Lustre* on ZFS*

Jinshan Xiong

Software Engineer

High Performance Data Division

October 20, 2016

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.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

* Other names and brands may be claimed as the property of others.

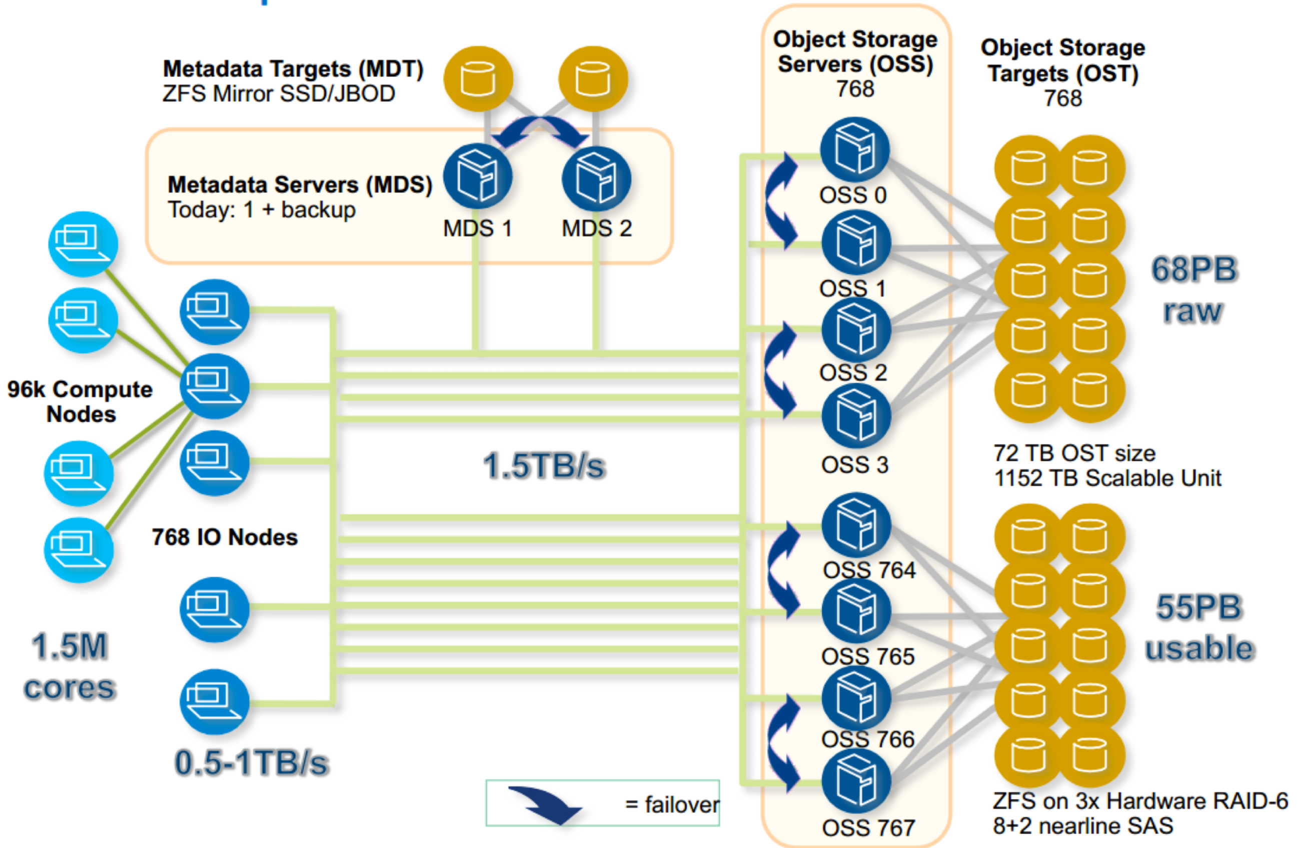
Agenda

- Motivations
- Lustre on ZFS Implementation
- Lustre on ZFS I/O Performance
- Lustre on ZFS Metadata Performance
- Intel Contributions to ZFS
- Key Takeaways

