LAD'17: LUSTRE* GENERATIONAL PERFORMANCE IMPROVEMENTS & NEW FEATURES

ADAM ROE
HPC SOLUTIONS ARCHITECT

HIGH LEVEL ABSTRACT

Lustre* has had a number of compelling new features added in recent releases; this talk will look at those features in detail and see how well they all work together from both a performance and functionality perspective. Comparing some of the numbers from last year we will see how far the Lustre* filesystem has come in such a short period of time (LAD'16 to LAD'17), comparing the same use cases observing the generational improvements in the technology.

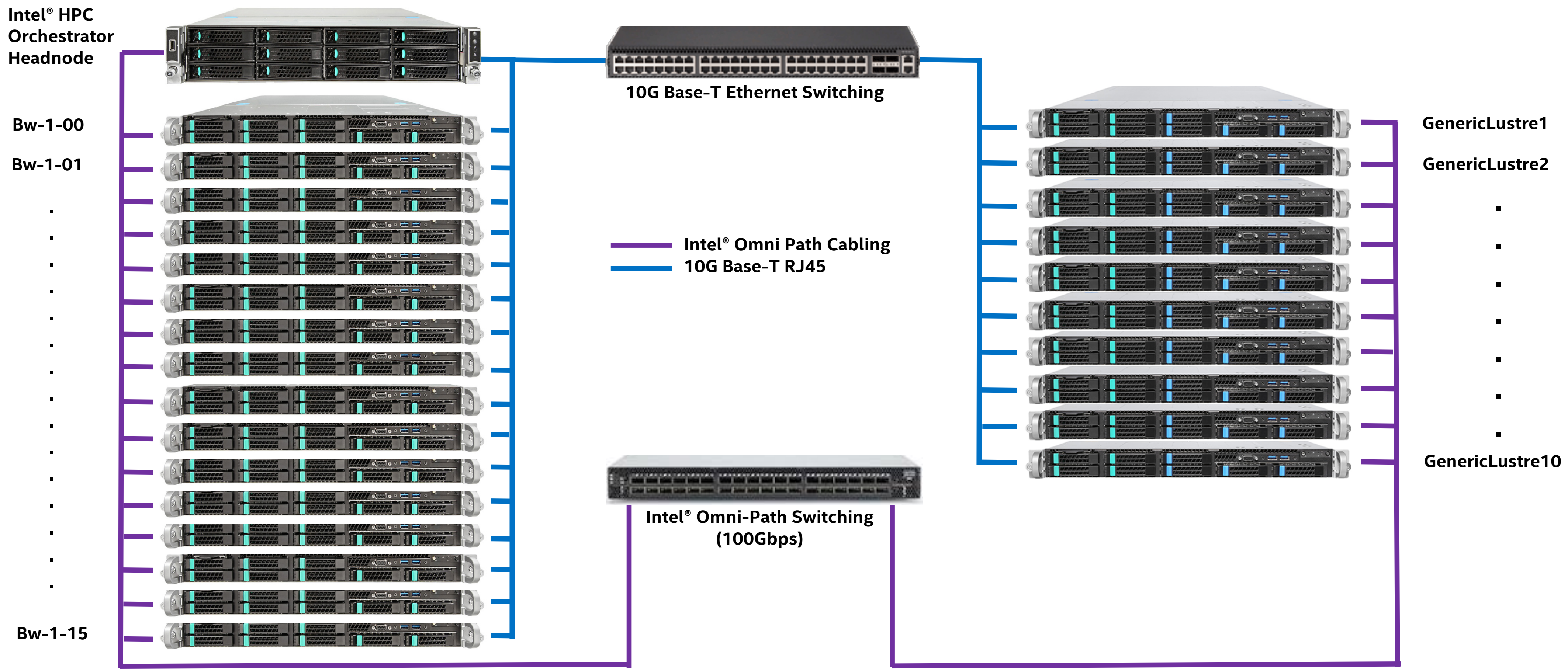
AGENDA

- Hero Numbers: Generational Performance March 2016 – today
  - Generational Metadata performance improvements
  - Small file performance on OpenZFS (no Data-on-MDT)
  - Has LDISKFS changed since last year, how does performance look today
- Scaling with DNE Phase 2
- How does PFL effect Performance

SUMMARY OF LAST YEARS TALK

- **No LDISKFS Numbers with DNE:**
  - Stability issues I observed have been resolved, see some new DNE 2 numbers with LDISKFS in this talk
- **DNE Phase 2 Scalability:**
  - Scalability was reasonable before, do new Lustre release demonstrate better scalability
- **Using DNE 2 In Production:**
  - Yes, I am still using DNE2 in production successfully for 18 months

TESTBED ARCHITECTURE



TESTBED ARCHITECTURE (CONT.)

