GridPP

UK Computing for Particle Physics

Lustre

Christopher J. Walker

Queen Mary
University of London

Overview

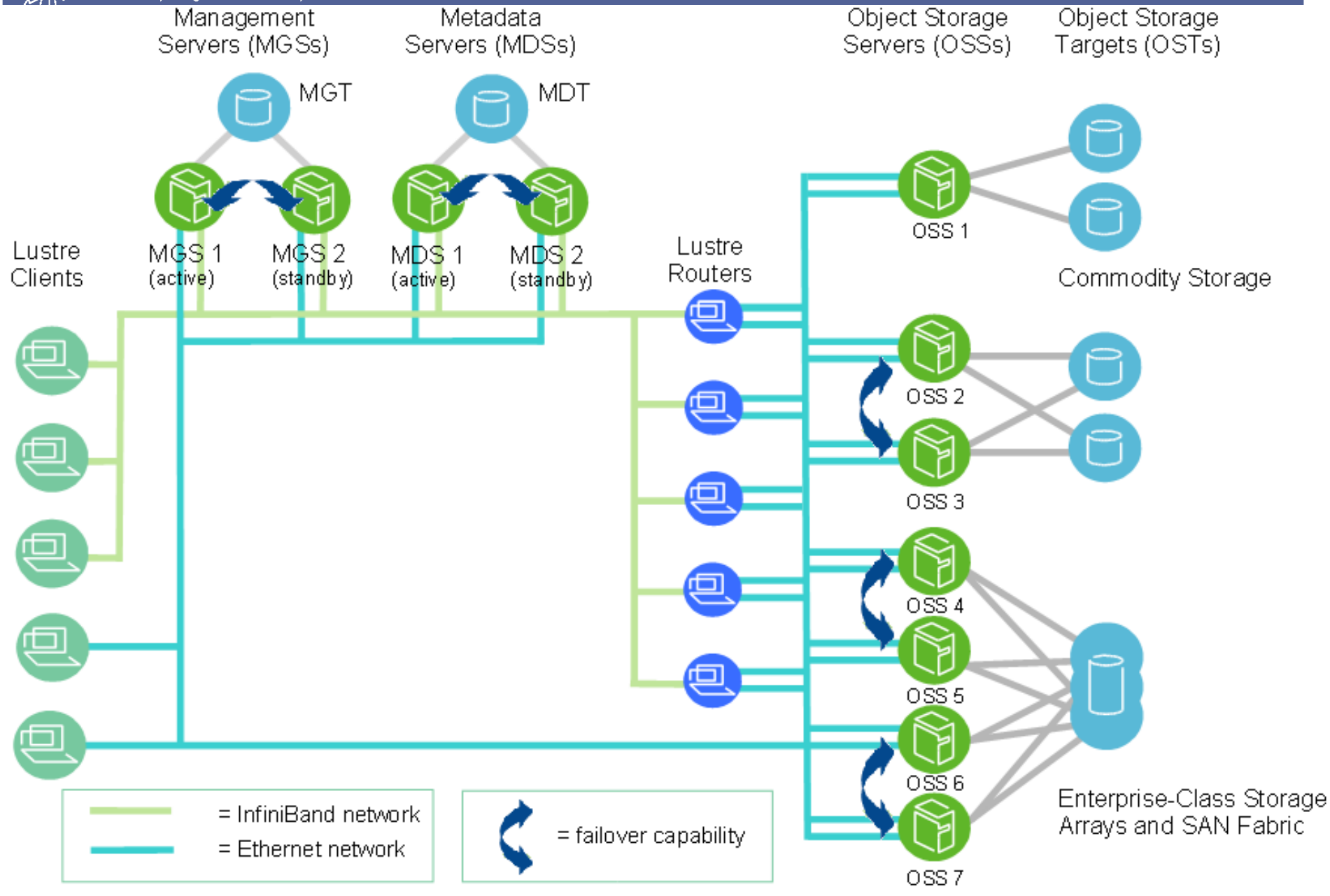
- Lustre Overview
- Lustre Roadmap
- QMUL Network
- QMUL benchmarks
- Conclusions

