Experiences with HP SFS / Lustre in HPC Production

Roland Laifer

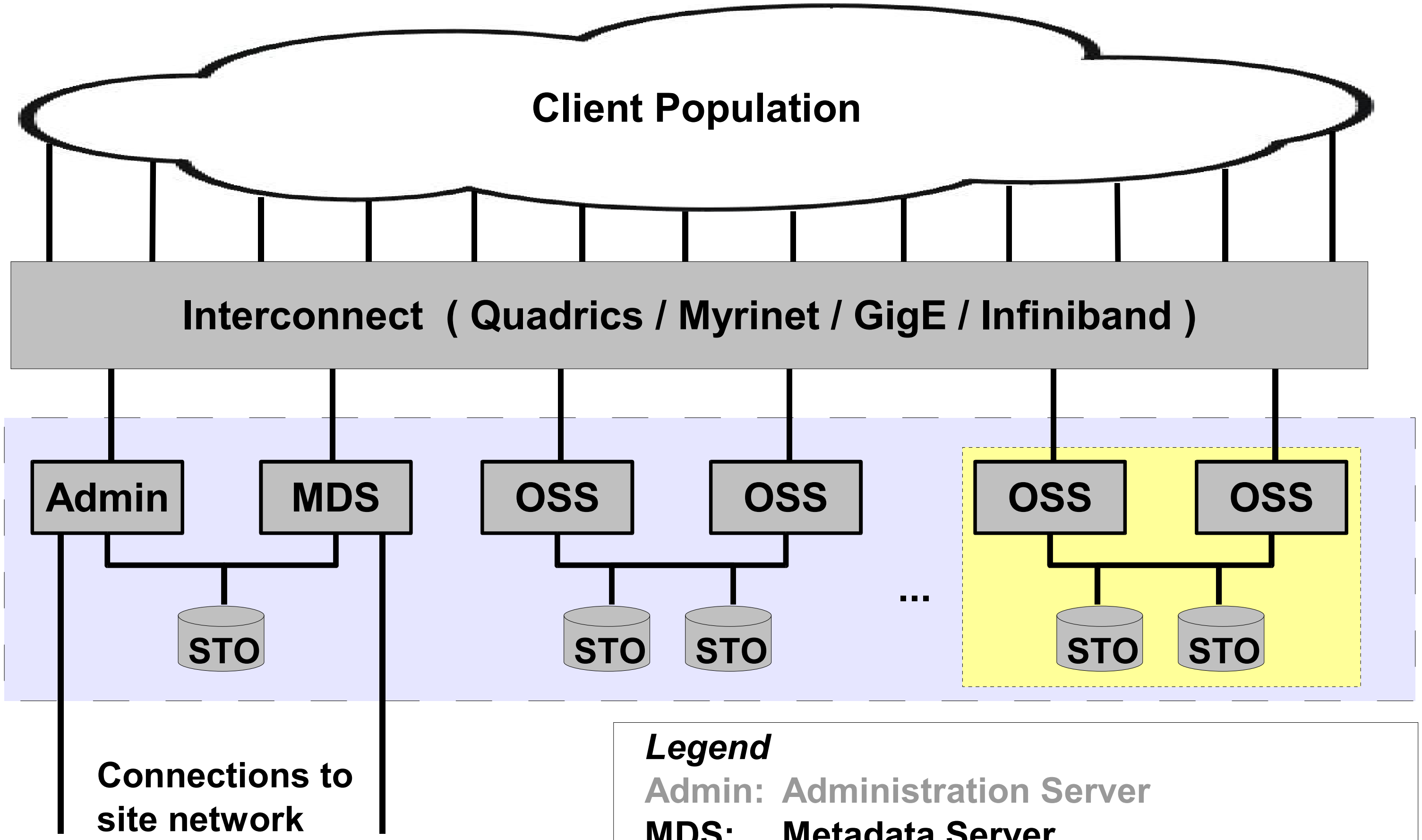
Computing Centre (SSCK)
University of Karlsruhe

Laifer@rz.uni-karlsruhe.de

Outline

- **What is HP StorageWorks Scalable File Share (HP SFS)?**
  - **A Lustre product from HP**
    - **available since December 2004**
- **Performance measurements**
  - **depending on underlying hardware at SSCK**
- **Experiences with HP SFS**
  - **SSCK has one of the first Lustre production installations in Europe**

HP SFS system architecture


Client Population
Interconnect (Quadrics / Myrinet / GigE / Infiniband)
Admin
MDS
OSS
OSS
OSS
OSS
STO
STO
STO
STO
STO
...
Connections to site network

