

# Benchmarking Lustre

## Setting Realistic Performance Expectations



**Torben Kling Petersen PhD**

Distinguished Technologist  
Lead HPC Storage Architect - EMEA & APAC

**John Fragalla**

Distinguished Technologist  
Lead HPC Storage Architect - Americas



# Problem Statement

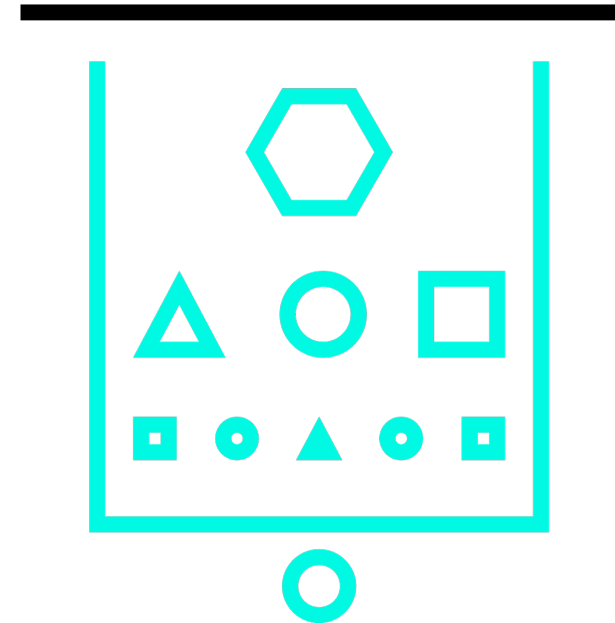
---

## Hardware complexity

- AMD Rome architecture means complex NUMA settings
  - CPU mapping to NVMe drives
  - CPU to Network adapters
  - CPU to SAS HBAs
- Network options
  - HDR InfiniBand
  - 200G Ethernet (both TCP and RoCE)
  - 200G SlingShot
  - (OmniPath)
- Client hardware

## Software

- Lustre release
  - Point release, backports, tunables
  - Ldiskfs vs OpenZFS
- OS release(s)



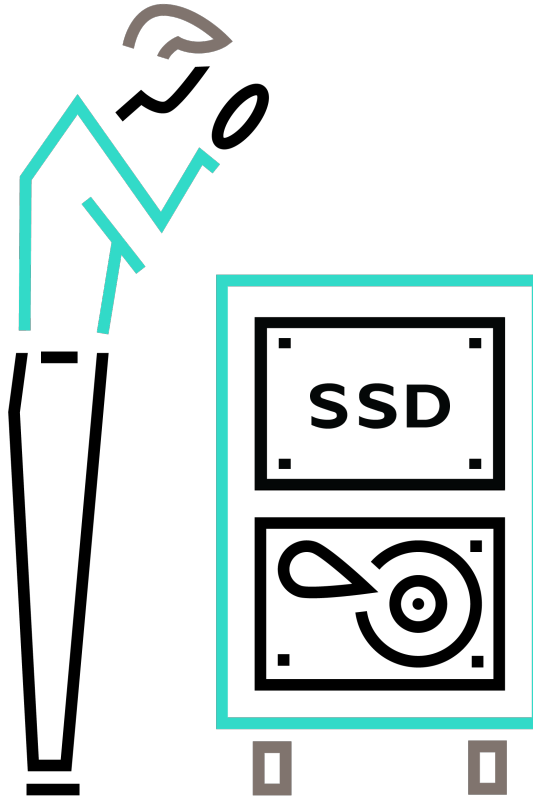
# Benchmarking methodology ??

## Benchmark purpose:

- Marketing
- Specific customer requirements
- Product consistency
- Other ??

## IO500

- Full
- 10 node challenge
- Issues
  - Not limited to production systems
  - Weighting between runs skews results
  - Limited tunings/modification
  - Consistency over time



## “Old School”

- IOR
- MDtest
- FIO
- IOzone
- Issues
  - Consistency over time

# Test Environments

## • Ethernet and HDR Testing

- 21 Clients
  - 1 MLX HCA in HDR Mode or Ethernet Mode
  - CentOS 8.4 (kernel 4.18.0-305.25.1.el8\_4.x86\_64) Lustre 2.12 and 2.15 Clients
- E1000 HDR-200 System
  - 1 MDU 2 GridRAID Flash Unit 1 Flash-10 Unit 1 D2 ( LDISKFS)
- E1000 HW RoCE / TCP 200GigE System
  - 1 MDU 1 GridRAID Flash Unit 1 D2 (LDISKFS or dRAID OpenZFS)

## • Slingshot-11

- 21 Clients
  - CentOS 8.4 (kernel 4.18.0-305.25.1.el8\_4.x86\_64) Lustre 2.12 and 2.15 Clients
  - 1 CXI HCA Adapter
  - 1 MLX HCA in Ethernet Mode
- E1000 Cassini/KFI System
  - 1 MDU 1 GridRAID Flash Unit 1 D2 LDISKFS (klibfabirc)

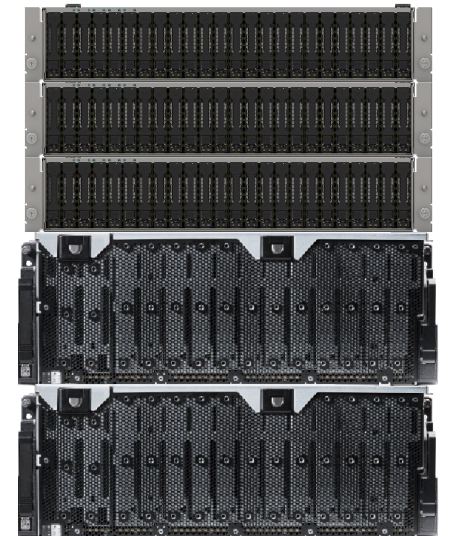
### Lustre LNET drivers used:

