Optimizing Lustre Throughput in a Software RAID Environment: Configuration tips and Performance Insights

ABOUT XINNOR

Most Innovative Flash Memory Customer Implementation

- Founded in Haifa, Israel, May 2022
- Background: 10+ years of experience with software RAID design and mathematical research
- Mission: to be the fastest RAID Engine
- Team: Around 40 people; >30 are accomplished mathematicians and industry talents from Global Storage OEMs
- >20 selling partners worldwide
- >100PB of end-customers data

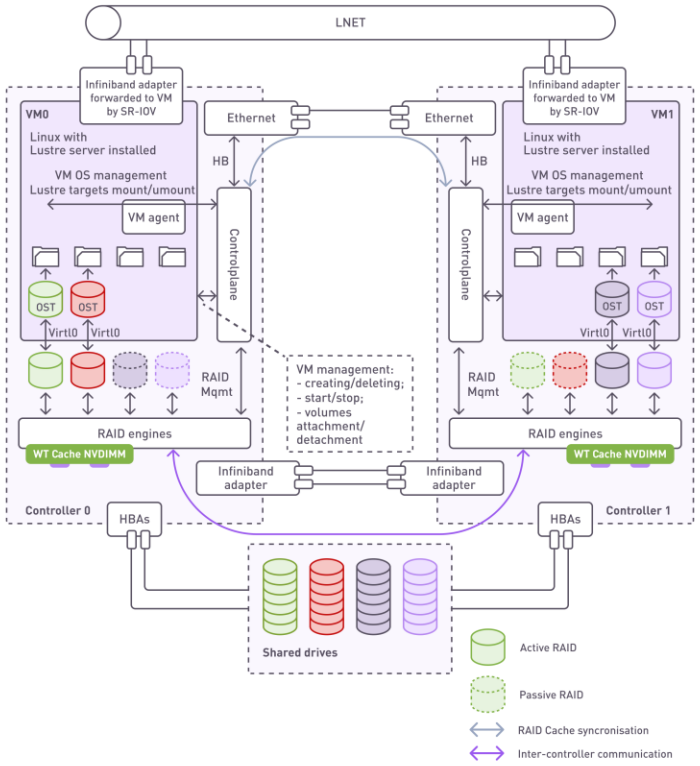
Technology partners

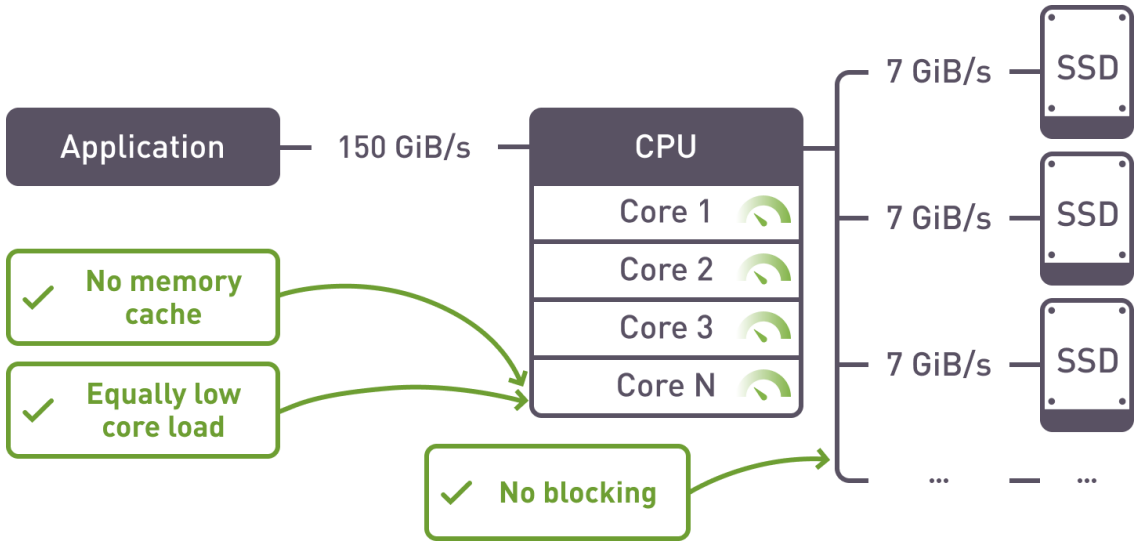
ATTO, Western Digital, ScaleFlux, KIOXIA, BeeGFS, ingrasys, TUXERA, LINBIT, NVIDIA, DapuStor

