
JOHANNES GUTENBERG
UNIVERSITÄT MAINZ
JG|U

LUSTRE-INSTALLATION AN DER JGU MAINZ

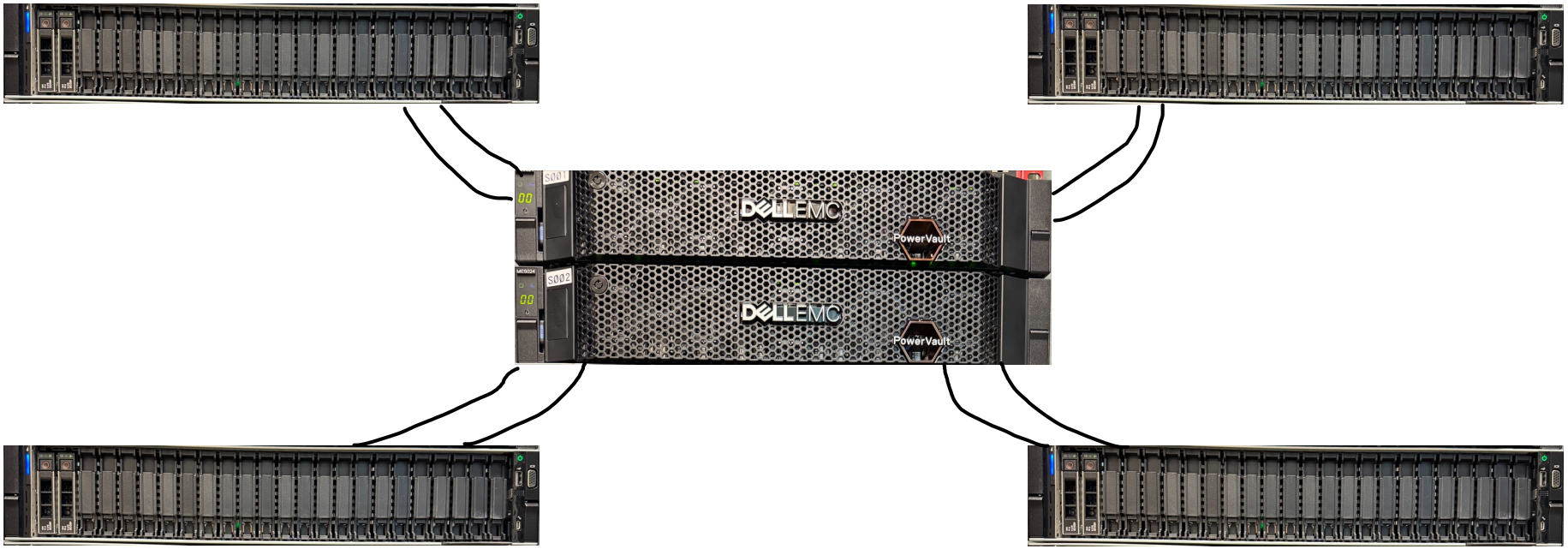
Sergey Noskov, Markus Tacke

27.02.2025

