Locality Aware dCache & Discussion on Sharing Storage

USATLAS Facilities Meeting

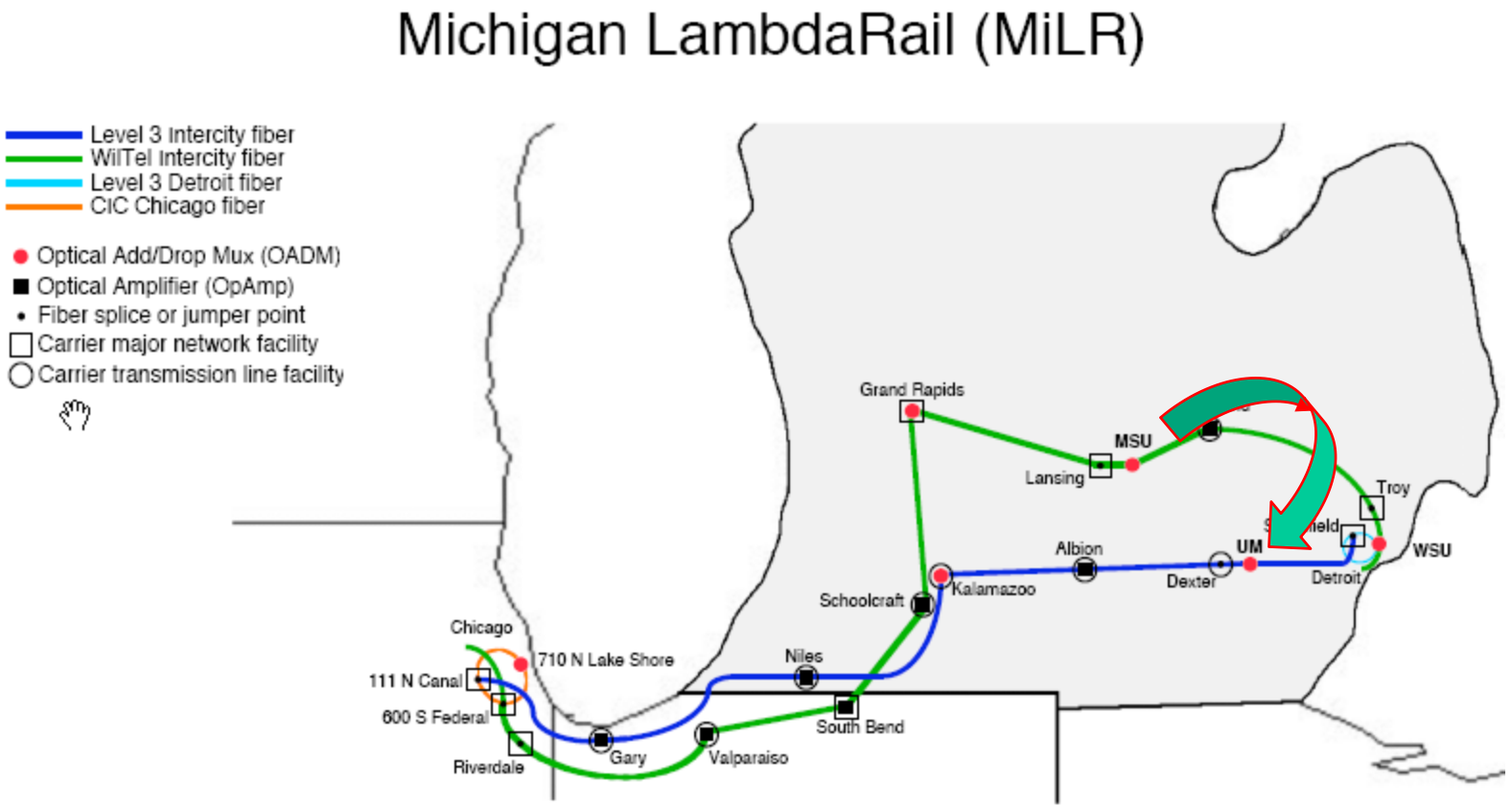
SMU

October 12, 2011

dCache and Locality-Awareness

- For AGLT2 we have seen significant growth in the amount of storage and compute-power at each site.
- We currently have a single 10GE connection used for inter-site transfers and it is becoming strained.
  - Given 50% of resources at each site, 50% of file access will be on the intersite link
- Cost for an additional link is **$30K/year** + equipment
- Could try traffic engineering to utilize the other direction on the MiLR triangle BUT this would compete with WAN use
- This got us thinking: we have seen pCache works OK for a single node but the hit rate is relatively small. What if we could "cache" our dCache at each site and have dCache use "local" files? **We don't want to halve our storage though!**

10GE Protected Network for ATLAS



- We have two "/23" networks for the AGL-Tier2 but a single domain: aglt2.org
  - Currently 3 10GE paths to Chicago for AGLT2. Another 10GE DCN path also exists (BW limited)
- Our AGLT2 network has three 10GE wavelengths on MiLR in a "triangle"
  - Loss of any of the 3 waves doesn't impact connectivity for both sites. VRF to utilize 4th wave at UM