Legend

Admin: Administration Server

MDS: Metadata Server

OSS: Object Storage Server

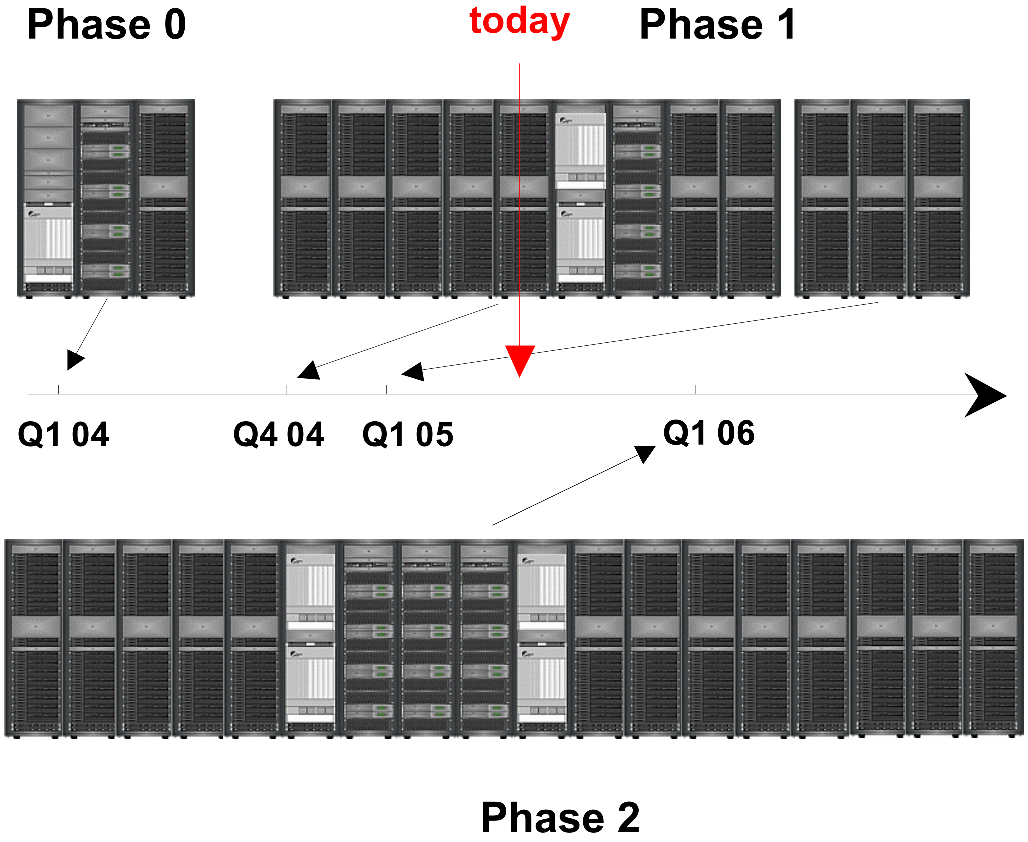
STO: Dual connected storage subsystem:

- either EVA3000 storage arrays
- or SFS20 storage arrays

Value added of HP SFS compared to free Lustre

» HP
- selects appropriate hardware
  - Only dedicated hardware is supported on server side
- selects a current Lustre version
  - Freely available Lustre releases become available with up to 1 year delay
- adds additional software for failover and management
  - Both components are not part of free Lustre
  - Management software supplies central point of administration
- runs additional tests and puts patches on top of the code
- delivers software, documentation, and licences
  - Software includes client rpm packages for XC clusters
- supplies support

HP XC 6000 Cluster installation schedule at SSCK



Phase 0 (Q1 2004), Development

- 16 two-way nodes
  - 12 Integrity rx2600
  - 4 ProLiant DL360 G3
  - Single rail QsNet II
- 2 TB storage system

Phase 1 (Q4 2004), Production

- 116 two-way nodes
  - 108 Integrity rx2600
  - 8 ProLiant DL360 G3
  - Single rail QsNet II
- 11 TB storage system

