LAD24 - Love and hate, 15 years with Lustre

NRIS

Norwegian research infrastructure services

Some context: Who we are

- Tier 0; Eurohpc
- Tier 1; Sigma2 AS :
  - Procures, owns and manage HPC/storage/AI resources on national level.
  - Resources and training is operated by NRIS, joint collaboration between four universities and Sigma2
  - Fram, Betzy, Saga, NIRD, Olivia
- Tier 2; Universities (NTNU, UIO, UIT, UIB)
  - IDUN @ NTNU
  - FOX @UIO

- Present here today:
  - Einar Næss Jensen (einar.nass.jensen@ntnu.no)
  - Andreas Kalvå (andreas.kalva@ntnu.no)

NTNU

- Largest University in Norway
- Unbroken line of HPC systems since Cray X-MP
- Partner with Sigma2
  - for large national systems as part of NRIS
- Current HPC system: IDUN
- Housing for Betzy and Saga
- ~1.8 megawatt heating for the campus. No hotwater import between April and November

University of Bergen: Hexagon

- Lustre FS v1 as SCRATCH file system on Cray in 2008
- Later upgraded to v2
- Later introduced additiona; project space shared between Cray, cluster and smaller computing servers
- Used quotas, specialle group quotas for projects
- Robinhood when it came was used for data reporting and policies, like autocleanup of scratch
- Robinhood was using changelog from the metadaserver
- Was in service until spring 2024
- Was use over Cray interconnect as well as over 100Gb ethernet

Cray XT4 2008 (#47 in Top500)
Cray XE6m-200

NTNU: Vilje 2012-2021 (2023)

#44 on top500

SGI ICE X, IB FDR hypercube

600 TB DDN storage

2 x sfa10K

300TB for scratch (separate mds)

170 TB for weather forecast (shared mds)

130 TB for home (shared mds)

Extremely successful machine. running without (paid) support. (thank you DDN and lustre email list)

SIGMA2: Fram (2017 - 2026?)

ca 80 on top500

Lenovo, EDR infiniband

2.5PB DDN storage

2 x sfa14KE + mdt

One large filesystem

Lots of lots of problems with hardware (not only this one but also on same hardware with GPFS)

SIGMA2: Betzy 2020-2028(?)

BullSequana XH2000

5.6 PB, 2x SFA18E + nv200

8 OSS,24 ost

4 MDS, 4 mdt

scratch has mdt0/1

home has mdt2

software has mdt3

using pools (scratch vs the rest)

Sigma2: Olivia (2025-????)

HPE Cray +slingshot

CPU partition + gpu partition

storage: clusterstore

1PB nvme

3.2PB capacity

(with promise of extreme performance numbers)

TBF rules

very interesting piece of software:
RH 1.5D
https://rh15d.readthedocs.io/en/latest/#
“Using more than 4000 cores and writing full output may cause I/O slowdowns and Lustre contention in some systems”

lctl set_param ost.OSS.ost_io.nrs_policies="tbf jobid"
lctl set_param ost.OSS.ost_io.nrs_tbf_rule="start globallimit jobid={*} rate=3000"

