Karlsruher Institut für Technologie

The parallel file system Lustre

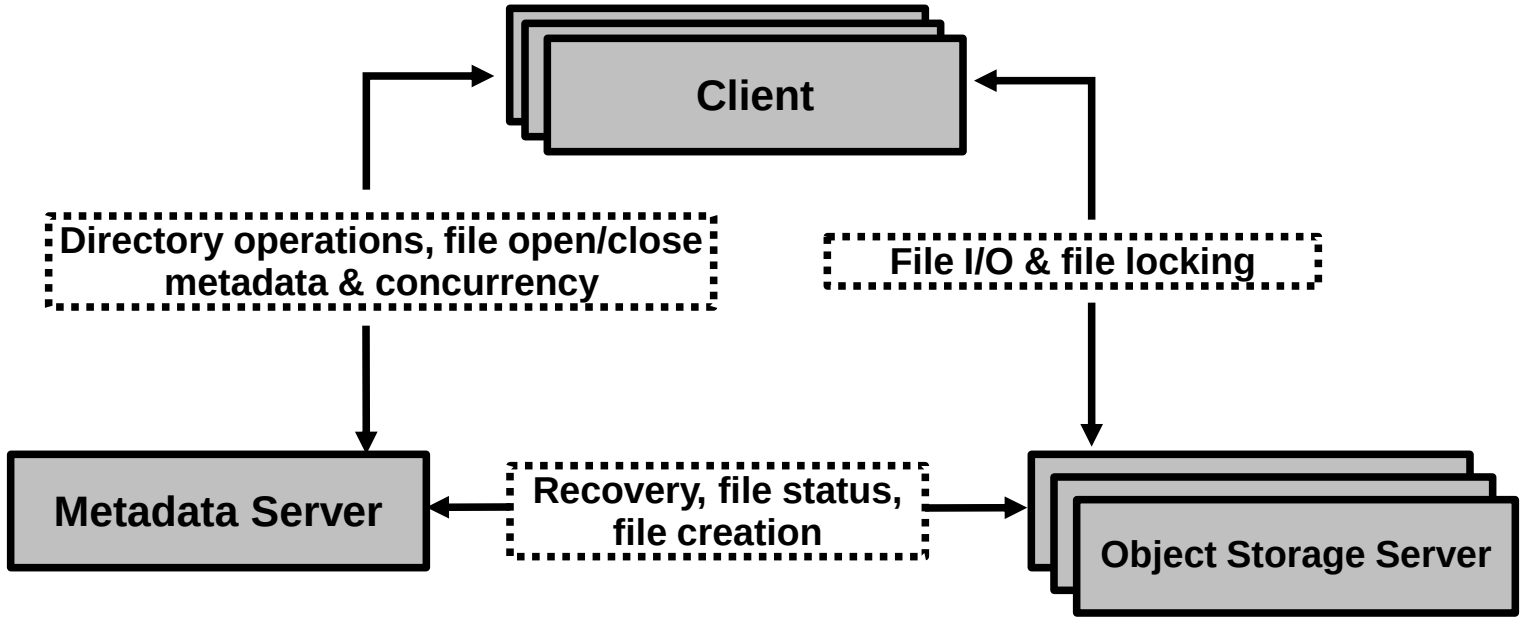
Roland Laifer

STEINBUCH CENTRE FOR COMPUTING - SCC

Overview

- Basic Lustre concepts
- Lustre status
  - Vendors
  - New features
  - Pros and cons
- Lustre systems at KIT
- Complexity of underlying hardware
- Remarks on Lustre performance

Basic Lustre concepts



- Lustre componets:
  - Clients offer standard file system API (POSIX)
  - Metadata servers (MDS) hold metadata, e.g. directory data, and store them on Metadata Targets (MDTs)
  - Object Storage Servers (OSS) hold file contents and store them on Object Storage Targets (OSTs)
  - All communicate efficiently over interconnects, e.g. with RDMA

Lustre status (1)

- Huge user base
  - about 70% of Top100 use Lustre
- Lustre HW + SW solutions available from many vendors:
  - DDN (via resellers, e.g. HP, Dell), Xyratex – now Seagate (via resellers, e.g. Cray, HP), Bull, NEC, NetApp, EMC, SGI
- Lustre is Open Source
  - Lots of organizational changes for main developers
    - CFS → Sun → Oracle → Whamcloud → Intel
  - OpenSFS/EOFS are mainly sponsoring Lustre development
    - Cray, LLNL, ORNL, Xyratex, DDN, EMC, Intel, NCSA, SNL, CEA
  - Many companies supplied code changes to Lustre 2.4:
    - Intel, LLNL, Xyratex, EMC, ORNL, Bull, TACC, CEA, Suse, NRL, S&C, Fujitsu, Cray, DDN, NASA
    - Main development is done at Intel HPDD

