Solving I/O Slowdown: DoM, DNE and PFL Working Together

John Fragalla

Bill Loewe

✉ [jfragalla,bloewe]@cray.com

CRAY™

Agenda

- Benchmark system configuration
- PFL baseline streaming performance
- Random 4K IO on small files: flash MDT vs flash OST with PFL
- ”Noisy Neighbor Problem” with PFL Small File Workload (random and sequential)
- MDTEST - DNE with and without DoM (Remote vs Sharded)
- Summary

System Setup

Hardware:

- Storage with EDR Server Nodes
  - 4 MDSs, each configured with a flash MDT RAID-10 – SAS SSDs
  - 2 OSS, each configured with a flash OSTs RAID-10 – SAS SSDs
  - 4 OSS, each configured with Parity Declustered RAID HDD OSTs (GridRAID)
- 64 Client nodes w/ FDR Connectivity
- EDR InfiniBand Non-Blocking Fabric

Software:

- Lustre 2.11.0 clients and server
- CentOS Linux release 7.5 (server and client)
- Spectre/Meltdown enabled kernels on Clients, disabled on Server
  - Client: 3.10.0-862.el7.x86_64
  - Server: 3.10.0-693.21.1.x3.1.9.x86_64

Disclaimer

- Results shared in this talk are intended to test various Lustre features with various I/O sizes to see relative results
- Performance results are not intended to show best results of the storage solution

Progressive File Layout (PFL) Base Streaming Performance

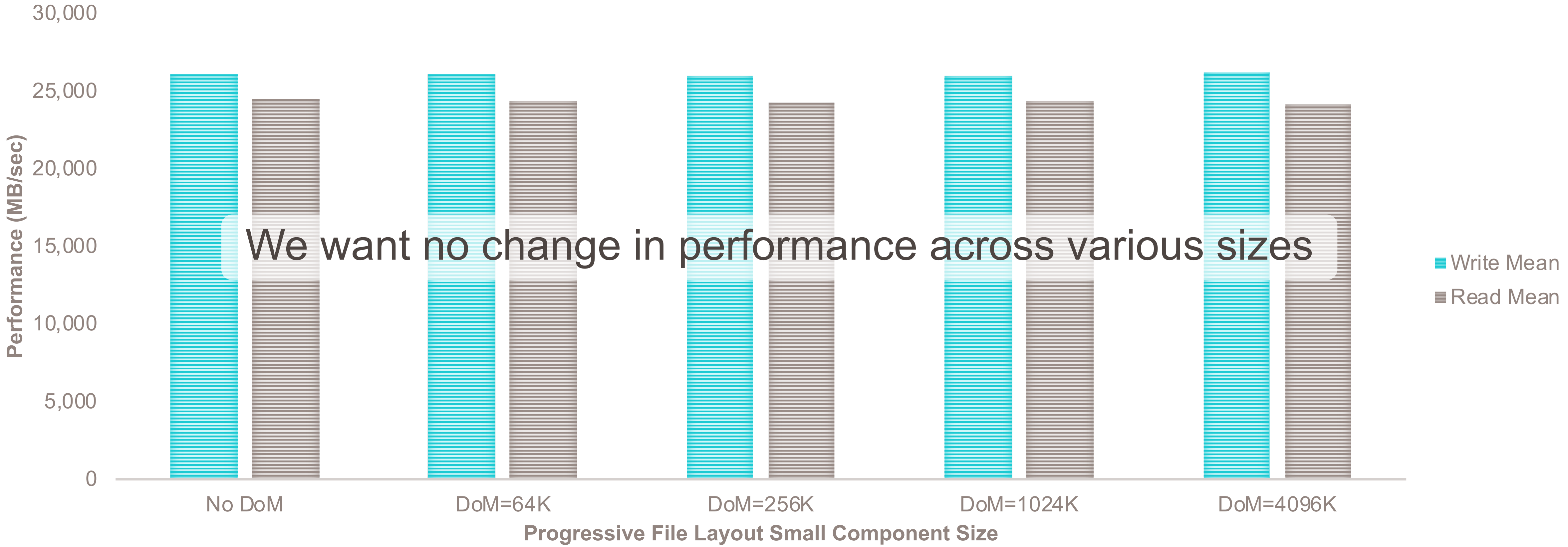
Sequential baseline results

- Measuring peak performance of 4 Disk OSTs with and without PFL, showing same peak throughput results
- Goal is demonstrate PFL with small file Layout to flash, large stream IO to disk has no effect on large streaming IO
- PFL Scheme
  - [0,1M] – DoM with Flash MDTs
  - [1M, EOF] – Disk OSTs
- IOR, DIO, 64m transfer, Larger IO, FPP, Stonewalling to measure peak throughput of L300N

