Disaster recovery of a ZFS based lustre object store

Sergey Noskov, Markus Tacke,
Jürgen Weiß, Carsten Allendörfer

Toulouse 01.10.2025
LAD 25

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

JG|U

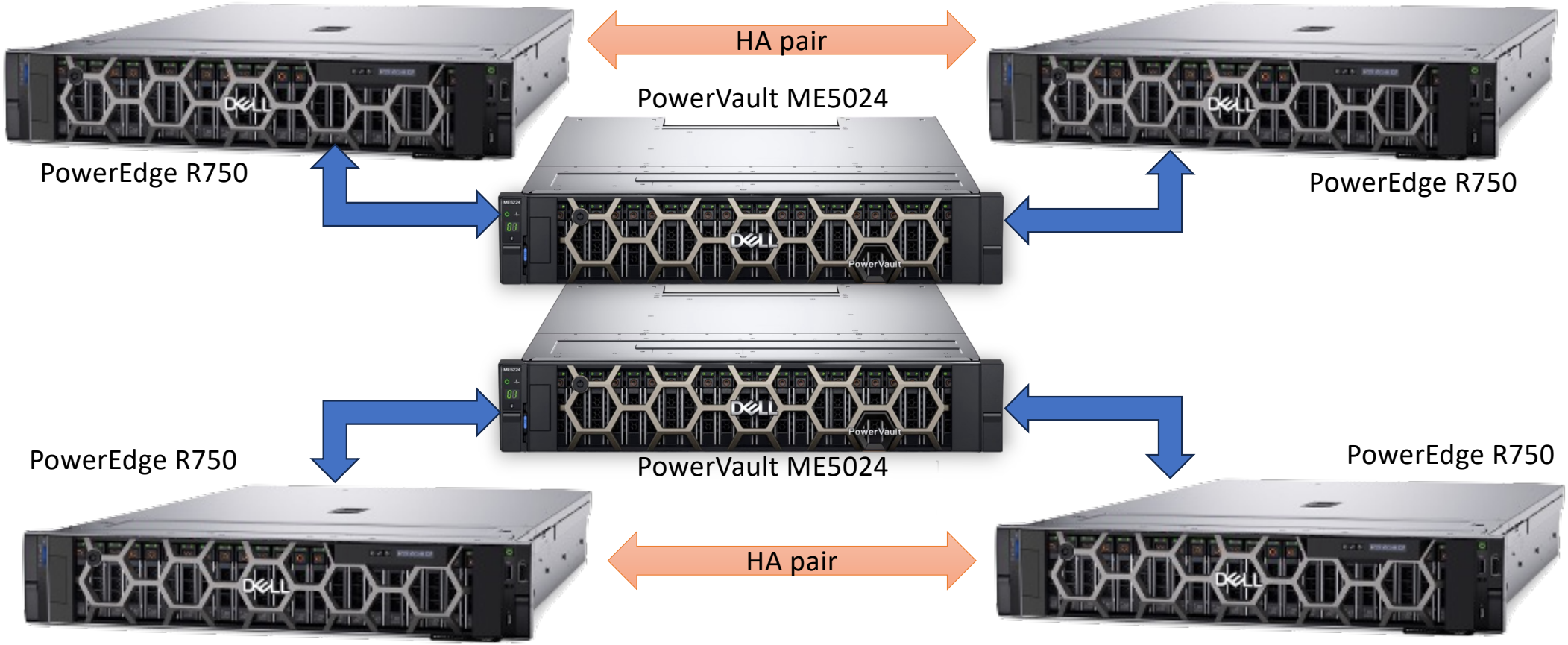
OUTLOOK FOR THE NEXT 20 MINUTES

- Lustre servers cluster overview
  - Hardware
  - Software
  - Tuning
- Failure incident on the ZFS-based object storage target (OST)
  - Events chronology
  - Pool state analysis
  - Data restore
  - Conclusion

