

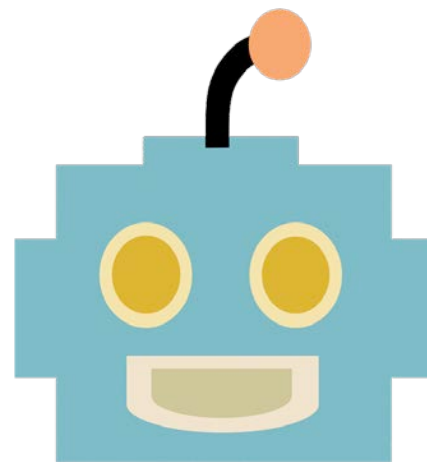


LUSTRE DEPLOYED 3 DIFFERENT WAYS

Scott Yockel, PhD | Senior HPC Team Lead | Harvard FAS Research Computing

Overview of Odyssey

- 150 racks spanning 3 data centers across 100 miles using 1 MW power
- 60k CPU cores, 100k+ GPU Cores
- 25 PB (Lustre, NFS, Isilon, Glustre)
- 10 miles of cat 5/6 + IB cabling
- 300k lines of Puppet code
- 300+ VMs
- 2015: 25.7 million jobs
240 million CPU hours

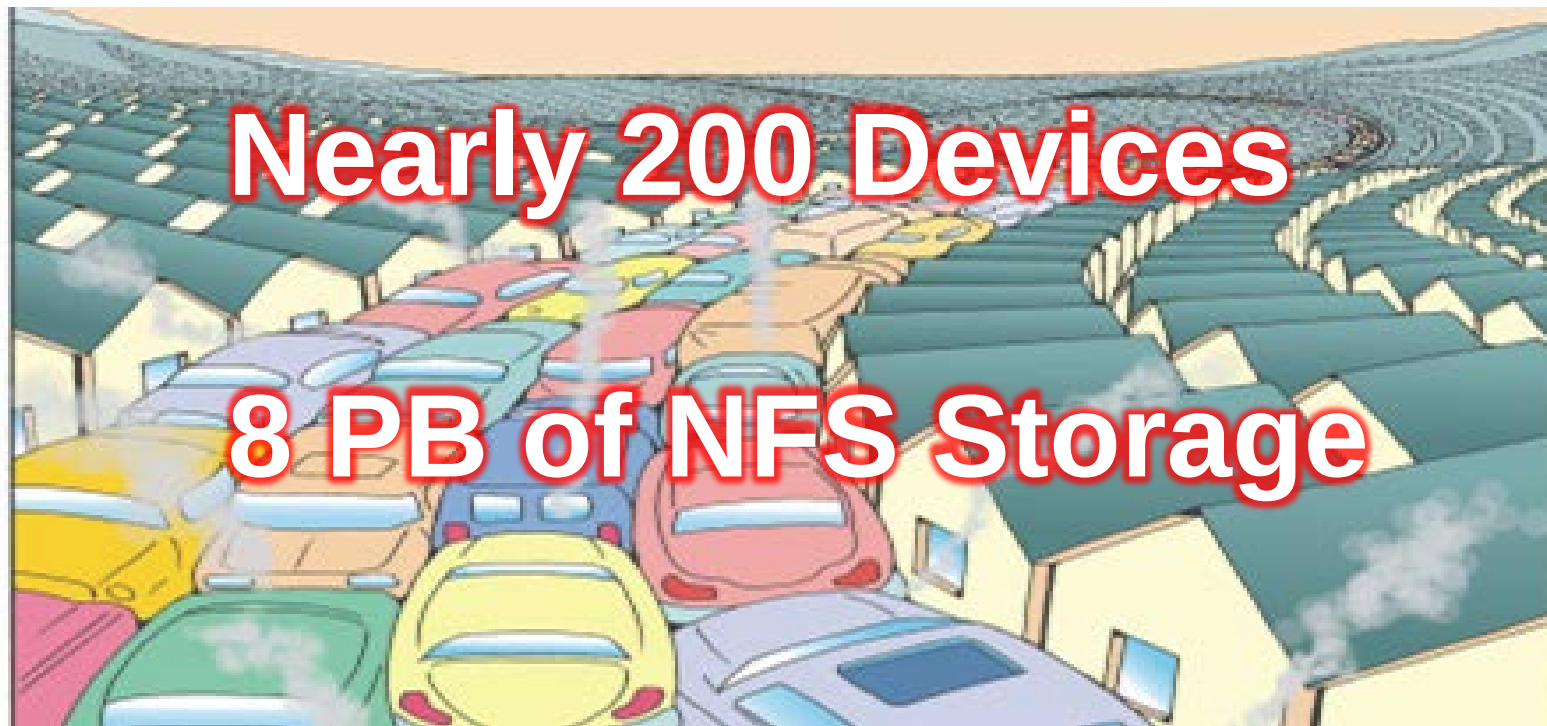




Outline

- Lustre Deployments in the past year
 - Two 5.6 PB IEEL (HOLYLFS, BOSLFS)
 - 64 TB all flash-based system (FLASHLFS)
 - 2.1 PB backup target (BULFS)
- Setup, Testing, Issues, ...

NFS Sprawl Anyone?

