Karlsruher Institut für Technologie

Lessons learned from Lustre file system operation

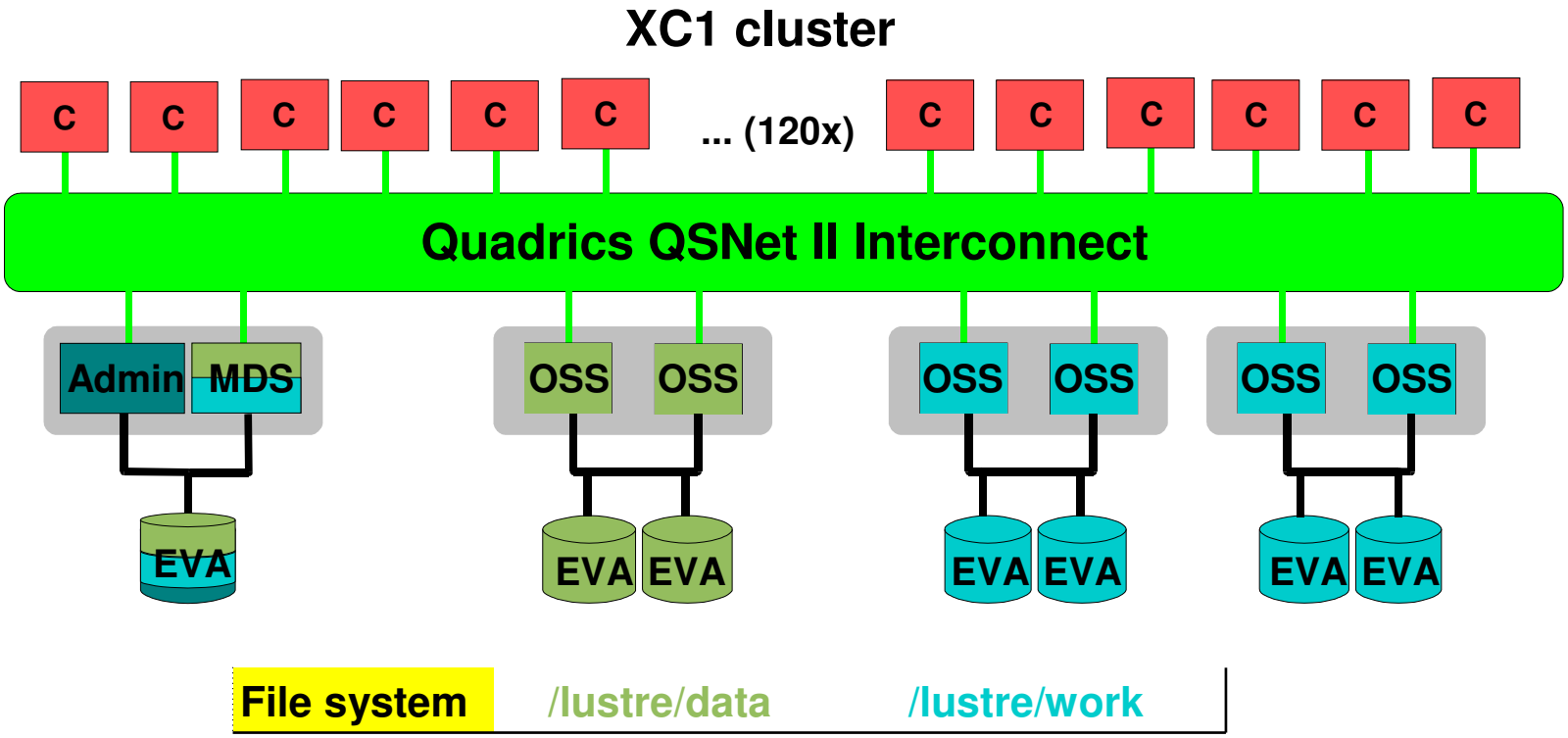
Roland Laifer

STEINBUCH CENTRE FOR COMPUTING - SCC

Overview

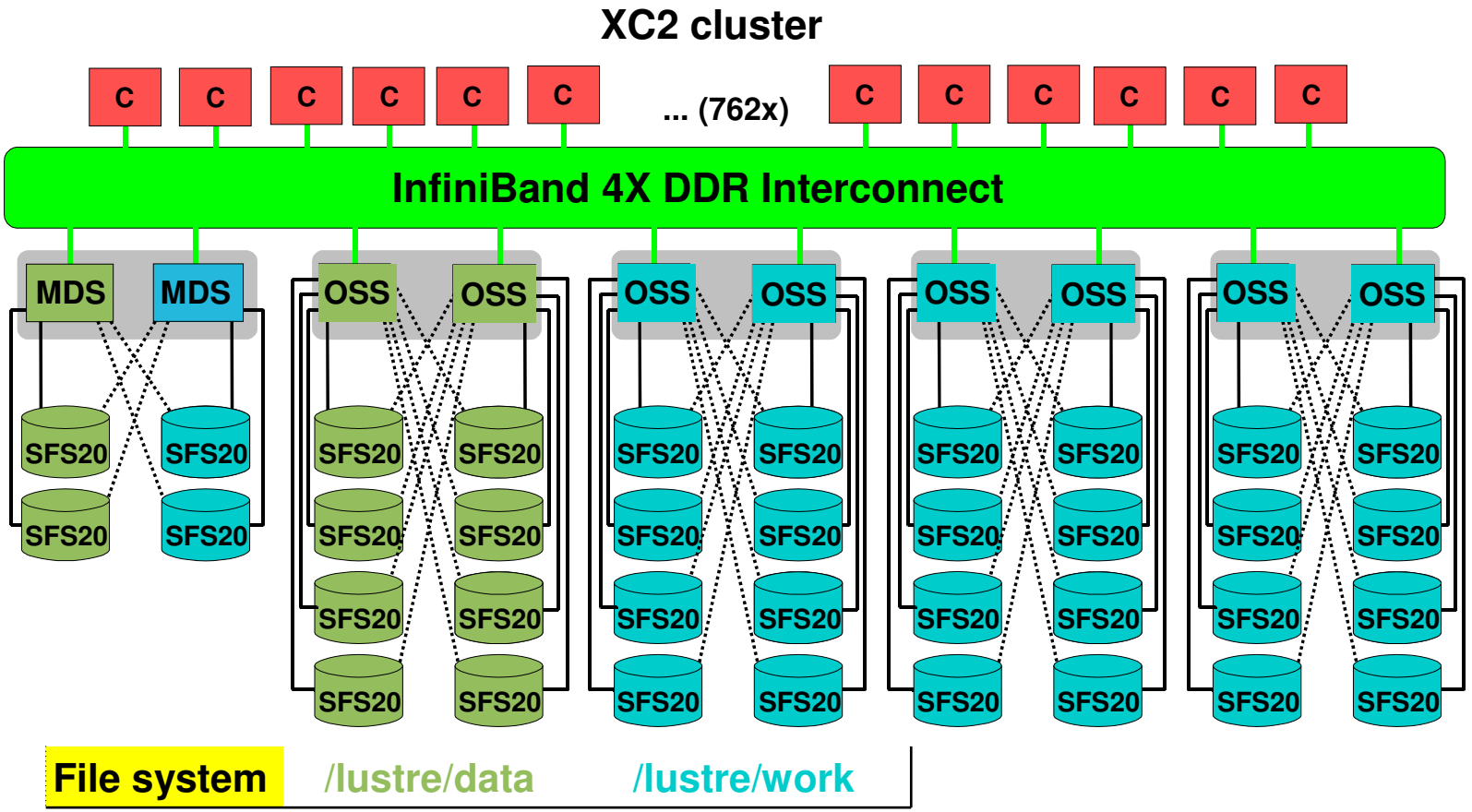
- Lustre systems at KIT
  - KIT is the merger of the University of Karlsruhe and Research Center Karlsruhe
    - with about 9000 employees and about 20000 students
- Experiences
  - with Lustre
  - with underlying storage
- Options for sharing data
  - by coupling InfiniBand fabrics
  - by using Grid protocols

