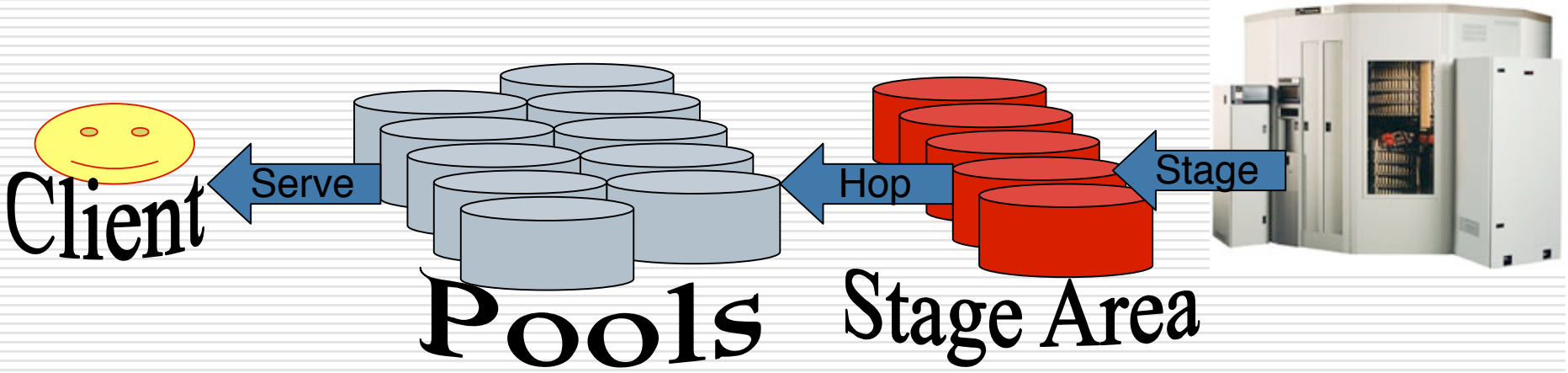
US CMS Tier 1 dCache

Timur Perelmutov,
Fermilab

dCache Workshop,
DESY, January 19, 2007

Stage Area

- Stage Area -11 nodes-10TB
  - Pools for staging files from tapes managed by dCache File Hopping
  - Pool-to-Pool copy to read pools
  - Limited resource tape drives running at full rate
  - Tape to Disk rate improved by 5 to 10 times !



Read/Write Area

- >100 nodes
- 700TB of Tape Backed pools
- Will Grow to 1.5 PETABYTE By September 2007
- One Gridftp server per node, used by SRM
- All pools allow both WAN and LAN access
- To improve reliability each pool has LAN and WAN queue
  - LAN Queue with 600 to 1800 active movers
  - WAN Queue with 5 to 15 active movers

Resilient Area On Worker nodes

- 2 Resilient Managers in the same dCache
- Worker Nodes Resilient Manager
  - PRECIOUS file
  - ~ 650 Worker nodes
  - More than 100TB
  - 3 copies of each file
- Precious Pools Resilient Manager
  - 55 TB of non-tape-backed PRECIOUS and RESILIENT pools for unmerged output
- Replica Monitoring is very useful

Central dCache Services

- Head Node functionaliy split between 11 nodes
  - 8 nodes run dCache Services
    1. Pnfs, PnfsManager, Dir
    2. LM, PoolManager, Broadcast,HsmController
    3. dCap Doors 0-2, gridFtp Door
    4. dCap Doors 3-6, gridFtp Door
    5. SRM,PinManager,SpaceManager
    6. dCap Doors 7-10, gridFtp Door
    7. Replic Manager, Replica Manager 2, gPlazma, Gridftp Door
    8. Billing, Httpd, InfoProvider, Statstics
  - 3 nodes run Monitoring, Controlling, scans

