OPENFABRICS ALLIANCE

14th ANNUAL WORKSHOP 2018

ACCELERATING CEPH WITH RDMA AND NVME-OF

Haodong Tang, Jian Zhang and Fred Zhang

Intel Corporation

{haodong.tang, jian.zhang, fred.zhang}@intel.com

April, 2018

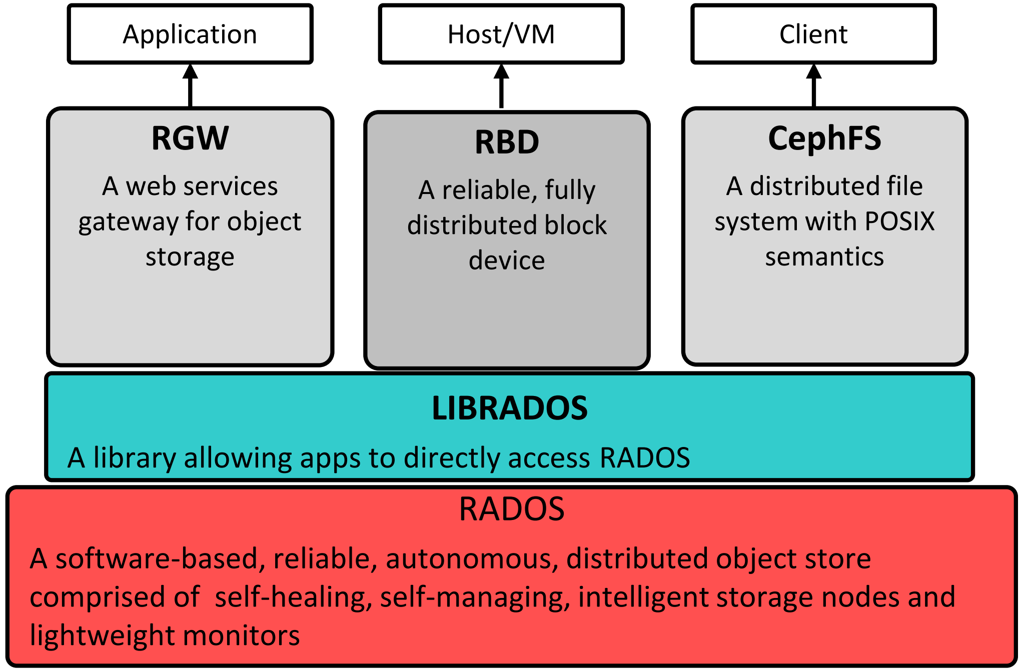
AGENDA

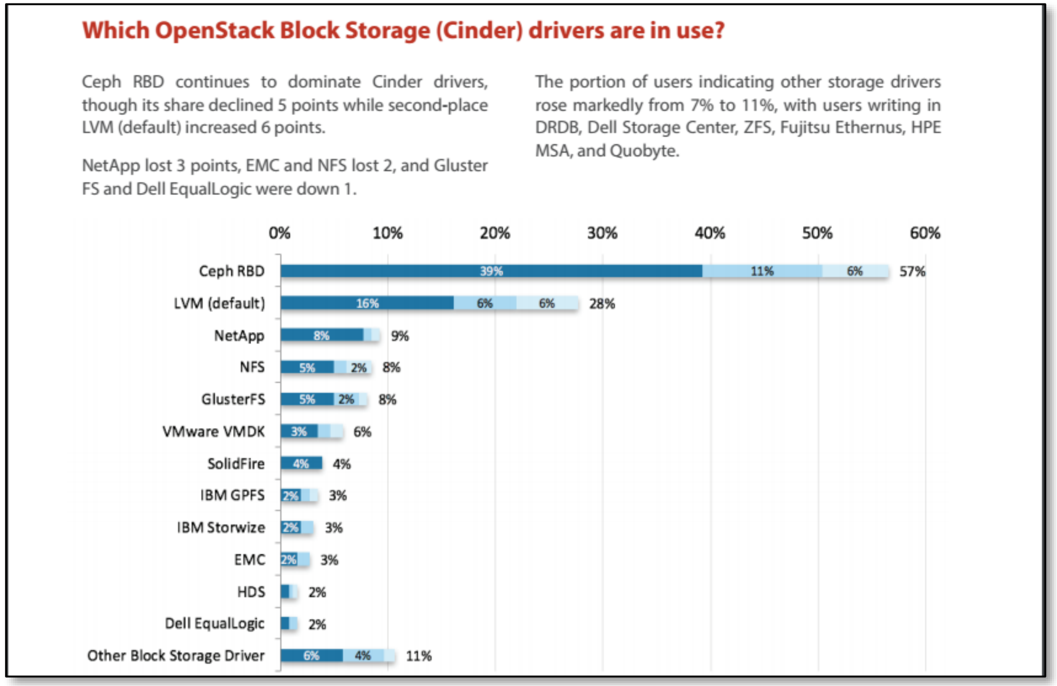
- Background and motivation
- RDMA as Ceph networking component
- RDMA as Ceph NVMe fabrics
- Summary & next step

OPENFABRICS ALLIANCE

BACKGROUND AND MOTIVATION

CEPH INTRODUCTION


Application
Host/VM
Client
RGW
A web services gateway for object storage
RBD
A reliable, fully distributed block device
CephFS
A distributed file system with POSIX semantics
LIBRADOS
A library allowing apps to directly access RADOS
RADOS
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors



Which OpenStack Block Storage (Cinder) drivers are in use?

Ceph RBD continues to dominate Cinder drivers, though its share declined 5 points while second-place LVM (default) increased 6 points.

NetApp lost 3 points, EMC and NFS lost 2, and Gluster FS and Dell EqualLogic were down 1.

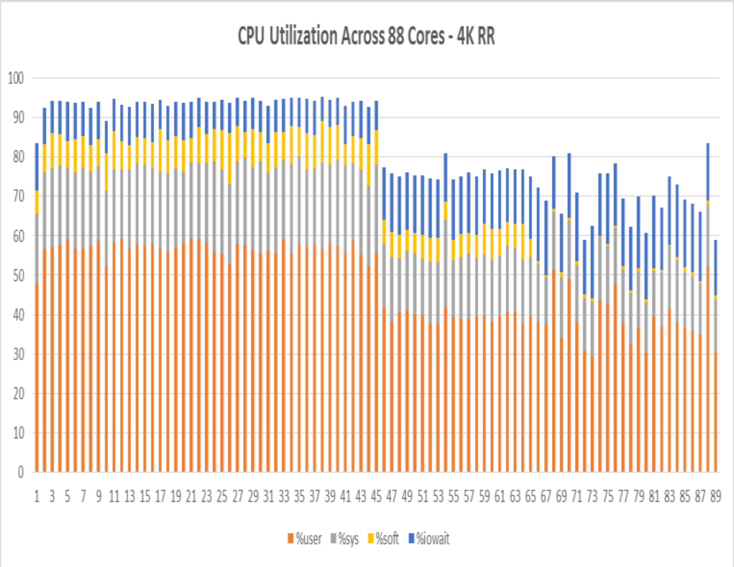
The portion of users indicating other storage drivers rose markedly from 7% to 11%, with users writing in DRDB, Dell Storage Center, ZFS, Fujitsu Ethernus, HPE MSA, and Quobyte.