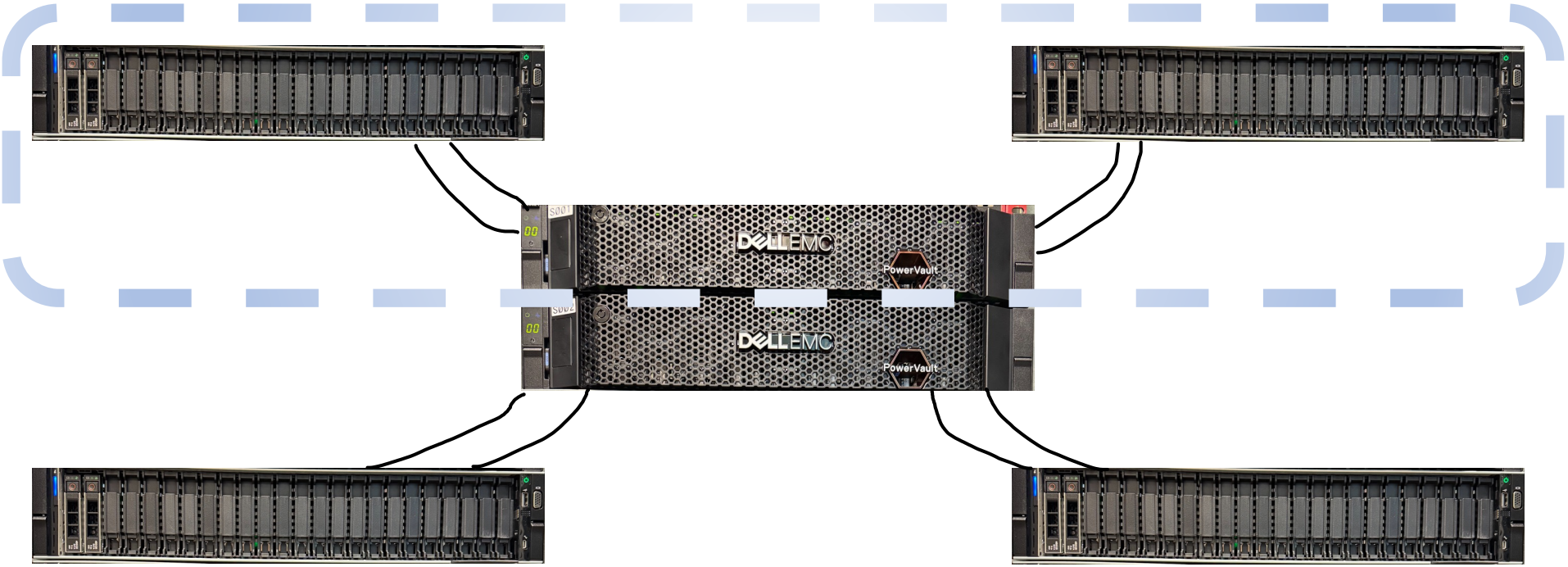
HARDWARE

Komponent	Anzahl
MDS Server	4
SSD-Enclosure x RAID-Kontroller für MD	2 x 2
OSS Server	10
Disk Enclosures für Object store	22