Specific lustre thingies

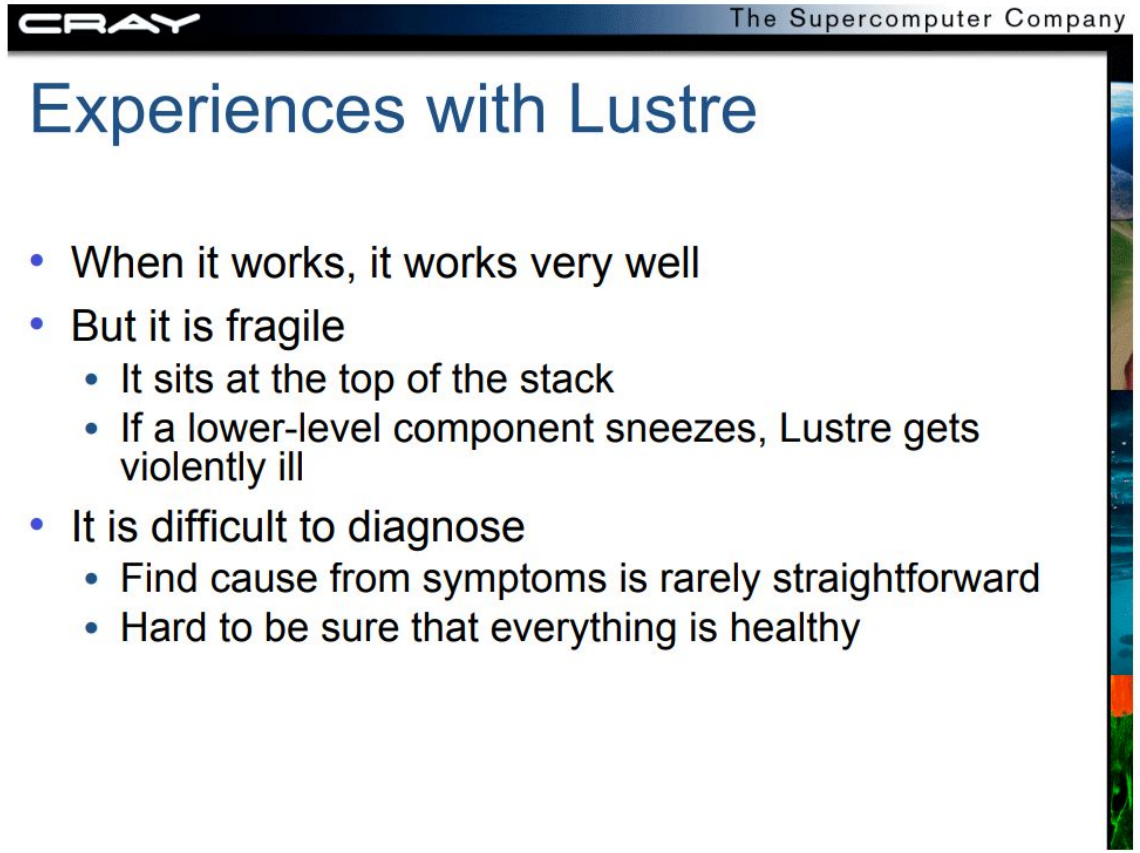
Sporadic attempts at running robinhood

- no full time lustre experts

```
lctl set_param
ost.OSS.ost_io.nrs_policies="tbf jobid"
ost.OSS.ost_io.nrs_tbf_rule="start globallimit
jobid={*} rate=3000"
```

Lustre impressions

- Performance is very good
- When something goes wrong, it can be hard to find culprit.
- Extremely dependent on other components like infiniband (specific mofed versions etc)
- not so easy to find or train people in Lustre
- Management tools are crude and different.



IDUN and BETZY IOR benchmarks

How we benchmark for best possible performance

```
# Create pool for each OST
lctl pool_new cluster.bench_OST0000
. . .
lctl pool_new cluster.bench_OST0017


# Add OST to the pool
lctl pool_add cluster.bench_OST0000 OST[0]
. . .
lctl pool_add cluster.bench_OST0017 OST[17]


# Created directory for the pool
mkdir /cluster/projects/nn9997k/benchmark-osts/OST0000
. . .
mkdir /cluster/projects/nn9997k/benchmark-osts/OST0017


# Associate a directory with an OST pool
lfs setstripe /cluster/projects/nn9997k/benchmark-osts/OST0000 --pool cluster.bench_OST0000
. . .
lfs setstripe /cluster/projects/nn9997k/benchmark-osts/OST0017 --pool cluster.bench_OST0017
```

IOR – each OSTs benchmark

Command example one OST:

```
$ mpirun -n 128 --host
b5102:16,b5103:16,b5104:16,b5105:16,b5106:16,b5107:16,b
5108:16,b5109:16
/cluster/home/pavlokh/ior-3.3.0/src/ior -a POSIX -v -F
-C -e -g -k -b 1m -t 1m -s 10000 -i 1 -w -r -o
/cluster/projects/nn9997k/benchmark-osts/OST0000/
```