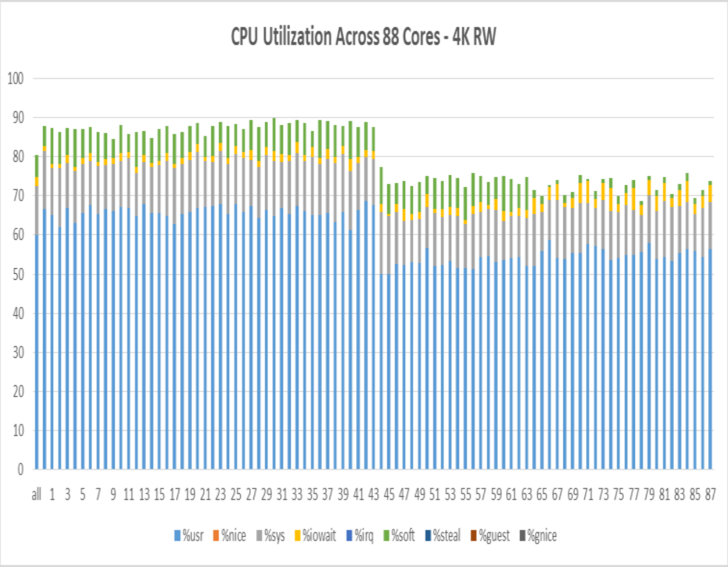
- Open-source, object-based scale-out storage
- Object, Block and File in single unified storage cluster
- Highly durable, available – replication, erasure coding
- Runs on economical commodity hardware
- 10 years of hardening, vibrant community

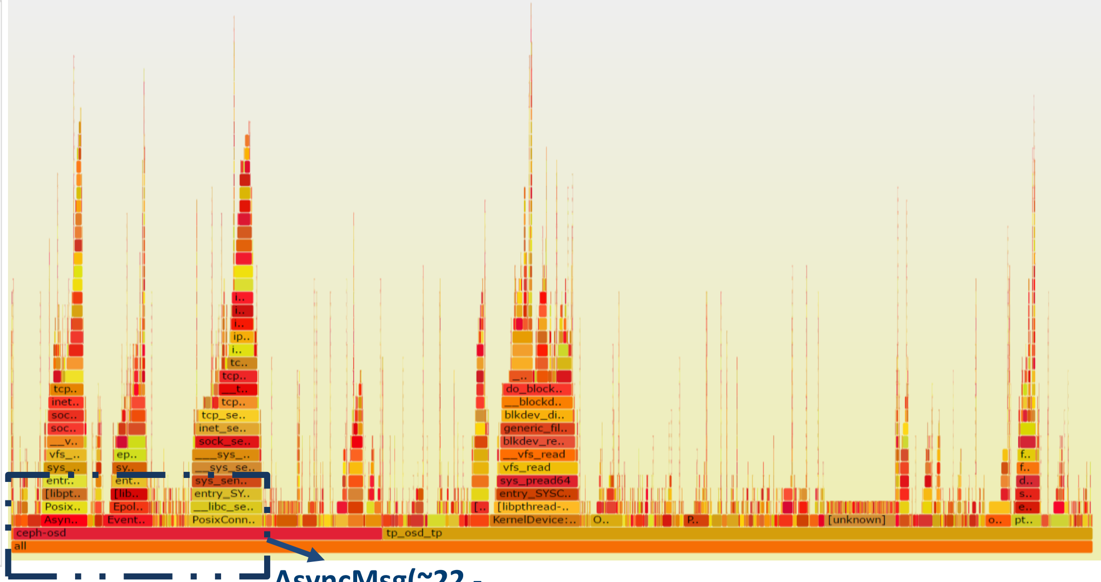
- Scalability – CRUSH data placement, no single POF
- Replicates and re-balances dynamically
- Enterprise features – snapshots, cloning, mirroring
- Most popular block storage for Openstack use cases
- Commercial support from Red Hat

- References: http://ceph.com/ceph-storage, http://thenewstack.io/software-defined-storage-ceph-way,

CEPH PERFORMANCE PROFILING







AsyncMsg(~22 - ~24%)

* This picture is from the Boston OpenStack Summit

- CPU is uneven distributed.
- CPU tend to be the bottleneck for 4K random write and 4K random read.
- Ceph networking layer consumes 20%+ CPU of the totally CPU used by Ceph in 4K random read workload.

MOTIVATION

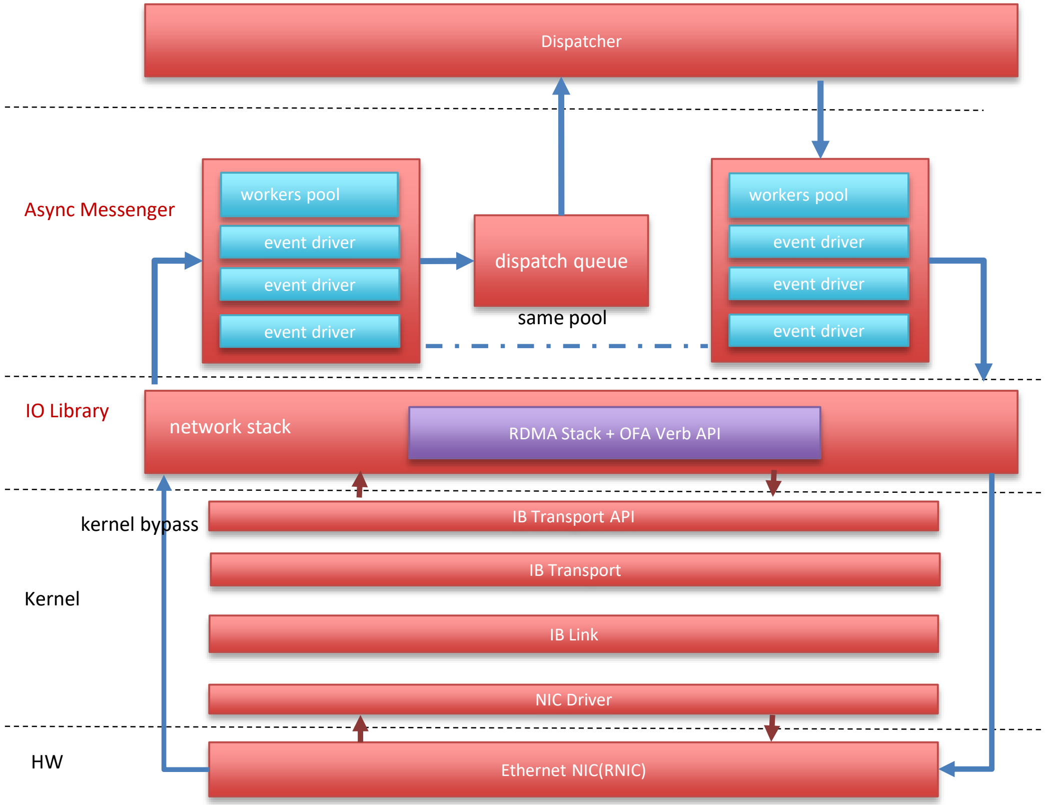
- **RDMA is a direct access from the memory of one computer into that of another without involving either one's operating system.**
- **RDMA supports zero-copy networking(kernel bypass).**
  - Eliminate CPUs, memory or context switches.
  - Reduce latency and enable fast messenger transfer.
- **Potential benefit for ceph.**
  - Better Resource Allocation – Bring additional disk to servers with spare CPU.
  - Lower latency - generated by ceph network stack.