Motivations

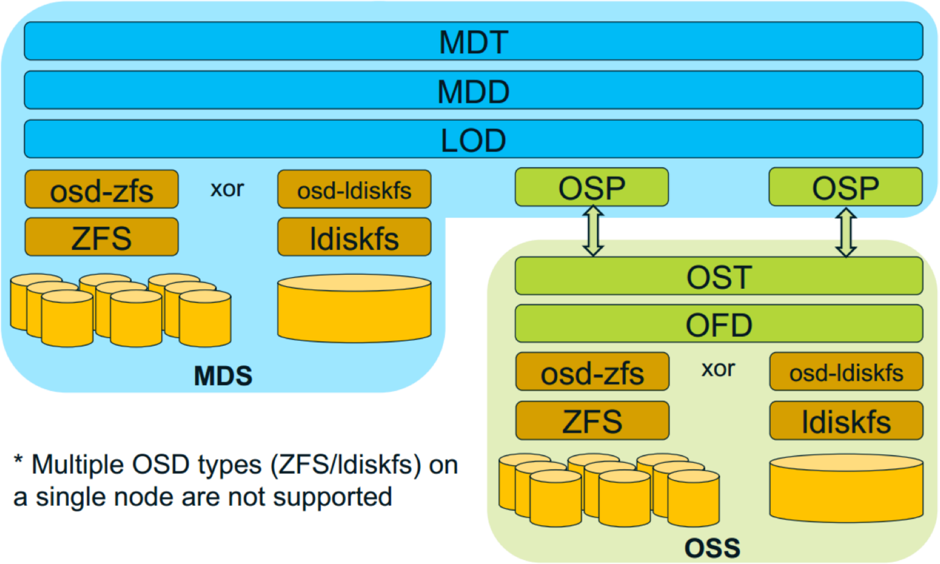
- Sequoia Requirements
  - 768 ZFS OSTs, 72TB each, 55PB capacity, 850GB/s I/O
  - Beyond the ability of ldiskfs
- ZFS Benefits
  - Superb write performance
  - Copy on write, always on-disk persistent
  - Built-in block checksum
  - Built-in disk management, RAIDZ, Mirror, etc
  - Built-in snapshot support
  - Scalable, online filesystem check/scrub/repair