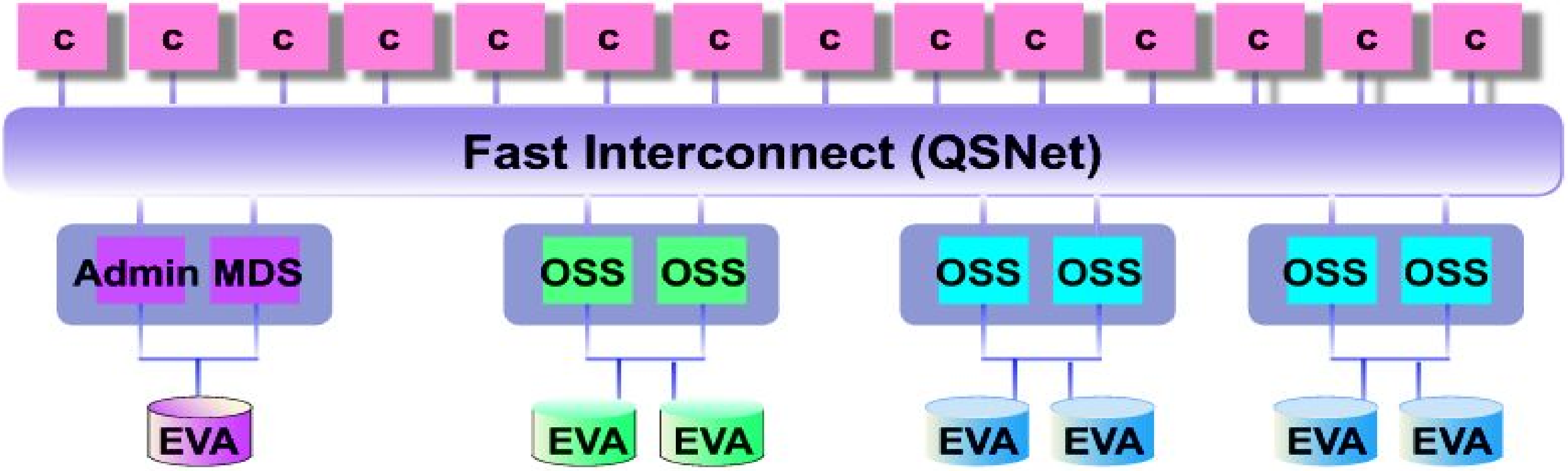
Phase 1 (Q1 2005), Production

- 12 eight-way nodes
  - 6 Integrity rx8620, two partitions
  - Single rail QsNet II

Phase 2 (Q1 2006), Production

- 218 four-way nodes
  - Two sockets
  - Dual core Montecito
  - Single or dual rail QsNet II
- 30 TB storage system