PFL Scheme

- lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/dom1024
- lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/dom1024
- lfs setstripe -E **1M -L mdt** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/dom1024

LUSTRE **PFL STREAMING PERFORMANCE**



We want no change in performance across various sizes

Progressive File Layout maintains peak performance for streaming workloads

Random 4K IO with small files with flash targets

Random 4K IO with small files with flash targets

- Workload: small file with random 4K I/O, FPP, IOR, Direct IO
- Writing/Reading 32KB, 128KB, 512KB, 2MB, or 8M Files in 4K random blocks
  - PFL scheme on flash targets <=[64K, 256K, 1M, 4M]
- Two Benchmark Setups
  - Compared results of flash MDTs with and without DOM/PFL
  - Compared results of flash OSTs with and without PFL

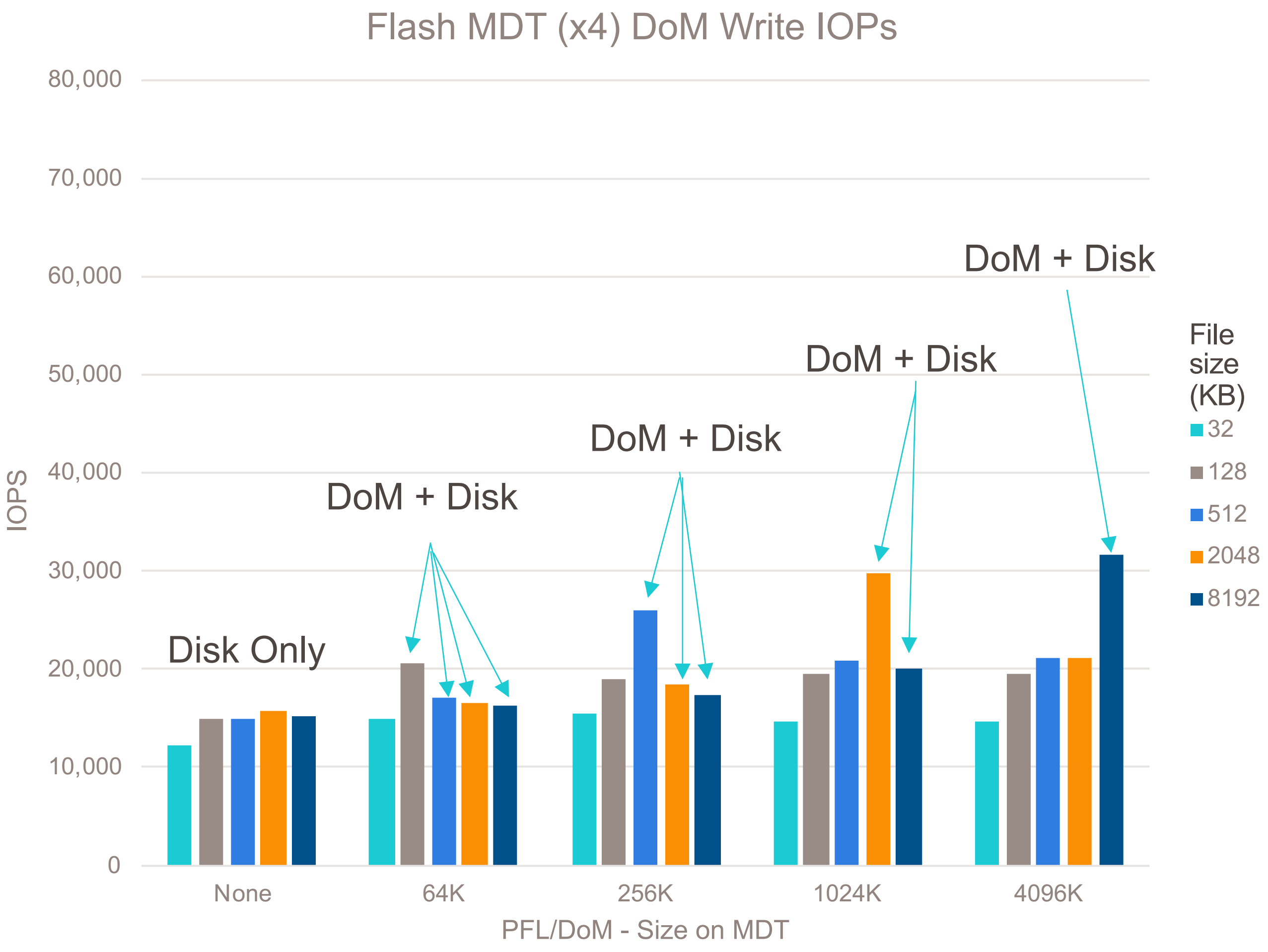
PFL scheme with DoM (4 MDTs)