- ko2iblnd
  - RDMA driver used for InfiniBand HDR and 200GigE RoCE
- ksocklnd
  - TCP/IP driver used for 200GigE
- kfilnd (klibfabric)
  - RDMA driver used on Cassini (CXI) adapters with Slingshot-11

# Scalable Storage Units Benchmark - Defined

	Metadata Unit (MDU)	Extreme Performance (SSU-F) Flash	IOPS Performance (SSU-F) Flash	HDD Performance (SSU-D2) HDD
LDISKFS RAID Layout	2 RAID-10 (11 drive) 2 Hot Spares	2xGridRaid 12[(8+2)+1]	2 RAID-10 (11 drive) 2 Hot Spares	4xGridRaid 53[(8+2)+2]
ZFS dRAID Layout	2x draid1:1d:12c:1s	2x draid2:9d:12c:1s	2x draid1:1d:12c:1s	4x draid2:53d:16c:2s
Network ports	4 x 200 Gbps	4 x 200 Gbps	4 x 200 Gbps	2 x 200 Gbps
Height Rack Units	2	2	2	10
Number of Lustre Servers	2 MDS Nodes	2 OSS Nodes	2 OSS Nodes	2 OSS Nodes
Number of Lustre Targets	2 MDTs	2 OSTs	2 OSTs	4 OSTs

- ClusterStor E1000 was launched in 2019 with LTS Lustre version **2.12**.
- Multiple software stack updates over the years (extract below used for this presentation):
  - Neo 4.1 CentOS **7.6** (kernel 3.10.0-957.1.3957) and Lustre 2.12.0.5
  - Neo 4.4 CentOS **7.6** (kernel 3.10.0-957.1.3957) and Lustre 2.12.4.3
  - Neo 6.x Rocky Linux **8.4** OS (kernel 4.18.0-305.10) and Lustre 2.15.0.3



# Benchmark Details - Standard

- IOR Throughput

- Direct-IO (DIO) and Buffered-IO (BIO)
- File-Per-Process (FPP) and Single Shared File (SSF)
- 64M transfer size for DIO 1M transfer size for BIO
- 16 ranks per node
- Fixed time results used (provided peak performance results across different protocols)
- Fixed data results collected for consistency
- Flush caches on the clients between writes and reads

- IOR IOPS

- Buffered-IO (BIO)
- 4K transfer size with random operation
- FPP using 8GB Files
- 64 ranks per node
- Fixed time results
- Flush caches on the clients between writes and reads

- MDTEST

- Unique directory operation
- 1 Million objects per MDT
- 16 ranks per node
- Directory and File operations
- Mean of 3 iterations
- OK and 32K File sizes
- Non-DOM Results

- obdfilter-survey

- Versions

- IOR: 3.3.0
- MDTEST: 1.9.3

# Standard Benchmark Sweep



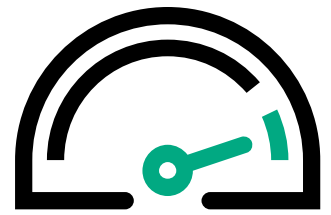
For each new release of patch the following benchmarks are collected:

- Throughput (both SSF and FPP)
  - IOR (R/W) on All Flash System (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) on Single JBOD (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) on 2x JBODs (4 OSTs) with GridRAID or OpenZFS
- IOR (R/W) - Single client single thread (with and without over-striping) and multi-thread
- IOR (R+W) – “Bi-directional” Write performance during simultaneous read (50% R / 50% W)
- IOPS
  - IOR (R/W) on All Flash System (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) - Single client single thread (with and without over-striping) and multi-thread
  - IOR (R/W) on Single JBOD (2 OSTs) with GridRAID or OpenZFS
  - IOR (R/W) on 2x JBODs (4 OSTs) with GridRAID or OpenZFS
- Metadata
  - MDtest – Full sweep on single and dual MDTs
    - 0K and 32K files without DoM
    - 0K and 32K files with DoM

A full sweep is run for every minor and major software release or patch delivered by engineering. Full sweep takes the better part of 2 days and is fully scripted for consistency. NB all page caches are flushed between each part (e.g. W and R).



