
# Unleashing the Power of Flash in Ceph DataStores:

## An All-NVMe™ Ceph Performance Deep Dive

Ryan Meredith

Micron Principal Solutions Engineer

# Micron Storage Solutions Engineering

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

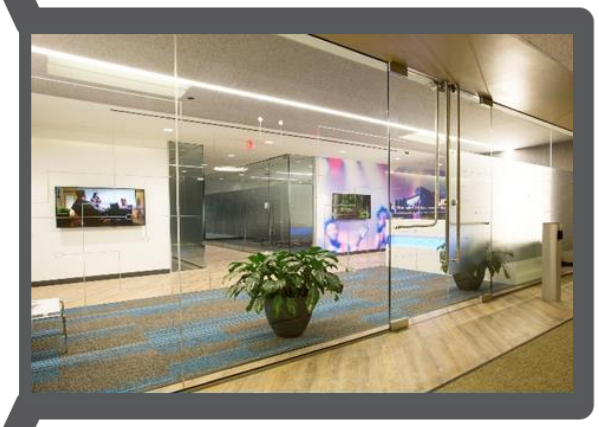
# How Micron Accelerated Solutions Are Born

Micron SSDs

Partner ISV SW

Partner OEM HW

Micron Memory



Micron Storage Solutions Center


Micron Reference Architecture

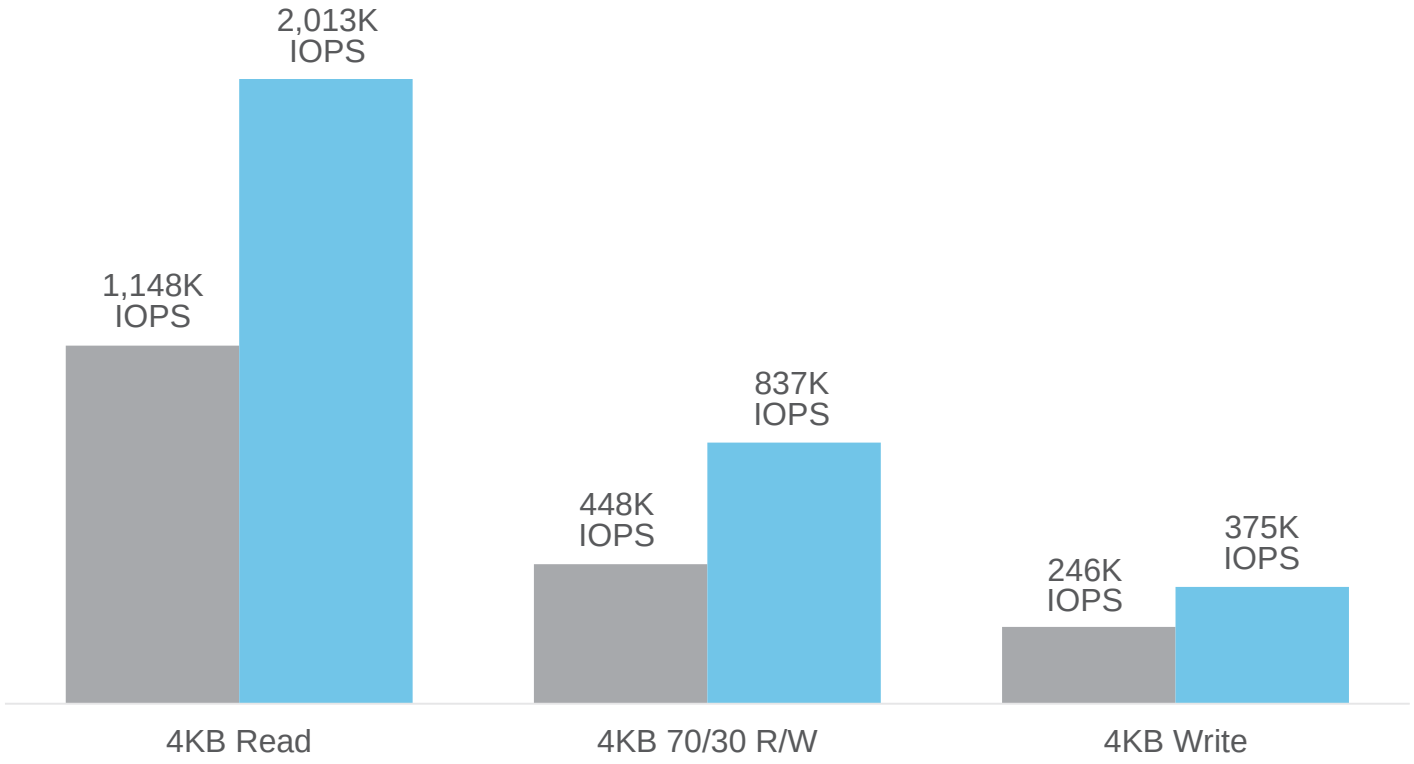
Pre-Validated Solution

DIY Solution

# Micron Accelerated Solutions

Micron
+ Red Hat
+ Supermicro
ALL-NVMe
Ceph RA

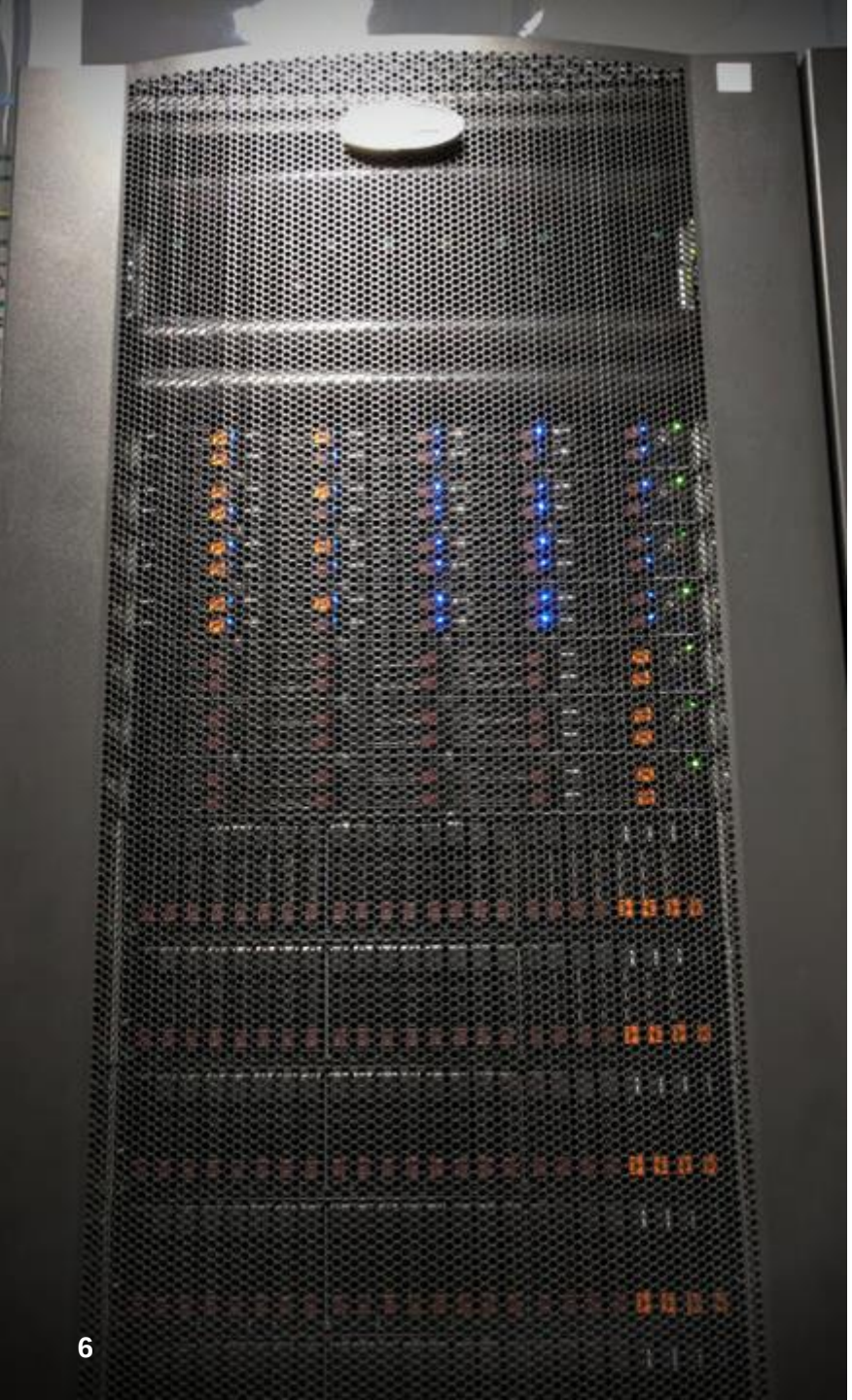
## Micron + Red Hat Ceph Storage Reference Architectures




