Case Study NRH@FAU: Architecture, Tunings and Data Protection of the Fastest Lustre Cluster in the IO500 List

Davide Villa – CRO, Xinnor
Daniel Landau – Solution Architect, Xinnor
Markus Hilger – Senior HPC Engineer, MEGWARE

Friedrich-Alexander-Universität (FAU)

FAU hosts the **Erlangen National High-Performance Computing Center** (NHR @ FAU), one of Germany's nine national HPC centers.

Recently, NHR@FAU has deployed a new GPU cluster named **Helma** made of:

- 192 dual-socket AMD EPYC 9554 "Genoa" compute nodes
- 768 NVIDIA H100/H200 GPUs

Helma is the most powerful university-owned AI supercomputer in Germany and #51 on the June 2025 TOP500 list.

Helma Storage Cluster

- Lustre 2.16.1
- 6.4PB raw and 4.9PB Net NVMe PCIe Gen5
- 10x Celestica SC6100 Storage Bridge Bay Systems
- Data protection assured by xiRAID:
  - MDT in RAID10 (2+2)
  - OST in RAID6 (8+2)
  - High availability between the 2 server nodes within Celestica system via Pacemaker and Corosync integration

IO500 Production ISC25 List

Production | 10 Node Production | Research | 10 Node Research | Full | Historical

#	INFORMATION							IO500		
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE	BW (GIB/S)	MD (KIOP/S)	REPRO.
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.90	10,066.09	102,785.41	
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	
3	ISC25	Erlangen National High Performance Computing Center	Helma	MEGWARE	Lustre	186	18,600	838.99	438.62	1,604.84	
4	ISC25	Samsung Electronics	SSC-24	WekaIO	WekaIO	291	16,005	826.86	248.67	2,749.41	
5	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	
6	SC24	MSKCC	IRIS	WekaIO	WekaIO	261	27,144	665.49	252.54	1,753.69	
7	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	
8	SC24	SoftBank Corp	CHIE-3	DDN	EXAScaler	240	26,880	500.20	331.66	754.41	
9	ISC25	Joint Center for Advanced High Performance Computing	Miyabi-G	DDN	Lustre	200	1,600	391.60	319.00	480.72	
10	SC24	Danish Centre for AI innovation AS	GEFION	DDN	EXAScaler	128	12,288	368.56	209.06	649.73	

The most efficient IO500 storage cluster

Efficiency reflects achieving high IO500 performance with fewer storage servers, thereby reducing power supplies, network cards, cooling requirements, and overall infrastructure costs per unit of performance delivered.

#	System	Solution (Vendor)	Score	Storage servers	Score / storage server
3	Helma	xiRAID + Lustre (Xinnor)	838.99	20	41.95
4	SSC-24	WekaFS (WekaIO)	826.86	40	20.67
5	Shaheen III	Lustre (HPE)	797.04	160	4.98
7	Leonardo	ExaScaler (DDN)	648.96	29	22.37
9	Miyabi-G	Lustre (DDN)	391.60	44	8.9

The #1 and #2 systems on the IO500 Production list are based on Intel DAOS. These results represent highly customized deployments and are not generally available as standard, off-the-shelf storage solutions.