Lustre file systems at XC1



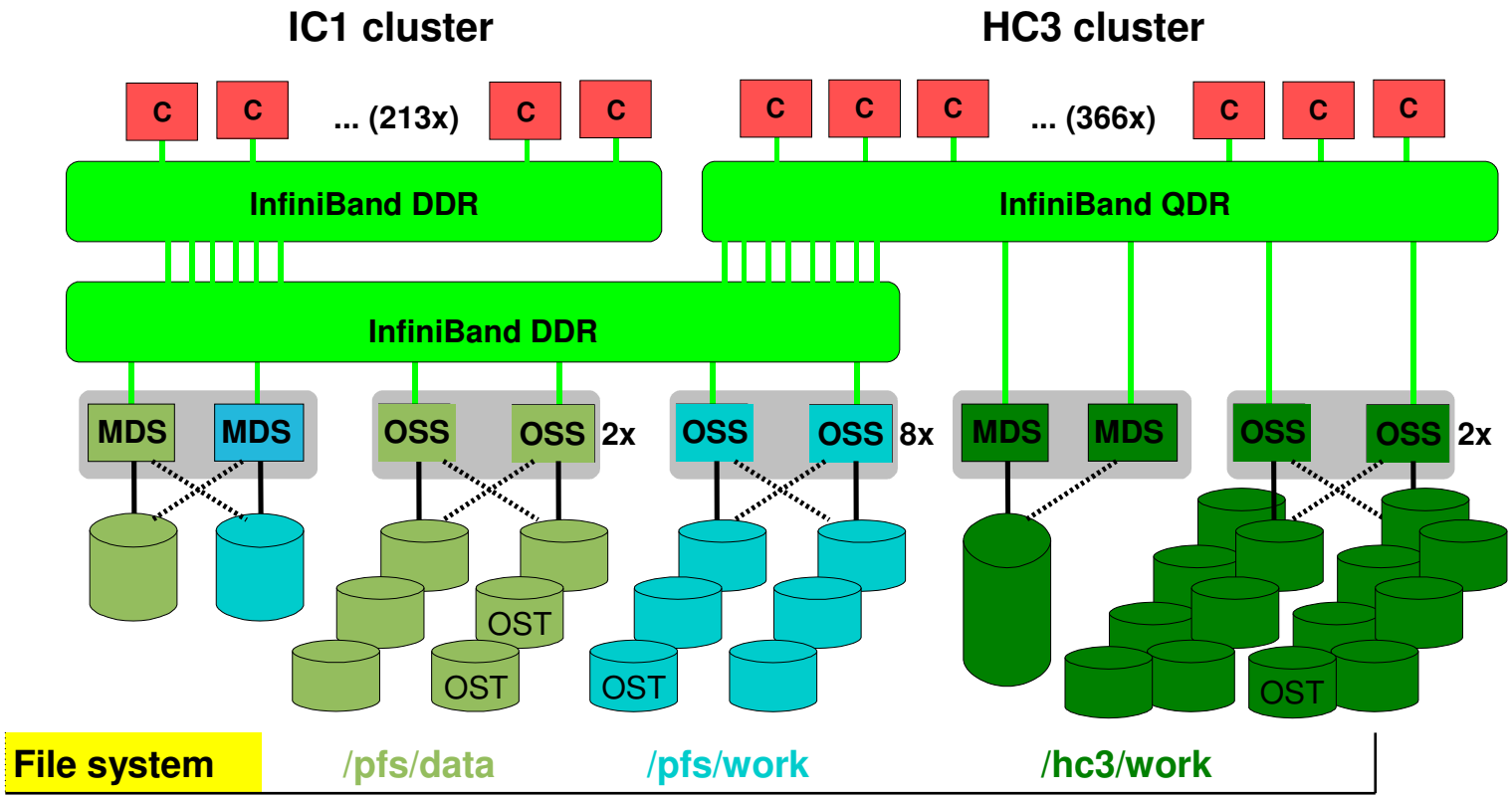
- HP SFS appliance with HP EVA5000 storage
- Production from Jan 2005 to March 2010