Server

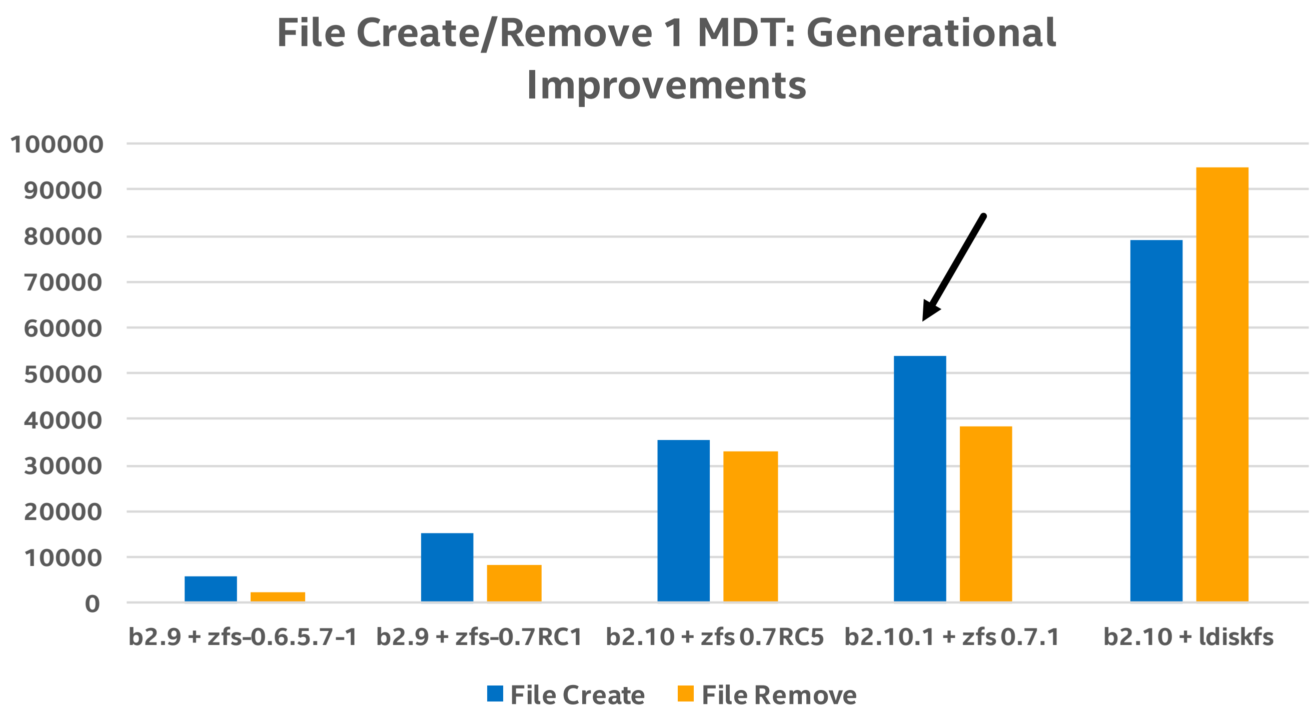
- 10x Generic Lustre servers with two slightly different configurations
  - Each System comprises of:
    - 2x Intel® Xeon E5-2697v3 (Haswell) CPU's
    - 1x Intel® Omni-Path x16 HFI
    - 128GB DDR4 2133MHz Memory
    - Eight of the nodes contain - 4x Intel P3600 2.0TB 2.5" (U.2) NVMe devices, while the other two have 4x Intel® P3700 800GB 2.5" (U.2) NVMe devices
    - One node equipped with 2x Intel® S3700 400GB's for MGT
- 16x 2S Intel® Xeon E5v4 (Broadwell) Compute nodes
  - 1x Intel® HPC Orchestrator (Beta 2) Headnode
  - Hardware Components:
    - 2x Intel® Xeon E5-2697v4 (Broadwell) CPU's
    - 1x Intel® Omni-Path x16 HFI
    - 128GB DDR4 2400MHz Memory
    - Local boot SSD
- 100Gbps Intel® Omni-Path Fabric
  - None-blocking fabric with single switch design.
  - Server side optimisations: "options hfi1 sge_copy_mode=2 krcvqs=4 wss_threshold=70"
    - Improve generic RDMA performance, only recommended on Lustre server side that do physically do any MPI

WHY?

GENERATIONAL PERFORMANCE IMPROVEMENTS
LUSTRE 2.9EA TO LUSTRE 2.10.1

METADATA PERFORMANCE

Lustre Metadata performance 2.9EA (Mid 2016) vs. 2.10.1 (today), MDTEST: Single MDT Performance.