HP SFS on SSCK's HP XC6000



MDS and Admin for $HOME and $WORK

- allows > 50 million files

$HOME

- 3.8 TB storage

$WORK

- 7.6 TB storage

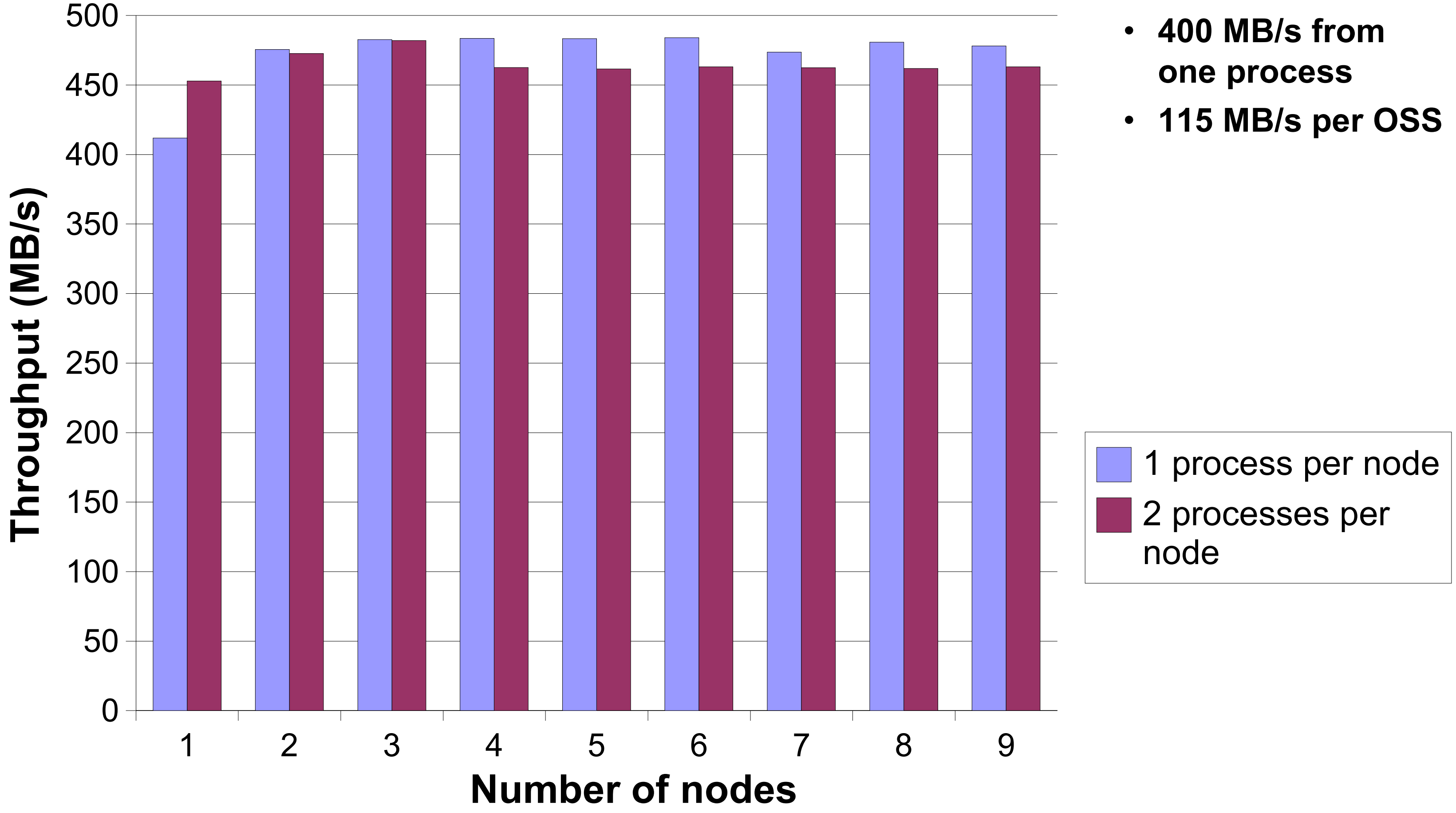
Legend

Admin:	Administration Server
MDS:	Metadata Server
OSS:	Object Storage Server
EVA:	EVA5000 storage array
C:	Client

Performance measurement environment

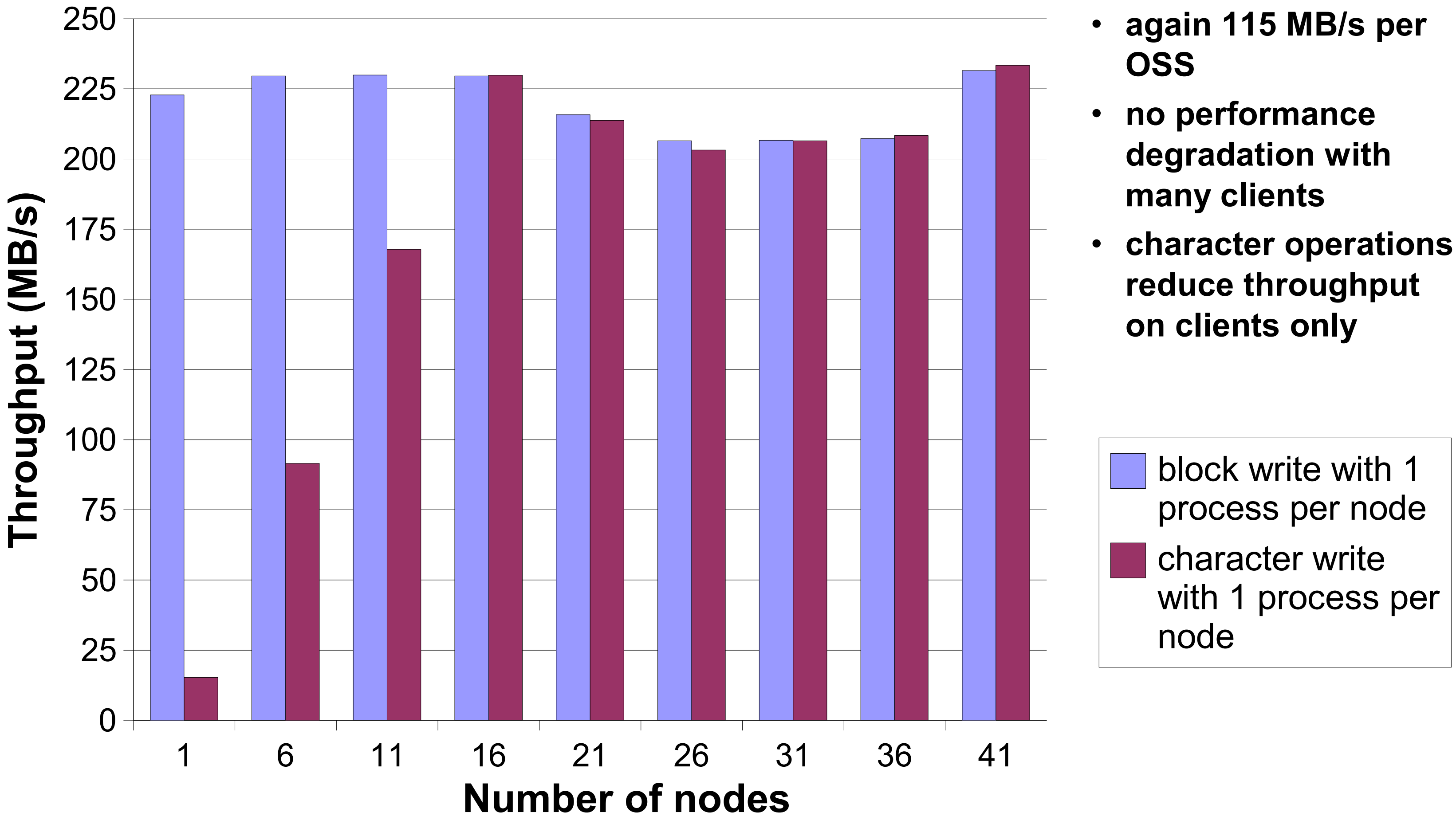
- Used HP SFS software version was 1.1-0
  - Is based on Cluster Filesystem's Lustre version 1.2.6
- Underlying hardware
  - Clients are IA64 systems (rx2600, 1.5 GHz, 2 CPUs, 12 GB memory)
  - Quadrics QsNet II (Elan4) interconnect
  - EVA5000 (not EVA3000) storage systems with 2 controllers
    - OSS disks are 146 GB 10K, MDS disks are 72 GB 15K
  - Servers are IA32 systems (DL360 G3, 3.2 GHz, 2 CPUs, 4/2 GB memory)
    - One file system ($HOME) with 2 OSS and 128 KB stripe size
    - One file system ($WORK) with 4 OSS and 1 MB stripe size
- Performance measurement details
  - Measurements were done in parallel to production
    - Visible impact should be low
  - Benchmarking software was bonnie++