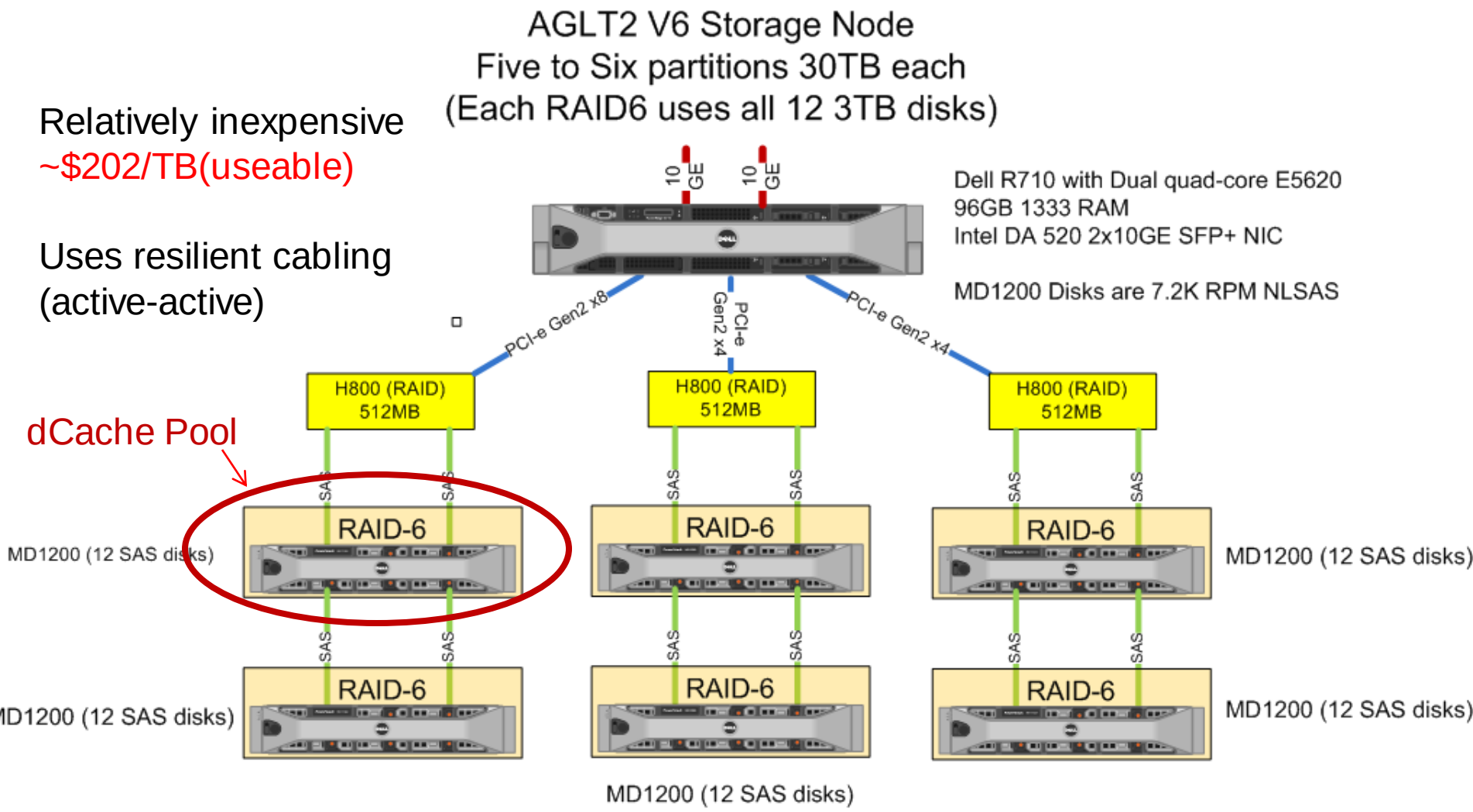
dCache Storage Organization/Access

- dCache organizes storage by pools. This is the "unit" of storage and maps to a device/partition on a node
- Pools may be grouped into pool-groups to organize storage. Typically this is done to group storage by owners/users.
- dCache defines links to control how pool groups are accessed. You can view links as a prioritized set of rules defining which pool-groups you might use to determine where you will read or write in dCache.

