CEPHALOCON APAC 2018
THE FUTURE OF STORAGE
22-23 March 2018 | BEIJING

Accelerating Ceph with RDMA

Haodong Tang,

Brien Porter,

March, 2018

Agenda

- Background and motivations
- Ceph with RDMA Messenger
- Ceph with NVMe-oF
- Summary & Next-step

ceph

CEPH

CEPHALOCON APAC 2018

THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

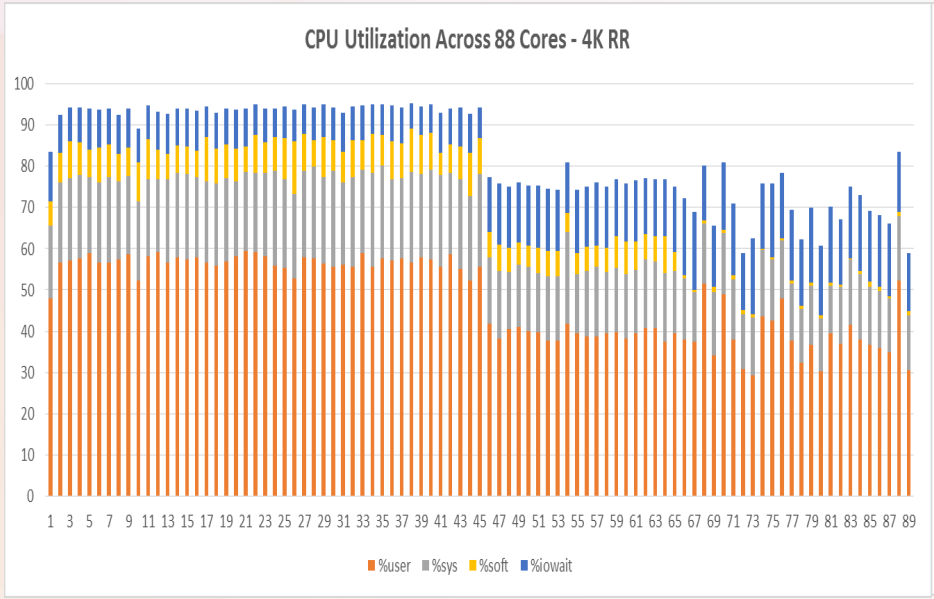
Background and motivations

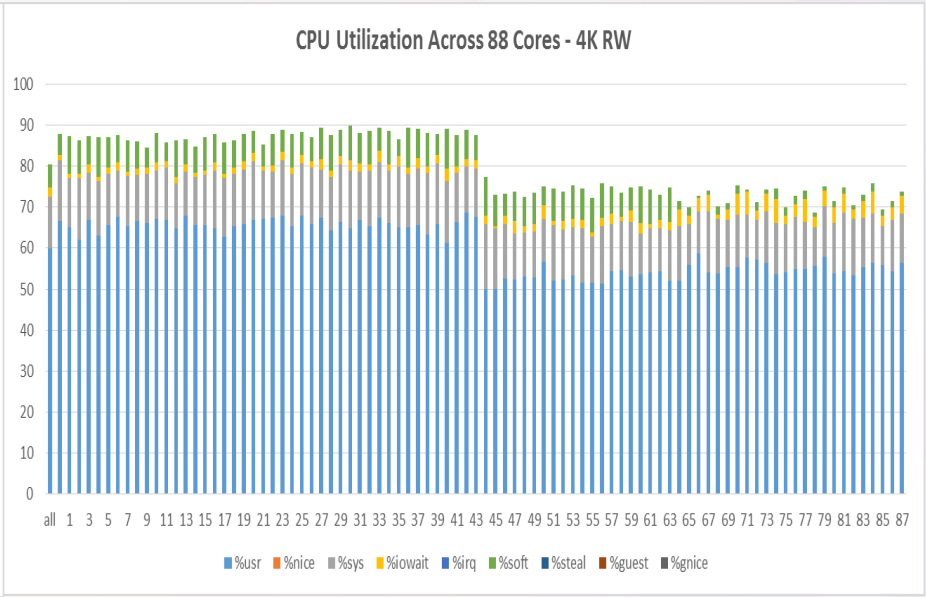
Background

- Past work (Boston OpenStack summit)
  - Optane based all-flash array was capable of delivering over 2.8M 4K random read IOPS with very low latency.
  - Optane as BlueStore DB drive dramatically improved the tail latency of 4K random write.
- In this session, we're talking about…
  - With the emergence of faster storage device (Optane/AEP), we need faster network stack (high BW, low CPU cost, low latency) to keep performance linear growth.

Uneven CPU distribution

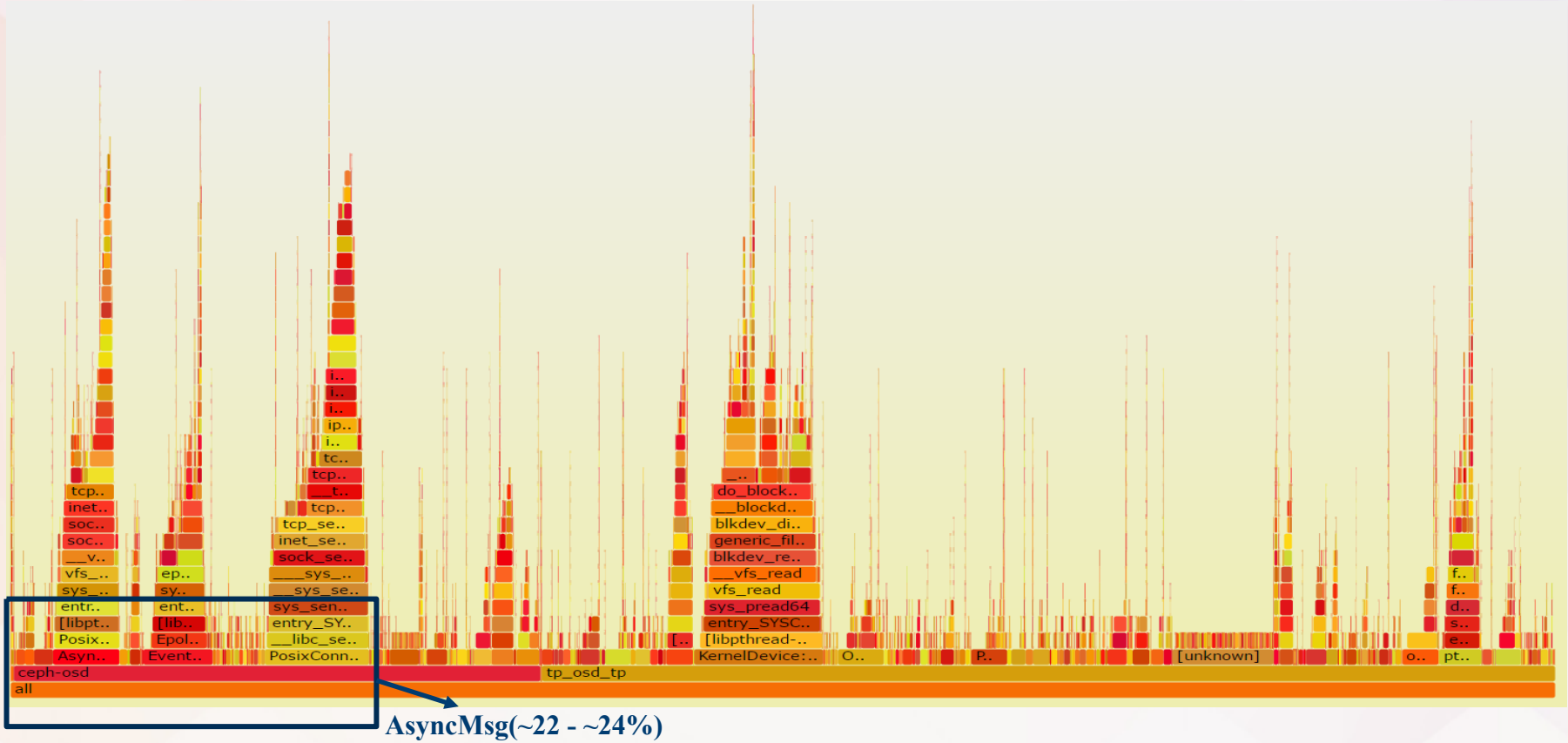
* This picture is from the Boston OpenS Summit