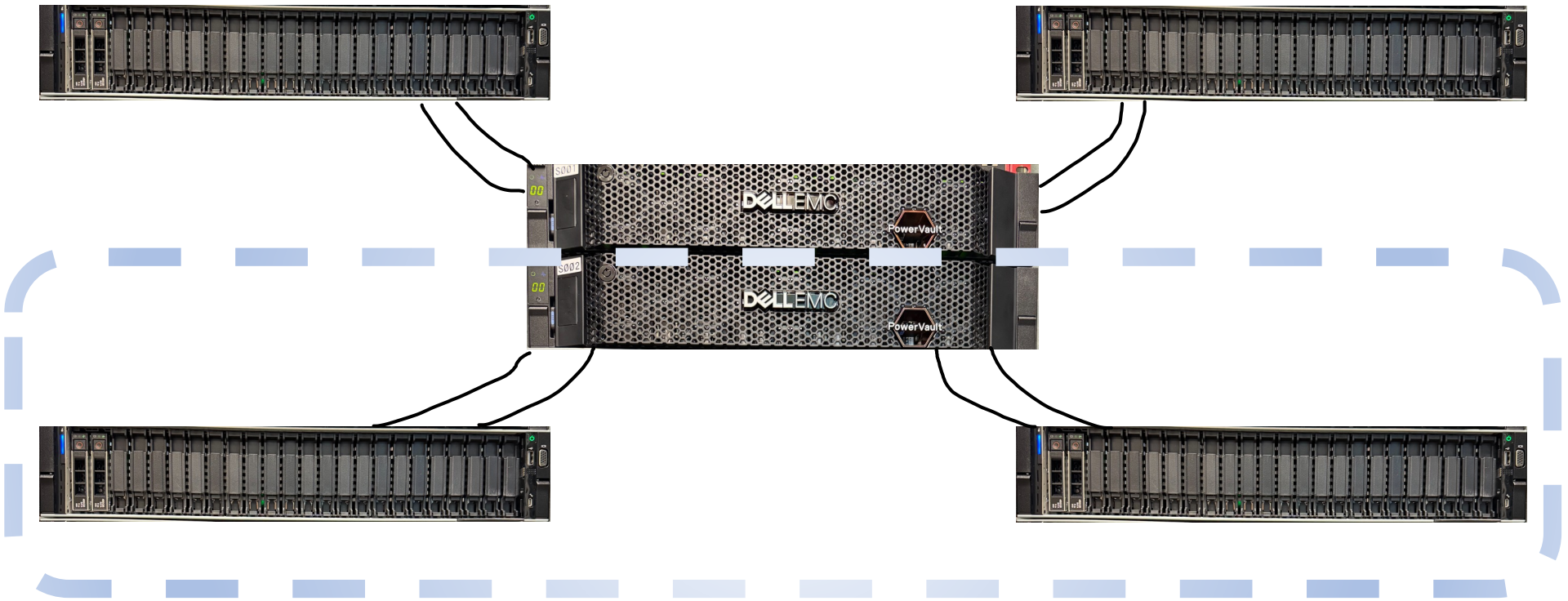
HARDWARE: LUSTRE METADATA STORAGE



HARDWARE: LUSTRE METADATA STORAGE



HARDWARE: LUSTRE METADATA STORAGE



HARDWARE: 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 affin. CPU1 Port1 -> Infiniband 100Gb Port2 -> Ethernet 100Gb (bond mit der 2.Karte)
SAS Adapter	2x Dell HBA355e PCIe gen.4 affin. CPU2
system volume	Raid PERC H745 with 2x 446.63 GB (mirror)