LLNL Sequoia Lustre Architecture



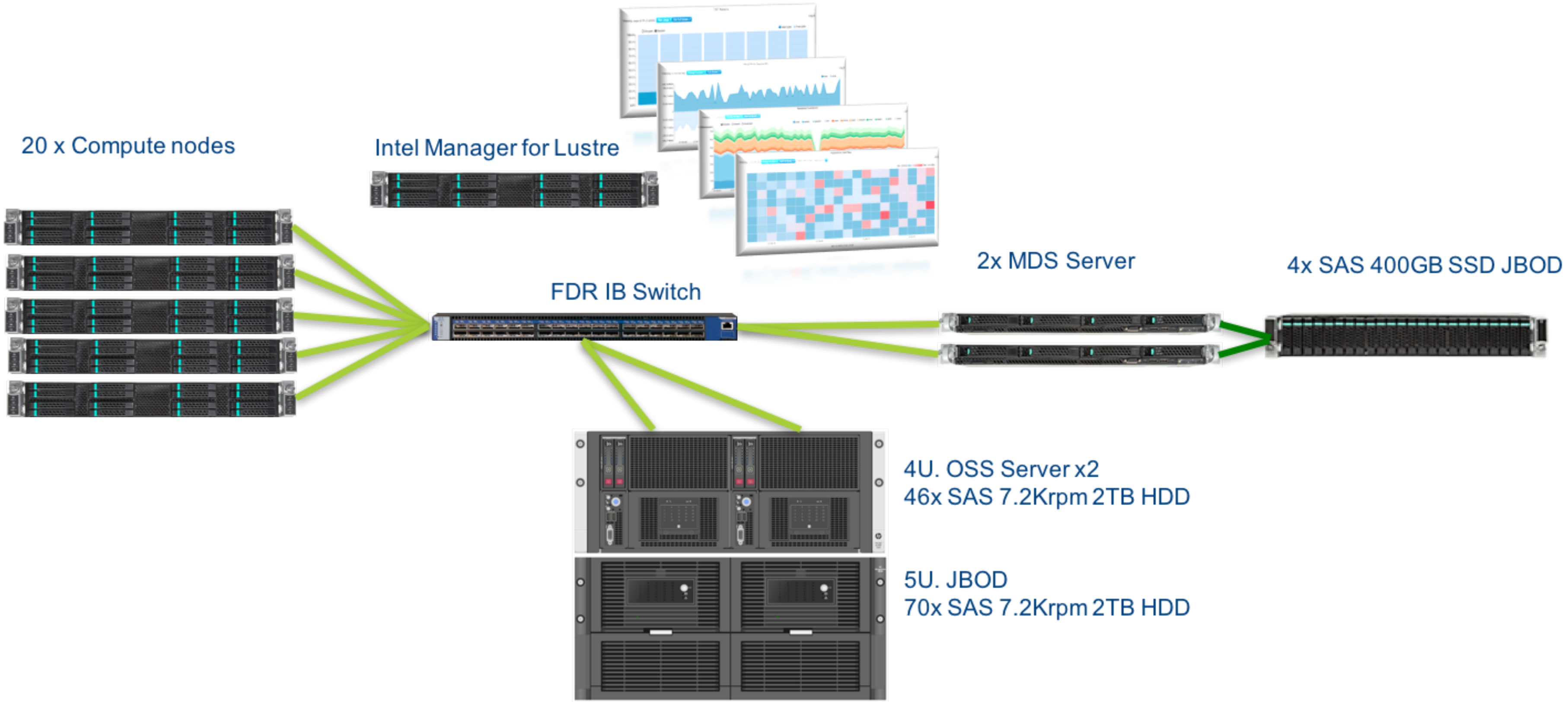
Implementation

- Use ZFS as a backend storage
- OSD-ZFS talks to DMU
  - ZPL compatible on-disk format
  - Can mount MDT/OST with ZFS kernel module
- Significant changes on Lustre side