PoolManager Configuration

Name	Partition	Preferences				Unit Groups				Pool Groups	Pools
		Read	Write	Cache	P2p						
LFSOnly-link	LFSOnly-section	20	20	0	20	any-protocol	world-net	LFSOnly	-	LFSOnlyPools	-
Resilient-link	Resilient-section	20	20	0	20	any-protocol	Resilient	world-net	-	ResilientPools	-
T1DiskEval-link	T1DiskEval	20	20	0	20	any-protocol	T1DiskEval	world-net	-	T1DiskEvalPools	-
read-link	read-section	10	10	0	10	any-protocol	any-store	world-net	-	readPools	-
stage-link	stage-section	0	0	100	(0)	any-protocol	any-store	world-net	-	stagePools	-
write-link	write-section	10	10	0	10	any-protocol	any-store	world-net	-	writePools	-

Optimizations

- LoginBroker in dCache domain
- gPlazma is fully integrated with GUMS, will work with SAZ soon
- Check/Update companion info 4times/day to keep it correct (1-3 files per day out of sync)
- We do Not use info-provider, prefer SRM info+scripts
- WAN dcap is disabled by iptables
- No offsite admin access

Monitoring, Scans

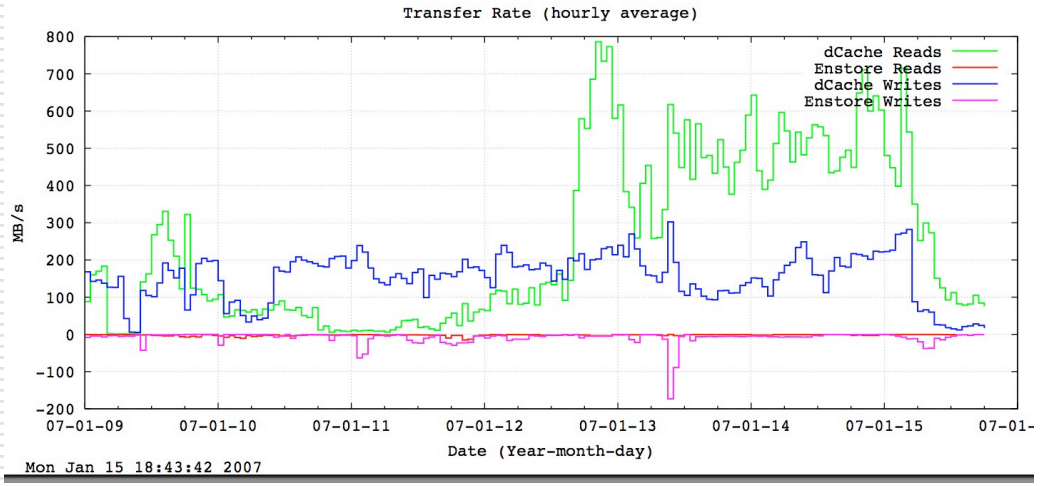
- Run Scans that check for
  - Files in precious pools are precious
  - 0-length files
  - E State files
  - Files not written to Enstore
  - Full crc scan every week
  - Many more scans, some minor errors are detected and fixed every day (to keep everything very clean)
- Monitoring
  - Network
  - Plots
  - SRMWatch
  - Replica Monitoring

Recommendations

- XEN works, but fails when stressed by high rates high volume services, not used
- Absolutely no pools or doors on Pnfs, Pool Manager, SRM
- To increase reliability
  - Identical services on separate nodes
  - Separate billing, http, statistics into separate JVMs
  - Use Pnfs Companion, Separate/Isolated Postgres Data areas/disks
- Deployment of each type of pool on many nodes key to success
- No non-dCache services on dCache nodes - great improvement in stability

