

A high-angle, low-key photograph of two technicians in a server room. One technician, wearing a blue t-shirt, is leaning over a server rack, while the other, in a light-colored shirt, is looking at a server component. The scene is dimly lit, with the primary light source coming from the server racks, creating a dramatic, industrial atmosphere.

# All-NVMe Performance Deep Dive Into Ceph

+ Sneak Preview of QLC + NVMe Ceph

Ryan Meredith – Sr. Manager, Storage Solutions Engineering

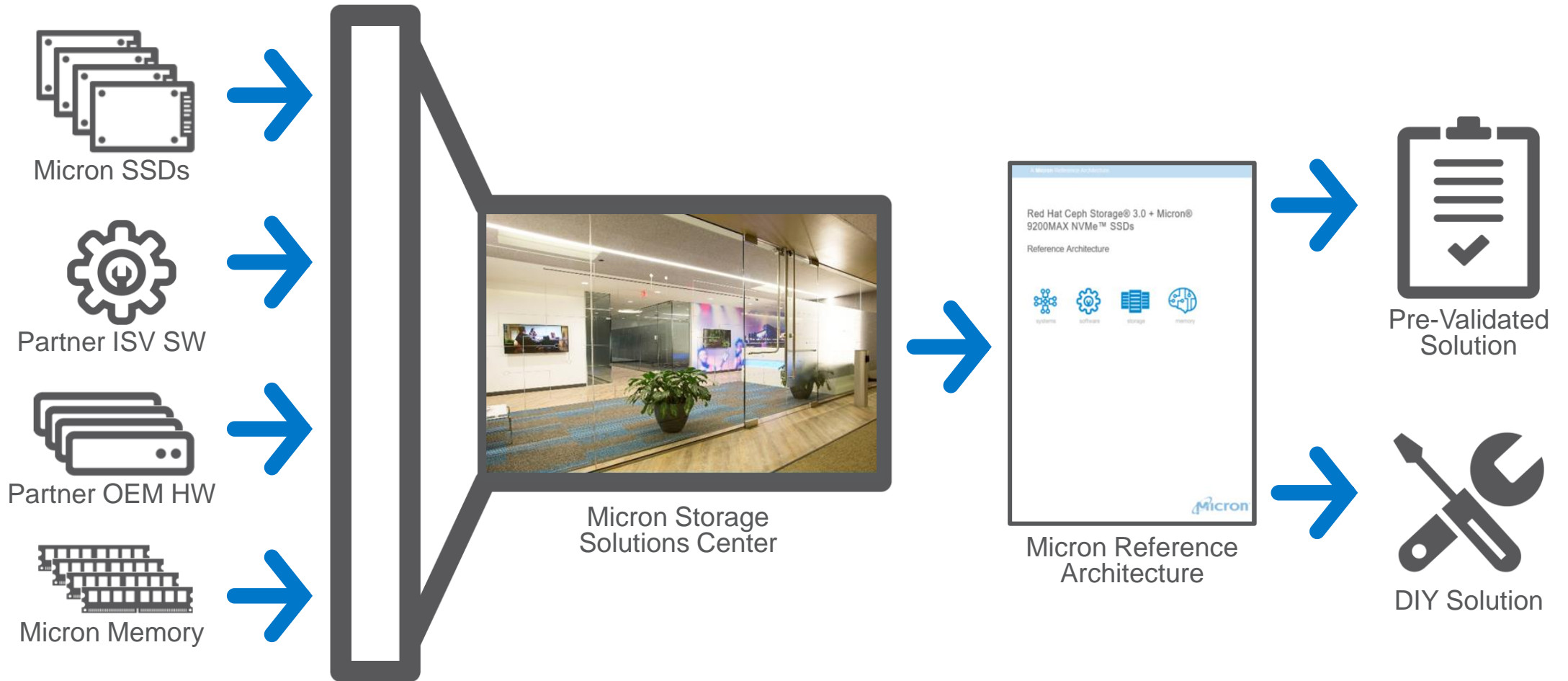
©2018 Micron Technology, Inc. All rights reserved. Information, products, and/or specifications are subject to change without notice. All information is provided on an "AS IS" basis without warranties of any kind. Statements regarding products, including regarding their features, availability, functionality, or compatibility, are provided for informational purposes only and do not modify the warranty, if any, applicable to any product. Drawings may not be to scale. Micron, the Micron logo, and all other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners.