WHAT DO WE DO?

xiRAID

the fastest flash-native SW RAID engine





xiSTORE

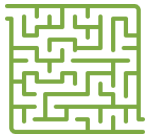
Integration of RAID engine with Parallel FileSystem Storage optimized for HPC and AI workloads

TOPICS

IOR throughput intensive workload from storage engineer perspective

Setting performance expectations for Software RAID

Simple tests that may help before running IOR



Linux impact on workload with large block size

Results with IOR

FIO WORKLOAD IS "EASY" FOR STORAGE SUBSYSTEM

Typical fio config:

```
[global]
ioengine=libaio
direct=1
numjobs=1
bs=1M
iodepth=32
rw=write
```

- IO size is 1 MB
  - likely no IO splits or merges in Linux IO stack
- Async IO with constant 32 IOs in-flight
  - We can control disk subsystem utilization by changing iodepth
- Application developers may mimic this pattern if it works well.
- SNIA PTS uses such pattern

IOR WORKLOAD IS MORE CHALLENGING FOR STORAGE SUBSYSTEM

DIO

- Large transfer size 64-256 MB
  - Large IO is split by stripe or RPC size
- Pros
  - Allows to control number of IOs in-flight
- Cons
  - workload is "bursty", fork-join model
  - Disk utilization is < 100%

BIO

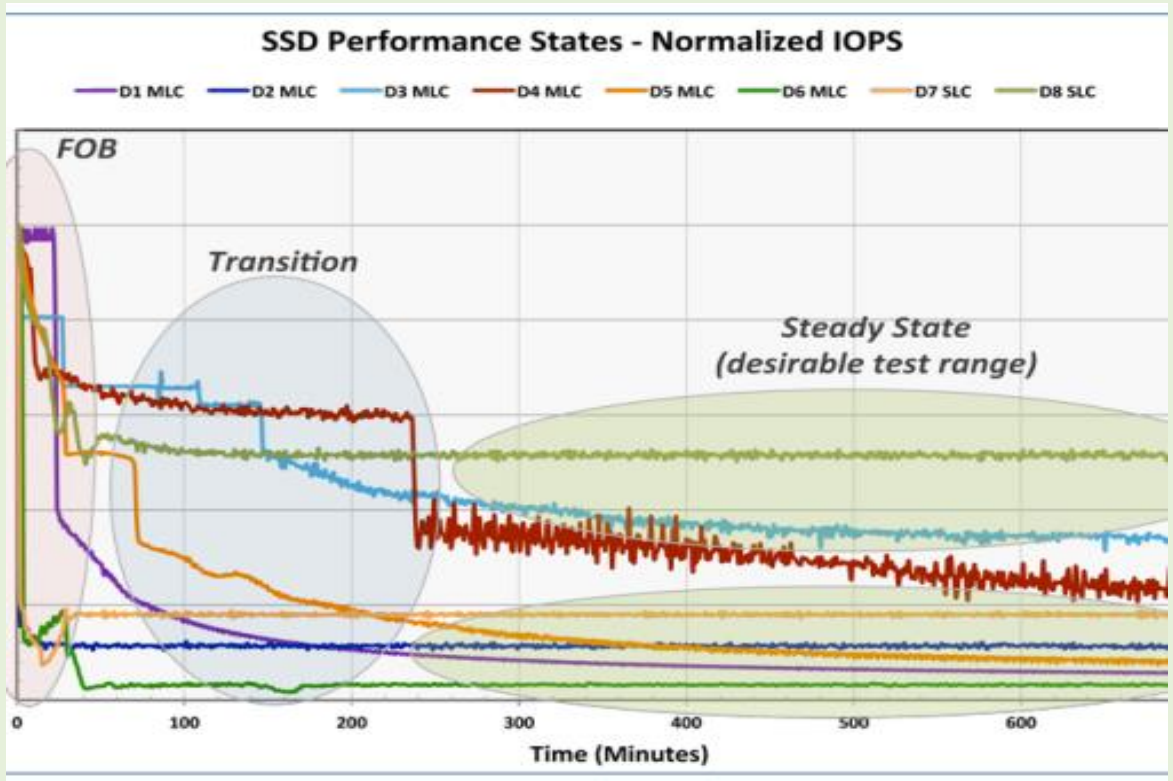
- Use client write-back cache
- Pros
  - Async IO
- Cons
  - No control for number of IOs-flight from client side
  - May overload disk subsystem