Transfer Rates

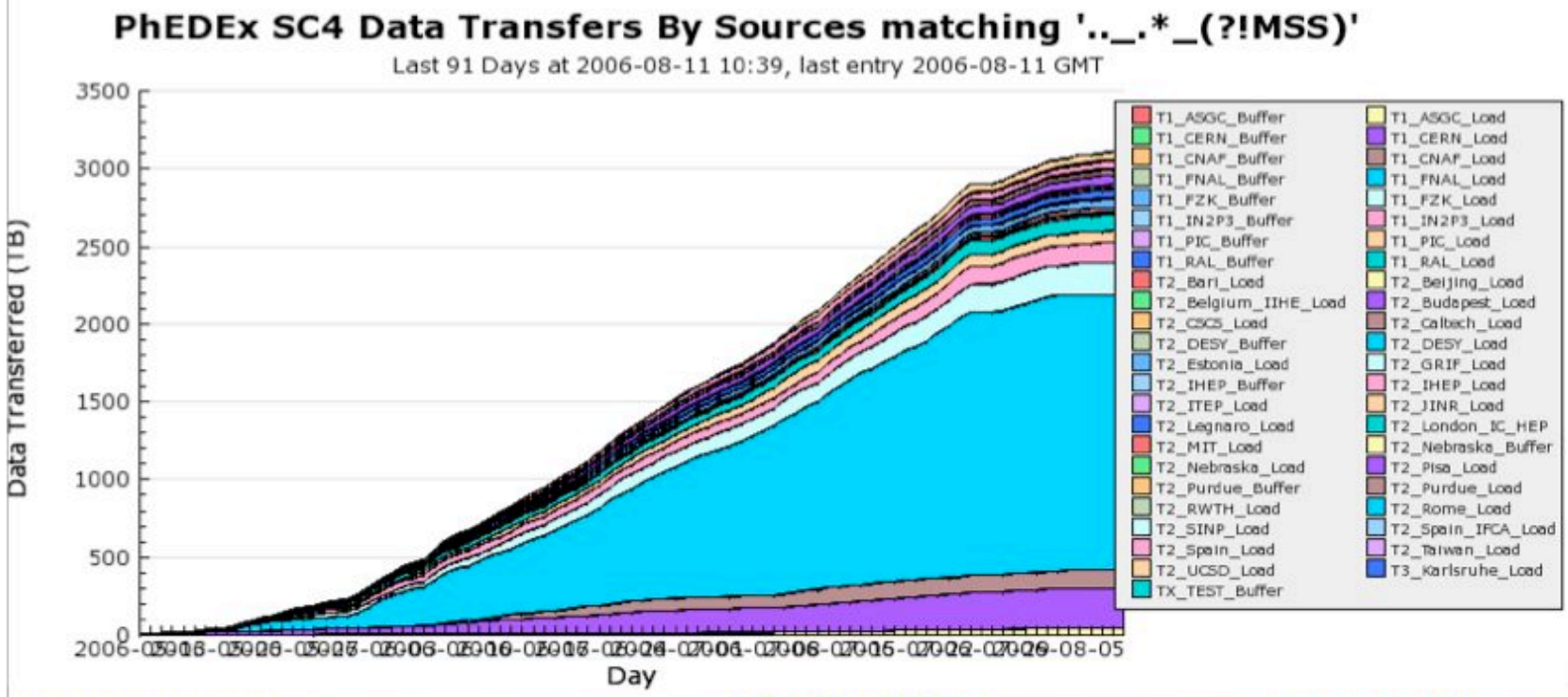
- Snapshot of the current Transfer rates
- Achieve needed CMS functionality at FNAL and target transfer rates without using SRM Storage Classes
- Almost no tape reads
- Peak Rate of 2.5 GB/s was achieved last July for a few hours



PhEDEx Transfers

Over the summer CMS was able to move over 1PB of data per month over the summer during the SC4 exercises

- CMS has averaged 1PB of data moved every month for the last three month
- FNAL and US Tier-2 centers have contributed significantly



Tier-2 Centers

Tier-2s also generally met the 50% milestone

- Sum of Tier-2 capacity is similar to the total Tier-1, as indicated in the model
- Tier-2 networking is in good shape

Site	CPU (kSI2K)	Disk (TB)	WAN (Gb/s)
Caltech	538	56	10
Florida	519	104	10
MIT	92	54	1
Nebraska	347	53	0.6
Purdue	743	184	10
UCSD	318	98	10
Wisconsin	547	119	10
TOTAL	3104	668	

References

- Organization: http://uscms.org
- US-CMS T1 dCache http://cmsdca.fnal.gov