Helma Storage Cluster at NHR@FAU

5PB HA storage cluster to serve 768 GPUs

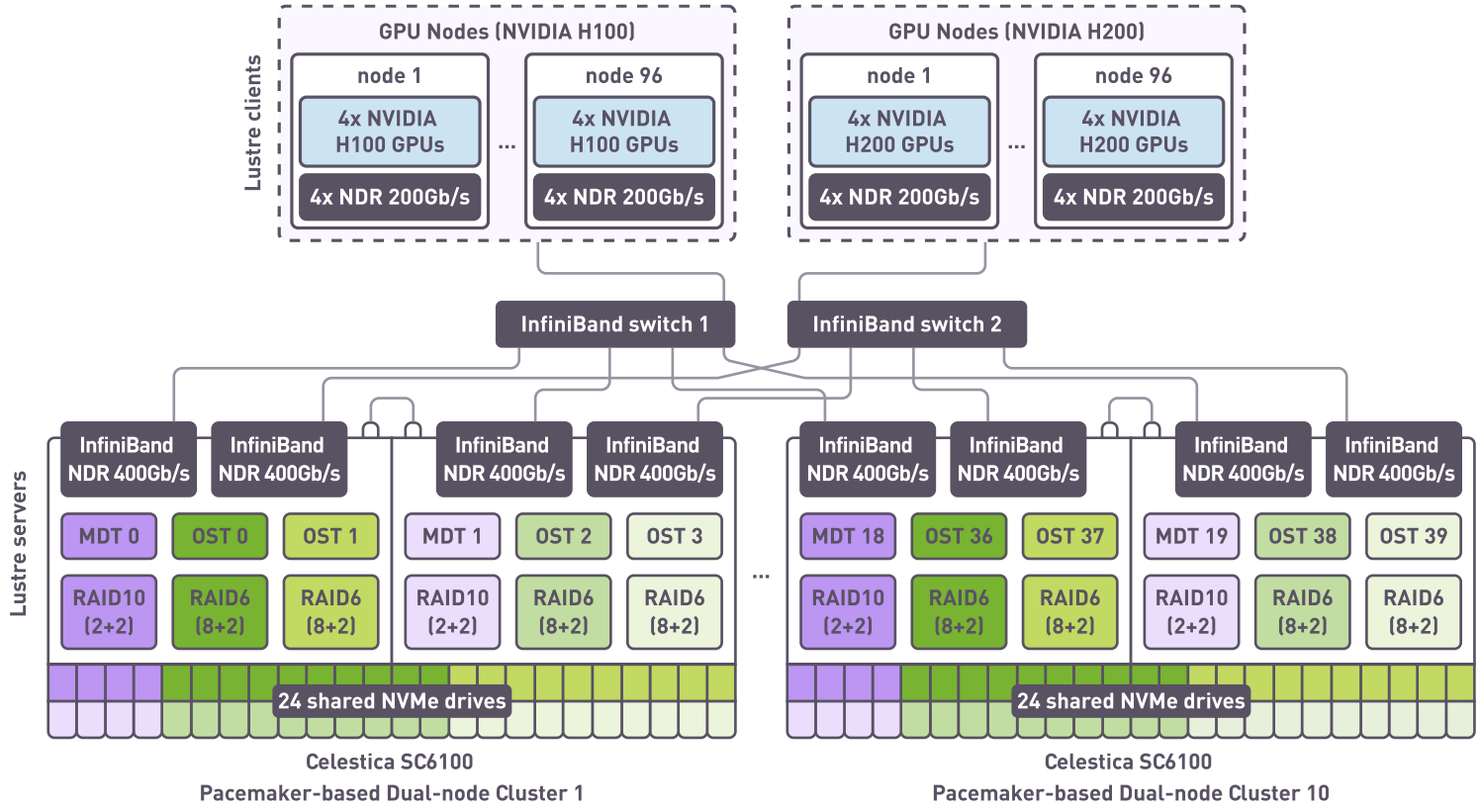
xiRAID + Lustre 2.16.1

10 Celestica SC6100 SBB systems, each made of 2 servers each with:

- AMD EPYC 9454P
- 384GB DDR5 RAM
- 2x NVIDIA CX-7 NDR400

Shared drives between 2 servers:

- 4x Phison Pascari PCIe5 X200E 6.4TB (3DWPD): 2x 2+2 namespaces xiRAID RAID10
- 20x Phison Pascari PCIe5 X200P 30.72TB (1DWPD): 4x 8+2 namespaces xiRAID RAID6 128K stripe size



Platform & OS Configuration

PCIe 5 SBB System

One 2U building block (e.g. Celestica SC6100)

- 12.8 TB Metadata (xiRAID-10)
- 490 TB Data (xiRAID-6)
- 2 NVMe namespaces for each drive for full PCIe 5 x4 throughput (2 MDTs, 4 OSTs)