Big Lustre Systems

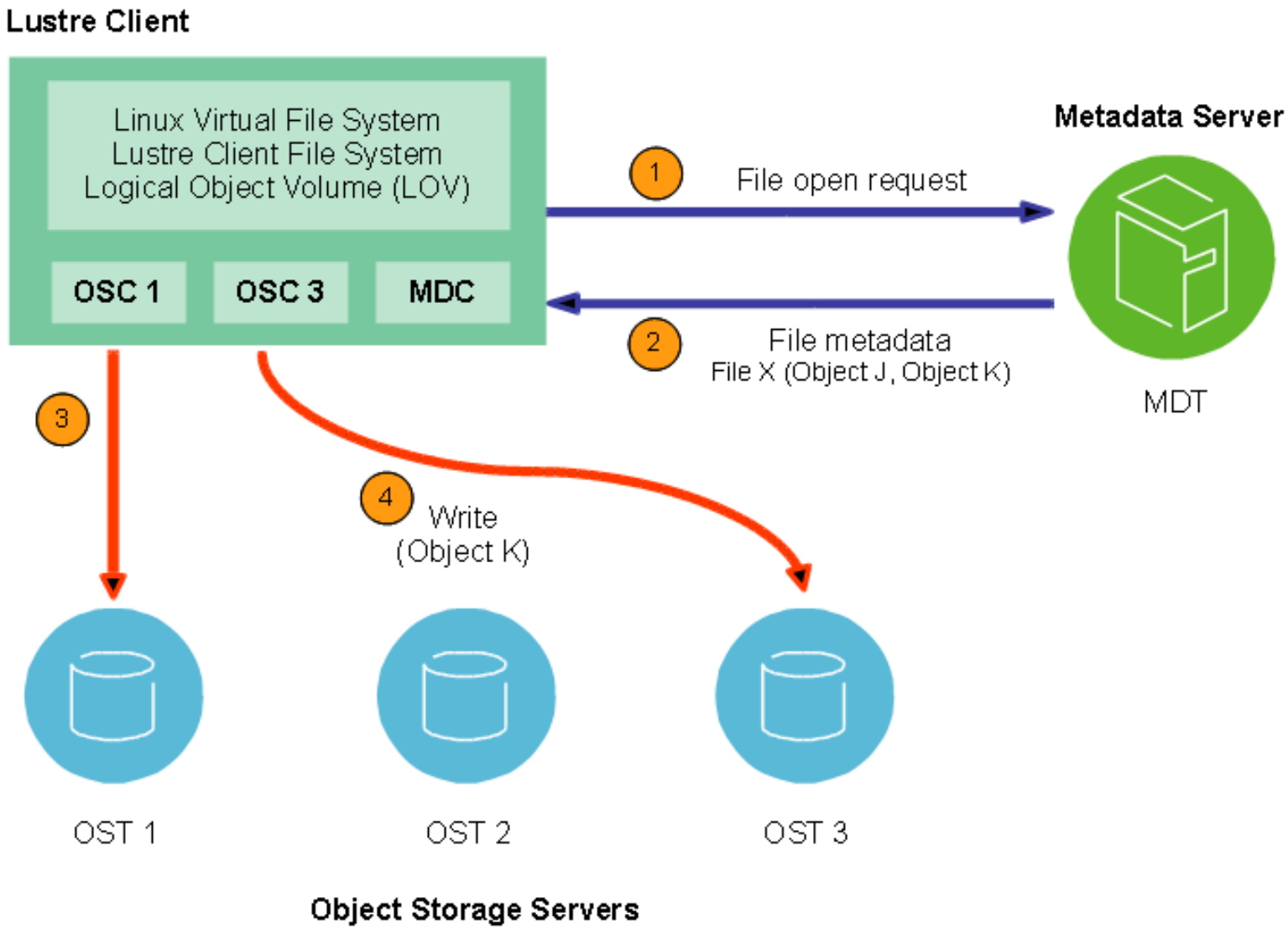
- Lustre stats – production system (practical)
  - 50000+ clients
  - 240 GB/sec I/O
  - 750 million files
  - 15000/s create operations,
  - 35000/s metadata stat operations
  - multi-TB max file size
  - 10 PB space,