HARDWARE


DELL EMC
PowerVault
DELL EMC
PowerVault
PowerVault
PowerVault

LUSTRE METADATA STORAGE



LUSTRE METADATA SERVER

Dell	PowerEdge R750
CPU(s)	2x Intel(R) Xeon(R) Gold 6336Y CPU @ 2.40GHz 24 cores
RAM	DDR4 256 GB (16 x 16GB, dual rank, 3200 MT/s)
Network	2x ConnectX-6 VPI PCIe gen.4 on the CPU1 Port1 -> Infiniband 100Gb Port2 -> Ethernet 100Gb (bond mit der 2.Karte)
SAS Adapter	2x Dell HBA355e PCIe gen.4 on the CPU2
system volume	Raid controller PERC H745; SSDs: 2x 446.63 GB (mirroring)

RAID-CONTROLLER FOR METADATA

Dell	PowerVault ME5024
Controllers count	2 (each 4x SAS)
SSD drives count	24
SSD drive size	7.6 TB

LUSTRE OBJECT STORAGE SERVICE

OSS Server count	Enclosures x Disks
8	16 x 84

LUSTRE OBJECT STORAGE SERVICE (HIM POOL)

OSS Server count	Enclosures x Disks
2	6 x 84

LUSTRE OSS SERVER

Dell	PowerEdge R750
CPU(s)	2x Intel(R) Xeon(R) Gold 6336Y CPU @ 2.40GHz 24 cores
RAM	DDR4 256 GB (16 x 16GB, dual rank, 3200 MT/s)
Network	2x ConnectX-6 VPI (MCX653106A-ECA_Ax) PCIe gen.4 conn. CPU1 Port1 -> Infiniband 100Gb Port2 -> Ethernet 100Gb
SAS Adapter	2x Dell HBA355e PCIe gen.4 conn. CPU2
system volume	Raid PERC H745 with 2x 446.63 GB (mirroring)

LUSTRE OSS SERVER (HIM PARTITION)

Dell	PowerEdge R750
CPU(s)	2x Intel(R) Xeon(R) Gold 6336Y CPU @ 2.40GHz 24 cores
RAM	DDR4 256 GB (16 x 16GB, dual rank, 3200 MT/s)
Network	2x ConnectX-6 VPI (MCX653106A-HDA_Ax) PCIe gen.4 conn. CPU1 Port1 -> Infiniband 200Gb Port2 -> Ethernet 100Gb
SAS Adapter	2x Dell HBA355e PCIe gen.4 conn. CPU2
system volume	Raid PERC H745 with 2x 446.63 GB (mirroring)

STORAGE ENCLOSURE

Dell	EMC Storage Expansion Enclosure ME484
SAS Moduls	2
12Gbit/s SAS connectors	4
Disk slots	84
Disk size	16 TB
total volume	1344 TB

