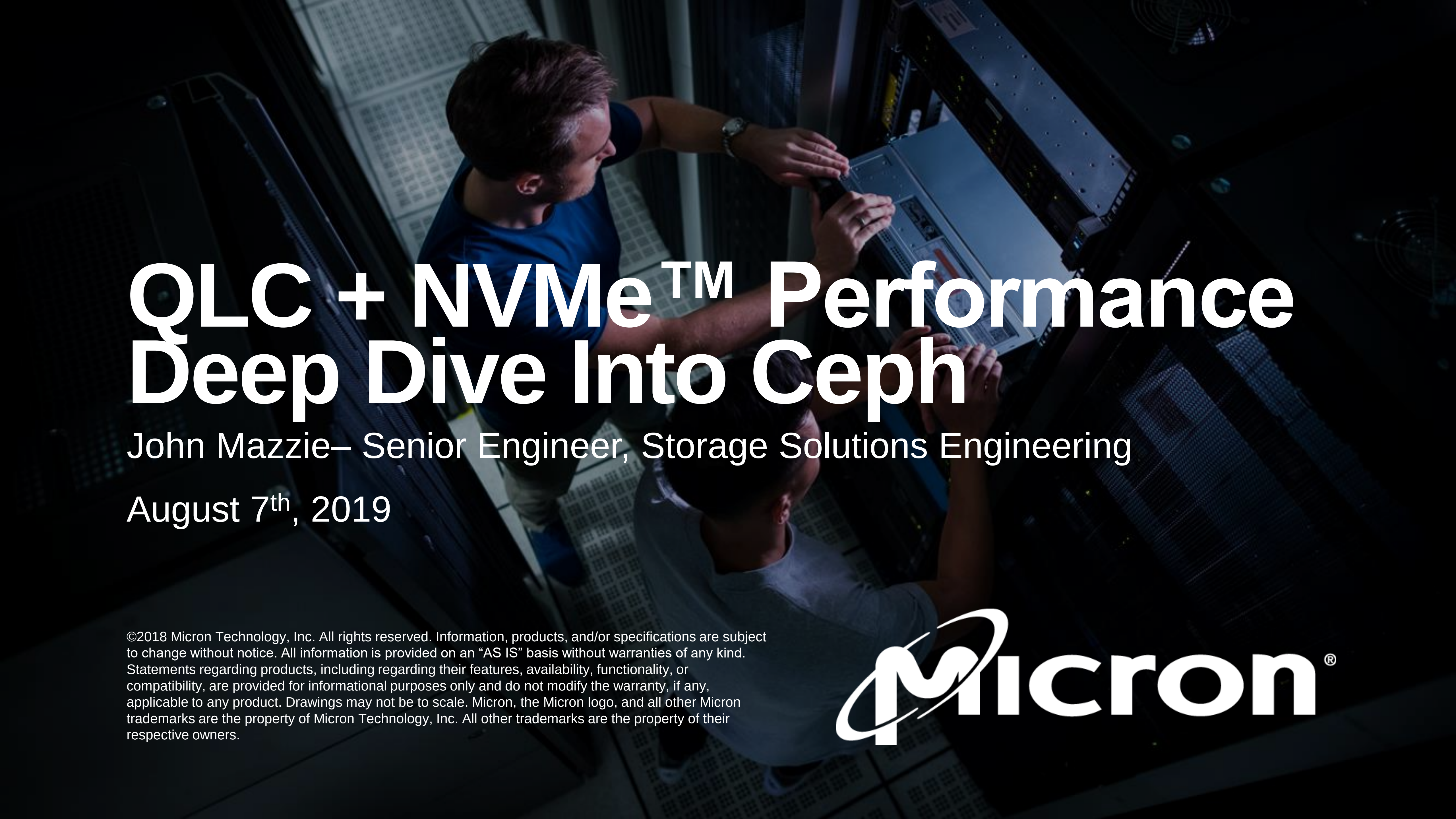


QLC + NVMe™ Performance Deep Dive Into Ceph

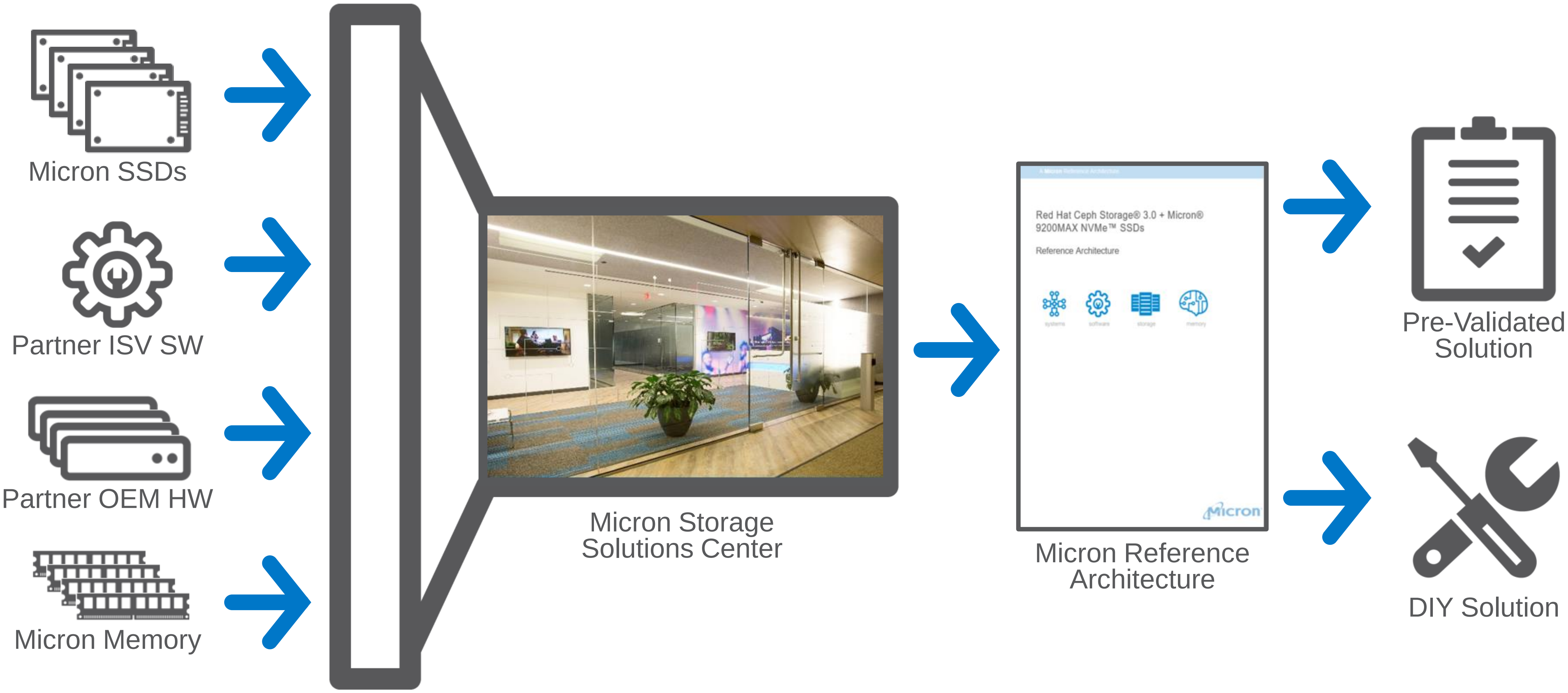
John Mazzie– Senior Engineer, Storage Solutions Engineering