OST	Write (MB/sec)	Read (MB/sec)	Production pool
OST0000	4124	5421	18k1
OST0001	4189	5479	18k1
OST0002	3859	3556	18k1
OST0003	3928	3218	18k1
OST0004	4014	3591	18k1
OST0005	4070	3438	18k1
OST0006	3910	3505	18k1
OST0007	3949	3673	18k1
OST0008	3963	3423	18k1
OST0009	3964	3403	18k1
OST000a	4131	3629	18k1
OST000b	4016	3575	18k1
OST000c	2771	4694	projects
OST000d	2888	4377	projects
OST000e	2891	4387	projects
OST000f	2950	4179	projects
OST0010	2926	4462	projects
OST0011	2897	4336	projects
OST0012	2701	4422	projects
OST0013	2954	4228	projects
OST0014	2816	4198	projects
OST0015	2961	4370	projects
OST0016	2908	4063	projects
OST0017	2980	4200	projects
SUM (Theoretical)	**82760**	**97827**	

IOR runs on all OSTs

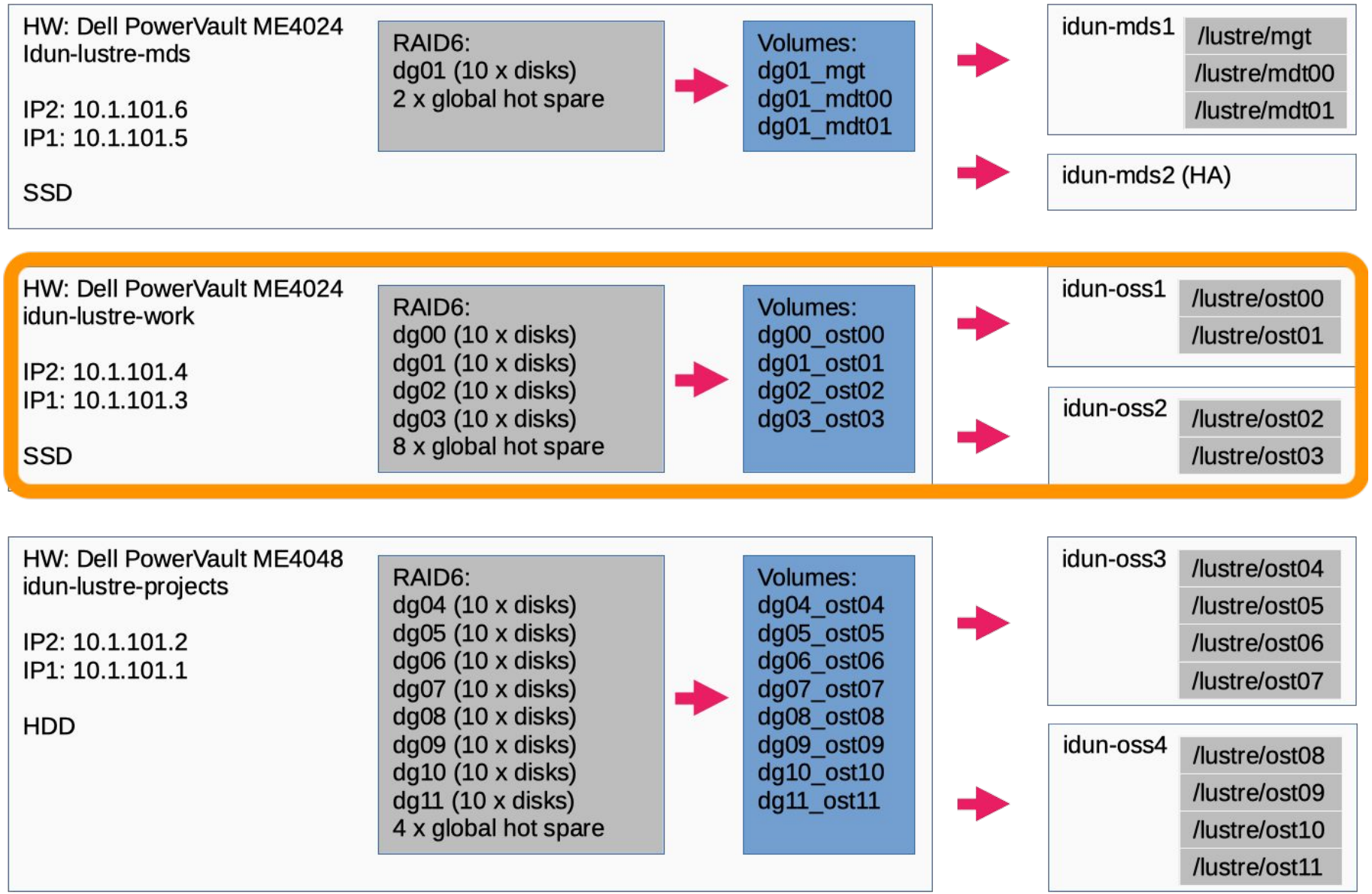
```
mpirun -n 96 --host
b5102:4,b5103:4,b5104:4,b5105:4,b5106:4,b5107:4,b5108:4,b5109:4,b5110:4,b5111:4,b5112:4,b5113:4,b5114:4,b5115:4,b5116:
4,b5117:4,b5118:4,b5119:4,b5120:4,b5121:4,b5122:4,b5123:4,b5124:4,b5125:4 /cluster/home/pavlokh/ior-3.3.0/src/ior -a
POSIX -v -F -C -e -g -k -b 1m -t 1m -s 100000 -i 1 -w -r -o
/cluster/projects/nn9997k/benchmark-osts/OST0000/f@
/cluster/projects/nn9997k/benchmark-osts/OST0001/f@