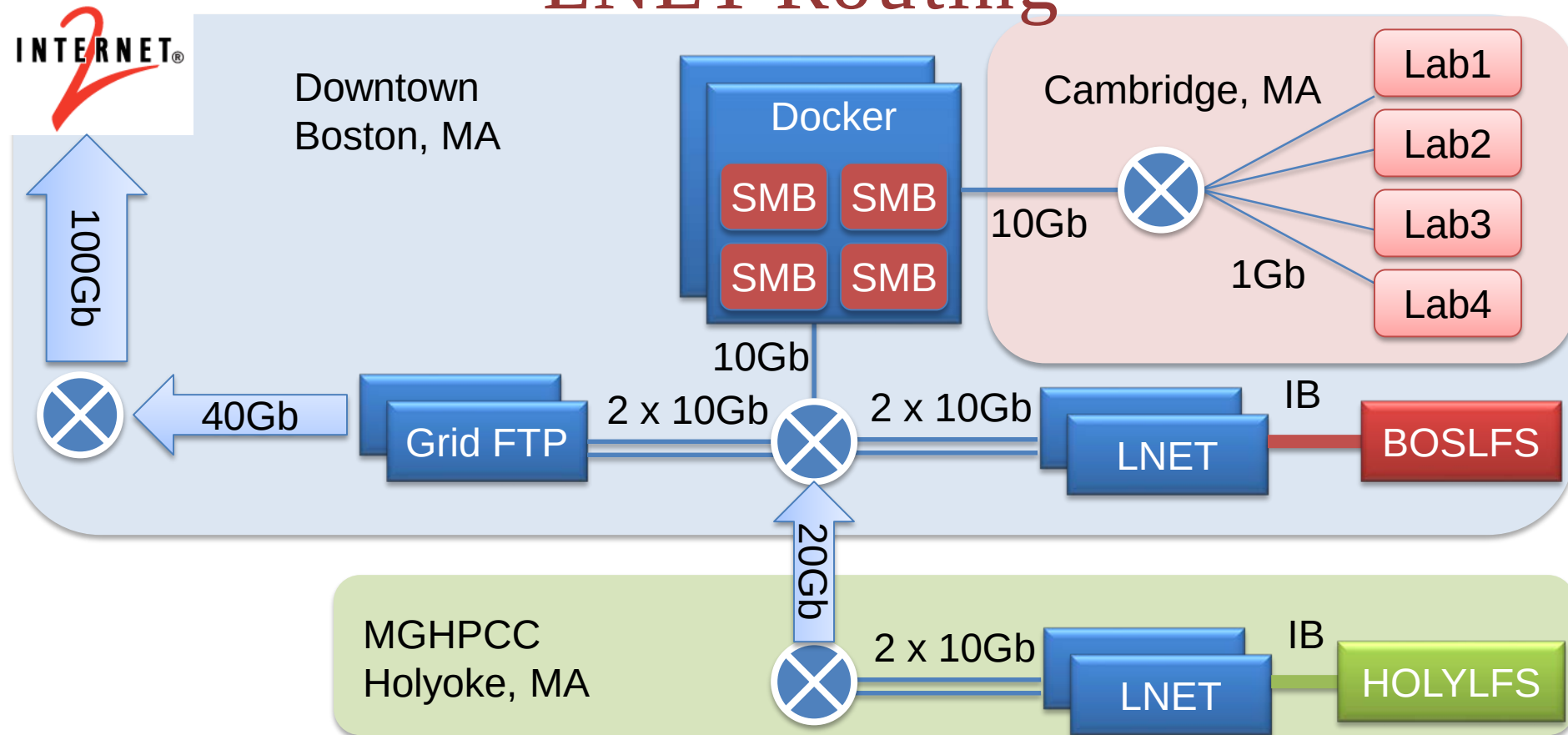




IEEL Deployment

- Two 5.6 PB systems
 - 2 x MDS (Dell R630s w/ MD3420)
 - 24 x 600GB 15K
 - two hot spares
 - 8 x OSS (Dell R630s w/ MD3460 + MD3060e)
 - 88 OSTs (8+2 RAID/6 with 6TB drives)
 - 1 entire OST as hot spare per HA set
- HOLYLFS – Persistent research data
- BOSLFS – Instrument data processing