August 7th, 2019

Micron Storage Solutions Engineering

- Austin, TX
- Big Fancy Lab
- Real-world application performance testing using Micron Storage & Memory
  - Ceph, vSAN, Storage Spaces Direct
  - Hadoop, Spark
  - Oracle, MSSQL, MySQL
  - Cassandra, MongoDB

Micron Reference Architectures

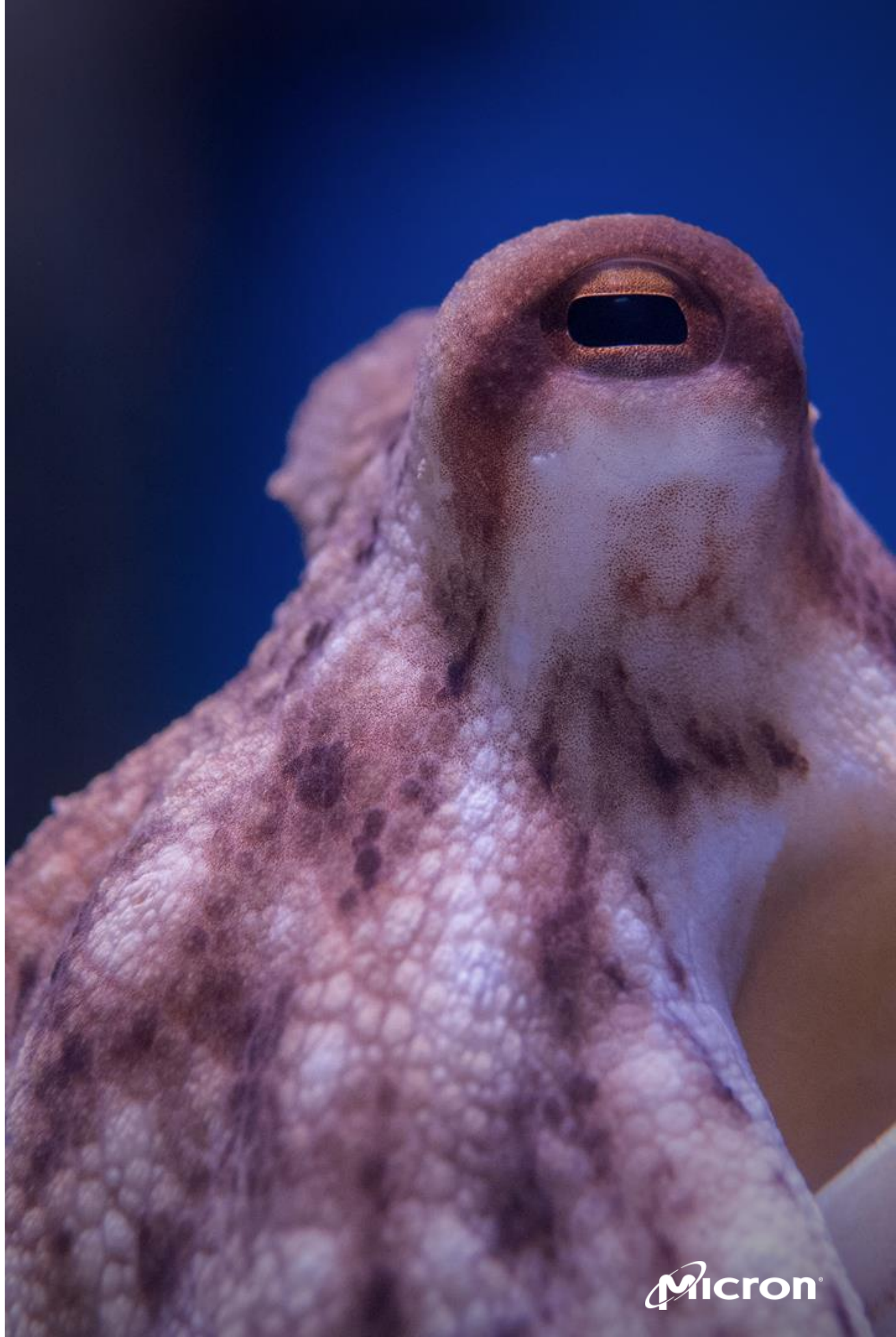


Micron SSDs
Partner ISV SW
Partner OEM HW
Micron Memory
Micron Storage Solutions Center
A Micron Reference Architecture
Red Hat Ceph Storage® 3.0 + Micron® 9200MAX NVMe™ SSDs
Reference Architecture
systems
software
storage
memory
Micron
Micron Reference Architecture
Pre-Validated Solution
DIY Solution

What is Ceph?

What is Ceph?

- Software Defined Storage
  - Uses off the shelf servers & drives
- Open Source
  - Dev team hired by Red Hat
  - Red Hat and other Linux vendors sell supported Ceph versions
- Scale Out
  - Add storage nodes to increase space & compute
  - Uses crc32c + replication or erasure coding for storage data protection