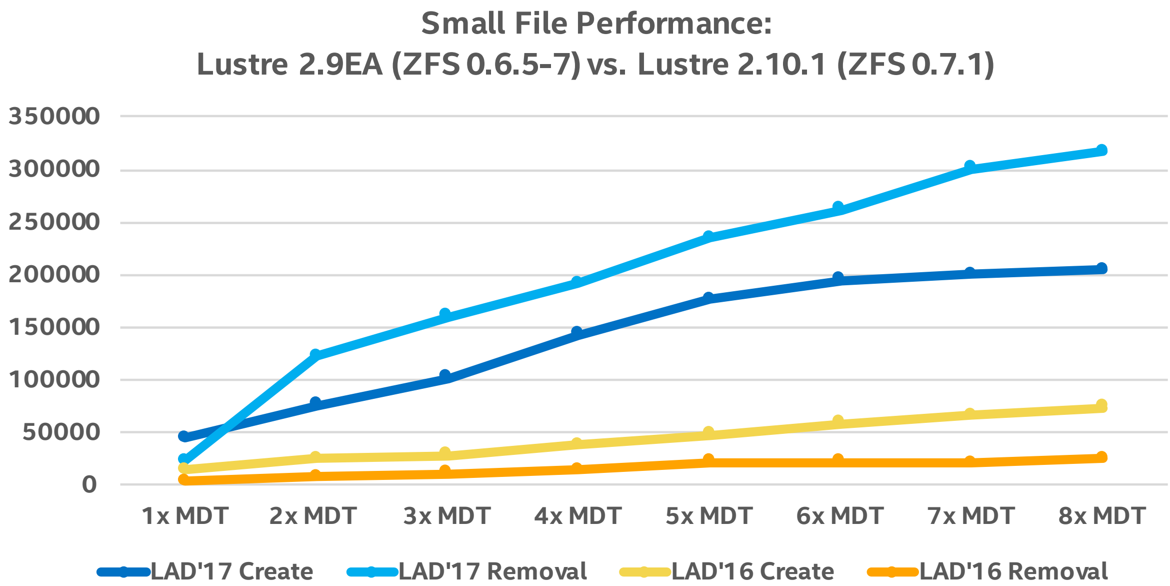


- LDISKFS quite close to ZFS for file create about 75%
- Removal still a way to go
- When testing on slower storage difference are marginal
- Demonstrates a good level of improvement.

/mnt/zlfs2/mdtest -i 3 -I 10000 -F -C -T -r -u -d /mnt/zlfs2/test1.out

SMALL FILE PERFORMANCE (4K)

Lustre Small file performance 2.9EA (Mid 2016) vs. 2.10.1 (today), MDTEST: Single MDT Performance. No Data-on-MDT used leveraging DNE Phase 2 up to 8 MDT's on Separate servers.



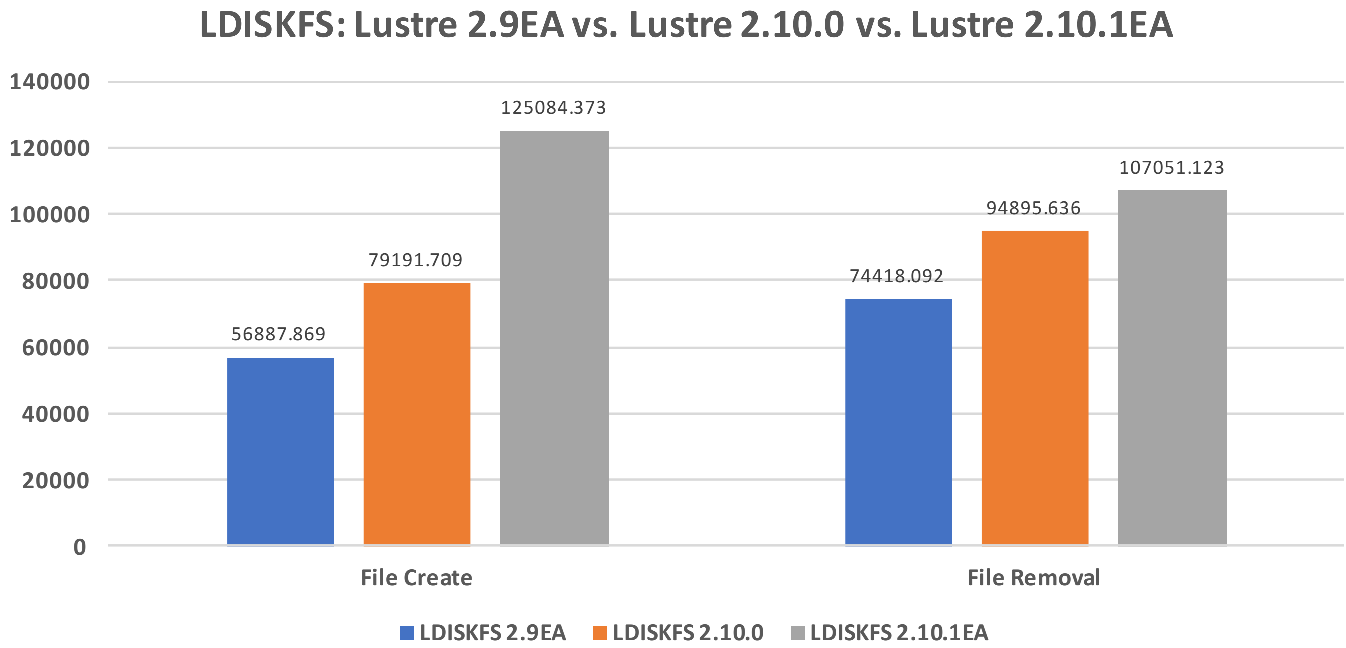
- Single MDT operation up **4x** compared to this time last year
- Scaling of DNE2 still not linear, but better
- Create performance trails off due to lack of clients
- Clear benefit versus the previous release