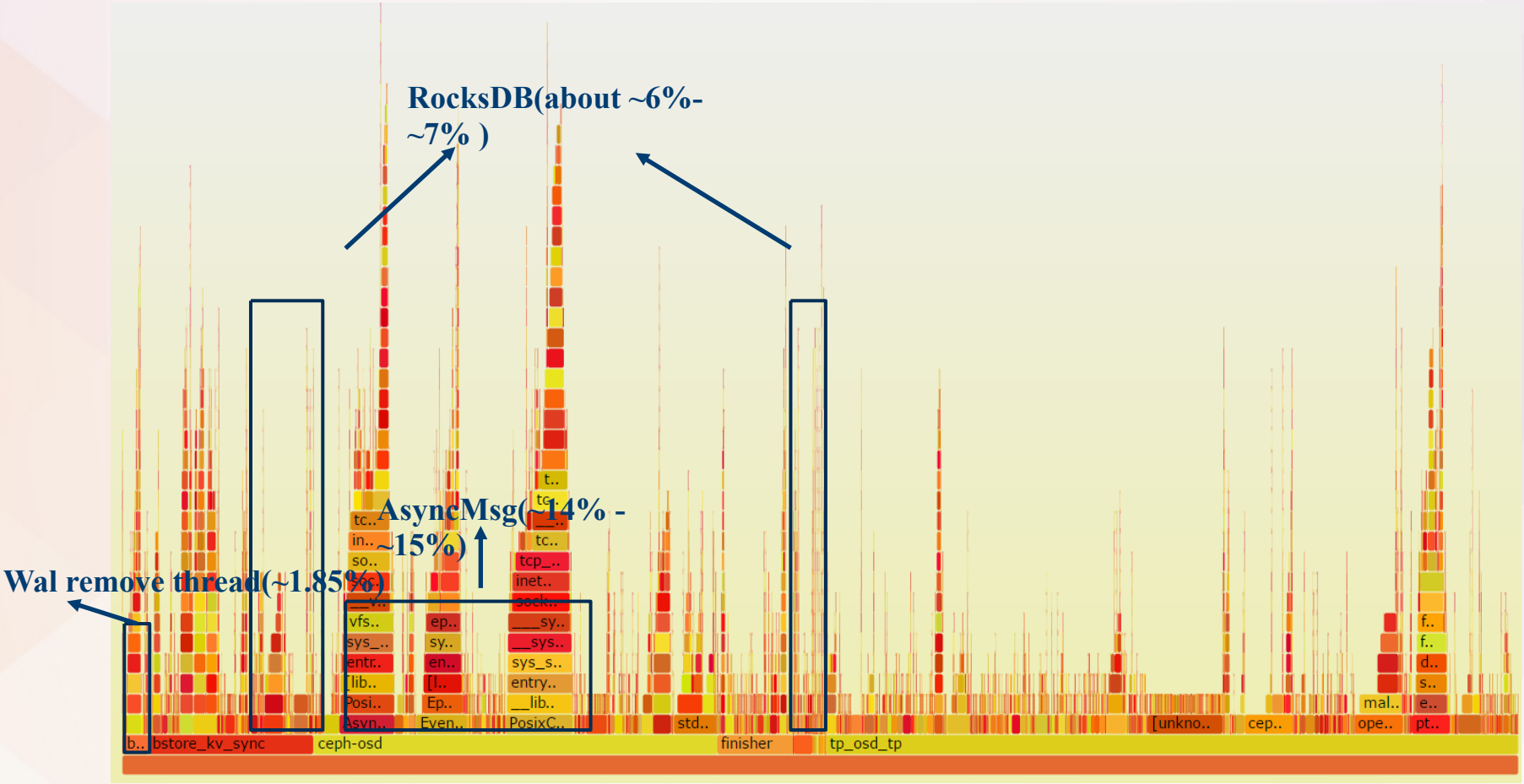


- Unbalanced CPU utilization.
- With single OSD per NVMe, Ceph can't take full advantage of NVMe performance. → to minimize latency of 4K RW.
- With multiple OSDs per NVMe, greatly improves 4K RW performance , but CPU tends to be the bottleneck. → to reduce CPU utilization.

CPU overhead – 4K random read



CPU overhead – 4K random write



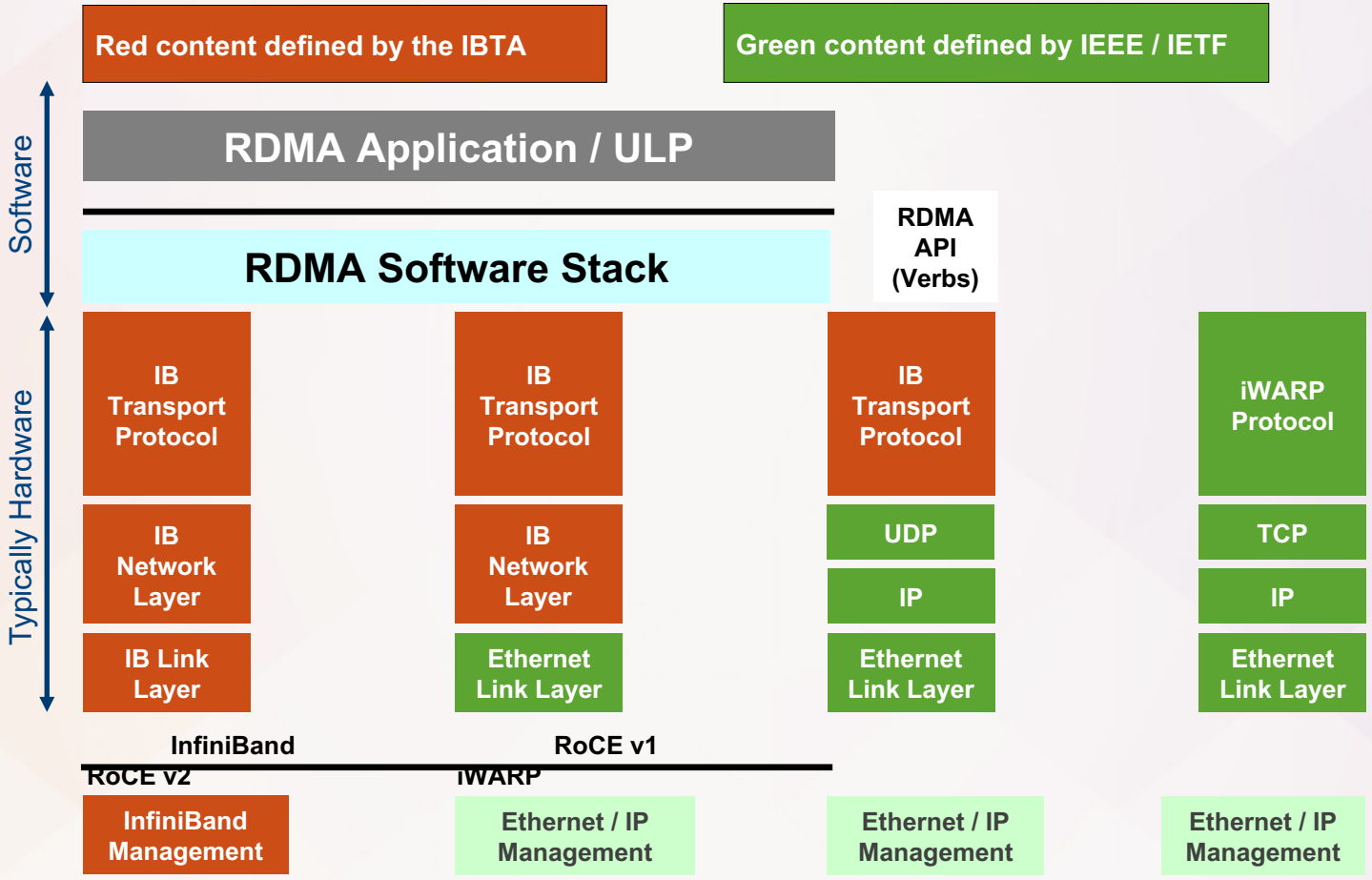
Motivations

- 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.
  - Reduce latency generated by ceph network stack.

CEPHALOCON APAC 2018
THE FUTURE OF STORAGE
22-23 March 2018 | BEIJING

Ceph with RDMA Messenger

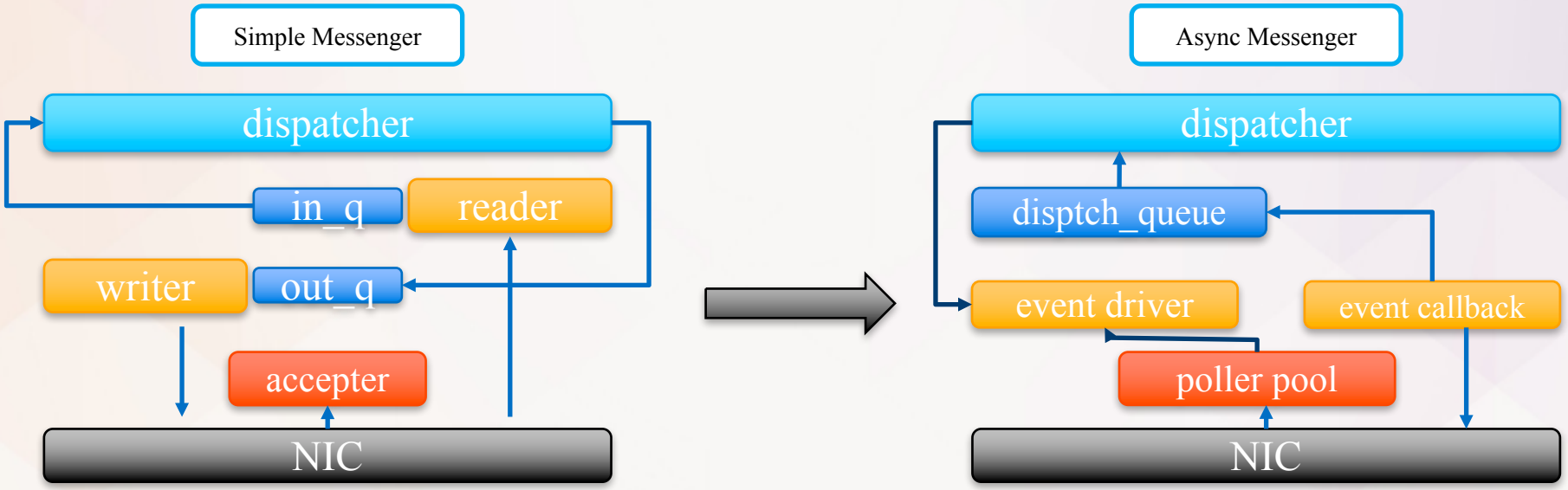
RDMA overview



RoCE portion adopted from “Supplement to InfiniBand Architecture Specification Volume 1 Release 1.2.1, Annex A17: RoCEv2”, September 2, 2014