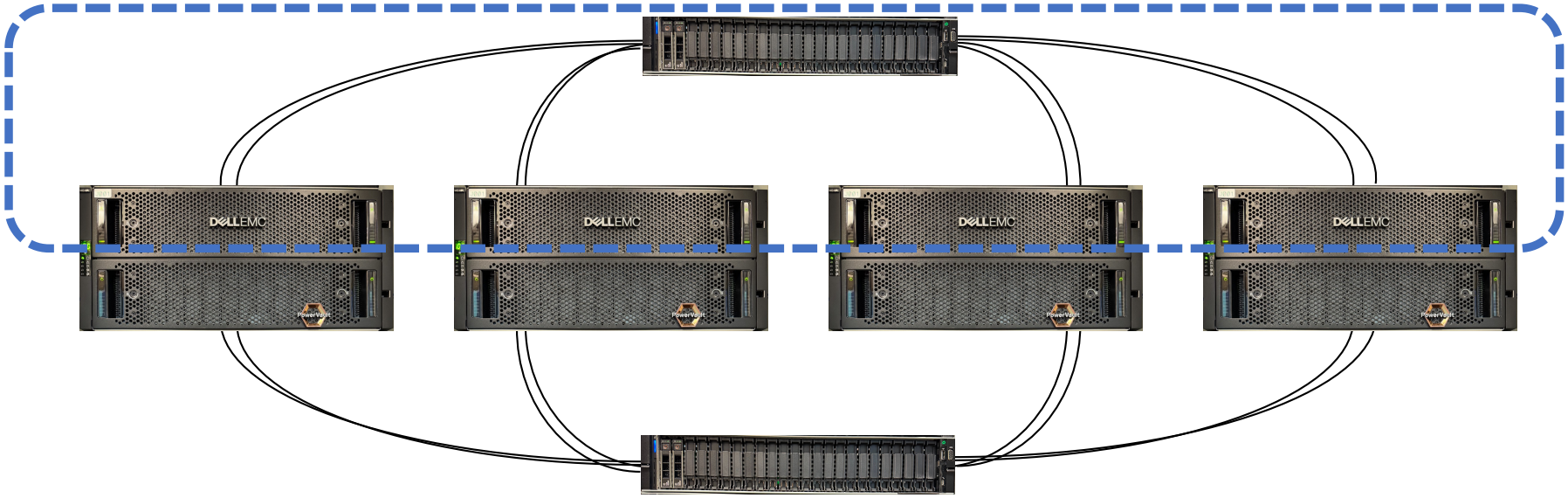
HARDWARE: RAID-ENCLOSURE FÜR METADATA

Dell	PowerVault ME5024
Anzahl der Kontroller	2 (jeweils 4x SAS Konnektors)
Anzahl der SSD	24
SSD Größe	7.6 TB
Disk Gruppen	61.4 TB + 61.4 TB (raid type: ADAPT)
Volumes enclosure 1	mdt000 = 61.4 TB mdt001 = 57.4 TB mgt = 3.9 TB
Volumes enclosure 2	mdt002 = 61.4 TB mdt003 = 61.4 TB

HARDWARE: LUSTRE OBJECT STORAGE

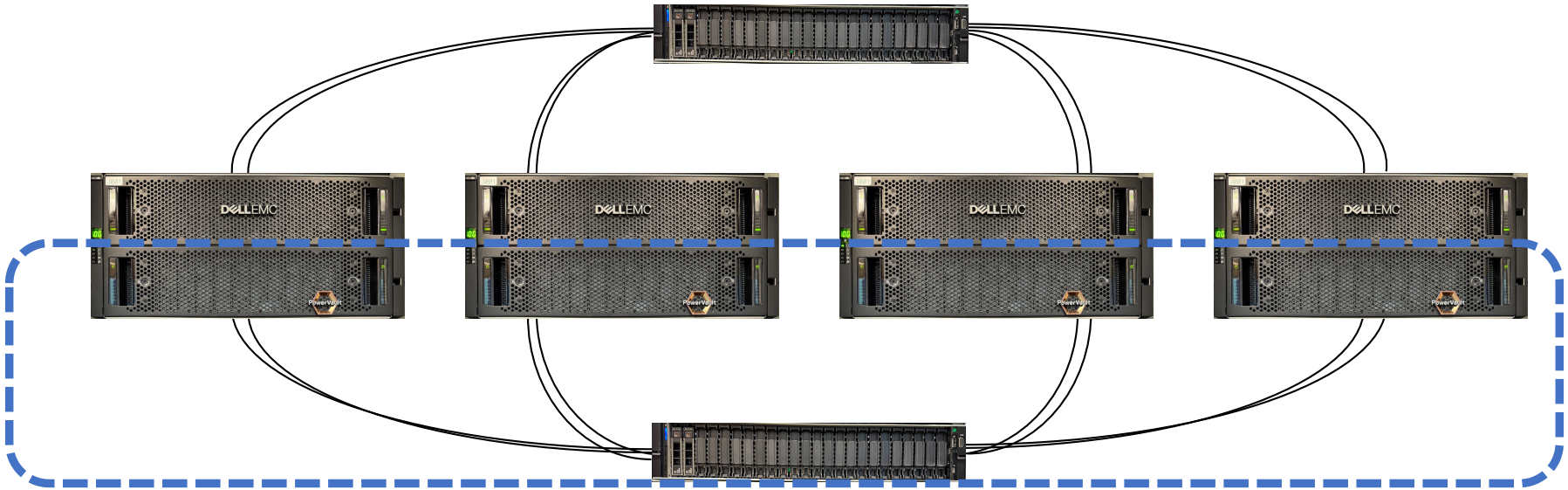
Anzahl OSS Server	Enclosures x Disks
8	16 x 84