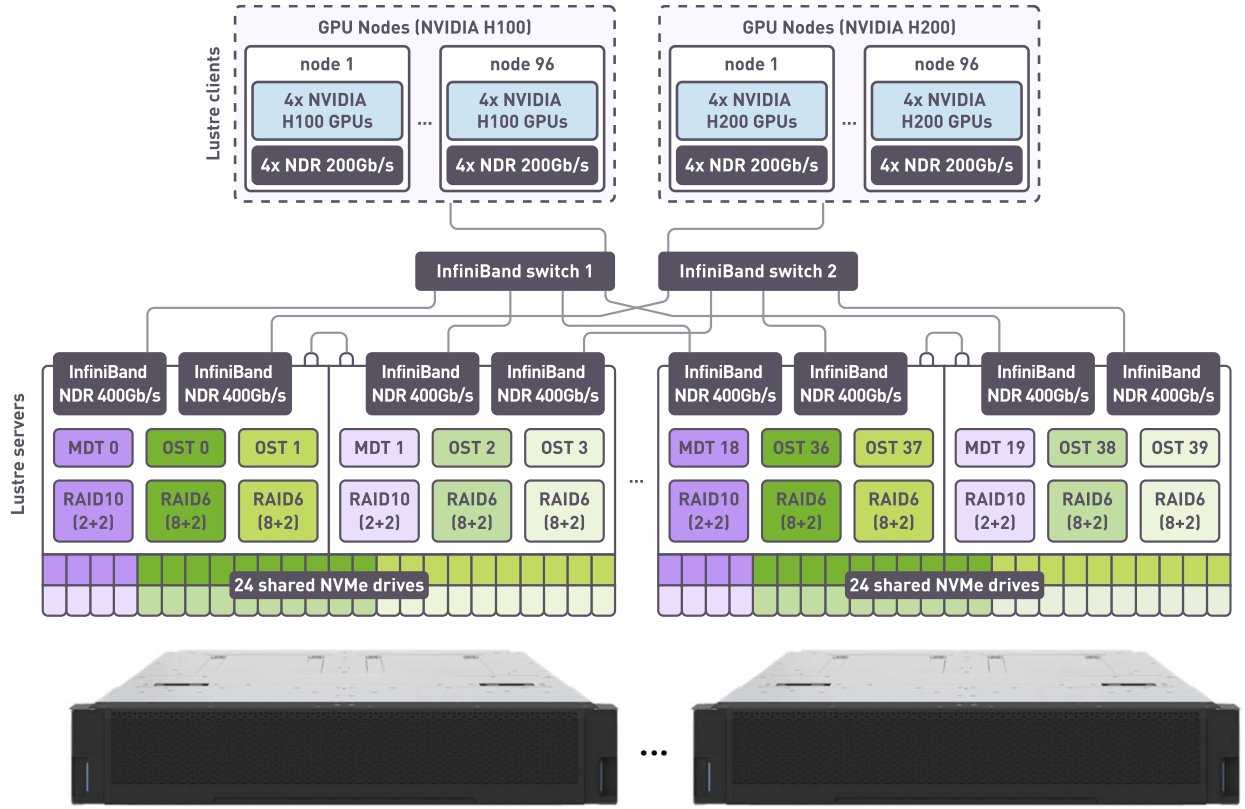
Dual-Port PCIe5 NVMe (e.g. Phison Pascari X200E 30TB)

- 14 GB/s read
- 7.5 GB/s write

Performance

- ~180 GB/s seq. read
- ~80 GB/s seq. write

HA Failover via Pacemaker/Corosync



Celestica SC6100

PCIe 5 SBB System: BIOS Settings

Disabled C/P-States

- Global C-state Control: Disabled
- APBDIS: 1
- DF Pstate: 0
- DF Cstates: Disabled

Disabled any virtualization and abstraction

- SVM | SEV-SNP | SEV Control: Disabled
- SMEE | TSME: Disabled
- SR-IOV Support: Disabled

IOMMU: Enabled

Local APIC Mode: x2APIC

NUMA setup

- NPS: 1 (1 NUMA node per socket)
- ACPI SRAT L3 Cache As NUMA Domain: Disabled

PCIe 5 SBB System: Kernel parameters

`iommu=pt`
- remove unnecessary abstraction layers

`nvme-core.multipath=N`
- Multipathing not needed for local disks

`pci=pcie_bus_perf`
- increases PCIe MaxPayload size (from 128b to 512b on this specific platform)
- required to get full PCIe NVMe bandwidth on this specific platform