OPENFABRICS ALLIANCE

RDMA AS CEPH NETWORKING COMPONENT

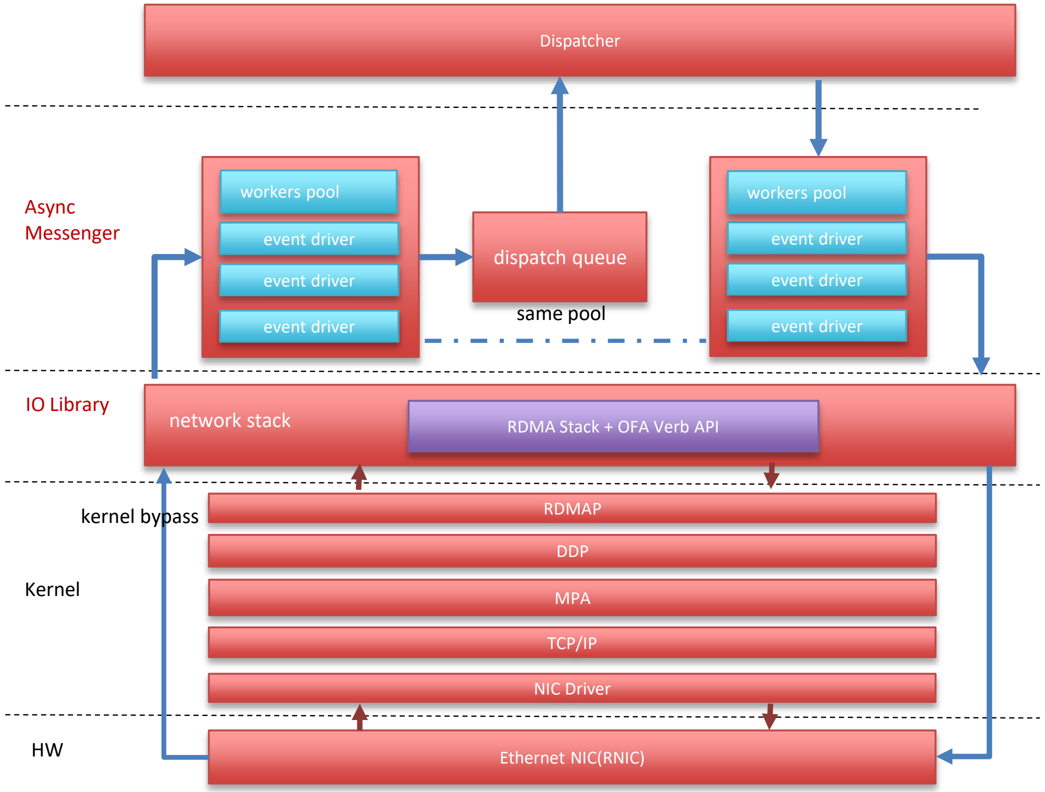
RDMA IN CEPH

- **XIO Messenger.**
  - Based on Accelio, seamlessly supporting RDMA.
  - Scalability issue.
  - Merged to Ceph master three years ago, no support for now.
- **Async Messenger.**
  - Async Messenger is compatible with different network protocol, like Posix, RDMA and DPDK.
  - Current RDMA implementation supports IB protocol.



RDMA OVER ETHERNET

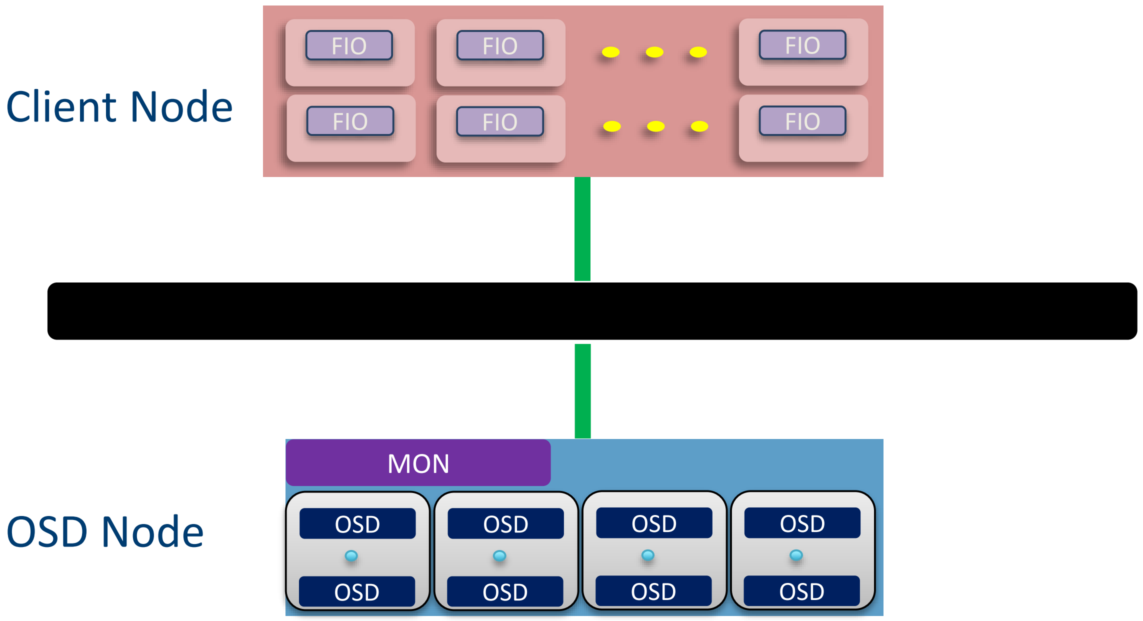
- **Motivation**
  - Leverage RDMA to improve performance (low CPU, low latency).
  - Leverage Intel RDMA NIC to accelerate Ceph.
  - RDMA over Ethernet provide is one of the most convenient and practical way for datacenter running Ceph over TCP/IP.
- **To-do**
  - Need introduce rdma-cm library.



IMPLEMENTATION DETAILS

- **Current implementation for Infiniband in Ceph:**
  - Connection management: Self-implemented TCP/IP based RDMA connection management
  - RDMA verbs: RDMA send, RDMA recv
  - Queue pairs: Shared receive queue (SRQ)
  - Completed Queue: All queue pair share one completed queue
- **iWARP protocol needs:**
  - Connection management: RDMA-CM based RDMA connection management
  - Queue pairs: centralized memory pool for recv queue (RQ)