# Test Permutations



## Client side

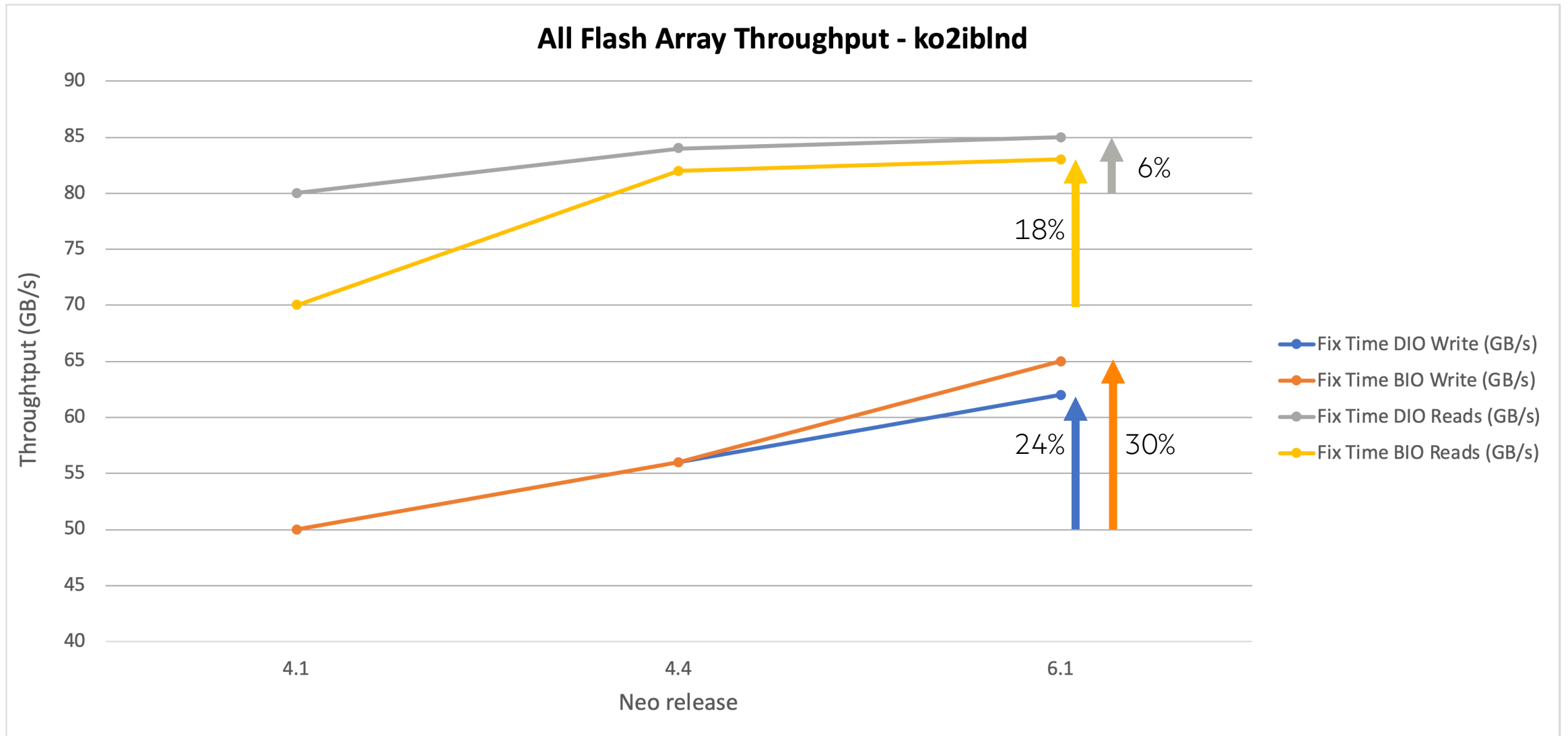
- Network Checksums
  - HPE use disabled ..
- Max RPCs in Flight
  - Default is 64
  - HPE use 256
- Max Dirty MB
  - Default is 2000
  - HPE use default
- Max Pages per RPC
  - Default is 256 (1MB),
  - HPE use 4MB for flash and 16MB for HDD based systems
- Max Read Ahead MB
  - Default is 64 MiB