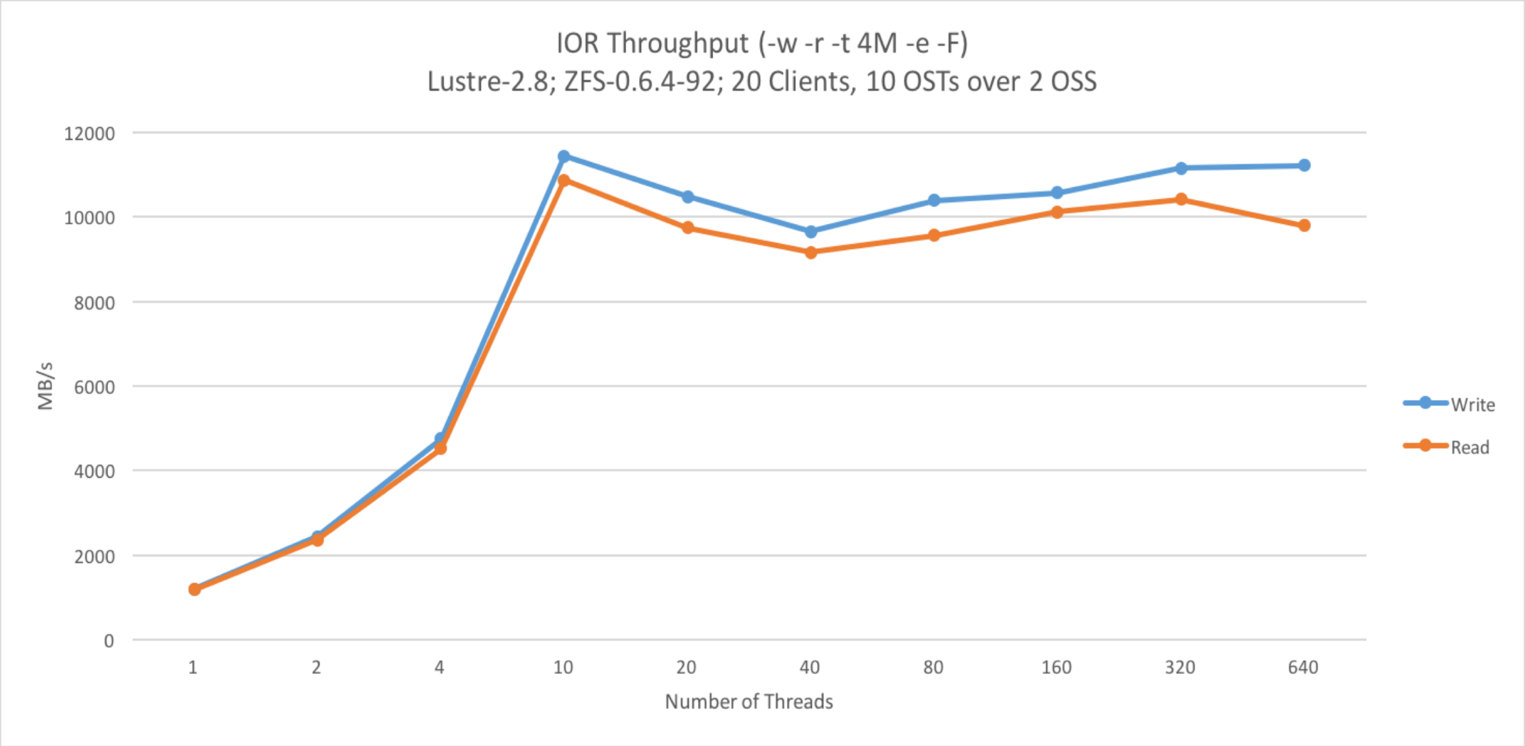


* Multiple OSD types (ZFS/ldiskfs) on a single node are not supported

Lustre I/O Performance on ZFS



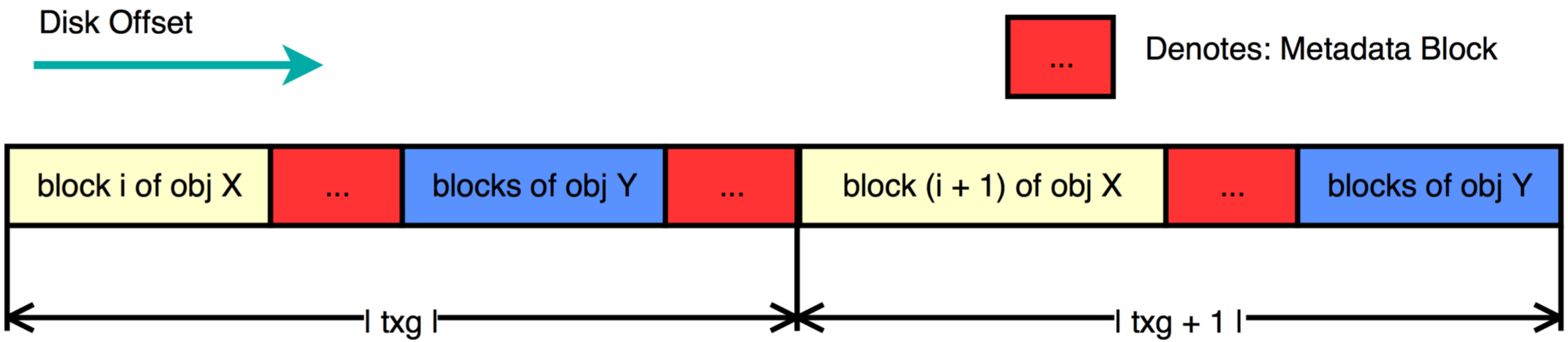
Lustre I/O Performance on ZFS - Results



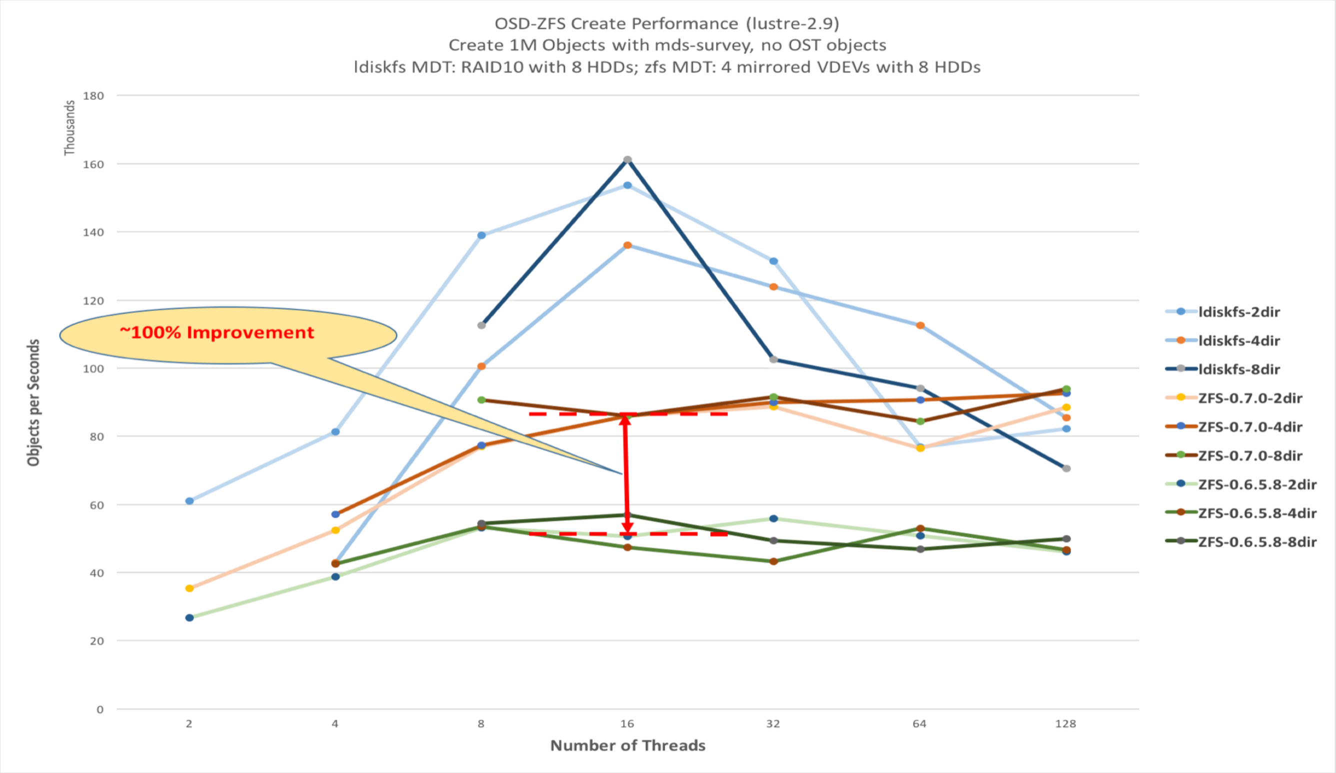
* Based on hardware configuration on slide #19

ZFS Read Problems