Sequential block write performance with 4 OSS



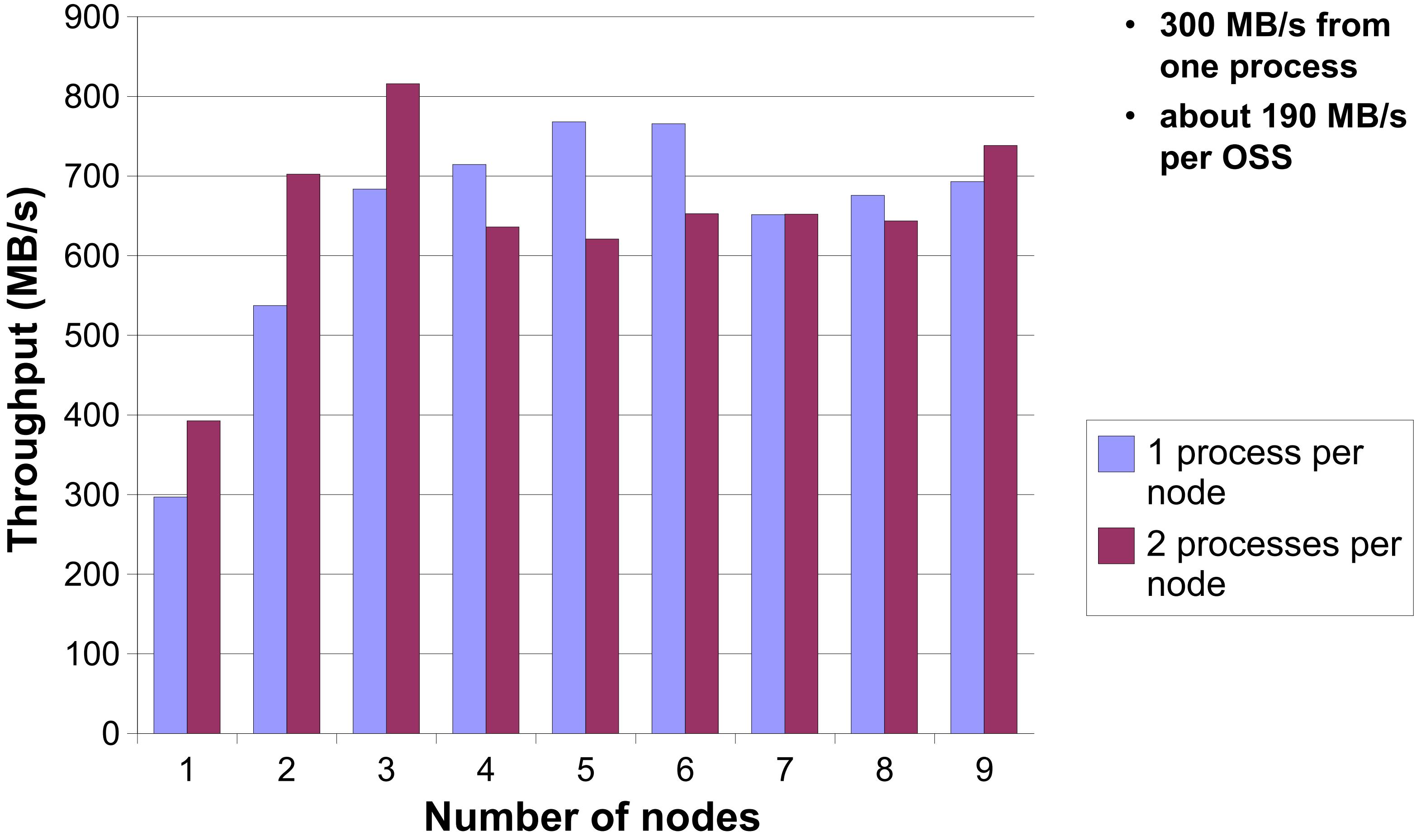
- **400 MB/s from one process**
- **115 MB/s per OSS**

