Experiences with SFS/Lustre in Multi-user Production Environments

Roland Laifer

Scientific Supercomputing Centre (SSCK)
University of Karlsruhe
Germany

Laifer@rz.uni-karlsruhe.de

Outline

SSCK's SFS production systems

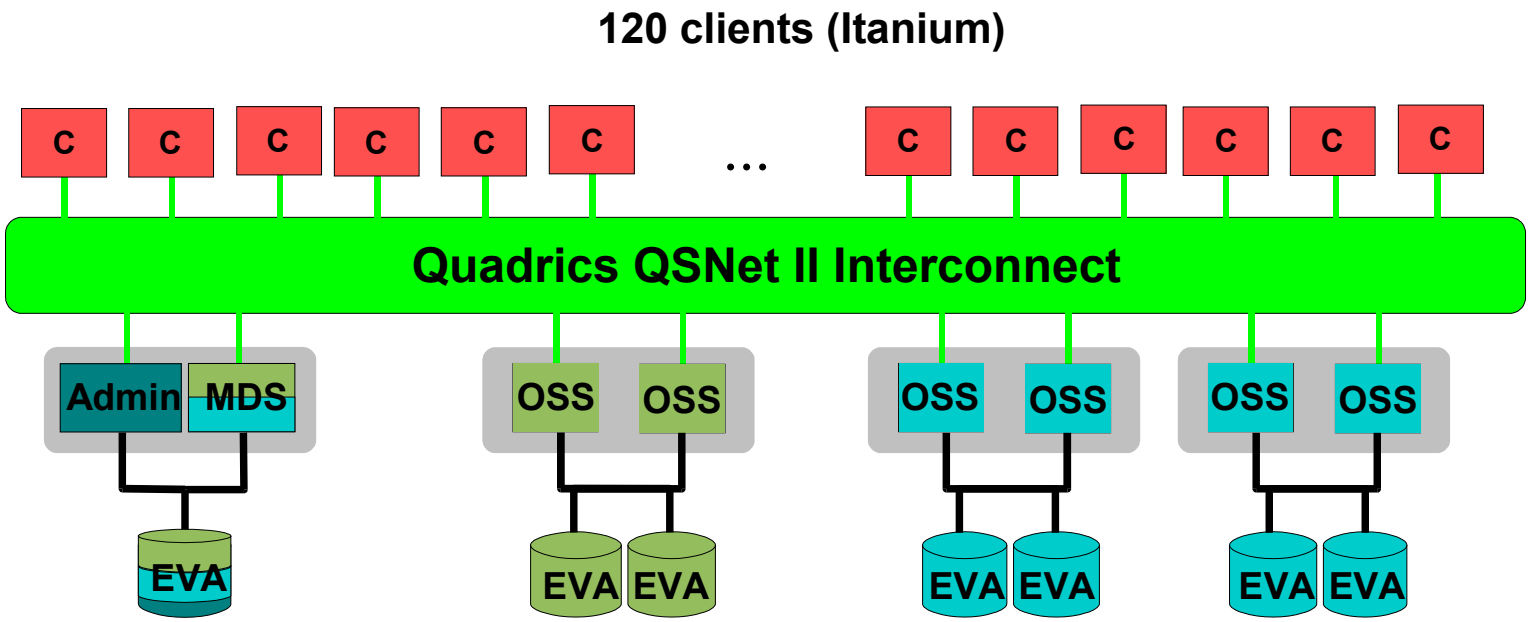
Performance monitoring graphs

Performance measurements

Operational experiences

Some open problems

Itanium production system (xc1)