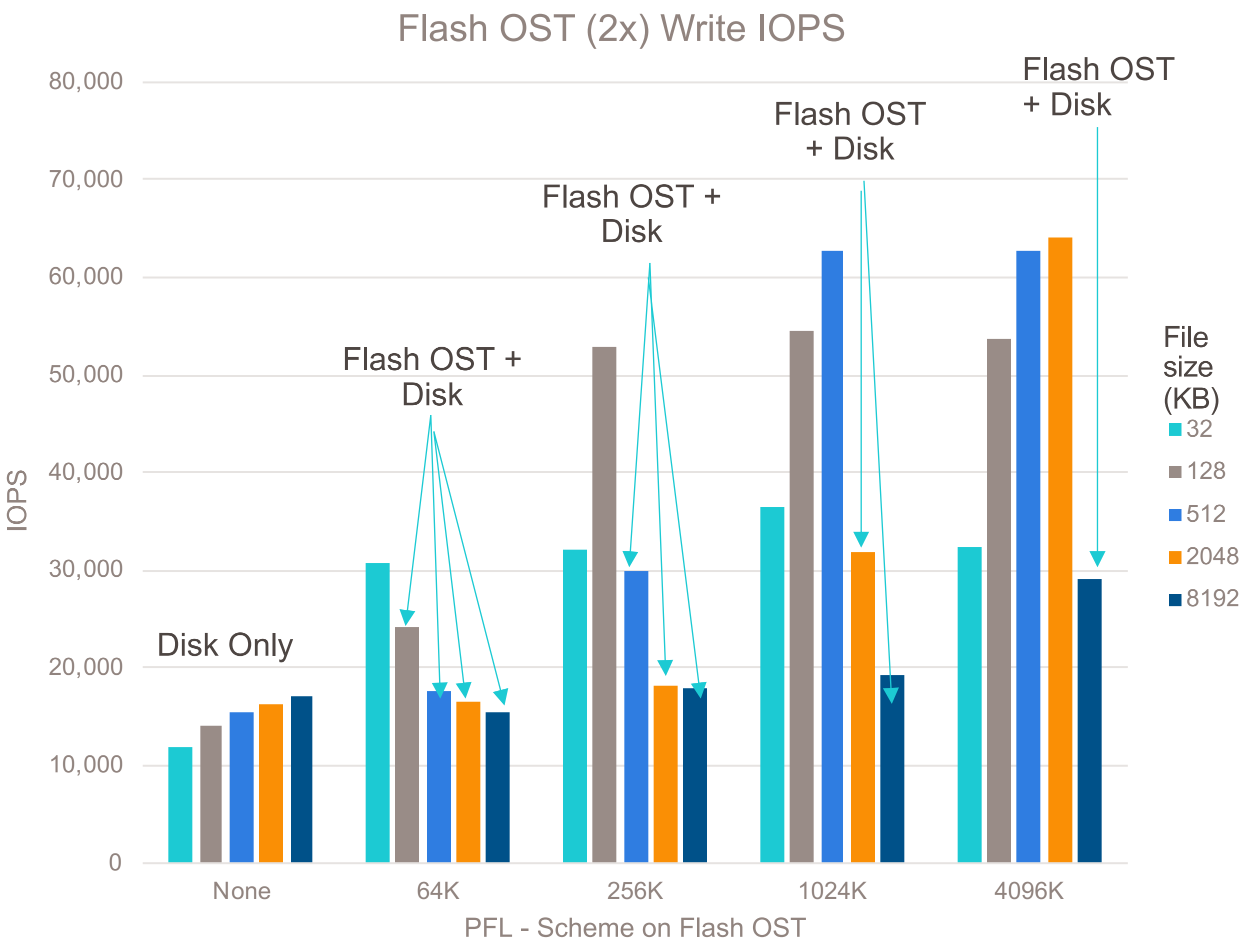
- **PFL with 0-64K land on MDTs >64K land on the HDD OSTs**
  - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/dom64
  - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/dom64
  - lfs setstripe -E **64K -L mdt** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/dom64
- **PFL with 0-256K land on MDTs > 256K land on HDD OSTs**
  - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/dom256
  - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/dom256
  - lfs setstripe -E **256K -L mdt** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/dom256
- **PFL with 0-1MB land on MDTs > 1MB land on HDD OSTs**
  - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/dom1024
  - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/dom1024
  - lfs setstripe -E **1M -L mdt** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/dom1024
- **PFL with 0-4MB land on MDTs, > 4MB land on HDD OSTs**
  - **mgs# lctl conf_param testfs-MDT000[0-3].lod.dom_stripesize=4M**
  - mgs# pdsh -g mds lctl get_param lod.*.dom_stripesize
  - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/dom4096
  - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/dom4096
  - lfs setstripe -E **4M -L mdt** -E -1 **-p testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/dom4096