SETTTING PERFORMANCE EXPECTATIONS FOR MEDIA

Example from the SSD datasheet:

- 128 KiB Sequential Write: 4,200 MB/s
- 4 KiB Random Write: 170 K IOPS = 680 MB/s

- Lustre FPP workloads are not 100% sequential, but also not 4 KiB random.
- Is fstrim the only option to get reproductible results?
- For HDD datasheets typically provides fio numbers with no seeks and measured at fast cylinders.



From SNIA PTS 2.0.2 which is used for most SSD tests

SIMPLE ESTIMATION OF SEQUENTIAL WRITE RAID PERFORMANCE

- For example: datasheet sequential write performance for HDD 260 MB/s
- RAID 60, 42 disks, 8D + 2P + 2 spare
- 4 groups [8d+2p] per RAID i.e 32 data disks
- Theoretical write performance 32 data disks *260 = 8320 MB/s

Typical performance is lower because of:

- Media latency deviation
  - 1 MB write IO to RAID turns into 10 IO 128 KB to media
- Software RAID overhead

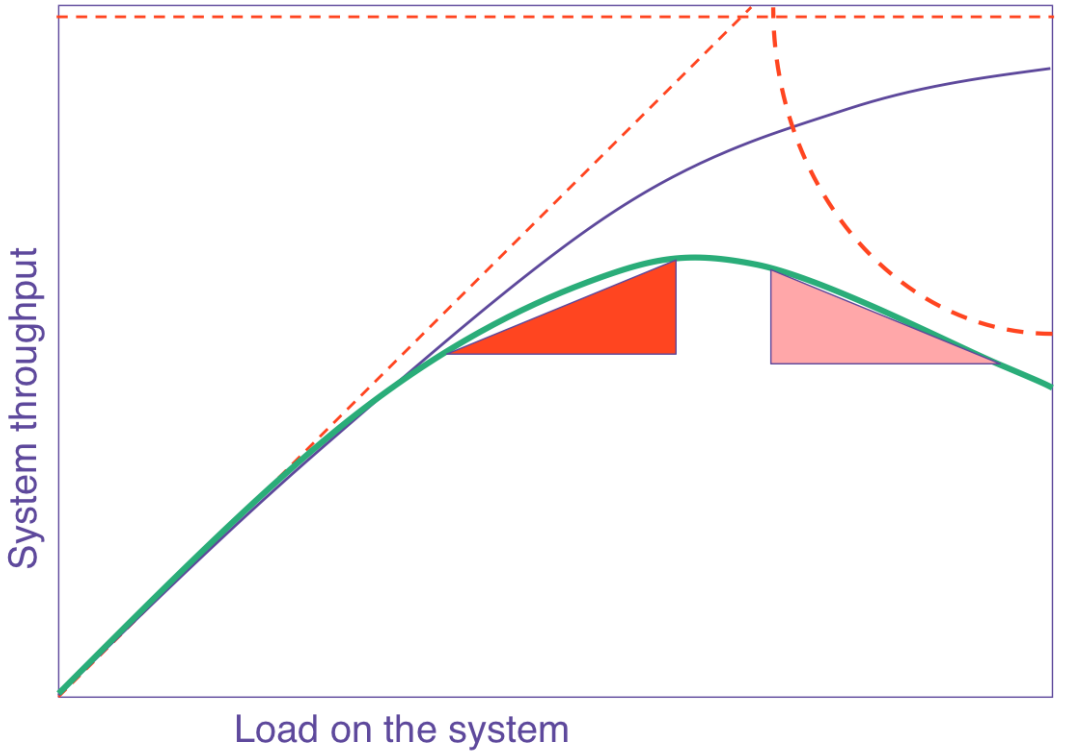
STRESS ALL DISK AT ONCE TO FIND OUT BOTTLENECKS IN STORAGE SUBSYSTEM

- Performance can be impacted by:
  - Server side limits: PCIe bus, NUMA, HBAs
  - SAN bottlenecks: ports, expanders, bad disks
- It is important to check that all disks show the same speed
  - RAID is as fast as slowest disk
- Good results helps to get proper motivation

```
[global]
ioengine=libaio
direct=1
numjobs=1
bs=128K
iodepth=8
rw=write
# one fio job per disk
[sda]
filename=/dev/sda
[sdb]
filename=/dev/sdb
```

RUN FIO TEST WITH DIFFERENT IODEPTH ON SINGLE RAID INSTANCE

- Queue depth can vary greatly with Lustre workloads