# 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

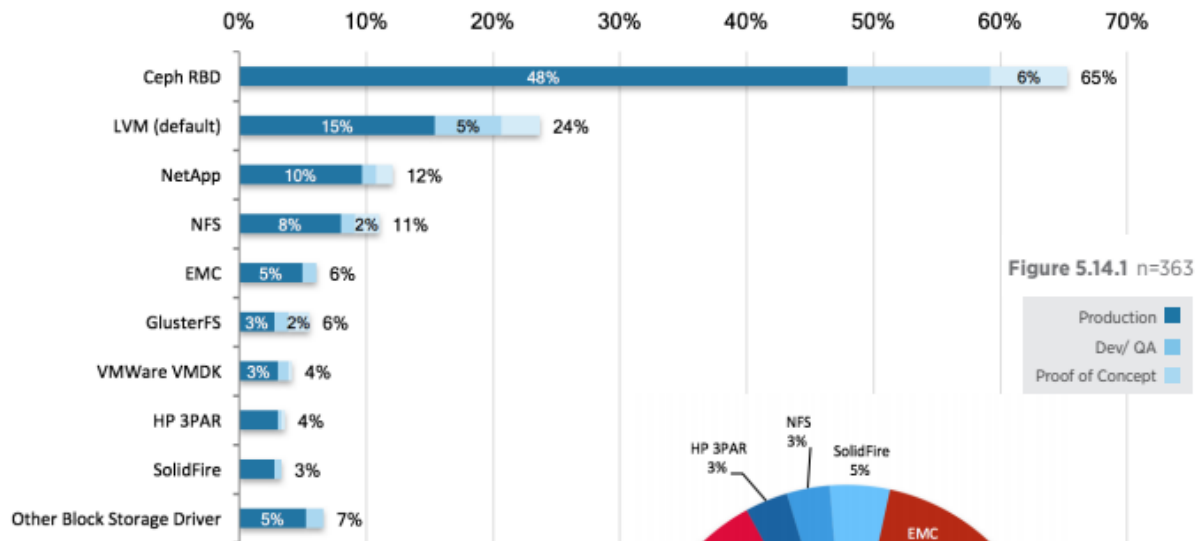
# Micron Reference Architectures



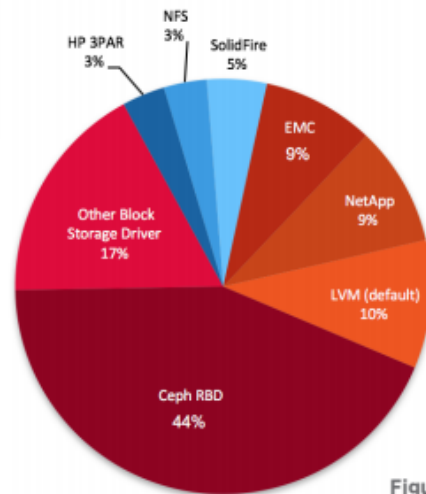


# What is Ceph?

# Ceph is the Primary Storage for OpenStack

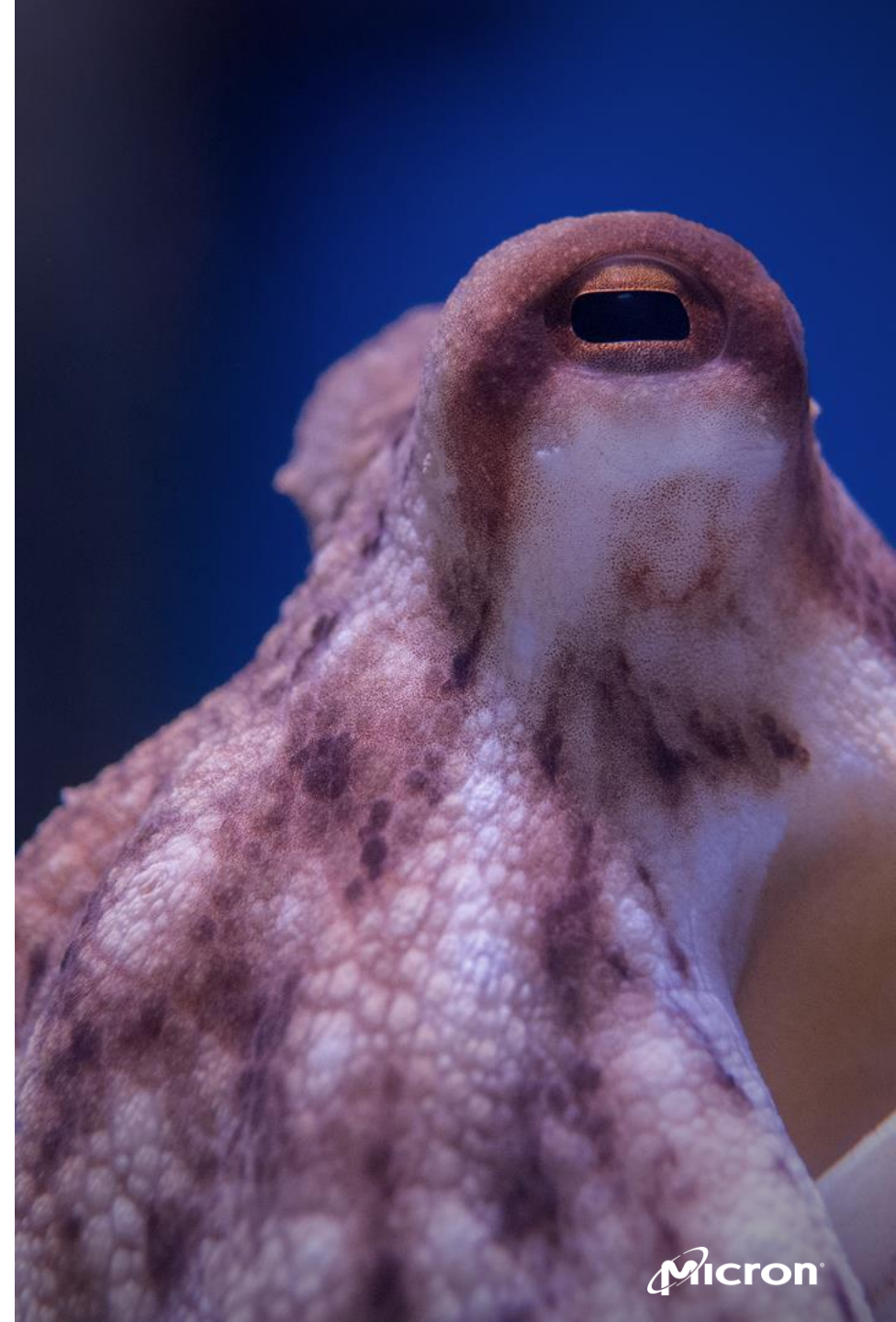


Among the largest clouds with 1,000 or more cores, Ceph RBD is still dominant, but not used by the majority, while other block storage drivers were also popular.



# 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



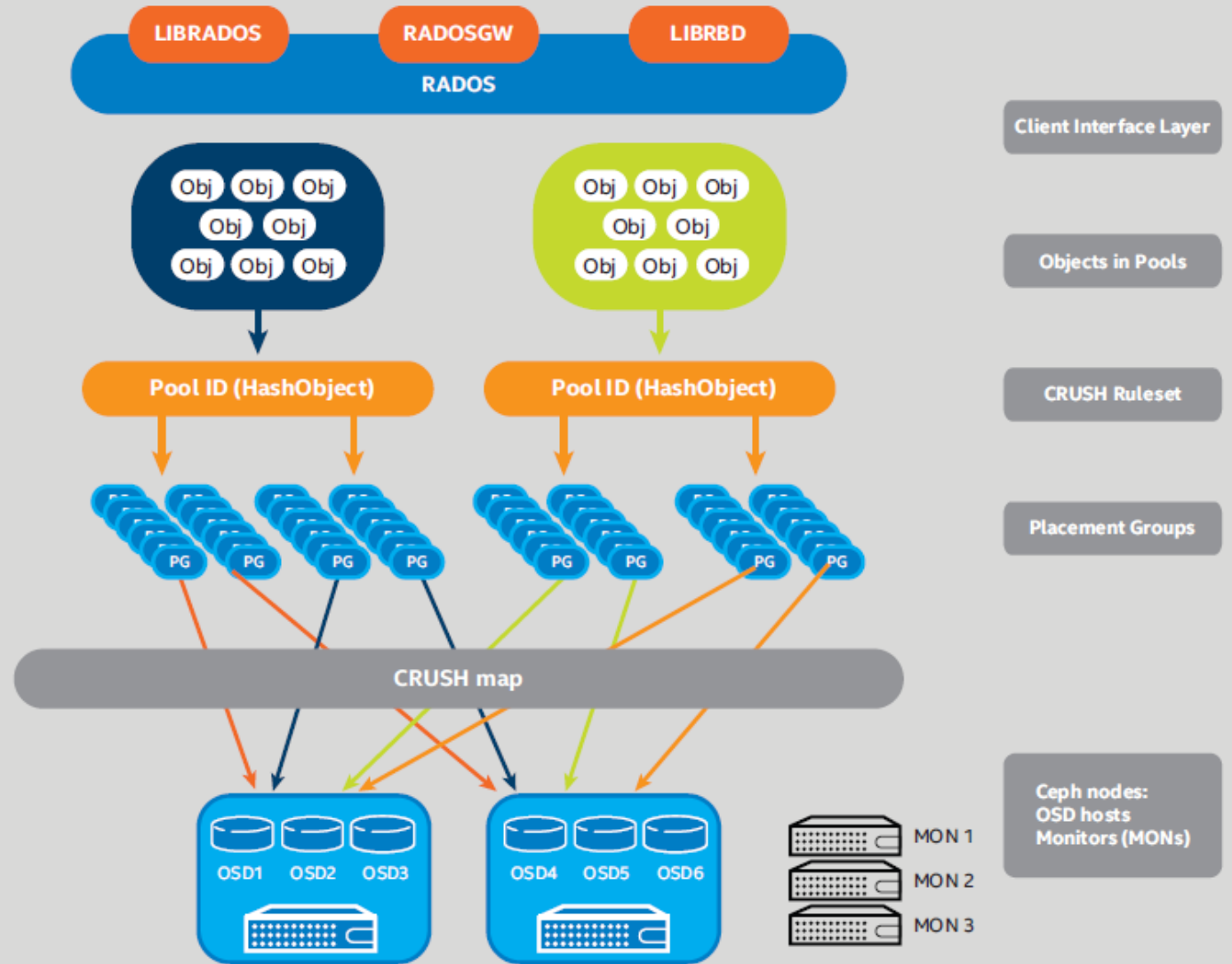
# What is Ceph?

- 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







# Ceph Luminous All-NVMe Performance



# Hardware Configuration

Micron + Red Hat + Supermicro ALL-NVMe Ceph

## Storage Nodes (x4)

- Supermicro SYS-1029U-TN10RT+
- 2x Intel 8168 24 core Xeon, 2.7Ghz Base / 3.7Ghz Turbo
- 384GB Micron High Quality Excellently Awesome DDR4-2666 DRAM (12x 32GB)
- 2x Mellanox ConnectX-5 100GbE 2-port NICs
  - 1 NIC for client network / 1 NIC for storage network
- 10x Micron 6.4TB 9200MAX NVMe SSD
  - 3 Drive Writes per Day Endurance
  - 770k 4KB Random Read IOPs / 270k 4KB Random Write IOPs
  - 3.15 GB/s Sequential Read / 2.3 GB/s Sequential Write
  - 64TB per Storage Node / 256TB in 4 node RA as tested



# Hardware Configuration

Micron + Red Hat + Supermicro ALL-NVMe Ceph

## Monitor Nodes (x3)

- Supermicro SYS-1028U-TNRT+ (1U)
  - 128 GB DRAM
  - 50 GbE Mellanox ConnectX-4

## 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

- Ceph Luminous 12.2.4
- Red Hat Enterprise Linux 7.4
- Mellanox OFED Driver 4.1

## Switch OS

- Cumulus Linux 3.4.1

## Deployment Tool

- Ceph-Ansible



# Performance Testing Methodology

Micron + Red Hat + Supermicro ALL-NVMe Ceph

- 2 OSDs per NVMe Drive / 80 OSDs total
- Ceph Storage Pool Config
  - 2x Replication: 8192 PG's, 100x 75GB RBD Images = 7.5TB data x 2
- FIO RBD for Block Tests (4KB Block size)
  - Writes: FIO at queue depth 32 while scaling up # of client FIO processes
  - Reads: FIO against all 100 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

- Ceph Luminous Community 12.2.4
  - Tested using Bluestore, a newer storage engine for Ceph
- Default RocksDB tuning for Bluestore in Ceph
  - Great for large object
  - Bad for 4KB random on NVMe
  - Worked w/ Mark Nelson & Red Hat team to tune RocksDB for good 4KB random 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_cache_kv_max = 200G
```

