StoRM and Lustre at QMUL

Christopher J. Walker
Alex Martin, Duncan Rand

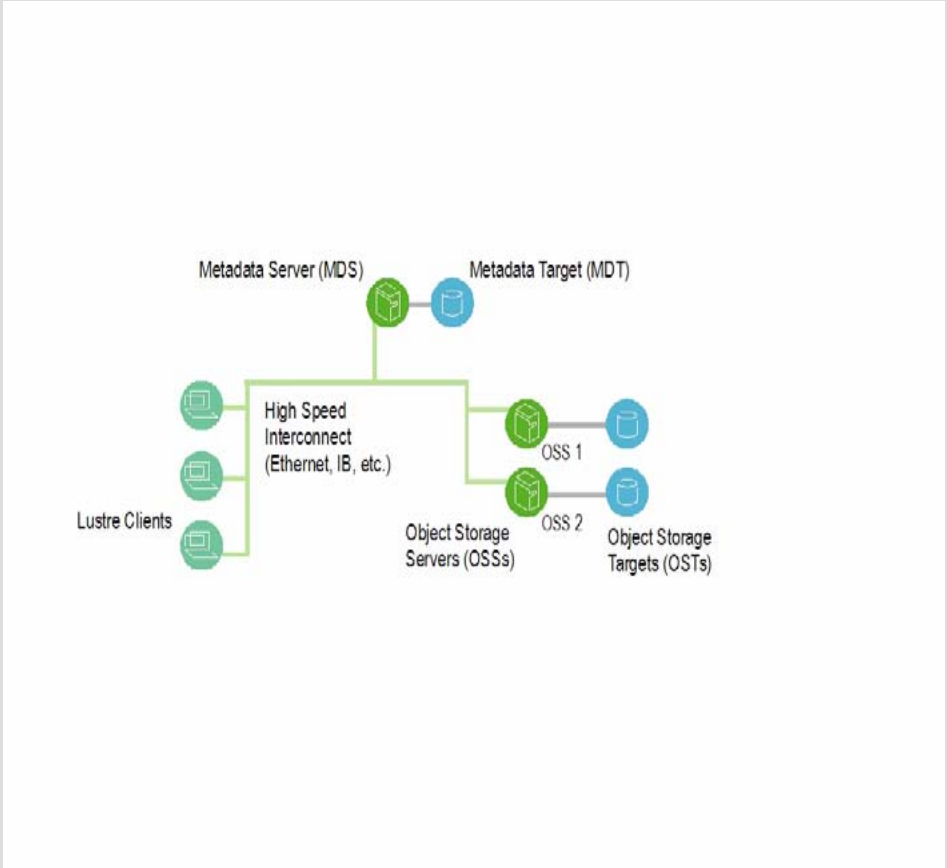
Overview

- Lustre
  - Concepts (How it works)
  - Implementation
    - QMUL's Network
    - Benchmarks
- Storm
  - Concepts
  - Implementation
- Storm / Lustre
  - Hammercloud results

Lustre

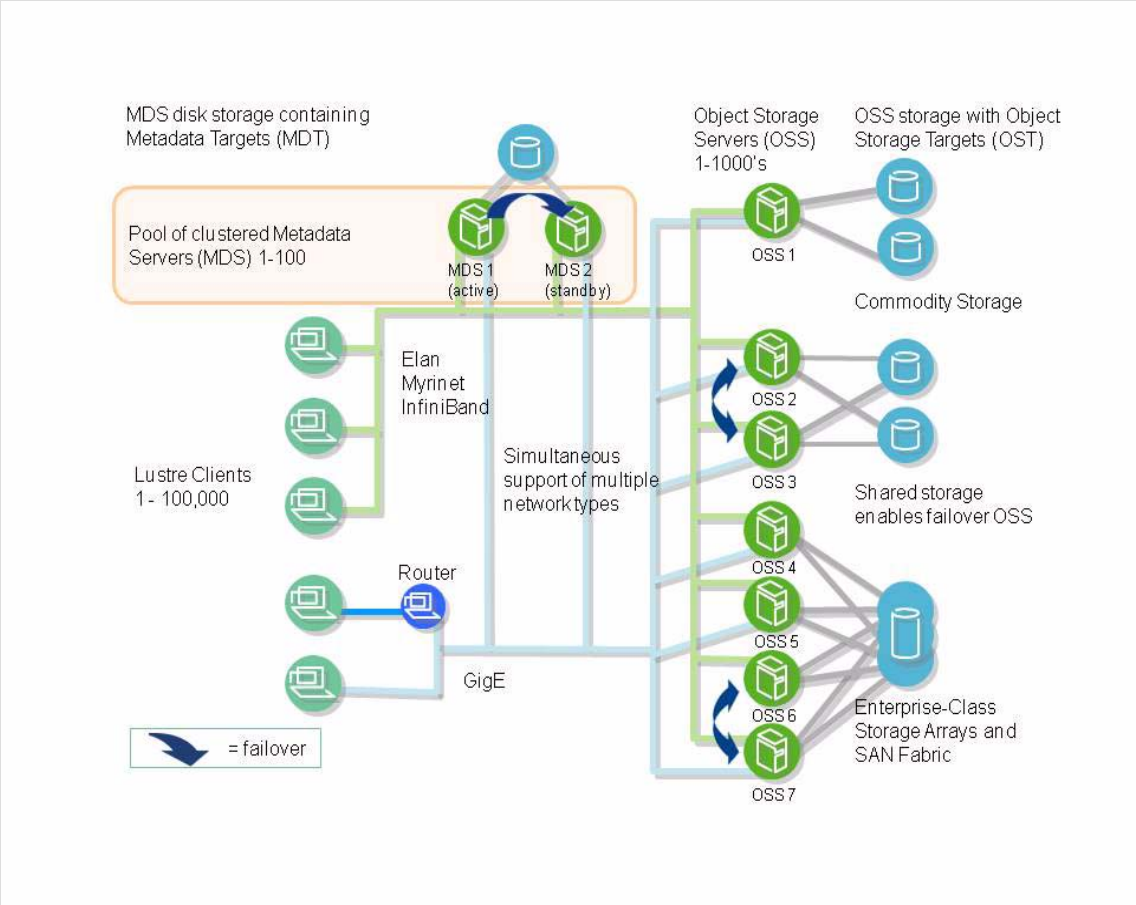
- Posix filesystem
- High performance
  - 7/10 of top supercomputers
- Scalable
  - Increasing OSTs increases performance
- Free (GPL)

Lustre Architecture



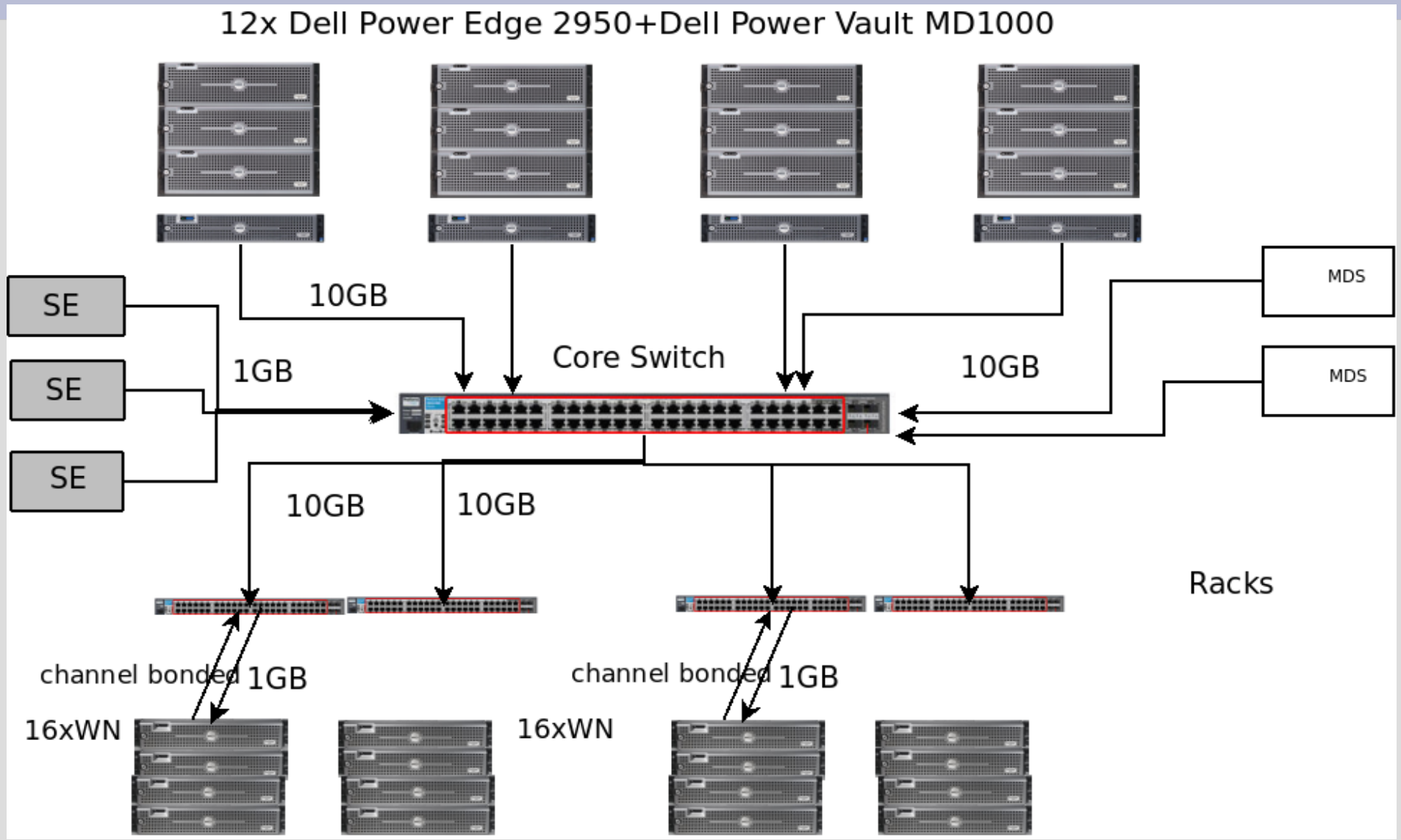
- Metadata server (MDS)
  - Holds file metadata
- OSS (object storage servers)
  - 10GigE
- OST (object storage target)
  - Raid 6 Disk