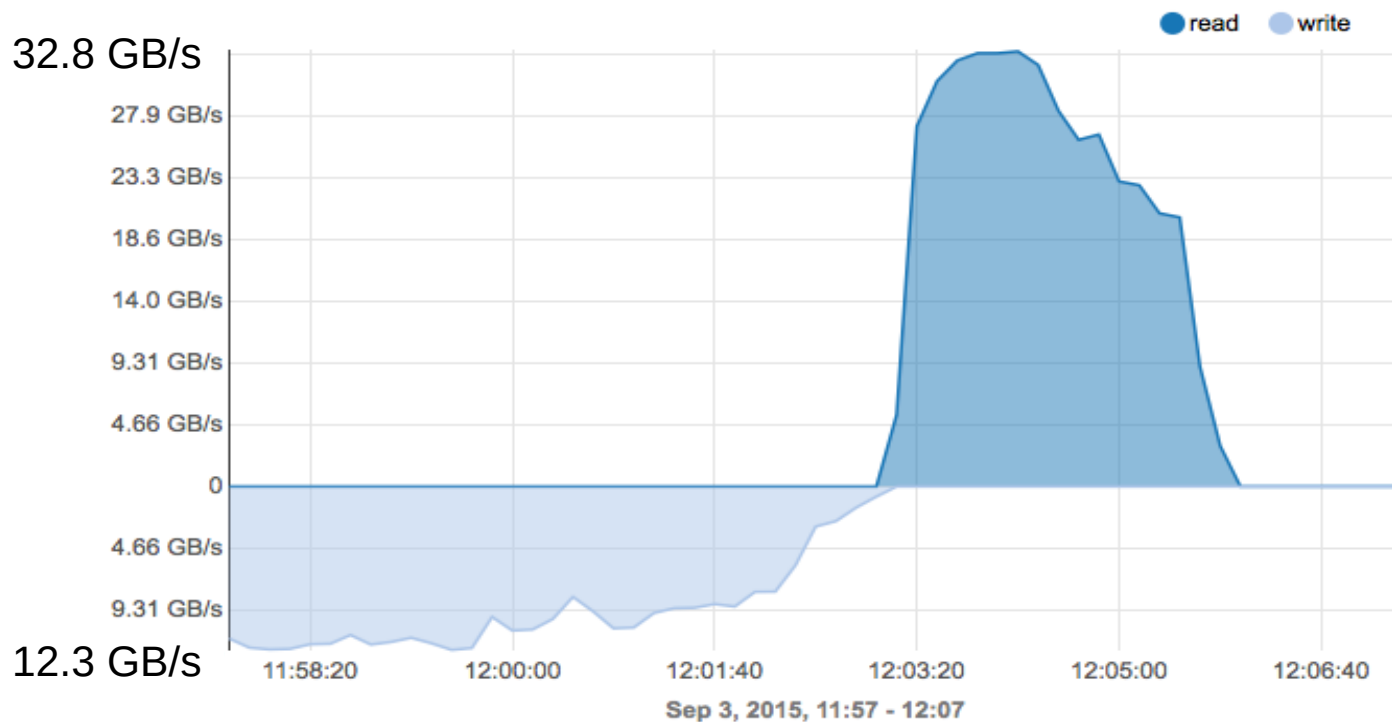
LNET Routing





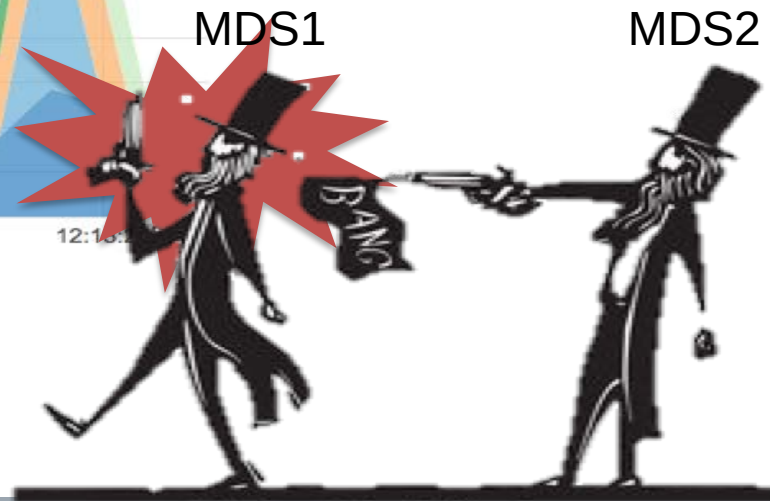
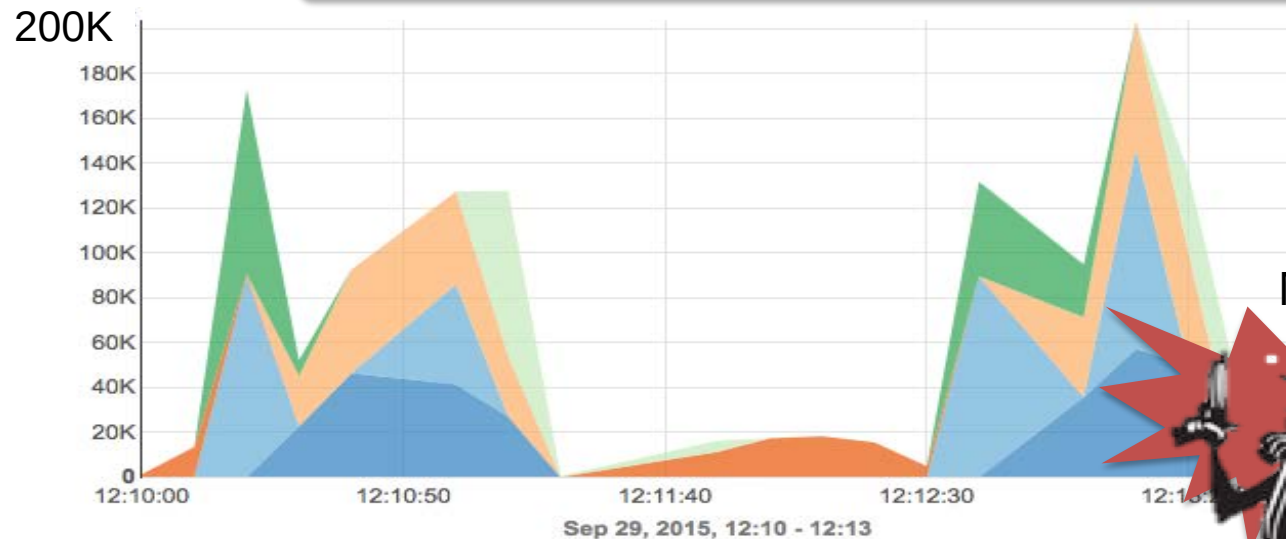
IEEL Testing

obdfilter-survey: thrhi=32 thrlo=2 nobjhi=32 nobjlo=1 size=64000



IEEL Testing

```
mdtest -z 2 -b 5 -l 20 -i 5 -d /n/holyfs/rc_admin/mdtest1
```





STONITH Shootout

Sep 24 21:07:25 holymds01 crmd[6780]: notice: throttle_handle_load: High CPU load detected: 90.419998

Sep 24 21:07:55 holymds01 crmd[6780]: notice: throttle_handle_load: High CPU load detected: 64.260002

Oleg Drokin pointed out:

Sep 24 21:28:02 holymds01 corosync[5268]: [TOTEM] Process pause detected for 39521 ms, flushing membership messages.

Sep 24 21:28:02 holymds01 corosync[5268]: [MAIN] Corosync main process was not scheduled for 39379.1172 ms (threshold is 4000.0000 ms). Consider token timeout increase.

Fix: corosync.conf
totem {
token: 10000



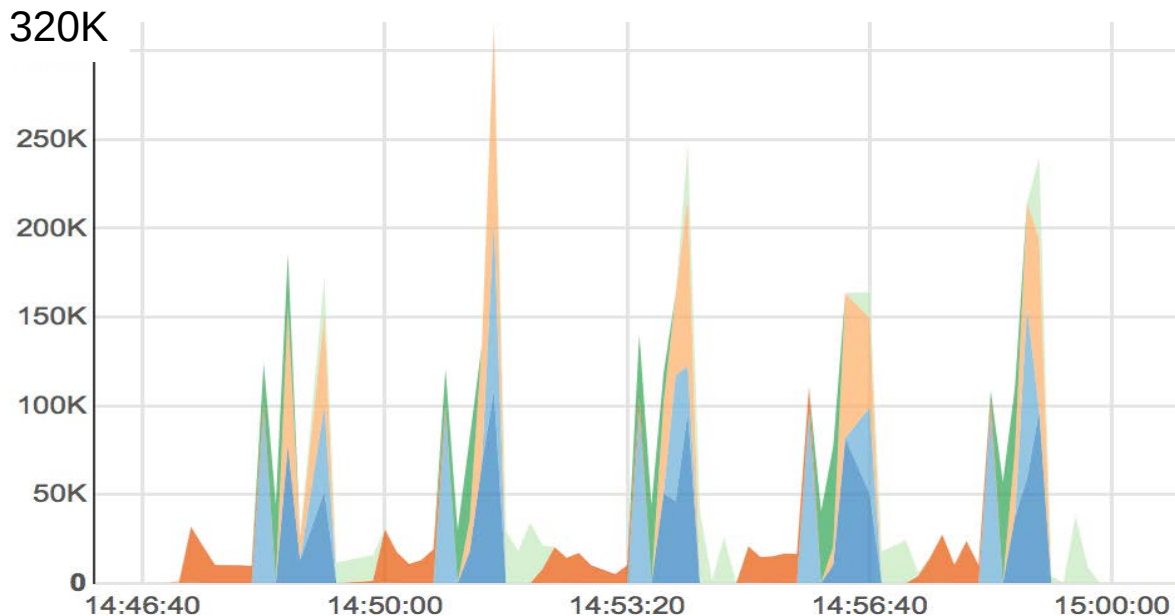
IEEL Testing

mdtest-1.9.3 was launched with 1560 total task(s) on 111 node(s)

mdtest -z 2 -b 5 -l 20 -i 5 -d /n/holyfs/rc_admin/mdtest12

FS: 3837.3 TiB Used FS: 0.1% Inodes: 3038.8 Mi Used Inodes: 0.0%

1560 tasks
967200 files/dir





Outline

- Lustre Deployments in the past year
 - Two 5.6 PB IEEL (HOLYLFS, BOSLFS)
 - 64 TB all Flash-based system (FLASHLFS)
 - 2.1 PB backup target (BULFS)
- Setup, Testing, Issues, ...