- No object aware block allocation
  - Blocks written sequentially may spread around the whole pool
  - Lots of disk seeks to read them back
- Read is usually slower than write
  - Bigger block size could mitigate this problem



Lustre on ZFS metadata performance



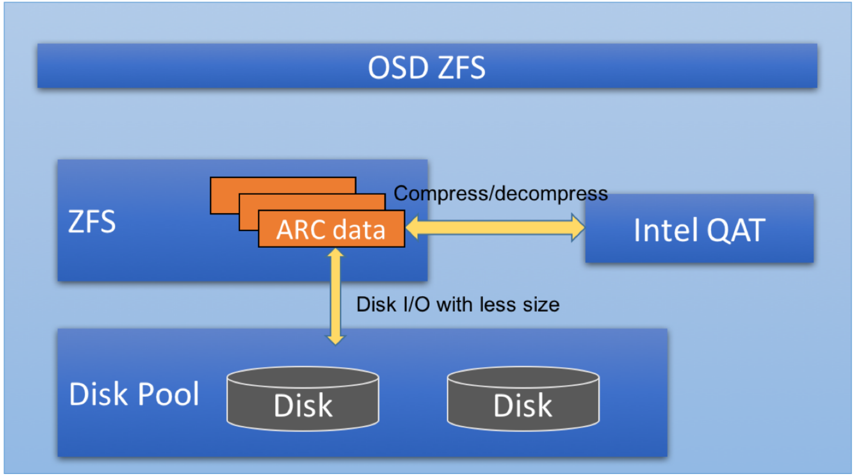
* Based on hardware configuration on slide #20