BENCHMARK METHODOLOGY – SMALL SCALE

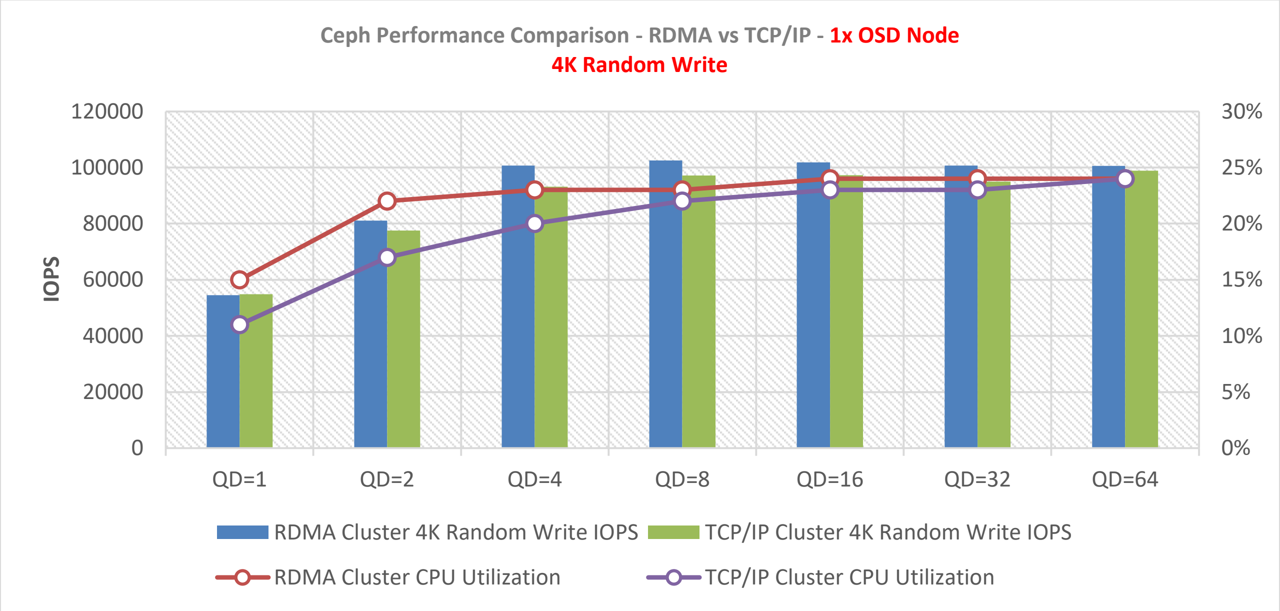


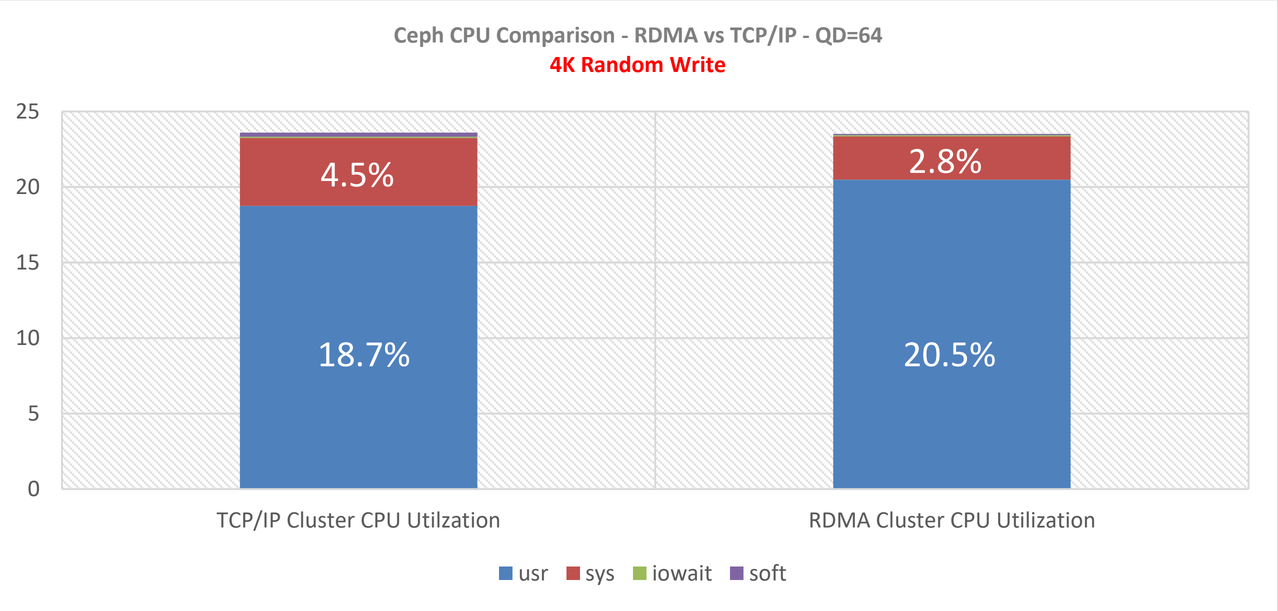
CPU	SKX Platform (112 cores)
Memory	128 GB
NIC	10 GbE Intel® Ethernet Connection X722 with iWARP
Disk distribution	4x P3700 as OSD drive, 1x Optane as DB driver
Software configuration	CentOS 7, Ceph Luminous (dev)
FIO version	2.17

The networking component protocol between OSD node and client node can be changed. We compared the Ceph performance w/ TCP/IP and it w/ RDMA protocol.

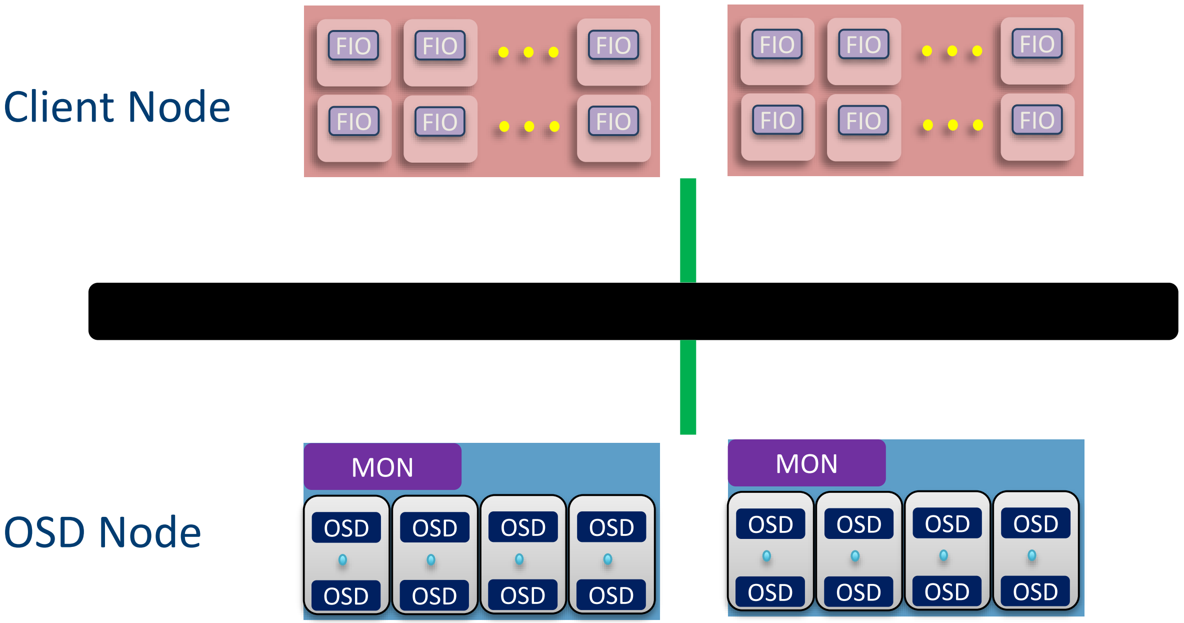
CEPH PERFORMANCE – TCP/IP VS RDMA – 1X OSD NDOE

- **Ceph w/ iWARP delivers higher 4K random write performance than it with TCP/IP.**
- **Ceph w/ iWARP generates higher CPU Utilization.**
  - Ceph w/ iWARP consumes more user level CPU.
  - Ceph w/ TCP/IP consumes more system level CPU.