HARDWARE: LUSTRE OBJECT STORAGE



Anzahl OSS Server	Enclosures x Disks
8	16 x 84

HARDWARE: LUSTRE OBJECT STORAGE

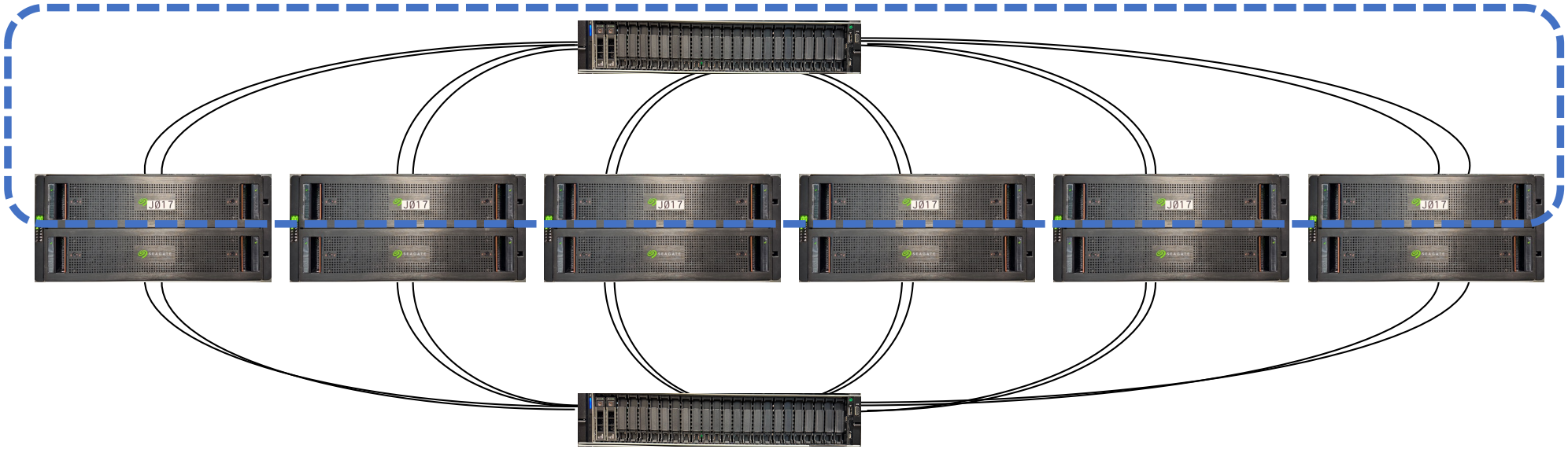


Anzahl OSS Server	Enclosures x Disks
8	16 x 84

HARDWARE: LUSTRE OBJECT STORAGE (GSI)

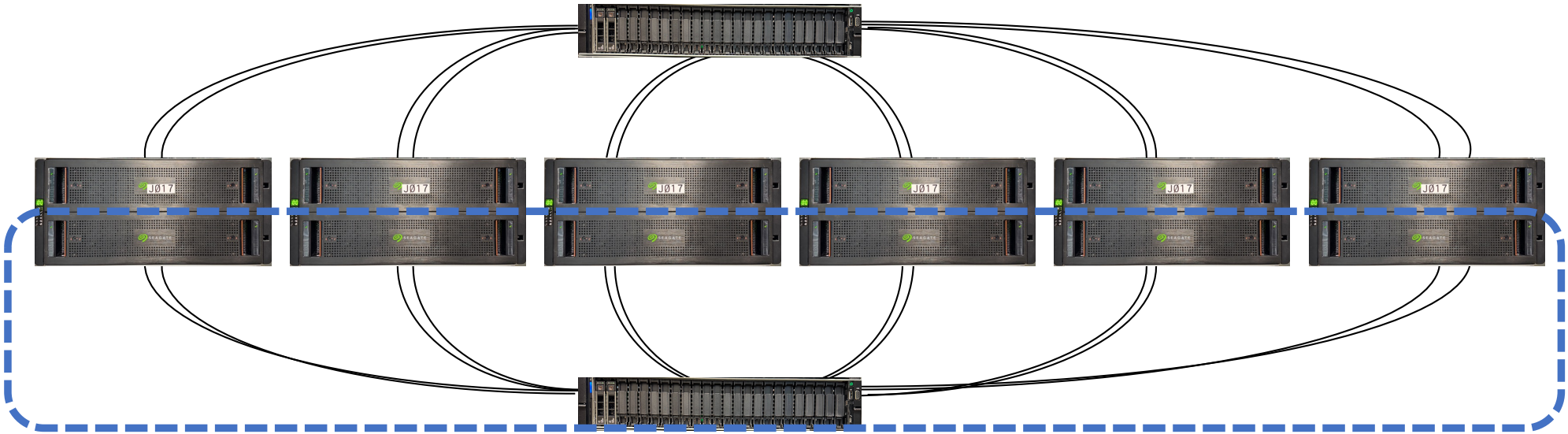
Anzahl OSS Server	Enclosures x Disks
2	6 x 84