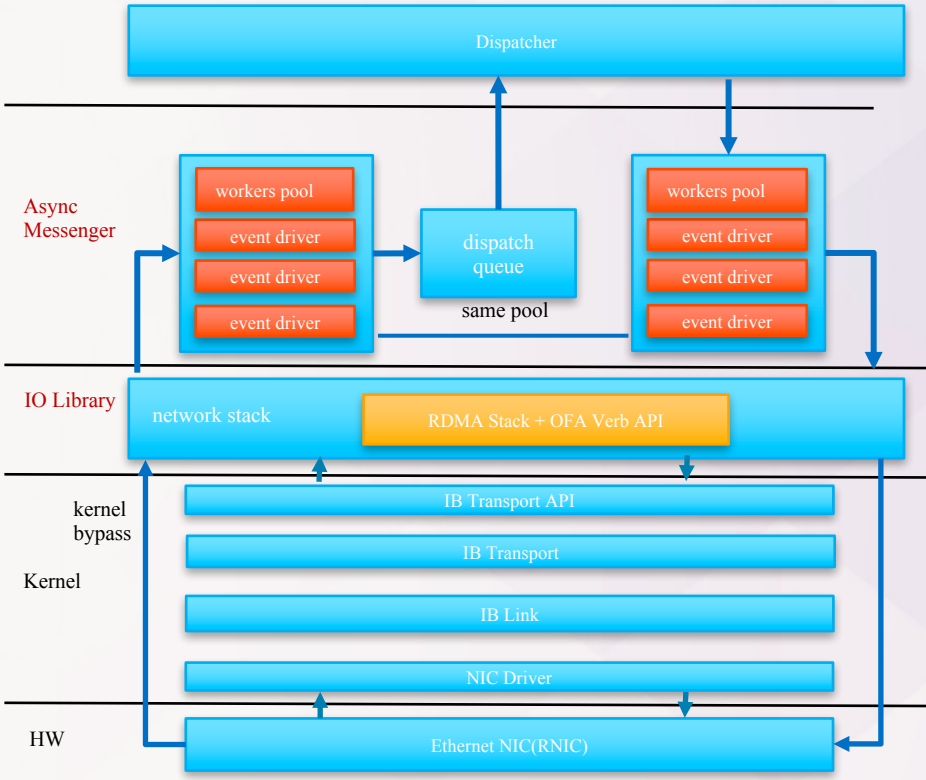
Ceph network layer

- Currently, the Async Messenger is the default network stack, which is subject to the message transfer between different dispatcher (client, monitor, OSD daemon).
- By test, Async Messenger bring ~8% CPU benefit (lower than Simple Messenger) , no throughput gain found.



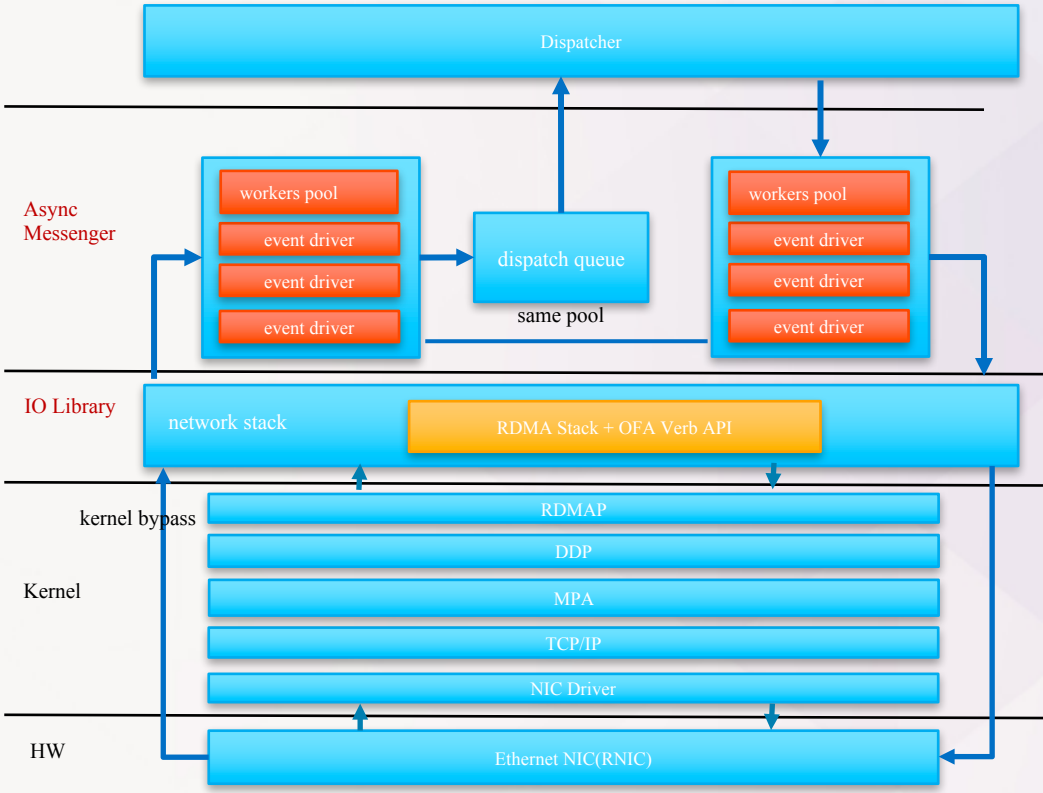
Ceph with RDMA

- XIO Messenger is based on Accelio, which seamlessly support RDMA. XIO Messenger was implemented in Ceph Hammer Release as Beta. **No support for now**.
- Async Messenger.
  - Async Messenger is compatible with different network protocol, like Posix, **RDMA** and DPDK.
  - Current Async Messenger RDMA support is based on IB protocol.
  - How about to integrate iwarp protocol ?



Ceph IWARP support

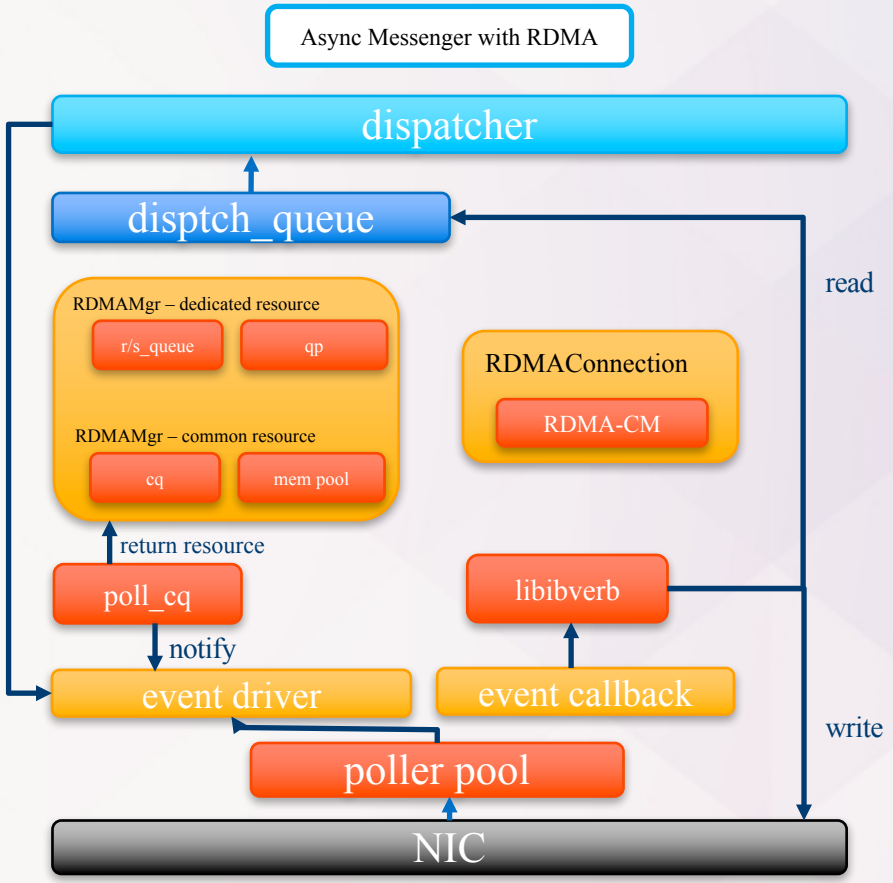
- Motivation
  - Leverage RDMA to improve performance (low CPU utilization, low latency) and improve drive scalability.
  - Leverage Intel network technology (NIC with IWARP support) to speed up Ceph.
- Prerequisite
  - Ceph AsyncMessenger provide asynchronous semantics for RDMA.
- To-do
  - **Need rdma-cm library.**



Ceph IWARP integration

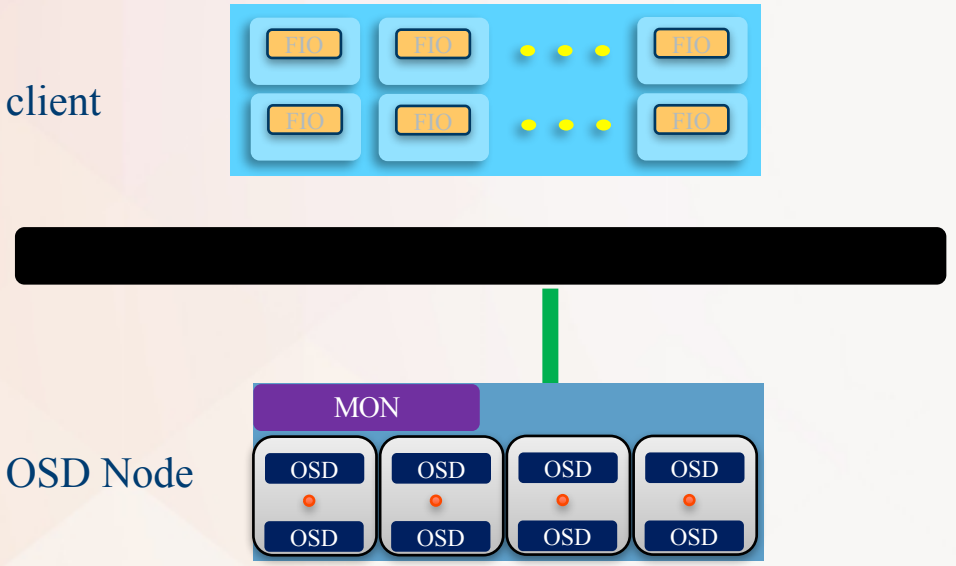
- Code link: iwarp enabling code

- Implementation
  - Every RDMA connection owns dedicated qp and recv/send queue.
  - All RDMA connection own common cq and memory pool.
  - One cq polling thread to get completed queue.
  - Use epoll to notify waiting event.