FLASHLFS

- NVMe Hardware
 - 16 Dell R730xd with 2 x 2TB Intel P3600
- Flexible Deployment (puppet + hiera)
 - Compute farm with stripped 4TB
 - Distributed Flash:
 - 1 MDT + 15 OSSs (30 OSTs)
 - IO Heavy:
 - Multiple MDTs and OSTs
 - Combinations of compute + Lustre



Build Challenges

- **Kernel Version:**
 - Intel P3600 series requires kernel 3.10
 - Lustre 2.7 server support listed at 2.6.32-504.8.1.el6
- **CentOS 7.2.1511 (3.10.0-229.14.1)**
 - Prep work: install dev-tools, remove libcom_err.i686, update e2fsprogs
 - Build lustre kernel
 - Clone kernel SRPM repo
 - Update SPEC file with kernel patch and configs
 - Build kernel rpm and install, reboot into kernel



Server Build

Build OFED

- `install kernel-devel-3.10.0-229.14.1_lustre_2.7.63`
- `mlnxofedinstall -k 3.10.0-229.14.1_lustre_2.7.63_5fda01f.el7.centos.x86_64 -add-kernel-support --force`

Build Lustre Server Stack

- `git clone git://git.whamcloud.com/fs/lustre-release.git`
- `sh ./autogen.sh`
- `sed -i 's/ldiskfsprogs/e2fsprogs/' lustre.spec.in`
- `./configure --with-linux=/usr/src/kernels/3.10.0-229.14.1_lustre_2.7.63_5fda01f.el7.centos.x86_64 --with-o2ib=/usr/src/ofa_kernel/default`

Puppetize build with created rpms and lay down configs



Single Node Write Test

```
time dd if=/dev/zero of=/mnt/test/cuff1.dat bs=1024k count=100000 oflag=direct
```

1 thread: (2.7 GB/s)

- 104857600000 bytes (105 GB) copied, 38.4266 s, 2.7 GB/s

2 threads: (3.4 GB/s)

- 104857600000 bytes (105 GB) copied, 62.3479 s, 1.7 GB/s

4 threads: (3.6 GB/s)

- 104857600000 bytes (105 GB) copied, 117.569 s, 892 MB/s

8 threads: (3.6 GB/s)

- 104857600000 bytes (105 GB) copied, 233.532 s, 449 MB/s



Single Node Read Test

```
time dd if=/dev/zero of=/mnt/test/cuff1.dat bs=1024k count=100000 oflag=direct
```

1 thread: (2.2 GB/s)

- 104857600000 bytes (105 GB) copied, 47.1717 s, 2.2 GB/s

2 threads: (5.5 GB/s)

- 104857600000 bytes (105 GB) copied, 29.8787 s, 3.5 GB/s, 2.0 GB/s

4 threads: (6.9 GB/s)

- 104857600000 bytes (105 GB) copied, 43.8152 s, 2.4 GB/s, 1.5 GB/s, ...

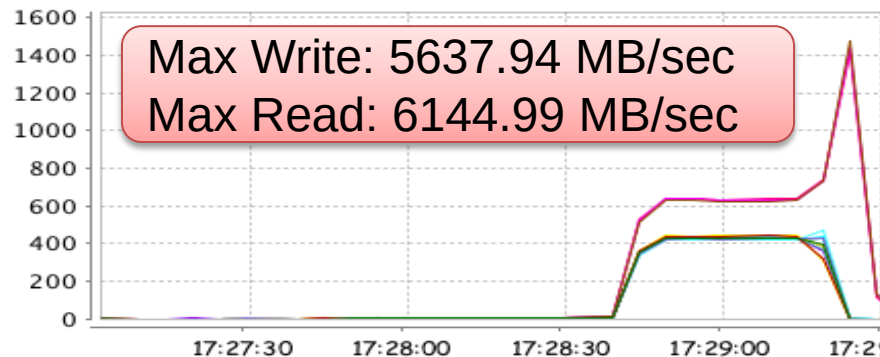
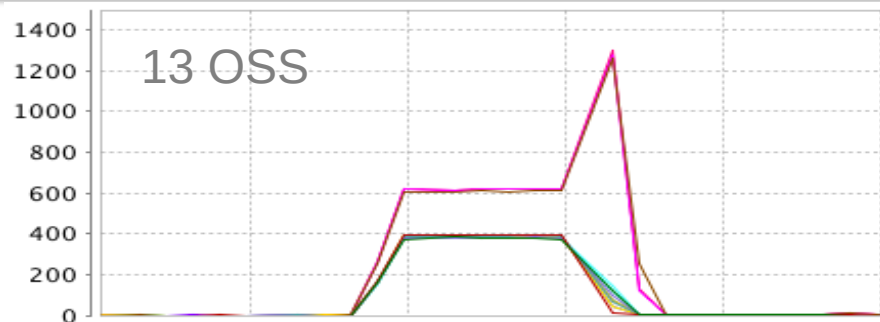
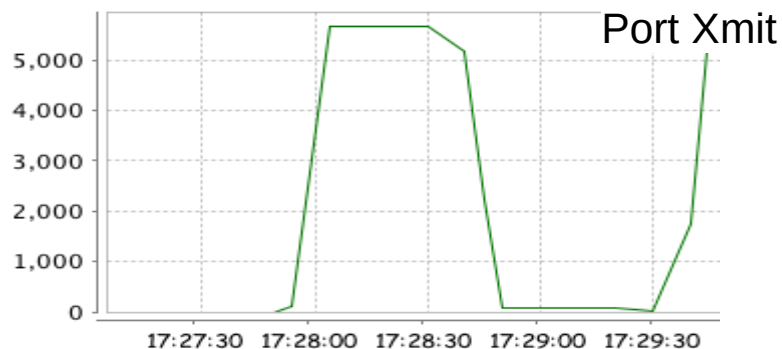
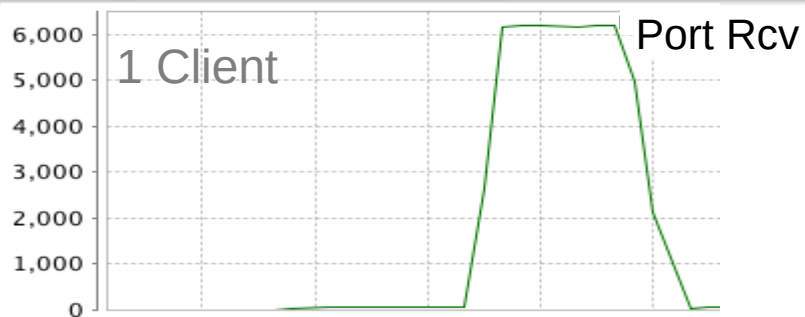
8 threads: (6.7 GB/s)

- 104857600000 bytes (105 GB) copied, 69.0972 s, 1.5 GB/s, 779 MB/s, 760 MB/s, ...