  - HPE use 512MiB
- Max Read Ahead per File MB
  - Default is 64 MiB
  - HPE use 512MiB

## Server side

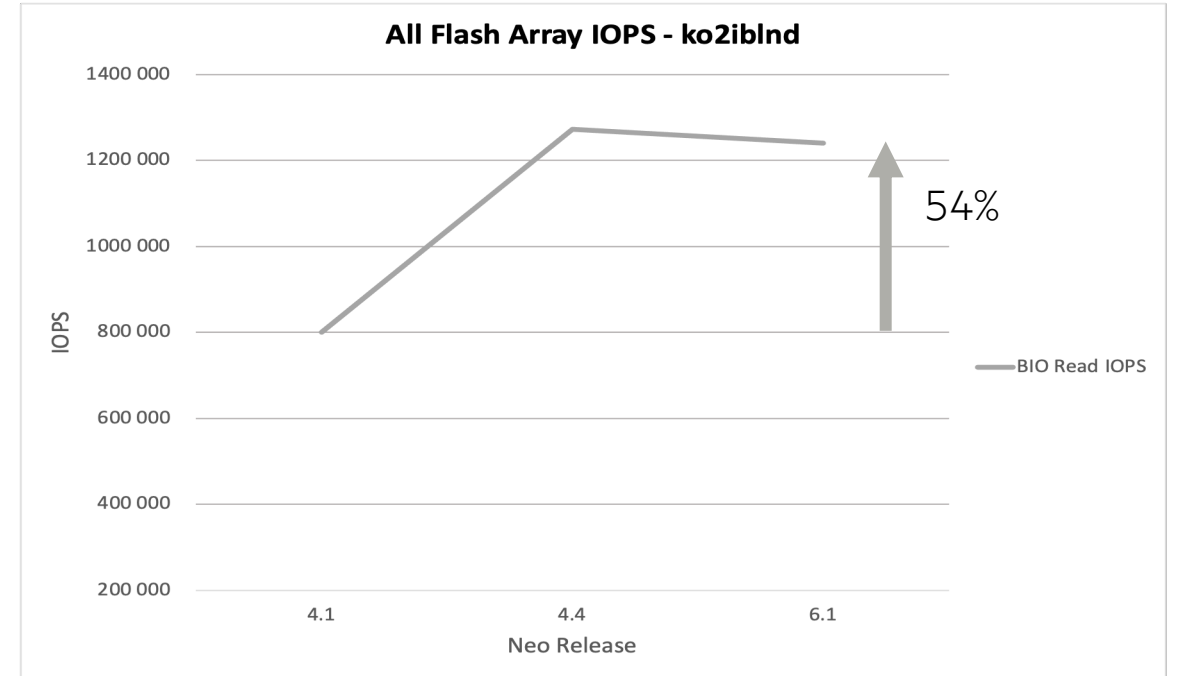
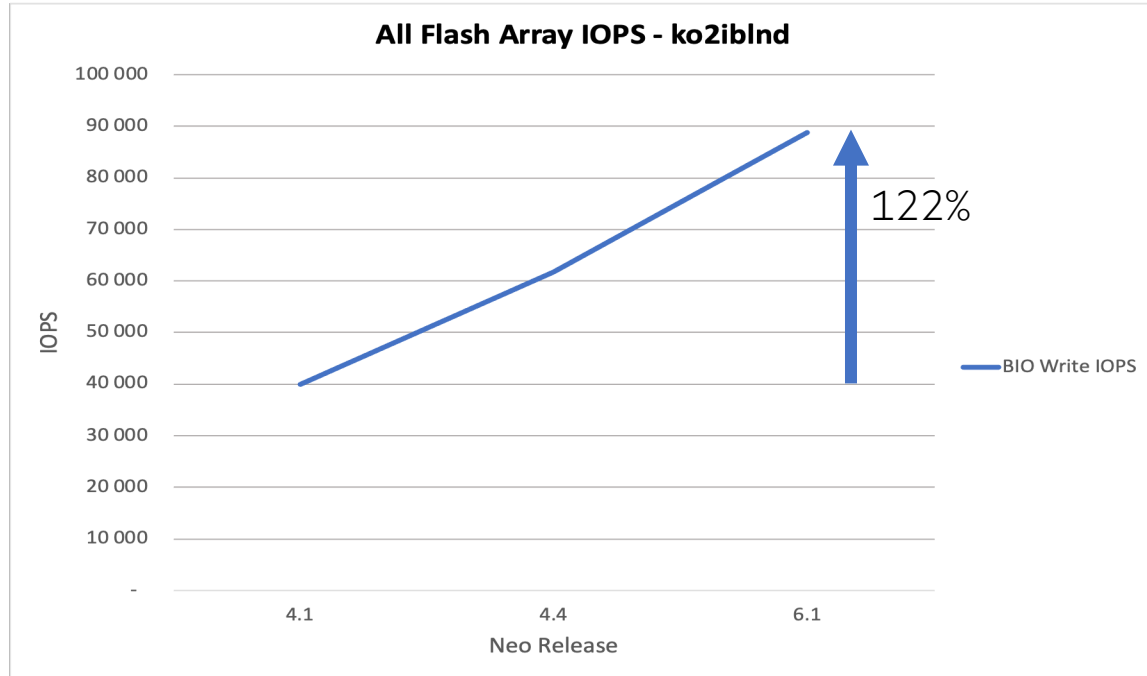
- System tunings
  - NPS (1, 2, 4)
  - HPE use
    - NPS4 for MDS
    - NPS2 for LDISKFS OSS
    - NPS1 for OpenZFS OSS
    - CPT=8 for all systems
- Failover testing