- Scalability test may show any contentions with high IO concurrency
- Simple test run fio with iodepth from 1 to 32



Source: http://www.perfdynamics.com/Manifesto/USLscalability.html

STRESS ALL RAIDs AT ONCE WITH SEQUENTIAL WORKLOADS

- Make sure that no mutual influence of RAID instances
- In case of virtual machines run test from host and VMs
- Software RAID is an application and needs monitoring and tuning (CPU, utilization, etc)
- Compare results with previous tests and theoretical expectations

```
[global]
ioengine=libaio
direct=1
numjobs=1
bs=1M
iodepth=32
rw=write
# one fio job per software raid
[raid1]
filename=<raid1 block device>

[raid2]
filename=<raid2 block device>
```

CHALLENGES WITH LARGE BLOCK I/O SIZE

- Lustre tunning 'brw_size=16' and 'max_pages_per_rpc=4096' showed best performance for us
- Linux kernel may split 16 MB IOs into smaller IOs,
  - For 4.18 RHEL kernel most of IOs aligned by 2 MB boundary
  - But IOs may be split by 4KB boundary not aligned by RAID stripe size -> **significant performance impact**.
- Use blktrace or 'perf' (perf record –e block:*) to trace physical IOs sent to block device

MISC CONSIDERATIONS

- Disabling Hyperthreading during testing simplifies configuration and tuning
- NUMA tuning – pin RAIDs and VMs to localities
- Virtual machine IO settings:
  - io='native' seems better for sequential HDD workloads. 'Threads' may cause more random pattern
  - 'scsi-blk' driver may be better for large IO size. 'virtio-blk' itself may cause IO splits and merges which may turn into IO not aligned by RAID stripe size