```
bluestore_cache_kv_ratio = 0.2
```

```
bluestore_cache_meta_ratio = 0.8
```

```
bluestore_cache_size_ssd = 18G
```

```
osd_min_pg_log_entries = 10
```

```
osd_max_pg_log_entries = 10
```

```
osd_pg_log_dups_tracked = 10
```

```
osd_pg_log_trim_min = 10
```

```
bluestore_rocksdb_options =  
compression=kNoCompression,max_write_buffer_number=64,min_write_buffer_number_to_merge=32,recycle_log_file_num=64,compaction_style=kCompactionStyleLevel,write_buffer_size=4MB,target_file_size_base=4MB,max_background_compactions=64,level0_file_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_readahead_size=2MB
```

# Ceph Luminous 12.2: 4KB Random Read

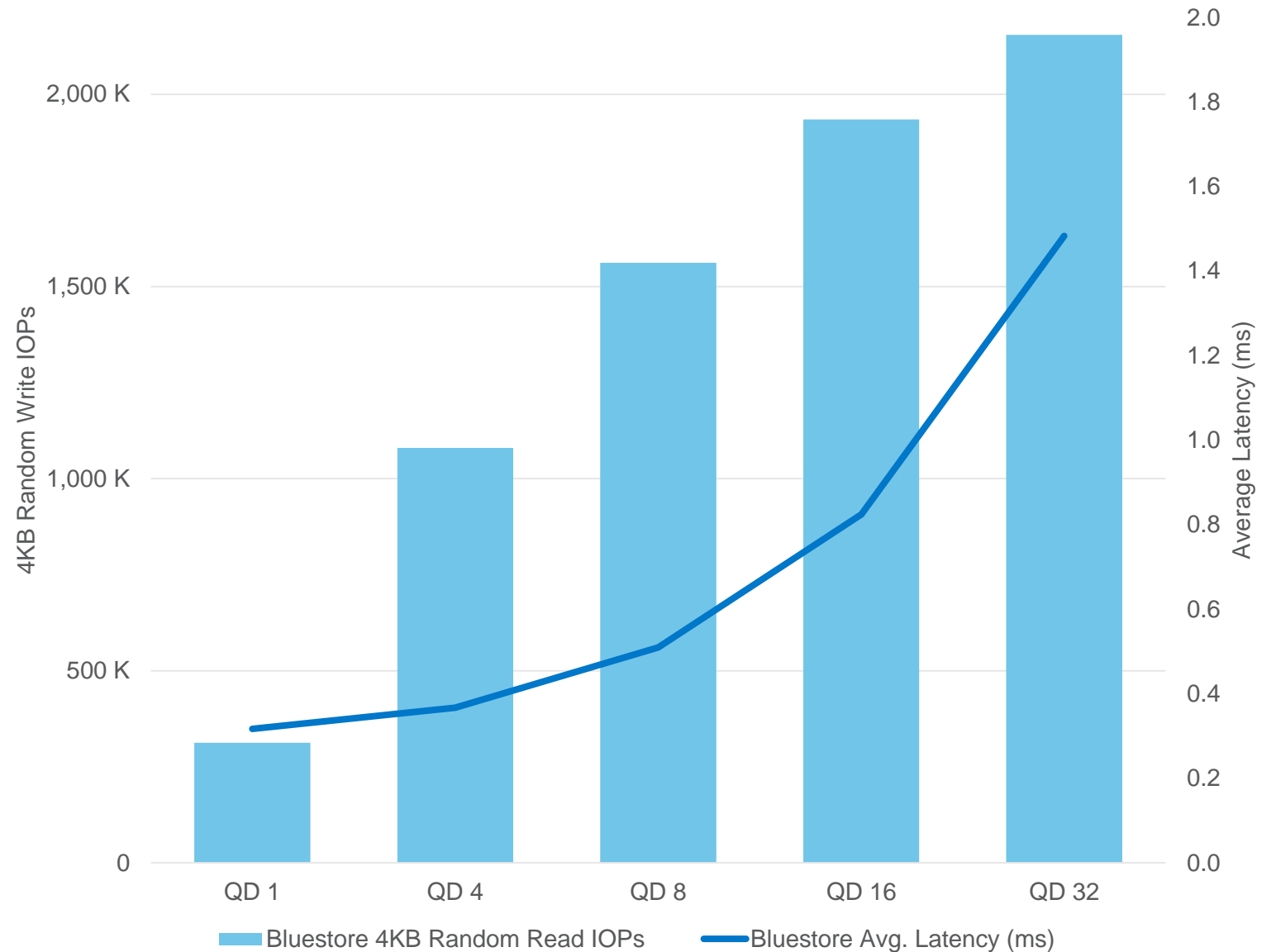
Micron + Red Hat  
+ Supermicro  
ALL-NVMe Ceph

## 4KB Random Reads:

- Queue Depth 32
  - 2.15 Million @ 1.5ms Avg. Latency

Tests become CPU  
Limited around Queue  
Depth 16

4KB Random Read: IOPs + Average Latency





# Ceph Luminous 12.2: 4KB Random Read

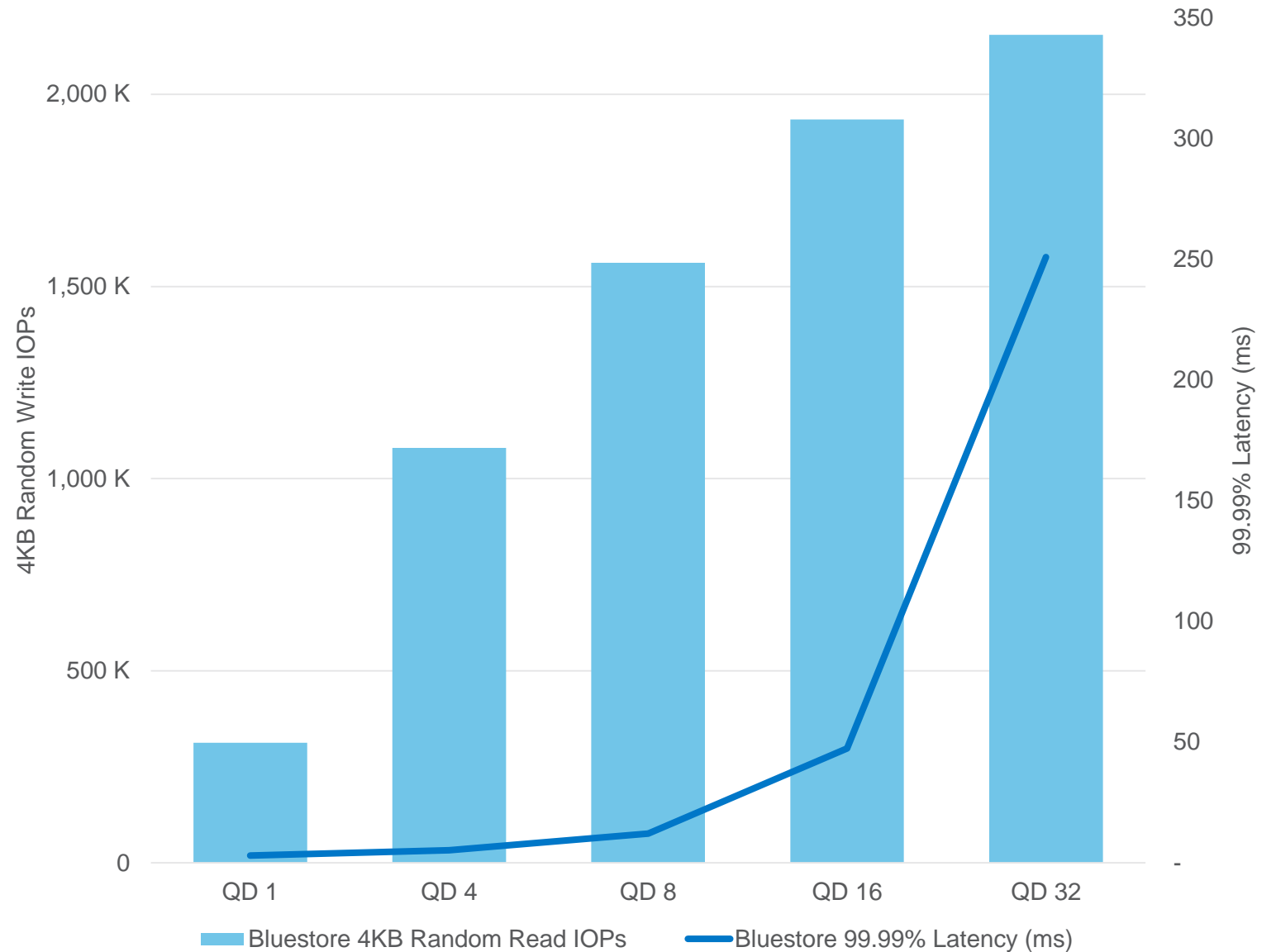
Micron + Red Hat  
+ Supermicro  
ALL-NVMe Ceph

## 4KB Random Reads:

- Queue Depth 32
  - Bluestore Tail Latency: 251 ms

Tail latency spikes as  
tests become CPU  
limited

4KB Random Read: IOPs + Tail Latency

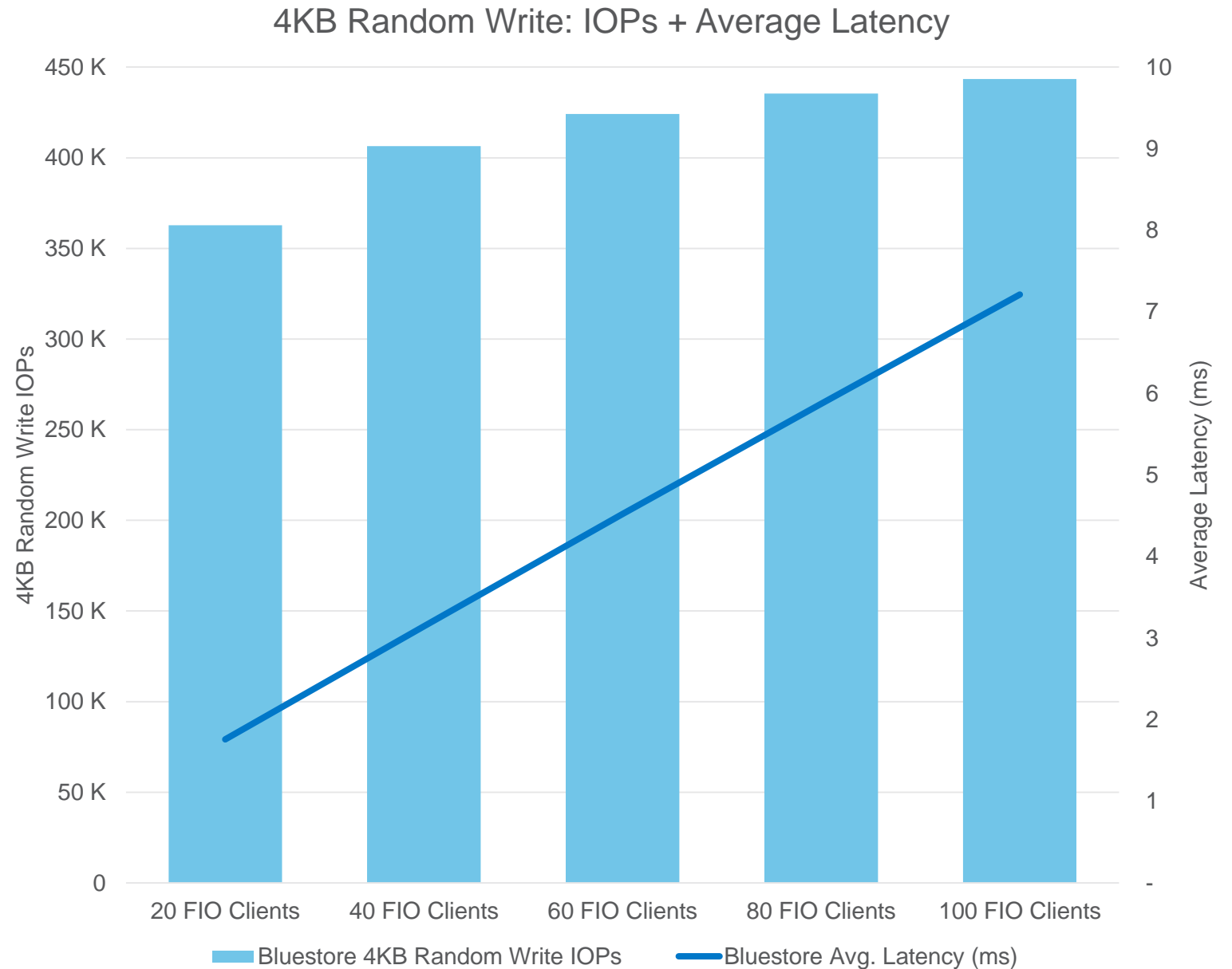


# Ceph Luminous 12.2: 4KB Random Write

Micron + Red Hat  