Lustre file systems at XC2



- HP SFS appliance (initially) with HP SFS20 storage
- Production since Jan 2007

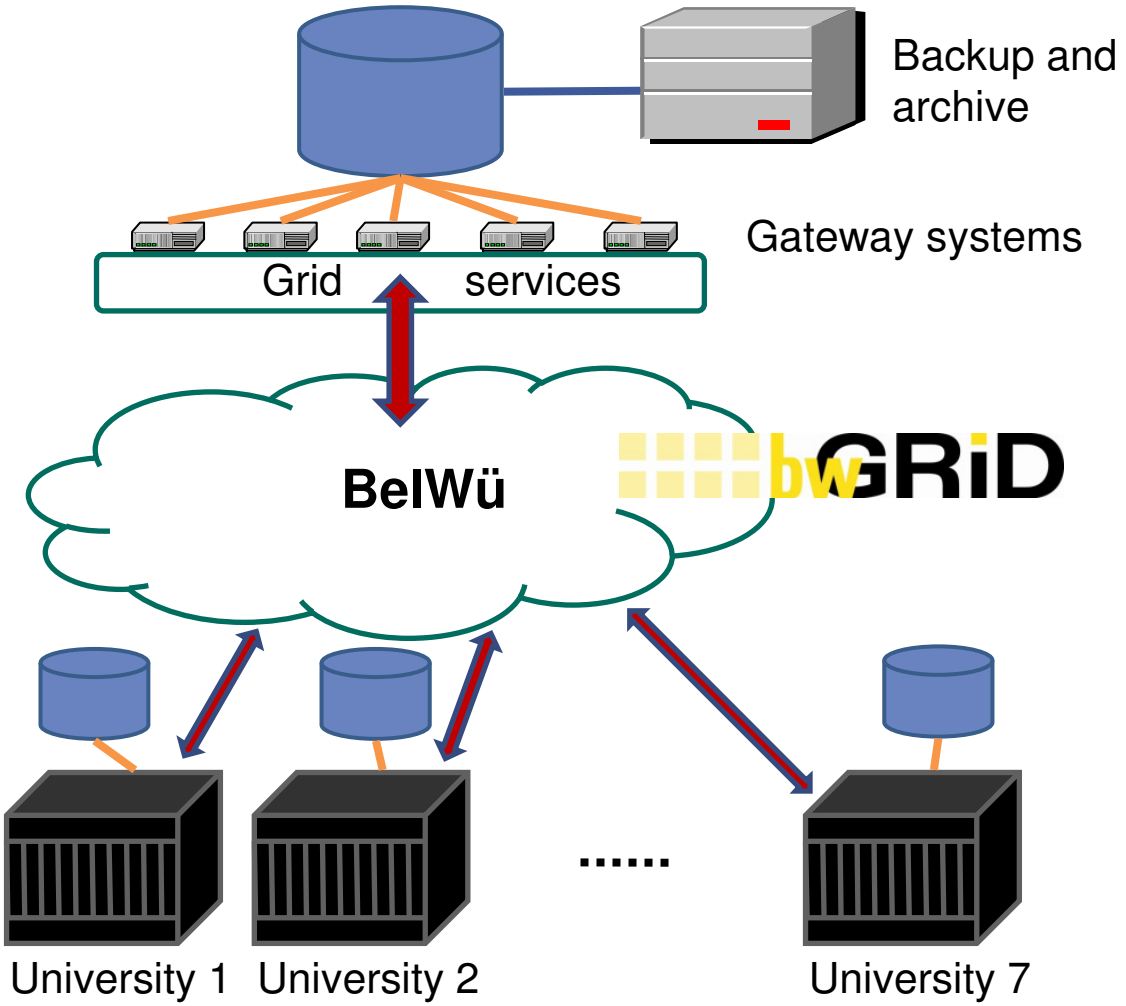
Lustre file systems at HC3 and IC1



- Production on IC1 since June 2008 and on HC3 since Feb 2010
- pfs is transtec/Q-Leap solution with transtec provigo (Infortrend) storage
- hc3work is DDN (HP OEM) solution with DDN S2A9900 storage

bwGRiD storage system (bwfs) concept

- Lustre file systems at 7 sites in state of Baden Württemberg
- Grid middleware for user access and data exchange
- Production since Feb 2009
- HP SFS G3 with MSA2000 storage



Summary of current Lustre systems

System name	hc3work	xc2	pfs	bwfs
Users	KIT	universities, industry	departments, multiple clusters	universities, grid communities
Lustre version	DDN Lustre 1.6.7.2	HP SFS G3.2-3	Transtec/Q-Leap Lustre 1.6.7.2	HP SFS G3.2-[1-3]
# of clients	366	762	583	>1400
# of servers	6	10	22	36
# of file systems	1	2	2	9
# of OSTs	28	8 + 24	12 + 48	7*8 + 16 + 48
Capacity (TB)	203	16 + 48	76 + 301	4*32 + 3*64 + 128 + 256
Throughput (GB/s)	4.5	0.7 + 2.1	1.8 + 6.0	8*1.5 + 3.5
Storage hardware	DDN S2A9900	HP SFS20	transtec provigo	HP MSA2000
# of enclosures	5	36	62	138
# of disks	290	432	992	1656