Advanced Lustre Architecture



- Failover MDS
- Striping
  - File or directory basis
  - Off by default
  - Hot files

QMUL Network



Lustre Filesystem Testing

- Measure performance
  - Identify bottlenecks / misconfigurations
  - Stress test filesystem
- Considerations
  - Files large enough to avoid caching
  - Network topology
  - Bonded interfaces

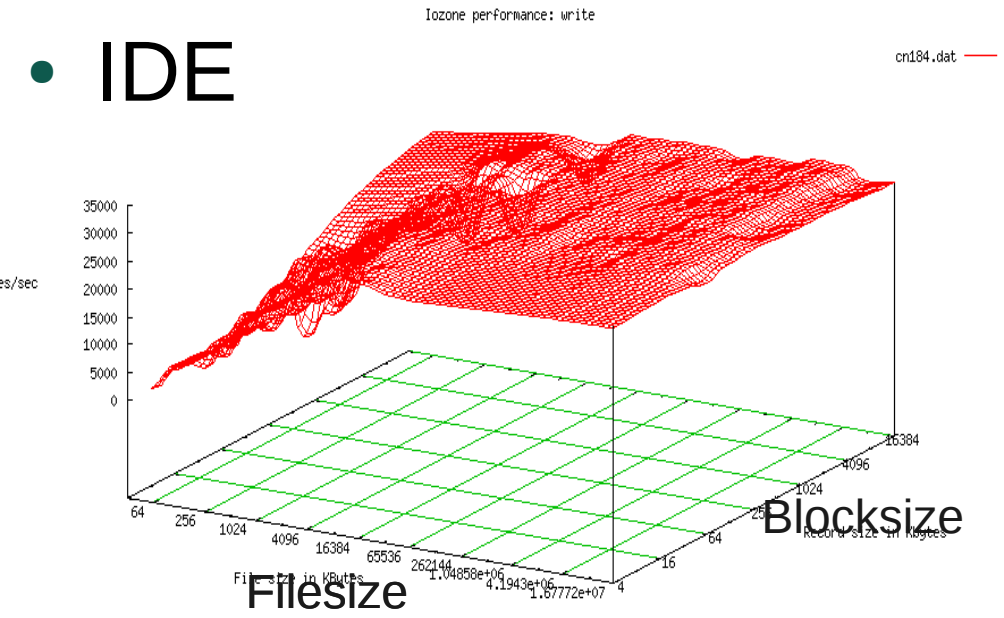
Performance Testing

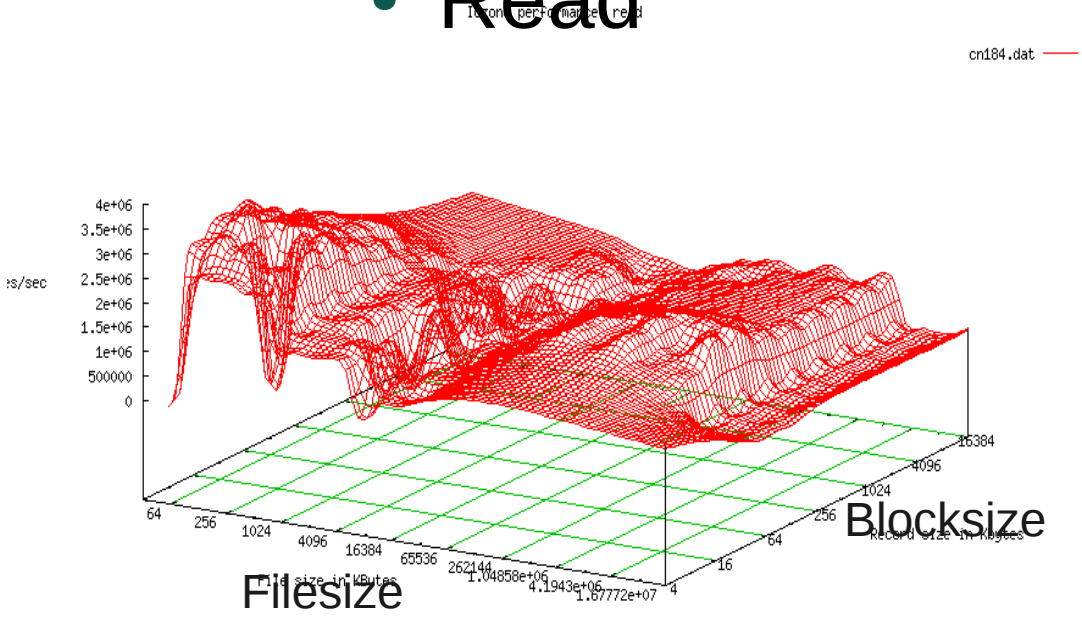
- Bonnie
  - Can't sync between machines
- Iozone
  - Used for these results
  - Hangs if it loses a UDP packet – likely on a heavily loaded network
- IOR
  - Not tried (requires MPI)