/mnt/zlfs2/mdtest -i 3 -I 10000 -z 1 -b 1 -L -u -F –w 4096 -d /mnt/zlfs2/*DNE-DIR*/test1.out

LDISKFS: GENERATIONAL METADATA IMPROVEMENTS

LDISKFS PERFORMANCE

Lustre 2.9EA vs. Lustre 2.10.0 vs. Lustre 2.10.1

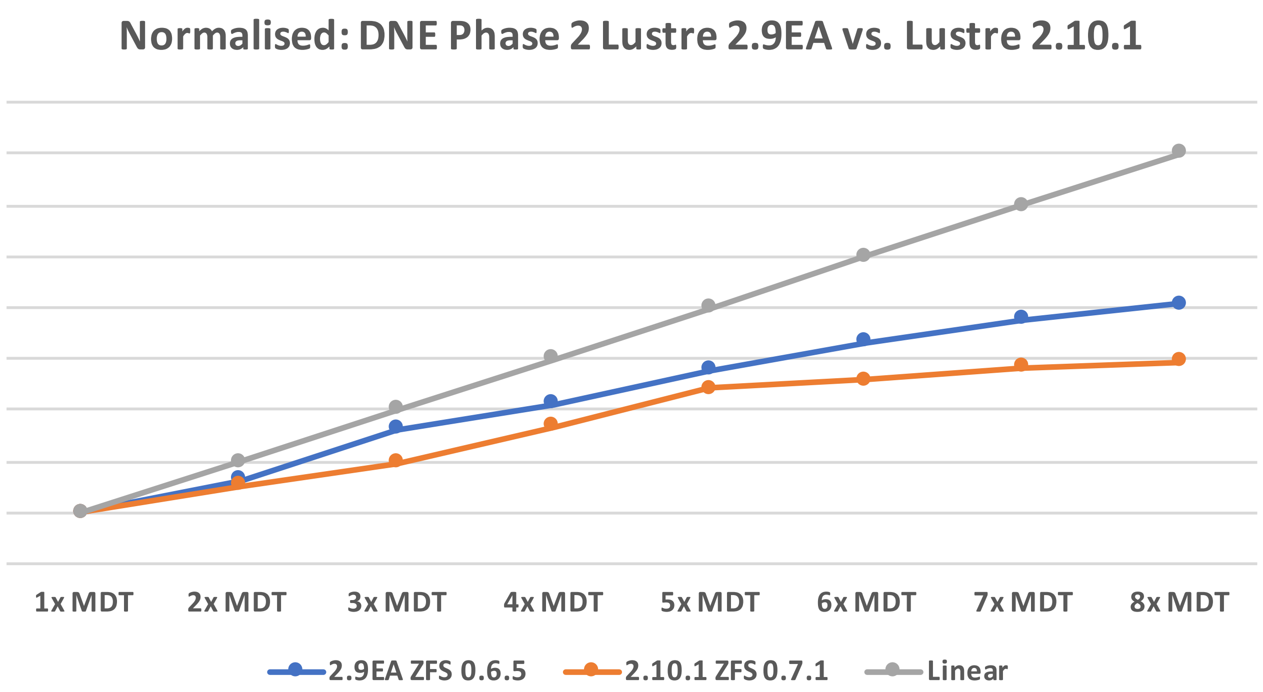


- Some performance boost was expected, but not this much
- Shows positive trend version to version
- 2.10 to 2.10.1 - LU-7899

