

Challenges in making Lustre systems reliable

Roland Laifer

STEINBUCH CENTRE FOR COMPUTING - SCC



Background and motivation

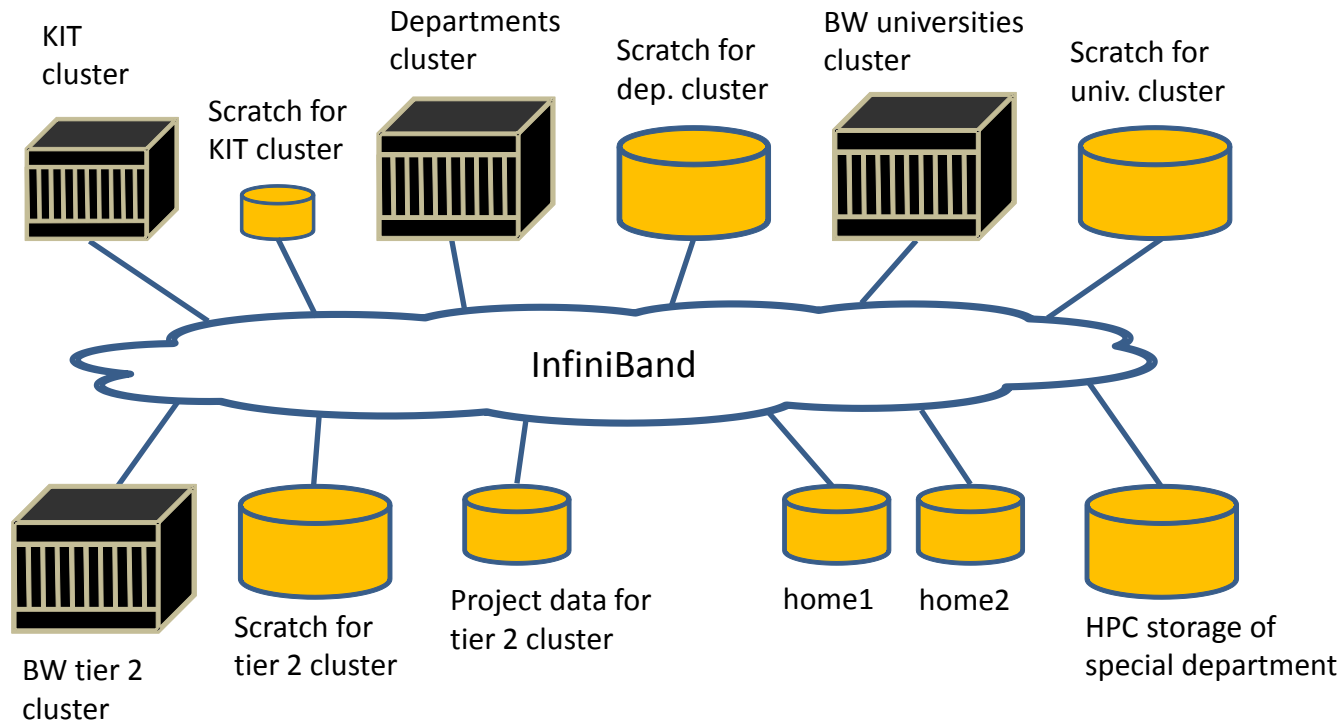
- Most of our compute cluster outages caused by Lustre
 - Probably this would be similar with other parallel file systems
- Even recently we had seen bad Lustre versions
 - Most bugs related to new client OS versions or quotas
- Frequent discussions with users about I/O errors
 - Often small changes allow to omit Lustre evictions
- We had silent data corruption caused by storage hardware
 - A huge damage needs restore of complete file systems

Overview

- Lustre systems at KIT
- Challenge #1: Find stable Lustre versions
- Challenge #2: Find stable storage hardware
- Challenge #3: Identify misbehaving applications
- Challenge #4: Recover from disaster

Lustre systems at KIT - overview

- Multiple clusters and file systems connected to same InfiniBand fabric
 - Good solution to connect Lustre to midrange HPC systems
 - Select appropriate InfiniBand routing mechanism and cable connections
 - Allows direct access to data of other systems without LNET routers



Lustre systems at KIT - details

System name	hc3work	pfs2	pfs3
Users	KIT, 2 clusters	universities, 4 clusters	universities, tier 2 cluster
Lustre server version	DDN Lustre 2.4.3	DDN Lustre 2.4.3	DDN Lustre 2.4.3
# of clients	868	1941	540
# of servers	6	21	17
# of file systems	1	4	3
# of OSTs	28	2*20, 2*40	1*20, 2*40
Capacity (TB)	203	2*427, 2*853	1*427, 2*853
Throughput (GB/s)	4.5	2*8, 2*16	1*8, 2*16
Storage hardware	DDN S2A9900	DDN SFA12K	DDN SFA12K
# of enclosures	5	20	20
# of disks	290	1200	1000