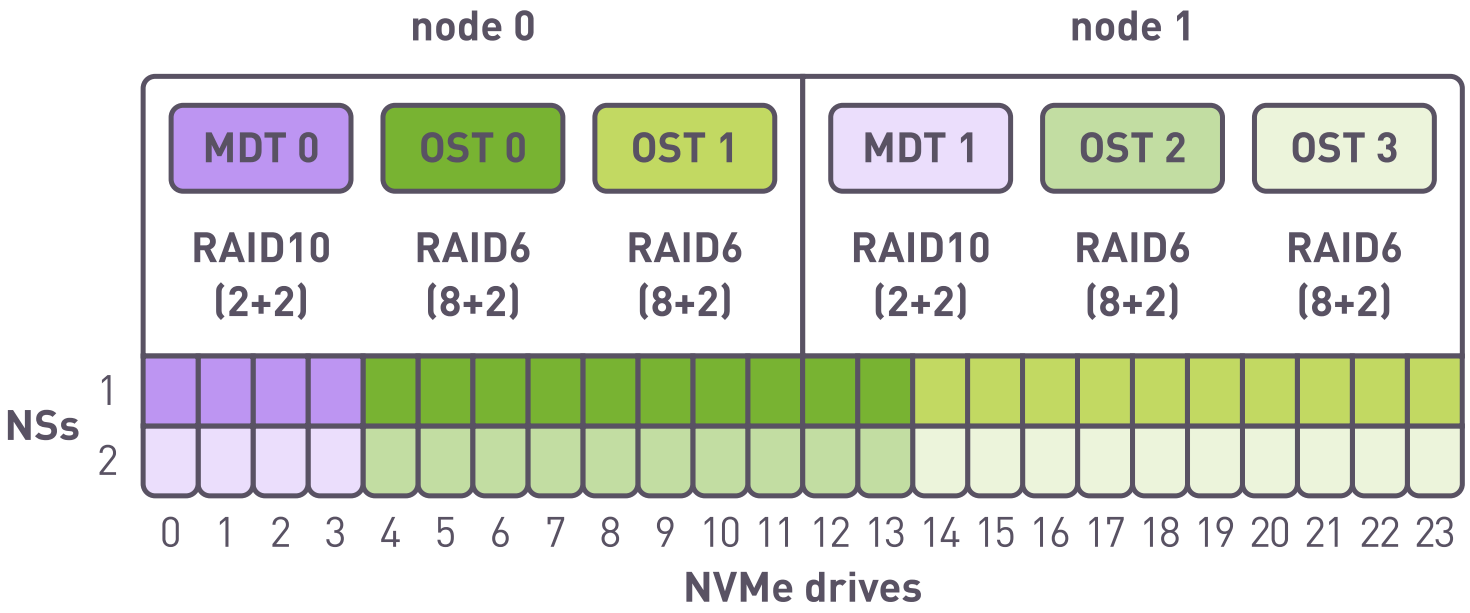
Other settings are default AlmaLinux 9 settings (no tuned, sysctl tunings etc.)

xiRAID and Cluster Configuration

xiRAID RAIDs configuration (1/2)

- RAID6 of 10 drives used for OSTs
- RAID10 of 4 drives used for MDTs
- Each NVMe drive of 24 in Celestica 6100 is connected to each server by 2 PCIe lines only.
- To get full performance of the drives we split them in two namespaces and use first namespaces on the first node and second namespace on the second one.

xiRAID RAIDs configuration (2/2)



RAID10 used for MDT, active on node 0

RAID10 used for MDT, active on node 1

RAID6 used for OST, active on node 0

RAID6 used for OST, active on node 1

RAID6 used for OST, active on node 0

RAID6 used for OST, active on node 1

xiRAID tunings

- Defaults are mostly fine
- Test minimum RAID stripe size for max. sequential performance
  - usually, 64k or 128k
- Be aware of your NUMA topology
- Test different settings for `--sched_enabled 0/1`
  - by default, xiRAID continues execution on currect core
  - distributing this load to other (multiple) cores can help if I/Os are low-threaded