DNE PHASE II SCALING
LUSTRE 2.9EA VS. 2.10.1

NORMALISED: SCALING LAD'16 TO LAD'17: DNE PHASE 2

Generational Performance and scalability of DNE Phase 2 on OpenZFS

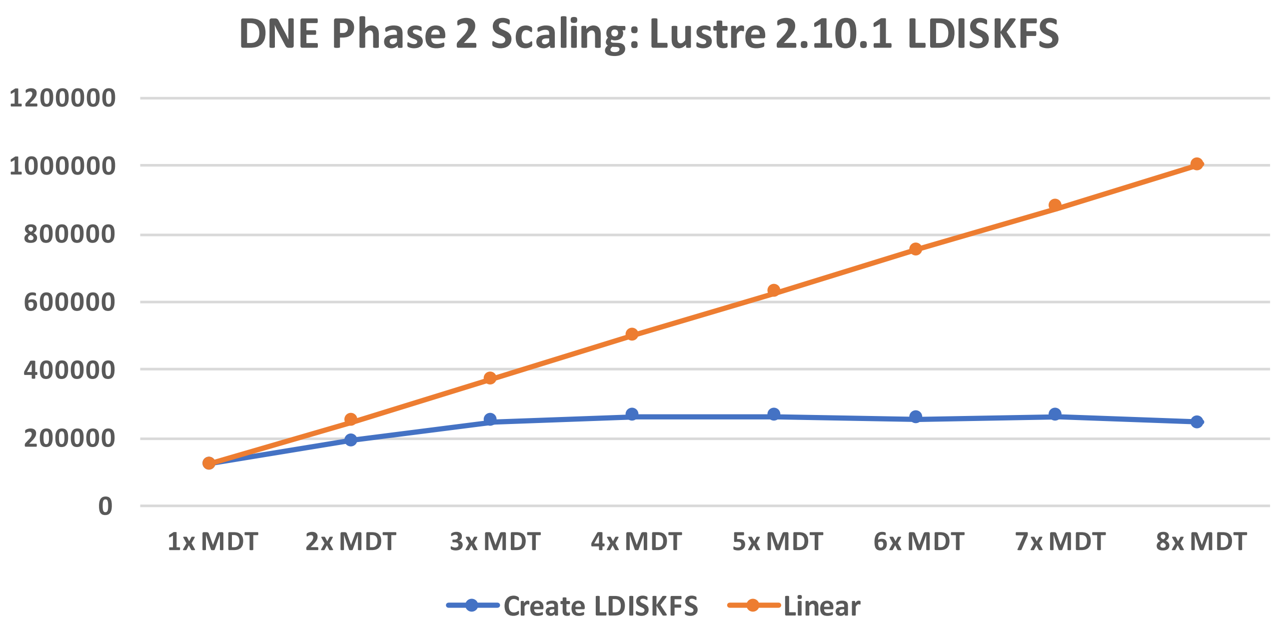


- Neither are linear
- Overall scalability dropped a little, but ultimate number is much higher
- Some work to do to get this close to OST scalability

DNE PHASE II ON LDISKFS

LDISKFS DNE PHASE 2 SCALING & FUNCTIONALITY

Following up from last year where I couldn't give DNE2 on LDISKFS.