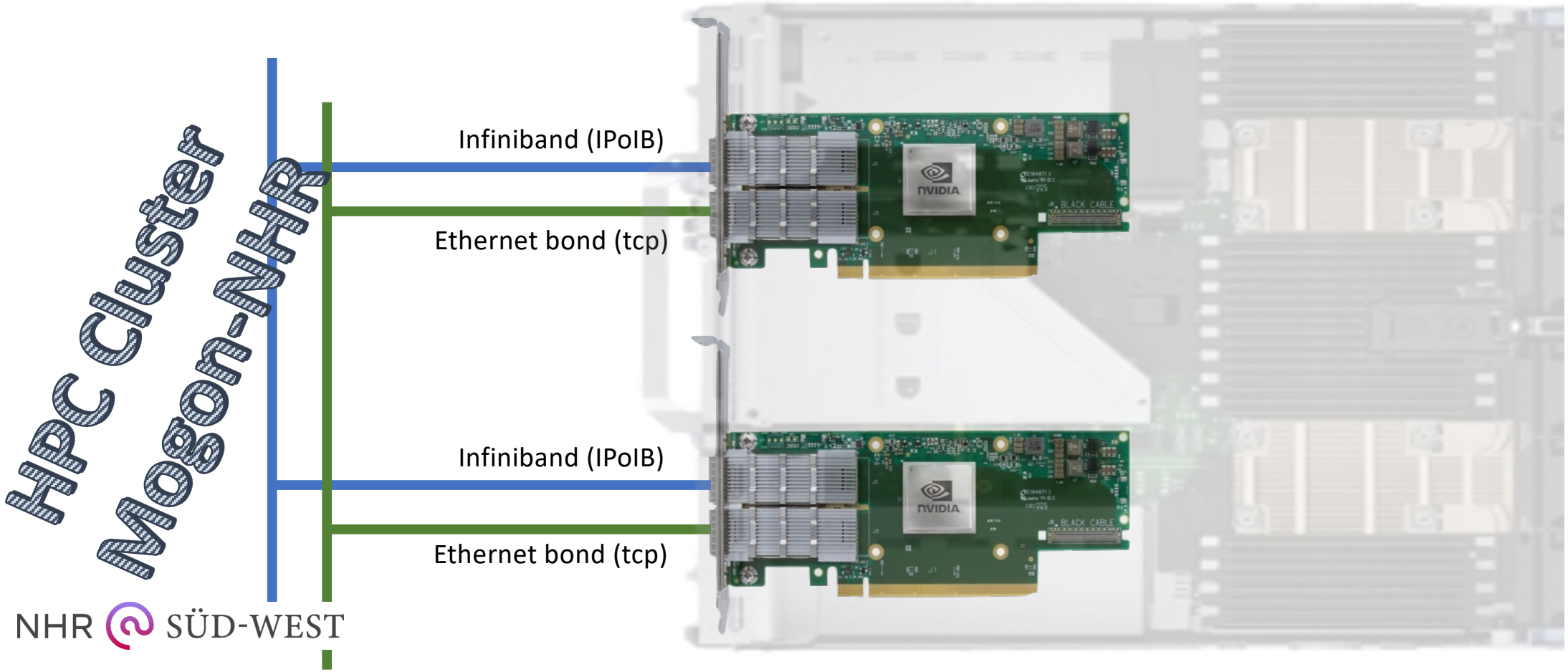
STORAGE ENCLOSURE (HIM PARTITION)

Seagate	OneStor™ SP-35840
SAS Moduls	2
12Gbit/s SAS connectors	4
Disk slots	84
Disk size	20 TB
total volume	1680 TB

CONNECTIVITY

HA-Paar	Targets	LNET Interfaces per server
MDS00[1/2] MDS00[3/4]	MGT MDT000{0..4}	2 x IPoIB 100GB/s 1 x TCP/IP EthernetBond(802.3ad) 2x 100Gb/s
OSS00[1/2] OSS00[3/4] OSS00[5/6] OSS00[7/8]	OST00{00..1f}	
OSS00[9/10]	OST00{100..10b}	2 x IPoIB 200GB/s 1 x TCP/IP EthernetBond(802.3ad) 2x 100Gb/s

CONNECTIVITY



NHR SÜD-WEST

JG|U

CLUSTER MOGON-NHR CONNECTION

- 700+ Nodes for CPU computations and few nodes with GPUs
- Inifiniband HDR100 (some HDR200)
- Topology Fat tree

SOFTWARE OF THE LUSTRE SERVERS

- Alma Linux 8.10
- Lustre 2.15.6(2.15.7) from Whamcloud
- ZFS 2.1.16(2.2.8)
- Pacemaker 2.1.7 rel.5.2.el8_10(5.3.el8_10)

ZFS POOLS AS LUSTRE STORAGE

HA-Paar	Targets	#	RAID typ	TB per target	Total, TB
MDS00[1/2]	mgt	1	Dell „ADAPT“	4	4
MDS00[1/2]	mdt0000 mdt0001	2	Dell „ADAPT“	56 52	220
MDS00[3/4]	mdt000[2/3]	2	Dell „ADAPT“	56	
OSS00[1/2]	OST00{00..07}	32	draid2:11d:42c:2s	474	15168
OSS00[3/4]	OST00{08..0f}				
OSS00[5/6]	OST00{10..17}				
OSS00[7/8]	OST00{18..1f}				
OSS0[09/10]	OST01{00..0b}	12	draid3:12d:42c:2s	513	6156