**2017 Micron Ceph RA**
Intel Broadwell 2699v4
256GB DRAM
50 GbE Networking
RHCS 2.1 (Jewel 10.2.3)

**2018 Micron Ceph RA**
Intel Purley 8168
384GB DRAM
100 GbE Networking
RHCS 3.0 (Luminous 12.2.1)

# Micron + Red Hat + Supermicro All-NVMe Ceph Reference Architecture 2018 Edition

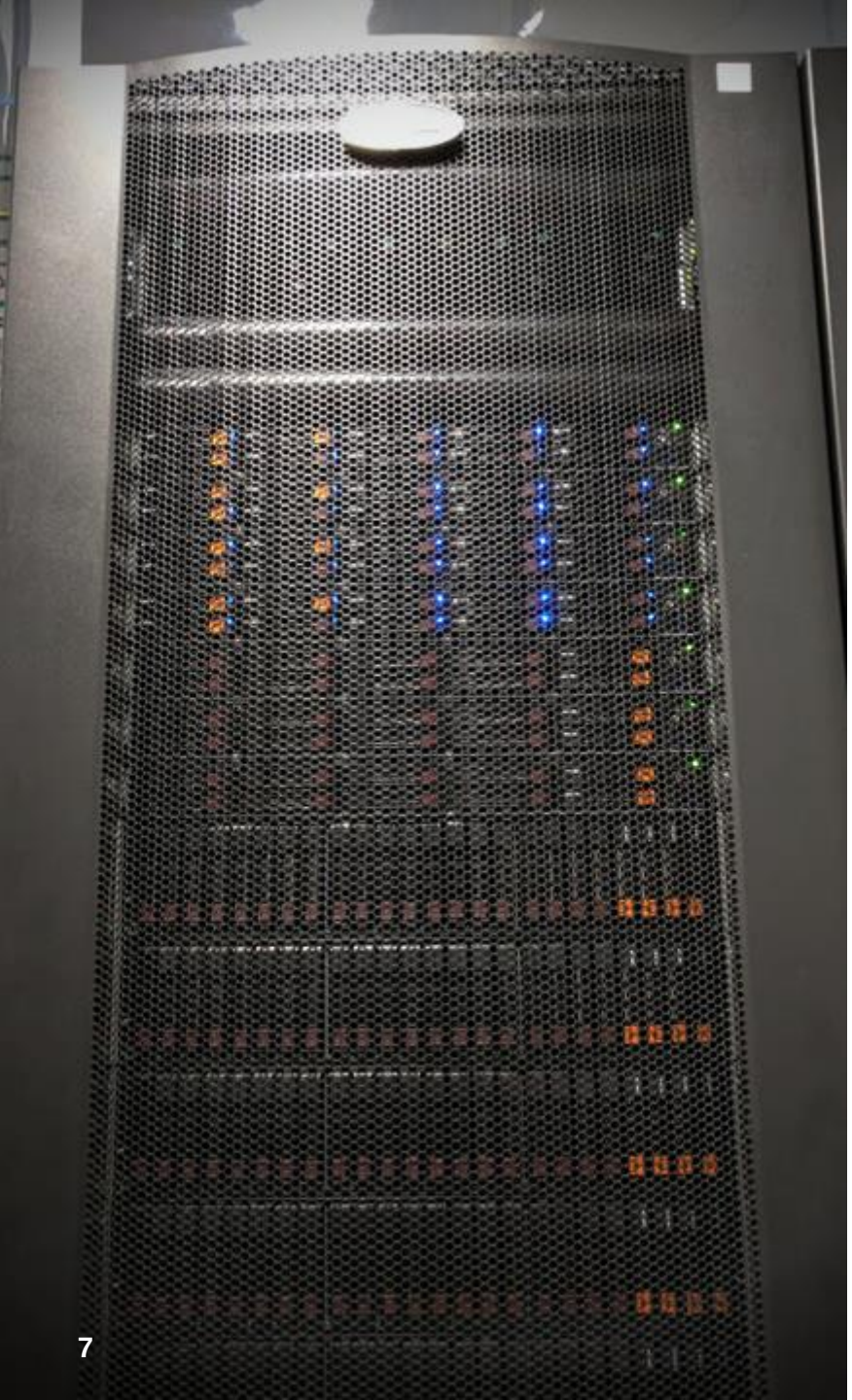


# Hardware Configuration

Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

## 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 RA

## 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 RA

## Storage + Monitor Nodes + Clients

- Red Hat Ceph Storage 3.0 (Luminous 12.2.1)
- 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 RA