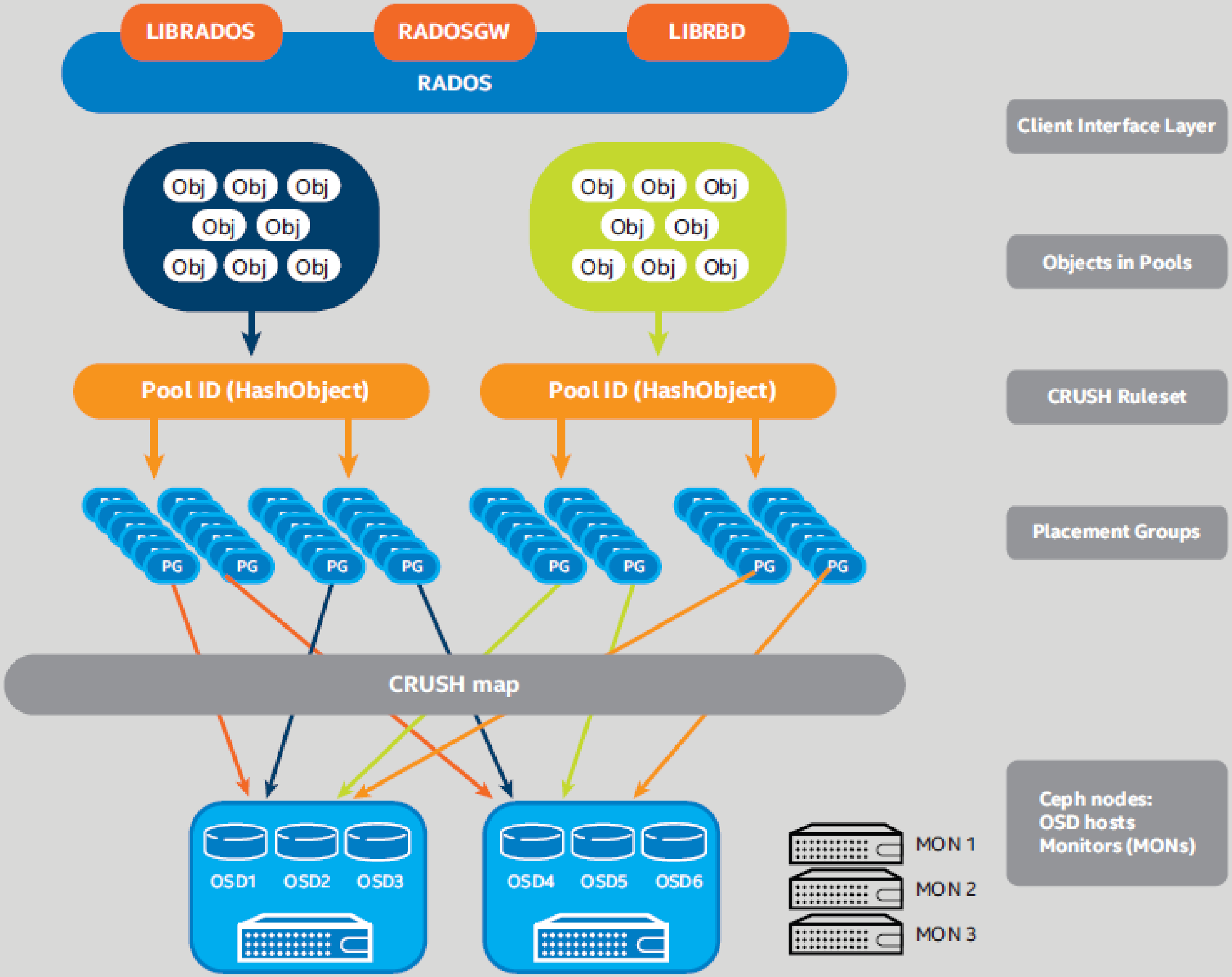


What is Ceph?

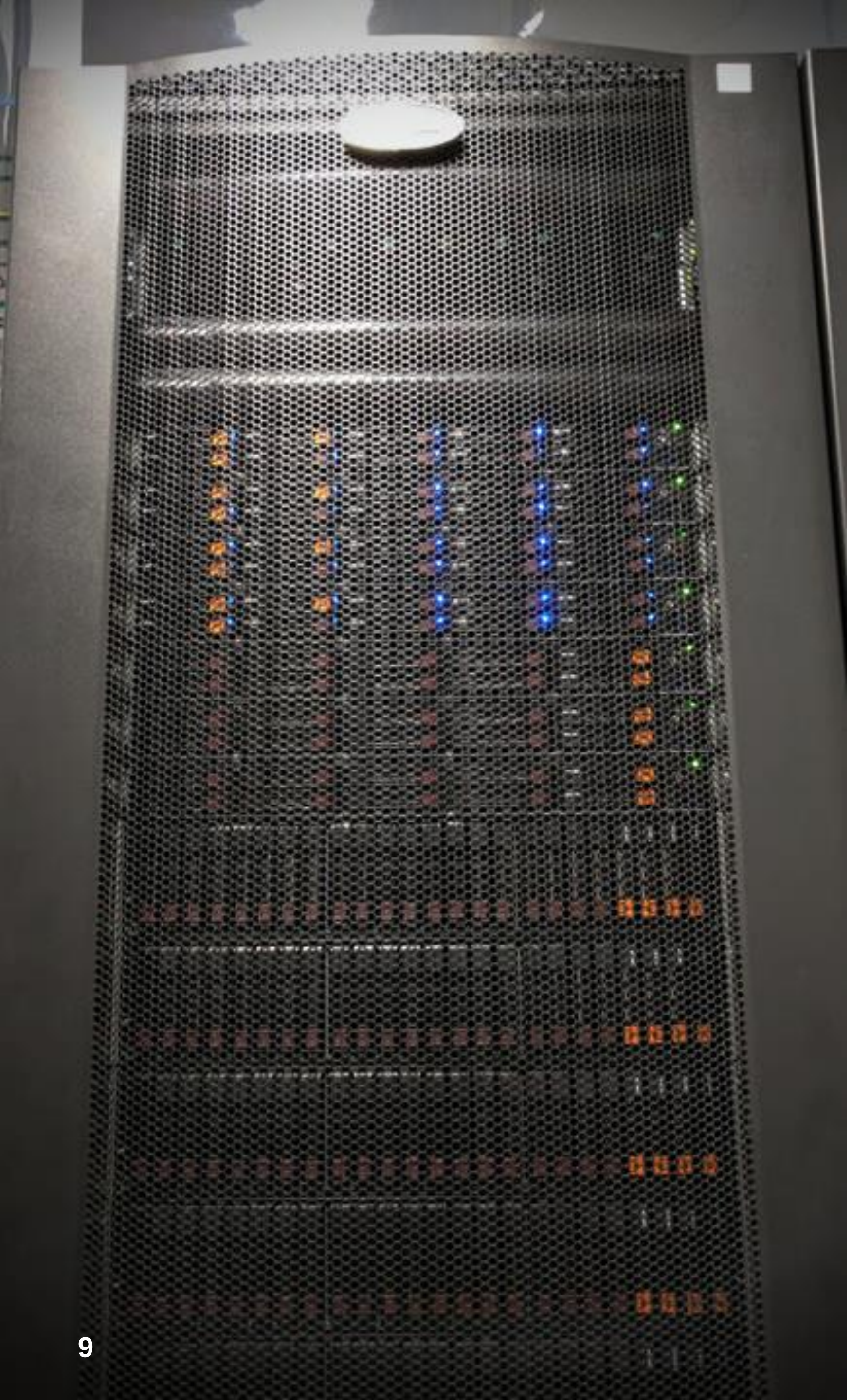
- Supports object, block, and file storage
  - Object: Native Rados API / Amazon S3 / Swift
  - Block: Rados Block Driver
    - Can present an image to a client as a standard block device.
    - Any Linux server with librbd installed can use as persistent storage
    - Tested using standard storage benchmark tools like FIO
  - File
    - POSIX compliant file system
    - Mount directly on Linux host
    - Single namespace

What is Ceph?