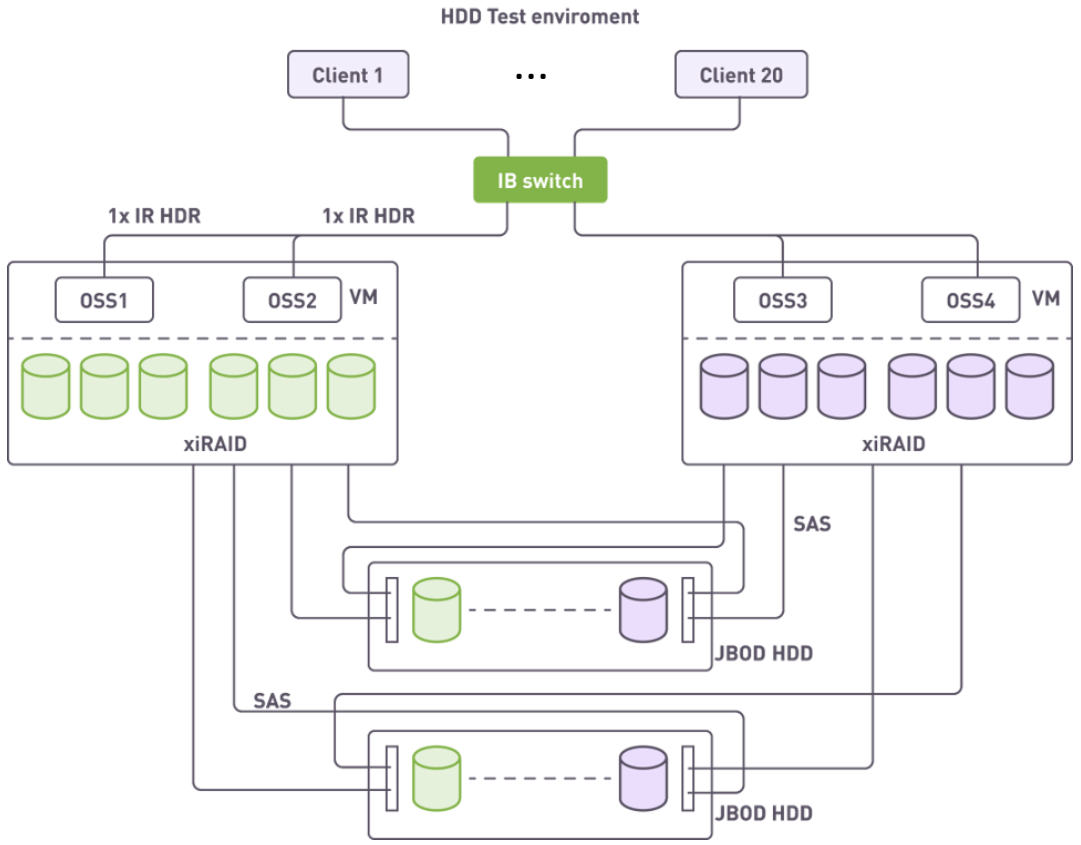
IOP FPP TEST WITH HDD xiRAID DCR

Servers

- 2 servers (1 socket, 64 core AMD)
- 4 OSS VM / 12 OST
- OST config xiRAID 60: dcr 42 disks (8d+2p), ldiskfs
- 1x IB-HDR per VM
- 7 JBODs connected via LSI 9500 HBAs
- VM OS: Rocky 8.7, Lustre 2.15.2

Clients

- 20 Lustre clients
- Oracle linux 8.8
- Lustre 2.15.3
- 1x IB-HDR per VM



IOR FPP TEST WITH HDD xiRAID DCR RAID 60 RESULTS

- fio test raw HDD **126 GB/s**
  - 128KiB block size, 1 thread/hdd, iodepth=8, 504 disks
  - Performance per HDD is 250 MB/s
- fio test RAID **80 GB/s**
  - 12 RAIDs from 4 VMs (384 data disks)
  - 1 MiB block size, 1 thread/RAID, iodepth=32
  - Performance per HDD is 206 MB/s
- IOR over **50 GB/s** for writes and reads
  - ior -F -w / -r -b 8g -t 1m -e
  - Flush client caches between writes and reads