Ceph w/ IWARP performance

- Test configuration

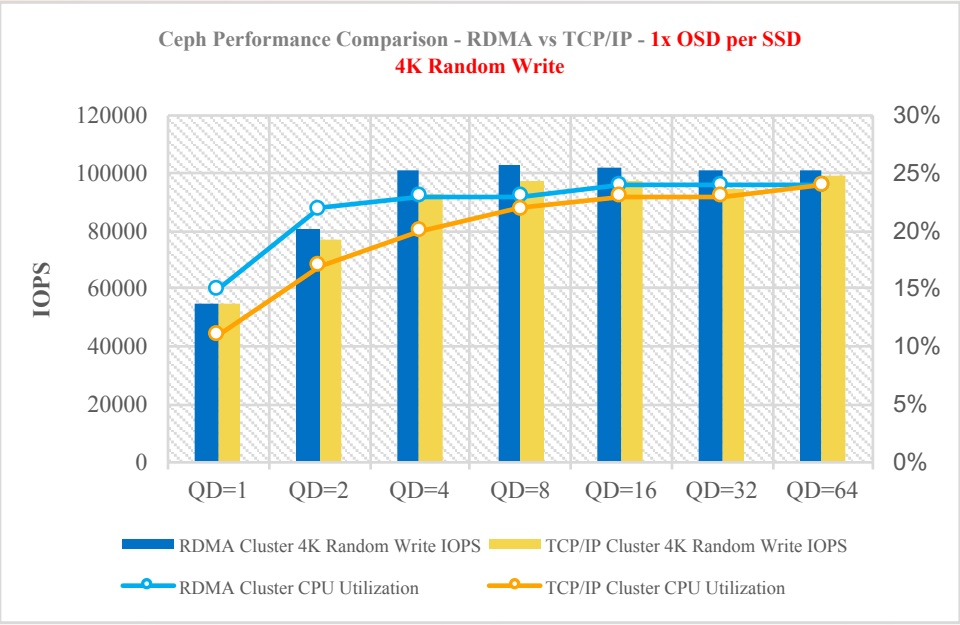


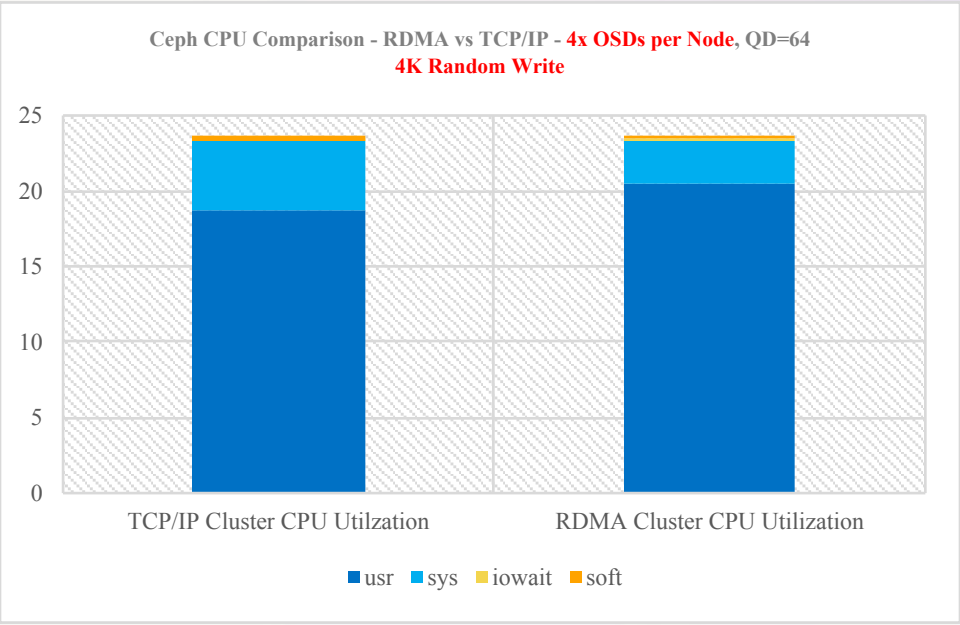
- Test Methodology
  - QD scaling: 1->128

CPU	SKX Platform
Memory	128 GB
NIC	10 Gb X722 NIC
Disk distribution	4x P3700 as OSD drive, 1x Optane as DB driver
Software configuration	CentOS 7, Ceph Luminous (dev)

Ceph w/ IWARP performance

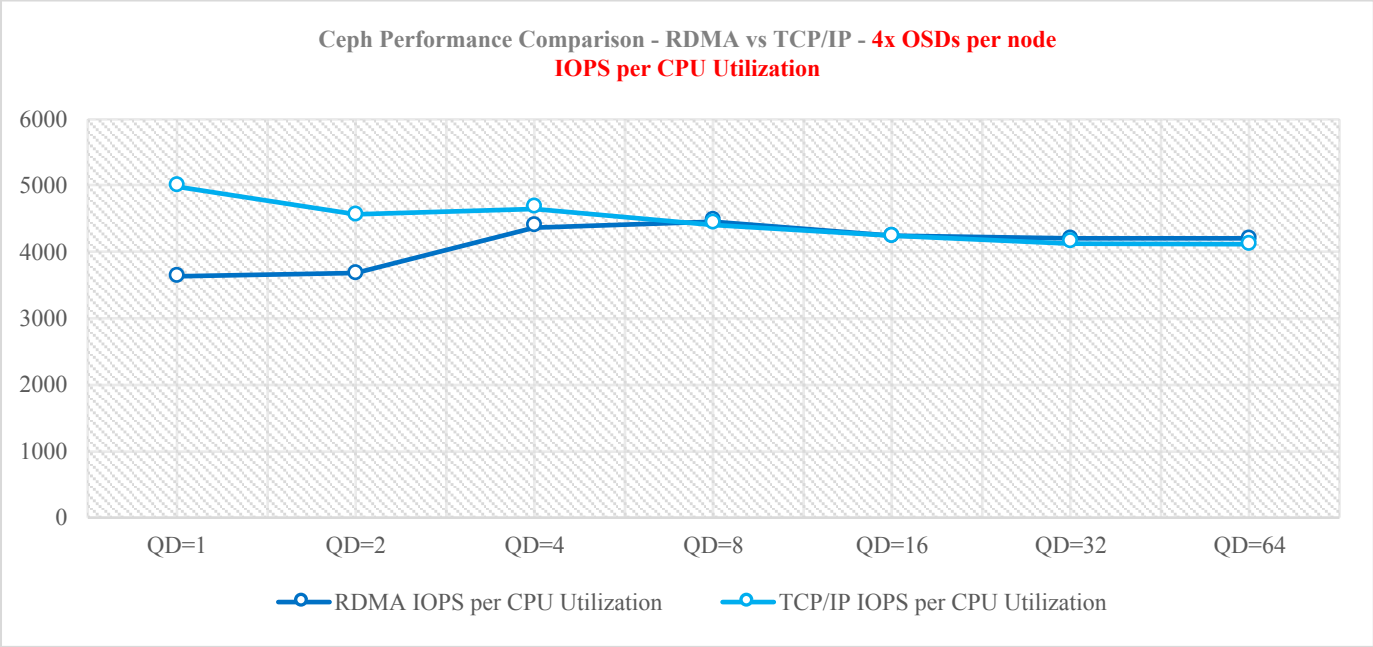
- Ceph w/ IWARP delivers higher 4K random write performance than TCP/IP.
- Ceph w/ IWARP generates higher CPU Utilization.
  - Ceph w/ IWARP consume more user level CPU, while Ceph w/ TCP/IP consumes more system level CPU.





Ceph w/ IWARP performance

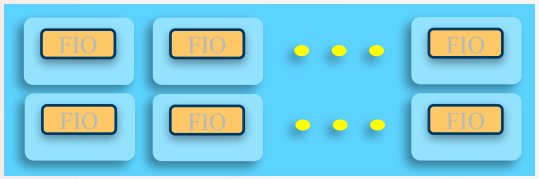
- With QD scaling up, the 4K random write IOPS per CPU utilization of Ceph w/ IWARP is catching up Ceph with TCP/IP.