Initial Single Node IOR Test

```
mpirun -np 16 OR -v -F -B -a POSIX -t 4m -b 16g -o /n/flashlfs/test/test.58424178  
stripe_count: -1 stripe_size: 1048576
```





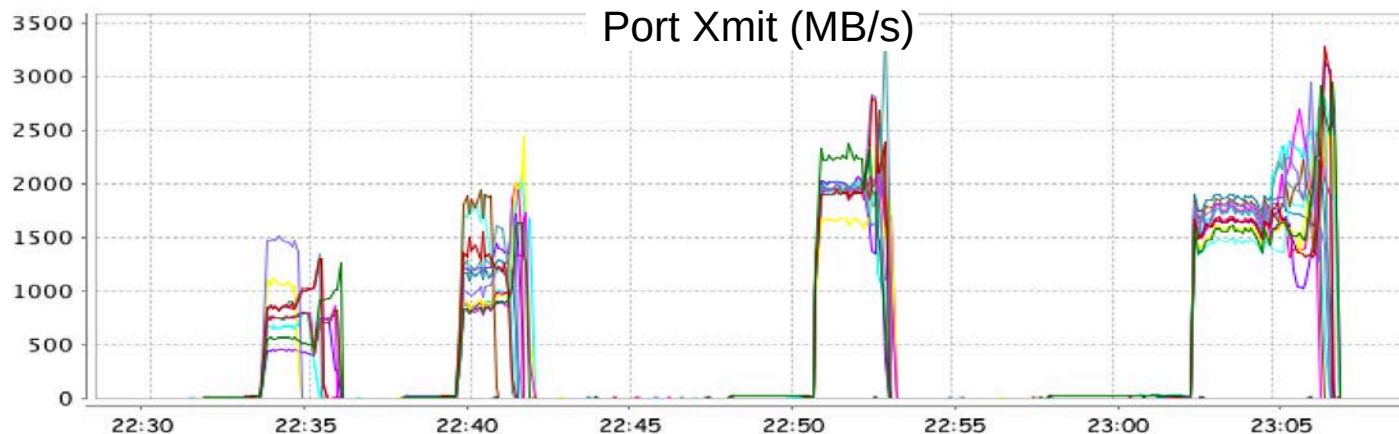
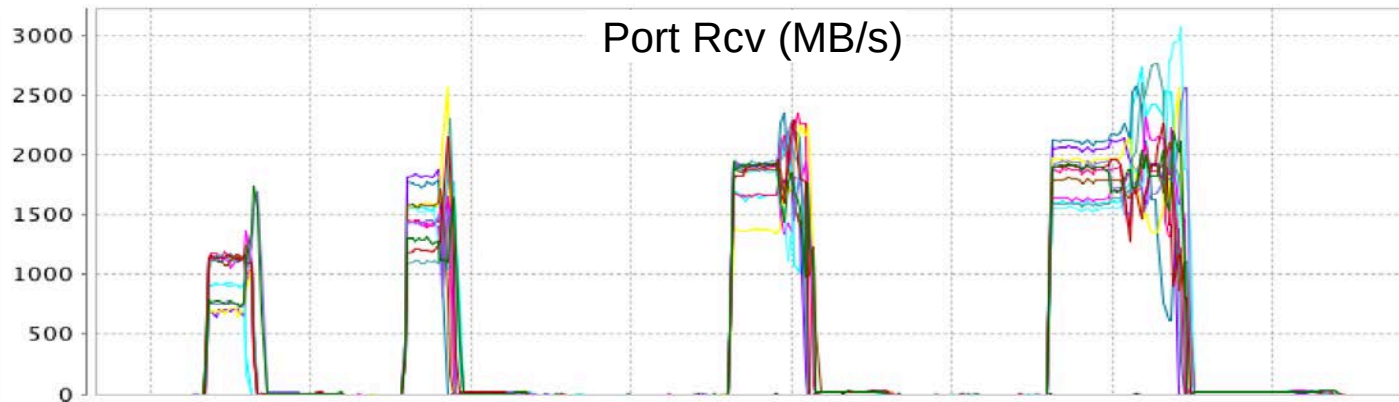
Initial Multi Node IOR Test

13 Clients
182 IOR threads
stripe_count: 1
stripe_size: 4M

13 OSS
(26 OSTs)

Read & Writes
19 – 22 GB/s

Max: (26) P3600
44.2 GB/s





Discussion with Zhiqi

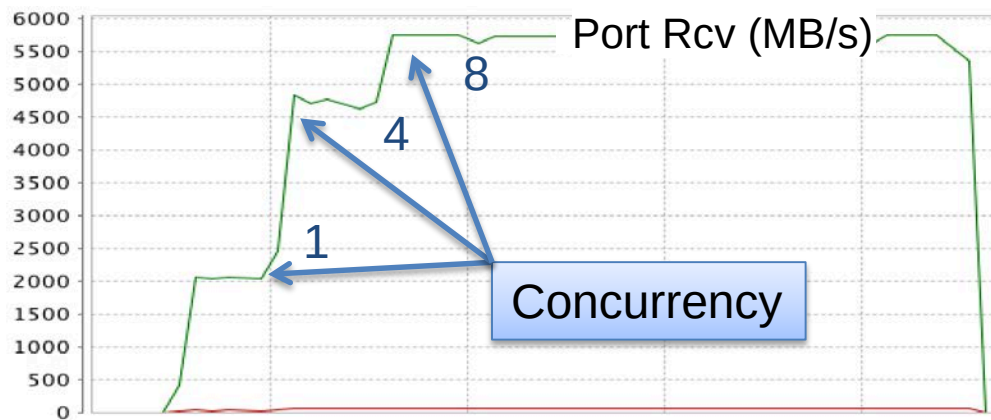
- lctl tweaks

```
lctl set_param osc.flashlfs-*.max_pages_per_rpc=1024
lctl set_param osc.flashlfs-*.max_rpcs_in_flight=64
lctl set_param osc.flashlfs-*.max_dirty_mb=1024
lctl set_param osc.flashlfs-*.checksums=0
lctl set_param llite.flashlfs-*.max_read_ahead_mb=1024
lctl set_param llite.flashlfs-*.max_cached_mb=81920
lctl set_param llite.flashlfs-*.max_read_ahead_per_file_mb=1024
lctl set_param subsystem_debug=0
```