GridPP
UK Computing for Particle Physics

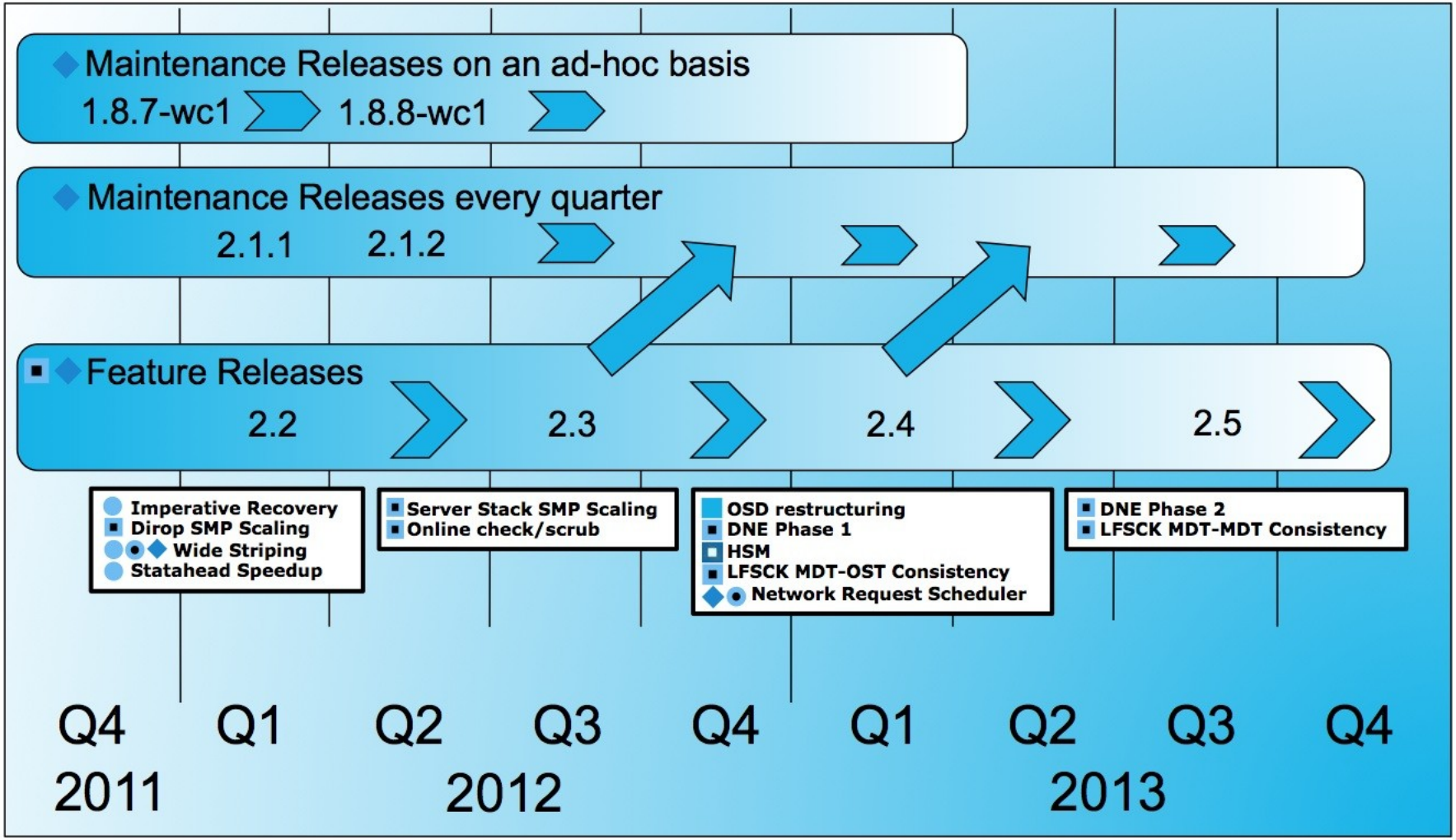
Lustre Overview



Lustre Overview



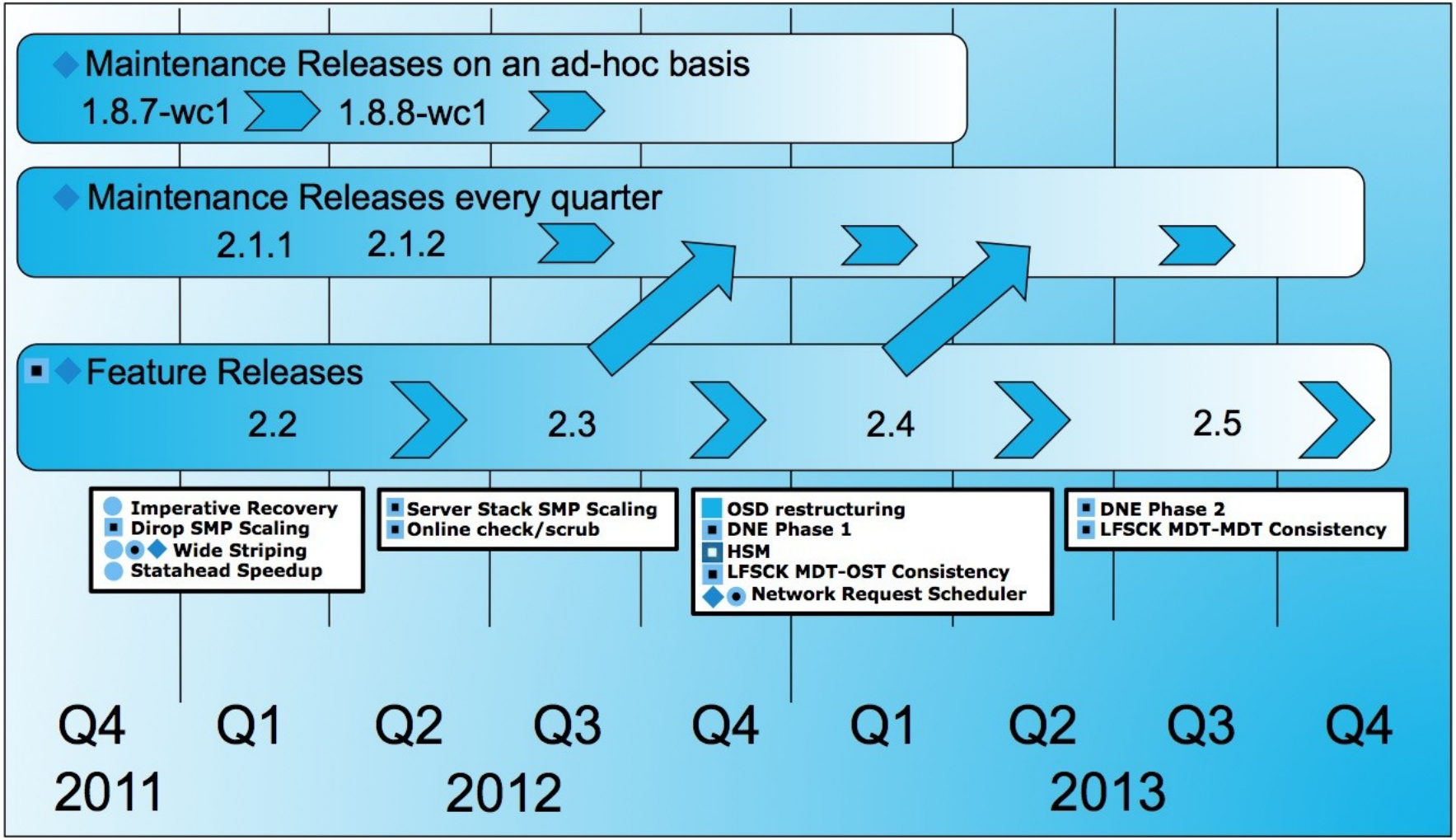