# All Flash Array with PD-RAID LDISKFS Performance

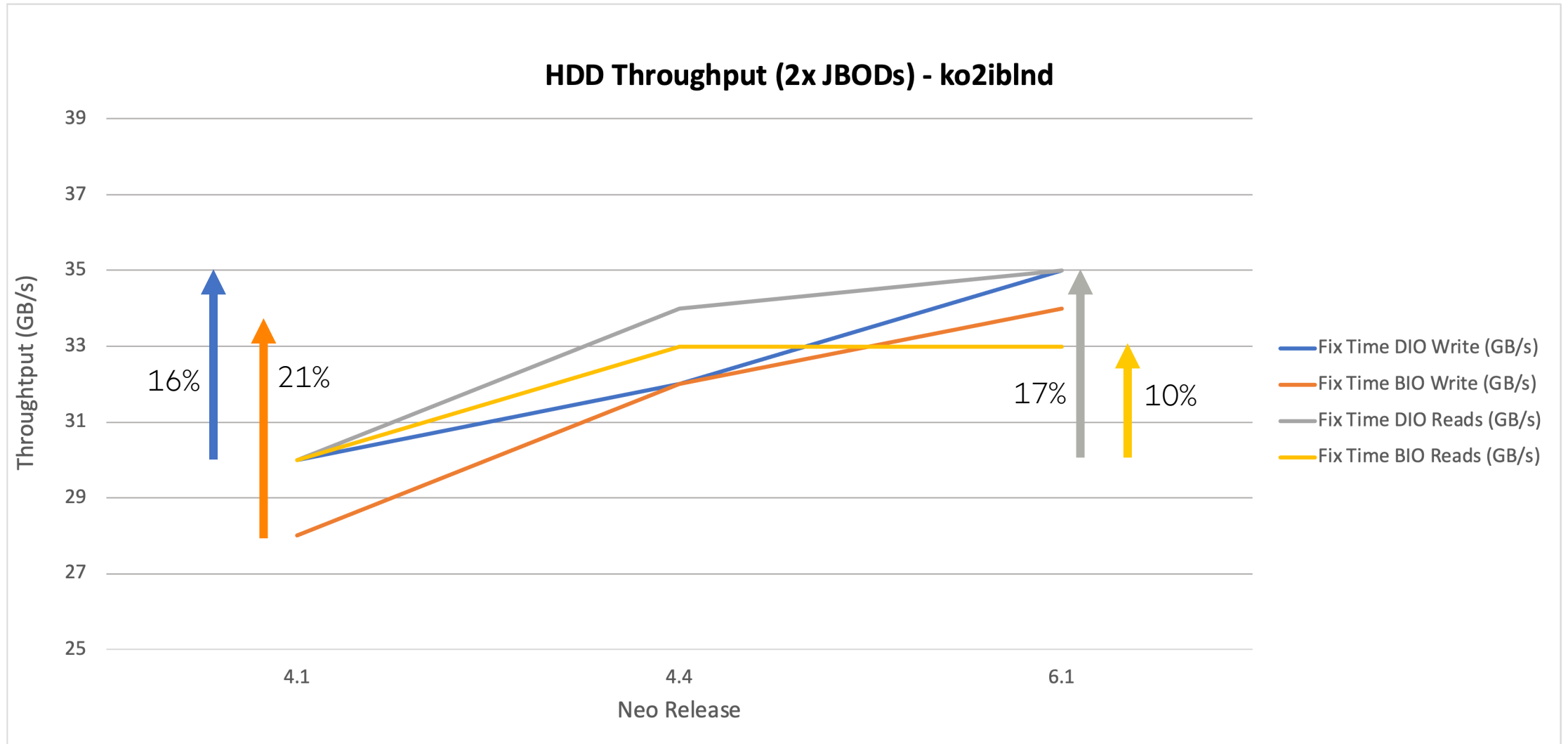


# Clusterstor E1000 SSU Flash Gridraid LDISKFS Performance



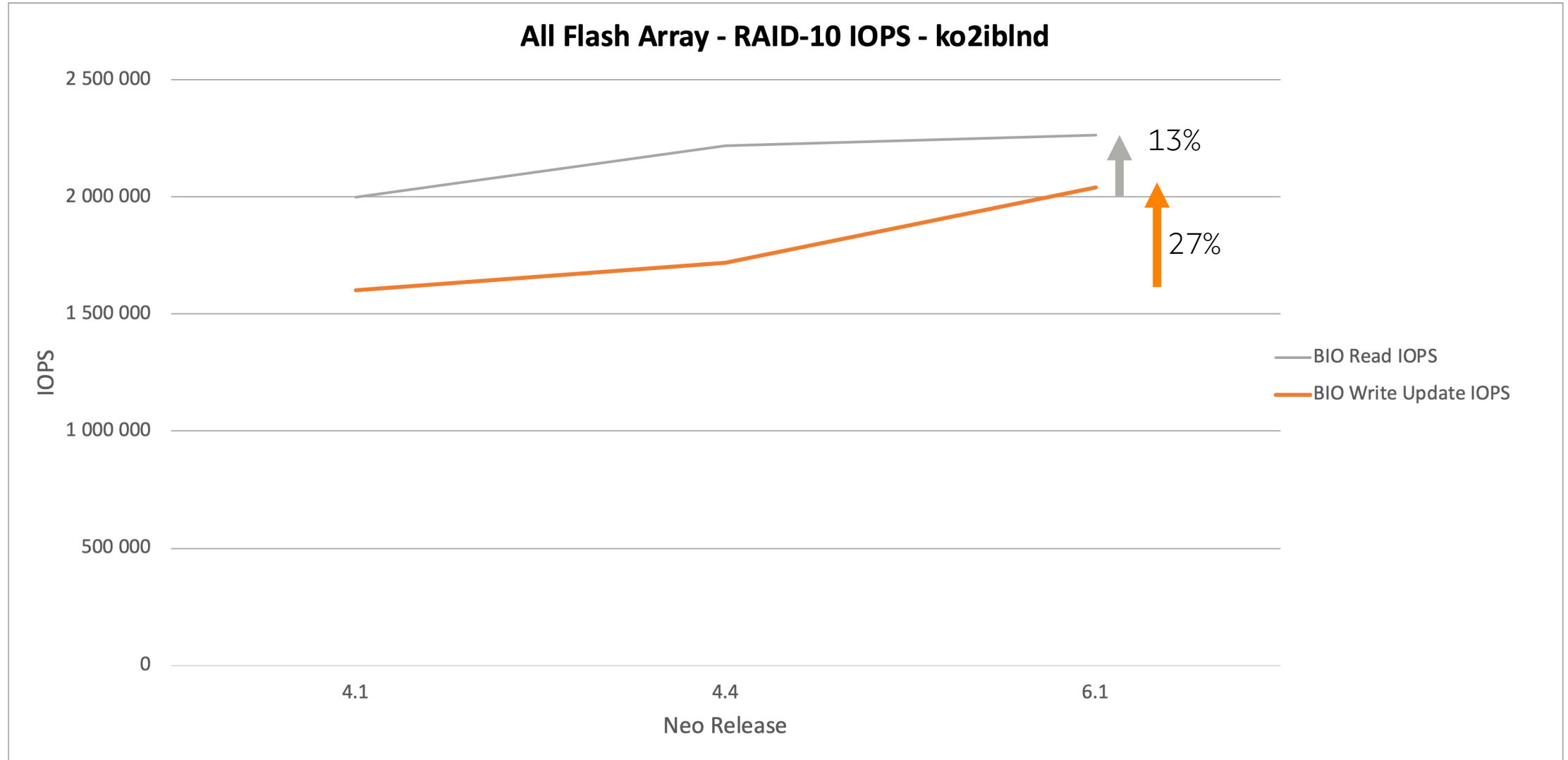
# HDD Performance – 2x JBODs PD-RAID on LDISKFS

Performance gain from new firmware BIOS Change from NPS4 to NPS2 Lustre 2.15 & new kernel.