- 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
  - 3x Replication: 8192 PG's, 100x 50GB RBD Images = 5TB x 3
- FIO RBD for Block Tests
  - 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
  - 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)

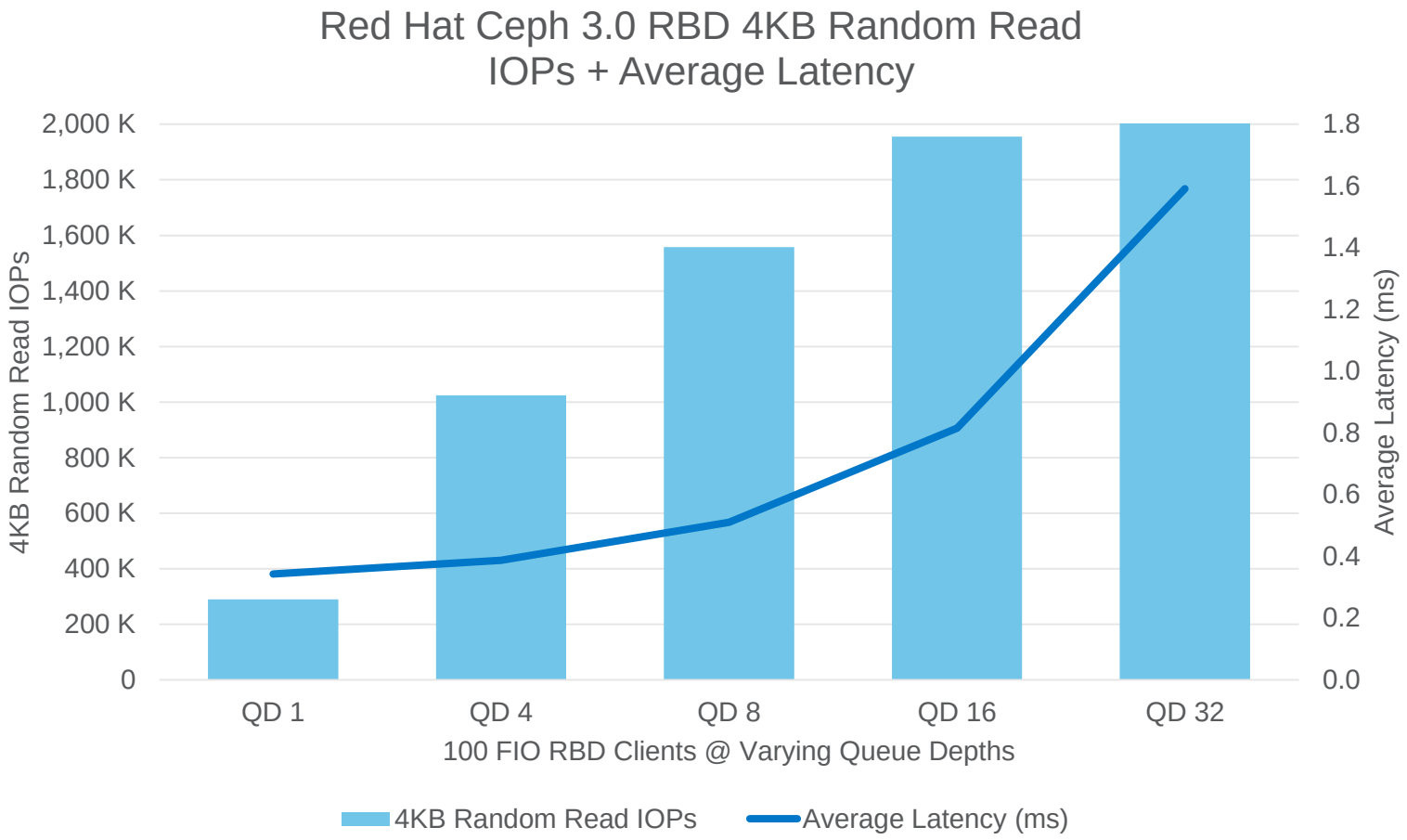
# FIO RBD 4KB Random Read Performance

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random ReadPerformance:

- 2 Million IOPs @ 1.6ms Avg. Latency
- 1.96 Million IOPs @ 0.8ms Avg. Latency

### CPU Limited



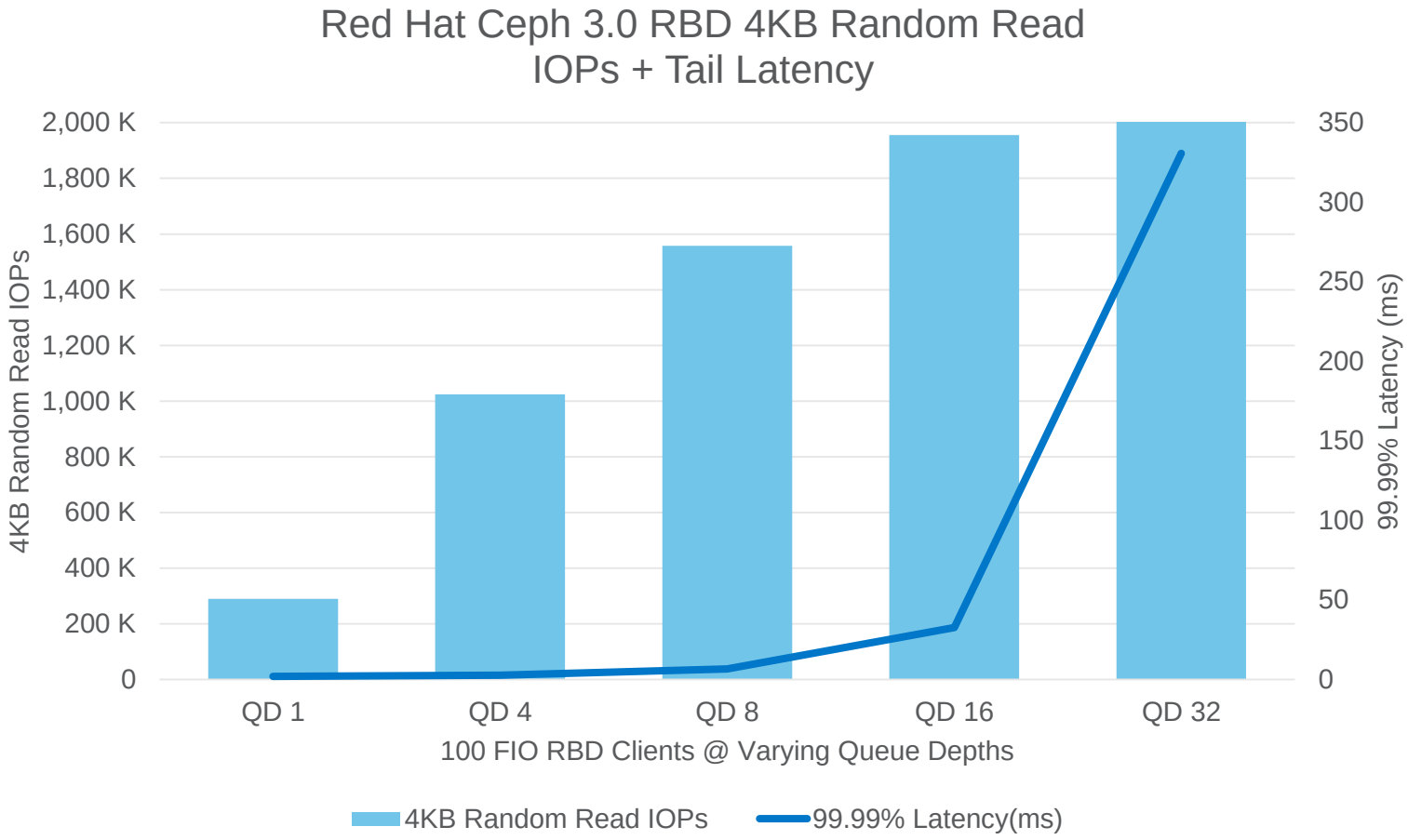
# FIO RBD 4KB Random Read Performance

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Read Performance:

- 1.96 Million 4KB Random Reads @ 33ms 99.99% Latency
- Tail latency spikes above Queue Depth 16

### CPU Limited



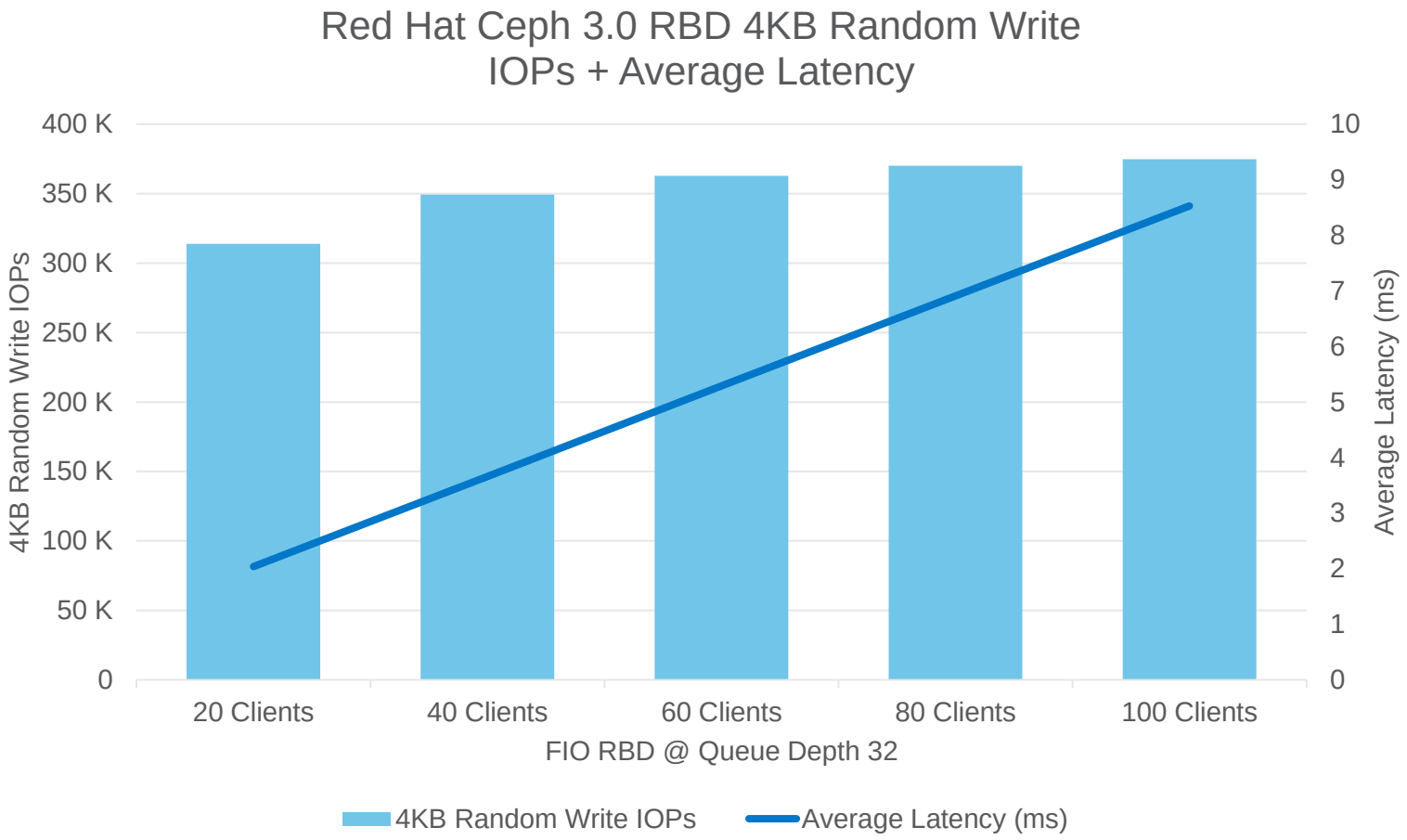
# FIO RBD 4KB Random Write Performance

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Write Performance:

- 375k IOPs @ 8.5ms Avg. Latency (100 clients)
- 363k IOPs @ 5.3ms Avg. Latency (60 Clients)

### CPU Limited



# FIO RBD 4KB Random Write Performance

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Write Performance:

- 362k 4KB Random Writes @ 50ms 99.99% Latency

### Tail latency spikes above 70 clients

### CPU Limited

# Rados Bench 4MB Object Read Performance

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

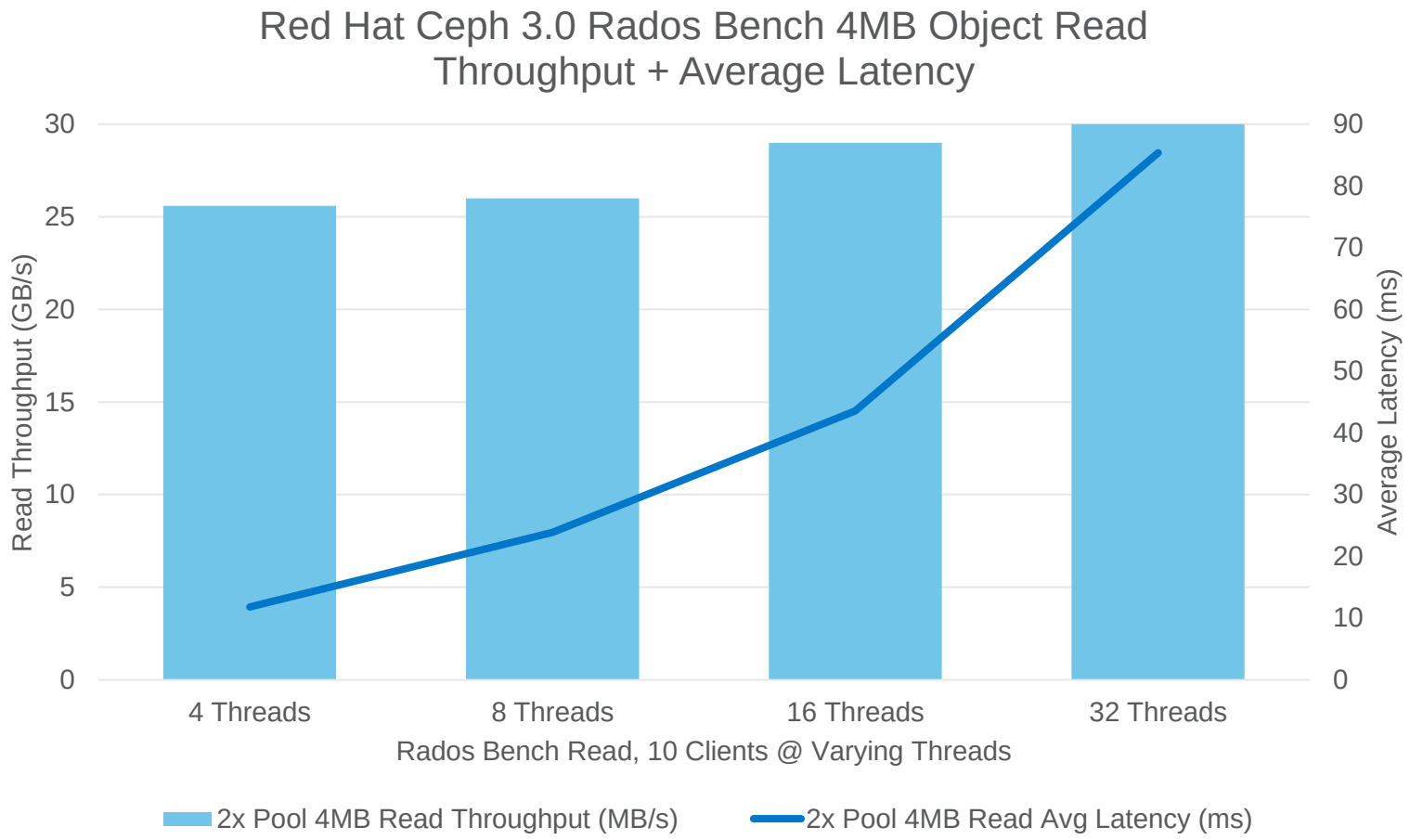
4MB Object Read Performance:

- 30 GB/s @ 85ms (32 Threads)
- 29 GB/s @ 44ms (16 Threads)