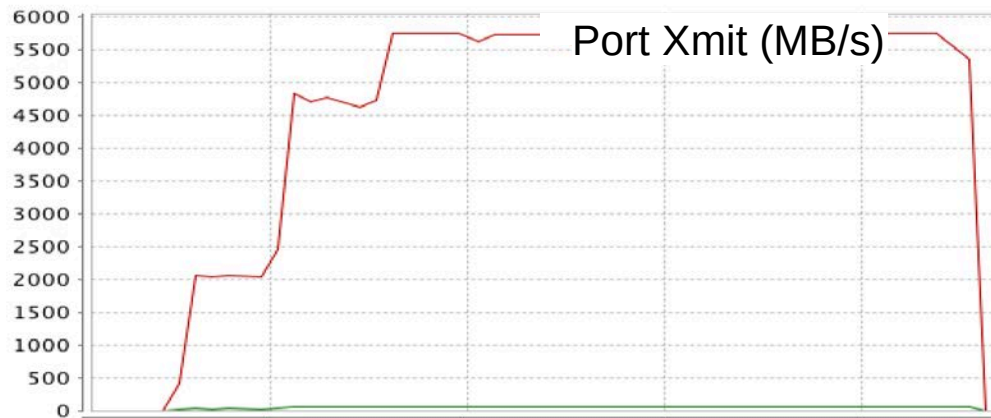
- lst with --concurrency={1 4 8 16}



1:1 Lnet Selftest

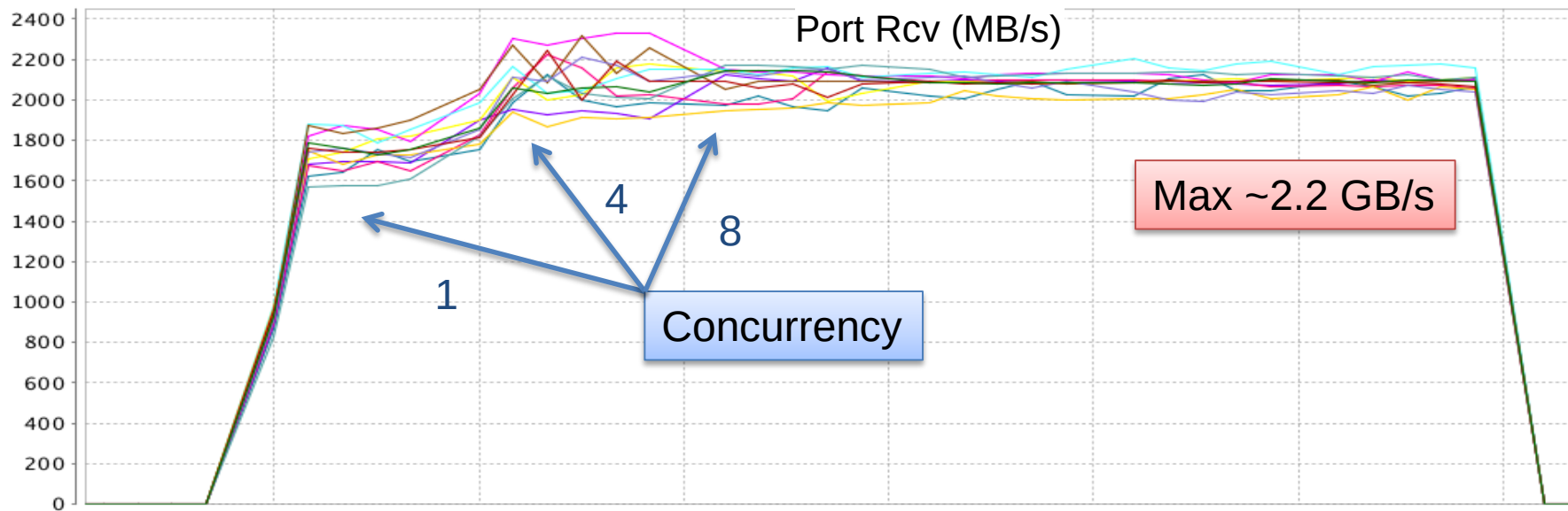


Max 5.8 GB/s





10 Client : 10 OSS Lnet Selftest

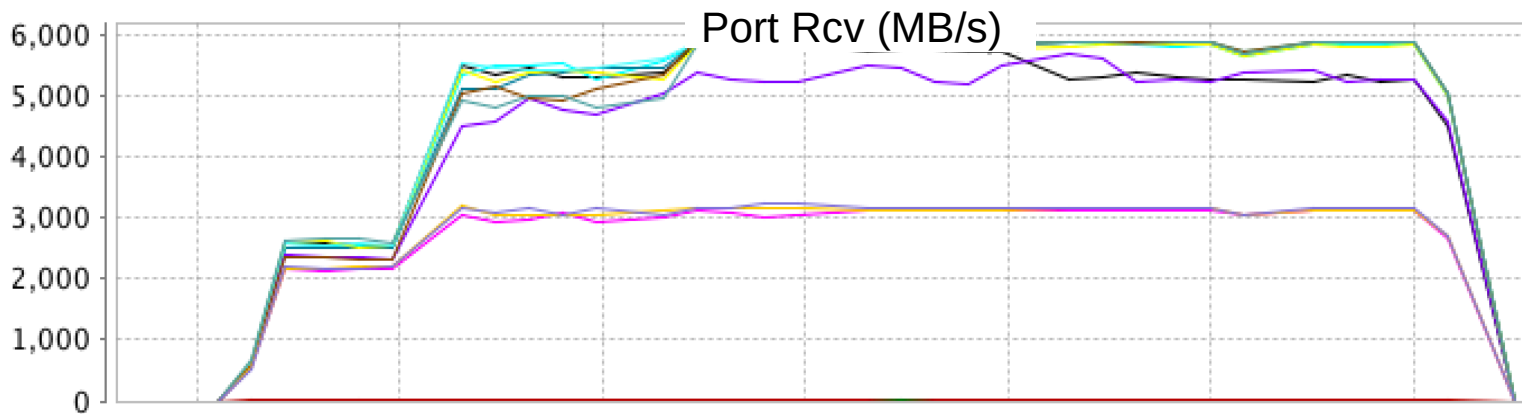
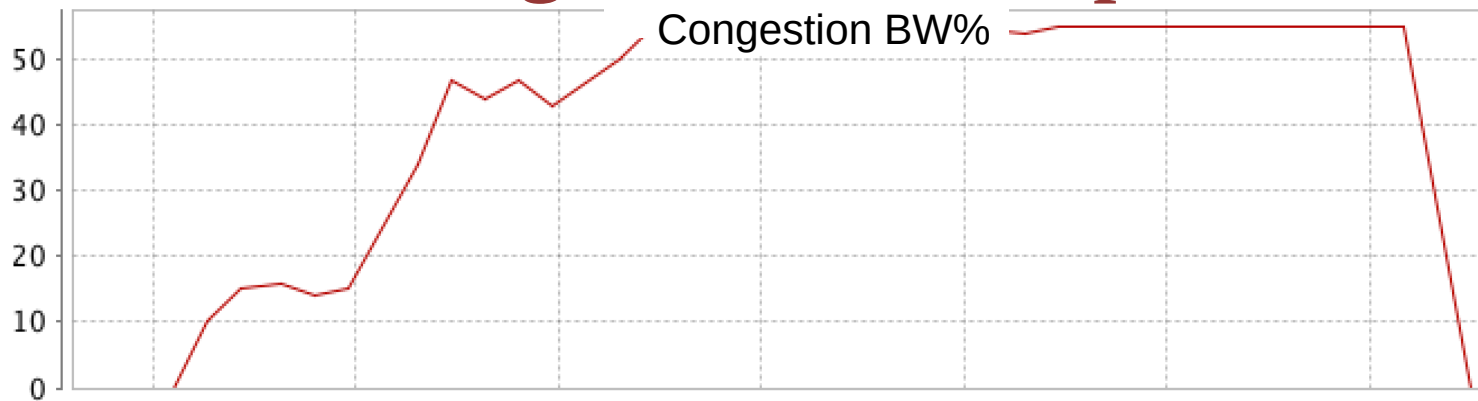