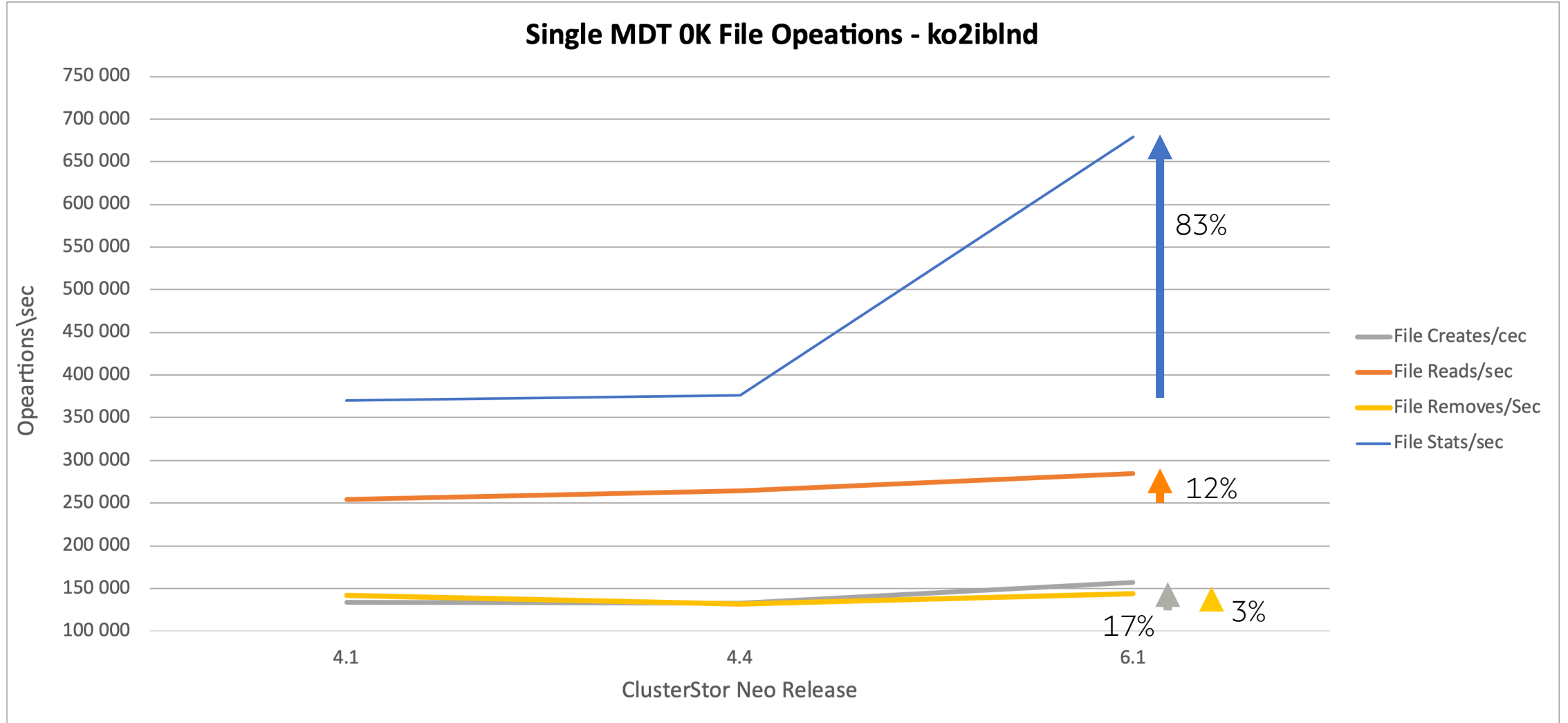
# All Flash RAID-10 LDISKFS IOPS Performance

Performance gain from Lustre 2.15 & new kernel.



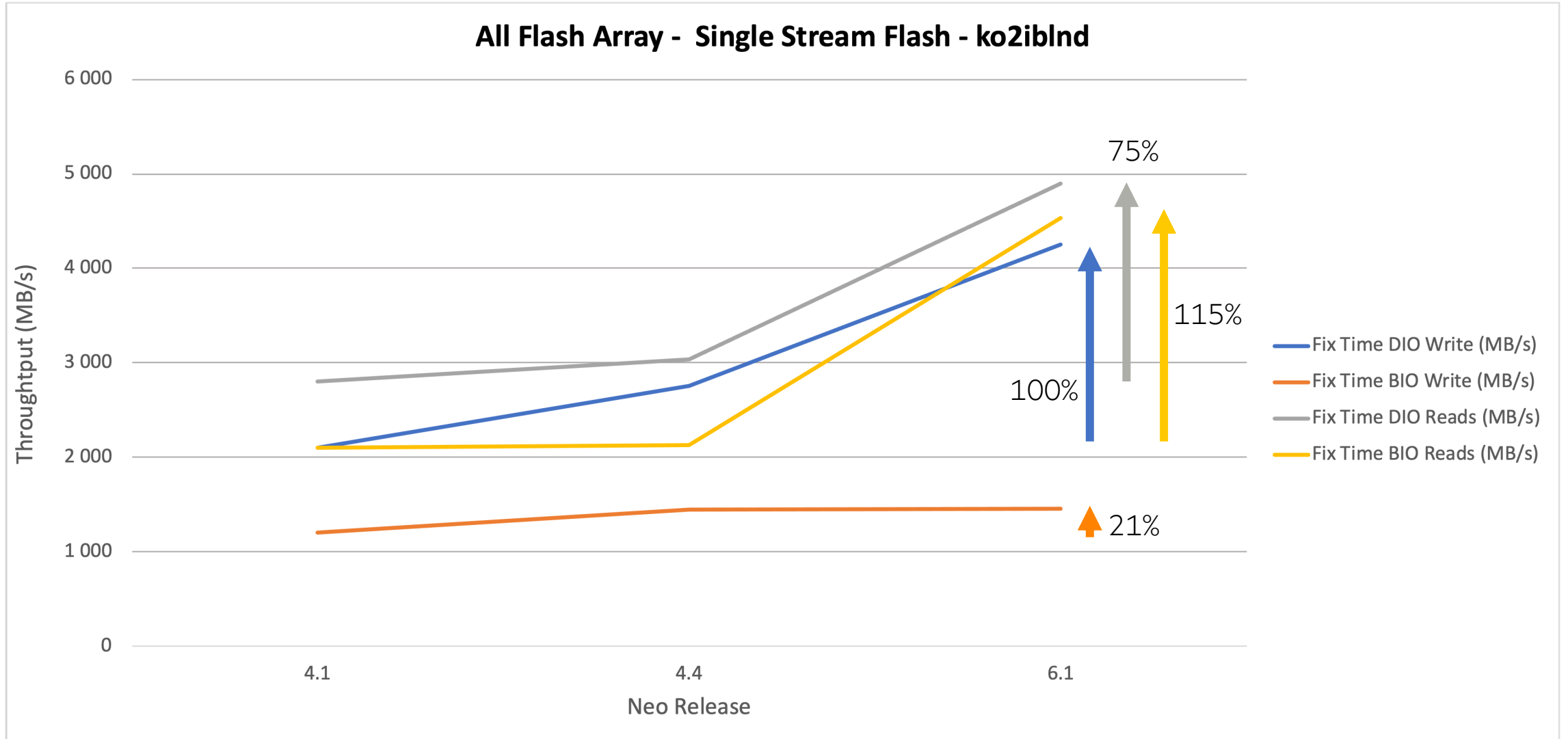
# Single All Flash MDT LDISKFS Performance