ZFS POOLS AS LUSTRE STORAGE

Targets	Create options
mgt	-o multihost=on -O canmount=off -O recordsize=128k -o cachefile=none
mdt000[0/1]	
mdt000[2/3]	
OST00{00..1f}	-o multihost=on -O canmount=off -O recordsize=1024k -o cachefile=none
OST01{00..0b}	

DATA DISTRIBUTION, PFL

pool	set stripe	OSTs	OS vol	Metadata	DOM
„JGU“	-E 1M -L mdt -E -1 -c 4 -p jgu /lustre/jgu	32	14.8 PB	220 TB	220 TB
„HIM“	-E -1 -c 1 -p him /lustre/him	12	6.0 PB		0

LUSTRE TUNING

LUSTRE NET TUNING

Parameter	o2ib3 interface	TCP3 interface
peer_timeout	180	180
peer_credits	16	8
credits	2560	
peercredits_hiw	31	
map_on_demand	256	
concurrent_sends	256	
fmr_pool_size	2048	
fmr_flush_trigger	1024	
fmr_cache	1	
conns_per_peer	4	4

ZFS TUNING

Parameter	Value
zfs_arc_max	134933297152 (1/2 RAM)
zfetch_max_distance	67108864
zfs_delay_scale	100000
zfs zfs_dirty_data_max	17179869184 (~2s)
zfs_vdev_aggregation_limit	1048576
zfs_vdev_async_read_min_active	4
zfs_vdev_async_read_max_active	16
zfs_vdev_async_write_min_active	5
zfs_vdev_async_write_max_active	10
zfs_dirty_data_sync_percent	50

FAILURE INCIDENT

FAILURE INCIDENT 08.MAI 2025

Server	
OSS009	OST0100, OST0102, OST0104, OST0106, OST0108, OST010a
OSS010	OST0101, OST0103, OST0105, OST0107, OST0109, OST010b

FAILURE INCIDENT 08.MAI 2025

May 8 10:07:44 oss010 kernel: sd 1:0:68:0: [sdbz] tag#6047 FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=7s

FAILURE INCIDENT 08.MAI 2025

- Starting at 10 a.m., problems on Lustre I/O operations
- At around 3 p.m., we were informed by the HPC group.
- noticed that zpool operations hanging.
- The kernel logs: a hard disk was causing I/O timeouts.
- The bad disk pulled. ZFS is operational again.
- To prevent colliding, the servers were switched to manual operation and restarted.

FAILURE INCIDENT 08.MAI 2025

„zpool import OST010b“ result:

cannot import 'OST010b': I/O error
Destroy and re-create the pool from
a backup source.

FAILURE INCIDENT 08.MAI 2025

```
  pool: OST010b
    id: 6548008278833985886
 state: FAULTED
status: One or more devices contains corrupted data.
action: The pool cannot be imported due to damaged devices
        or data.
   see: https://openzfs.github.io/openzfs-docs/msg/ZFS-8000-5E
config:

        OST010b                 FAULTED  corrupted data
          draid3:12d:42c:2s-0   DEGRADED
            j022-42             ONLINE
            j022-43             ONLINE
            j022-44             ONLINE
            j022-45             ONLINE
            j022-46             ONLINE
            j022-47             ONLINE
            j022-48             ONLINE
            j022-49             ONLINE
            j022-50             ONLINE
            j022-51             ONLINE
            j022-52             ONLINE
            j022-53             ONLINE
            j022-54             ONLINE
            j022-55             ONLINE
            j022-56             ONLINE
            j022-57             ONLINE
            j022-58             ONLINE
            j022-59             ONLINE
            j022-60             ONLINE
            j022-61             ONLINE
            j022-62             ONLINE
            j022-63             ONLINE
            spare-22            DEGRADED
              j022-64           UNAVAIL
              draid3-0-0        ONLINE
            j022-65             ONLINE
            j022-66             ONLINE
            j022-67             ONLINE
            j022-68             ONLINE
            j022-69             ONLINE
            j022-70             ONLINE
            j022-71             ONLINE
            j022-72             ONLINE
            j022-73             ONLINE
            j022-74             ONLINE
            j022-75             ONLINE
            j022-76             ONLINE
            j022-77             ONLINE
            j022-78             ONLINE
            j022-79             ONLINE
            j022-80             ONLINE
            j022-81             ONLINE
            j022-82             ONLINE
            j022-83             ONLINE
```