- RADOS: Reliable Autonomic Distributed Object Storage
  - LIBRADOS: Object API
  - RADOSGW: S3, Swift, API Gateway
  - LIBRBD: Block Storage
- OSD: Object Store Daemon
  - Process that manages storage
  - Usually 1 to 2 OSDs per Drive
- MON: Monitor Node
  - Maintains CRUSH map
  - 3 Mons minimum for failover



Red Hat Ceph Storage 3.2 QLC + NVMe

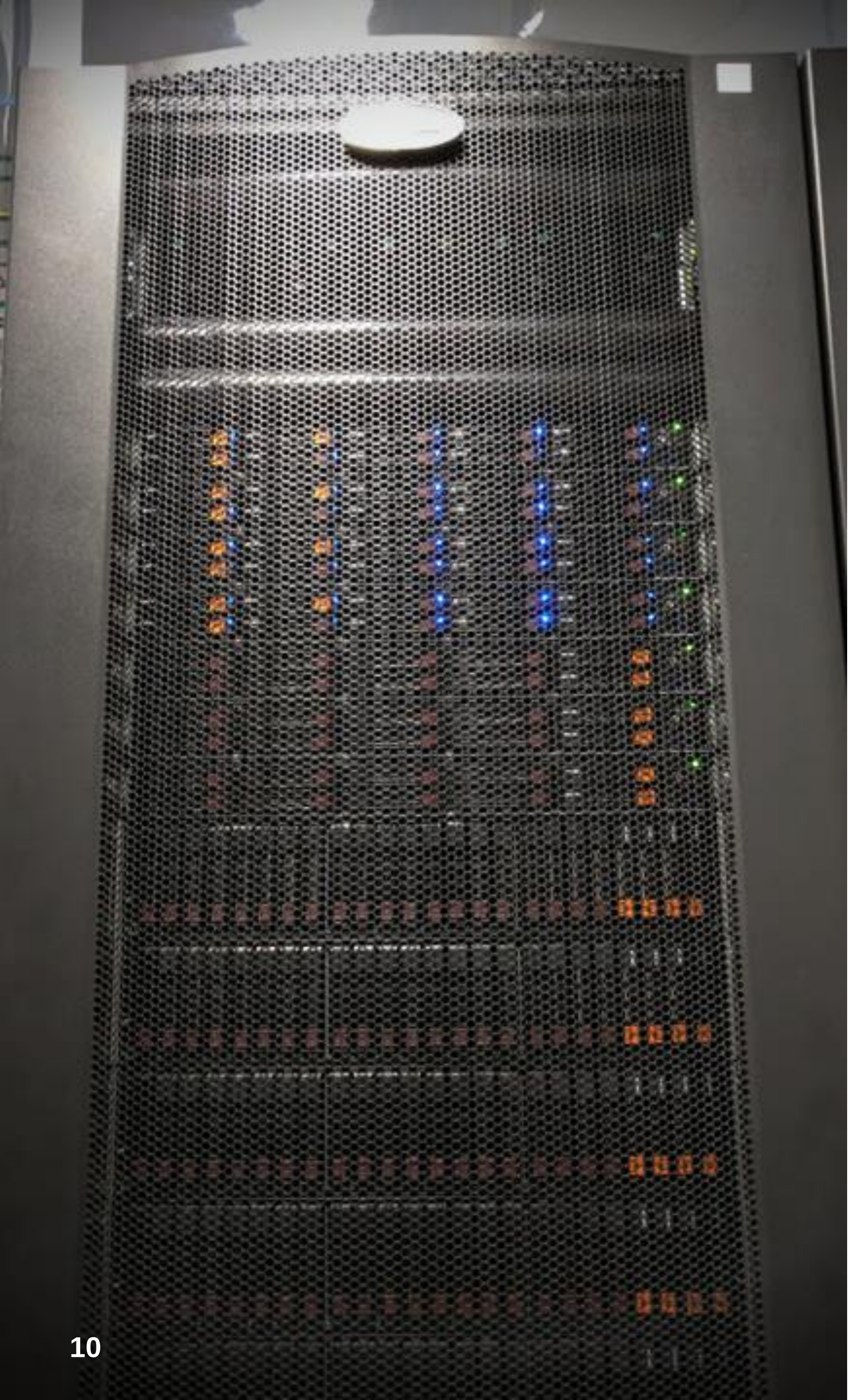


Hardware Configuration

Micron + Red Hat + Supermicro QLC + NVMe Ceph

Storage Nodes (x4)

- Supermicro A+ AS-2113S-WTRT
- 1x AMD EPYC 7551P 32 core, 2.0Ghz Base / 2.55Ghz Boost
- 256GB Micron High Quality Excellently Awesome DDR4-2666 DRAM (8x 32GB)
- 2x Mellanox ConnectX-5 100GbE 2-port NICs
  - 1 NIC for client network / 1 NIC for storage network
- Broadcom SAS 9305-24i HBA
- 12x Micron 3.84TB 5210 ION QLC SATA SSD
  - 83k 4KB Random Read IOPs / 6.5k 4KB Random Write IOPs
  - 540 MB/s Sequential Read / 83 MB/s Sequential Write
- 2x Micron 1.6TB 9200 MAX NVMe SSD
  - 680k 4KB Random Read IOPs / 255k 4KB Random Write IOPs
  - 3.5 GB/s Sequential Read / 1.9 GB/s Sequential Write



Hardware Configuration

Micron + Red Hat + Supermicro ALL-NVMe Ceph

Monitor Nodes (x1)

- Supermicro A+ AS-2113S-WTRT
  - 1x AMD EPYC 7551P 32 core, 2.0Ghz Base / 2.55Ghz Boost
  - 256GB Micron High Quality Excellently Awesome DDR4-2666 DRAM (8x 32GB)
  - Mellanox ConnectX-4 50GbE single-port

Network

- 2x Supermicro SSE-C3632SR, 100GbE 32-Port Switches
  - 1 switch for client network / 1 switch for storage network

Load Generation Servers (Clients)

- 10x Supermicro SYS-2028U (2U)
- 2x Intel 2690v4
- 256GB RAM
- 50 GbE Mellanox ConnectX-4

Software Configuration

Micron + Red Hat + Supermicro ALL-NVMe Ceph

Storage + Monitor Nodes + Clients

- Red Hat Ceph Storage 3.2
- Red Hat Enterprise Linux 7.6
- Mellanox OFED Driver 4.4-2.0.7.0