Block vs character write performance with 2 OSS



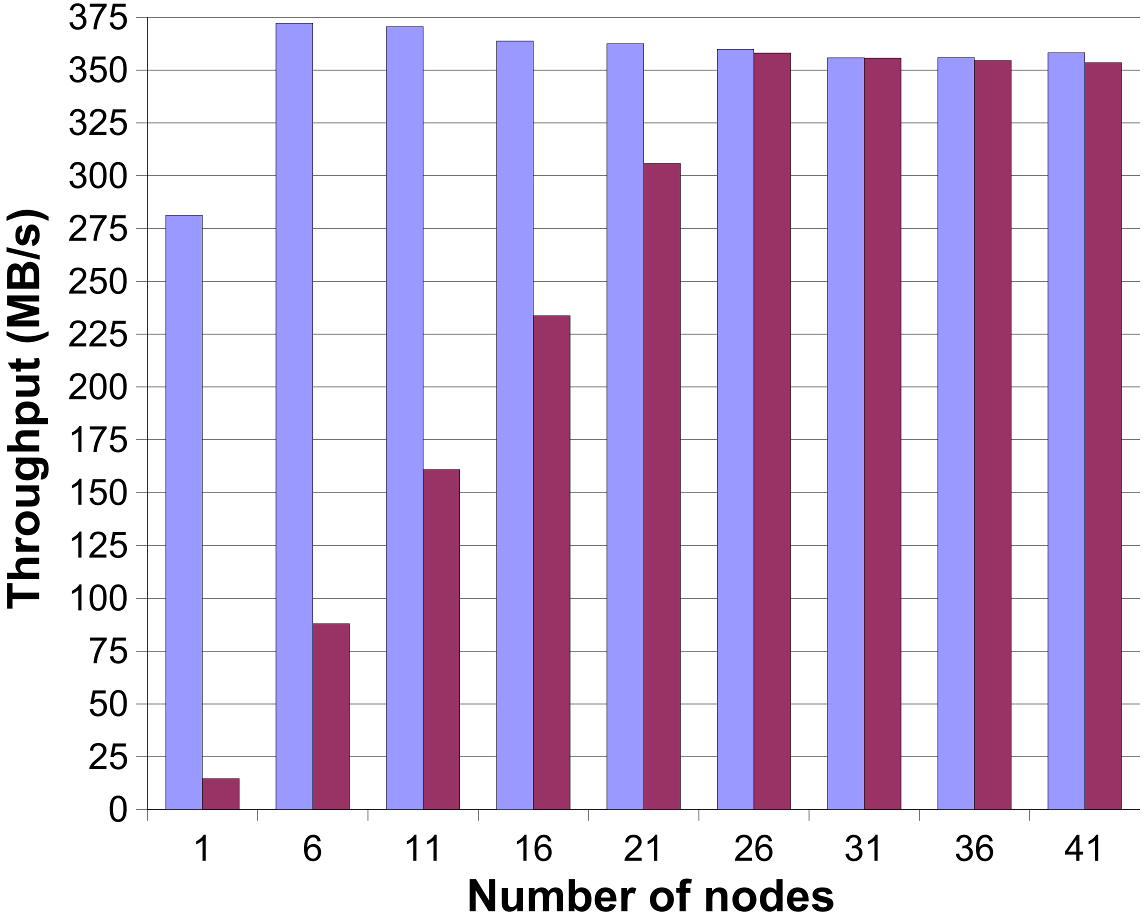
- again 115 MB/s per OSS
- no performance degradation with many clients
- character operations reduce throughput on clients only