So what happened?

Lots of IB Congestion



IB Congestion Example





Multi Node IOR Testing – Part 2

IB Switches	Nodes	Threads	Write (MB/s)	Read (MB/s)
2	64	256	34744	35643
2	64	512	37594	37296
9	16	256	36053	22676
9	20	240	35814	16133
22*	128	512	41499	42285

*FDR(10) connected AMD nodes

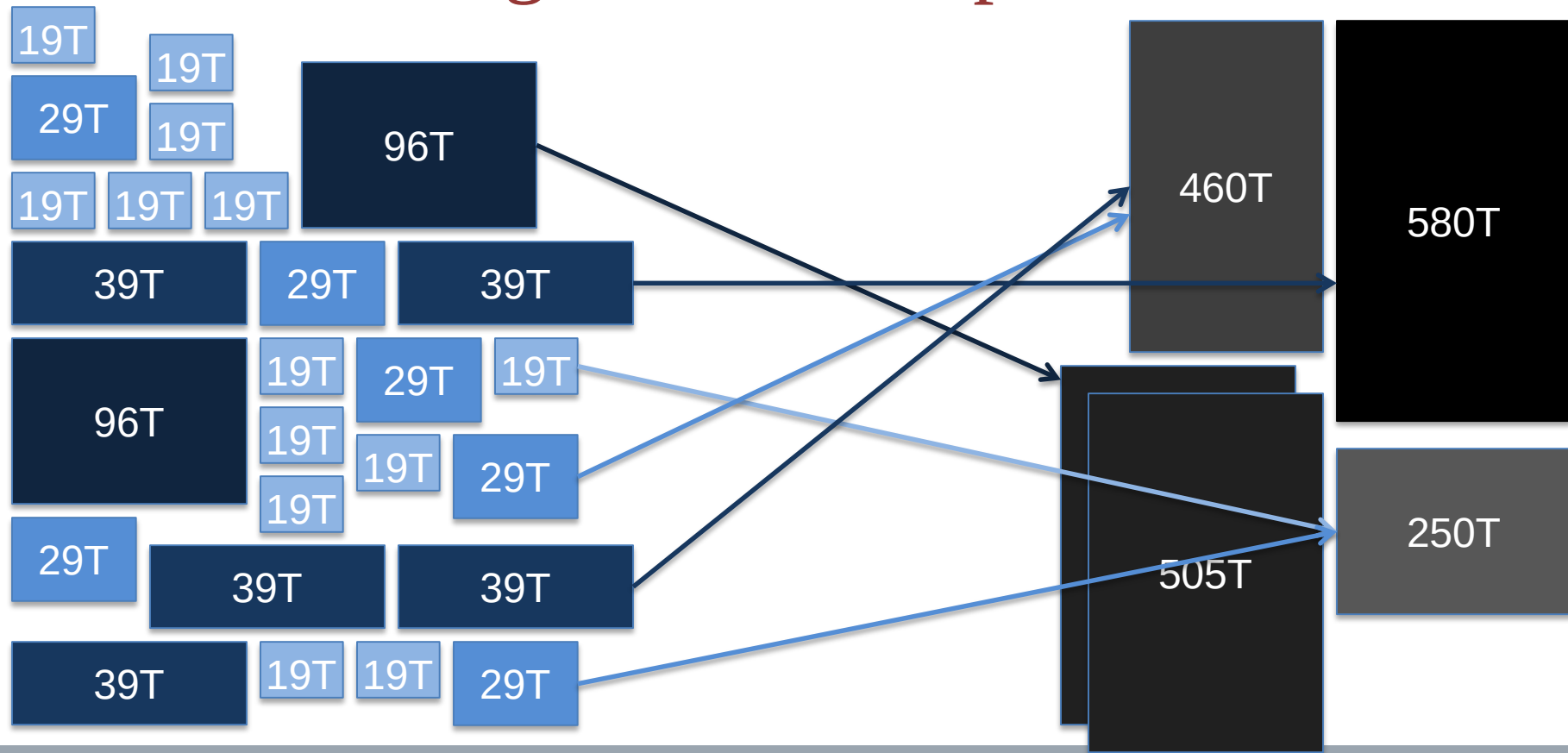
Max Write:
(26) P3600
44.2 GB/s



Outline

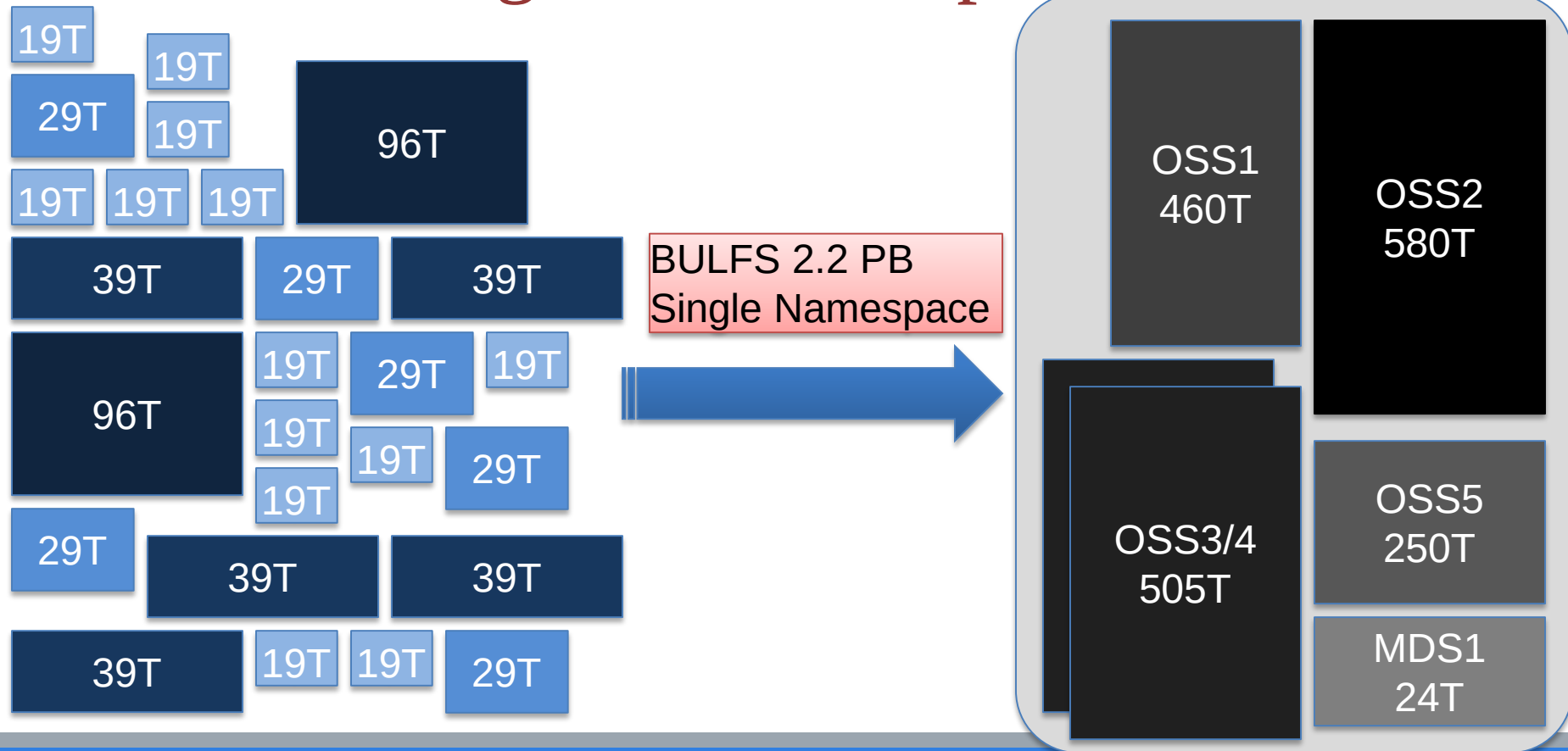
- Lustre Deployments in the past year
 - Two 5.6 PB IEEL (HOLYLFS, BOSLFS)
 - 64 TB all Flash-based system (FLASHLFS)
 - 2.1 PB backup target (BULFS)
- Setup, Testing, Issues, ...

Converting NFS Backups to Lustre





Converting NFS Backups to Lustre