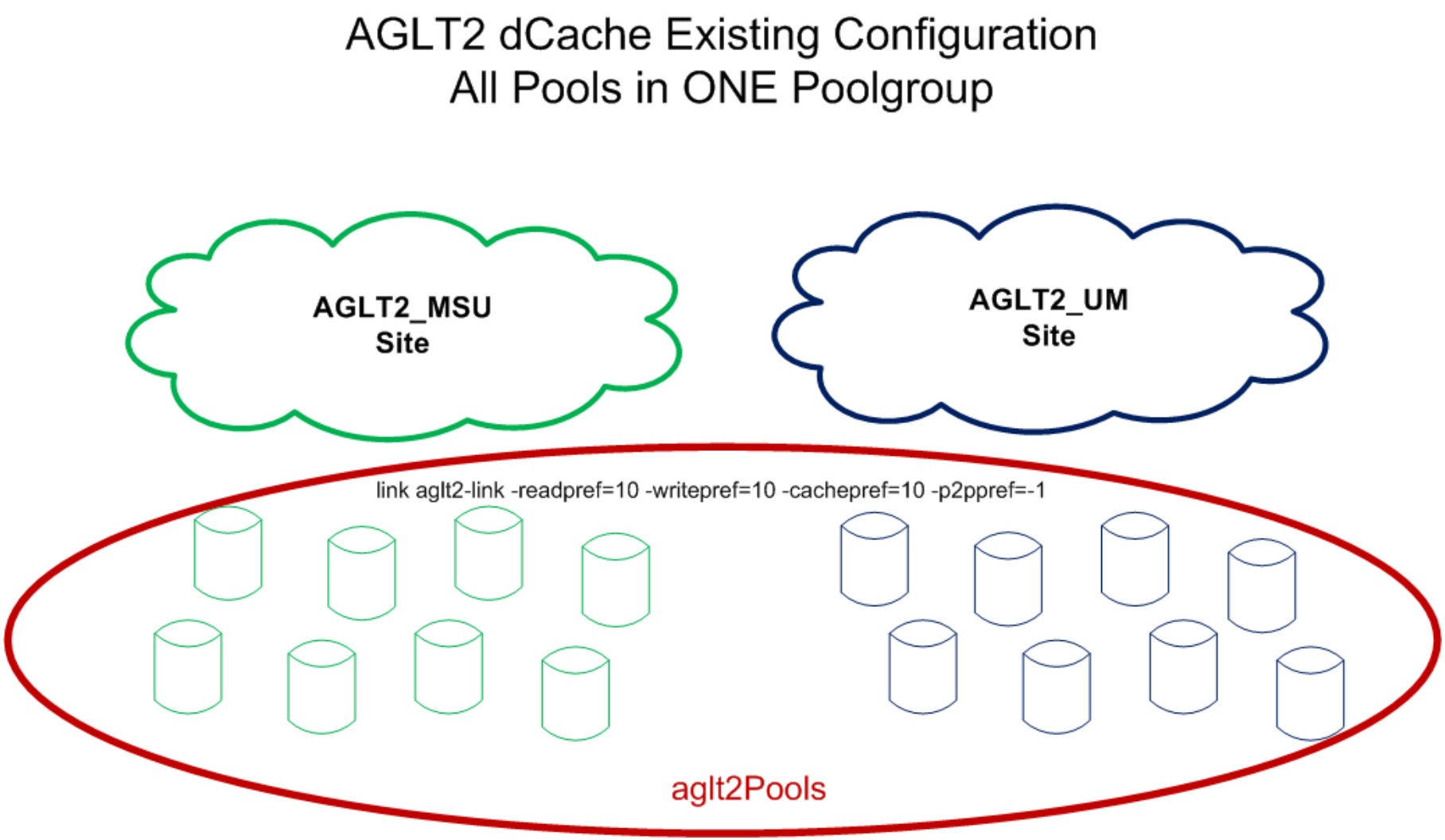
dCache Pool at AGLT2

Relatively inexpensive
~$202/TB(useable)

Uses resilient cabling
(active-active)