IDE/SATA Disk

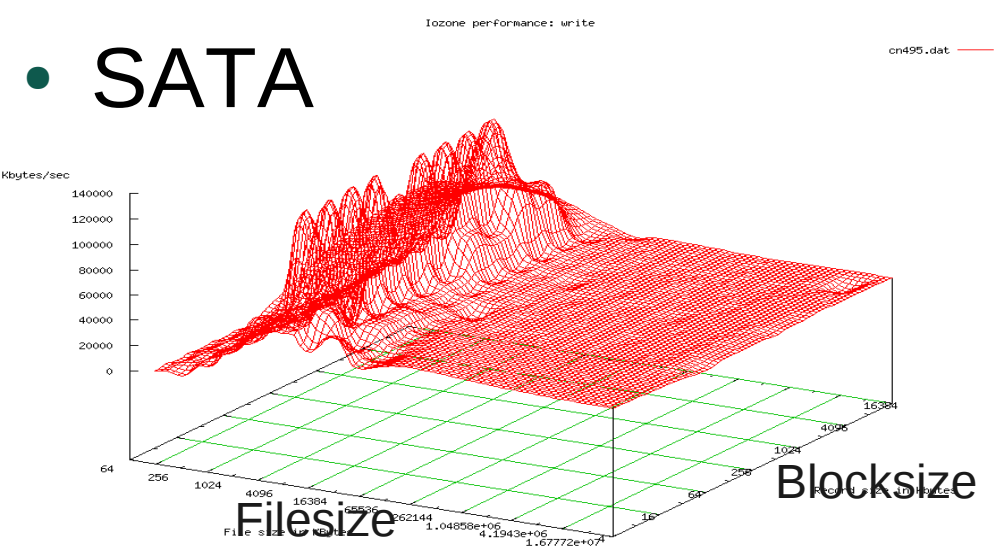
- Write
- Read

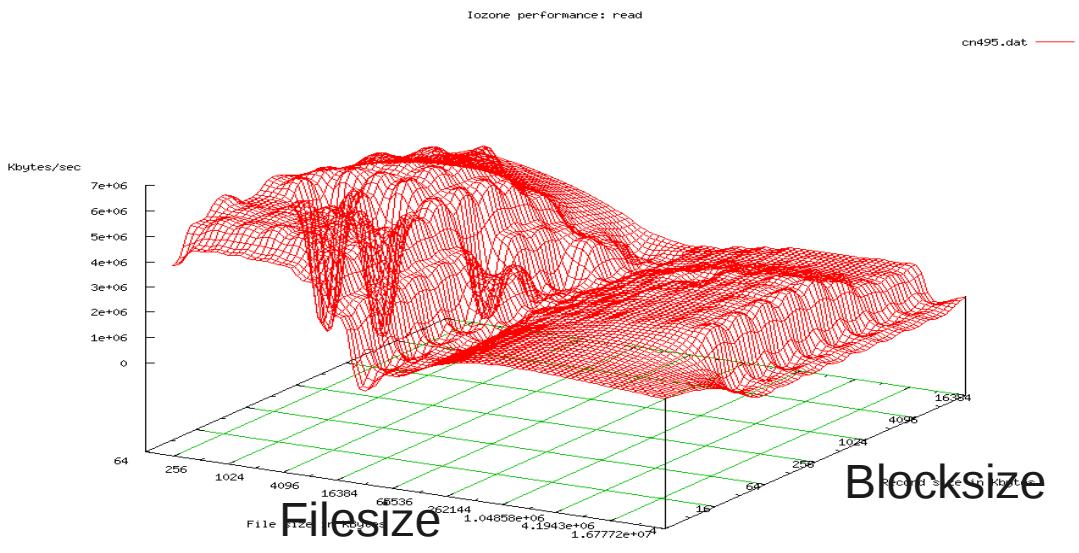
- IDE



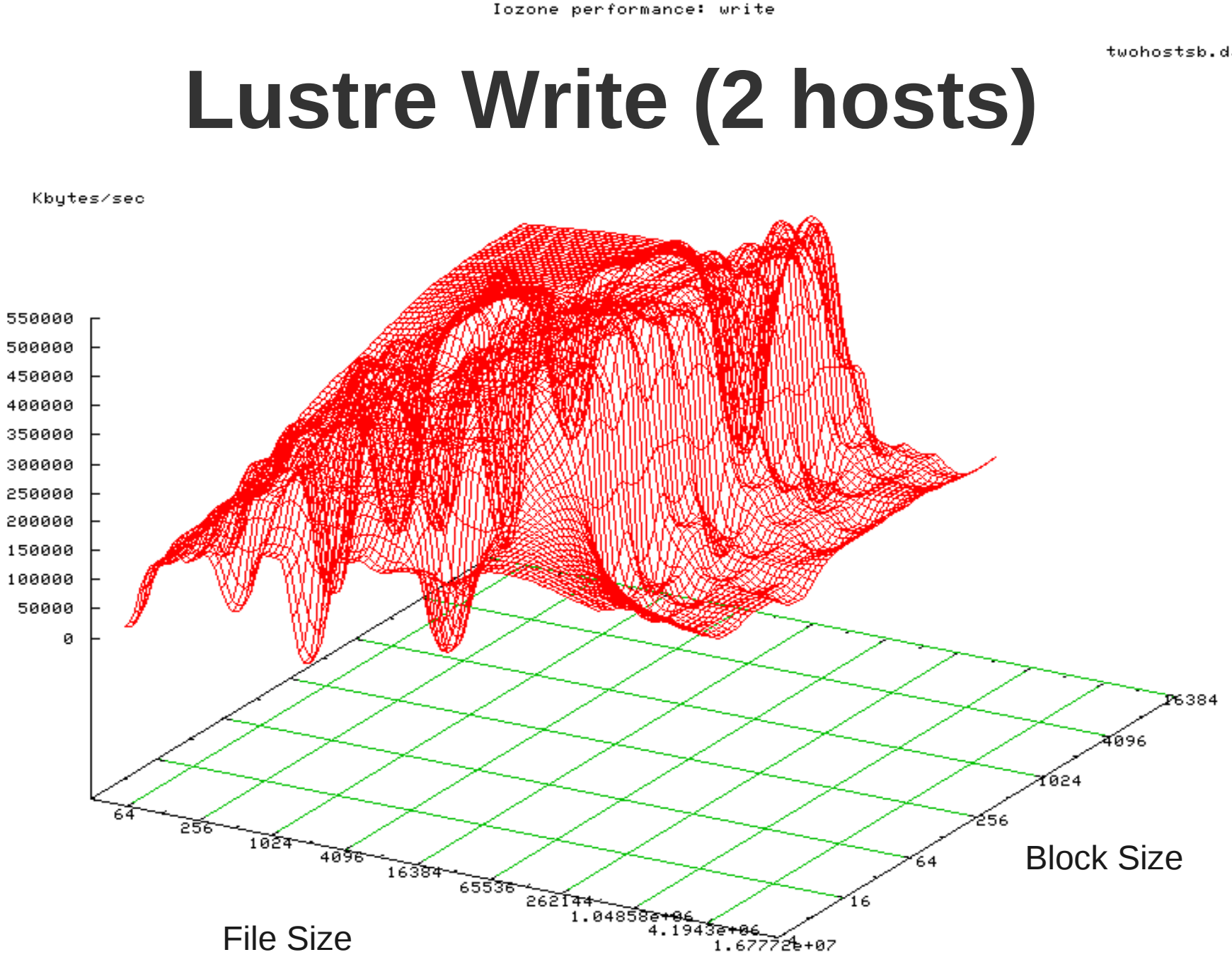


- SATA

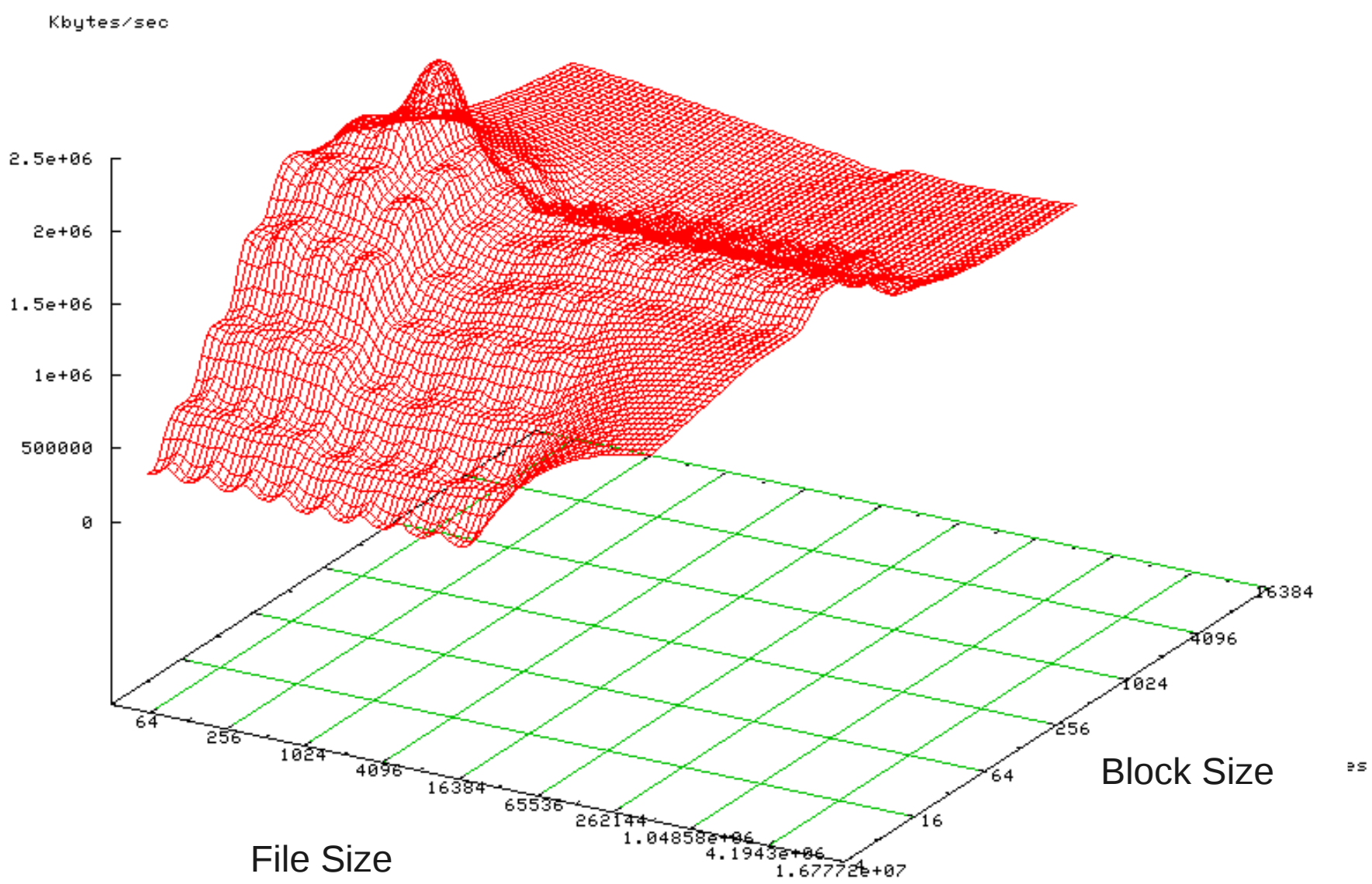




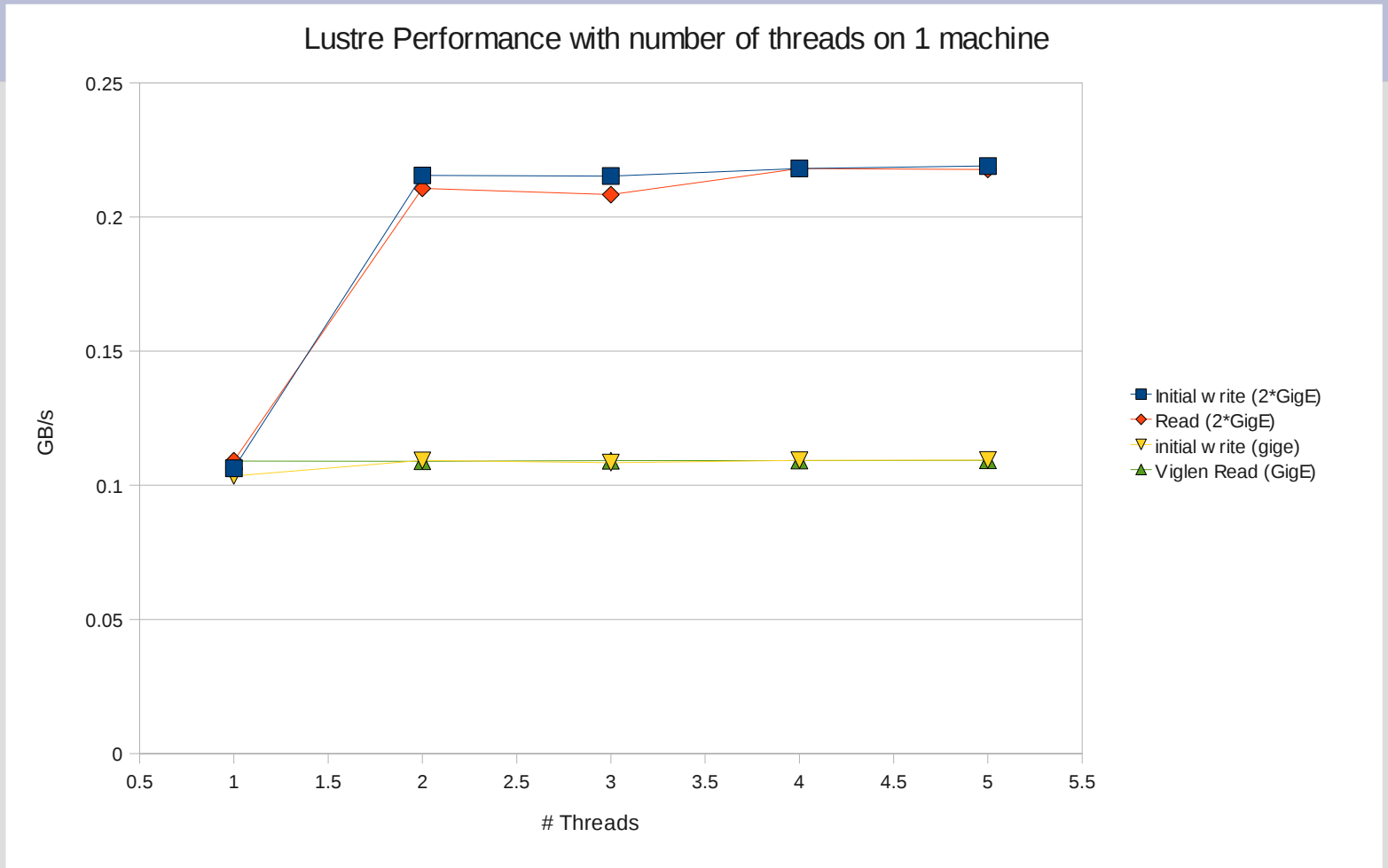
Lustre Write (2 hosts)