Switch OS

- Cumulus Linux 3.7.1

Deployment Tool

- Ceph-Ansible

Performance Testing Methodology

Micron + Red Hat + Supermicro ALL-NVMe Ceph

- 2 OSDs per NVMe Drive / 96 OSDs total
- Ceph Storage Pool Config
  - 2x Replication: 8192 PG’s, 50x 150GB RBD Images = 7.5TB data x 2
- FIO RBD for Block Tests (4KB Block size)
  - Writes: FIO at queue depth 64 while scaling up # of client FIO processes
  - Reads: FIO against all 50 RBD Images, scaling up QD
- RADOS Bench for Object Tests (4MB Objects)
  - Writes: RADOS Bench @ threads 16, scaling up # of clients
  - Reads: RADOS Bench on 10 clients, scaling up # of threads
- 10-minute test runs x 3 for recorded average performance results (5 min ramp up on FIO)

Ceph Bluestore & NVMe

The Tune-Pocalypse

- Red Hat Ceph Storage 3.2
  - Tested using Bluestore
  - Official support added in 3.2
- Default RocksDB tuning for Bluestore in Ceph
  - Great for large object
  - Not great for 4KB random on SSD
  - Modified tuning for 4KB performance

Bluestore & NVMe

The Tune-Pocalypse

Bluestore OSD Tuning for 4KB Random Writes:

- Set high `max_write_buffer_number` & `min_write_buffer_number_to_merge`
- Set Low `write_buffer_size`

```
[osd]
 bluestore_csum_type = none
 bluestore_extent_map_shard_max_size = 200
 bluestore_extent_map_shard_min_size = 50
 bluestore_extent_map_shart_target_size = 100
 osd_min_pg_log_entries = 10
 osd_max_pg_log_entries = 10
 osd_pg_log_dups_tracked = 10
 osd_pg_log_trim_min = 10
 osd_memory_target = 10737418240
 bluestore_rocksdb_options =
compression=kNoCompression,max_write_buffer_number=64,min_wr
ite_buffer_number_to_merge=32,recycle_log_file_num=64,compac
tion_style=kCompactionStyleLevel,write_buffer_size=4MB,targe
t_file_size_base=4MB,max_background_compactions=64,level0_fi
le_num_compaction_trigger=64,level0_slowdown_writes_trigger=
128,level0_stop_writes_trigger=256,max_bytes_for_level_base=
6GB,compaction_threads=32,flusher_threads=8,compaction_reada
head_size=2MB
```

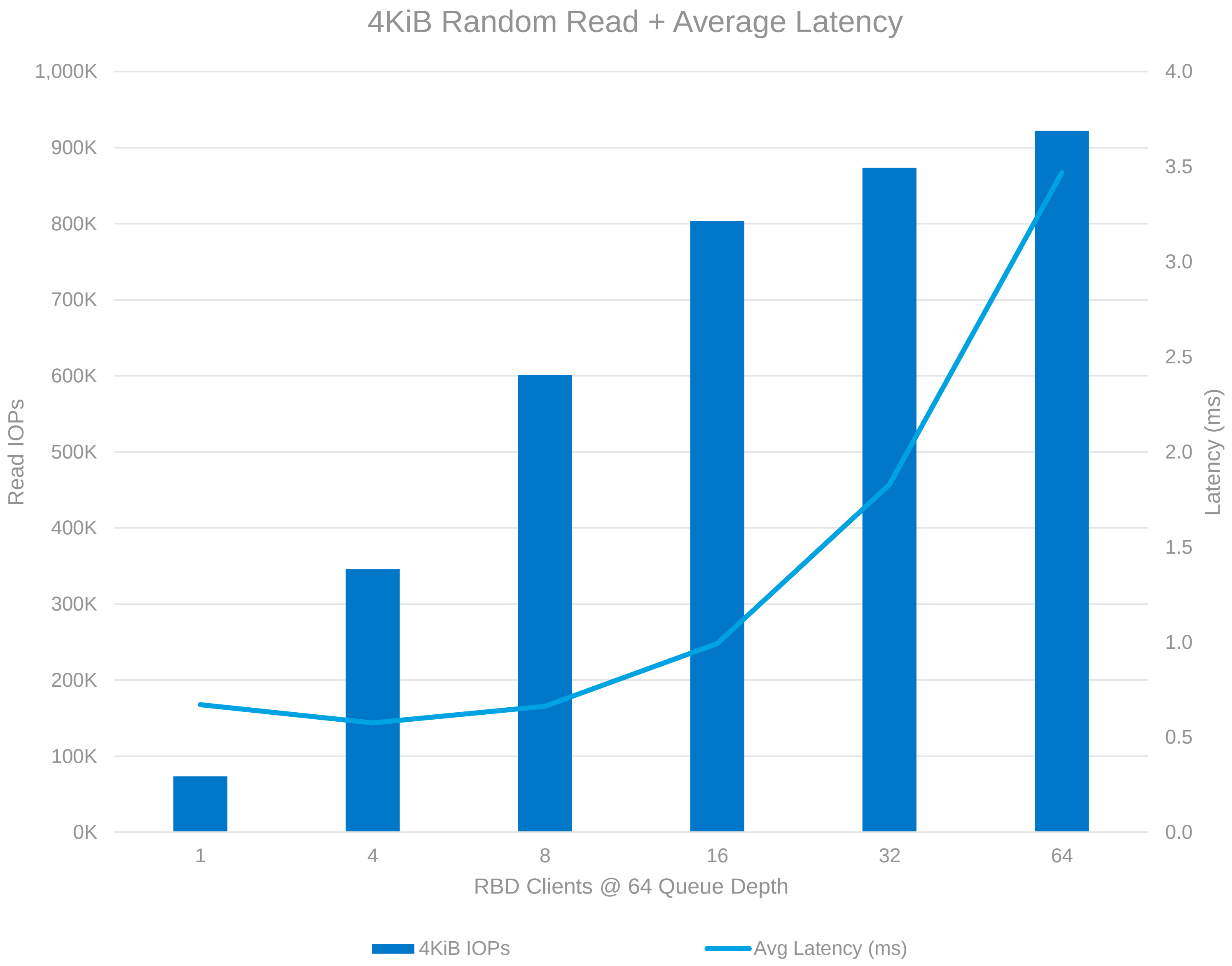
RHCS 3.2: 4KB Random Read

Micron + Red Hat
+ Supermicro
QLC + NVMe Ceph

4KB Random Reads:

- Queue Depth 32
  - 873K @ 1.8ms Avg. Latency

Tests become CPU Limited around Queue Depth 16



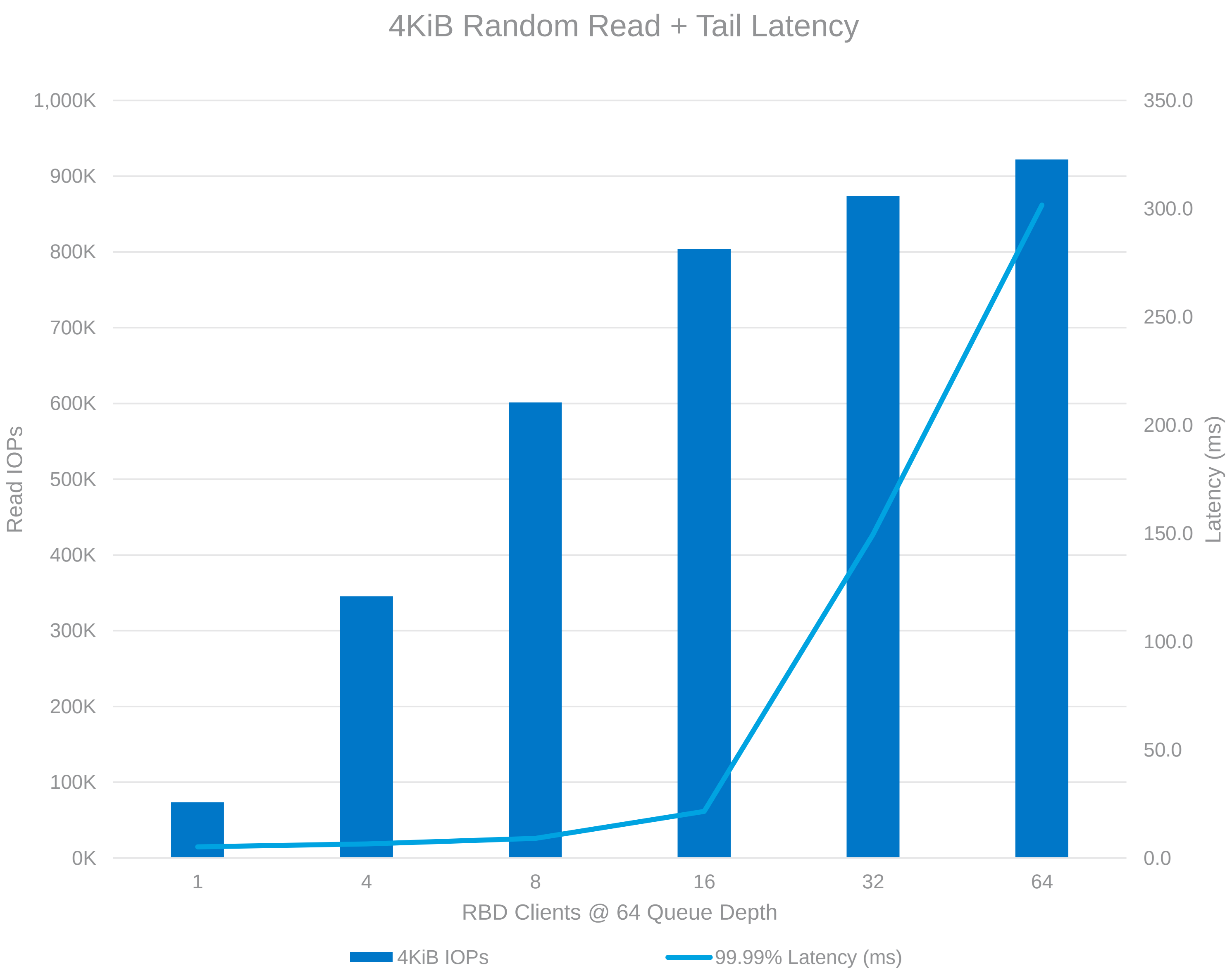
RHCS 3.2: 4KB Random Read

Micron + Red Hat + Supermicro QLC + NVMe Ceph

4KB Random Reads:

- Queue Depth 32
  - Bluestore Tail Latency: 149 ms

Tail latency spikes as tests become CPU limited

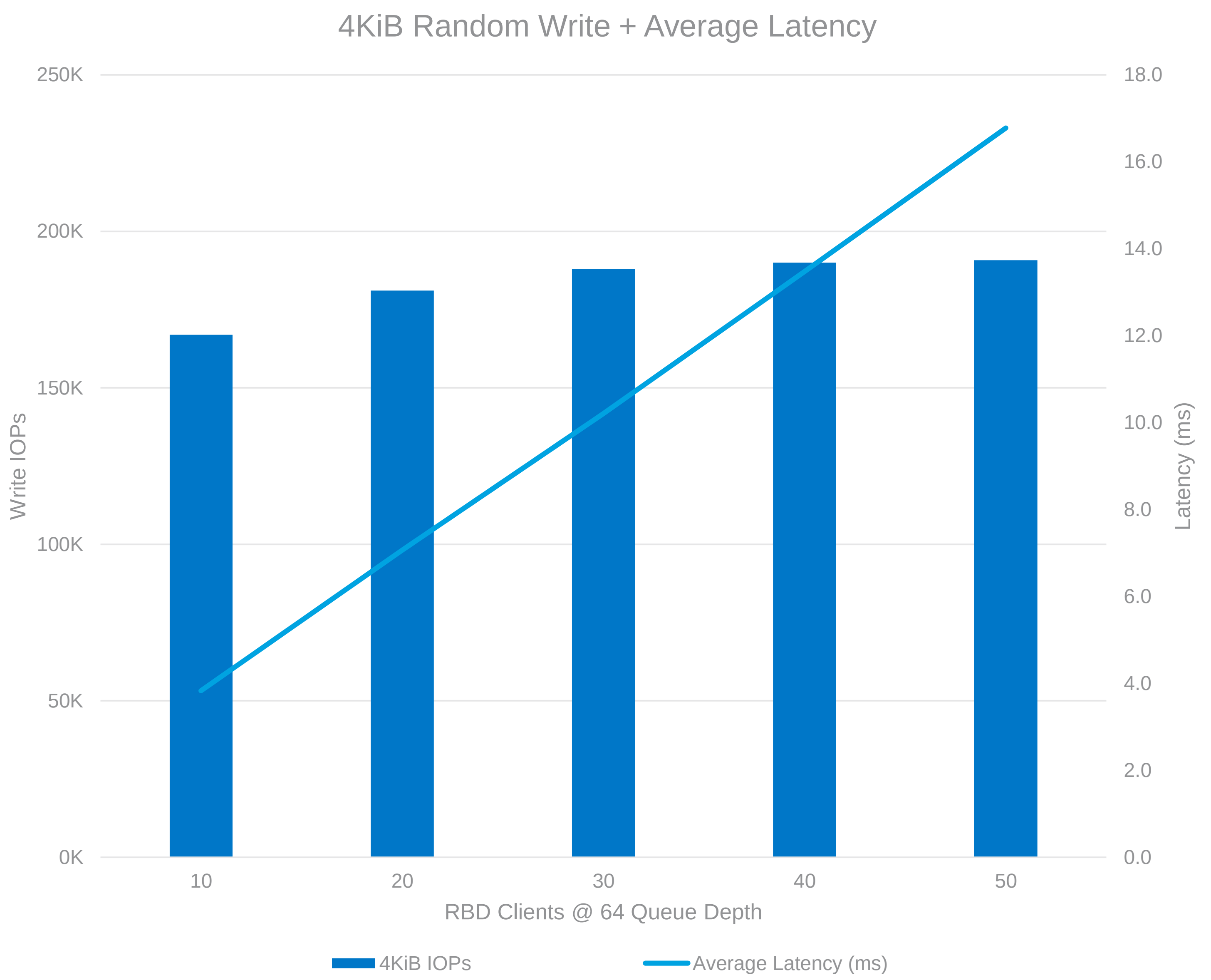


RHCS 3.2: 4KB Random Write

Micron + Red Hat
+ Supermicro
QLC + NVMe Ceph

4KB Random Writes:

- 40 Clients
  - 190k IOPs @ 13.5ms Avg Latency

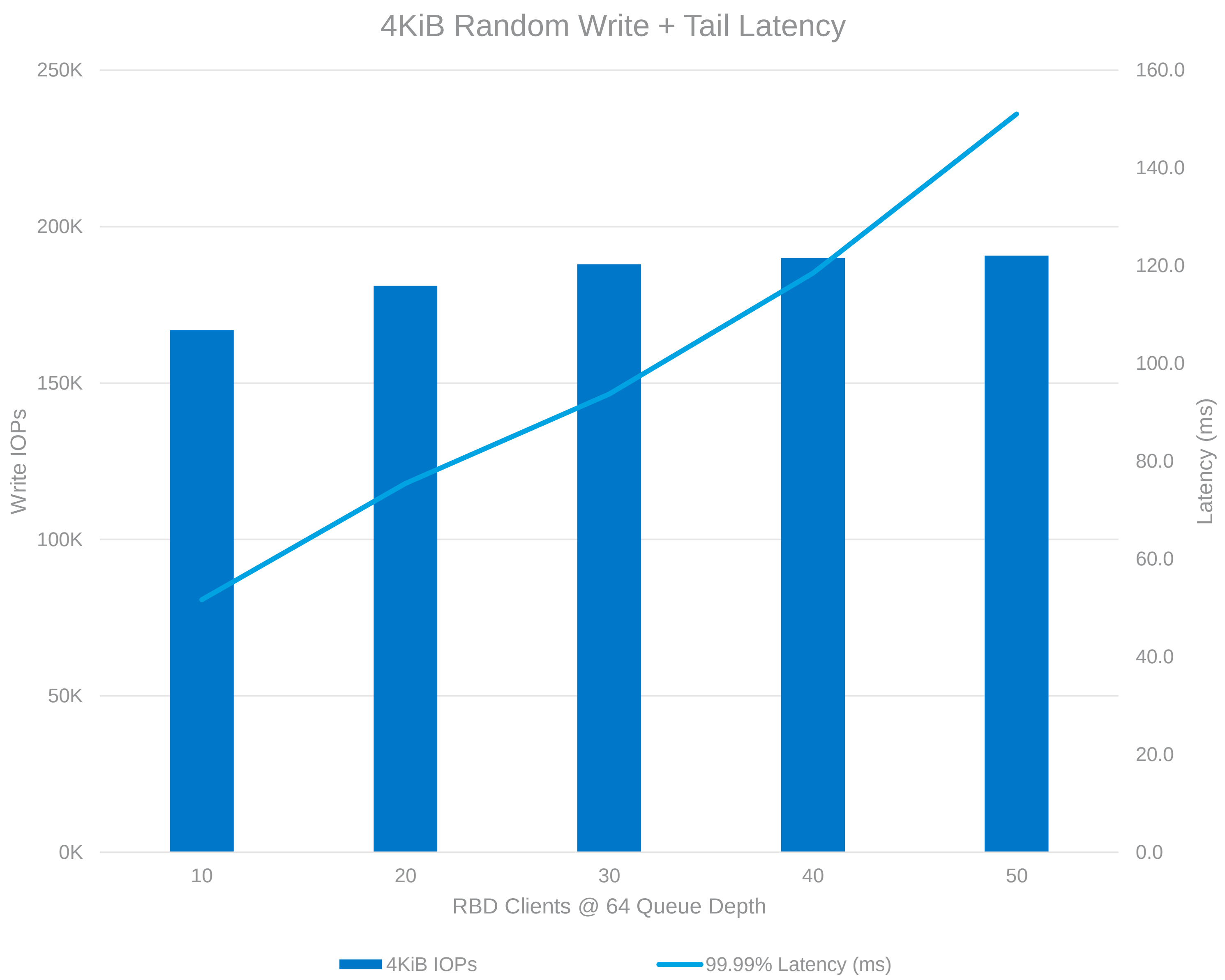


RHCS 3.2: 4KB Random Write

Micron + Red Hat + Supermicro QLC + NVMe Ceph

4KB Random Writes:

- 40 Clients
  - Tail Latency: 118.5ms

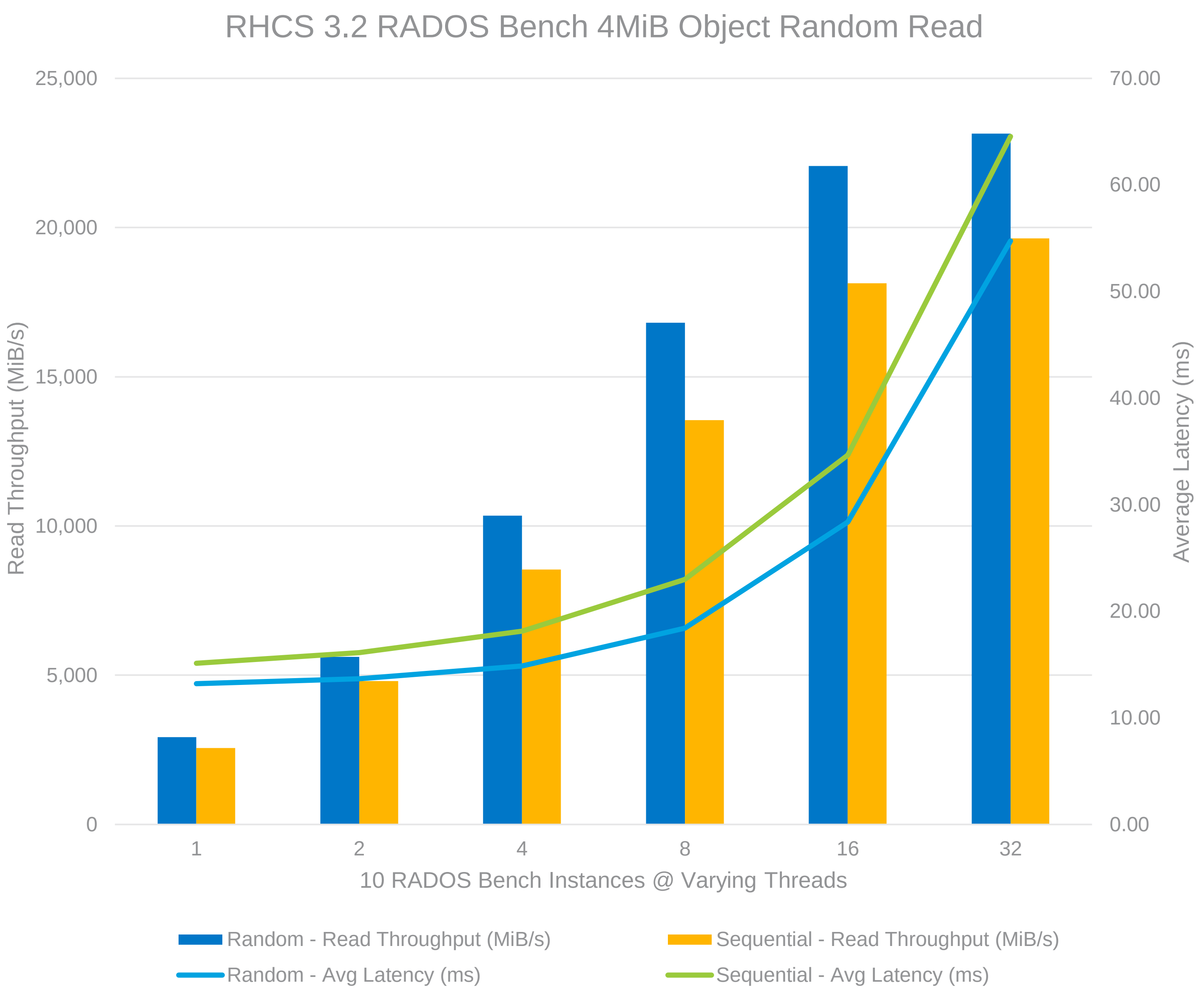


RHCS 3.2: 4MB Object Read

Micron + Red Hat + Supermicro QLC + NVMe Ceph

4MB Object Reads:

- 32 Threads:
  - Random Read
    - 22.4 GiB/s @ 55ms
  - Sequential Read
    - 19.0 GiB/s @ 65ms

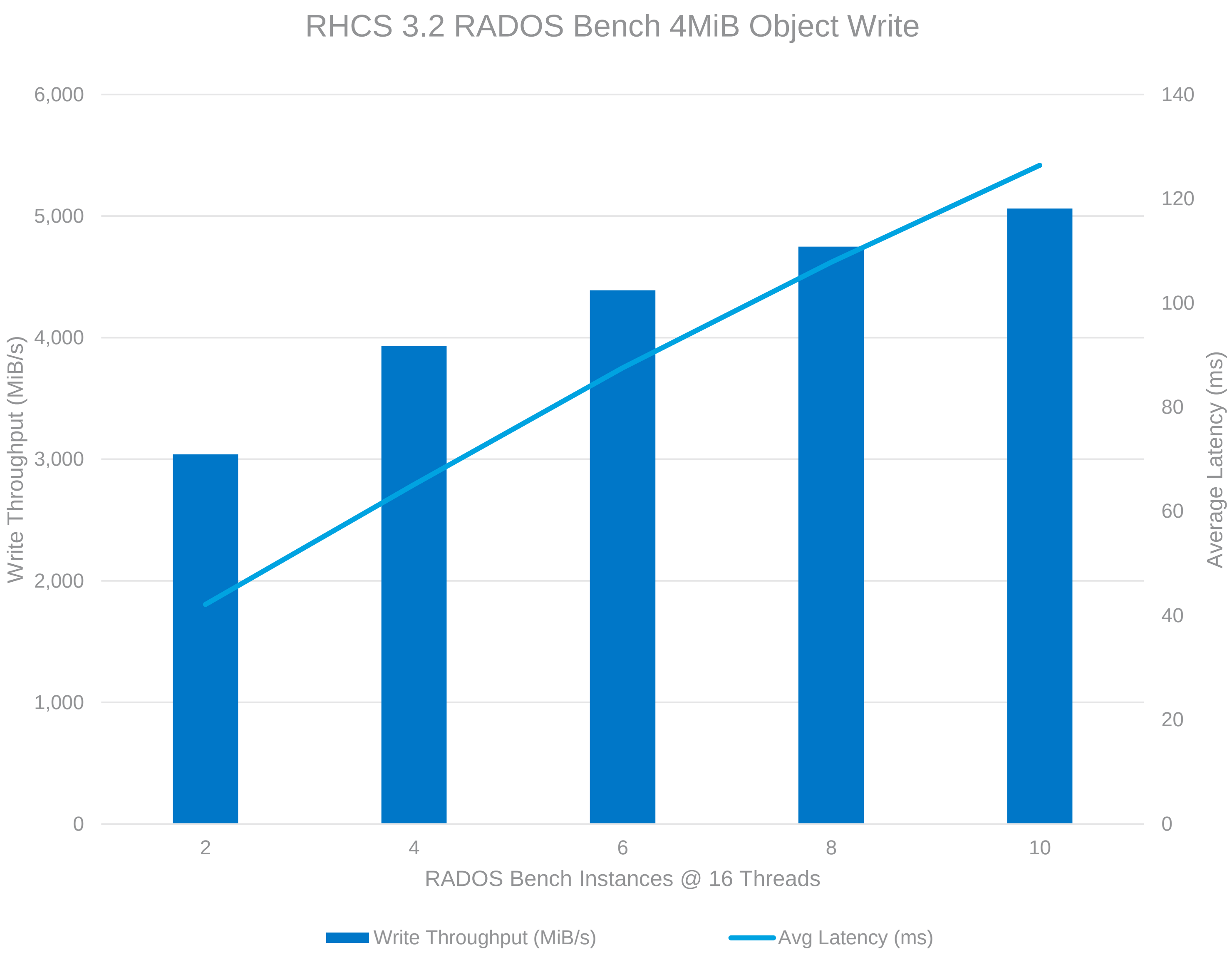


RHCS 3.2 : 4MB Object Write

Micron + Red Hat
+ Supermicro
QLC + NVMe Ceph

4MB Object Writes:

- 10 Clients:
  - 4.9 GiB/s @ 128ms



Would you like to know more?

Micron NVMe Reference Architecture:

https://www.micron.com/-/media/client/global/documents/products/other-documents/5210_9200_amd_ceph_reference_architecture.pdf

Micron Storage Blogs: Ceph

https://www.micron.com/about/blog/2019/march/ceph-bluestore-to-cache-or-not-to-cache-that-is-the-question

Thanks All

Micron®