BENCHMARK METHODOLOGY – LARGER SCALE

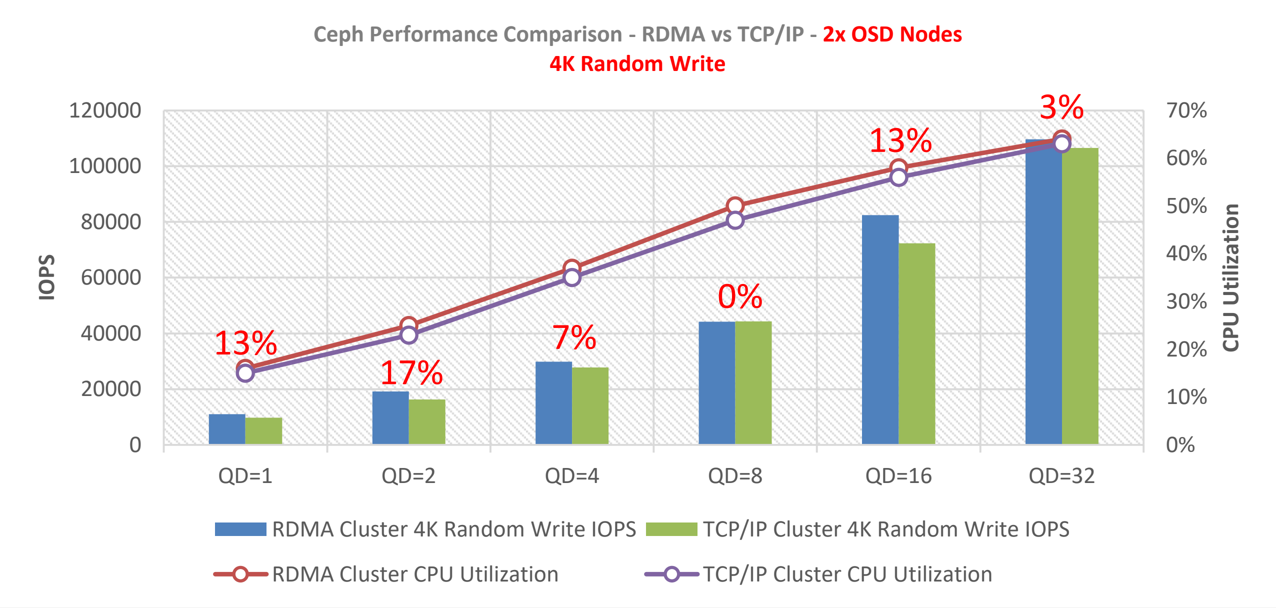


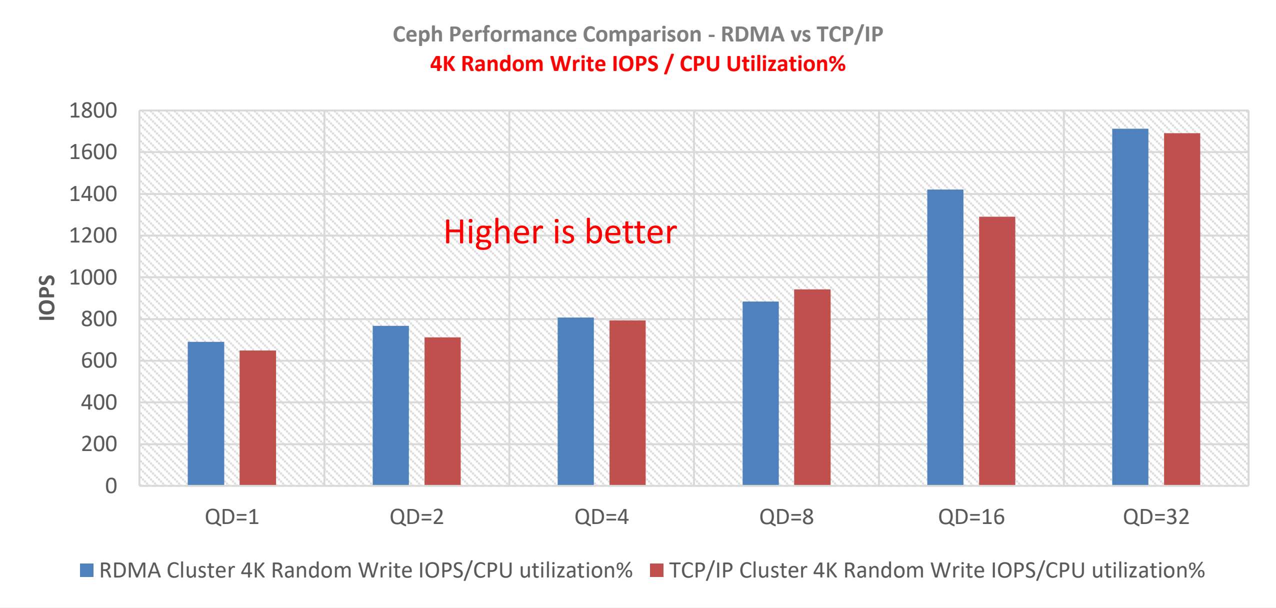
CPU	SKX Platform (72 cores)
Memory	128 GB
NIC	10 GbE Intel® Ethernet Connection X722 with iWARP
Disk distribution	4x P4500 as OSD/DB drive
Software configuration	Ubuntu 17.10, Ceph Luminous (dev)
FIO version	2.12

We scale the OSD node to verify the RDMA protocol scale-out ability.

CEPH PERFORMANCE – TCP/IP VS RDMA – 2X OSD NODES

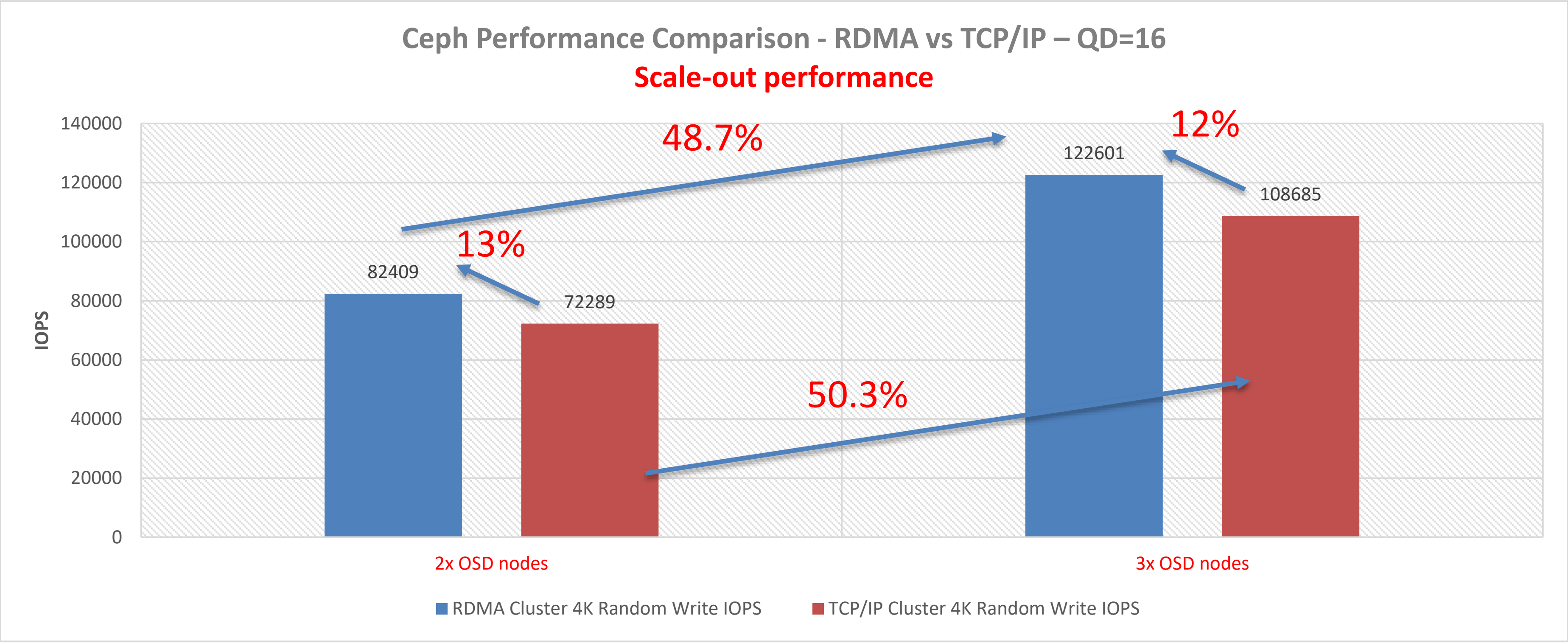
- Ceph w/ iWARP delivers up to 17% 4K random write performance benefit than it w/ TCP/IP.
- Ceph w/ iWARP is more CPU efficient.





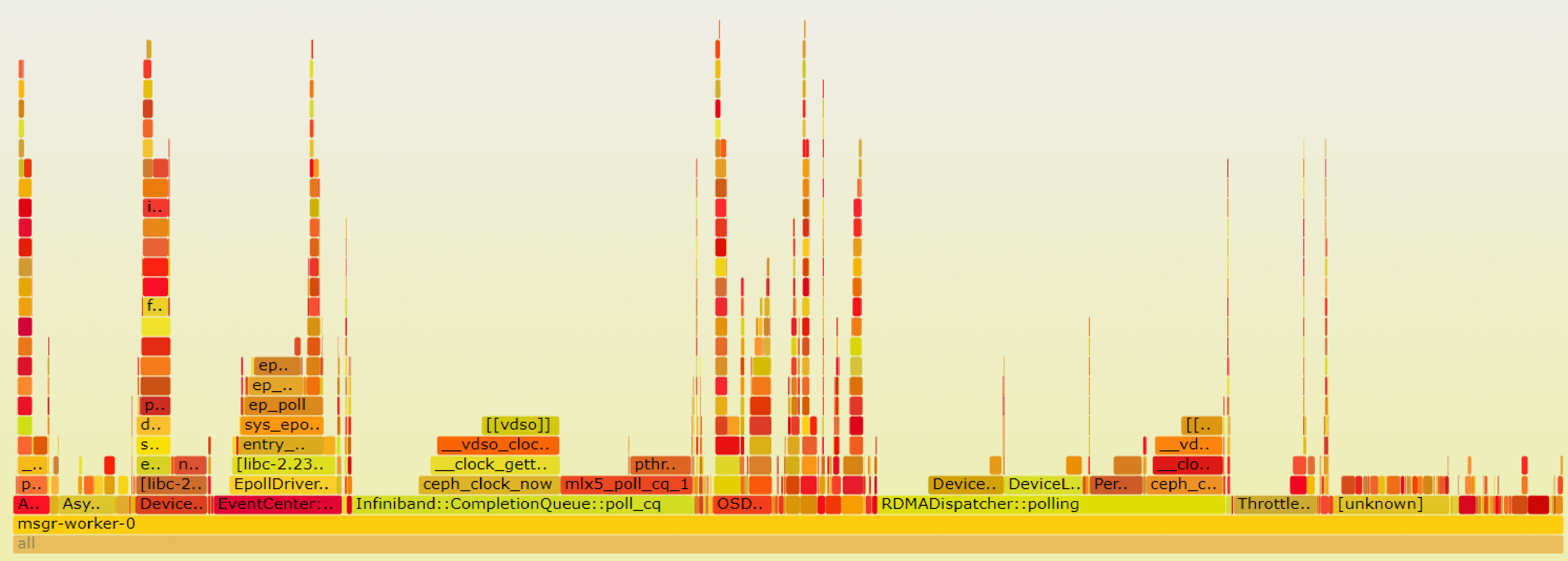
CEPH PERFORMANCE – TCP/IP VS RDMA – 3X OSD NODES