PFL scheme with 2x flash OSTs

- **PFL with 0-64K is land on Flash OSTs > 64K land on the disk OSTs**
  - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/flash64
  - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/flash64
  - lfs setstripe -E **64K** -p **testfs.flash -c 1 -S 64K** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/flash64
- **PFL with 0-256K is land on Flash OSTs > 256K hit the disk OSTs**
  - lfs mkdir -c 4 -i 0,1,2,3 /mnt/lustre/benchmark/flash256
  - lfs mkdir -c 4 -D /mnt/lustre/benchmark/flash256
  - lfs setstripe -E **256K** -p **testfs.flash -c 1 -S 256K** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/flash256
- **PFL with 0-1MB is land on Flash OSTs > 1MB land on disk OSTs**
  - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/flash1024
  - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/flash1024
  - lfs setstripe -E **1M** -p **testfs.flash -c 1 -S 1m** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/flash1024
- **PFL with 0-4MB is land on Flash OSTs > 4MB land on disk OSTs**
  - mgs# **lctl conf_param testfs-MDT000[0-3].lod.dom_stripesize=4M**
  - mgs# pdsh -g mds lctl get_param lod.*.dom_stripesize
  - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/flash4096
  - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/flash4096
  - lfs setstripe -E **4M** -p **testfs.flash -c 1 -S 1m** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/flash4096

4KB IOPs writes: flash comparison





”Noisy Neighbor Problem” with PFL

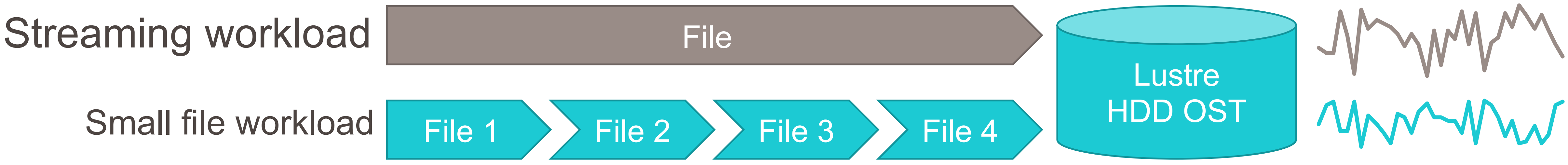
Small file competing workload

”Noisy neighbor problem” with PFL sequential small file workload

- Two Competing Benchmarks Writing to the same PFL Layout
  - Foreground **Measured** Benchmark: Large Sequential IOR measuring L300N Streaming Performance
    - Competing benchmark “Noisy Neighbor”: Small Files using MDTEST (and IOR Random 4K) Workload
- PFL scheme
  - Layout 2: [<=1M, 4M] to Flash Targets using PFL, rest of the data to Disk
- File Sizes: Writing/Reading 1MB or 4MB Files with MDTEST (and IOR Random 4K)
  - Noisy Neighbor Benchmark used 1MB or 4MB Files to show the performance effects of the Foreground Benchmark for this particular benchmark setup