Community Lustre Roadmap



Sponsor for Whamcloud Development and Releases: ORNL OpenSFS LLNL Whamcloud

Third Party Development: CEA Xyratex

Community Lustre Roadmap

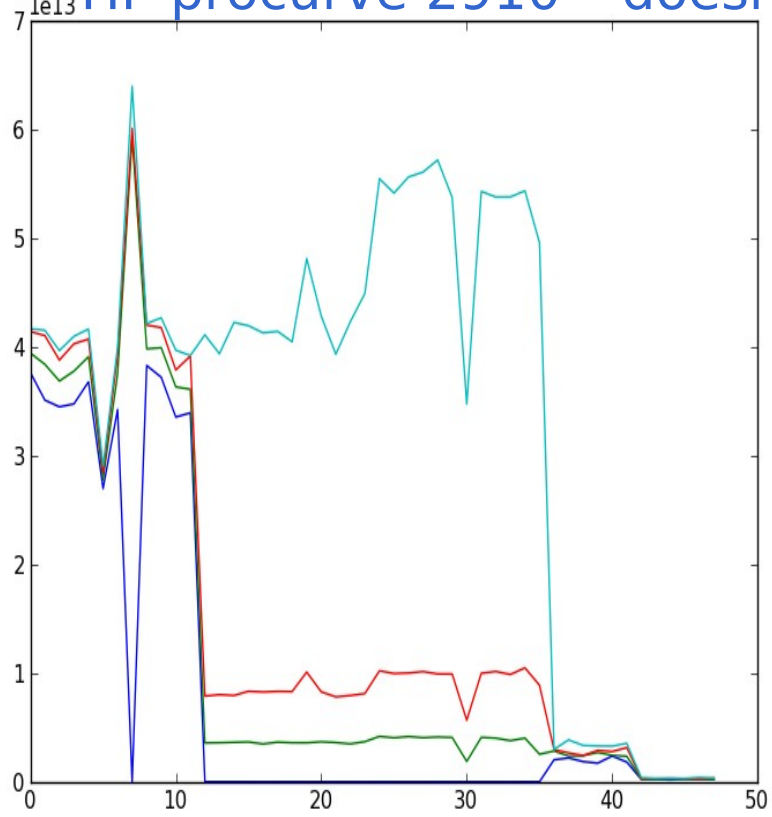


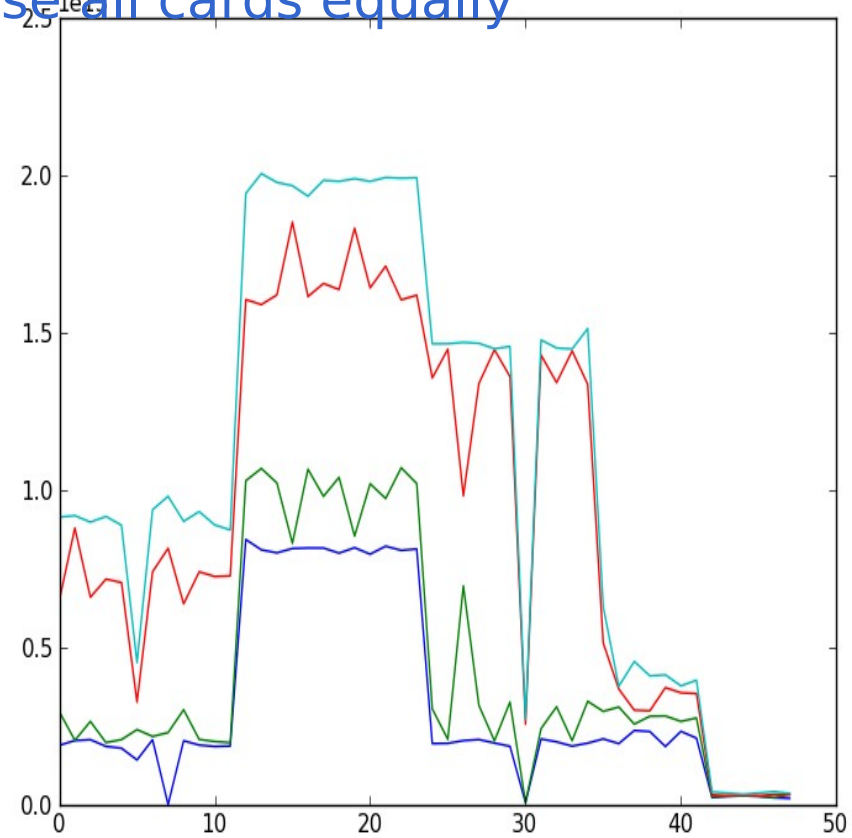
Sponsor for Whamcloud Development and Releases: ORNL OpenSFS LLNL Whamcloud