- **Ceph node scaling out: RDMA vs TCP/IP - 48.7% vs 50.3% ➔ scale out well.**
- **When QD is 16, Ceph w/ RDMA shows 12% higher 4K random write performance.**



PERFORMANCE RESULT DEEP ANALYSIS

CPU Profiling

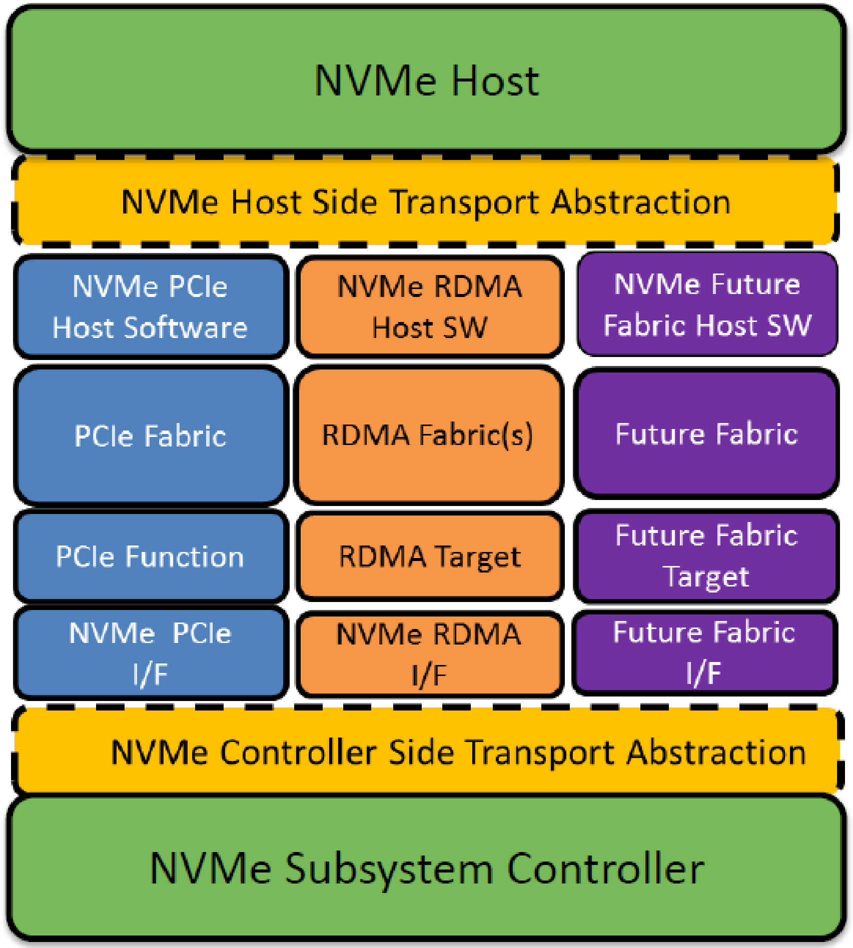


- Two polling thread: Ceph Epoll based Async driver thread + RDMA polling thread.
- Not really zero-copy: there’s one copy from RDMA recv buffer to Ceph Async driver buffer.

RDMA AS CEPH NVME FABRICS

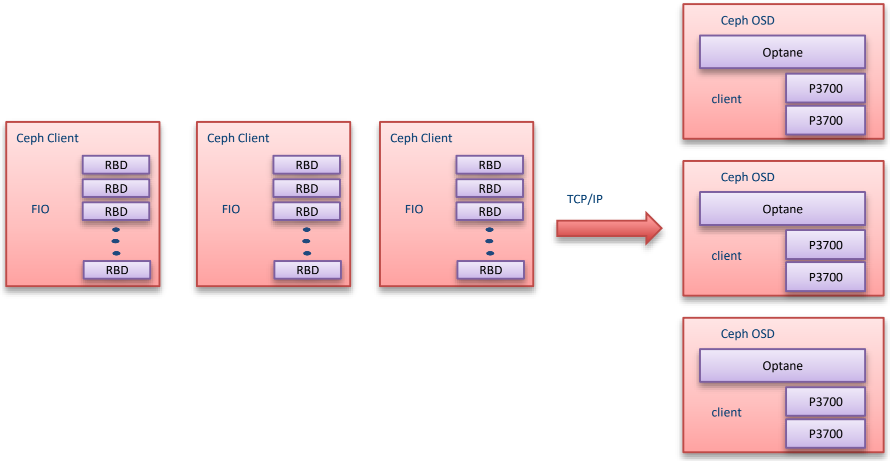
RDMA AS CEPH NVME FABRICS

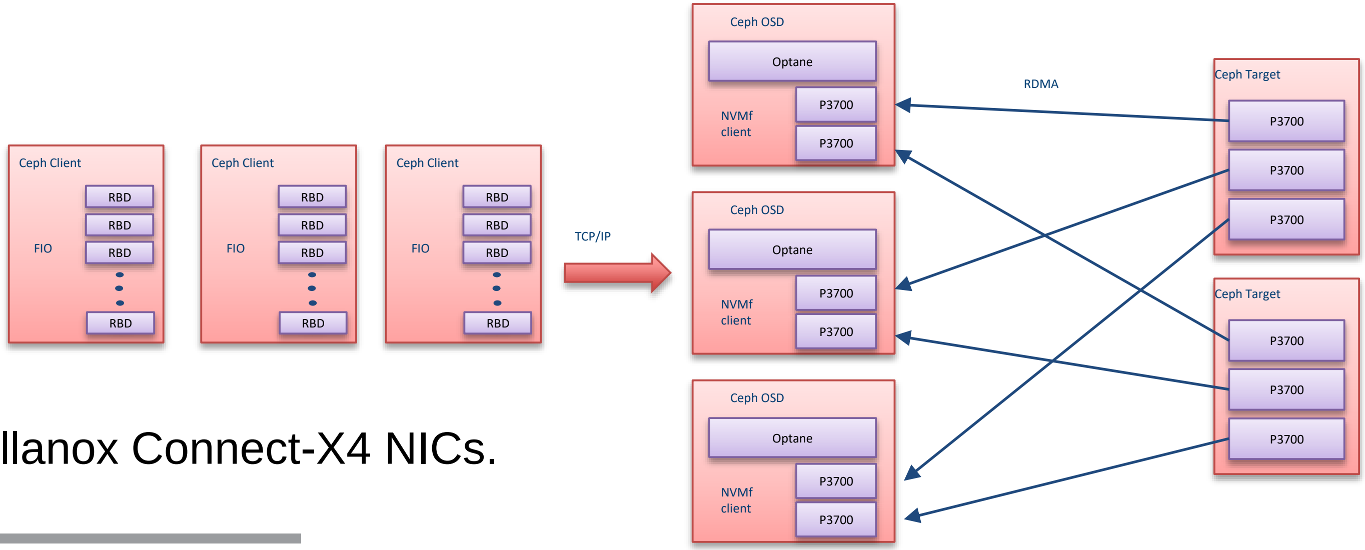
- **NVMe is a new specification optimized for NAND flash and next-generation solid-state storage technologies.**
- **NVMe over Fabrics enables access to remote NVMe devices over multiple network fabrics.**
  - Supported fabrics
    - RDMA – InfiniBand, IWARP, RoCE
    - Fiber Channel
    - TCP/IP
- **NVMe-oF benefits**
  - NVMe disaggregation.
  - Delivers performance of remote NVMe on-par with local NVMe.



RDMA AS CEPH NVME FABRICS

- **Baseline and comparison**
  - The baseline setup used local NVMe.
  - The comparison setup attaches remote NVMe as OSD data drive.
    - 6x 2T P3700 are among 2x Storage nodes.
    - OSD nodes attach the 6x P3700 over RoCE V2 fabric.
    - Set NVMe-oF CPU offload on target node.
- **Hardware configuration**
  - 2x Storage nodes, 3x OSD nodes, 3x Client nodes.
  - 6x P3700 (800 GB U.2), 3x Optane (375 GB)
  - 30x FIO processes worked on 30x RBD volumes.
  - All these 8x servers are BRW, 128 GB memory, Mellanox Connect-X4 NICs.