HARDWARE: LUSTRE OBJECT STORAGE (GSI)



Anzahl OSS Server	Enclosures x Disks
2	6 x 84

HARDWARE: LUSTRE OBJECT STORAGE (GSI)



Anzahl OSS Server	Enclosures x Disks
2	6 x 84

HARDWARE: 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 PCIe gen.4 affin. CPU1 Port1 -> Infiniband 100Gb Port2 -> Ethernet 100Gb (bond mit der 2.Karte)
SAS Adapter	2x Dell HBA355e PCIe gen.4 affin. CPU2
system volume	Raid PERC H745 with 2x 446.63 GB (mirror)

DATEISYSTEM GRÖßE

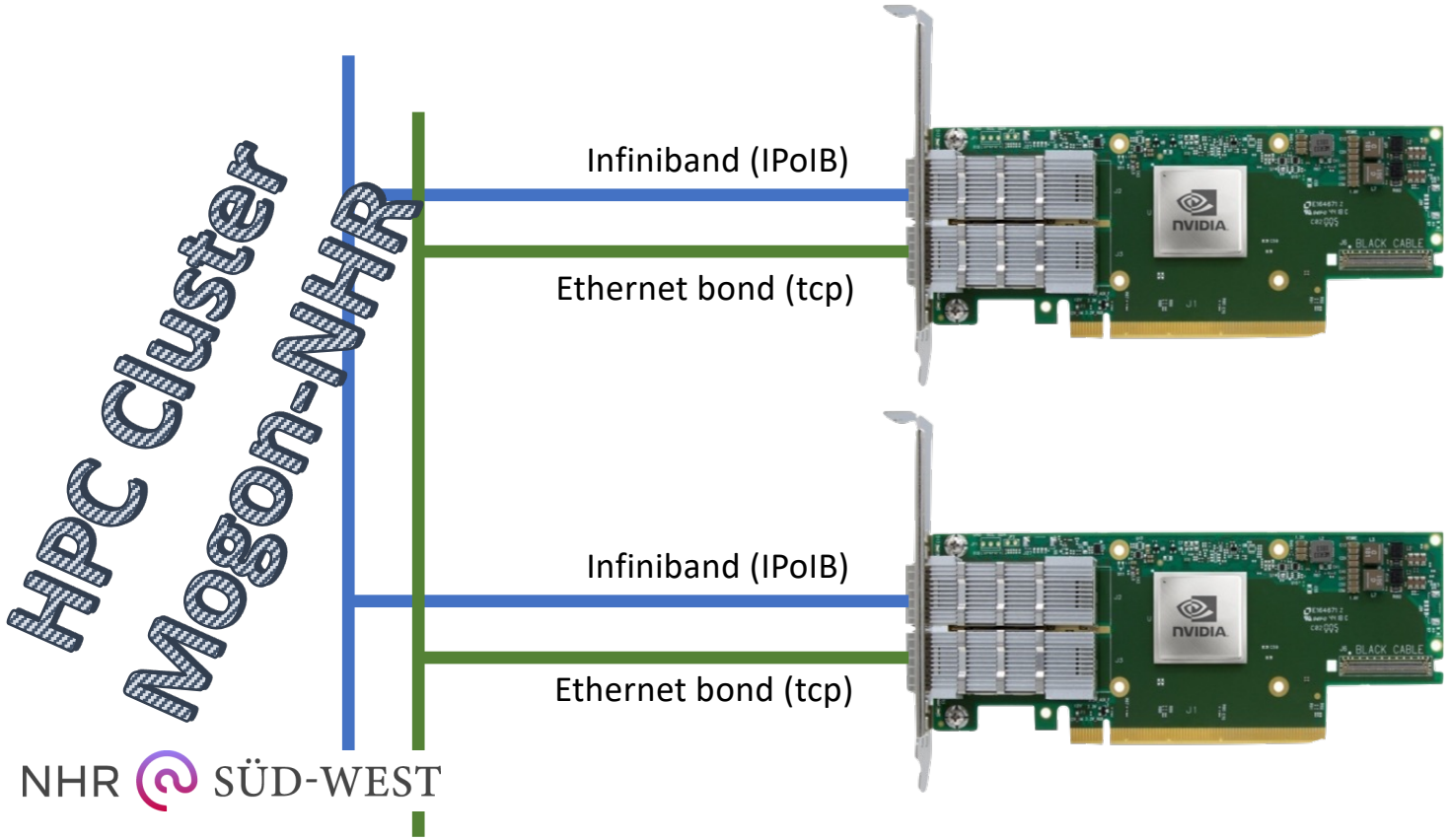
HA-Paar	Targets	RAID typ	Größe, TB	Pool
l1mds001 l1mds002	mgt mdt0000 mdt0001	Dell „ADAPT“	3.62 55.3 51.7	- (DOM)
l1mds003 l1mds004	mdt0002 mdt0003	Dell „ADAPT“	55.3 55.3	- (DOM)
l1oss001 l1oss002	OST000{0..7}	draid2:11d:42c:2s	8x 473 = 3784	allgemein
l1oss003 l1oss004	OST000{8..15}	draid2:11d:42c:2s	8x 473 = 3784	allgemein
…	…	draid2:11d:42c:2s	…	allgemein
l1oss009 l1oss010	OST00{32..43}	draid3:12d:42c:2s	12x 513 = 6156	HIM(GSI)

