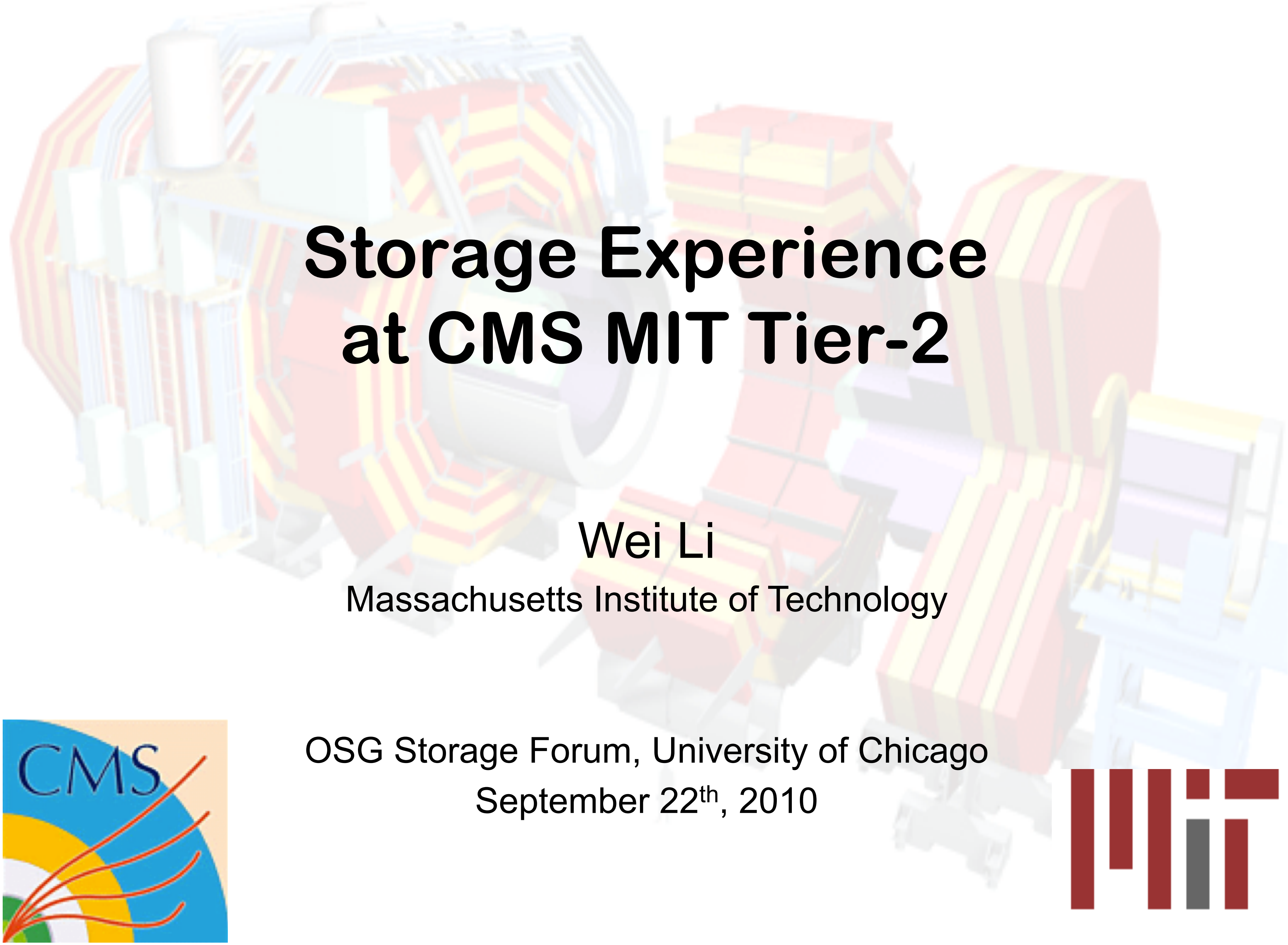


Storage Experience at CMS MIT Tier-2

Wei Li

Massachusetts Institute of Technology

OSG Storage Forum, University of Chicago

September 22th, 2010

Location of the Tier-2 Facility

Bates Linear Accelerator Center:

- hosted by MIT, 22 mi from Cambridge, MA
- network managed by MIT IS&T
- racks, power, infrastructure managed by Bates
- overseen 7x24
- UPS backup for all servers (4 racks)
- 30 water cooled racks, rack – 40 U and 10 kW