+ Supermicro  
ALL-NVMe Ceph

## 4KB Random Writes:

- 100 Clients
  - 443k IOPs @ 7ms

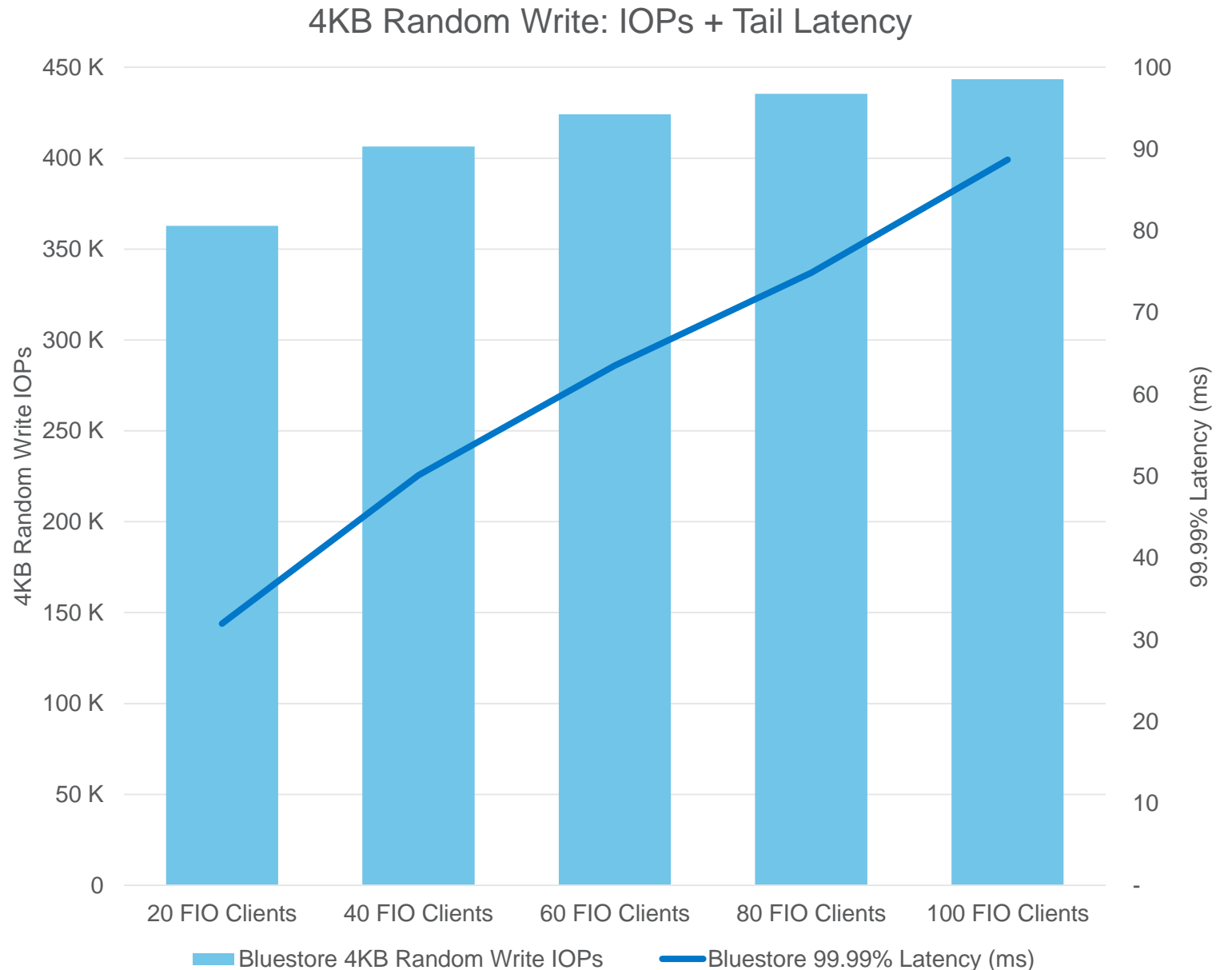


# Ceph Luminous 12.2: 4KB Random Write

Micron + Red Hat  
+ Supermicro  
ALL-NVMe Ceph

## 4KB Random Writes:

- 100 Clients
  - Tail Latency: 89ms



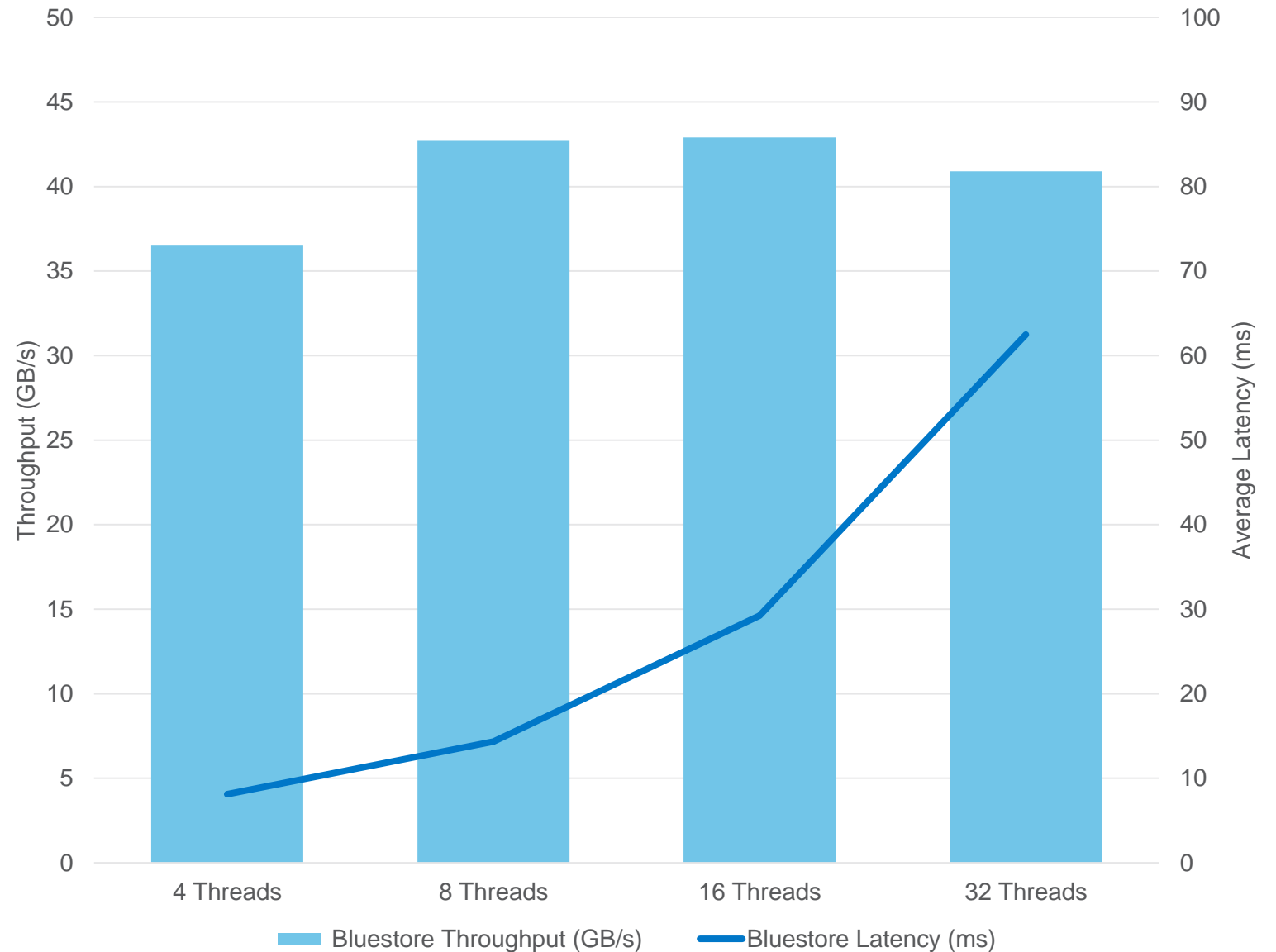
# Ceph Luminous 12.2: 4MB Object Read

Micron + Red Hat  
+ Supermicro  
ALL-NVMe Ceph

## 4KB Random Writes:

- 16 Threads:
  - 42.9 GB/s @ 29ms

4MB Object Read: Throughput + Average Latency

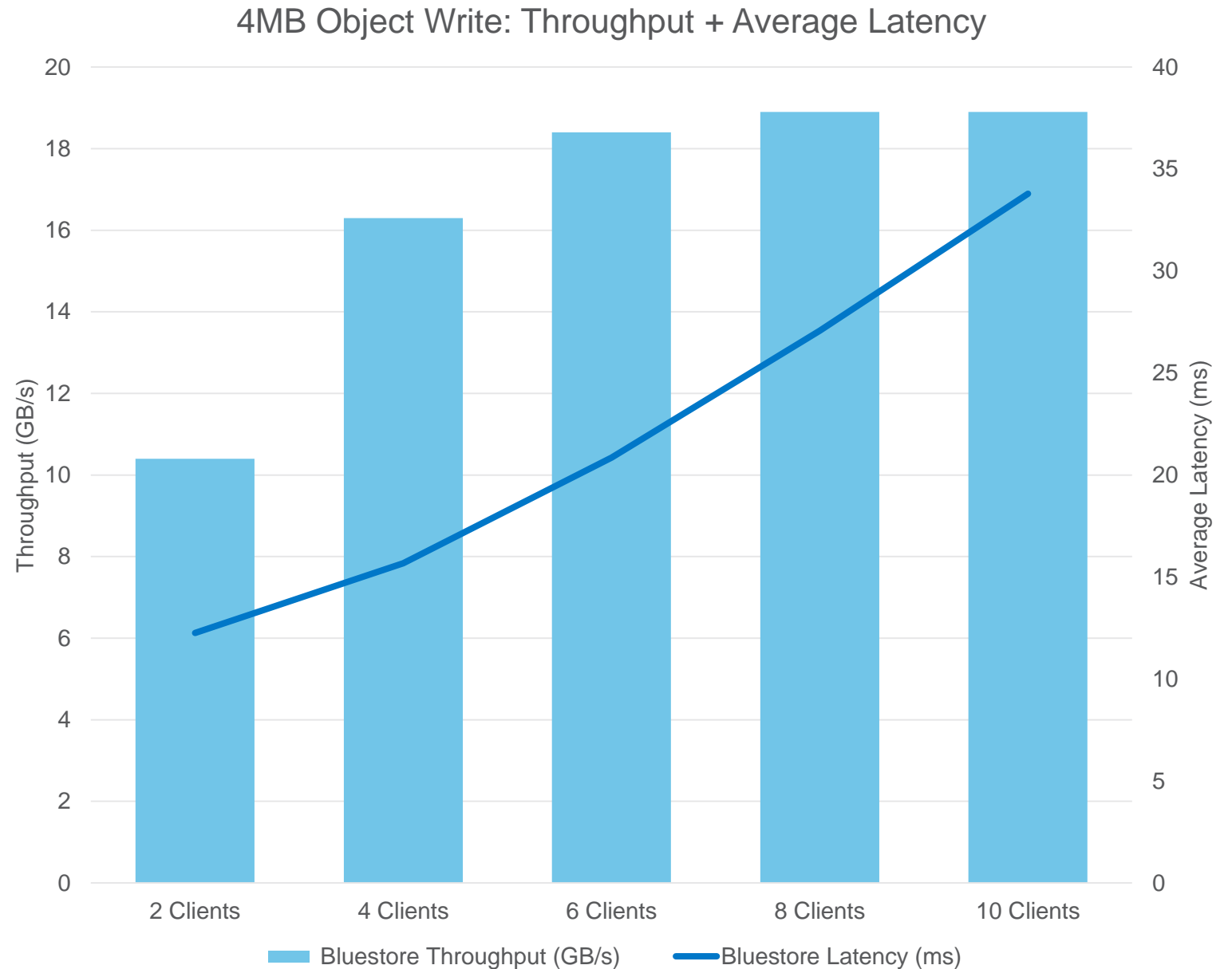


# Ceph Luminous 12.2: 4MB Object Write

Micron + Red Hat  
+ Supermicro  
ALL-NVMe Ceph

## 4MB Object Writes:

- 6 Clients:
  - 18.4 GB/s @ 21ms



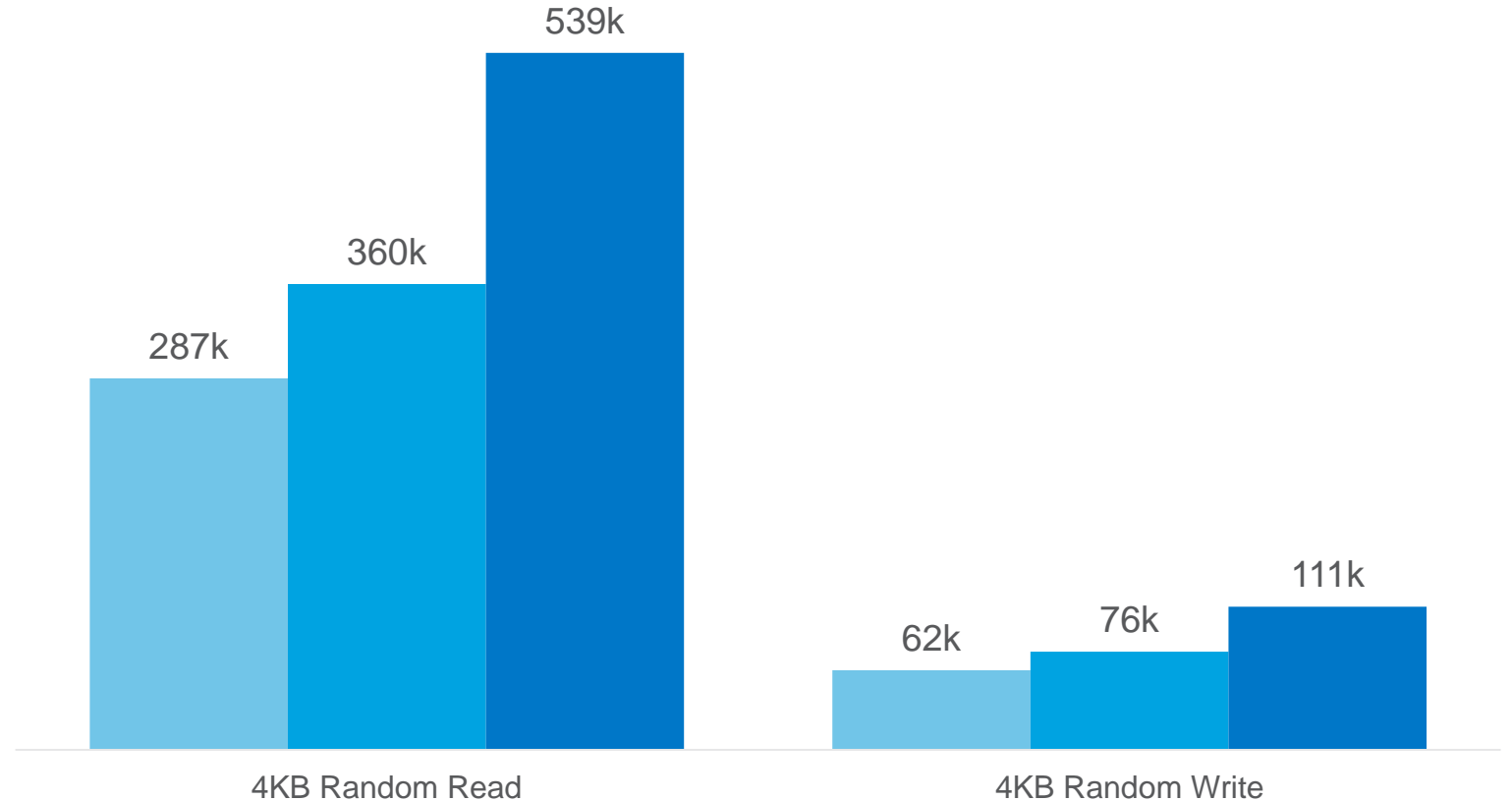
# Performance Comparison: 4KB Random Block

2017 Micron RA vs. 2017 Competitor vs. 2018 Micron + Bluestore

## 4KB Random IOPs

- 4KB Random Reads
  - 1.75X IOPs
- 4KB Random Writes
  - 1.9X IOPs

4KB Random IO Performance / Node



■ 2017 Micron Ceph RA Filestore RHCS 2.1/9100MAX

■ 2018 Micron Ceph RA Bluestore Luminous 12.2/9200MAX

■ 2017 Competitor RA Bluestore Luminous 12.0/P4800X+P4500 NVMe

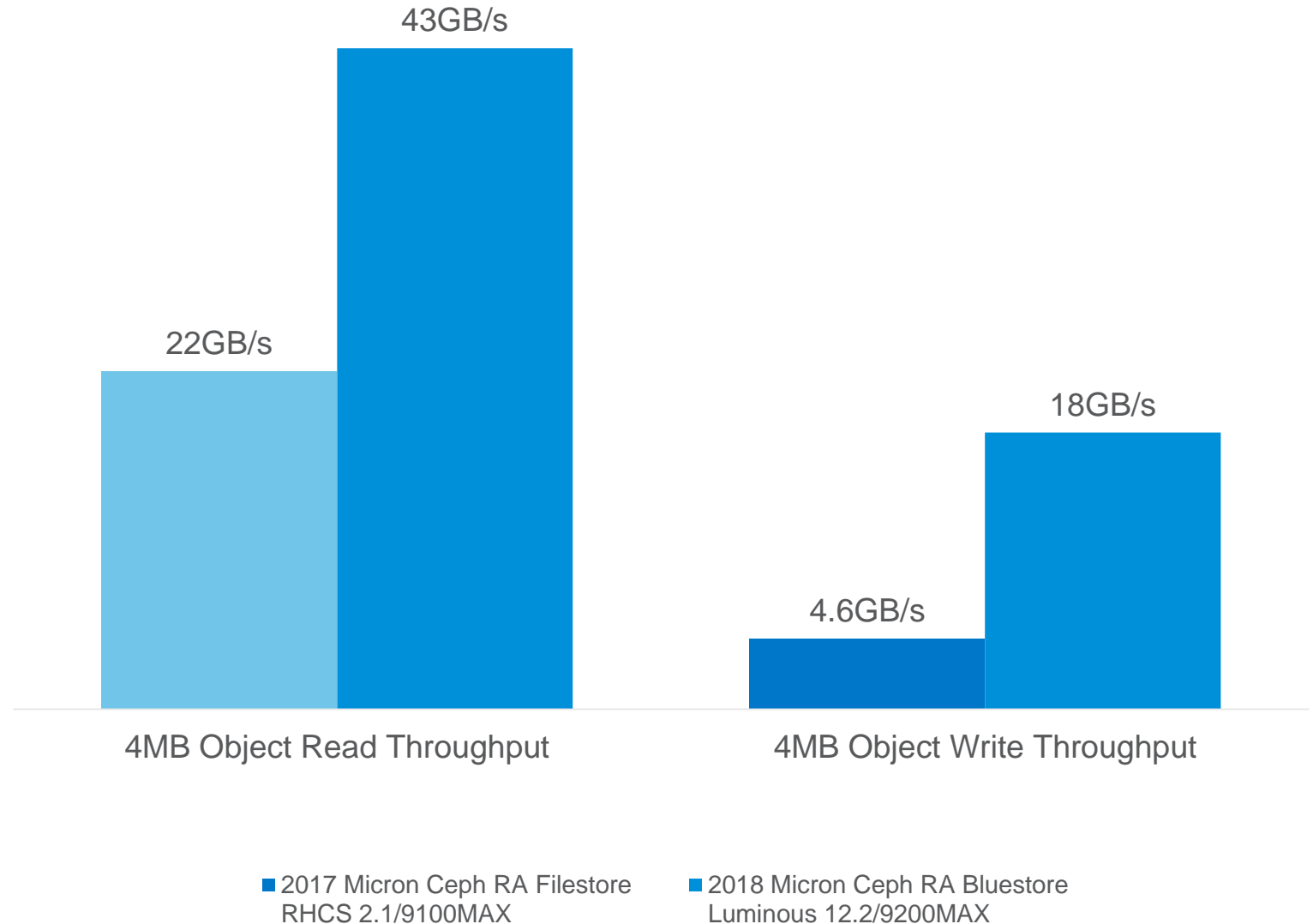
# Performance Comparison: 4MB Object

2017 Micron RA vs. 2018 Micron RA + Bluestore