General Lustre experiences (1)

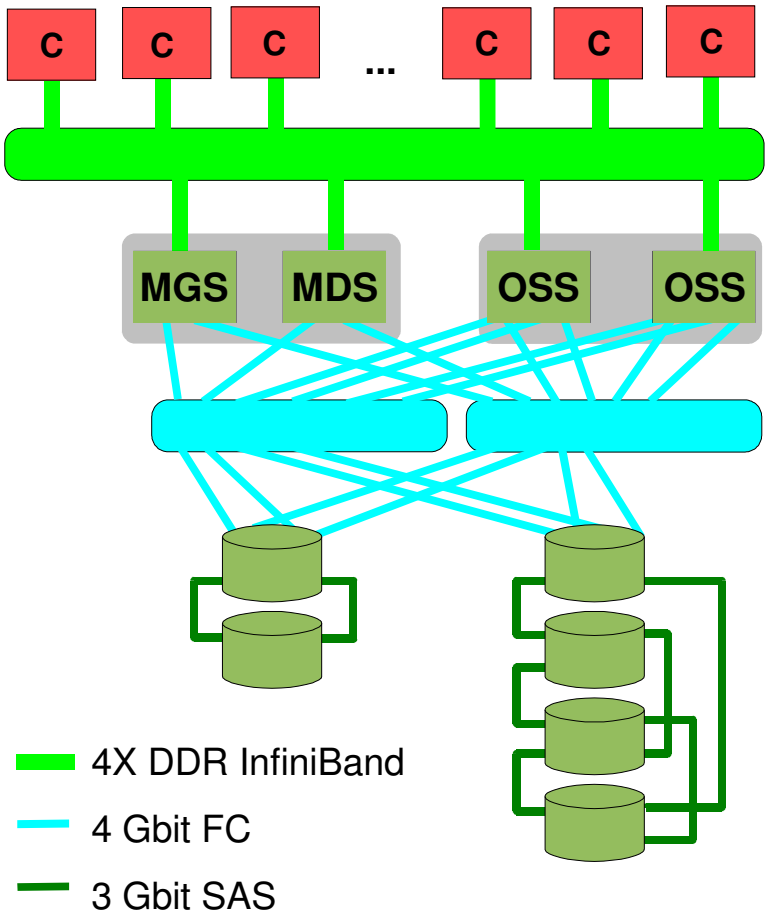
- Using Lustre as home directories works
  - Problems with users creating 10000s of files per small job
    - Convinced them to use local disks (we have at least one per node)
  - Problems with unexperienced users using home for scratch data
    - Also puts high load on backup system
  - Enabling quotas helps to quickly identify bad users
    - Enforcing quotas for inodes and capacity is planned
  - Restore of home directories would last for weeks
    - Idea is to restore important user groups first
    - Luckily up to now complete restore was never required

General Lustre experiences (2)

- Monitoring performance is important
  - Check performance of each OST during maintenance
    - We use **dd** and **parallel_dd** (own perl script)
  - Find out which users are heavily stressing the system
    - We use **collectl** and and script attached to bugzilla 22469
    - Then discuss more efficient system usage, e.g. striping parameters
- Nowadays Lustre is running very stable
  - After months MDS might stall
    - Usually server is shot by heartbeat and failover works
  - Most problems are related to storage subsystems

Complexity of parallel file system solutions (1)

- Complexity of underlying storage
  - Lots of hardware components
    - Cables, adapters, memory, caches, controllers, batteries, switches, disks
    - All can break
  - Firmware or drivers might fail
  - Extreme performance causes problems not seen elsewhere
    - Disks fail frequently
    - Timing issues cause failures



Complexity of parallel file system solutions (2)