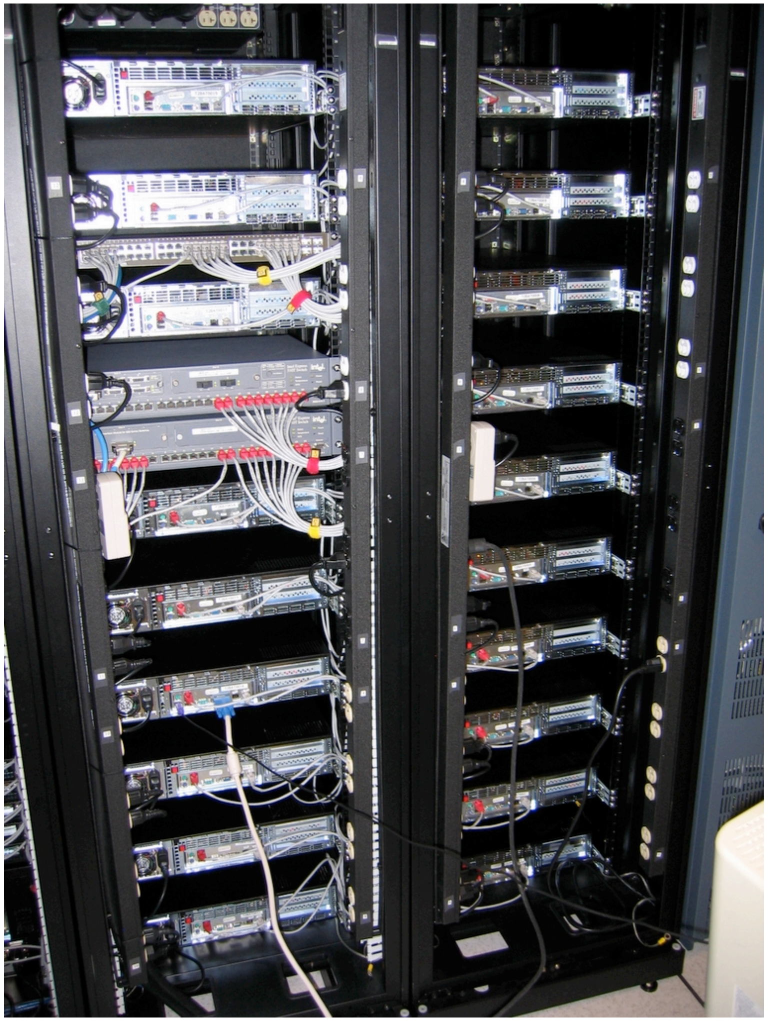
Storage Hardware

Storage system at CMS MIT Tier-2:

- Various types of machines, total of about 400 nodes
  - Intel vs AMD
  - Dell (PowerEdge 2950 and R710) vs Thinkmate
  - Different CPU and memory capability
- 2U with 6 - 8 disks on each node for storage
- dCache on individual node with RAID 5
  - Software vs Hardware raid

Storage Hardware

Water-cooled racks



Storage capability summary

Storage figures:

Tier2 resources	564 TB
Other (CDF/HI)	28 TB
Total	592 TB

Run with Raid 5, more raw storage space

Future purchase:
additional 240TB in the next couple of years

CPU capability summary

CPU power and batch slots:

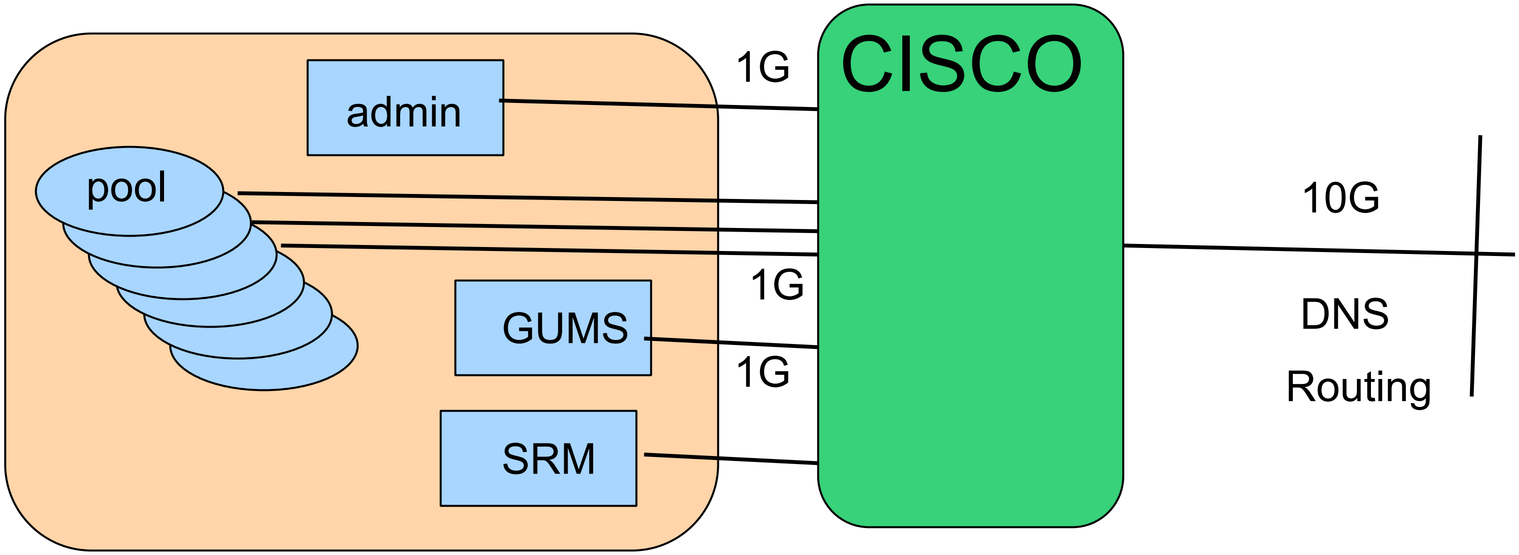
Tier2 resources	Batch slots CPU power	1922 14002 **HS06**
CDF	Batch slots CPU power	240 782 **HS06**
Total	Batch slots CPU power	2162 14784 **HS06**

Storage Software

Upgraded dCache from 1.9.3 to 1.9.5 yesterday:

- Admin server:
  - Intel 2.4GHz, 2x4-core, 24GB memory
  - services: lm, dCache, dir, admin, httpd, gPlazma, infoProvider, info, dcap
- Pnfs server:
  - Intel 2.4GHz, 2x4-core, 24GB memory
  - services: pnfs, utility
- Srm server:
  - Intel 2.4GHz, 1x4-core, 16GB memory
  - services: srm
- Pool nodes:
  - variety of compositions
  - services: pool, gridftp, gsidcap