FAILURE INCIDENT 08.MAI 2025

```
  pool: OST010b
        id: 6548008278833985886
     state: FAULTED
status: One or more devices contains corrupted data.
  action: The pool cannot be imported due to damaged devices or data.
     see: https://openzfs.github.io/openzfs-docs/msg/ZFS-8000-5E
  config:
...
```

TIME LINE LOG FILES ANALYSIS

Event on OSS010
Disk j022-64 fails. The Linux Kernel reports multiple I/O errors and timeouts at various levels
The pacemaker umounted the affected OST and tried to export the zfs pool
Time out of the zfs pool export; OSS010 is fenced

```
May  8 10:08:25 oss010 kernel: sd 1:0:68:0: [sdbz] tag#1334 FAILED Result: hostbyte=DID_TIME_OUT driverbyte=DRIVER_OK cmd_age=30s
May  8 10:07:44 oss010 kernel: zio pool=l1fs-OST010b vdev=/dev/mapper/j022-64 error=61 type=1 offset=5876068593664 size=90112 flags=180880
May  8 10:07:51 oss010 zed[2663015]: eid=32 class=io pool='l1fs-OST010b' vdev=j022-64 size=90112 offset=5876068593664 priority=0 err=61
flags=0x180880 delay=7458ms bookmark=387:6001105:0:0
May  8 10:08:25 oss010 multipathd[20267]: sdbz: mark as failed
May  8 10:08:25 oss010 multipathd[20267]: j022-64: remaining active paths: 1
May  8 10:08:26 oss010 multipathd[20267]: j022-64: sdbz - tur checker timed out
May  8 10:08:26 oss010 multipathd[20267]: checker failed path 68:208 in map j022-64
May  8 10:08:26 oss010 multipathd[20267]: j022-64: sdes - tur checker timed out
May  8 10:08:26 oss010 multipathd[20267]: checker failed path 129:64 in map j022-64
May  8 10:08:26 oss010 multipathd[20267]: j022-64: remaining active paths: 0
May  8 10:08:26 oss010 kernel: device-mapper: multipath: 253:82: Failing path 129:64.
May  8 10:08:26 oss010 multipathd[20267]: sdbz: mark as failed
May  8 10:08:26 oss010 multipathd[20267]: sdes: mark as failed
```

TIME LINE LOG FILES ANALYSIS

Event on OSS009	Event on OSS010
The Pacemaker started to import all ZFS pools from OSS010 on OSS009 (odd pools)	…booting… System start. ZFS pool import target started to import own odd pools.
Time out of the import of the every odd pool	
The pacemaker tried to export the pools: Success only by OST0103 and the own even pools	
Time out of the export of the OST0101, OST0105, OST0107, OST0109, OST010b	
OSS009 is fenced	
…restarting…	The pacemaker tried to import all pools. Time out of the importing. The pacemaker tried to export all pools. Only OSS0100, OSS0102, OSS0104, OSS0109 successfully
System started	
The pacemaker tried to import all pools.	

TIME LINE LOG FILES ANALYSIS

Event on OSS009	Event on OSS010
The pacemaker tried to import all pools.	
Time out of the importing. The pacemaker tried to export all pools : successfully	
„shutdown –r now“ manually	Any zpool command hangs Smartctl request for J022-64 stalled J022-64 is manually pulled(!!! While two servers attempted to import the same ZFS pool ???)
	„shutdown –r now“ manually
Started.	Started. „zpool import OST010b“ result: cannot import 'OST010b': I/O error Destroy and re-create the pool from a backup source.