/cluster/projects/nn9997k/benchmark-osts/OST0002/f@
/cluster/projects/nn9997k/benchmark-osts/OST0003/f@
/cluster/projects/nn9997k/benchmark-osts/OST0004/f@
/cluster/projects/nn9997k/benchmark-osts/OST0005/f@
/cluster/projects/nn9997k/benchmark-osts/OST0006/f@
/cluster/projects/nn9997k/benchmark-osts/OST0007/f@
/cluster/projects/nn9997k/benchmark-osts/OST0008/f@
/cluster/projects/nn9997k/benchmark-osts/OST0009/f@
/cluster/projects/nn9997k/benchmark-osts/OST000a/f@
/cluster/projects/nn9997k/benchmark-osts/OST000b/f@
/cluster/projects/nn9997k/benchmark-osts/OST000c/f@
/cluster/projects/nn9997k/benchmark-osts/OST000d/f@
/cluster/projects/nn9997k/benchmark-osts/OST000e/f@
/cluster/projects/nn9997k/benchmark-osts/OST000f/f@
/cluster/projects/nn9997k/benchmark-osts/OST0010/f@
/cluster/projects/nn9997k/benchmark-osts/OST0011/f@
/cluster/projects/nn9997k/benchmark-osts/OST0012/f@
/cluster/projects/nn9997k/benchmark-osts/OST0013/f@
/cluster/projects/nn9997k/benchmark-osts/OST0014/f@
/cluster/projects/nn9997k/benchmark-osts/OST0015/f@
/cluster/projects/nn9997k/benchmark-osts/OST0016/f@
/cluster/projects/nn9997k/benchmark-osts/OST0017/f
```