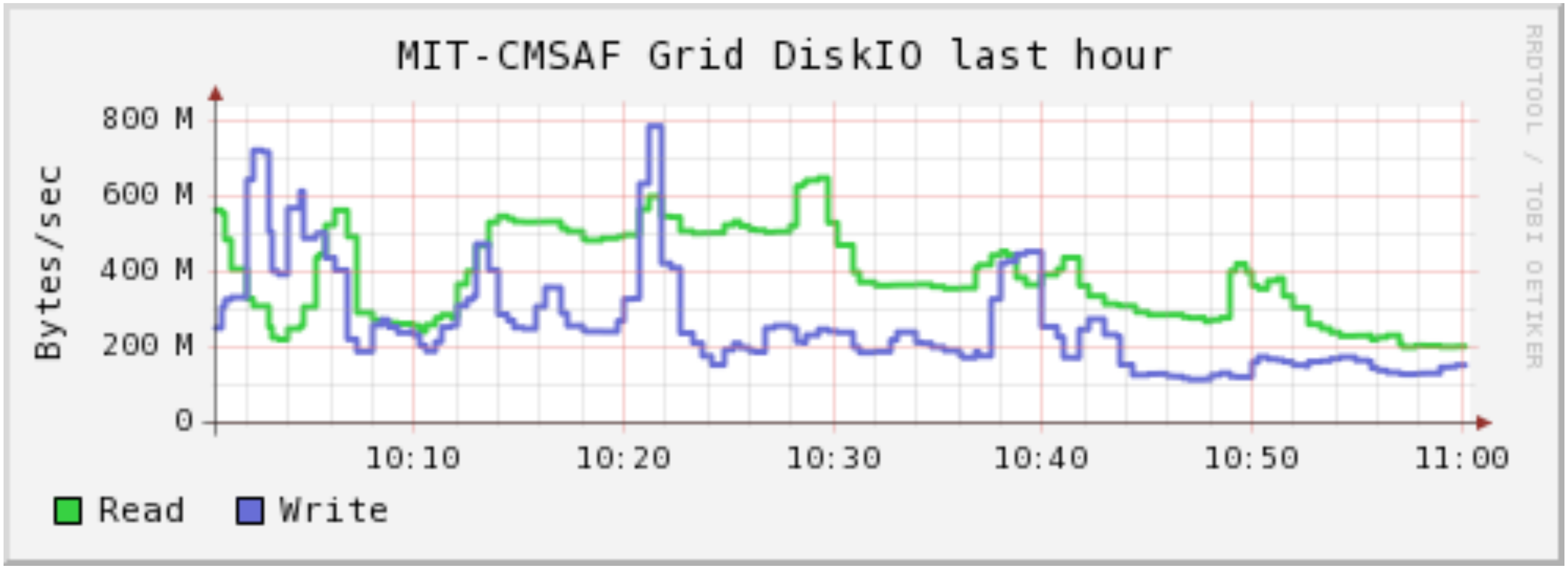
Network Configuration



- Leased for 5 years CISCO Nexus 7016
- Machines can talk at 1Gb through copper cable links
- dCache – each pool sits on public network and serves as gridFTP

Grid Disk I/O Performance

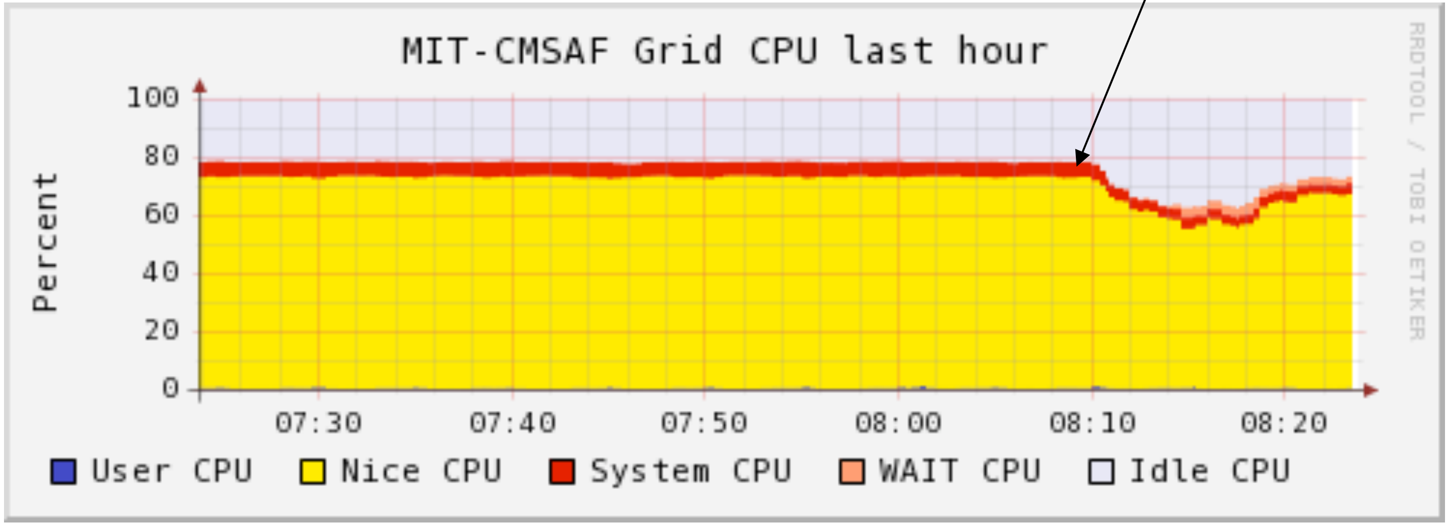
Typical disk I/O on the grid:



Major issues and challenges

Disk I/O issue with analysis jobs:

- Very high I/O load on the pool when many analysis jobs are running
- I/O time even longer than actual CPU time
- Cluster becomes very inefficient
- Current workaround:
  - ➢ deploy replica manager
  - ➢ uniformly distribution dataset over many pools



Space allocation for groups

- Officially hosting 4 groups in CMS for researches: QCD, JetMET, Higgs and Heavy Ion

- >10TB NFS based scratch area for local CMS users

- ~ 10TB NFS area and ~ 100 job slots for non-CMS users from neutrino, dark matter etc. experiments at MIT

Summary

- MIT Tier-2 is now in its final location. Plenty of room for expansion
- The center has been operating very well including the storage system
- Large dCache storage capability (600TB) with plan of 240TB additional space soon
- Provide strong supports to several analysis groups
- We are working to have higher performance and reliability of the storage system