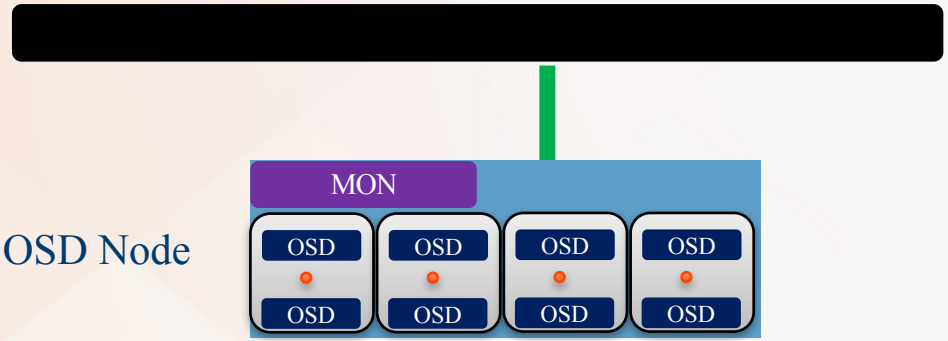


Ceph w/ RoCE performance

- Test configuration

client





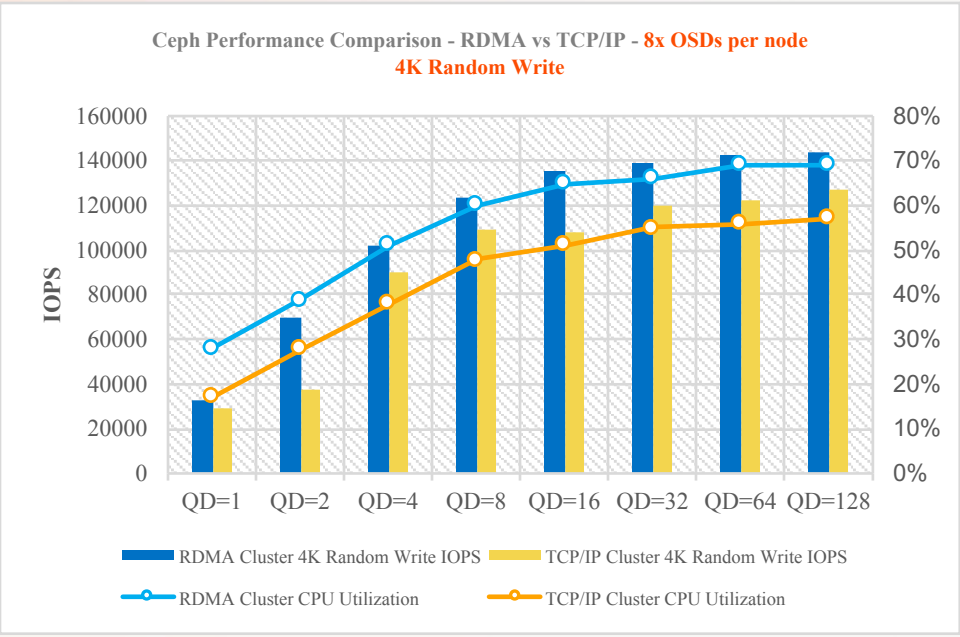
OSD Node

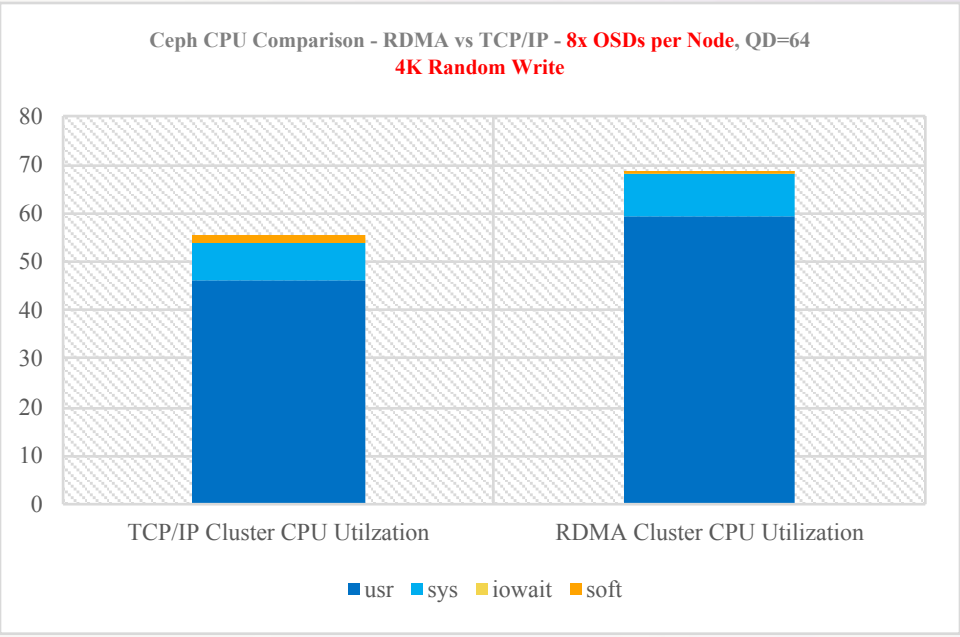
- Test Methodology
  - QD scaling: 1->128

CPU	Broadware 88x Linux cores
Memory	128 GB
NIC	40 Gb Mellanox NIC
Disk distribution	4x P3700 as OSD drive, 1x Optane as DB driver
Software configuration	Ubuntu 14.04, Ceph Luminous (dev)

Ceph w/ RoCE performance on 8x OSDs cluste

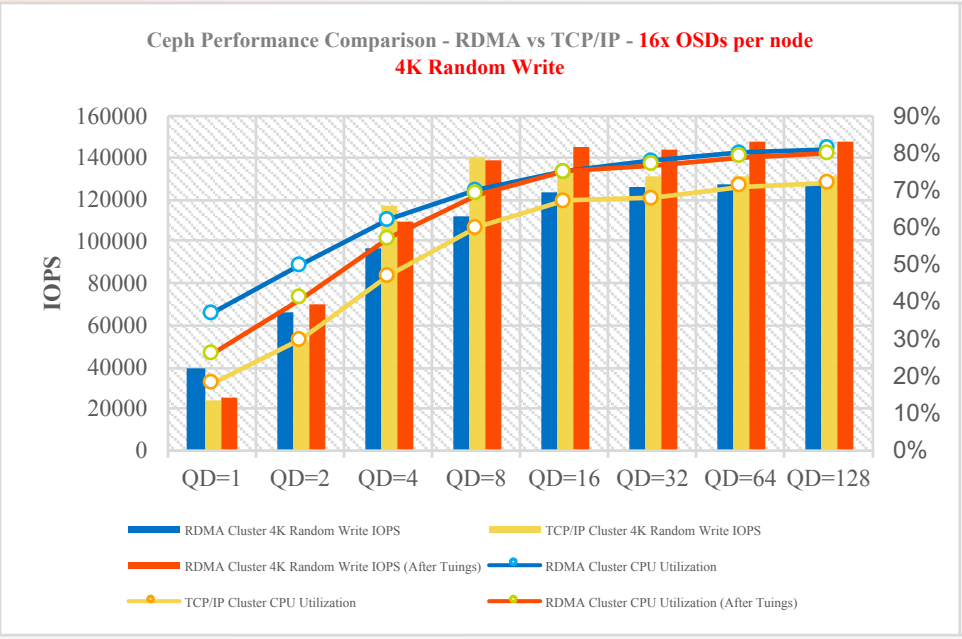
- The performance of Ceph w/ RoCE is ~11% to ~86% higher than TCP/IP.
- The total CPU utilization of Ceph w/ RoCE cluster is ~14% higher than TCP/IP.
- The user level CPU utilization of Ceph w/ RoCE cluster is ~13% higher than TCP/IP.

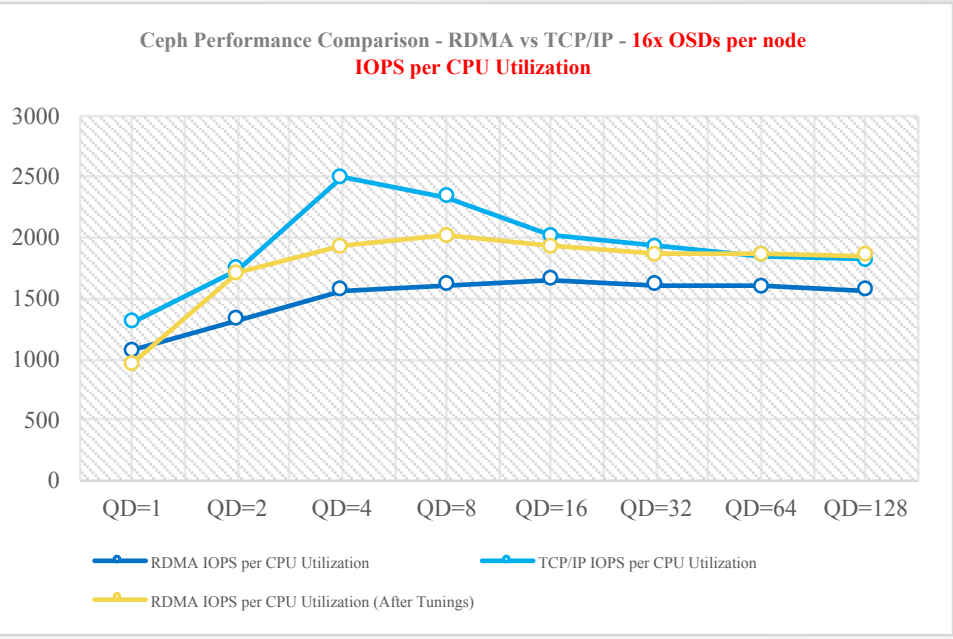




Ceph w/ RoCE performance (after tunings) on 16x OSDs cluster

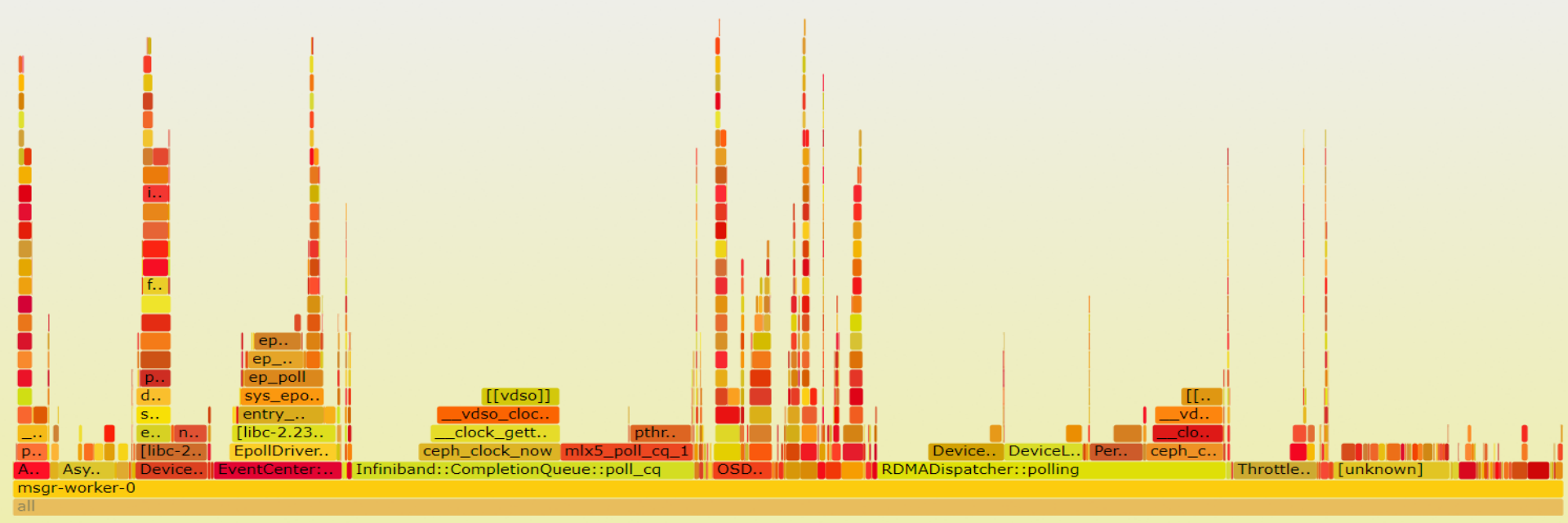
- With Ceph tunings, the performance of Ceph w/ RoCE is higher than TCP in high QD workload.
  - The IOPS per CPU of Ceph w/ RoCE cluster is higher than TCP cluster.
  - But still lower in low QD workload.
- Tunings:
  - Increase RDMA completed queue depth.
  - Decrease Ceph RDMA polling time.