Intel Contributions to ZFS

- Compute fletcher-4 with vector instructions
  - 4 times faster than original version
  - https://github.com/zfsonlinux/zfs/pull/4330
- ZFS dnode quota
  - Significant metadata performance improvement
  - https://github.com/zfsonlinux/zfs/pull/3983
- Intel is collaborating with Delphix* to improve ZFS metadata performance
  - The work will be landed to ZoL and OpenZFS*

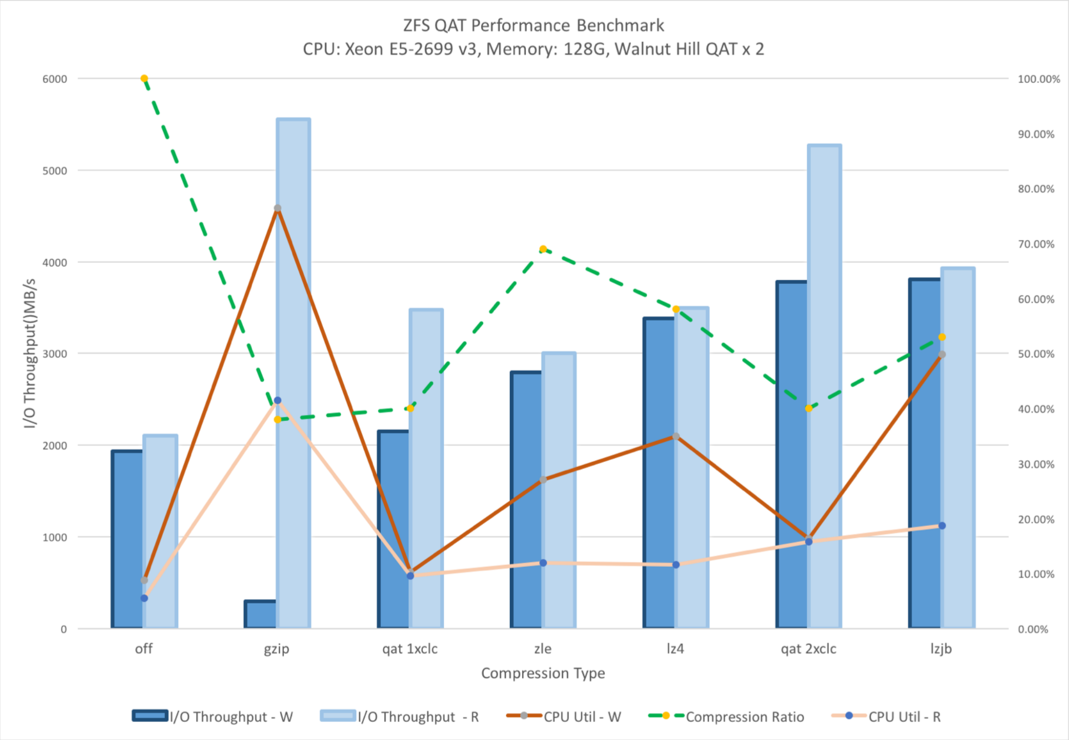
Intel Contributions - QAT

- QAT: Intel Quick Assistant Technology
  - Dedicated PCI-e card to offload compression, encryption workload to reduce CPU utilization
- ZFS with QAT
  - Enable data compression
  - Boot I/O performance & reduce disk utilization
- ZFS native encryption is on the way
  - https://github.com/zfsonlinux/zfs/pull/4760