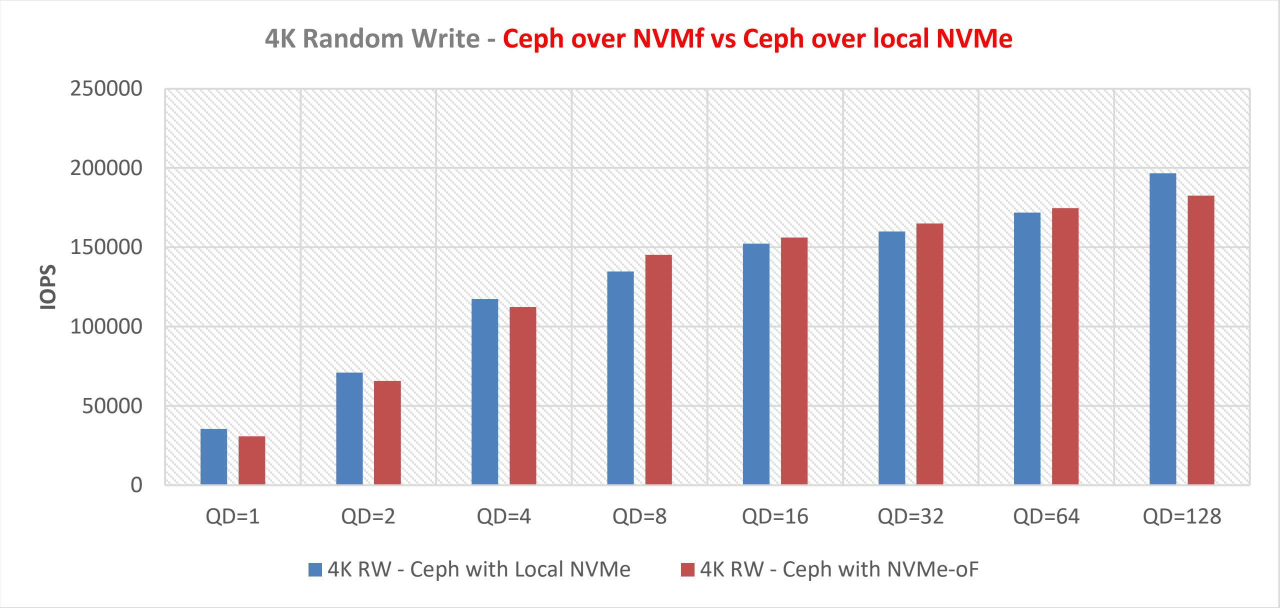
EXPECTATION BEFORE POC

- **Expectations and questions before POC.**
  - Expectations: According to the benchmark from the first part, we’re expecting
    - on-par 4K random write performance.
    - on-par CPU utilization on NVMe-oF host node.
  - Questions:
    - How many CPU will be used on NVMe-oF target node ?
    - How is the behavior of tail latency(99.0%) latency with NVMe-oF ?
    - Does NVMe-oF influence the Scale-out ability of Ceph ?

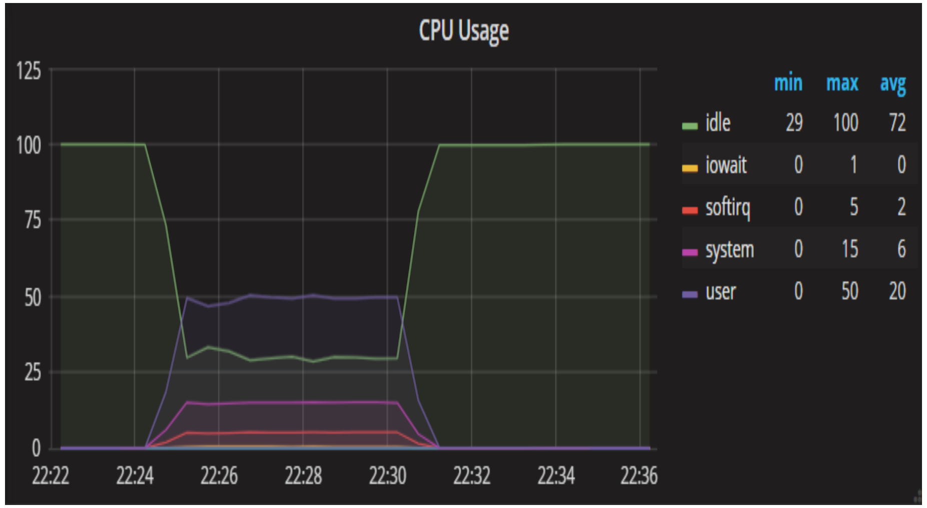
RDMA AS CEPH NVME FABRICS

Client side performance comparison

- **On-par 4K random write performance**
- **Running Ceph with NVMe-oF brings <1% CPU overhead on target node.**
- **CPU is not the bottleneck on the host node.**