Challenge #1: Find stable Lustre versions (1)

- Lustre 1.8.4 / 1.8.7 was running very stable
 - Needed to upgrade to version 2.x for SLES11 SP2 client support
- Lustre 2.1.[2-4] on servers was pretty stable
 - Clients with version 2.3.0 caused trouble, needed for SLES11 SP2
 - SLES11 SP2 clients with version 2.4.1 were stable
 - Needed to upgrade servers to 2.4.1 for SP3/RH6.5 client support
- Lustre 2.4.1 on servers caused some problems
 - Bad failover configuration caused outages (LU-3829/4243/4722)
 - Needed to disable ACLs after security alert (LU-4703/4704)
 - User and group quota values were wrong (LU-4345/4504)
- Lustre 2.5.1 clients on RH6.5 were bad
 - LBUGs and clients hanging in status evicted (LU-5071)

Challenge #1: Find stable Lustre versions (2)

- Lustre 2.4.3 on servers is stable
 - SLES11 SP3 clients with Lustre 2.4.3 cause no problems
 - RH6.5 clients with Lustre 2.5.2 are now stable, too
 - User and group quotas are still forged (LU-4345/4504/5188)
 - Not yet fixed in current maintenance release(?)
 - Objects on OSTs sometimes created with arbitrary UID/GID
 - Recommended way to fix bad quotas is hardly usable

- ➔ Choose stable maintenance release
- ➔ Stay with old stable versions
- ➔ Bad dependency: OS upgrade → new Lustre client version required → new Lustre server version required

Challenge #2: Find stable storage hardware

- We had silent data corruption on Infortrend and HP RAIDs
 - More details given at my talk at ELWS'11
 - Infortrend systems provided different data when reading twice
 - This happened once per year on 1 of 60 RAID systems
 - HP MSA2000 G2 had problem with cache mirroring
 - After PCIE link failed messages e2fsck sometimes showed corruption
- Some hardware features greatly improve stability
 - Features which reduce rebuild times
 - Declustered RAID or partial rebuilds using bitmaps
 - We have seen triple disk failures on RAID6 arrays multiple times
 - Features which detect silent data corruption
- ➔ High end storage systems are most likely more stable
- ➔ During procurements give stability features high weighting

Challenge #3: Identify misbehaving apps

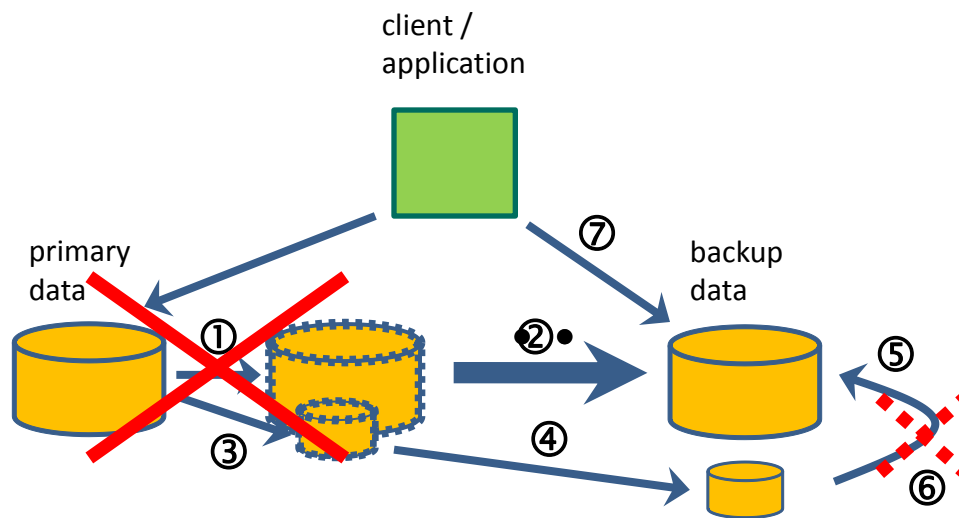
- Find out which file is causing an eviction
 - On version 2.x LustreError messages like this might appear:
 - LustreError: 45766:0:(osc_lock.c:817:osc_ldlm_completion_ast()) lov sub@...: [0 ... W(2):[0, ...]@[0x200001ea7:0x18682:0x0]]
 - Use FID and file system name to identify the badly accessed file
 - lfs fid2path pfs3wor4 [0x200001ea7:0x18682:0x0]
 - Unfortunately, sometimes only the root of the file system is reported
- Use performance monitoring to check what users are doing
 - For good instructions see Daniel Kobras' talk at LAD'12
 - E.g. this shows bad scratch usage of home directory file systems
 - Frequently users do not know how their I/O ends up on Lustre
- ➔ Enhancing misbehaving applications and reducing load stabilizes the file system and makes it more responsive