Lustre status (2)

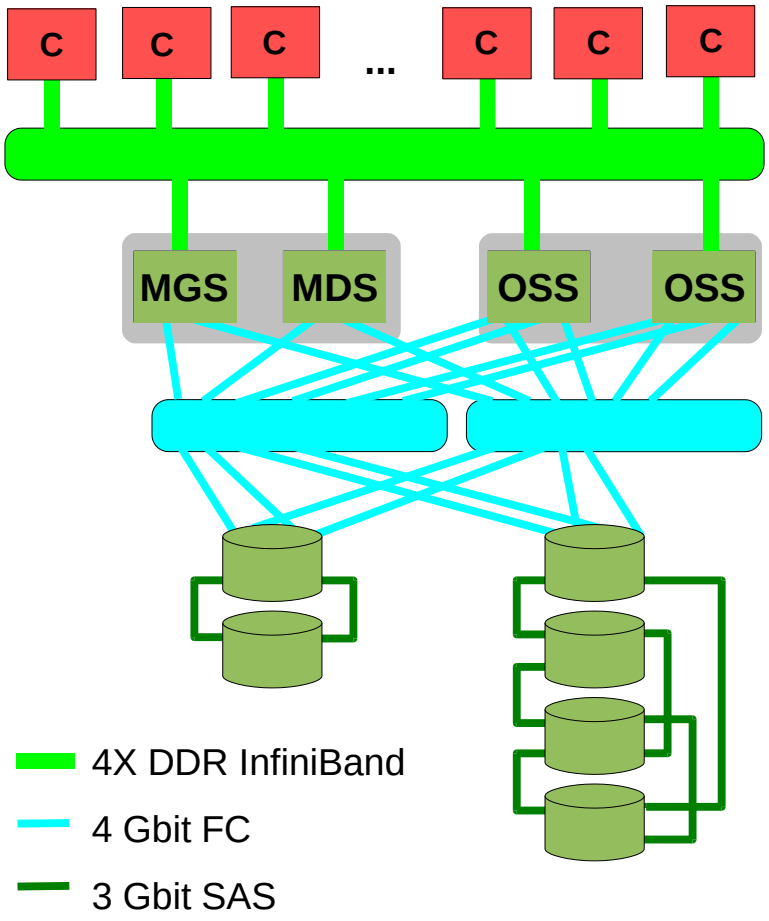
- Maintenance and feature releases
  - Current maintenance release versions: 2.1.6, 2.4.3, 2.5.1
  - Vendor releases with backported patches on top
- New Lustre features in Lustre 2.4/2.5:
  - ZFS as optional underlying file system (pushed by LLNL)
  - Support for multiple MDS (currently not for same directory)
  - HSM support (pushed by CEA)
  - Lustre is in staging area of Linux kernel (pushed by EMC)
- Pros and Cons
  - \+ Usually runs very stable, failover capabilities
  - \+ Open source, open bug tracking
  - \+ Scalable up to 10000s of clients
  - \+ High throughput with multiple network protocols and LNET routers
  - \- Limited in its features, e.g. no data replication or snapshots