Max Write: 92704.91 MB/sec
Max Read: 101013.93 MB/sec

Aggregate filesize : 9.16 TiB
Time: 00:03:29

Performance issues with PowerVault

4 x RAID6 (10 drives), 8 x dedicated spares


HW: Dell PowerVault ME4024
Idun-lustre-mds
IP2: 10.1.101.6
IP1: 10.1.101.5
SSD
RAID6:
dg01 (10 x disks)
2 x global hot spare
Volumes:
dg01_mgt
dg01_mdt00
dg01_mdt01
idun-mds1
/lustre/mgt
/lustre/mdt00
/lustre/mdt01
idun-mds2 (HA)
HW: Dell PowerVault ME4024
idun-lustre-work
IP2: 10.1.101.4
IP1: 10.1.101.3
SSD
RAID6:
dg00 (10 x disks)
dg01 (10 x disks)
dg02 (10 x disks)
dg03 (10 x disks)
8 x global hot spare
Volumes:
dg00_ost00
dg01_ost01
dg02_ost02
dg03_ost03
idun-oss1
/lustre/ost00
/lustre/ost01
idun-oss2
/lustre/ost02
/lustre/ost03
HW: Dell PowerVault ME4048
idun-lustre-projects
IP2: 10.1.101.2
IP1: 10.1.101.1
HDD
RAID6:
dg04 (10 x disks)
dg05 (10 x disks)
dg06 (10 x disks)
dg07 (10 x disks)
dg08 (10 x disks)
dg09 (10 x disks)
dg10 (10 x disks)
dg11 (10 x disks)
4 x global hot spare
Volumes:
dg04_ost04
dg05_ost05
dg06_ost06
dg07_ost07
dg08_ost08
dg09_ost09
dg10_ost10
dg11_ost11
idun-oss3
/lustre/ost04
/lustre/ost05
/lustre/ost06
/lustre/ost07
idun-oss4
/lustre/ost08
/lustre/ost09
/lustre/ost10
/lustre/ost11

RAID Controller PERC H755 Front (Embedded), PowerEdge R660, RAM 512GB

KIOXIA, KPM6WRUG3T84 Write 2450 MB/s, Read 4150 MB/s, https://apac.kioxia.com/en-apac/business/ssd/enterprise-ssd/pm6-r.html

	Non-RAID (MB/s)	RAID6 (4 drives)	RAID6 (6 drives)	RAID6 (8 drives)
Write ior	1127	2002	4235	4478
Read ior	1130	4500	6766	9034

SAMSUNG, MZILT3T8HBLS0D3, Write 2000 MB/s, Read 2100 MB/s,
https://semiconductor.samsung.com/ssd/enterprise-ssd/pm1643-pm1643a/mzilt3t8hbls-00007/

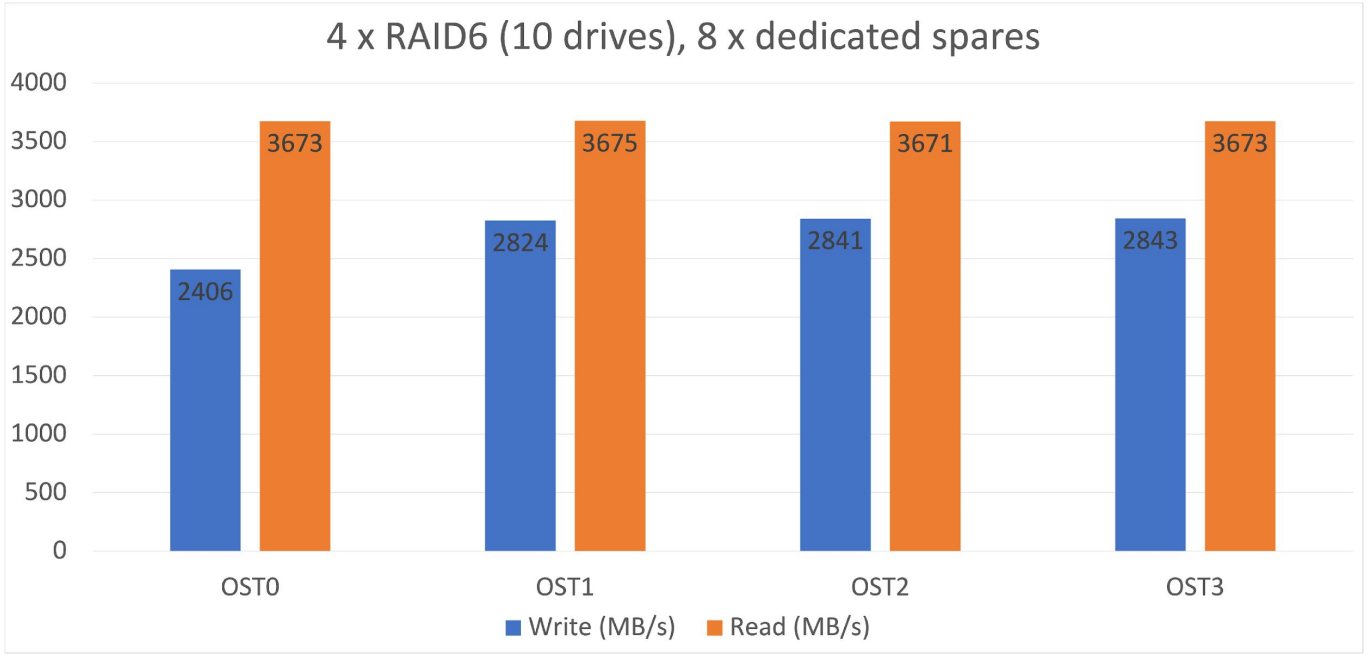
	Non-RAID (MB/s)	RAID6 (4 drives)	RAID6 (6 drives)	RAID6 (8 drives)
Write ior	1114	2236	4248	4480
Read ior	1123	4483	6737	8986