Third Party Development: CEA Xyratex

Hashing

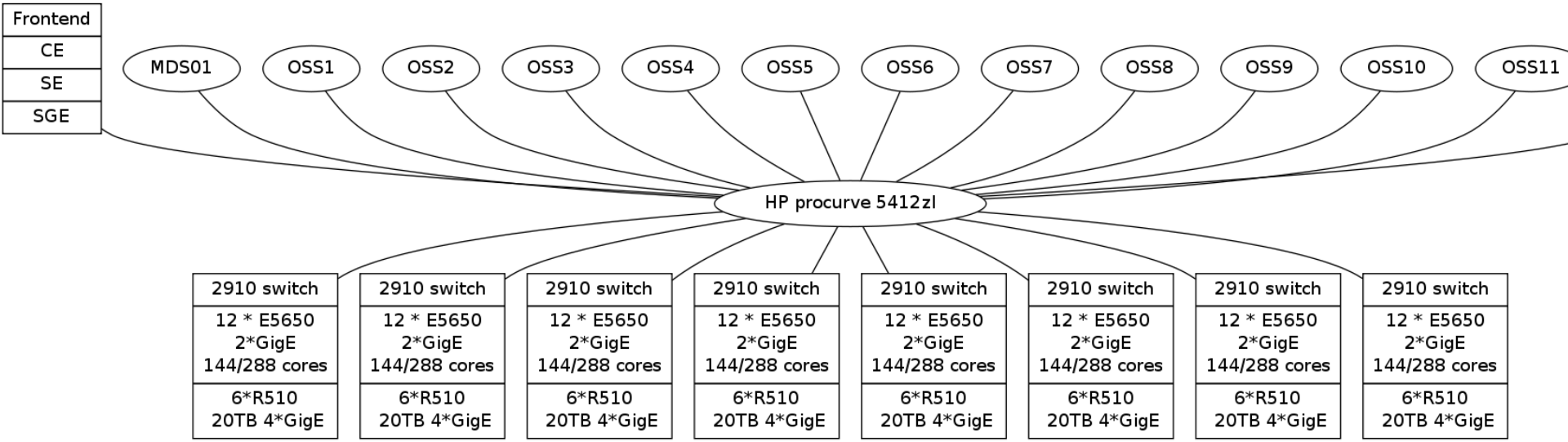
- Bonding: LACP Mode 4:
  - Need xmit_hash_policy=3+4
    - eth0 even on all machines – default doesn't work
  - HP procurve 2910 – doesn't use all cards equally