Performance gain from Lustre 2.15 & new kernel.



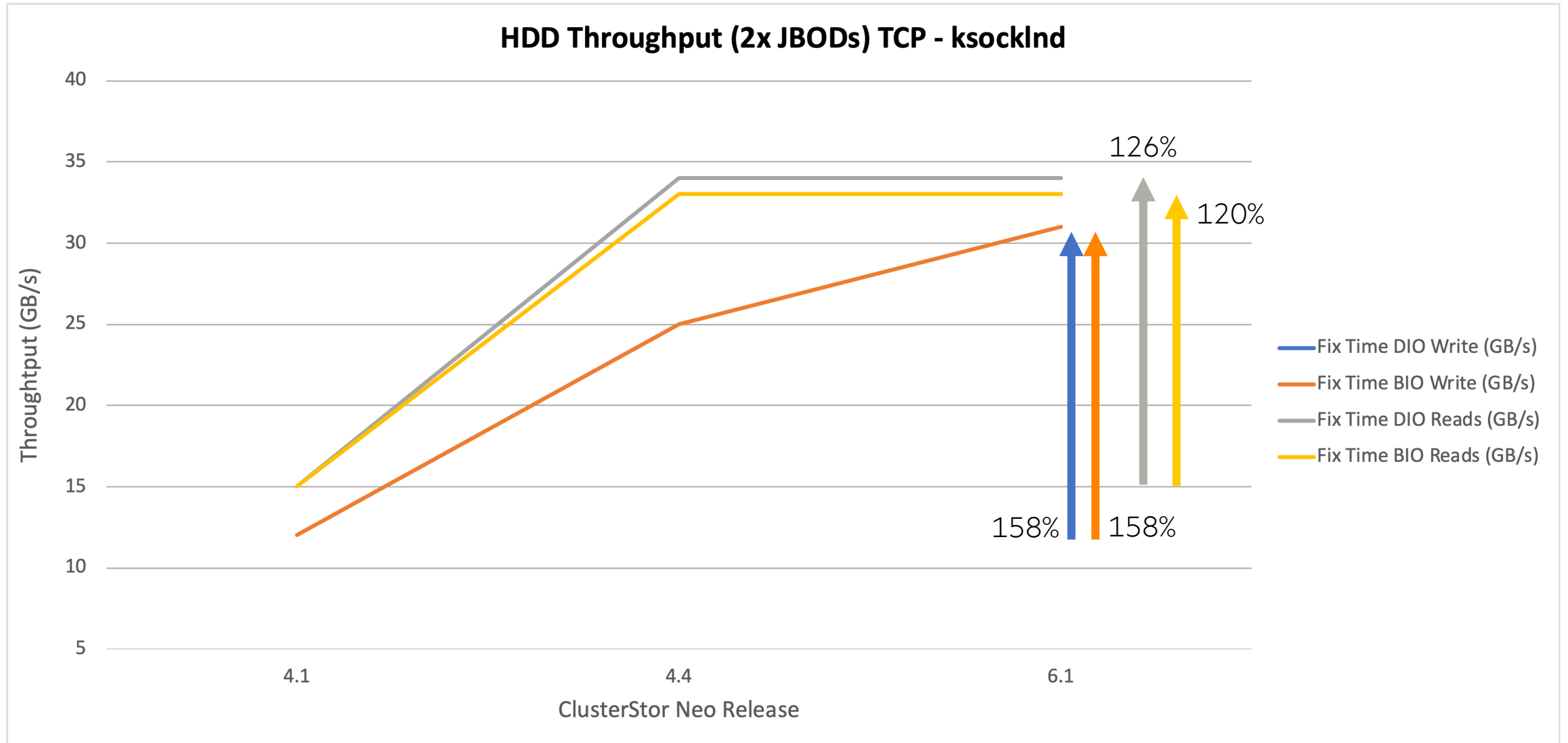
# All Flash OST - Single Stream IO Performance

Performance gain from Lustre 2.15 & new kernel.




# HDD TCP Performance - TCP on 2x JBODs with LDISKFS

Performance gain from BIOS Change from NPS4 to NPS2







# Comparing RDMA Fabrics Protocols

# PERFORMANCE UPDATE E1000 SSU-F

- Samsung PM1733
- 20+ clients, Stonewalling IOR, GridRAID-12, Lustre 2.15 etc ....