Original AGLT2 dCache Config



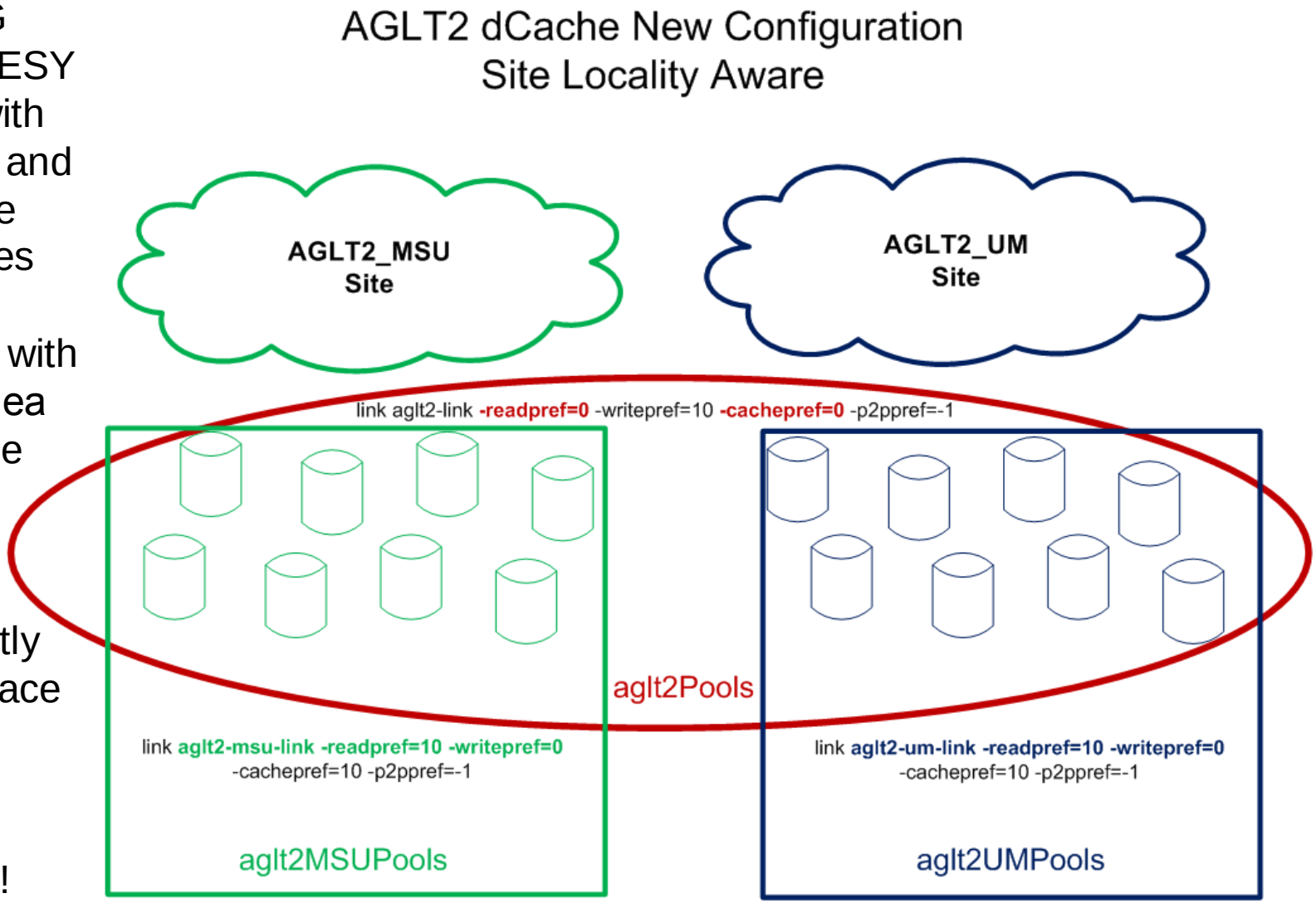
dCache and Locality-Awareness

At the WLCG meeting in DESY we worked with Gerd, Tigran and Paul on some dCache issues

We came up with a 'caching' idea that has some locality awareness

It transparently uses pool space for cached replicas

Working well!



Configuration Discussion

- The dCache configuration on the previous slide relies on the fact that dCache allows pools to participate in more than one pool-group.
- All pools at AGLT2 are members of aglt2Pools and a "local" group like aglt2UMPools or aglt2MSUPools
- Reads **must** come from a local pool. Writes can go to any pool.
- Read requests for files that are not on a local pool cause a Pool-to-Pool (P2P) transfer from another pool-group which has the requested file. The remote pool-group is treated analogously to "tape"
- The resulting local copy is a "cached" replica. Cached replicas don't show up as using space in dCache and can use "unused" space.
- Cached replicas are cleaned via LRU algorithms when space is needed

Storage Sharing:Can This Be Extended?

- The dCache configuration at AGLT2 allows us to better optimize our storage use for ATLAS while minimizing the required inter-site network traffic.
- Could something like this be generalized to allow cross-Tier-2 transparent sharing of files?
  - An idea might be to treat other dCache instances as “tape” somehow
  - Having a single dCache instance spanning two sites would be another way but with obvious issues in implementation/use.
  - Could dCache utilize Federated Xrootd as a “tape” source?
  - Other ideas for how to share storage? Between dCache sites? More generally (Xrootd <-> dCache)?

Regional Sharing?

- Rob and I have discussed options to better inter-connect MWT2 and AGLT2.
  - We are "close" in our WAN peering locations
  - Invest in a regional "meet-me" switch/device in Chicago?
    - Advantage is inter-site traffic is isolated from other traffic (uses the meet-me switch)
  - LHCONE could also be a means of better managing inter-site connectivity
- One advantage of regional sharing is a reduced need for storage…can rely upon other sites for some of your storage.
- Ideas/discussion about possibilities here?

Discussion…

Discussion? Options?