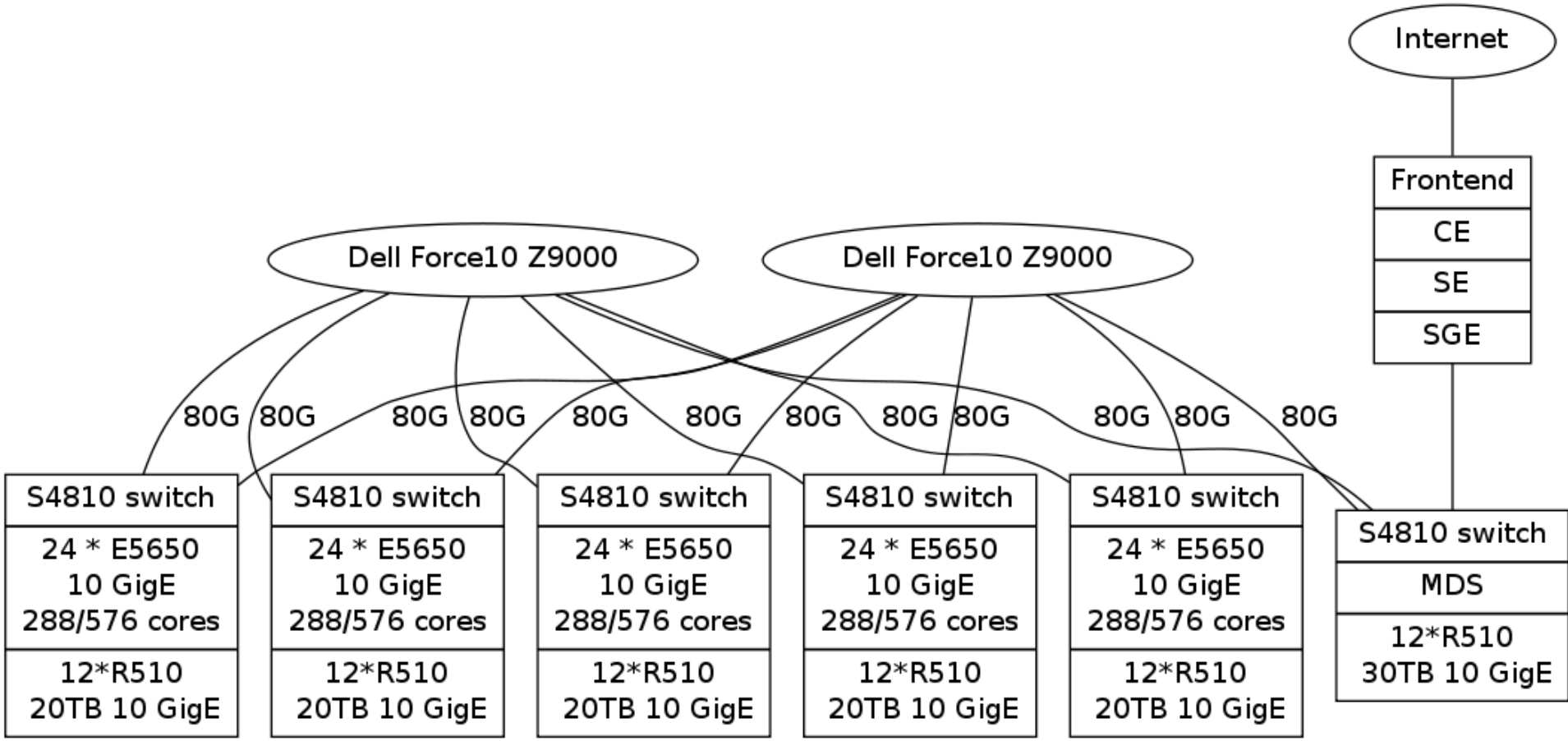
Old QMUL Network

- 48 *1Gig per switch – 10Gig uplink
- 12 compute nodes (2*1Gig)
- 6 Storage nodes (4*1Gbit)



New QMUL Network

- 5* Lustre “Brick”
  - Force10: S4810 Switch



Bonding and NIC ordering

- SL5
  - NIC ordering nondeterministic
    - Even between boots
    - Use udev rules

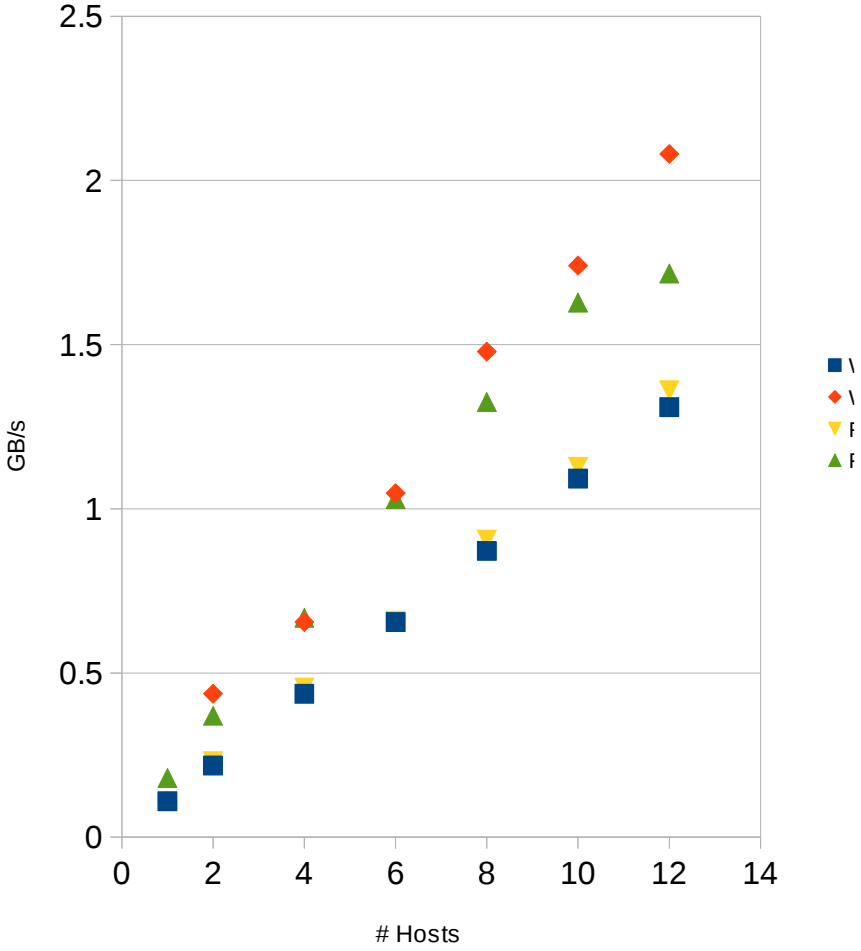
```
[root@cn636 ~]# cat /etc/udev/rules.d/70-persistent-net.rules
KERNEL=="eth*",  ID=="0000:01:00.0", NAME="eth0"
KERNEL=="eth*",  ID=="0000:01:00.1", NAME="eth1"
KERNEL=="eth*",  ID=="0000:04:00.0", NAME="eth2"
KERNEL=="eth*",  ID=="0000:04:00.1", NAME="eth3"
```