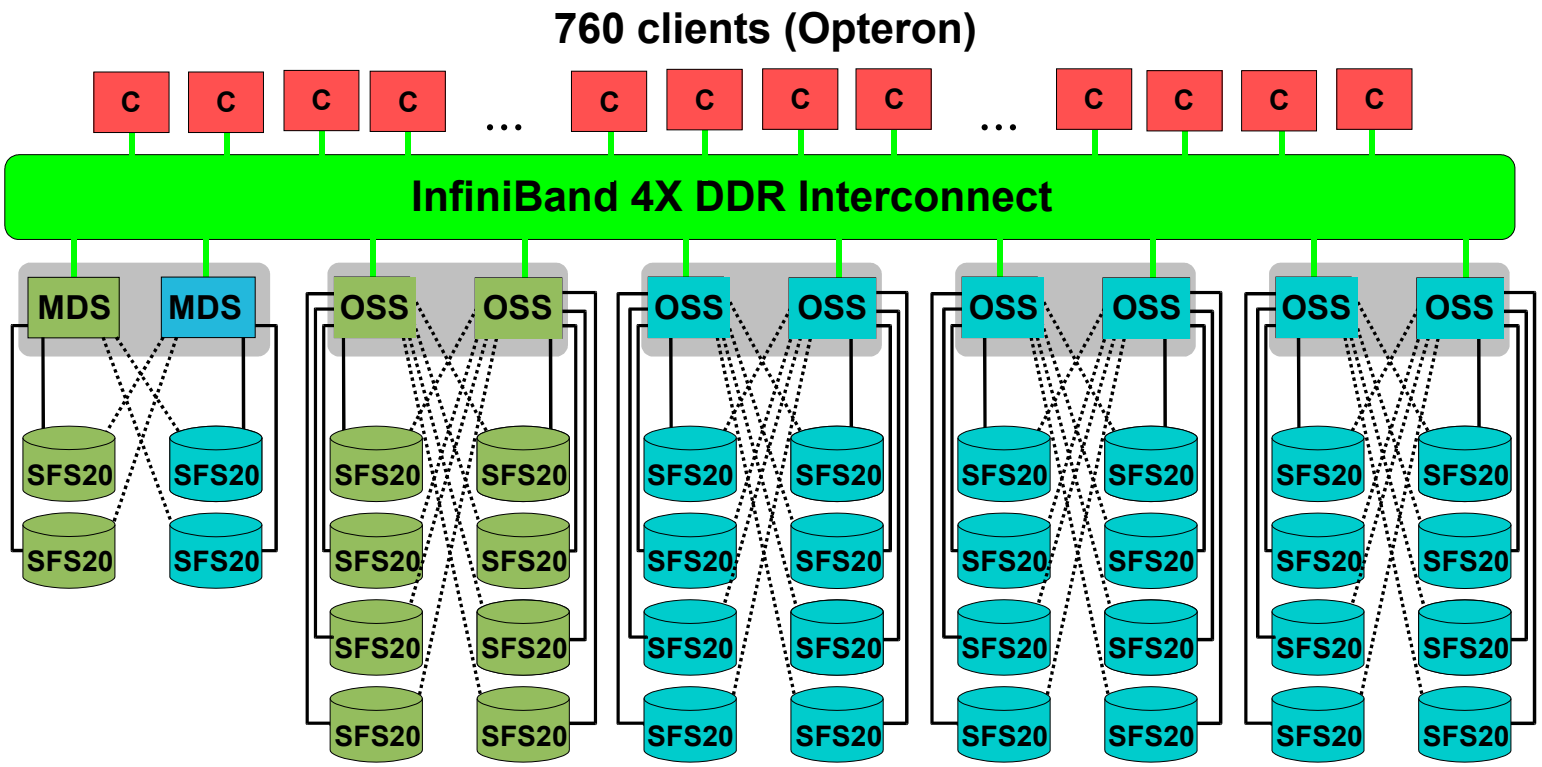


	$HOME	$WORK
Capacity	3.8 TB	7.6 TB
Write performance	280 MB/s	560 MB/s
Read performance	380 MB/s	760 MB/s

Notes:

- **In production since January 2005**
- **Good experiences using Lustre for home directories**

Opteron production system (xc2)