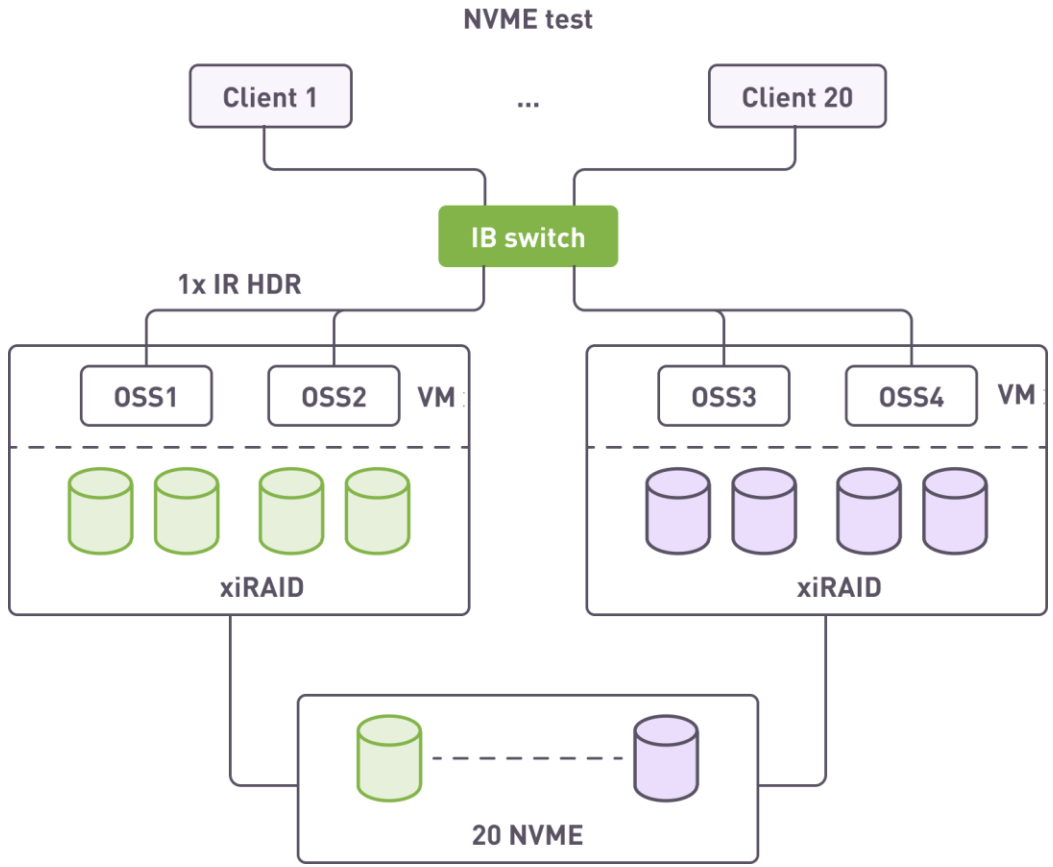
IOP FPP TEST NO SSD xiRAID CONFIGURATION

Servers

- 2 servers (1 socket, 64 core AMD)
- 4 OSS VM / 4 OST
- OST config xiRAID 6: 10 disks (8d+2p), ldiskfs
- Kioxia KCM61RUL3T84
- 2 namespaces per NVMe, to avoid PCIe x2 limits
- 7 JBODs connected via LSI 9500 HBAs
- Rocky 8.7, Lustre 2.15.2

Clients

- 20 Lustre clients
- Oracle linux 8.8
  - Lustre 2.15.3



IOR FPP TEST WITH SSD xiRAIDS RESULTS

- Theoretical write performance:
  - 16 data disks * 4200 MB/s (KIOXIA datasheet) = 67.2 GB/s
- fio RAID write test 59.5 GB/s (full capacity)
- IOR write 63 GB/s, read 89 GB/s
  - ior -F -w / -r -b 8g -t 1m -e
- Flush client caches between writes and reads

CONCLUSIONS

Lustre IOR workloads can be very different and more challenging for storage than fio microbenchmarks.

fio raw device microbenchmarks are still valuable to set expectations and tune before IOR test

There are 3 types of lies: lies, damn lies and ~~statistics~~ benchmarks

XINNOR

THANK YOU!

Give us a try:
https://xinnor.io/