CPU profiling

- ~10% of the total CPU used by Ceph is consumed by RDMA polling thread.
- Both Ceph RDMA and TCP/IP code are based on Epoll, while RDMA polling thread requires extra CPU cycle.



ceph

CEPH

CEPHALOCON APAC 2018

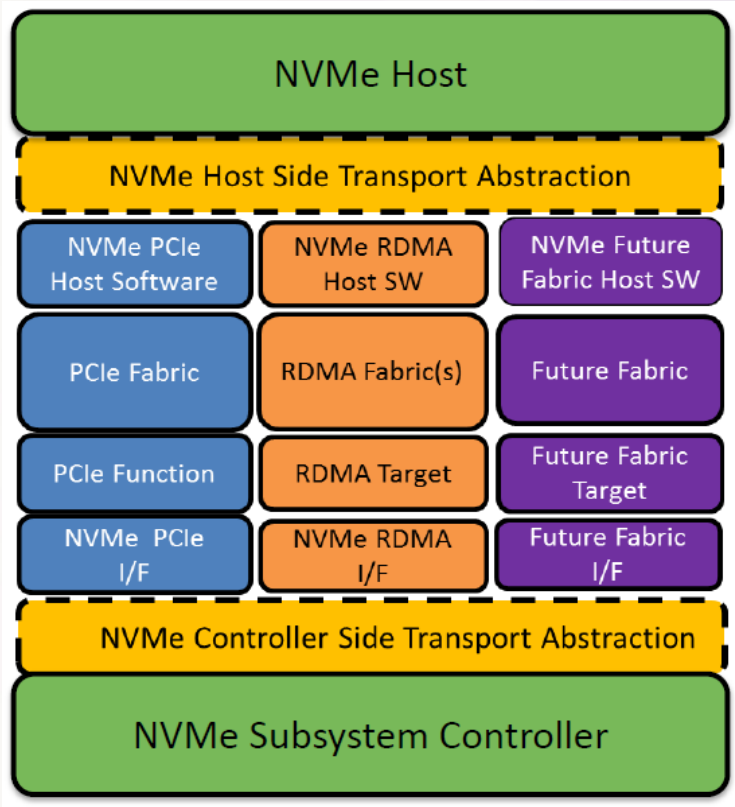
THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

Ceph with NVMe-oF

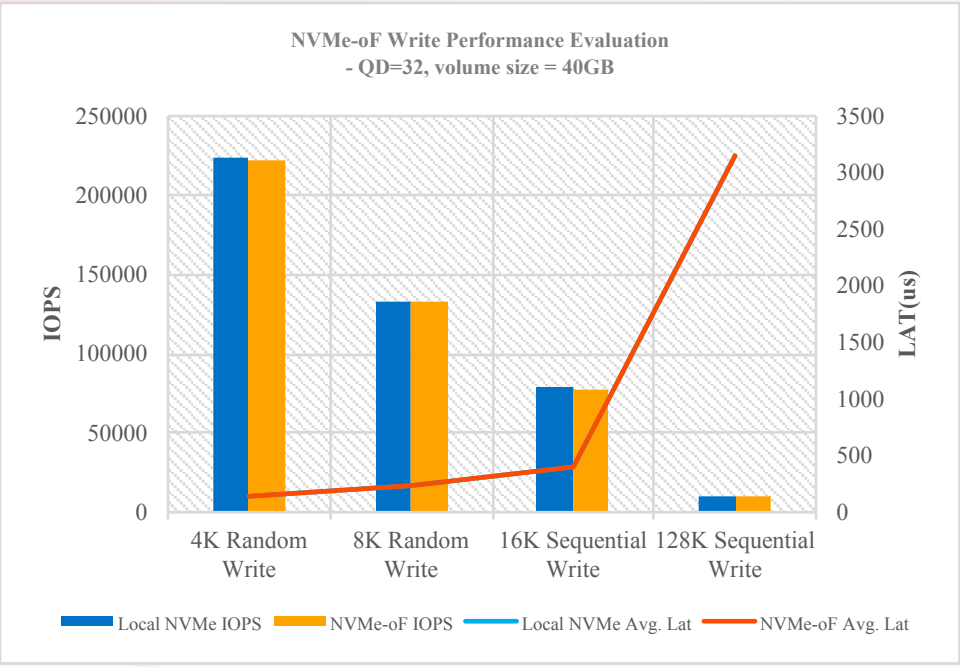
NVMe over Fabrics (NVMe-oF)

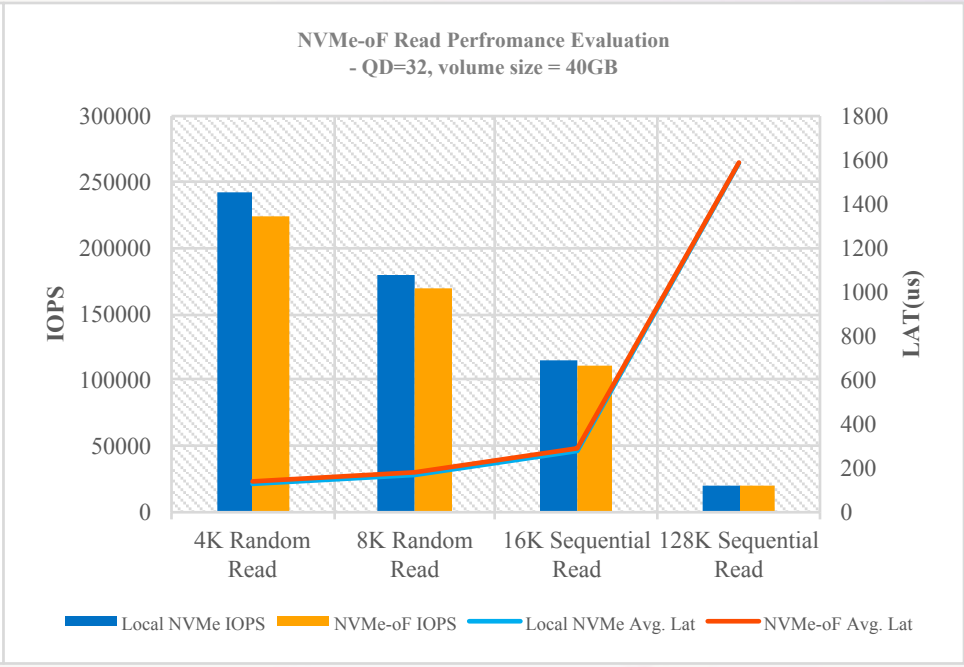
- NVMe is a new specification optimized for NAND flash and next-generation solid-state storage technologies.
- NVMe over Fabrics enables access to remote NVMe devices over multiple network fabrics.
  - Supported fabrics
    - RDMA – InfiniBand, IWARP, RoCE
    - Fiber Channel
    - TCP/IP
- NVMe-oF benefits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



FIO performance

- NVMe-oF added negligible performance overhead for write IO (< 1%)
- NVMe-oF added up to ~8.7% performance gap for read IO.



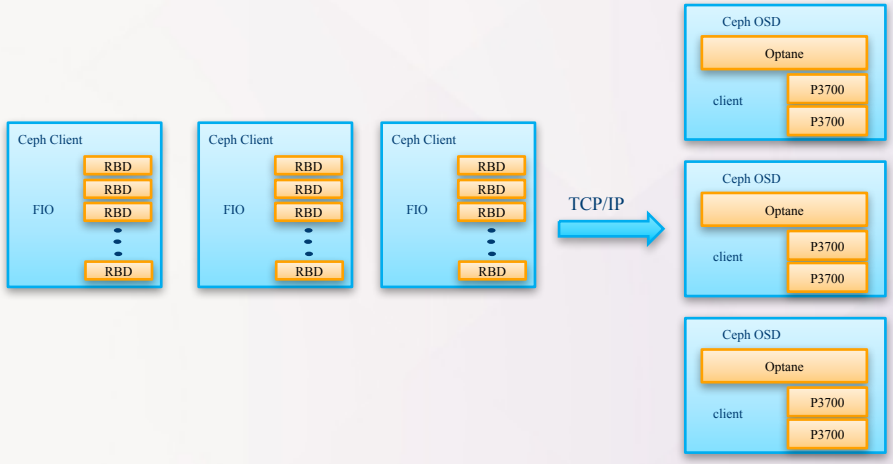


Ceph over NVMf

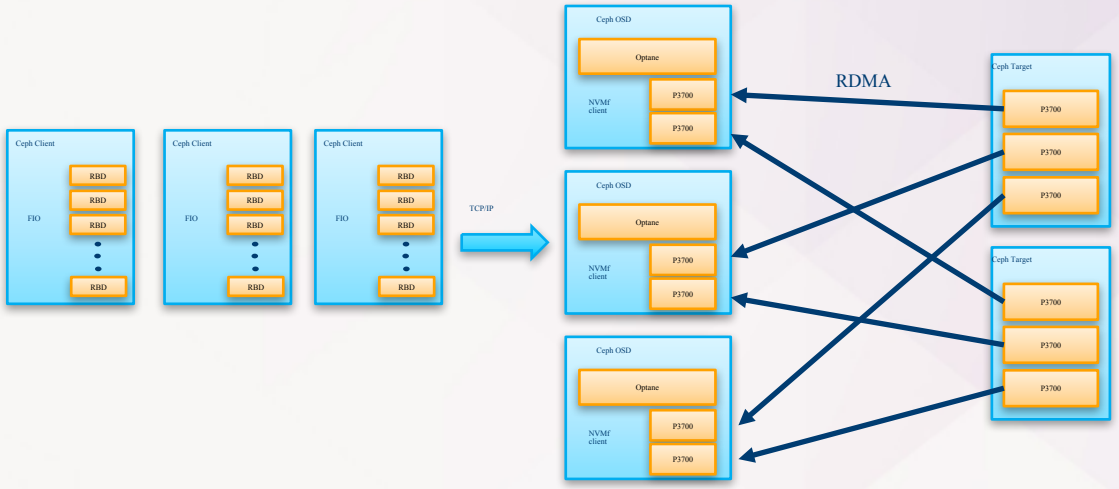
- Expectations and questions before POC.
  - Expectations: According to the benchmark from the first part, we're expecting
    - on-par 4K random write performance with NVMe-oF for Ceph.
    - on-par CPU utilization on NVMe-oF host node.
  - Questions:
    - How many CPU will be used on NVMe-oF target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

Benchmark methodology

- 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.