Lustre PFL “noisy neighbor” isolation

Two competing workloads on same HDD resources



Streaming workload

File

Small file workload

File 1

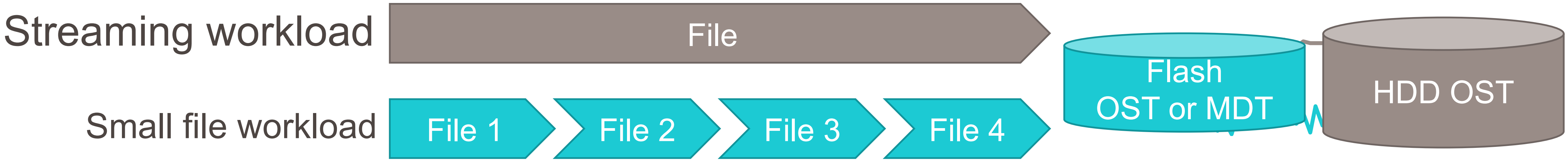
File 2

File 3

File 4

Lustre
HDD OST

Two competing workloads with PFL scheme



Streaming workload

File

Small file workload

File 1

File 2

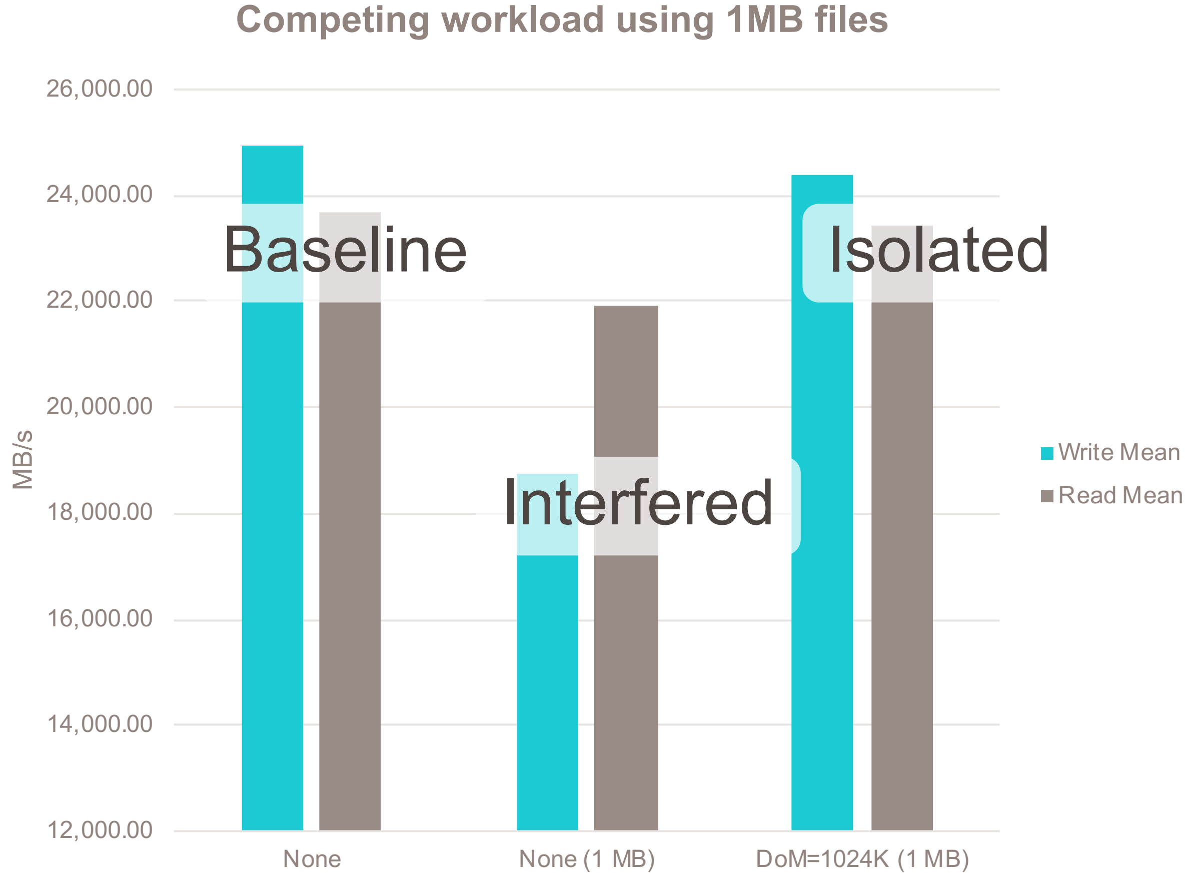
File 3

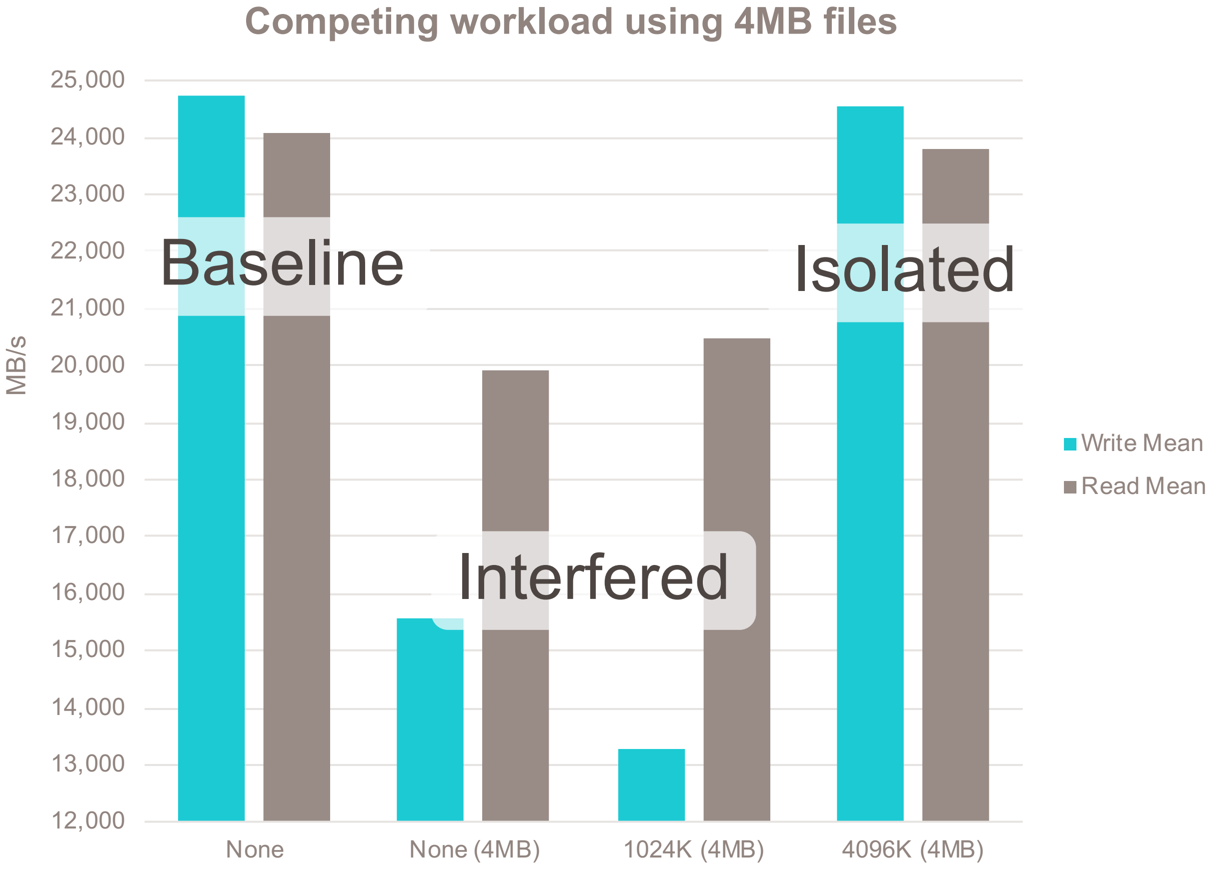
File 4

Flash
OST or MDT

HDD OST

Lustre PFL “noisy neighbor” isolation
Flash tier (OST or DoM) -> HDD OST tier





X-Axis Legend
PFL Size on Flash (Noisy Neighbor File Size)

PFL isolation for small I/O from streaming I/O improves performance

MDTEST DNE with and without DoM

MDTEST - DNE with and without DoM

- Evaluated DNE Remote Directory vs DNE Sharded Directory with up to 4 MDT Flash Targets
- MDTEST with 0KB and 32KB Files with and without DoM, unique and shared Directory
- DNE Remote Directory provide near linear scaling for Metadata operations
- Sharded Directory improves single directory Metadata operations and allows more inodes in a single directory