## 4MB Object Throughput

- 1.95X Read Throughput
- 3.9X Write Throughput

4MB Object Performance: RADOS Bench





# Ceph Luminous QLC + NVMe Performance





# Hardware Configuration

QLC + NVMe Ceph

## Storage Nodes (x3)

- Supermicro SYS-1029U-TR25M
- 2x Intel 6148 20-core Xeon, 2.4Ghz Base / 3.7Ghz Turbo
- 384GB Micron High Quality Excellently Awesome DDR4-2666 DRAM (12x 32GB)
- 25GbE Network
  - 1 port for client network / 1 port for storage network
- 8x Micron 7.8TB 5210 ION QLC SSD
  - ~64TB per Storage Node / 192TB in 3 node cluster as tested
- 2x Micron 1.6TB 9200MAX NVMe SSD
  - Used for WAL + RocksDB instances (Write coalescing + acceleration)



# Performance Testing Methodology

Micron + Red Hat + Supermicro ALL-NVMe Ceph

- 2 OSDs per 5210 / 48 OSDs total
- Ceph Storage Pool Config
  - 2x Replication: 8192 PG's, 50x 115GB RBD Images = 5.8TB data x 2
- FIO RBD for Block Tests (16KB Block Size)
  - Mixed Workloads + 100% reads: FIO against all 50 RBD Images, scaling up QD
- RADOS Bench for Object Tests (1MB Object Size)
  - Reads & Writes: 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 Luminous:

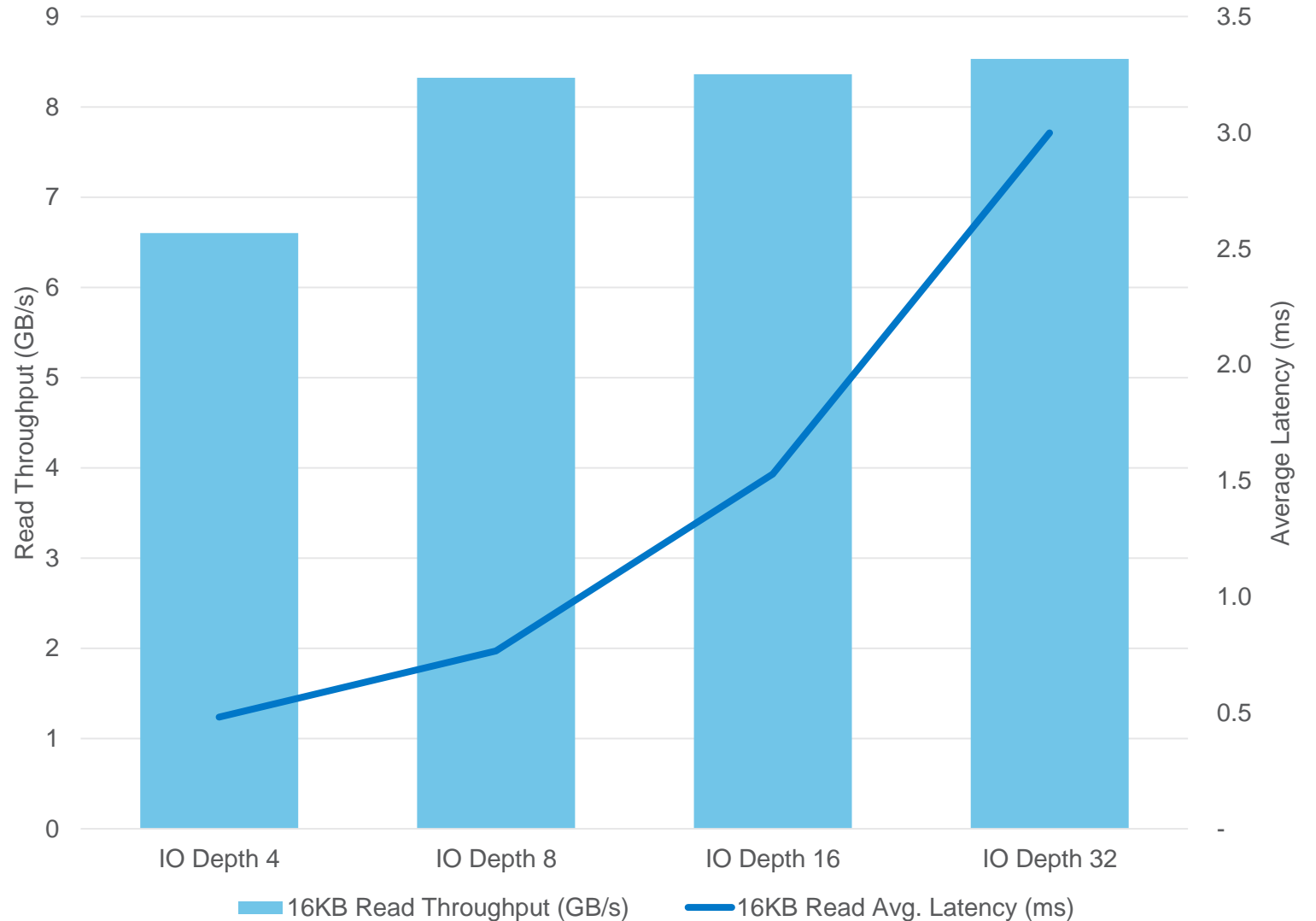
## 16KB 100% Read

Micron QLC + NVMe

### 16KB Reads:

- 8+ GB/s at QD 8
  - Network limited on 3x 25GbE