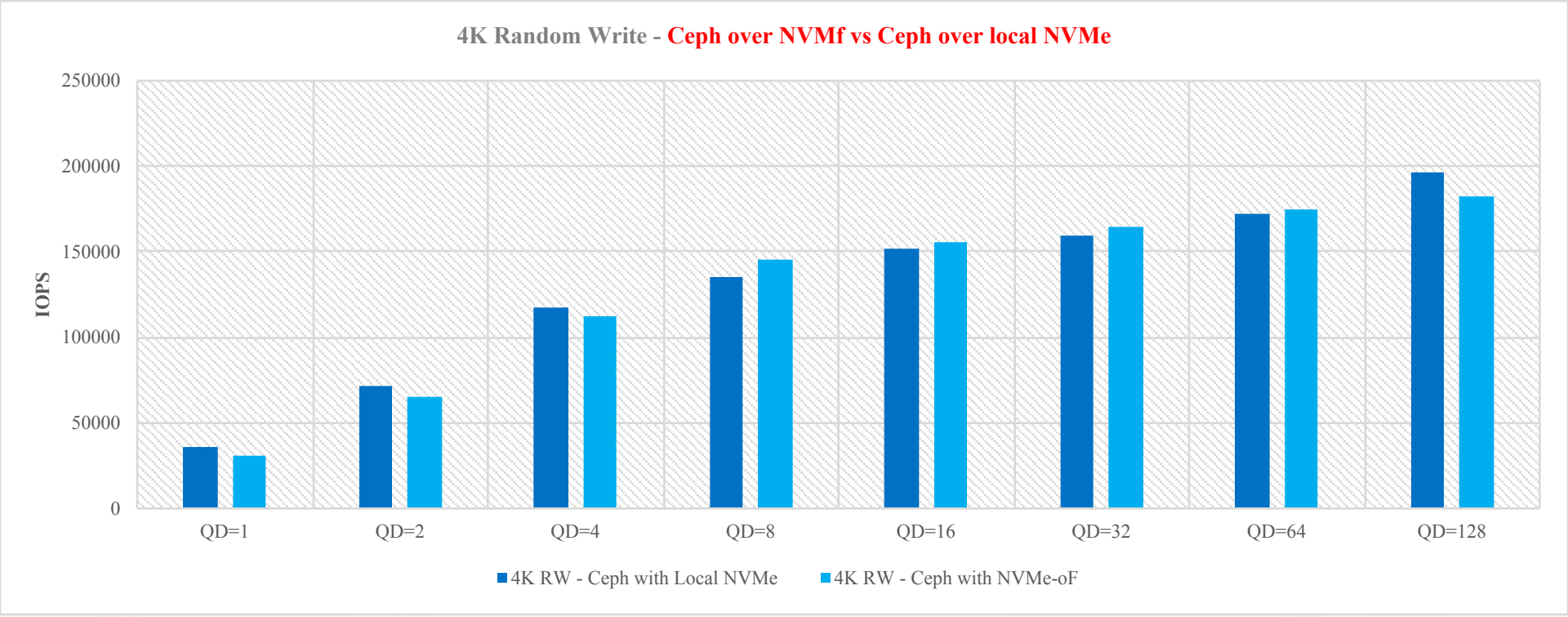


- Hardware configuration
  - 2x Storage nodes, 3x OSD nodes, 3x Client nodes.
  - 6x P3700 (800 GB U.2), 3x Optane (375 GB)
  - 30x FIO processes worked on 30x RBD volumes.
  - All these 8x servers are BRW, 128 GB memory, Mellanox Connect-X4 NICs.



Ceph over NVMe-oF – 4K random write

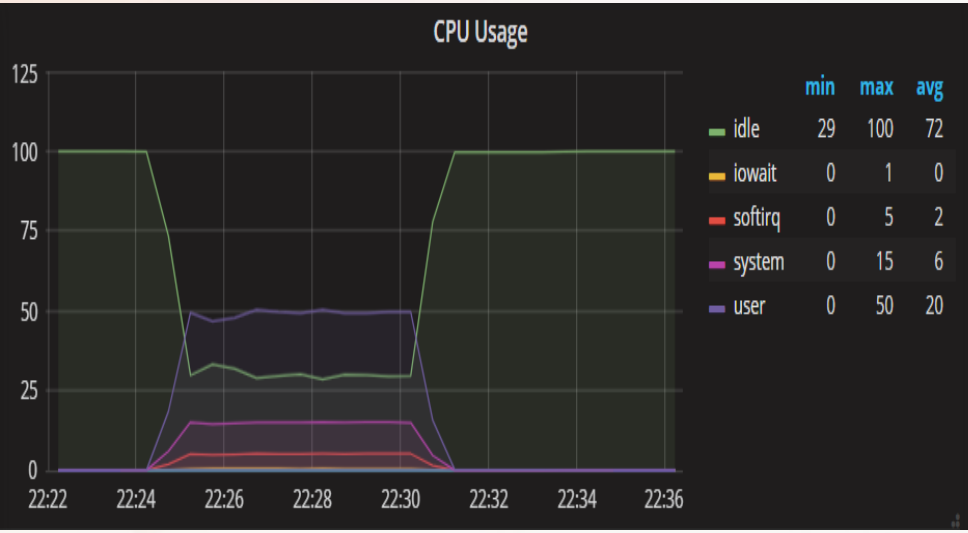
- Compared with traditional setup, running Ceph over NVMf didn't degrade 4K random write IOPS.



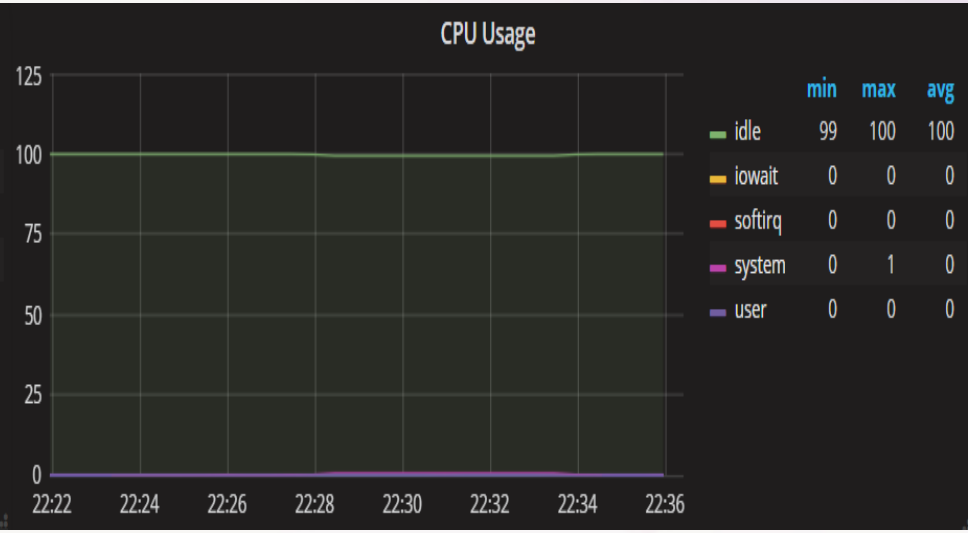
Ceph over NVMe-oF – CPU overheads

- Running Ceph over NVMe-oF add < 1% CPU overheads on target node.
- Running Ceph over NVMe-oF didn't add extra CPU overheads on host(OSD) node.