Unique directory: DNE1 and DNE2 with Flash MDTs (with and without DoM)

	0KB Files - Unique Directory				
DNE Striping	**Files/MDT**	**File Create/s**	**File Stat/s**	**File Read/s**	**File Unlink/s**
Remote Directory – 1x MDT **(No DoM)**	1 048 576	85 142	310 410	150 618	94 711
Remote Directory – 4x MDTs **(No DoM)**	1 048 576	**261 318**	**754 905**	615 785	389 527
Sharded Directory – 4x MDTs **(No DoM)**	1 048 576	167 611	753 885	602 834	346 796
Sharded Directory – 4x MDTs **(64K DoM)**	1 048 576	**352 809**	**1 053 564**	787 548	373 597

	32KB Files - Unique Directory				
DNE Striping	**Files/MDT**	**File Create/s**	**File Stat/s**	**File Read/s**	**File Unlink/s**
Remote Directory – 1x MDT **(No DoM)**	1 048 576	83 007	**315 608**	151 369	37 000
Remote Directory – 4x MDTs **(No DoM)**	1 048 576	174,833	1,222,748	606,567	20,694
Sharded Directory – 4x MDTs **(No DoM)**	1 048 576	**159 109**	1 210 448	596 610	20 532
Sharded Directory – 4x MDTs **(64K DoM)**	1 048 576	**89,266**	1,164,580	778,803	**191,191**

Shared directory: DNE1 and DNE2 with Flash MDTs (with and without DoM)

	0KB Files - Shared Directory				
DNE Striping	**Files/MDT**	**File Create/s**	**File Stat/s**	**File Read/s**	**File Unlink/s**
Remote Directory 1x MDT ***(No DoM)***	1 048 576	**76 578**	181 320	152 441	80 390
Sharded Directory - 4x MDTs ***(No DoM)***	1 048 576	148 974	428 402	605 334	187 857
Sharded Directory - 4x MDTs **(64K DoM)**	1 048 576	**174 572**	332 047	**823 025**	189 968

	32KB Files - Shared Directory				
DNE Striping	**Files/MDT**	**File Create/s**	**File Stat/s**	**File Read/s**	**File Unlink/s**
Remote Directory 1x MDT ***(No DoM)***	1 048 576	76 515	180 198	151 425	35 700
Sharded Directory - 4x MDTs ***(No DoM)***	1 048 576	128 437	354 109	590 935	19 995
Sharded Directory - 4x MDTs **(64K DoM)**	1 048 576	**80,747**	346,724	501,908	**98,762**

Remote and Sharded DNE Setup

- **DNE2 Sharded Directory with DoM**
  - PFL with 0-64K land on MDTs/DoM > 64K land on HDD OST
    - lfs mkdir **-c 4 -i 0,1,2,3** /mnt/lustre/benchmark/dom64
    - lfs mkdir **-c 4 -D** /mnt/lustre/benchmark/dom64
    - lfs setstripe -E **64K -L mdt** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/dom64
- **DNE1 Remote Directories with DoM**
  - lfs mkdir **-i 0** /mnt/lustre/benchmark/mdt0
  - lfs setstripe -E **64K -L mdt** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/mdt0
  - lfs mkdir -i 1 /mnt/lustre/benchmark/mdt1
  - lfs setstripe -E **64K -L mdt -E -1** -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/mdt0
  - lfs mkdir -i 2 /mnt/lustre/benchmark/mdt2
  - lfs setstripe -E **64K -L mdt -E -1** -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/mdt0
  - lfs mkdir -i 3 /mnt/lustre/benchmark/mdt3
  - lfs setstripe -E **64K -L mdt** -E -1 -p **testfs.disk -c 1 -S 1m** /mnt/lustre/benchmark/mdt0

Summary

Summary

- Lustre PFL validated sequential performance was not affected
- Lustre PFL is a good solution to isolate small I/O (random/sequential) on Flash to not affect performance of sequential I/O
- Lustre PFL allows transparent use of Flash and HDDs
- Flash on Metadata or OSTs is a good solution for small I/O
- Sharded Directory better at automated optimization than DNE1.
- Automated striping, Sharded Directory is preferred, and scales higher than single MDT, but lower than peak performance
- DoM with MDTEST improves read performance
- Sharded Directory allows more files in a single directory than DNE1

THANK YOU

QUESTIONS?

cray.com