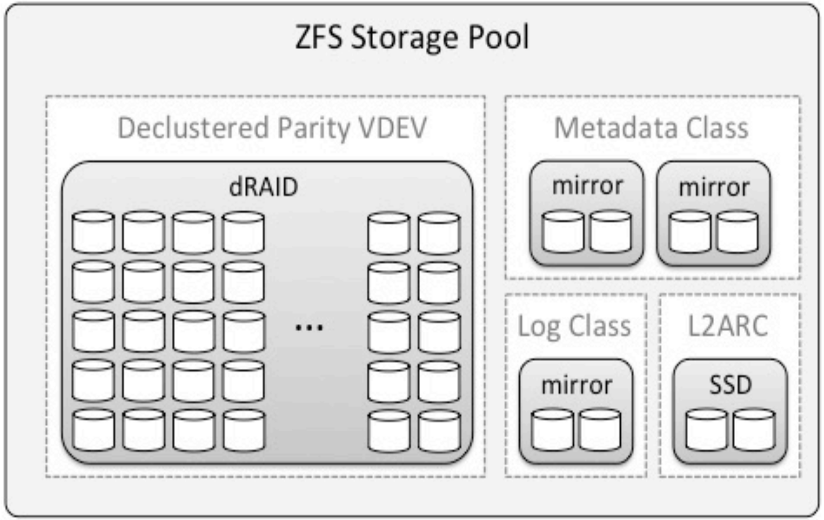
Intel Contributions – QAT Performance

- Test environment
  - CPU: Intel® Xeon® E5-2699 v3 x 2
  - Memory: 128GB
  - Walnut Hill QAT card x 2
  - SSD x 6 contribute to 2.4GB/s I/O bandwidth
- ZFS 0.6.5.6 with data compression enabled
- FIO parameters: libaio, bs=64k, size=2G, numjobs=32, thread=1



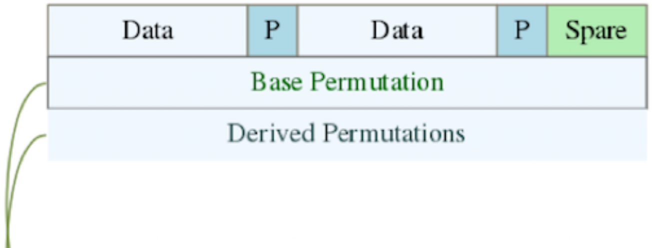
Intel Contributions – Metadata allocation class

- Metadata blocks are with smaller size, and accessed more frequently
- A dedicated VDEV with high IOPS drives to store metadata
  - SSD or NVRAM
  - Mirrored for redundancy
- Better use of SSD than L2ARC



Intel Contributions - DRAID

- Faster rebuild/resilver time
  - Spare blocks are distributed over all disks
  - Short time leads to less risk on data loss
    - 2nd or 3rd disk failure during rebuild time
- Reasonable throughput in degraded mode
  - Lost one disk -> lose 1/N disk bandwidth
- Permutation development based on randomly generated initial permutation
- Rebuild/Resilver is 6 times faster than RAIDZ
  - No block pointer tree traversal



Permutation Group 0

4	3	10	7	2	11	9	1	0	6	5	8
5	4	11	8	3	0	10	2	1	7	6	9
6	5	0	9	4	1	11	3	2	8	7	10
7	6	1	10	5	2	0	4	3	9	8	11
8	7	2	11	6	3	1	5	4	10	9	0
9	8	3	0	7	4	2	6	5	11	10	1
10	9	4	1	8	5	3	7	6	0	11	2
11	10	5	2	9	6	4	8	7	1	0	3
0	11	6	3	10	7	5	9	8	2	1	4
1	0	7	4	11	8	6	10	9	3	2	5
2	1	8	5	0	9	7	11	10	4	3	6
3	2	9	6	1	10	8	0	11	5	4	7

Key Takeaways

- ZFS has great features that can benefit Lustre
  - Lustre snapshot is based on ZFS snapshot
- Lustre on ZFS Performance
  - I/O performance is good, and it can saturate disk bandwidth in my test
  - Metadata performance has great improvement recently
- Intel is contributing to ZFS community
  - Lots of features are being developed at Intel, and they will be open sourced and landed to ZoL
- ZFS and Lustre are both open sourced and free to download

Q&A

谢谢