- Sub-millisecond average latency

16KB Read Performance  
Ceph Luminous 12.2.5  
Micron 5210 + 9200MAX



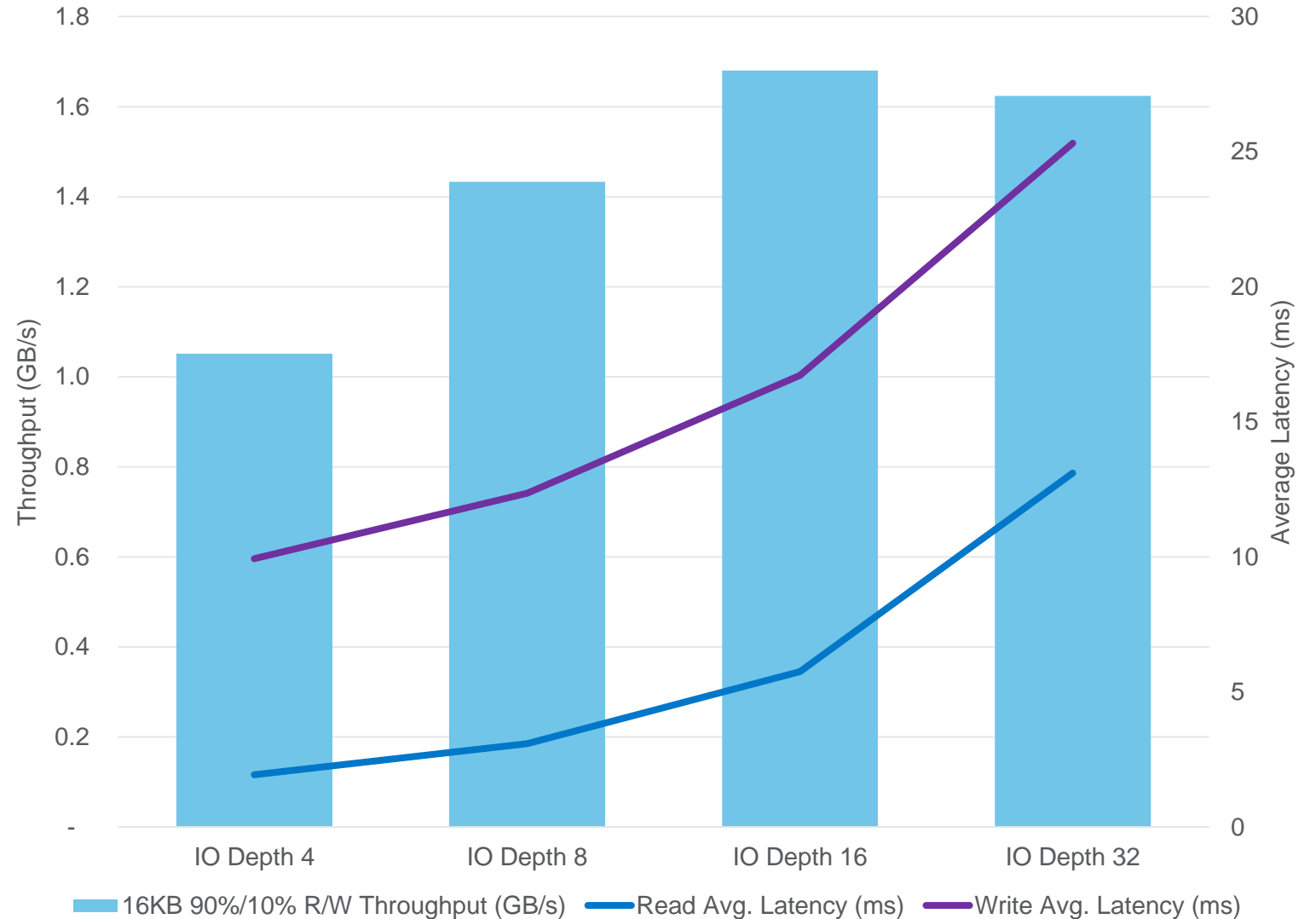
# Ceph Luminous: 16KB 90%/10% R/W

Micron QLC + NVMe

## 16KB 90%/10% Reads/Writes:

- 1.7 GB/s at QD 16
  - Drive Limited
- 5.8ms Average Read Latency
- 16.7ms Average Write Latency

16KB 90%/10% Read/Write Performance  
Ceph Luminous 12.2.5  
Micron 5210 + 9200MAX



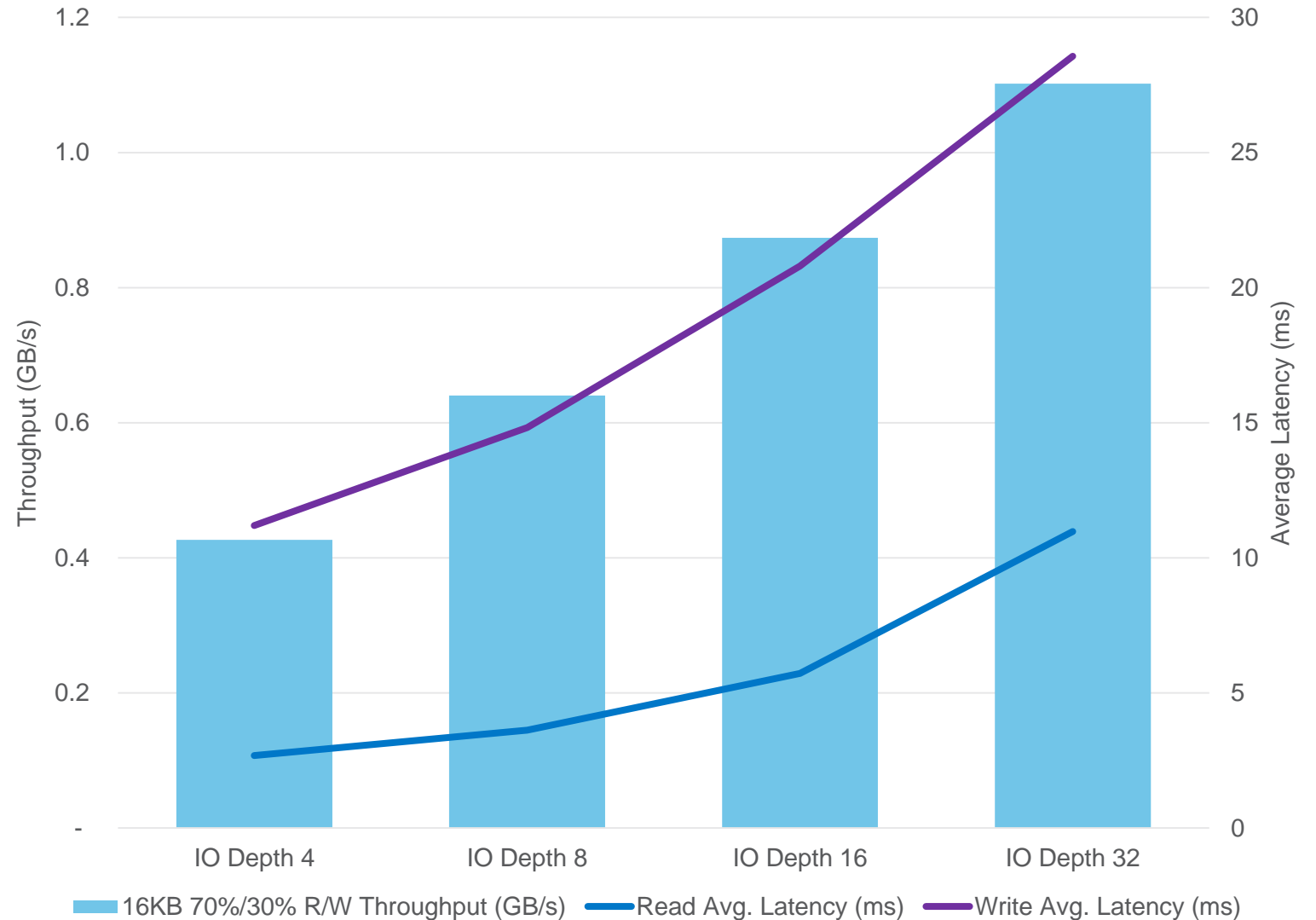
# Ceph Luminous: 16KB 70%/30% R/W

Micron QLC + NVMe

## 16KB 70%/30% Reads/Writes:

- 1.1 GB/s at QD 32
  - Drive Limited
- 11ms Average Read Latency
- 28ms Average Write Latency

16KB 70%/30% Read/Write Performance  
Ceph Luminous 12.2.5  
Micron 5210 + 9200MAX



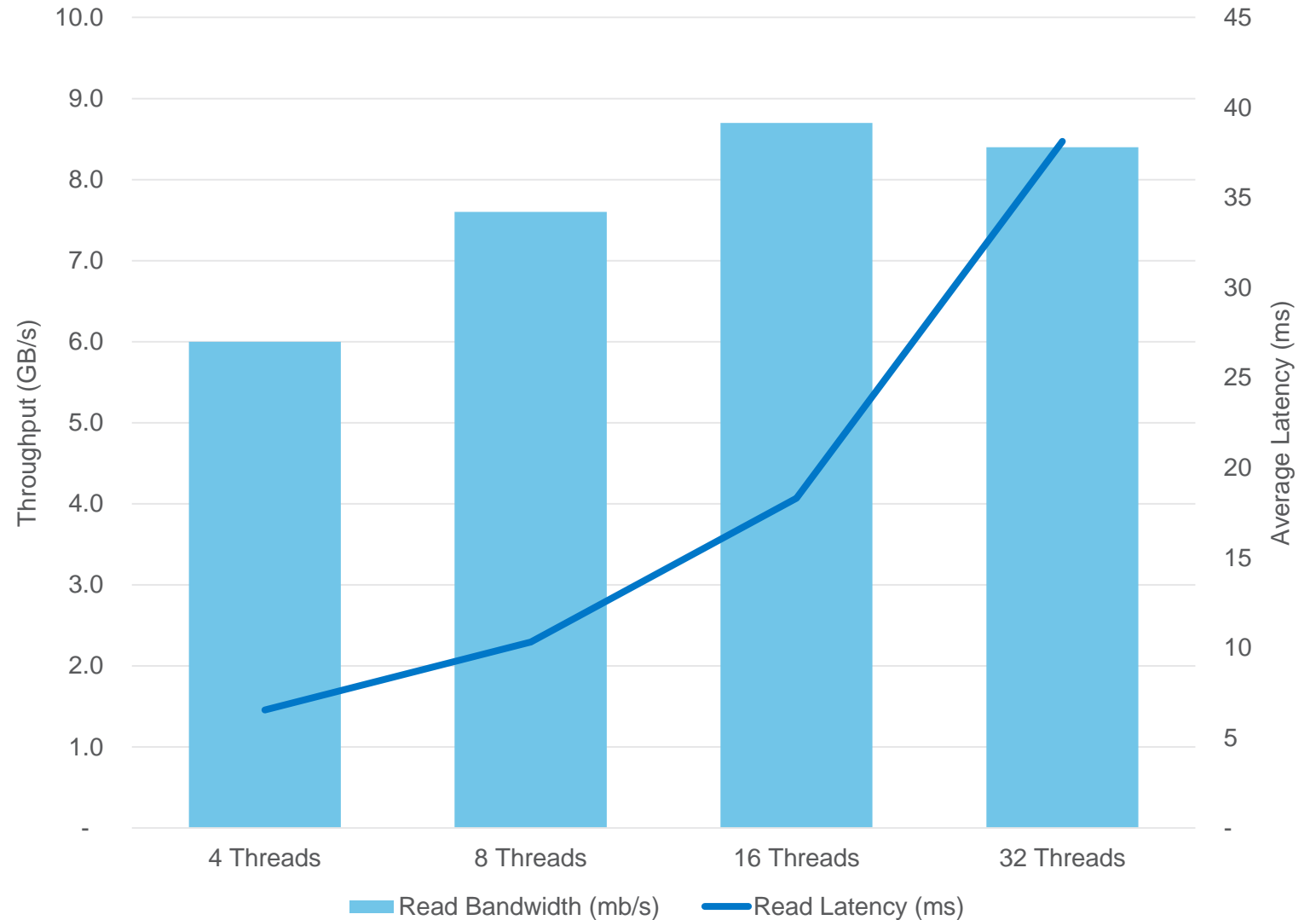
# Ceph Luminous 12.2: 1MB Object Read

Micron QLC + NVMe

## 1MB Object Reads:

- 8+ GB/s at 16 Threads
  - Network limited on 3x 25GbE
- 18ms average latency

1MB Object Read Performance  
Ceph Luminous 12.2.5  
Micron 5210 + 9200MAX



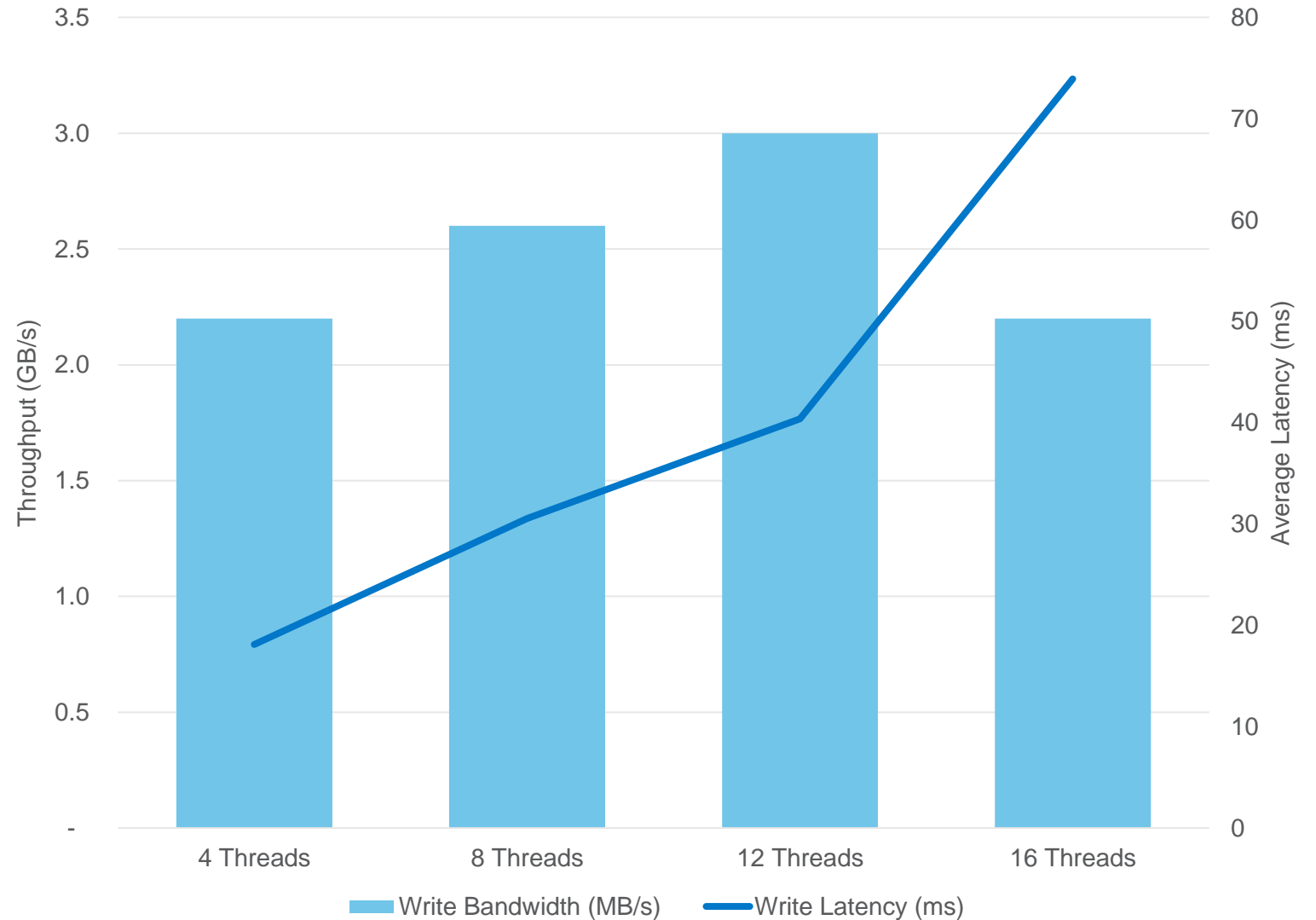
# Ceph Luminous 12.2: 1MB Object Write

Micron QLC + NVMe

## 1MB Object Writes:

- 3+ GB/s at 12 Threads
  - Drive Limited
- 40ms average latency

1MB Object Write Performance  
Ceph Luminous 12.2.5  
Micron 5210 + 9200MAX



# Would you like to know more?

Today 8/7