xiRAID settings

MDT0001	size: 5901 GiB level: 10 group_size: 2 strip_size: 16 block_size: 4096 sparepool: - active: True config: True	online initialized	--------gr:01-------- 0 /dev/nvme21n2 online 1 /dev/nvme8n2 online --------gr:02-------- 2 /dev/nvme20n2 online 3 /dev/nvme19n2 online	---- 100% 100% ---- 100% 100%	-- 0% 0% -- 0% 0%	---------- 91F00411_2 91F003Q8_2 ---------- 91F0040K_2 91F003X2_2	init_prio : 100 memory_limit_mb : 0 memory_prealloc_mb : 2048 recon_prio : 100 request_limit : 0 restripe_prio : 100 sched_enabled : 1 cpu_allowed : 36-47,84-95	memory_usage_mb : -
OST0002	size: 113298 GiB level: 6 strip_size: 128 block_size: 4096 sparepool: - active: True config: True	online initialized	0 /dev/nvme0n2 online 1 /dev/nvme2n2 online 2 /dev/nvme1n2 online 3 /dev/nvme3n2 online 4 /dev/nvme6n2 online 5 /dev/nvme7n2 online 6 /dev/nvme9n2 online 7 /dev/nvme5n2 online 8 /dev/nvme10n2 online 9 /dev/nvme11n2 online	100% 100% 100% 100% 100% 100% 100% 100% 100% 100%	0% 0% 0% 0% 0% 0% 0% 0% 0% 0%	91H00341_2 91H003F5_2 91H00310_2 91H002SX_2 91H0030E_2 91H003F0_2 91H0030Y_2 91H00312_2 91H0030L_2 91H0032G_2	init_prio : 100 recon_prio : 100 memory_limit_mb : 0 memory_prealloc_mb : 2048 merge_read_enabled : 0 merge_write_enabled : 0 merge_read_wait_usecs : 300 merge_read_max_usecs : 1000 merge_write_wait_usecs: 300 merge_write_max_usecs : 1000 resync_enabled : 1 sched_enabled : 1 request_limit : 0 restripe_prio : 100 cpu_allowed : 18-35,66-83 adaptive_merge : 0	memory_usage_mb : -

xiRAID High Availability Architecture

- Lustre HA approach is based on Pacemaker clustering
- Cluster with shared storage requires properly configured STONITH (`fence_ipmilan` was used)
- xiRAID Classic includes Pacemaker agent for HA cluster integration