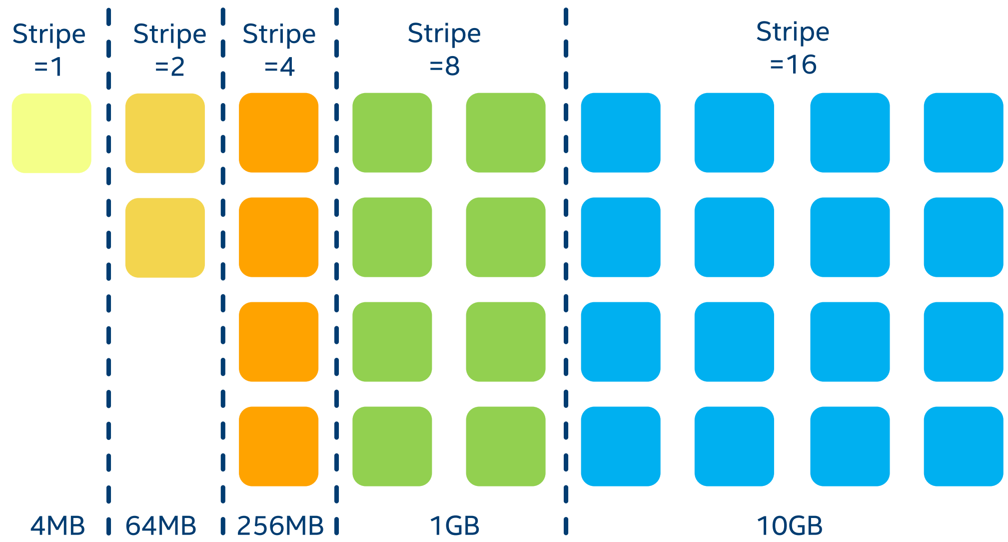


- Totally stable, versus pervious testing
- Scaling stops after 2 MDT's
  - Clients unable to push that much I/O
  - You can see from the previous slide 200 – 250k is my HW limit

PROGRESSIVE FILE LAYOUT

PROGRESSIVE FILE LAYOUT

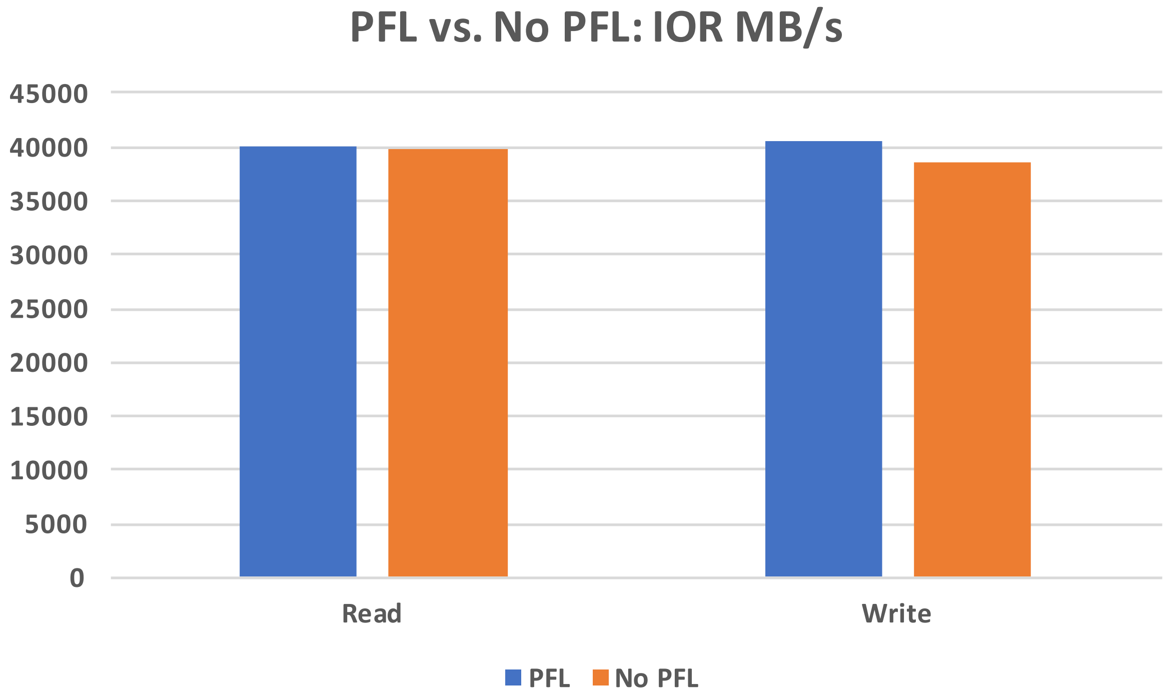
- Example Layout
- lfs setstripe -E 4M -c 1 -E 64M -c 2 -E 256M -c 4 -E 1G -c 8 -E 10G -c 16 -E -1 -c -1 /mnt/zlfs2/pfl_test01/



```
/mnt/zlfs2/pfl_test01/
  lcm_layout_gen:  0
  lcm_entry_count: 6
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 0
    lcme_extent.e_end:   4194304
      stripe_count:  1       stripe_size:   1048576       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 4194304
    lcme_extent.e_end:   67108864
      stripe_count:  2       stripe_size:   1048576       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 67108864
    lcme_extent.e_end:   268435456
      stripe_count:  4       stripe_size:   1048576       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 268435456
    lcme_extent.e_end:   1073741824
      stripe_count:  8       stripe_size:   1048576       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 1073741824
    lcme_extent.e_end:   10737418240
      stripe_count:  16       stripe_size:   1048576       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 10737418240
    lcme_extent.e_end:   EOF
      stripe_count:  -1       stripe_size:   1048576       stripe_offset: -1
```

PFL PERFORMANCE WHEN USING IOR

Lustre 2.10.1; IOR Performance, file per process (256 files, 16GB per file) mean performance MB/s. PFL as described before versus traditional -1 stripe.