ANALYSIS OF EVERYTHING THAT IS AFFECTED

PFL pool JGU (on OST00{00..1f})	PFL pool HIM (on OST01{00..0b})
/lustre/jgu	/lustre/him

- There are files in both PFL directories that are located on all OSTs (stripe -c -1)
  - typically log files, or redirected output like "command >> file" created with "open(O_APPEND)“ => default -1-stripe and no pool
  - configurable via "mdd.*.append_stripe_count" and "mdd.*.append_pool“

NEXT STEPS TO END THE DOWN TIME

PFL pool JGU (on OST00{00..1f})	PFL pool HIM (on OST01{00..0b})
/lustre/jgu	/lustre/him

- Disconnected all Lustre-clients
- Moved all affected files as a tree into a separated directory
- Disable the corrupted OST
- Configure "mdd.*.append_stripe_count" and "mdd.*.append_pool“

DEBUGGING ON ALMALINUX 8.10

- zpool import −f −o readonly=on OST010b
- zpool import −f -F −o readonly=on OST010b
- zpool import −f -F -X −o readonly=on OST010b
- cat /proc/spl/kstat/zfs/dbgmsg

-F : Recovery mode for a non-importable pool. Attempt to return the pool to an importable state by discarding the last few transactions. Not all damaged pools can be recovered by using this option. If successful, the data from the discarded transactions is irretrievably lost. This option is ignored if the pool is importable or already imported.
-X : Used with the -F recovery option. Determines whether extreme measures to find a valid txg should take place. This allows the pool to be rolled back to a txg which is no longer guaranteed to be consistent. Pools imported at an inconsistent txg may contain uncorrectable checksum errors. For more details about pool recovery mode, see the -F option, above. WARNING: This option can be extremely hazardous to the health of your pool and should only be used as a last resort.

```
1747321800   spa_misc.c:404:spa_load_failed(): spa_load(l1fs-OST010b, config untrusted): FAILED: unable to retrieve MOS config
1747321800   spa_misc.c:419:spa_load_note(): spa_load(l1fs-OST010b, config untrusted): UNLOADING
```

DEBUGGING ON FREEBSD 14.3 AND DEBIAN 13

modified ZDB and ZFS library(spa_misc.c, vdev.c)

- zdb −e −AF −o readonly=on OS010b
- zdb −e −AAF −o readonly=on OS010b
- zdb −e −AAF −t <txg> −o readonly=on OS010b
- zdb −e −AAAFX −o readonly=on OS010b

If successfull then zdb −−backup ...

results:

„spa_misc.c:404:spa_load_failed(): spa_load(l1fs-OST010b, config trusted): FAILED: error loading spares nvlist

spa_misc.c:419:spa_load_note(): spa_load(l1fs-OST010b, config trusted): UNLOADING

ZFS_DBGMSG(zdb) END“

FINDINGS

- ZFS Pool OST010b: vdev configuration might be corrupted
  - Because ZFS multihost protection seems to have failed due to strange hardware errors
- The failed disk j022-64 is not readable
- The magn. surface of the failed disk j022-64 may still contain the healthy ZFS metadata
- New plan:
  - Recovery the data from the failed disk(sector by sector copy if possible)
  - Replace the failed disk with a new one as a clone
  - Try to import or make a backup with zdb

RECOVERY OF THE FAILED DISK(S)

```
…
=== START OF INFORMATION SECTION ===
Vendor:               SEAGATE
Product:              ST20000NM002D
…
  Pending defect count:2632 Pending Defects: index, LBA and accumulated_power_on_hours follow
     1:  0x20           ,  10731
     2:  0x21           ,  10731
    …
  1006:  0x4ad          ,  10731
  1007:  0x4ae          ,  10731

[GLTSD (Global Logging Target Save Disable) set. Enable Save with '-S on']
Self-test execution status:                    100% of test remaining
SMART Self-test log
Num  Test              Status                 segment  LifeTime  LBA_first_err [SK ASC ASQ]
     Description                              number   (hours)
# 1  Reserved(3)      Failed in segment -->      -   10730       2136693681 [-   -    -]

Long (extended) Self-test duration: 104460 seconds [29.0 hours]
```