MEGWARE NINNOR

xiRAID High Availability Architecture

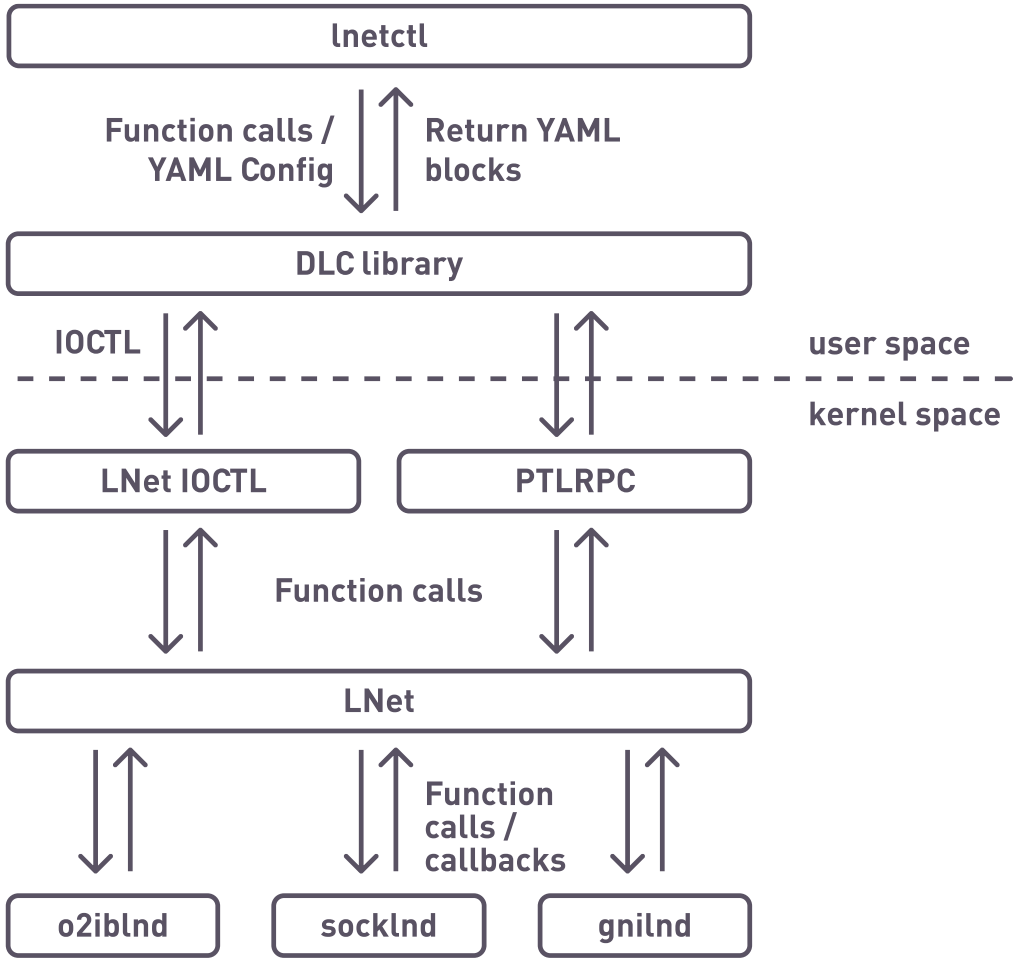
```
Node List:
  * Online: [ hnvme01 hnvme02 ]

Full List of Resources:
  * stonith-hnvme01     (stonith:fence_ipmilan):       Started hnvme02
  * stonith-hnvme02     (stonith:fence_ipmilan):       Started hnvme01
  * Clone Set: healthLUSTRE-clone [healthLUSTRE]:
    * Started: [ hnvme01 hnvme02 ]
  * Clone Set: healthLNET-ibs1-clone [healthLNET-ibs1]:
    * Started: [ hnvme01 hnvme02 ]
  * Clone Set: healthLNET-ibs2-clone [healthLNET-ibs2]:
    * Started: [ hnvme01 hnvme02 ]
  * Clone Set: healthLNET-ens4np0-clone [healthLNET-ens4np0]:
    * Started: [ hnvme01 hnvme02 ]
  * Resource Group: hnvme01-MDT0000:
    * xiraid-hnvme01-MDT0000  (ocf:xraid:raid):  Started hnvme01
    * lustre-hnvme01-MDT0000  (ocf:lustre:Lustre):     Started hnvme01
  * Resource Group: hnvme01-OST0000:
    * xiraid-hnvme01-OST0000  (ocf:xraid:raid):  Started hnvme01
    * lustre-hnvme01-OST0000  (ocf:lustre:Lustre):     Started hnvme01
  * Resource Group: hnvme01-OST0001:
    * xiraid-hnvme01-OST0001  (ocf:xraid:raid):  Started hnvme01
    * lustre-hnvme01-OST0001  (ocf:lustre:Lustre):     Started hnvme01
  * Resource Group: hnvme02-MDT0001:
    * xiraid-hnvme02-MDT0001  (ocf:xraid:raid):  Started hnvme02
    * lustre-hnvme02-MDT0001  (ocf:lustre:Lustre):     Started hnvme02
  * Resource Group: hnvme02-OST0002:
    * xiraid-hnvme02-OST0002  (ocf:xraid:raid):  Started hnvme02
    * lustre-hnvme02-OST0002  (ocf:lustre:Lustre):     Started hnvme02
  * Resource Group: hnvme02-OST0003:
    * xiraid-hnvme02-OST0003  (ocf:xraid:raid):  Started hnvme02
    * lustre-hnvme02-OST0003  (ocf:lustre:Lustre):     Started hnvme02
```

Lustre Setup for IO500 Tests

LNet tunings

```
o2ib20 (2x400G IB NDR):
            tunables:
                  peer_timeout: 180
                  peer_credits: 16 (default 8)
                  peer_buffer_credits: 0
                  credits: 512 (default 256)
            lnd tunables:
                  peercredits_hiw: 8 (default 4)
                  concurrent_sends: 16
                  map_on_demand: true
                  fmr_pool_size: 512
                  fmr_flush_trigger: 384
                  fmr_cache: 1
                  ntx: 512
                  conns_per_peer: 1
                  timeout: 49
                  tos: -1
tcp20 (100 GbE): defaults
```



Lustre tunings (for io500)