Sequential block read performance with 4 OSS

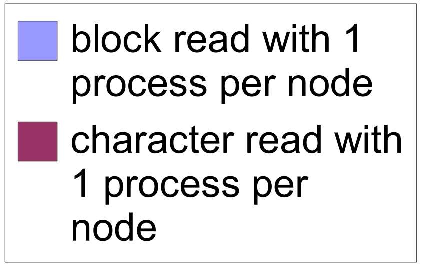


- **300 MB/s from one process**
- **about 190 MB/s per OSS**

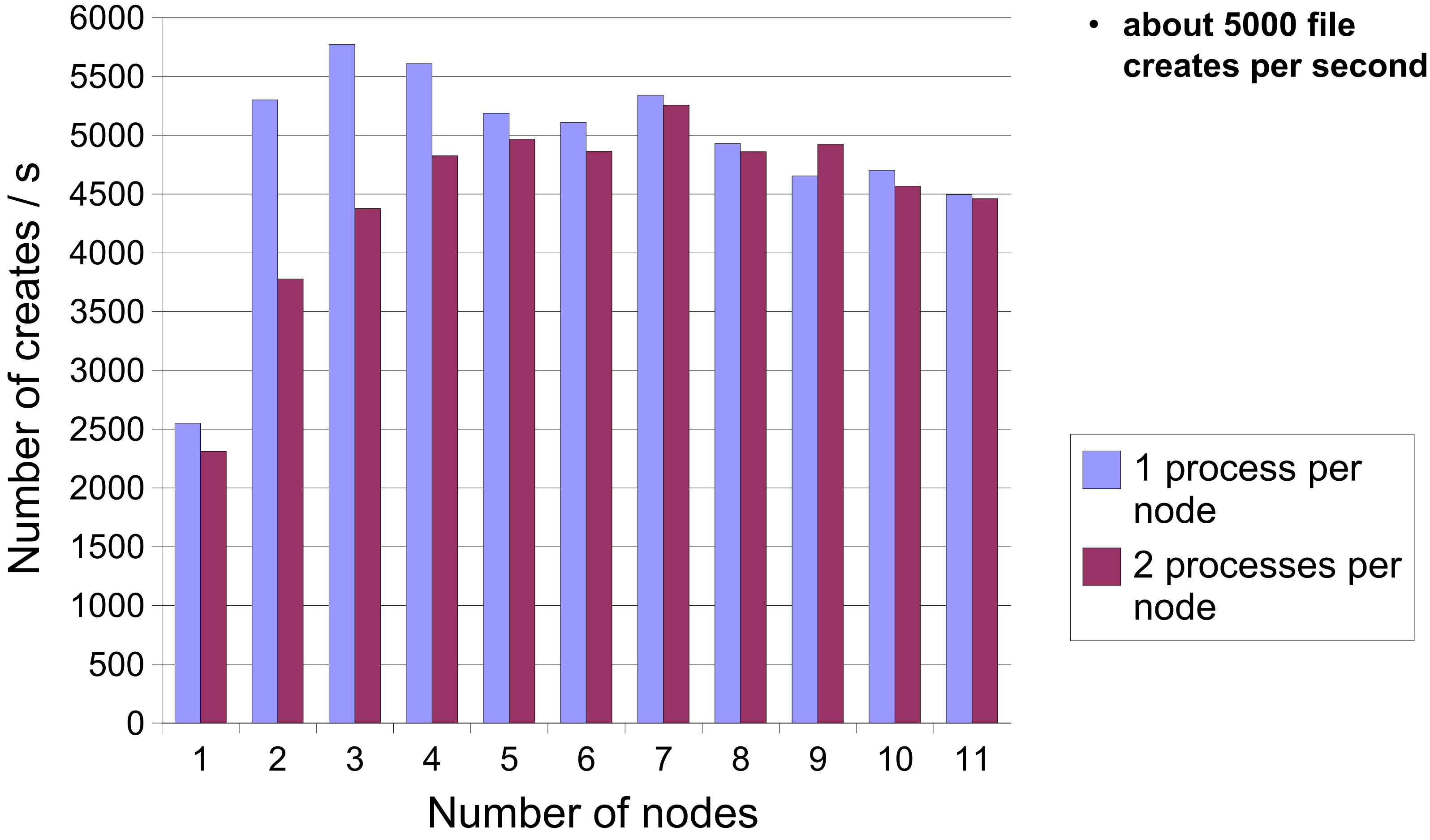
Block vs character read performance with 2 OSS



- **again 190 MB/s per OSS**
- **again no impact of character operations on server performance**