PowerVault

	RAID1 (first 2 drives)	RAID6 (4 drives)	RAID6 (6 drives)	RAID6 (8 drives)	RAID6 (10 drives)	RAID6 (12 drives)	RAID1 (last 2 drives)
Write ior	1330	786	1494	1956	2457	2480	443
Read ior	3310	3667	3665	3665	3667	3667	3318

```
Benchmark commands:
mpirun -n 64 ior -a POSIX -v -F -C -e -g -k -b 1m -t 1m -s 20000 -i 1 -w -r -o /mnt/disk1/
mpirun -n 64 ior -a POSIX -v -F -C -e -g -k -b 1m -t 1m -s 20000 -i 1 -w -r -o /mnt/disk1/
```

4 x RAID6 (10 drives), 8 + dedicated spares

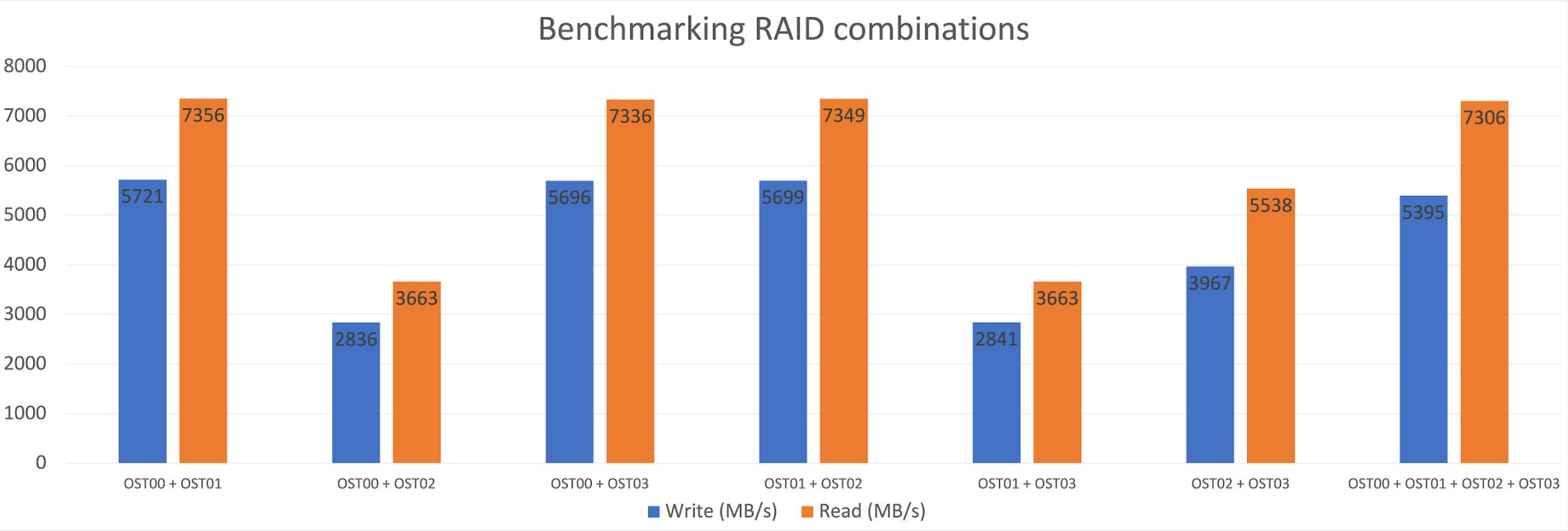


	OST0	OST1	OST2	OST3
Write ior	2378	2384	2378	2354
Read ior	3670	3670	3670	3668

Benchmark commands:

```
mpirun --allow-run-as-root -np 32 /root/ior-3.3.0/src/ior -a POSIX -v -F -C -e -g -k -b 1m -t 1m -s 20000 -i 1 -w -r -o /mnt/ost00/
```

Benchmarking OSTs combinations



	OST00 + OST01	OST00 + OST02	OST00 + OST03	OST01 + OST02	OST01 + OST03	OST02 + OST03	OST00 + OST01 + OST02 + OST03
Write ior	3100	2804	3740	3030	2801	3539	5223
Read ior	7337	3669	7323	7338	3668	5149	7281

Benchmark command OST00 + OST01

```
mpirun --allow-run-as-root -np 32 /root/ior-3.3.0/src/ior -a POSIX -v -F -C -e -g -k -b 1m -t 1m -s 20000 -i 1 -w -r -o