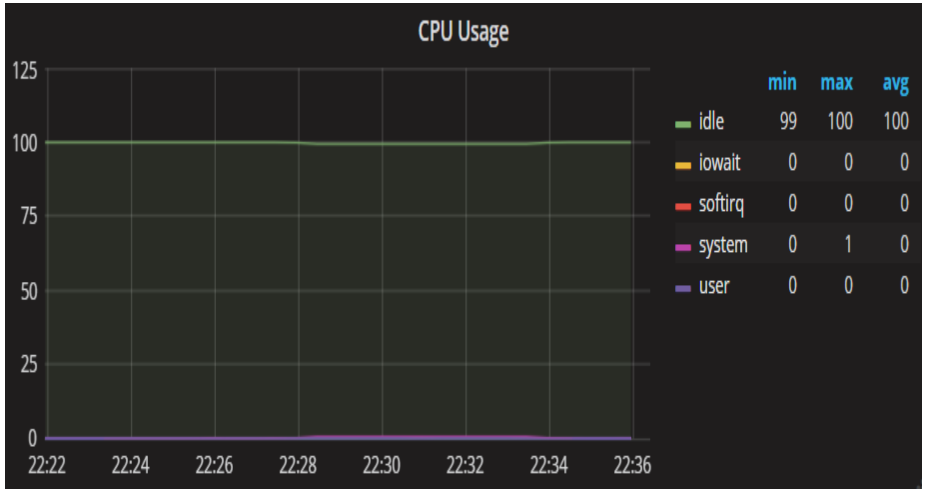


CPU Utilization on OSD Node



	min	max	avg
idle	29	100	72
iowait	0	1	0
softirq	0	5	2
system	0	15	6
user	0	50	20

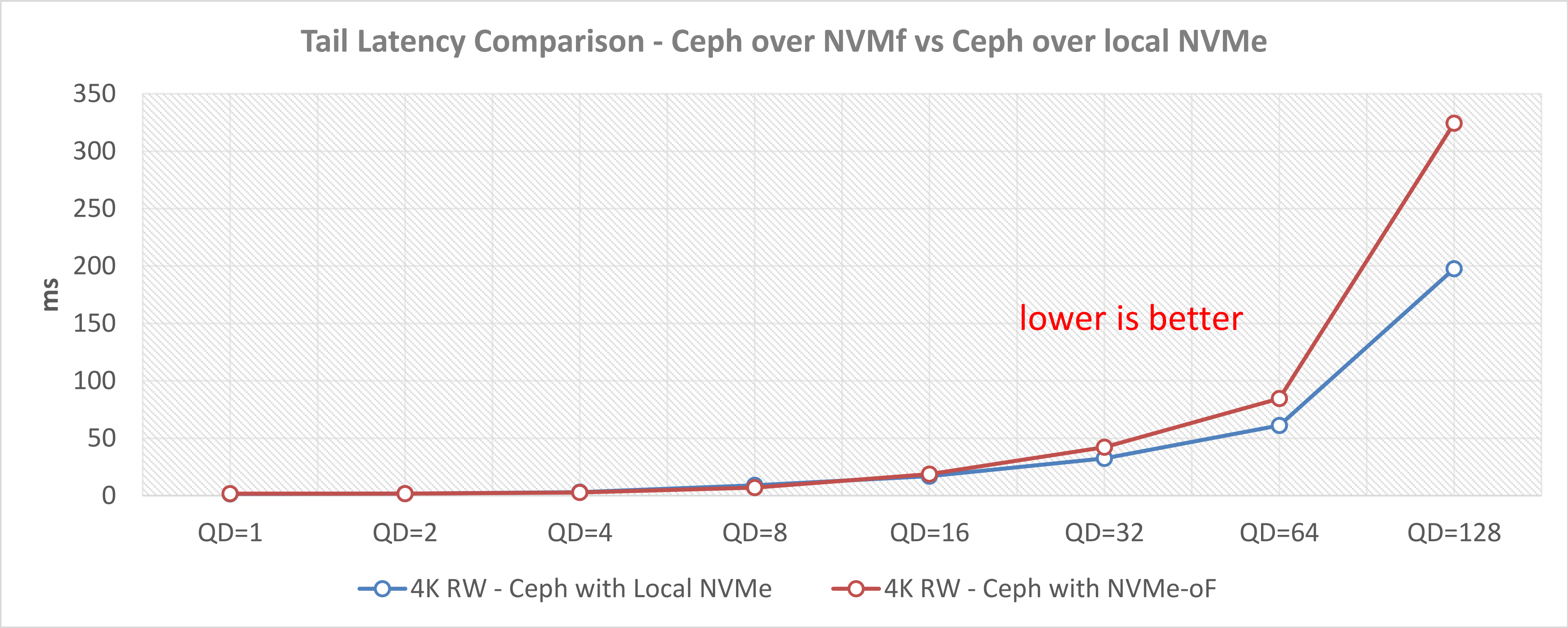
CPU Utilization on Target Node



	min	max	avg
idle	99	100	100
iowait	0	0	0
softirq	0	0	0
system	0	1	0
user	0	0	0

CEPH TAIL LATENCY

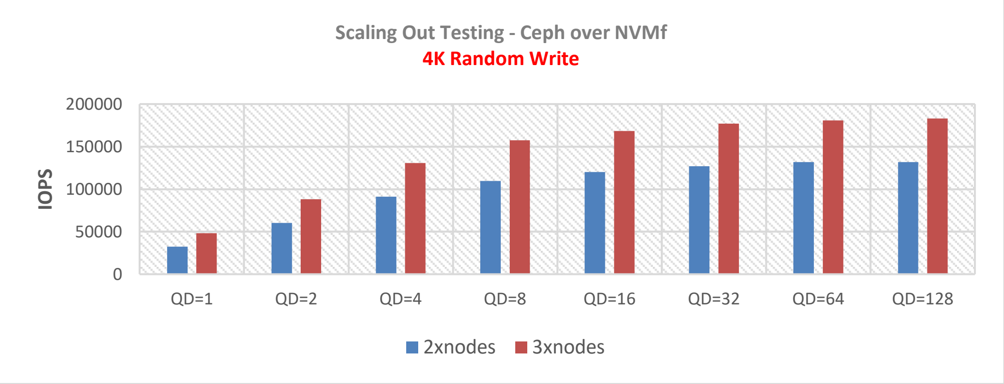
- **When QD is higher than 16, Ceph with NVMe-oF shows higher tail latency (99%).**
- **When QD is lower than 16, Ceph with NVMe-oF on-par with Ceph over local NVMe.**

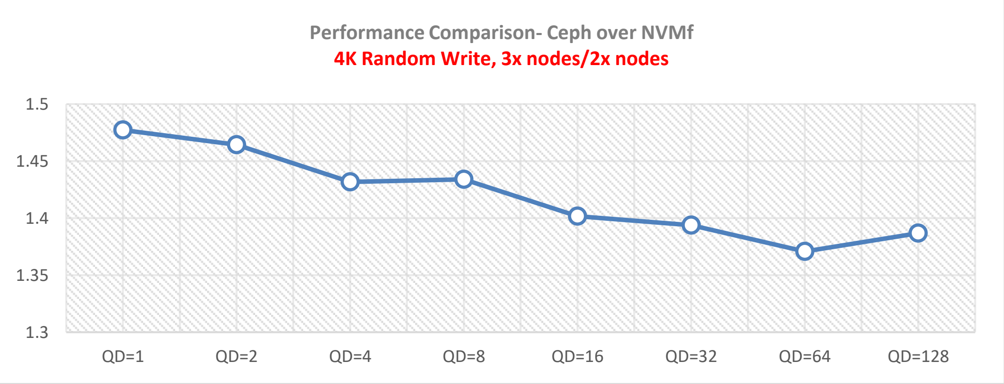


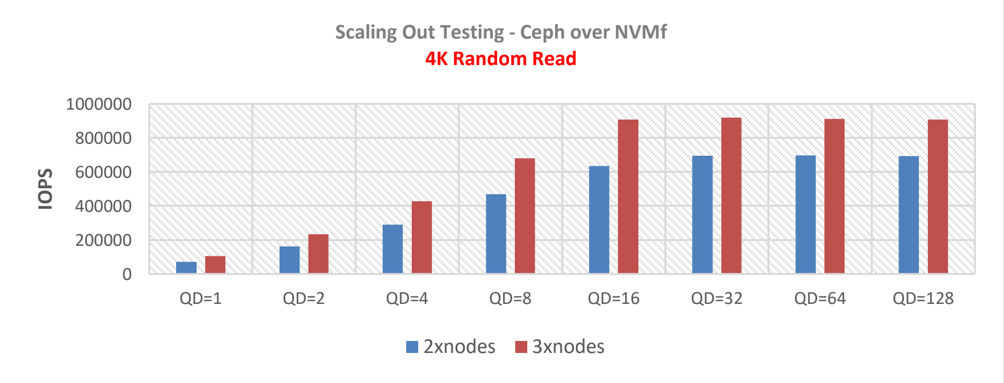
RDMA AS CEPH NVME FABRICS

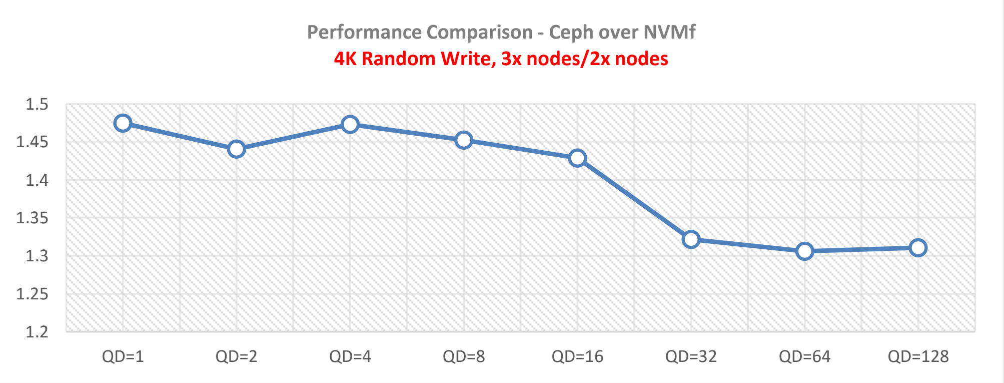
Scaling out performance

- **Running Ceph over NVMe-oF didn’t limit the Ceph OSD node scaling out.**
  - For 4K random write/read, the maximum ratio of 3x nodes to 2x nodes is 1.47, closing to 1.5 (ideal value).


Scaling Out Testing - Ceph over NVMf
4K Random Write


Performance Comparison- Ceph over NVMf
4K Random Write, 3x nodes/2x nodes


Scaling Out Testing - Ceph over NVMf
4K Random Read


Performance Comparison - Ceph over NVMf
4K Random Write, 3x nodes/2x nodes

SUMMARY

SUMMARY & NEXT-STEP

- **Summary**
  - RDMA is critical for future Ceph AFA solutions.
    - Ceph with RDMA messenger provides up to ~17% performance advantage over TCP/IP.
    - Ceph with RDMA messenger shows great scale-our ability.
  - As network fabrics, RDMA performs well in Ceph NVMe-oF solutions.
    - Running Ceph on NVMe-oF does not appreciably degrade Ceph write performance.
    - Ceph with NVMe-oF brings more flexible provisioning and lower TCO.
- **Next-step**
  - Ceph RDMA networking component optimization based on previous analysis.
  - leverage NVMe-oF with the high density storage node for lower TCO.

LEGAL DISCLAIMER & OPTIMIZATION NOTICE

- **Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.**
- **INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.**
-

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

OPENFABRICS ALLIANCE

14th ANNUAL WORKSHOP 2018

THANK YOU!

Haodong Tang, Jian Zhang and Fred Zhang

Intel Corporation

{haodong.tang, jian.zhang, fred.zhang}@intel.com

April, 2018

intel®