Challenge #4: Recover from disaster (1)

- A disaster can be caused by
 - hardware failure, e.g. a triple disk failure on RAID6
 - silent data corruption caused by hardware, firmware or software
 - complete infrastructure loss, e.g. caused by fire or flood
- Timely restore of 100s TB does not work
 - Transfer takes too long and rates are lower than expected
 - Bottlenecks often in network or at backup system
 - Metadata recreation rates can be limiting factor
 - We restored a 70 TB Lustre file system with 60 million files
 - With old hardware and IBM TSM this took 3 weeks

Challenge #4: Recover from disaster (2)

- Solution on other huge storage systems:
- ➔ Do not restore but switch on client or application to backup copy !



Note: Data created after last good incremental snapshot is lost.

Backup:

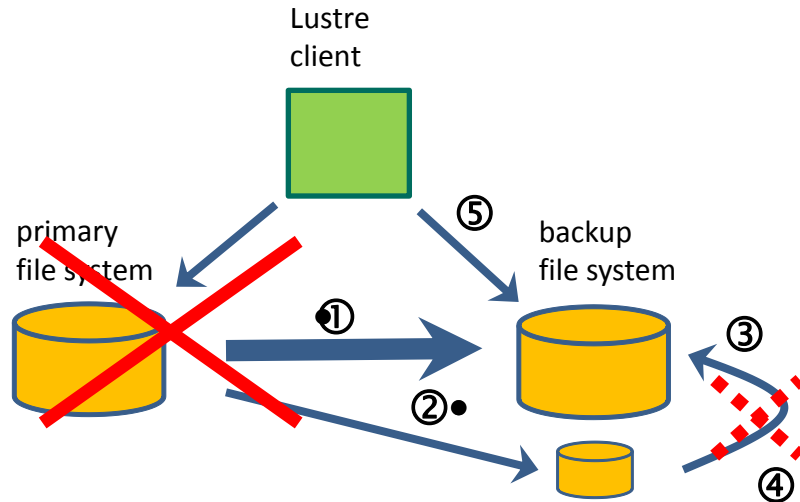
1. Complete snapshot
2. Transfer snapshot to backup system
3. Incremental snapshot
4. Transfer incr. snapshot
5. Install incr. snapshot on backup system

Disaster recovery:

6. Remove bad incr. snapshot if required
7. Redirect application or client to backup system

Challenge #4: Recover from disaster (3)

- Transfer previous solution to Lustre:
- ➔ rsnapshot uses rsync to create copy and creates multiple copies by using hard links for unchanged data



Note: Data created after last good rsync is lost.

Backup:

1. Use rsnapshot (rsync) to transfer all data to backup file system
2. Use rsnapshot (rsync + hard links) to transfer new data
3. rsnapshot possibly removes old copies

Disaster recovery:

4. Use good rsnapshot copy and move directories to desired location
5. Adapt mount configuration and reboot Lustre clients

Challenge #4: Recover from disaster (4)

■ Details of our solution

- 3 file systems with permanent data (home, project, software)
 - Each holds production data of some groups and backup data of others

■ Experiences

- Did not yet need the disaster recovery on production systems
 - User requested restores have been done and are just a copy
- Backup done twice per week on one client with 4 parallel processes
 - For 100 mill. files and with 5 TB snapshot data this takes 26 hours
- In addition, we still use IBM TSM backup
 - This allows users to restore by themselves
 - Alternatively, rsnapshot copies could be exported read-only via NFS

Challenge #4: Recover from disaster (5)

- Restrictions of the solution
 - Slow silent data corruption might pollute all backup data
 - Same problem for other backup solutions
 - We did not yet see this case, i.e. OSS go pretty fast in status read-only
 - Recovery does not work if both file systems have critical Lustre bug
 - Different Lustre versions on primary and backup file system might help
 - Using `lustre_rsync` instead of `rsync` would omit file system scans
 - We plan to investigate if and how this would work

Further information

- rsnapshot
 - <http://www.rsnapshot.org>
- All my talks about Lustre
 - <http://www.scc.kit.edu/produkte/lustre.php>
- roland.laifer@kit.edu