Insgesamt ~ 20.8 PB

KONNEKTIVITÄT

HA-Paar	Targets	Adapter
l1mds001 l1mds002	mgt mdt0000 mdt0001	2x ConnectX-6 VPI adapter card; H100Gb/s (HDR100; EDR IB and 100GbE)
l1mds003 l1mds004	mdt0002 mdt0003	2x ConnectX-6 VPI adapter card; H100Gb/s (HDR100; EDR IB and 100GbE)
l1oss001 l1oss002	OST000{0..7}	2x ConnectX-6 VPI adapter card; H100Gb/s (HDR100; EDR IB and 100GbE)
l1oss003 l1oss004	OST000{8..15}	2x ConnectX-6 VPI adapter card; H100Gb/s (HDR100; EDR IB and 100GbE)
…	…	2x ConnectX-6 VPI adapter card; H100Gb/s (HDR100; EDR IB and 100GbE)
l1oss009 l1oss010	OST00{32..43}	2x ConnectX®-6 InfiniBand/Ethernet adapter card, **HDR IB** (200Gb/s) and 200GbE, **dual**-port QSFP56

KONNEKTIVITÄT



CLUSTER MOGON-NHR ANBINDUNG

- Inifiniband HDR100
- Topology Fat tree, **aber** GPU partition und Fileserver in Level2 **1/3 Verbindungen**

LUSTRE FEATURES

pool	DOM	PFL	ZFS Komprimierung
„allgemein“	Bis 1MB	Ab 1 MB c=4	-
„HIM“	-	c=1, ab 1MB c=4	-

LUSTRE TUNNING

LUSTRE NET TUNNING

Parameter	Wert (IB)	Wert(tcp)
peer_timeout	180	180
peer_credits	16	8
peer_buffer_credits	0	
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 TUNNING

Parameter	Wert
zfs_arc_max	197656233 (¾RAM)
zfetch_max_distance	67108864
zfs_delay_scale	100000
zfs zfs_dirty_data_max	17179869184 (~2s)
zfs_vdev_aggregation_limit	1048576
vdev_max_active	336
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

ZUKUNFTSWÜNSCHE

- Tuneables optimieren
- mehr MDS / eventuell getrennte MDS für das Pool des HIMs
- Netzwerk Struktur optimieren
- Cluster-Anbindung optimieren (mehrere Cluster?)

DANKE FÜR IHRE AUFMERKSAMKEIT

Wir danken auch allen, die beitragen, dass es besser wird.