Lustre Read (2 hosts)

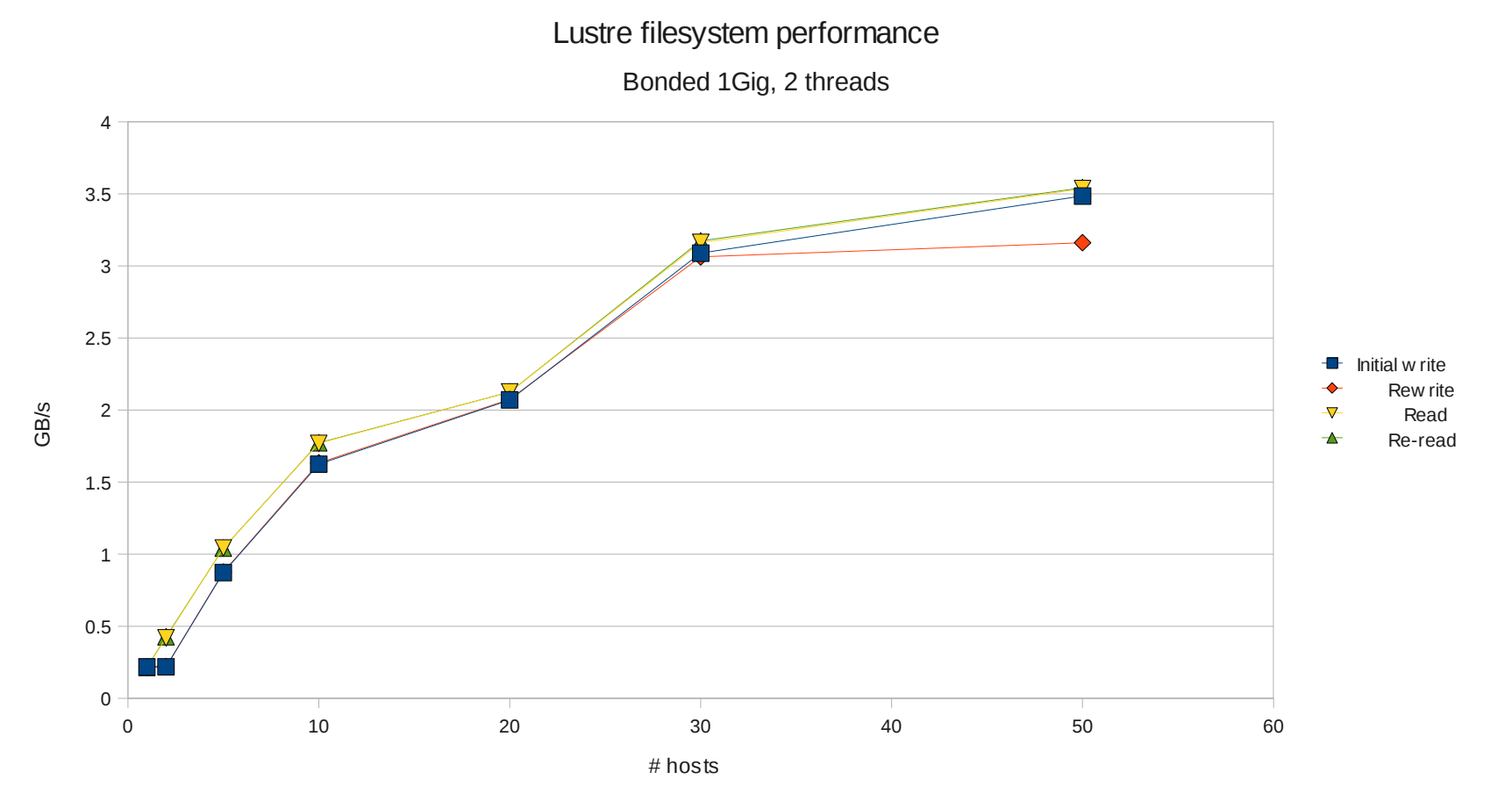


Number of Threads



- 1MB transfer size
- 0.2 GB/s max transfer (network limited)

Number of machines



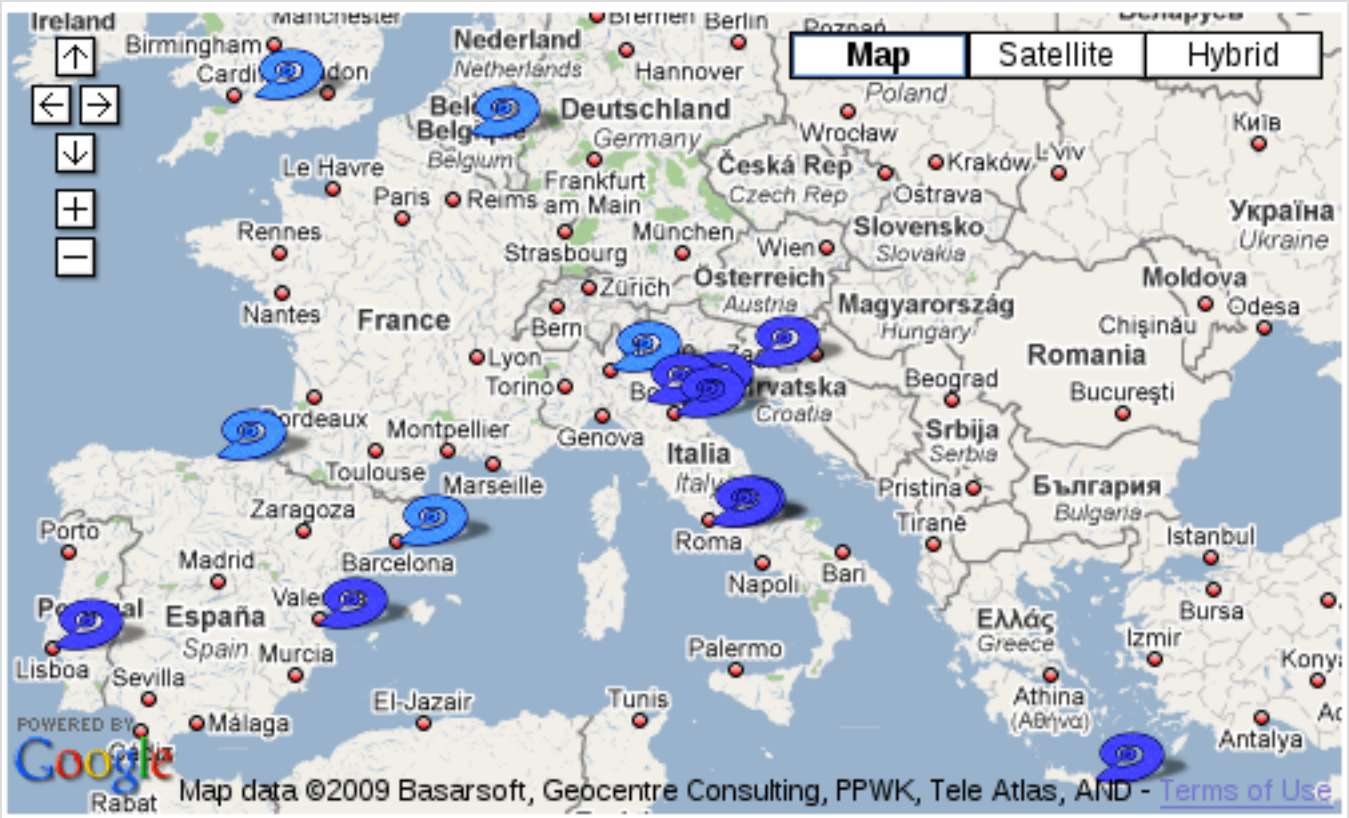
- 2 Threads, 1MB block size
- 3.5 GB/s max transfer
- Probably limited by network to racks used