1:50pm Micron Keynote – Derek Dicker

3:40pm QLC is the Best Way to Replace Enterprise HDDs

3:40pm Flash-Memory Based Architectures

3:40pm New Directions in Security

4:55pm Meeting the Storage Needs of 5G Networks

4:55pm Encryption for Data Protection

7:00pm Micron Mixer in the Terra Courtyard


Wednesday 8/8

8:30am New Flexible Form Factors for Enterprise SSDs

Thursday 8/9

2:10pm The Next Great Breakthrough in NAND Flash





# Would you like to know more?

Micron NVMe Reference Architecture:

[https://www.micron.com/~media/documents/products/technical-marketing-brief/micron\\_9200\\_ceph\\_3,-d-,0\\_reference\\_architecture.pdf](https://www.micron.com/~media/documents/products/technical-marketing-brief/micron_9200_ceph_3,-d-,0_reference_architecture.pdf)

Micron Storage Blogs: Ceph

<https://www.micron.com/about/blogs/authors/ryan-meredith>

A large, detailed octopus is the central focus, set against a dark, atmospheric underwater background. The octopus's tentacles are spread out, some curled and some reaching towards the viewer. The lighting is dramatic, highlighting the texture of the octopus's skin and the suction cups on its tentacles. The overall mood is mysterious and serene.

# Thanks All

# Micron Ceph Appendix

## ALL-NVMe Ceph RA: Bluestore Ceph.conf