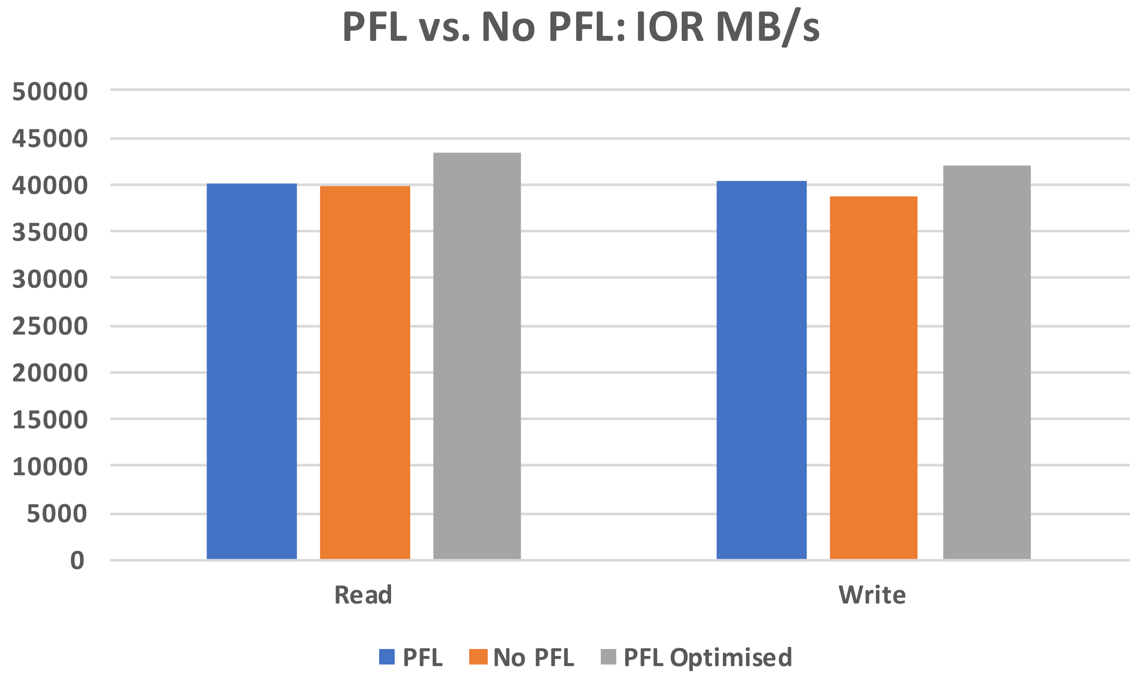
- Each files stripe dynamically grows based on file size
- Write performance up 4.6%, read within margin of error
- Certainly not detrimental to performance

/mnt/zlfs2/IOR -wr -C -F -i 3 -t 1m -b 1m -s 16384 -a MPIIO -o /mnt/zlfs2/pfl_new/testme1.file

```
/mnt/zlfs2/pfl_new1/
  lcm_layout_gen:  0
  lcm_entry_count: 6
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 0
    lcme_extent.e_end:   4194304
      stripe_count:  1       stripe_size:   1048576       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 4194304
    lcme_extent.e_end:   67108864
      stripe_count:  2       stripe_size:   2097152       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 67108864
    lcme_extent.e_end:   268435456
      stripe_count:  4       stripe_size:   16777216       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 268435456
    lcme_extent.e_end:   1073741824
      stripe_count:  8       stripe_size:   33554432       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 1073741824
    lcme_extent.e_end:   10737418240
      stripe_count:  16       stripe_size:   134217728       stripe_offset: -1
    lcme_id:             N/A
    lcme_flags:          0
    lcme_extent.e_start: 10737418240
    lcme_extent.e_end:   EOF
      stripe_count:  -1       stripe_size:   268435456       stripe_offset: -1
```

PFL PERFORMANCE WITH USING IOR (CONT.)