Lustre systems at KIT

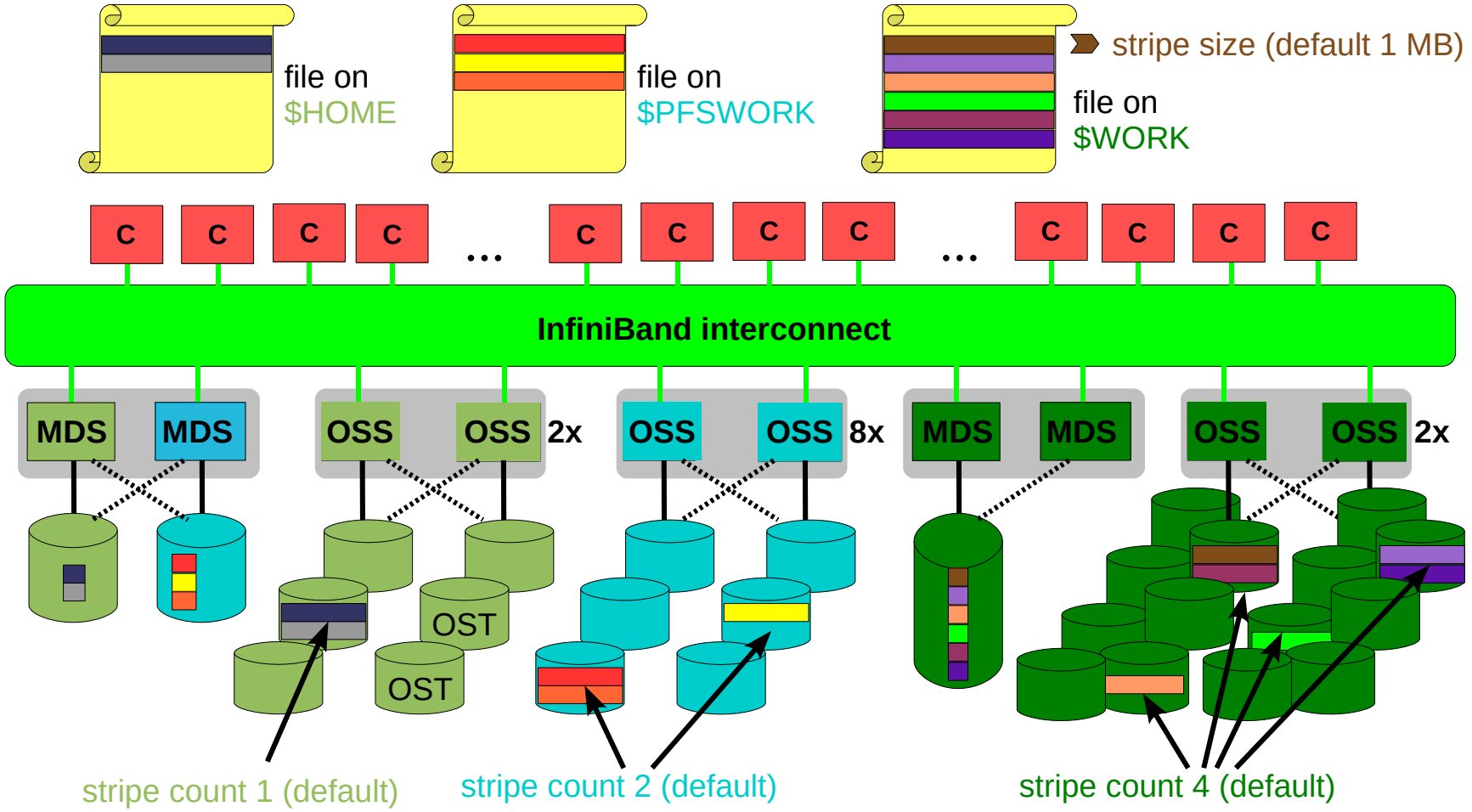
System name	pfs1	hc3work	pfs2	pfs3 (June 2014)
Users	KIT, 2 clusters	KIT, 2 clusters	universities, 3 clusters	universities, research cluster
Lustre server version	Q-Leap Lustre 2.1.3	DDN Lustre 2.4.1	DDN Lustre 2.4.1	DDN Lustre 2.4.3
# of clients	863	863	1395	538
# of servers	18	6	21	17
# of file systems	1	1	4	3
# of OSTs	48	28	2*20, 2*40	1*20, 2*40
Capacity (TB)	301	203	2*427, 2*853	1*427, 2*853
Throughput (GB/s)	6.0	4.5	2*8, 2*16	1*8, 2*16
Storage hardware	transtec provigo	DDN S2A9900	DDN SFA12K	DDN SFA12K
# of enclosures	50	5	20	20
# of disks	800	290	1200	1000

Complexity of underlying hardware

- 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
- Challenges:
  - Silent data corruption
  - Analyze performance problems



How does Lustre striping work?



- Parallel data paths from clients to storage

Remarks on Lustre performance

- Scalable throughput performance
  - Linear increase with number of OSTs
  - Files are automatically distributed across OSTs
  - For few large files increase stripe count
- Metadata performance
  - With one MDT restricted
    - 10000s creates, stats, deletes with empty files
  - Objects are created previously on OSTs
  - Deletes are slower than creates because of seeks on disks
- IOPS performance
  - Not good since many Lustre operations happen on 1 MB blocks
  - Locking conflicts/timeouts for I/O to same area from many clients
    - Lustre guarantees data consistency

Further information

- Lustre wiki at Intel
  - https://wiki.hpdd.intel.com/display/PUB/HPDD+Wiki+Front+Page
- Lustre manuals
  - http://lustre.opensfs.org/documentation/
- Latest Lustre User Group talks
  - http://www.opensfs.org/lug-2014-presos/
- Latest Lustre Admins & Developers talks
  - http://www.eofs.eu/?id=lad13
- SCC talks about Lustre
  - http://www.scc.kit.edu/produkte/lustre.php
- Good white paper from Intel
  - http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/architecting-lustre-storage-white-paper.pdf