File creation performance



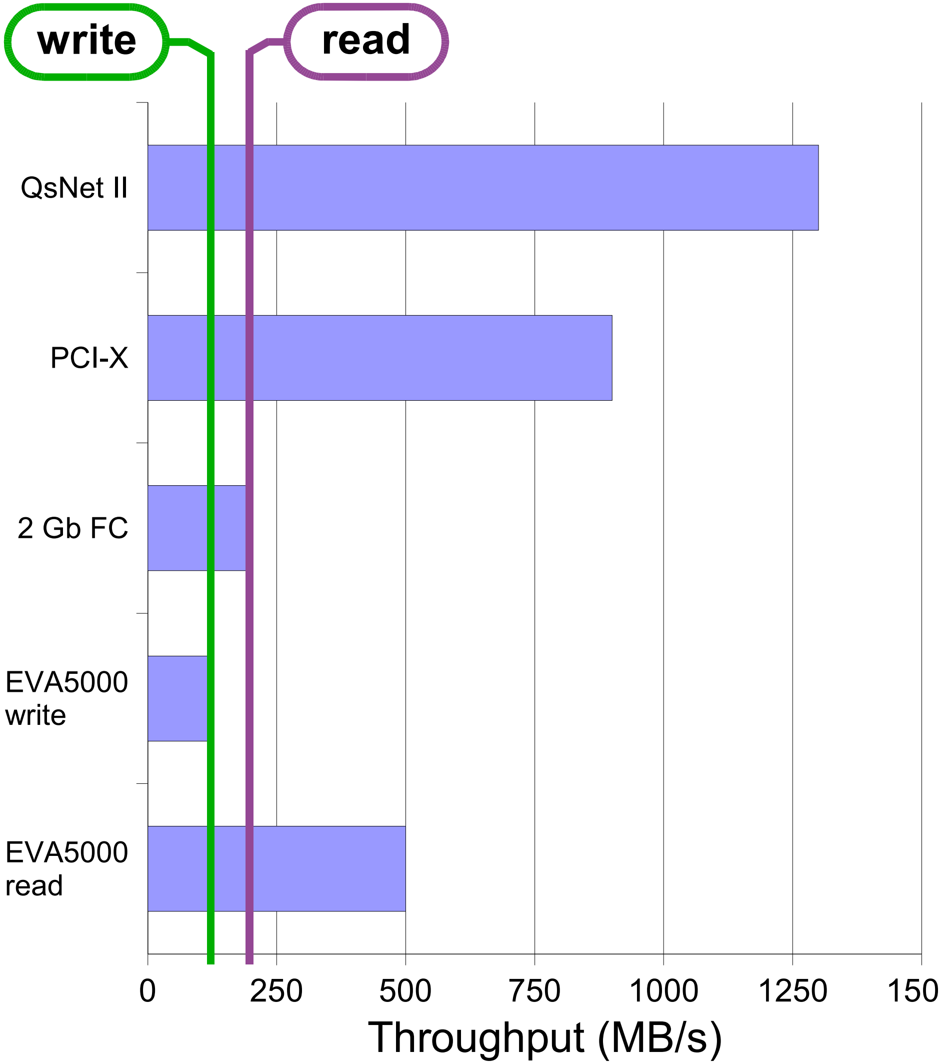
- **about 5000 file creates per second**

Performance measurement summary

- RAW lun performance using 2 controllers on 1 EVA5000 in parallel
  - showed about 120 MB/s for writes and about 195 MB/s for reads
- Main benchmarking results
  - Write performance is about 115 MB/s per OSS
  - Read performance can reach 190 MB/s per OSS
- Possible results per OSS with 4 SFS20 storage arrays:
  - About 400 MB/s for writes and about 580 MB/s for reads
    - SFS20 was not yet available when SSCK’s hardware was delivered
- Performance mainly depends on installed hardware
  - Linear scaling with number of OSS

Performance of OSS components

- **Quadrics Elan4**
  - Internally about 1300 MB/s
  - Only PCI-X adapters exist
- **PCI-X bus on servers**
  - About 900 MB/s
- **Dual-ported FC adapter**
  - About 195 MB/s
  - Actually only 1 port is used
- **EVA5000 storage array**
  - About 120 MB/s for writes
  - Nearly 500 MB/s for reads



Experiences with HP SFS 1.1-0

- Works pretty stable when everything is up and running
  - Production server system usually runs for weeks without problems
    - MDS threads got blocked after about 4 weeks, solved with a patch
- Filesystem operations continue after a problem is repaired
  - Usually batch jobs continue to run
- Understanding the system behaviour is not easy:
  - Some Lustre error messages are critical and some are normal
  - Status of clients can have influence on servers
    - e.g. takeover is faster if all clients can be reached
  - Timing has an influence
    - e.g. takeover only occurs if failover server is up for more than 10 minutes
- After dumps check local disk space
  - Filesystem /local on OSS is hidden and not visible by the df command

Conclusion

- We are working together with HP to reach a highly reliable system
  - Parallel file systems are very complex
    - Hence it is normal to have critical software bugs with new file systems
- HP SFS has the most important features of a parallel file system
  - Performance, resilience, scalability, and ease of administration
  - Additional features are needed for using file systems from two clusters
    - e.g. support for different Lustre versions between clients and servers
- HP SFS and Lustre are very interesting and promising products
  - It works and is heavily used at SSCK's production system
- Now it's the right time to start using HP SFS / Lustre !