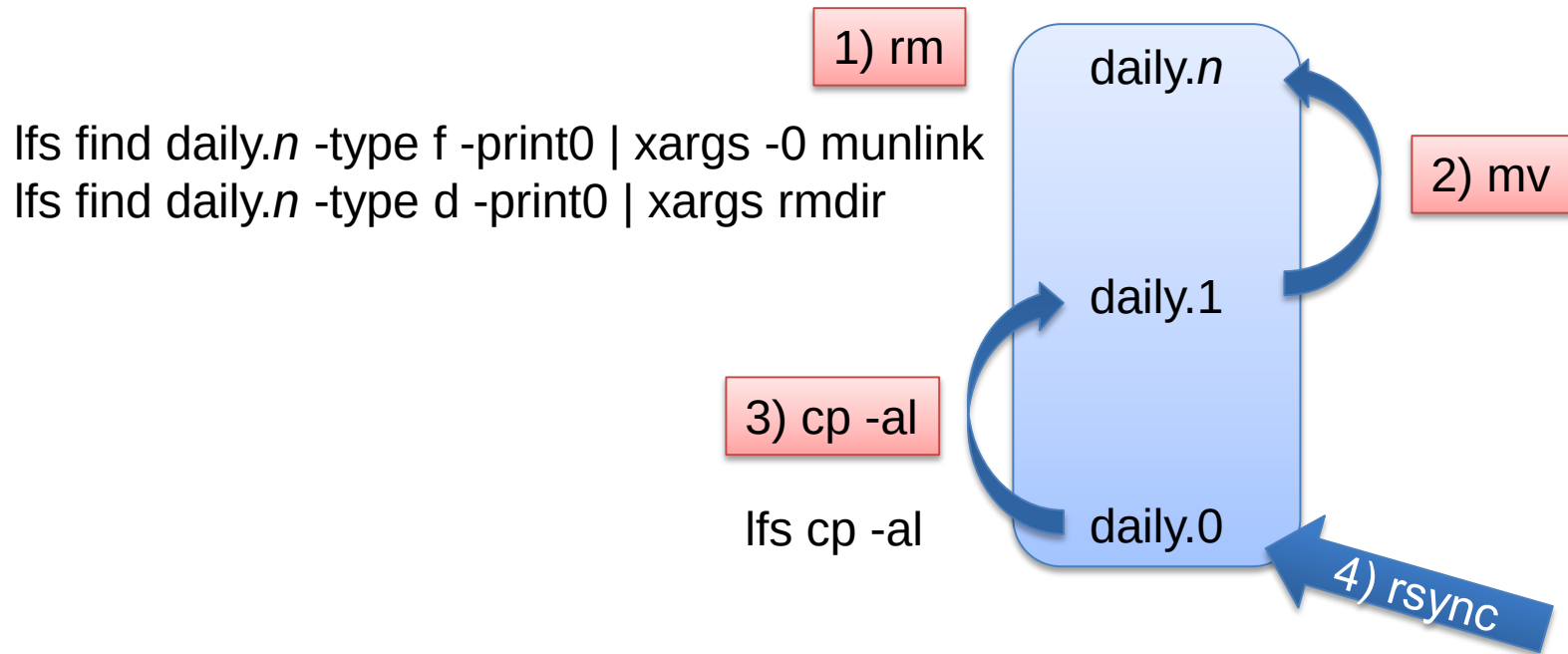




Backup Challenges

- Disaster Recovery Backups 100 miles away
- Backup of every NFS filesystem every day
- Entire MDT must be parsed everyday
- Rsnapshot performance
- Loads of tiny files, 10k file directories
- Monitoring IO bottlenecks

Rsnapshot





Monitoring

Python capture
feed into Graphite

- mdt: md_stat
- ost: stats,
brw_stats

Dashboard in
Grafana





Summary

- Lustre Deployments in the past year

- Two 5.6 PB IEEL (HOLYLFS, BOSLFS)
- 64 TB all Flash-based system (FLASHLFS)
- 2.1 PB backup target (BULFS)

Testing is Key

IB Topology Matters

Let Metrics Drive Design

- Setup, Testing, Issues, ...