MEGWARE XINNOR

Enable Lustre 16MB bulk RPCs:

```
lctl set_param -P obdfilter.*.brw_size=16
lctl set_param -P obdfilter.*.osc.max_pages_per_rpc=4096
lctl set_param -P osc.*.max_pages_per_rpc=4096
```

Allow more RPCs in flight

```
lctl set_param -P mdc.*.max_rpcs_in_flight=128
lctl set_param -P osc.*.max_rpcs_in_flight=128
lctl set_param -P mdc.*.max_mod_rpcs_in_flight=127
lctl set_param -P mdt.*.max_mod_rpcs_in_flight=128 *
```

* new in 2.16, previously mdt module option `max_mod_rpcs_per_client`

Disable memory and wire checksums on clients (~20% performance hit)

```
lctl set_param -P llite.*.checksum_pages=0
lctl set_param -P llite.*.checksums=0
lctl set_param -P osc.*.checksums=0
lctl set_param -P mdc.*.checksums=0
```

Limit locks per client to improve ior-hard and metadata performance a lot

```
lctl set_param ldlm.namespaces.*.lru_size=12800
lctl set_param ldlm.namespaces.*.lru_max_age=1000
```

Miscellaneous

```
lctl set_param -P obdfilter.*.precreate_batch=1024
lctl set_param llite.*.max_read_ahead_mb=2048
lctl set_param osc.*.short_io_bytes=65536
```

IO500 settings

Use "`API = POSIX --posix.odirect`" for ior-easy and new ior-rnd4K-easy-read test

io500.sh settings:

```
lfs setstripe -c 1 $workdir
mkdir $workdir/ior-easy $workdir/ior-hard
mkdir $workdir/mdtest-easy $workdir/mdtest-hard

# Use DNE3 for mdtest-easy, it’s faster than the default DNE2 set by io500.sh!
lfs setdirstripe -D --max-inherit-rr 2 $workdir/mdtest-easy
lfs setdirstripe -D -c -1 $workdir/mdtest-hard

lfs setstripe -c 1 -S 16M $workdir/ior-easy
local osts=$(lfs df --ost $workdir | grep -c OST)
# Try overstriping for ior-hard to improve scaling
lfs setstripe -C $((osts * 5)) -S 16M $workdir/ior-hard
# DoM for mdtest
lfs setstripe -E 1M -L mdt $workdir/mdtest-easy
lfs setstripe -E 1M -L mdt $workdir/mdtest-hard
```

IO500 results


```
IO500 version io500-isc25_v1 (standard)
[RESULT]       ior-easy-write      811.333074 GiB/s : time 455.060 seconds
[RESULT]    mdtest-easy-write     1819.160034 kIOPS : time 411.094 seconds
[      ]            timestamp        0.000000 kIOPS : time 0.002 seconds
[RESULT]       ior-hard-write       60.521807 GiB/s : time 404.143 seconds
[RESULT]    mdtest-hard-write      387.630357 kIOPS : time 385.449 seconds
[RESULT]                 find     3016.995633 kIOPS : time 295.969 seconds
[RESULT]        ior-easy-read     1798.770433 GiB/s : time 205.397 seconds
[RESULT]     mdtest-easy-stat     8221.826113 kIOPS : time 92.063 seconds
[RESULT]        ior-hard-read      419.039776 GiB/s : time 59.941 seconds
[RESULT]     mdtest-hard-stat     3358.073167 kIOPS : time 47.000 seconds
[RESULT]   mdtest-easy-delete     1420.241286 kIOPS : time 527.692 seconds
[RESULT]     mdtest-hard-read     2236.327682 kIOPS : time 68.171 seconds
[RESULT]   mdtest-hard-delete      235.844820 kIOPS : time 634.046 seconds
[      ]  ior-rnd4K-easy-read       87.658845 GiB/s : time 300.833 seconds
[SCORE ] Bandwidth 438.617055 GiB/s : IOPS 1604.836213 kiops : TOTAL 838.992571
```

MEGWARE XINNOR

Read the full case study:

High-Availability NVMe Storage with xiRAID + Lustre for the Helma Supercomputer

xinnor.io/case-studies/helma

MEGWARE XINNOR

Prove it yourself:

xinnor.io

megware.com