/mnt/ost00/file01@/mnt/ost00/file02@/mnt/ost00/file03@/mnt/ost00/file04@/mnt/ost00/file05@/mnt/ost00/file06@/mnt/ost00/file07@/mnt/ost00/file08@/mnt/ost00/file09@/mnt/ost00/file10@/mnt/ost00/file11@/mnt/ost00/file12@/mnt/ost00/file13@/mnt/ost00/file14@/mnt/ost00/file15@/mnt/ost00/file16@/mnt/ost01/file01@/mnt/ost01/file02@/mnt/ost01/file03@/mnt/ost01/file04@/
mnt/ost01/file05@/mnt/ost01/file06@/mnt/ost01/file07@/mnt/ost01/file08@/mnt/ost01/file09@/mnt/ost01/file10@/mnt/ost01/file11@/mnt/ost01/file12@/mnt/ost01/file13@/mnt/ost01/file14@/mnt/ost01/file15@/mnt/ost01/file16
```

Benchmark command OST00 + OST01 + OST02 + OST03

```
mpirun --allow-run-as-root -np 32 /root/ior-3.3.0/src/ior -a POSIX -v -F -C -e -g -k -b 1m -t 1m -s 60000 -i 1 -w -r -o
/mnt/ost00/file01@/mnt/ost00/file02@/mnt/ost00/file03@/mnt/ost00/file04@/mnt/ost00/file05@/mnt/ost00/file06@/mnt/ost00/file07@/mnt/ost00/file08@/mnt/ost01/file01@/mnt/ost01/file02@/mnt/ost01/file03@/mnt/ost01/file04@/mnt/ost01/file05@/mnt/ost01/file06@/mnt/ost01/file07@/mnt/ost01/file08@/mnt/ost02/file01@/mnt/ost02/file02@/mnt/ost02/file03@/mnt/ost02/file04@/
mnt/ost02/file05@/mnt/ost02/file06@/mnt/ost02/file07@/mnt/ost02/file08@/mnt/ost03/file01@/mnt/ost03/file02@/mnt/ost03/file03@/mnt/ost03/file04@/mnt/ost03/file05@/mnt/ost03/file06@/mnt/ost03/file07@/mnt/ost03/file08
```

Default “Adaptive” create issues for multithreaded sequential read