- Failover bonding (10Gig default)

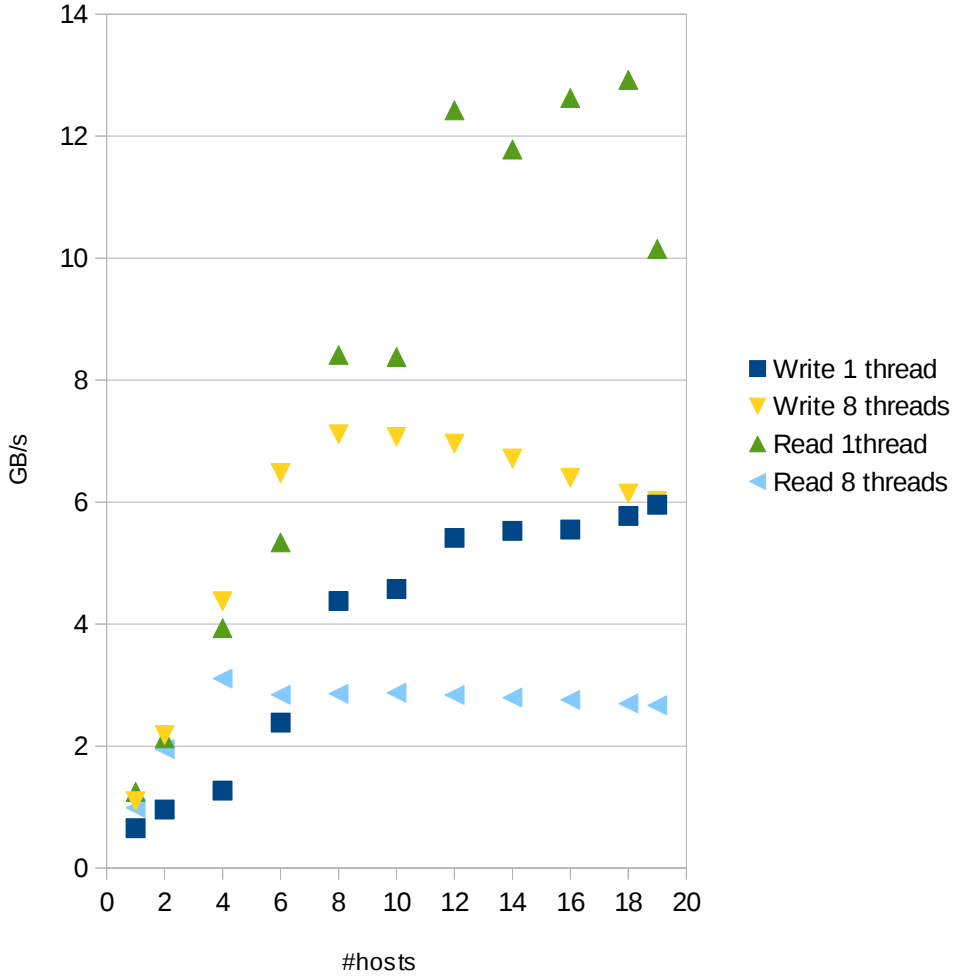
```
# cat /etc/modprobe.conf
alias bond0 bonding
options bond0 miimon=100 mode=1 updelay=200 downdelay=200 primary=eth2
```

Iozone results

Gbit: 6*Storage Server



10Gig: 10 Storage servers



Misc interesting links

- Hadoop on Lustre
  - http://wiki.lustre.org/index.php/Running_Hadoop_with_Lustre
- Lustre on Amazon
  - http://www.cloudscale.com/index.php/technology/lustre-on-aws-

Conclusions

- Future looks bright for Lustre
  - Especially if it gets into the kernel
- Good performance
  - Bonded Gbit: use xmit_hash_policy
  - 10Gbit: Excellent performance
- Scales well