Storm

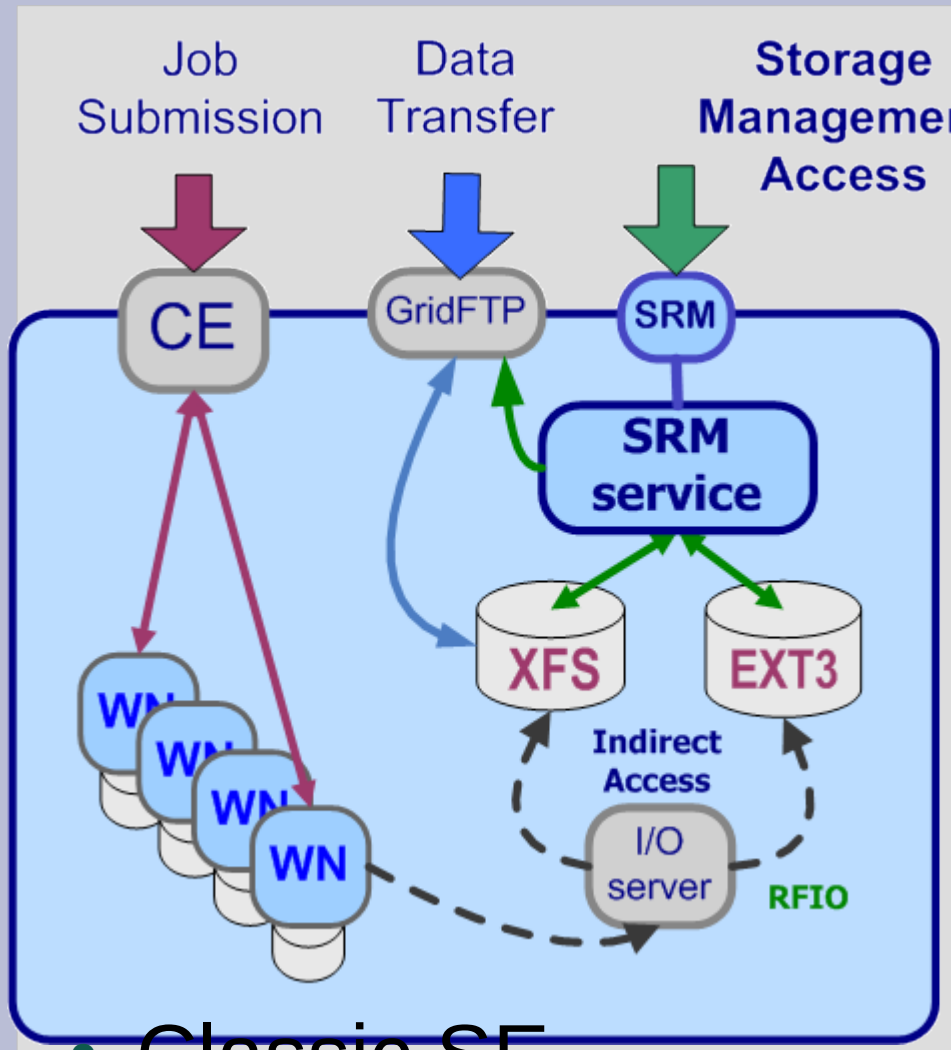
- SRM implementation
  - Light
  - Scalable
  - Flexible
  - High-performance
  - file system independent
- file:// Protocol
  - Cluster filesystem performance

Storm sites

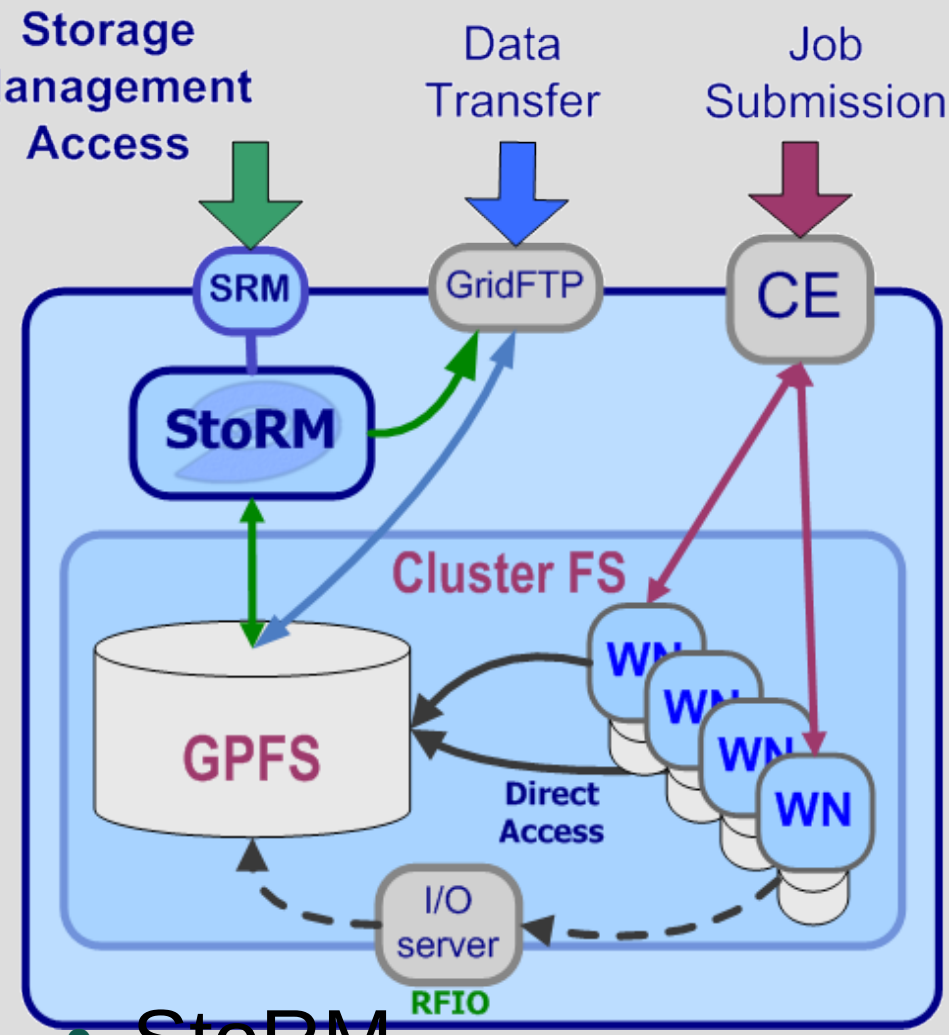
- 22 sites
  - T1 CNAF
  - 19 WLCG T2 + 2 others
    - 14 Italy
    - 2 UK
    - 2 Portugal
    - 2 Spain
    - 1 Israel
    - 1 Greece
  - Lustre and GPFS