Lets optimise the stripe size this time, assuming the a larger stripe size for the higher stripe counts
lfs setstripe -E 4M -S 1M -c 1 -E 64M -S 2M -c 2 -E 256M -S 16M -c 4 -E 1G -S 32M -c 64 -E 10G -S 128M -c 16 -E -1 -S 256M -c -1 /mnt/zlfs2/pfl_new/



- PFL is giving us the opportunity to optimise the stripe relative to data type
- Write up 8.7% on base results and and 4.1% relative to test one
- Reads get a 9.2% boost with larger stripe sizes

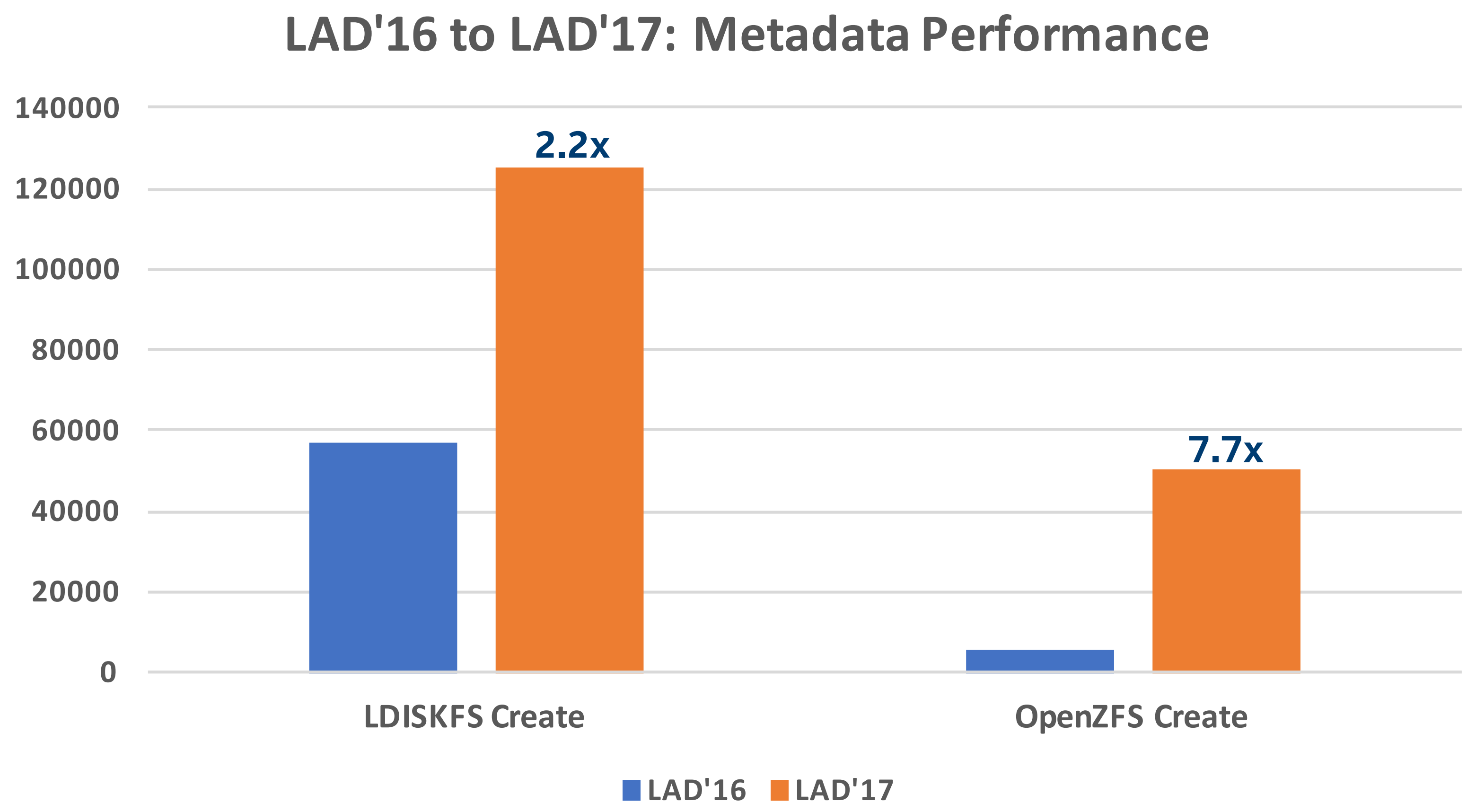
/mnt/zlfs2/IOR -wr -C -F -i 3 -t 1m -b 1m -s 16384 -a MPIIO -o /mnt/zlfs2/pfl_new/testme1.file

GENERATIONAL SUMMARY
LAD'16 TO LAD'17

KEY TAKEAWAYS

- Small file / metadata performance across the board is simply just better
- Amazing work done on OpenZFS to get 0.7.1 to where it is today
  - Performance is comparable to LDISKFS of previous releases
- Overall DNE Phase 2 scalability is very similar to what we have seen before
  - Overall usage and stability feels better, but was good before
- Optimising Striping layouts with PFL is essential, striping is done for you and can be configured for best performance

IF YOU TAKE ANYTHING AWAY FROM THIS TALK...



LEGAL INFORMATION

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html.

Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

experience
what's inside™