## GridRAID

	IO	IB (HDR) 6.1-010.39	Ethernet (TCP) 6.1-010	Ethernet (RoCE) 6.1-010	KFI 6.1-010.40
DIO 64PPN	Write	58.9	35.7	57.9	54.2
	Read	85.3	81.2	85.4	82.2
BIO 64PPN	Write	63.4	36.5	64.4	59.1
	Read	83.3	70.7	83.1	82.1
IOR Buffered IO 4K Random IOPS	Write	84,389	83,747	85,590	81,857
	Re-Write	53,085		53,903	48,542
	Read	1,217,062	676,030	1,217,216	735,866

# Throughput Comparison HDD based OSTs

20+ clients, Stonewalling IOR, GridRAID-53, Lustre 2.15 etc ....

<b>Single JBOD (2 OSTs)</b>	<b>IO</b>	<b>IB (HDR) 6.1-010.39</b>	<b>Ethernet (RoCE) 6.1-010.48</b>	<b>KFI 6.1-010.40</b>
DIO 64PPN	Write	18.7	18.5	21.4
	Read	19.2	18.9	20.2
BIO 64PPN	Write	17.7	17.8	20.3
	Read	16.2	16.2	17.9
<b>2x JBODs (4 OSTs)</b>	<b>IO</b>	<b>IB (HDR) 6.1-010.39</b>	<b>Ethernet (RoCE) 6.1-010.48</b>	<b>KFI 6.1-010.40</b>
DIO 64PPN	Write	35.9	35.8	33.9
	Read	35.9	34.7	33.0
BIO 64PPN	Write	34.9	34.9	36.7
	Read	33.6	31.2	36.2

# E1000 SINGLE MDT RAID-10

- 20+ clients, Stonewalling IOR, GridRAID-53, Lustre 2.15 etc ....

MDT0  
OK files  
Non-DOM  
Unique Directory  
Files Only

<b>MDtest (Single MDT)</b>	<b>IB (HDR) 6.1-010.39</b>	<b>Ethernet (RoCE) 6.1-010.48</b>	<b>KFI 6.1-010.40</b>
File Creates per Second	155,771	156,851	100,699
File Stats per Second	682,054	683,682	524,510
File Reads per Second	301,053	293,466	210,164
File Removes per Second	143,286	152,239	113,314

MDT0  
OK files  
Non-DOM  
Unique Directory  
Directory+Files

<b>Single MDT</b>	<b>IB (HDR) 6.1-010.39</b>	<b>Ethernet (RoCE) 6.1-010.48</b>	<b>KFI 6.1-010.40</b>
Directory Creates/sec	109,971	101,232	83,437
Directory Stats/sec	406,336	406,419	351,709
Directory Removes/sec	175,605	180,357	136,213
File Creates/sec	156,606	159,457	101,253
File Stats/sec	679,259	683,807	526,747
File Reads/sec	284,603	255,884	212,517
File Removes/sec	143,981	155,846	122,003

# Lustre MDT performance - TCP/IP vs HDR IB RAID-10 LDISKFS

**MDT0 0K Files Unique Directory 1M objects**

Operation	ksocklnd	ko2iblnd	Diff
Dir creation	91 841	109 971	84%
Dir stats	319 945	406 336	79%
Dir removes	140 888	175 605	80%
File creation	108 772	156 606	69%
File stats	326 335	679 259	48%
File reads	184 085	284 603	65%
File removes	115 581	143 981	80%

**MDT0 32K Files Unique Directory 1M objects**

Operation	ksocklnd	ko2iblnd	Diff
Dir creation	97 151	109 271	89%
Dir stats	318 725	408 859	78%
Dir removes	139 571	193 565	72%
File creation	107 753	152 572	71%
File stat	339 420	698 665	49%
File read	182 009	216 228	84%
File Removes	112 215	149 733	75%

ksocklnd IOPS limited by TCP/IP latency and CPU utilization.

# Process and Lessons Learnt

---

## Switching from Intel to AMD Rome

- Started with BIOS setting with 4 NUMA domains (NPS4) and changed to 2 NUMA domains (NPS2) plus tuning Lustre CPU Partition Tables to 8 provided an increase in performance on throughput and IOPS

## Not all NVMe drives are the same

- Different NVMe drive vendors do not perform equal despite similar specs (e.g. Samsung PM1733 and Kioxia CM6)
  - New NVMe Firmware improved performance

## Keeping up with Linux enhancements takes a lot of work

- Moving from RHEL based 7.8 kernel with Lustre 2.12 to RHEL based 8.4 Kernel (e.g. Rocky Linux 8.4) with Lustre 2.15 provided additional improvement on DIO path
- Significant experimentation of Lustre tunables required

## New versions (client, server, ofed, OS etc.) often introduce regressions

- Repeated baseline testing is paramount to deliver consistency

# SUMMARY

---

- Since 2019 ClusterStor E1000 has improved **all** facets of performance **up to 400%** depending on the I/O operation
- Changes were due to constant tunings on the platform improvements with new kernels or adopting Lustre 2.15.
- ko2ibIpd performance is identical for HDR InfiniBand and HW RoCE 200 GigE
- ksockIpd performance is limited on peak writes or due to TCP/IP Latency
- Lustre 2.15 brings big **performance improvement** on MDT File Stats/s and single stream performance
  
- RDMA based protocols perform essentially the same regardless of type (e.g., Infiniband, RoCE or SlingShot) ...
- Continuous benchmarking is important for any product
  - Performance regressions can and will occur frequently....
  - The ability to confidently propose sizing to meet a future deployment is key
- Peak performance is repeatable but **ONLY** in the lab
  - No NOT expect hero numbers in customer environments (but we can get close) ...





THANK YOU

(for listening to a madmans ramblings ....)

tkp@hpe.com