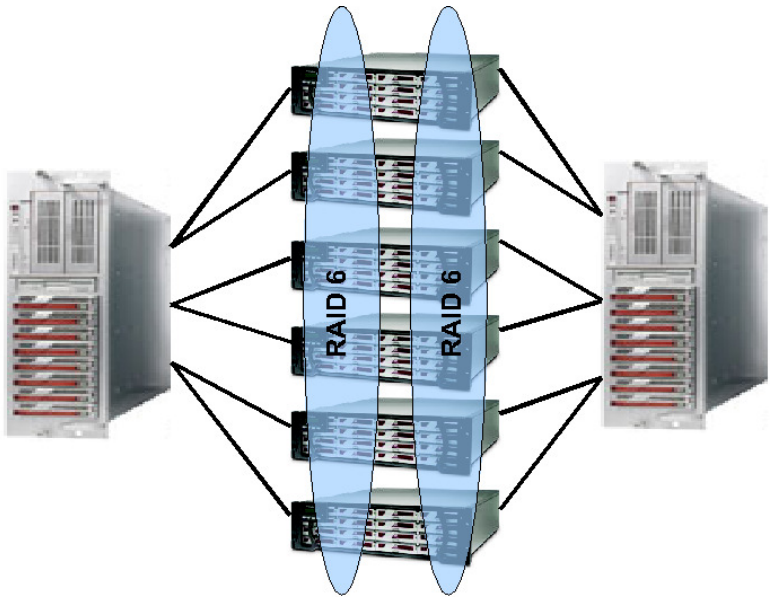
- Complexity of parallel file system (PFS) software
  - Complex operating system interface
  - Complex communication layer
  - Distributed system: components on different systems involved
    - Recovery after failures is complicated
    - Not easy to find out which one is causing trouble
  - Scalability: 1000s of clients use it concurrently
  - Performance: low level implementation required
    - Higher level solutions loose performance
- Expect bugs in any PFS software
- Vendor tests at scale are very important
- Lots of similar installations are benefical

Experiences with storage hardware

- HP SFS20 arrays hang after disk failure under high load
  - Happened at different sites for years
  - System stalls, i.e. no file system check required
- Data corruption at bwGRiD sites with HP MSA2000
  - Firmware of FC switch and of MSA2000 most likely reason
  - Largely fixed by HP action plan with firmware / software upgrades
- Data corruption with transtec provigo 610 RAID systems
  - File system stress test on XFS causes RAID system to hang
  - Problem is still under investigation by Infortrend
- SCSI errors and OST failures with DDN S2A9900
  - Caused by single disk with media errors
  - Happened twice, new firmware provides better bad disk removal
- Expect severe problems with midrange storage systems

Interesting OSS storage option

- OSS configuration details
  - Linux software RAID6 over RAID systems
    - RAID systems have hardware RAID6 over disks
    - RAID systems have one partition for each OSS
- No single point of failure
  - Survives 2 broken RAID systems
  - Survives 8 broken disks
- Good solution with single RAID controllers
  - Mirrored write cache of dual controllers often is bottleneck



Future requirements for PFS / Lustre

- Need better storage subsystems
- Fight against silent data corruption
  - It really happens
    - Finding responsible component is a challenge
  - Checksums quickly show data corruption
    - Provide increased probability to avoid huge data corruptions
  - Storage subsystems should also check data integrity
    - E.g. by checking the RAID parity during read operations
- Support efficient backup and restore
  - Need point in time copies of the data at different location
  - Fast data paths for backup and restore required
  - Checkpoints and differential backups might help

Sharing data (1): Extended InfiniBand fabric

- Examples:
  - IC1 and HC3
  - bwGRiD clusters in Heidelberg and Mannheim (28 km distance)
    - InfiniBand coupled with Obsidian Longbow over DWDM
- Requirements:
  - Select appropriate InfiniBand routing mechanism and cabling
  - Host based subnet managers might be required
- Advantages:
  - Same file system visible and usable on multiple clusters
  - Normal Lustre setup without LNET routers
  - Low performance impact
- Disadvantages:
  - InfiniBand possibly less stable
  - More clients possibly cause additional problems

Sharing data (2): Grid protocols

- Example:
  - bwGRiD
    - gridFTP and rsync over gsiSSH to copy data between clusters
- Requirements:
  - Grid middleware installation
- Advantages:
  - Clusters usable during external network or file system problems
  - Metadata performance not shared between clusters
  - User ID unification not required
  - No full data access for remote root users
- Disadvantages:
  - Users have to synchronize multiple copies of data
  - Some users do not cope with Grid certificates

Further information

- Talks about Lustre administration, performance, usage
  - http://www.scc.kit.edu/produkte/lustre.php
- roland.laifer@kit.edu