Summary

- There are many components and complex interactions possible in our sites.
- We need to understand our options (frequently site specific) to help create robust, high-performing infrastructures
- Hardware choices can give resiliency and performance. Need to tweak for each site.
- Reminder: **test changes** to make sure they actually do what you want (and not something you don’t want!)

Questions / Discussion?