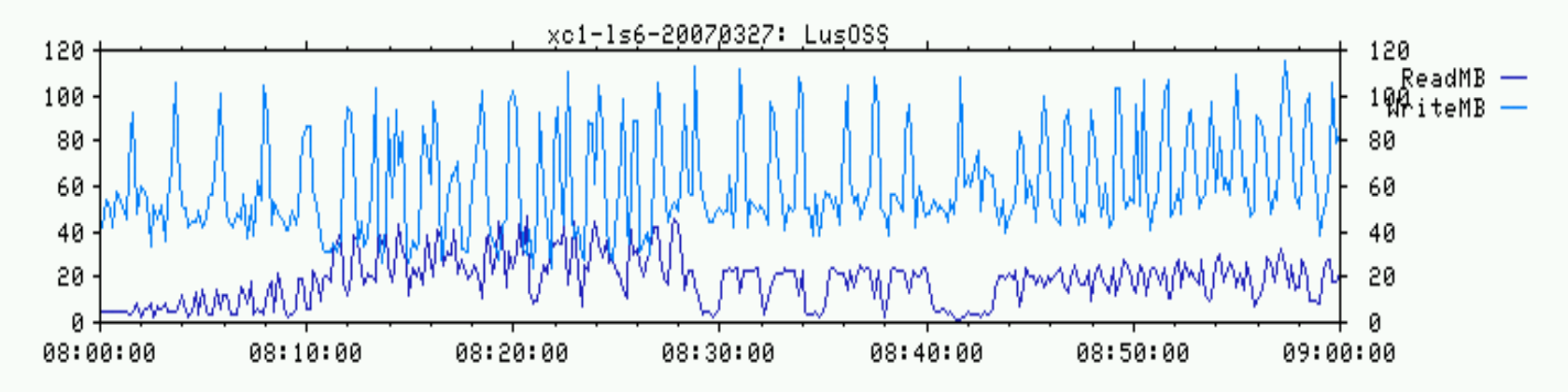
	$HOME	$WORK
Capacity	8 TB	48 TB
Write performance	360 MB/s	2100 MB/s
Read performance	800 MB/s	3000 MB/s

Notes:

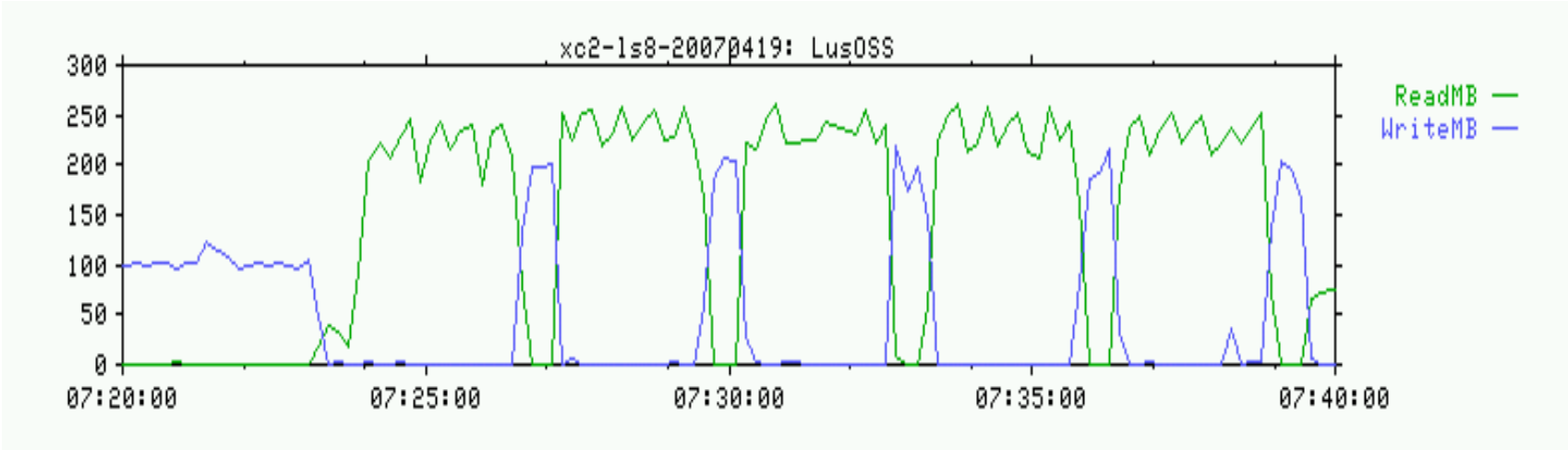
- **In production since January 2007**
- **$HOME file system uses mirrored OST luns**
- **Performance values were measured with new file systems**

Performance monitoring graphs for one OSS

xc1:

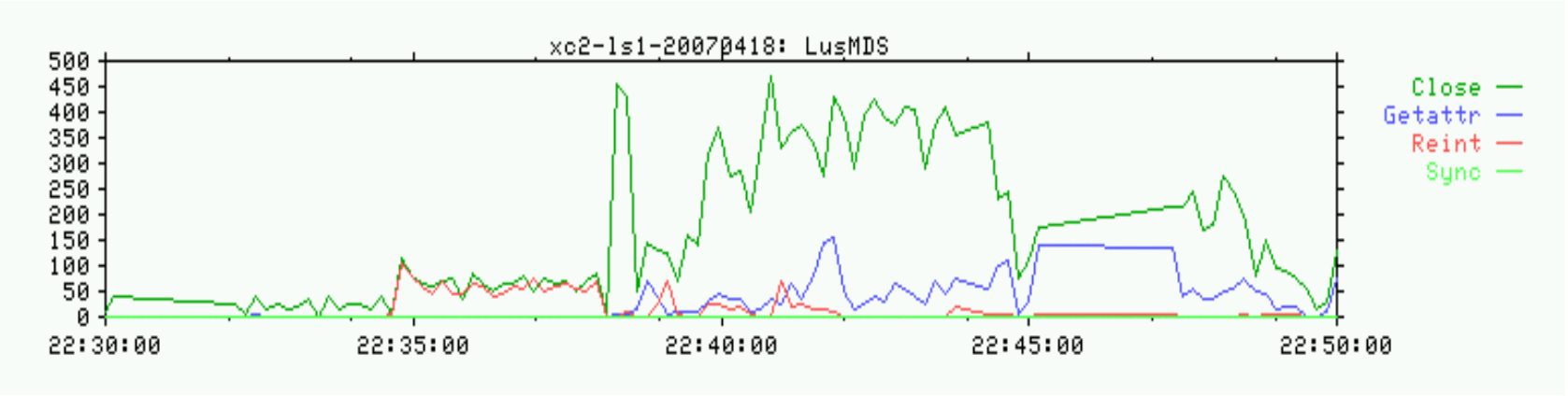


xc2:

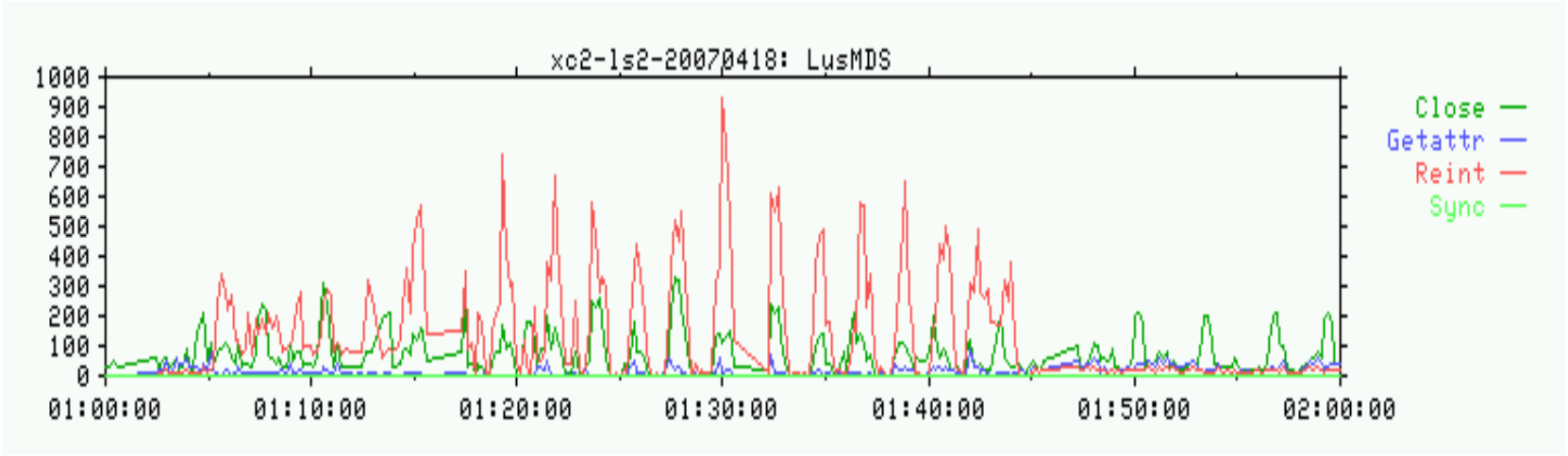


Performance monitoring graphs for one MDS

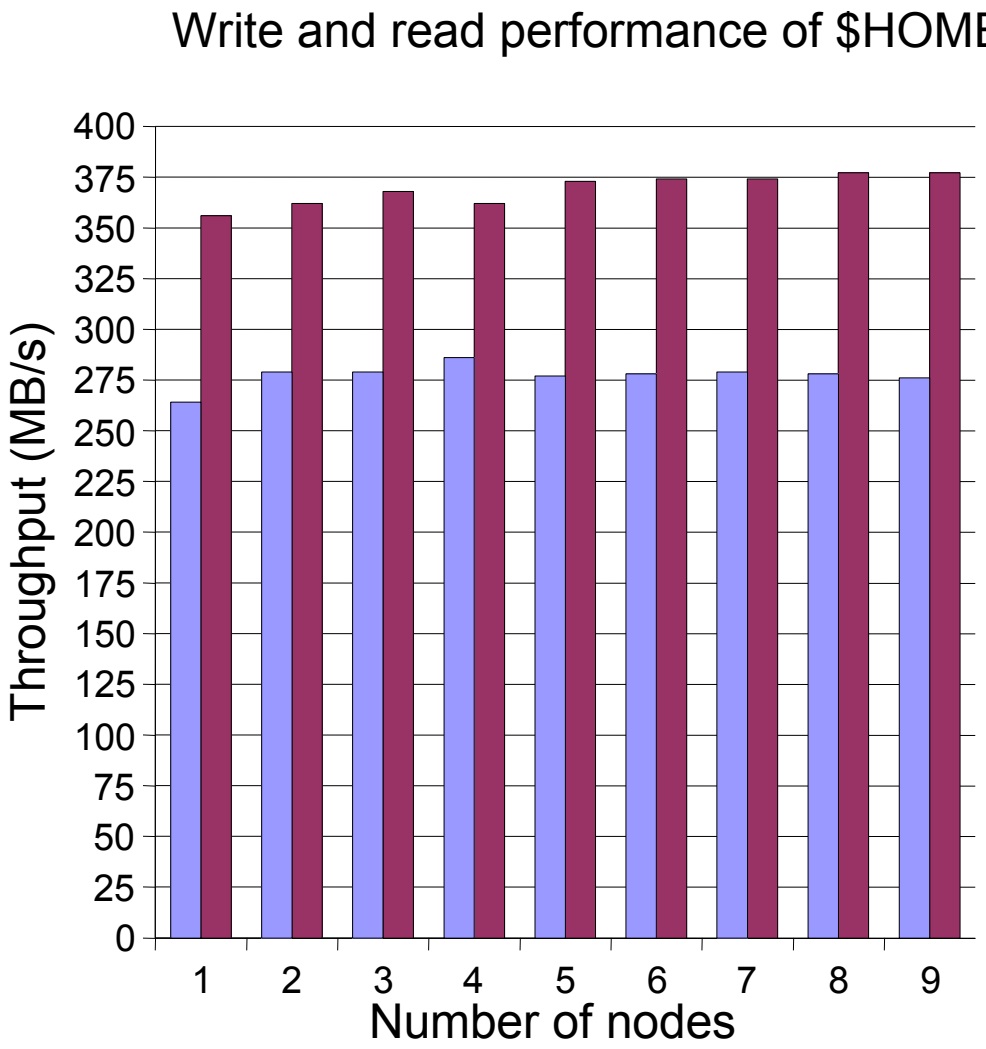
MDS for $HOME on xc2:



MDS for $WORK on xc2:



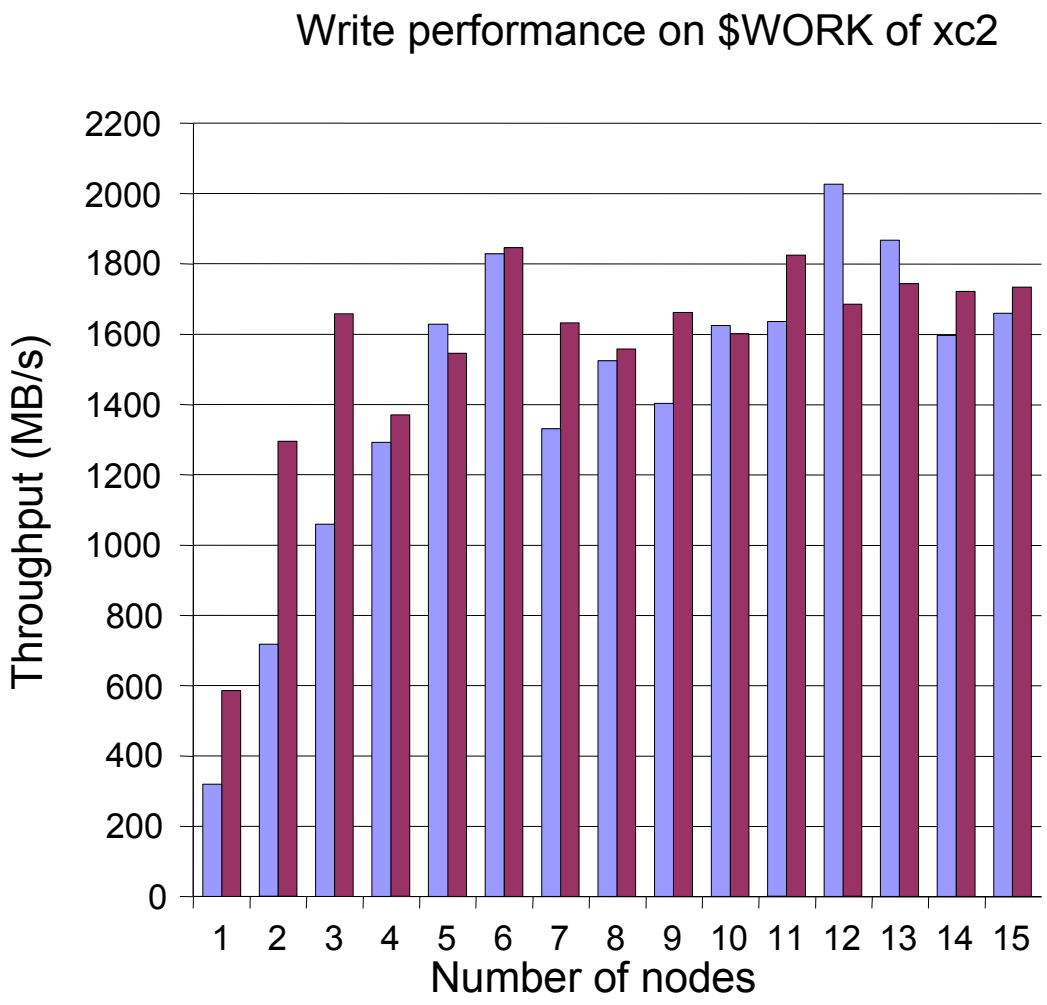
Performance measurement with parallel dd on xc1



Note:

- **Omit caching effects by reading a file which was created on a different node**