Storm Architcture I

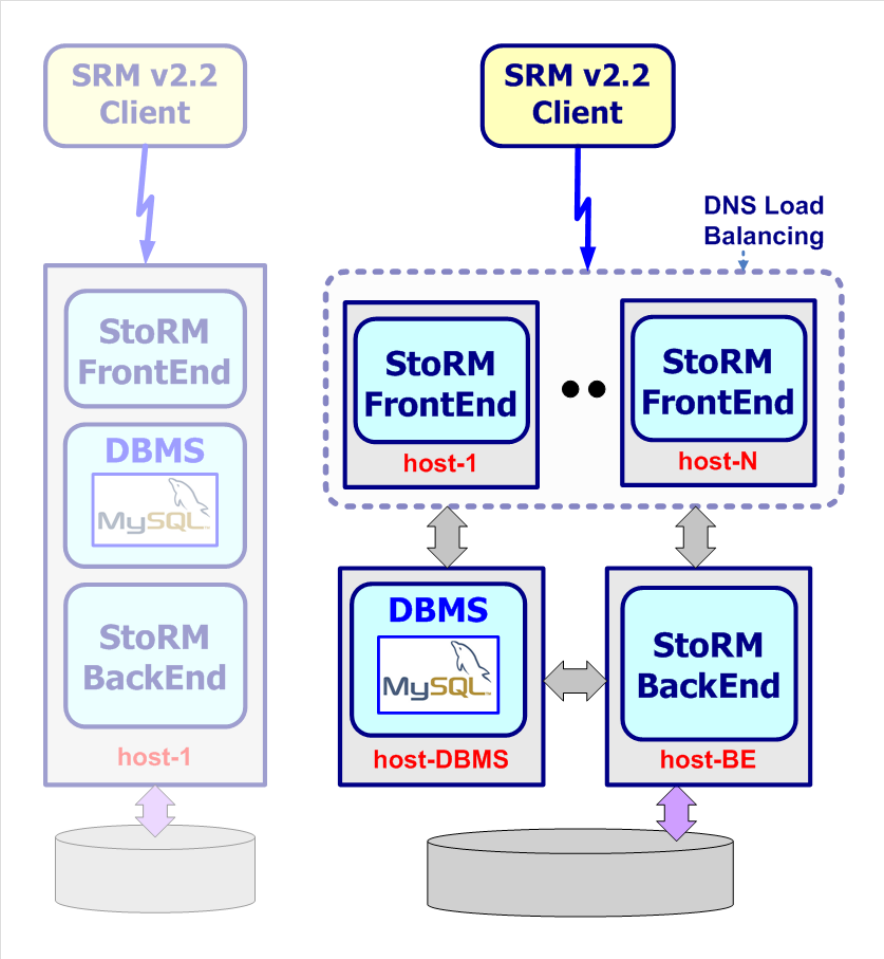


- Classic SE



- StoRM

Storm Architecture II

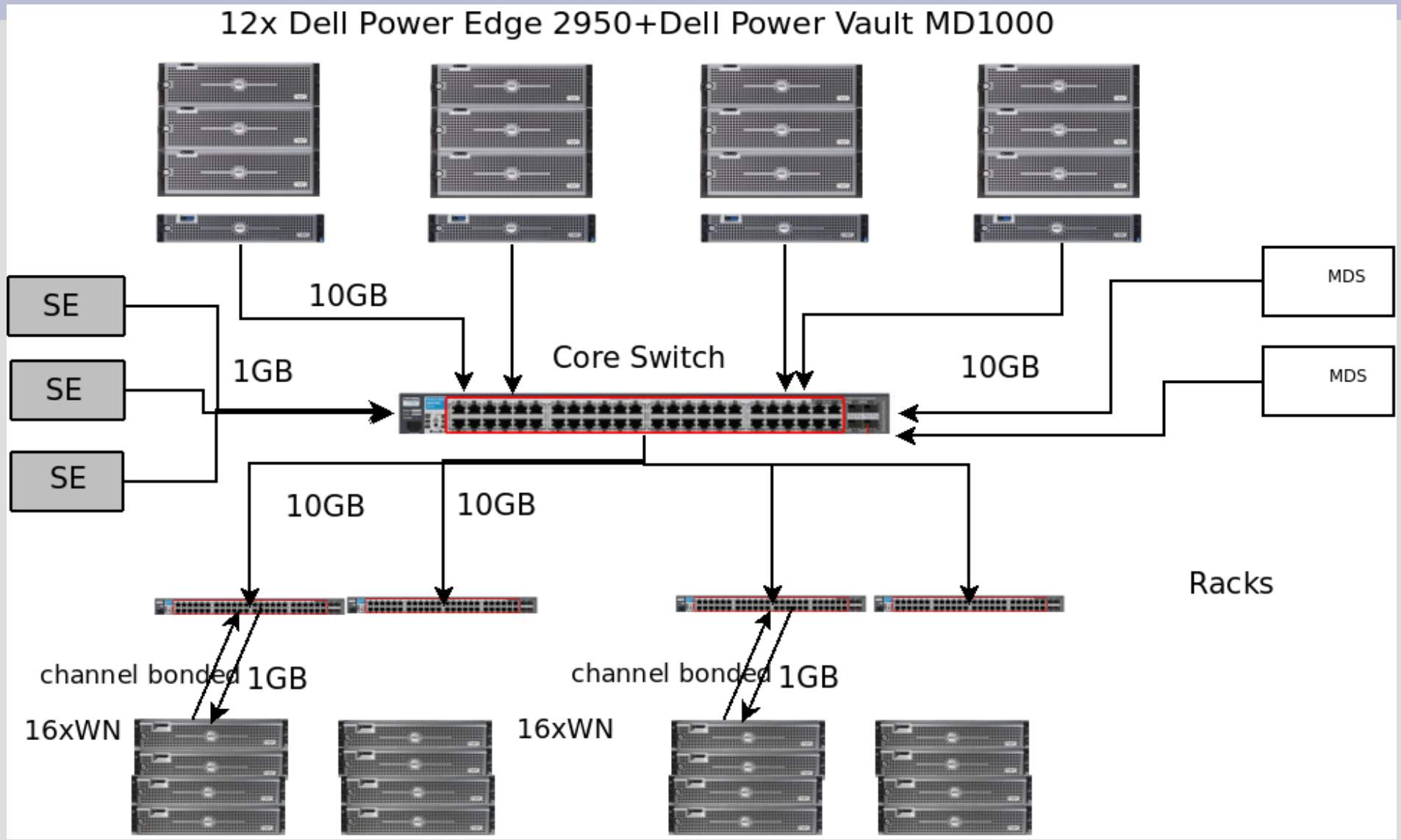


- Servers
  - Multiple Frontends
  - Multiple GridFTP
  - Single backend
  - Single database
- Access control
  - Posix ACLs
    - AoT (LCG)
    - JiT (Finance)

Space Authorisation

- **Space Auth DB**
  - ace.2=dn:/O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan:S:ALLOW
  - ace.3=fqan:EVERYONE:RQ:ALLOW
  - ace.4=fqan:EVERYONE:S:DENY
- **Code Name**
  - D RELEASE SPACE
  - U UPDATE SPACE
  - R READ FROM SPACE
  - W WRITE TO SPACE
  - S STAGE TO SPACE
  - C REPLICATE FROM SPACE
  - P PURGE FROM SPACE
  - Q QUERY SPACE

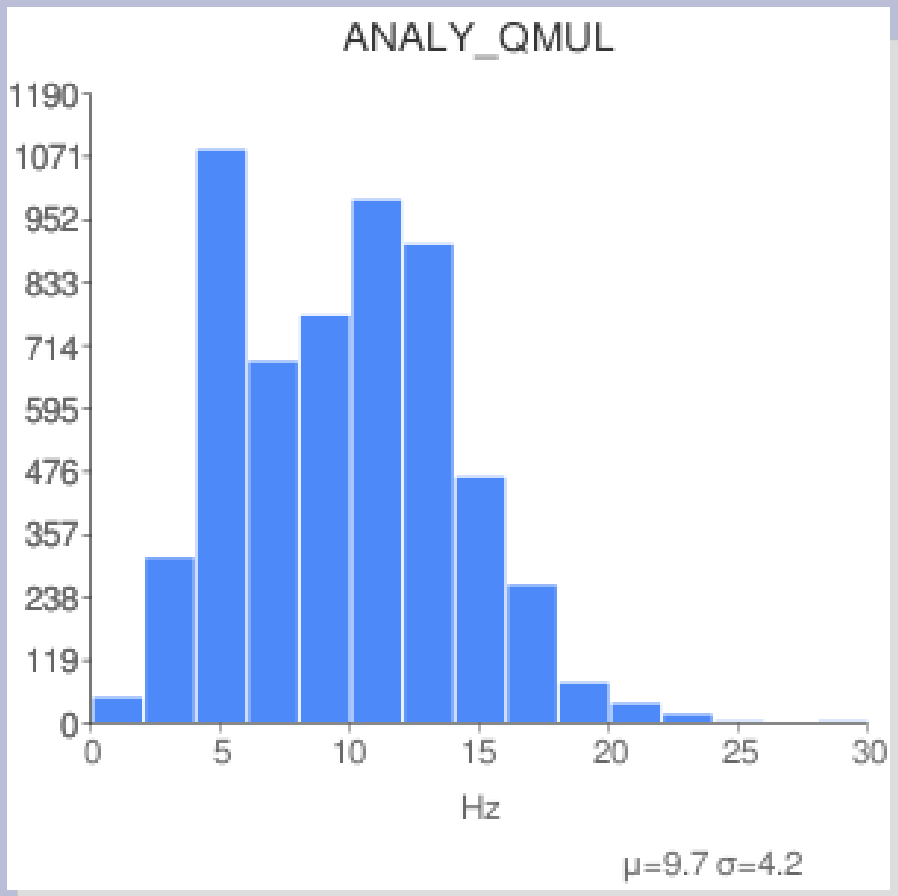
QMUL Network

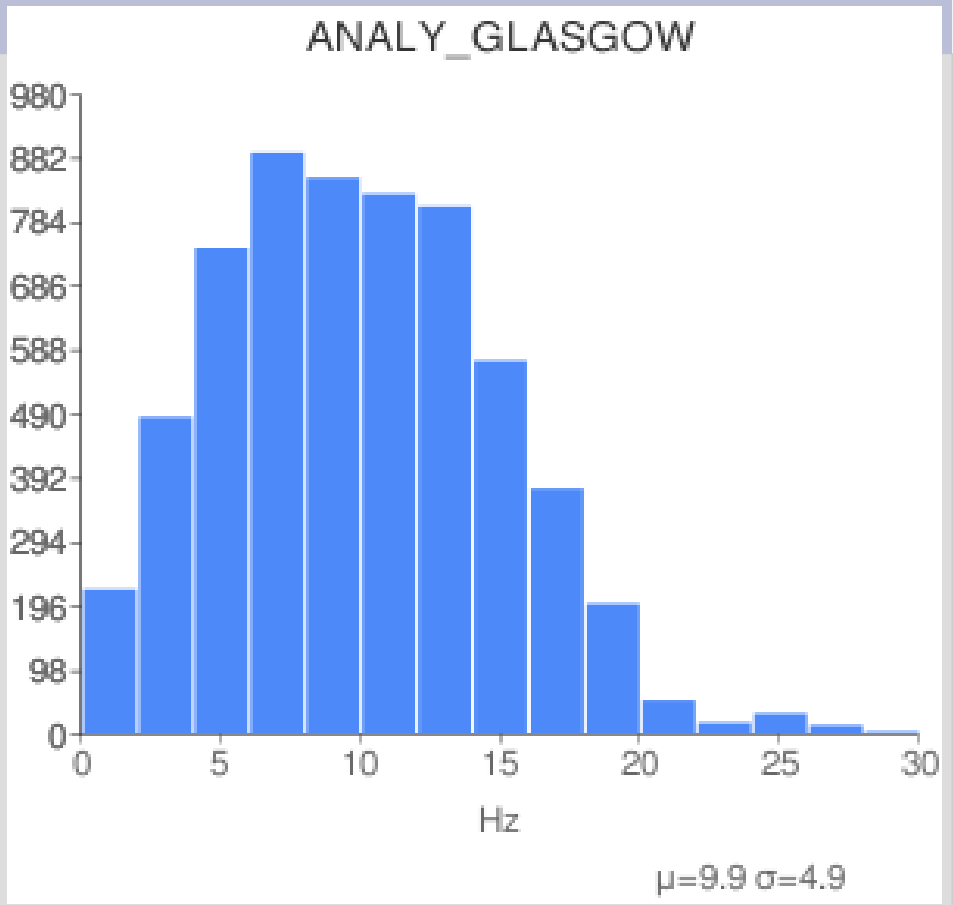


QMUL's experience

- Consistent DN->userid mapping
- Easy to install
- Reliable
- file:// (necessary for internal traffic)
  - Enforceable in future