```
[global]
auth client required = none
auth cluster required = none
auth service required = none
auth supported = none
mon host = xxxxxxxx
osd objectstore = bluestore
cephx require signatures = False
cephx sign messages = False
mon_allow_pool_delete = true
mon_max_pg_per_osd = 800
mon_pg_warn_max_per_osd = 800
ms_crc_header = False
ms_crc_data = False
ms_type = async
perf = True
rocksdb_perf = True
osd_pool_default_size = 2

[mon]
mon_max_pool_pg_num = 166496
mon_osd_max_split_count = 10000

[client]
rbd_cache = false
rbd_cache_writethrough_until_flush = false

[osd]
bluestore_csum_type = none
bluestore_cache_kv_max = 200G
bluestore_cache_kv_ratio = 0.2
bluestore_cache_meta_ratio = 0.8

bluestore_cache_size_ssd = 18G
bluestore_extent_map_shard_min_size = 50
bluestore_extent_map_shard_max_size = 200
bluestore_extent_map_shard_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

bluestore_rocksdb_options =
compression=kNoCompression,max_write_buffer_
number=64,min_write_buffer_number_to_merge=3
2,recycle_log_file_num=64,compaction_style=k
CompactionStyleLevel,write_buffer_size=4MB,t
arget_file_size_base=4MB,max_background_comp
actions=64,level0_file_num_compaction_trigge
r=64,level0_slowdown_writes_trigger=128,leve
l0_stop_writes_trigger=256,max_bytes_for_lev
el_base=6GB,compaction_threads=32,flusher_th
reads=8,compaction_readahead_size=2MB
```

The Micron logo features a stylized white 'M' with two white elliptical orbits around it, positioned to the left of the word 'micron' in a white, lowercase, sans-serif font. A registered trademark symbol (®) is located at the top right of the word.

**micron**®