Performance measurement with parallel dd on xc2

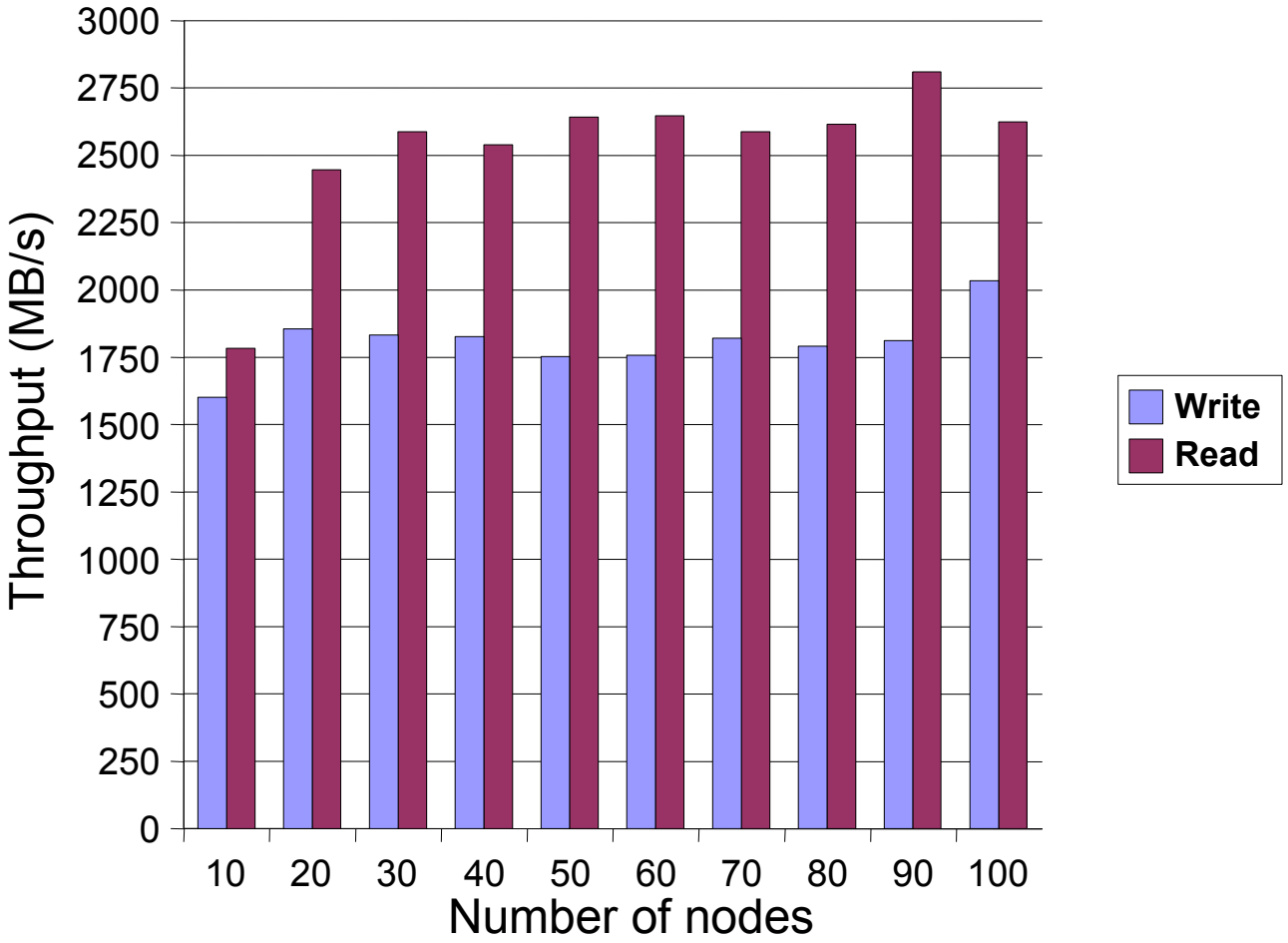


Notes:

- **Each process used its own file with stripe count 4**
- **Peaks when all 24 OST luns were used**

Performance scalability of $WORK on xc2

Write and read with 2 processes per node on $WORK of xc2



Operational experiences

Very high I/O usage rates

- File system work on xc2 was saturated for months
  - Several new users use I/O throughput with different applications
- Some users have millions of files
  - Metadata intensive commands might be slow due to high load on the OSS

HP SFS runs pretty stable

- We never lost data
  - Use RAID6 whenever possible
- Systems usually ran for months without a problem
  - After months Lustre bugs and a hanging I/O subsystem appeared
- Lustre works pretty good on an unstable network
  - During initial xc2 system test InfiniBand was not very stable

Some open problems

Performance degradation on xc2

- After 6 months of production we lost half of the file system performance
  - Problem is under investigation by HP
  - We had a similar problem on xc1 which was due to fragmentation
  - Current solution for defragmentation is to recreate file systems

Quotas not decreased after deleting files

- Happens sometimes after setting quota limit to a too low value
  - Problem is under investigation by HP, recreation is not easy
  - We also had other problems with quotas which were fixed

Summary

Lustre provides a scalable and stable parallel file system

HP SFS supplies additional features

- **This makes it a real product**

Investigation of performance problems is not easy

Further experiences with HP SFS:

- **http://www.rz.uni-karlsruhe.de/dienste/lustretalks**