RECOVERY OF THE FAILED DISK(S)

Try to initialize the relocation of bad pending blocks using SeaChestUtilities

```
# SeaChest_SMART -d /dev/sdcg --seagateClean --errorLimit 1000l --confirm
I-understand-this-command-may-erase-single-sectors-if-they-are-already-
unreadable
```

RECOVERY OF THE FAILED DISK(S)

https://www.gnu.org/software/ddrescue/ddrescue.html

- ddrescue -d -r 10 /dev/mapper/j022-64 j022-64.img j022-64.mapfile
- Replace the failed disk with a new one
- ddrescue -f j022-64.img /dev/mapper/j022-64 j022-64rest.mapfile

RECOVERY OF THE FAILED OST

- ddrescue images of all 42 disks of the OST(found and replaced two more bad disks)
- zpool import −f −FX −o readonly=on OST010b (a test)
- zpool import −f −FX OST010b
  - Success after 77 hours!

Is this OST healthy?

MOVE ALL DATA FROM OST010B

- mkdir /lustre/recovered
- lfs setstripe -O 256-266 /lustre/recovered
- Copy the tree
  - **lfs find** /lustre/ -O 267 -print0|parallel -j 64 rsync ...
- Check copy...
- Delete files(?)
  - **lfs find** /lustre/ -O 267 -type f -print0|parallel -j 64 rm -f

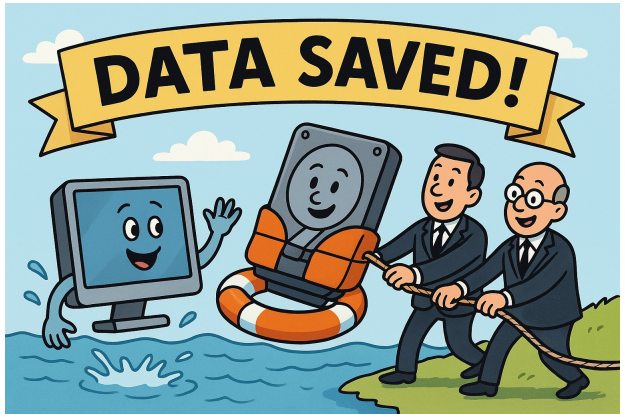
O. Tange (2018): GNU Parallel 2018, March 2018, https://doi.org/10.5281/zenodo.1146014.

DISABLE THE OLD OST010B IN THE LUSTRE

- On every MDS Server with the corresponding MDT:
  - lctl --device OST010b-osc-MDT000${n} deactivate
  - lctl set_param -P osc.OST010b-osc-MDT000${n}.active=0
  - lctl set_param -P osp.OST010b-osc-MDT000${n}.max_create_count=0

RECOVERY OF THE FAILED OST

With the old OST name	With a new OST
• Backup lustre data/variables from OST010b using zfs send	• Destroy the old zfs pool OST010b
• Destroy and recreate the zfs pool OST010b	• Create a new target OST010c
• Restore lustre data/variables on the new OST010b using zfs receive	• format using mkfs.lustre as a new OST OST010c with a new index
• Enable and mount the new OST010b	• Enable and mount the new OST010c



CONCLUSION

- ZFS multihost protection seems to have failed due to strange hardware errors
  - As a suggestion: An implementation of the ZFS multihost protection using SCSI Persistent Reservation
- ZFS is a (the most?) powerful and robust file system
- No data was lost.
- There are several scenarios for recovering data

THANK YOU FOR YOUR ATTENTION

We would like to express our sincere thanks to the

- Lustre OSD ZFS maintainers team in the **Livermore Computing division at LLNL**, specially to **Tony Hutter** for the help by the analysis of the affected ZFS pool.
- Lustre Discuss Forum, specially to **Andreas Dilger** for explaining the magical secrets of the Lustre file system
- **ABC Systems** for the quick delivery of the required spare parts.
- **Seagate Technology** for providing with the necessary information