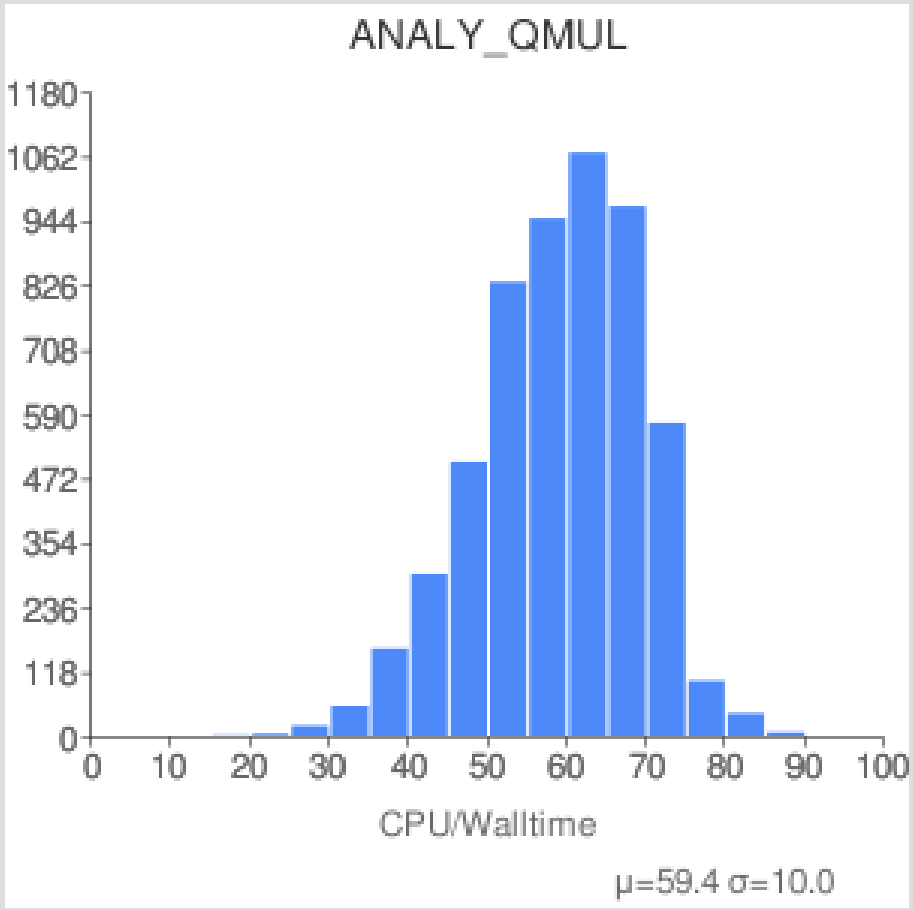
Hammercloud I

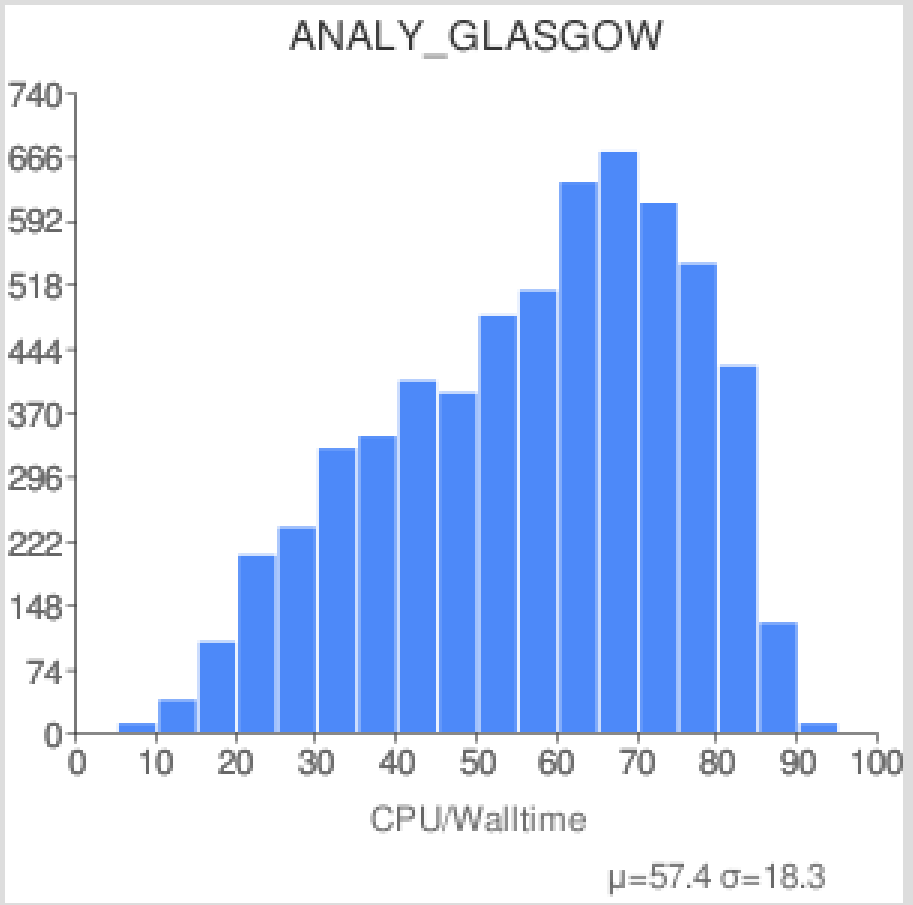




Hz limited by job submission

Hammercloud tests Ib





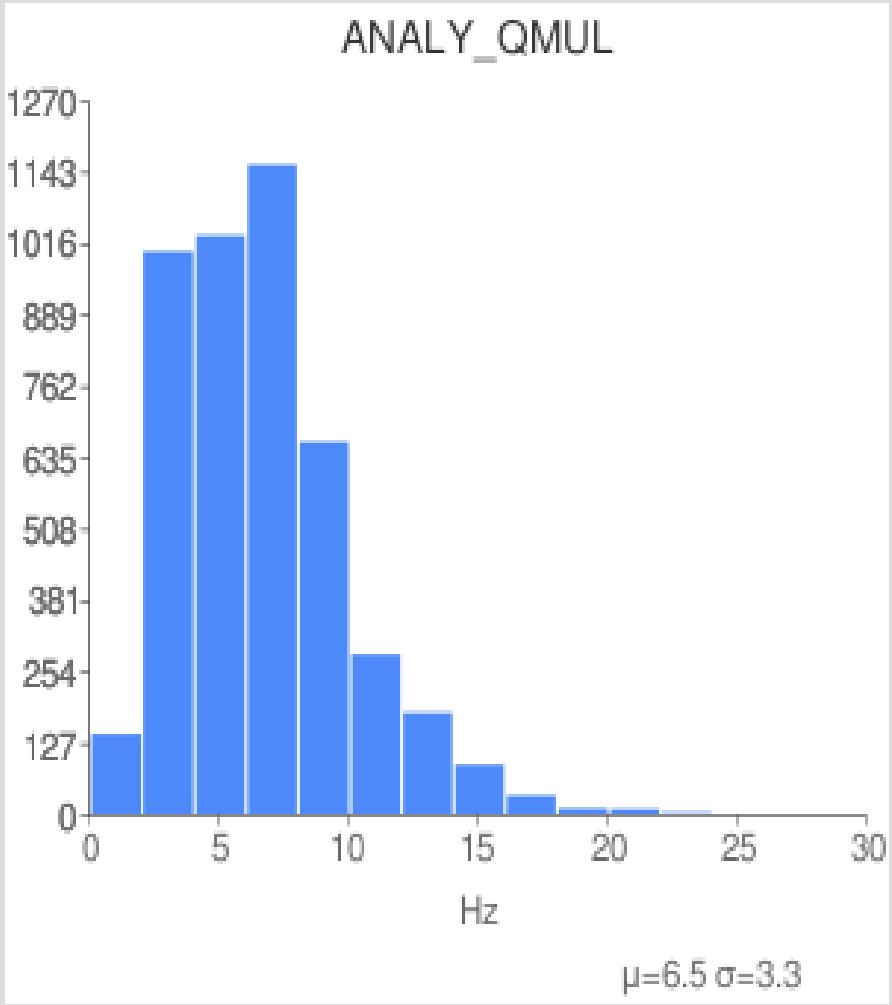
Hammercloud I Files

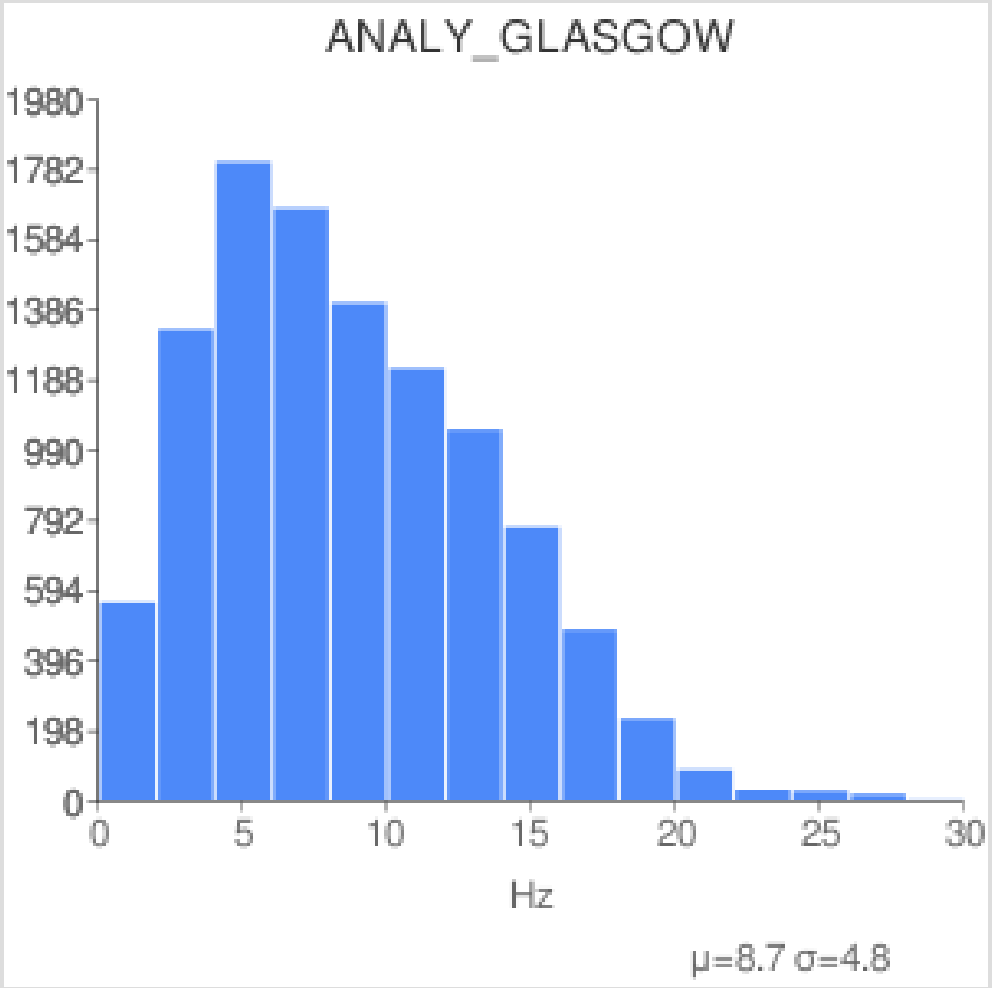
SITE	# PROCESSED	# EXPECTED
ANALY_BHAM	812	812
ANALY_CAM	4320	4320
ANALY_GLASGOW	22520	22520
ANALY_LANCS	2372	2372
ANALY_LIV	8638	8638
ANALY_MANC1	378	378
ANALY_MANC2	729	729
ANALY_QMUL	22163	22163
ANALY_RALPP	2165	2165
ANALY_RHUL	8338	8338
ANALY_SHEF	3611	3611

Hammercloud I Events

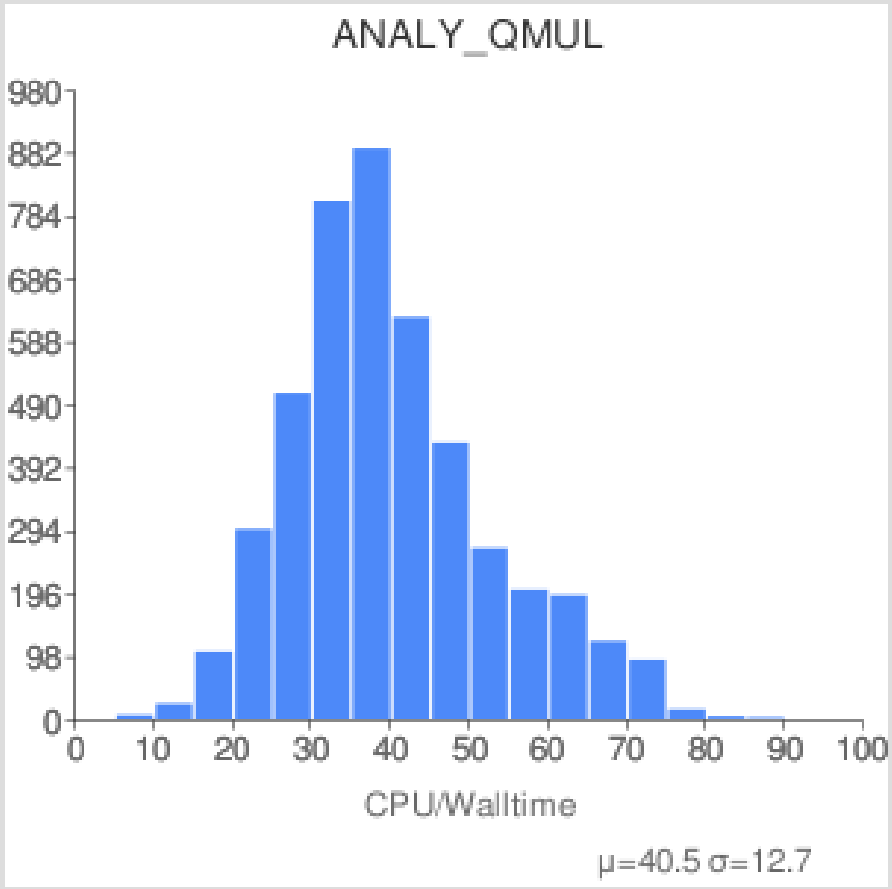
- ANALY_BHAM 7722240
- ANALY_CAM 42554378
- ANALY_GLASGOW 2054139902
- ANALY_LANCS 21403061