CPU Utilization on OSD Node

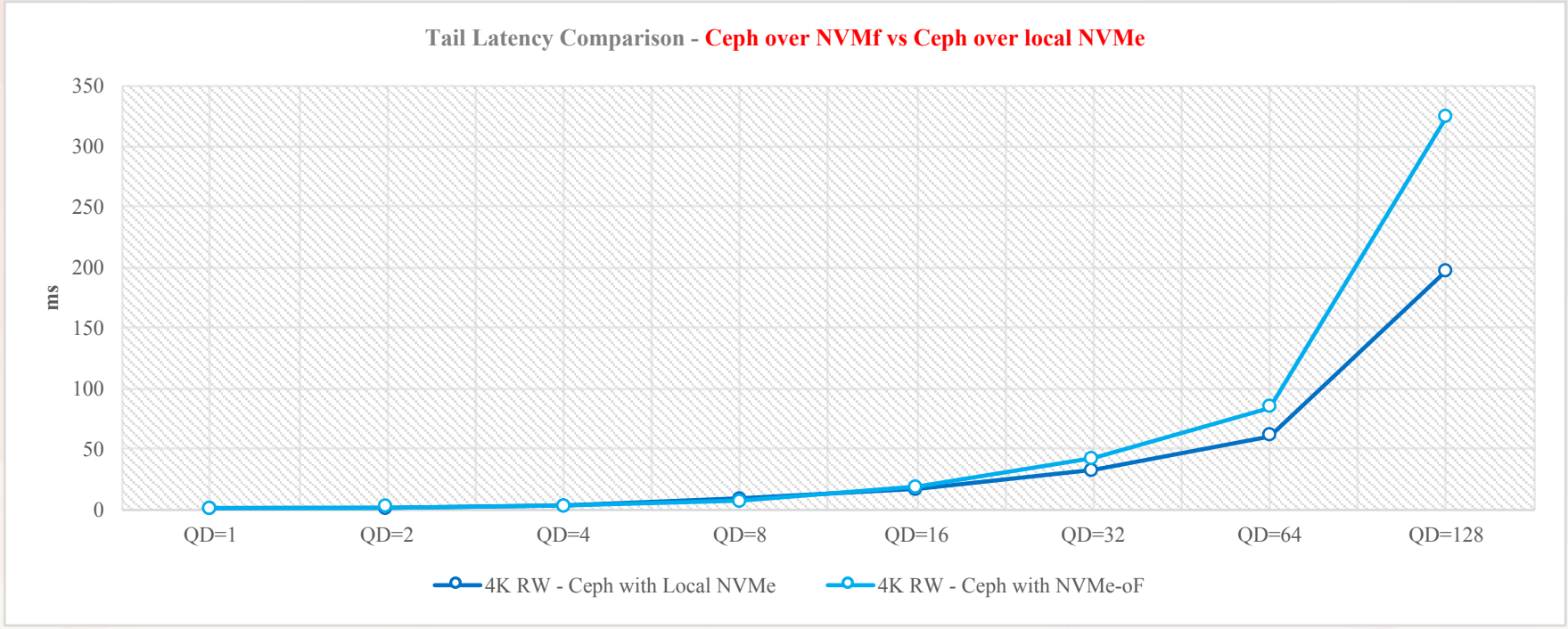


CPU Utilization on Target Node



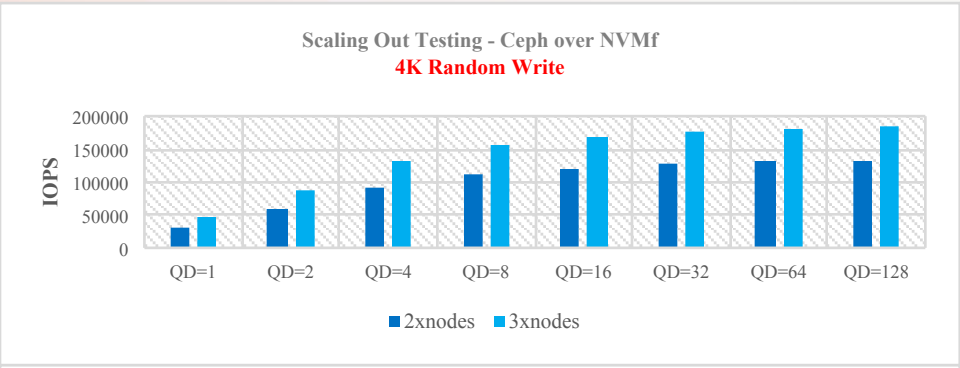
Ceph over NVMe-oF – tail latency

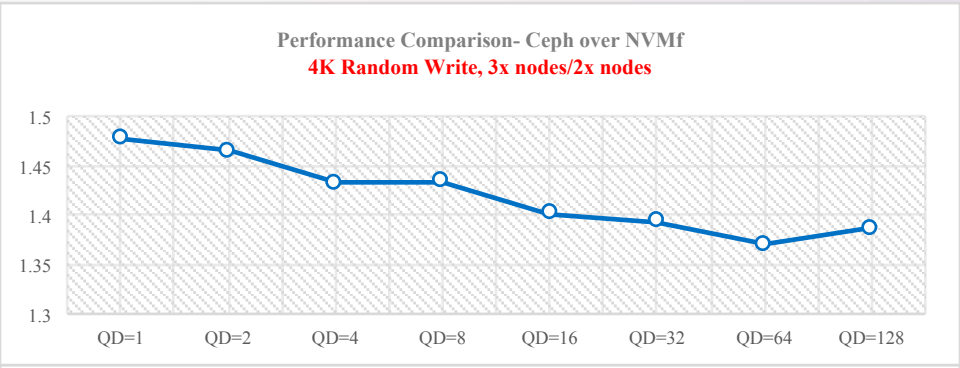
- When QD is higher than 16, Ceph over NVMf shows higher tail latency (99%).
- When QD is lower than 16, Ceph over NVMf on-par with Ceph over local NVMe.

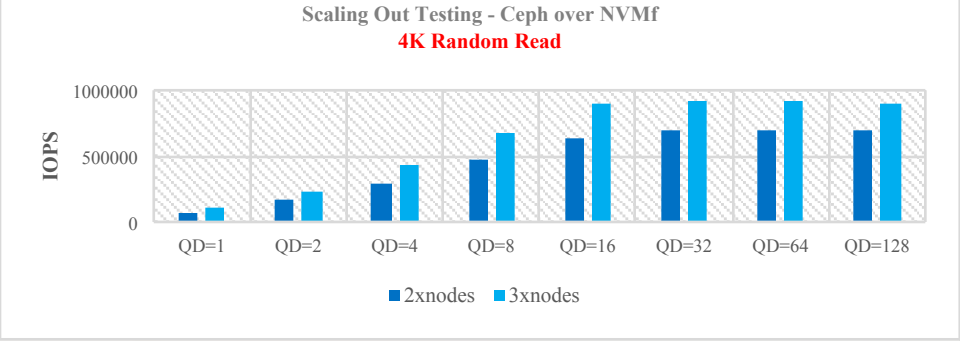


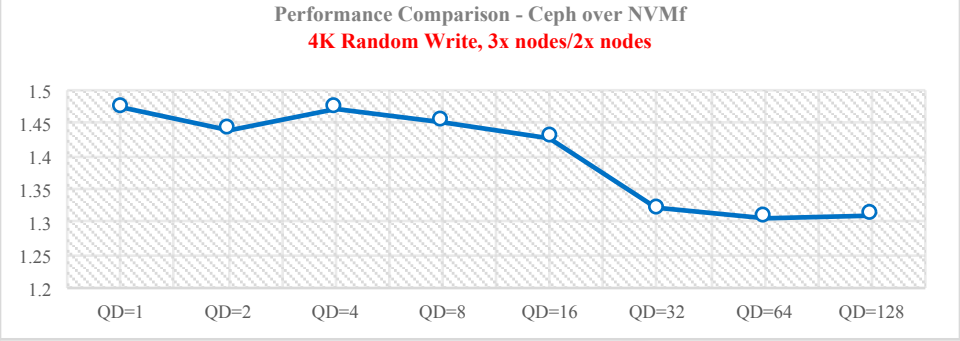
Ceph over NVMe-oF – OSD node Scaling out

- 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).



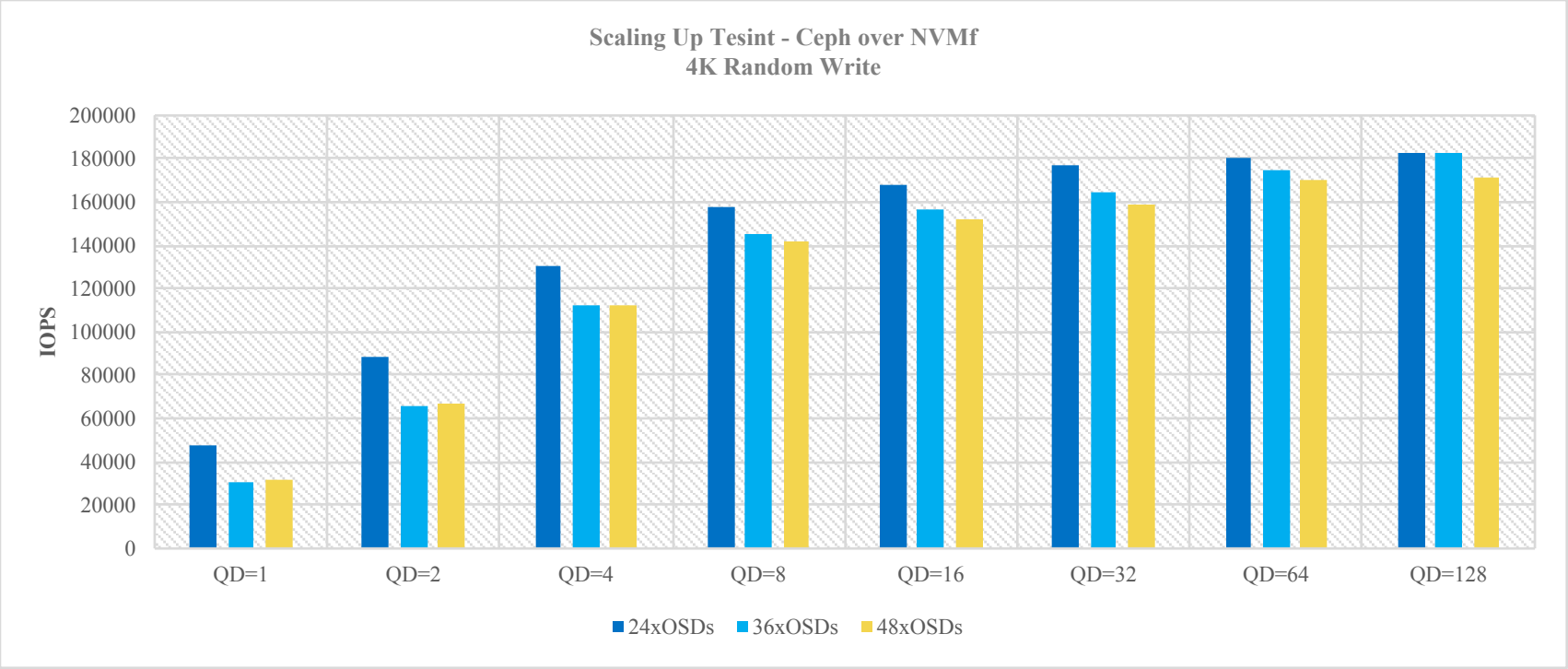






Ceph over NVMe-oF – OSD Scaling up

- The OSD scalability per OSD node depends on Ceph architecture.
- Running Ceph over NVMe-oF didn't improve the OSD scalability.



ceph

CEPH

CEPHALOCON APAC 2018

THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

Summary & next-step

Summary & Next-step

- Summary
  - RDMA is critical for future Ceph AFA solutions.
    - Ceph with RDMA messenger provides up to ~86% performance advantage over TCP/IP in low queue depth workload.
  - 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
  - Expand Ceph iWARP cluster scale, to 5 or 10 ODS node with 5 client node.
  - 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