- ANALY_LIV 80665930
- ANALY_MANC1 3262760
- ANALY_MANC2 5949309
- ANALY_OX 0
- ANALY_QMUL 9264112625L
- ANALY_RALPP 20288369
- ANALY_RHUL 274296367
- ANALY_SHEF 31567802

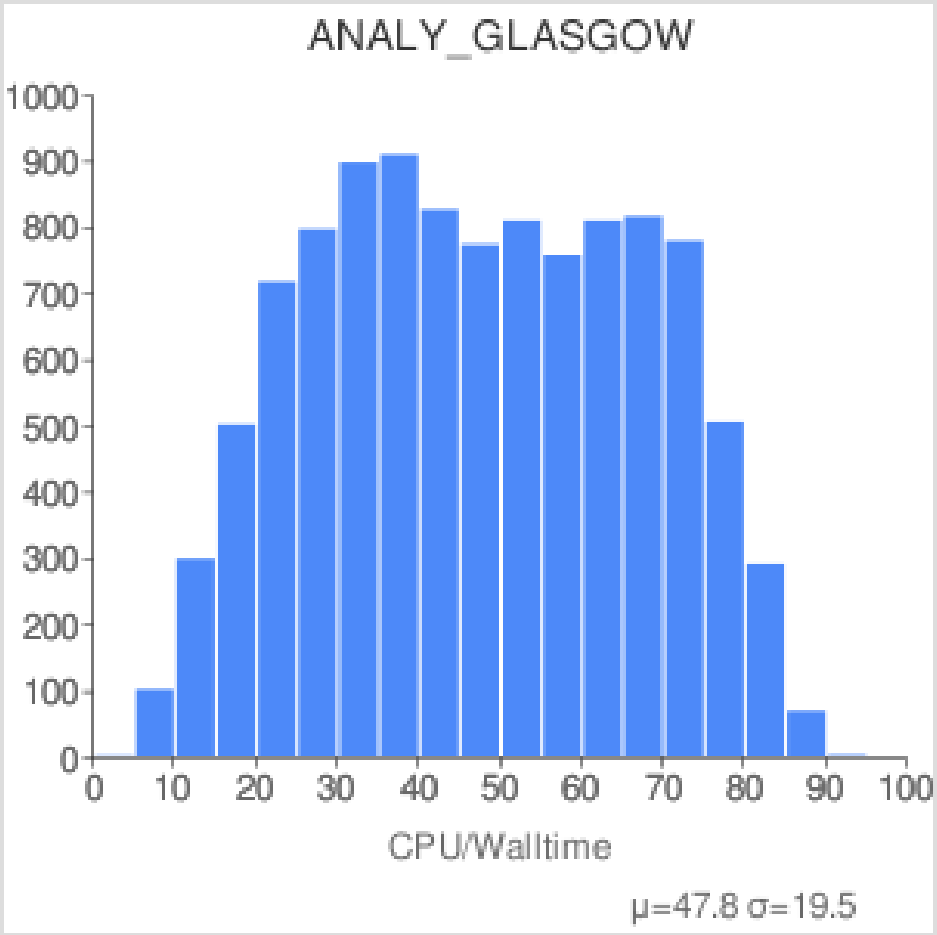
Hammercloud Tests II





Hammercloud Tests IIb





Hammercloud summary

- Preliminary results good
- Performance below iozone tests
- Potential improvements
  - Copy to WN?
  - Lustre 1.8?
  - Different disk layout
- Open Questions
  - Hot file identification
  - What to do if OSS is down.

Conclusions

- Lustre Performance excellent
- StoRM/Lustre – Good performance
- Need to use file:// protocol
  - Easier in future versions of StoRM
- More measurements necessary