Ceph RA is tuned for small block performance

Object Read is Software Limited

Should be possible to tune ceph.conf for higher object read performance



# Rados Bench 4MB Object Write Performance

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

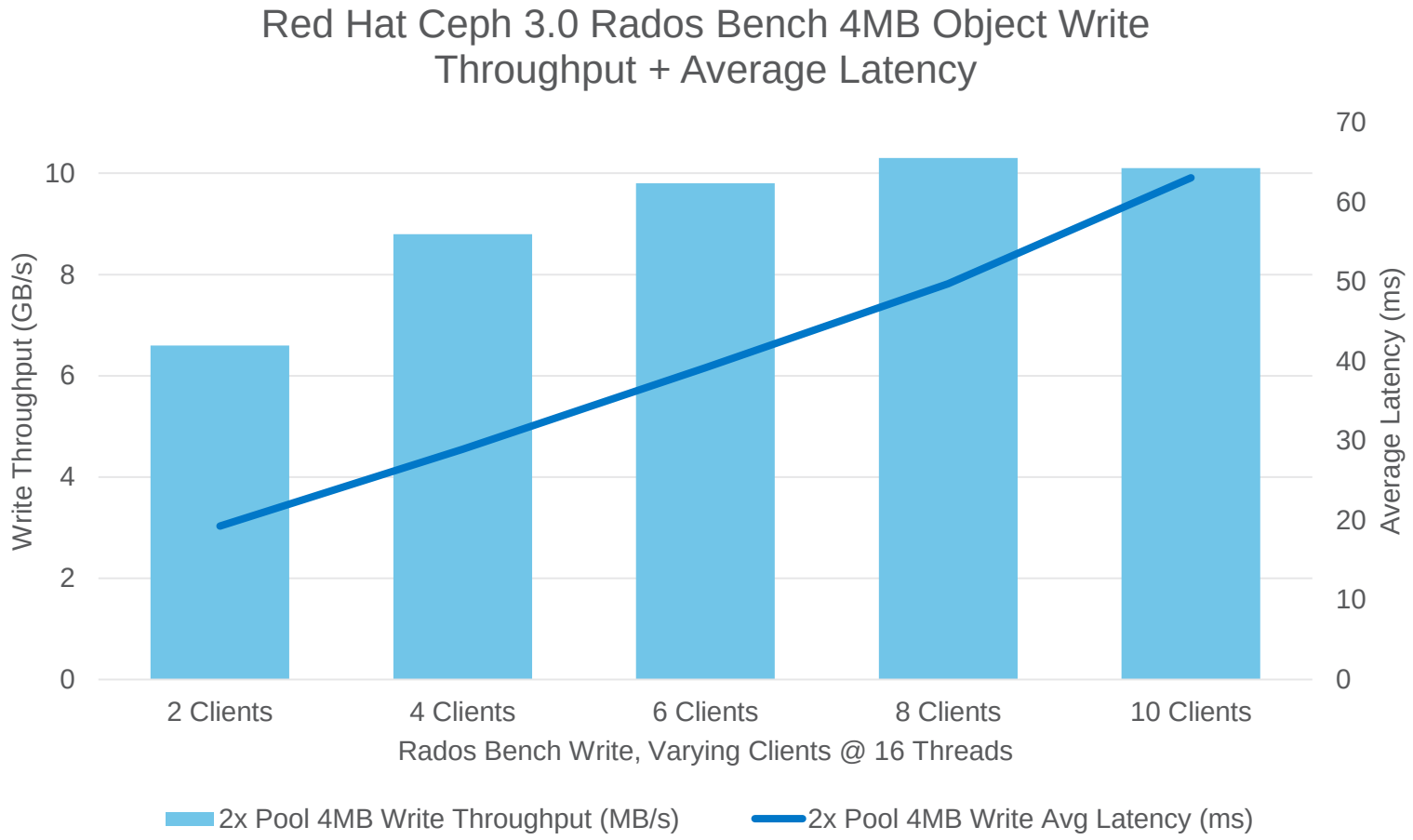
4MB Object Write Performance:

- 10.3 GB/s @ 50ms (8 Clients)
- 9.8 GB/s @ 39ms (6 Clients)

Ceph RA is tuned for small block performance

Object Write is Software Limited

May be possible to tune ceph.conf for higher object write performance



# Bluestore vs. Filestore

# Bluestore & NVMe

## The Tune-Pocalypse

- Red Hat Ceph 3.0 currently supports Filestore
  - Will support Bluestore in upcoming release
- Ceph Luminous Community 12.2.4
  - Tested Bluestore on the same RA hardware
- 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
```



Micron

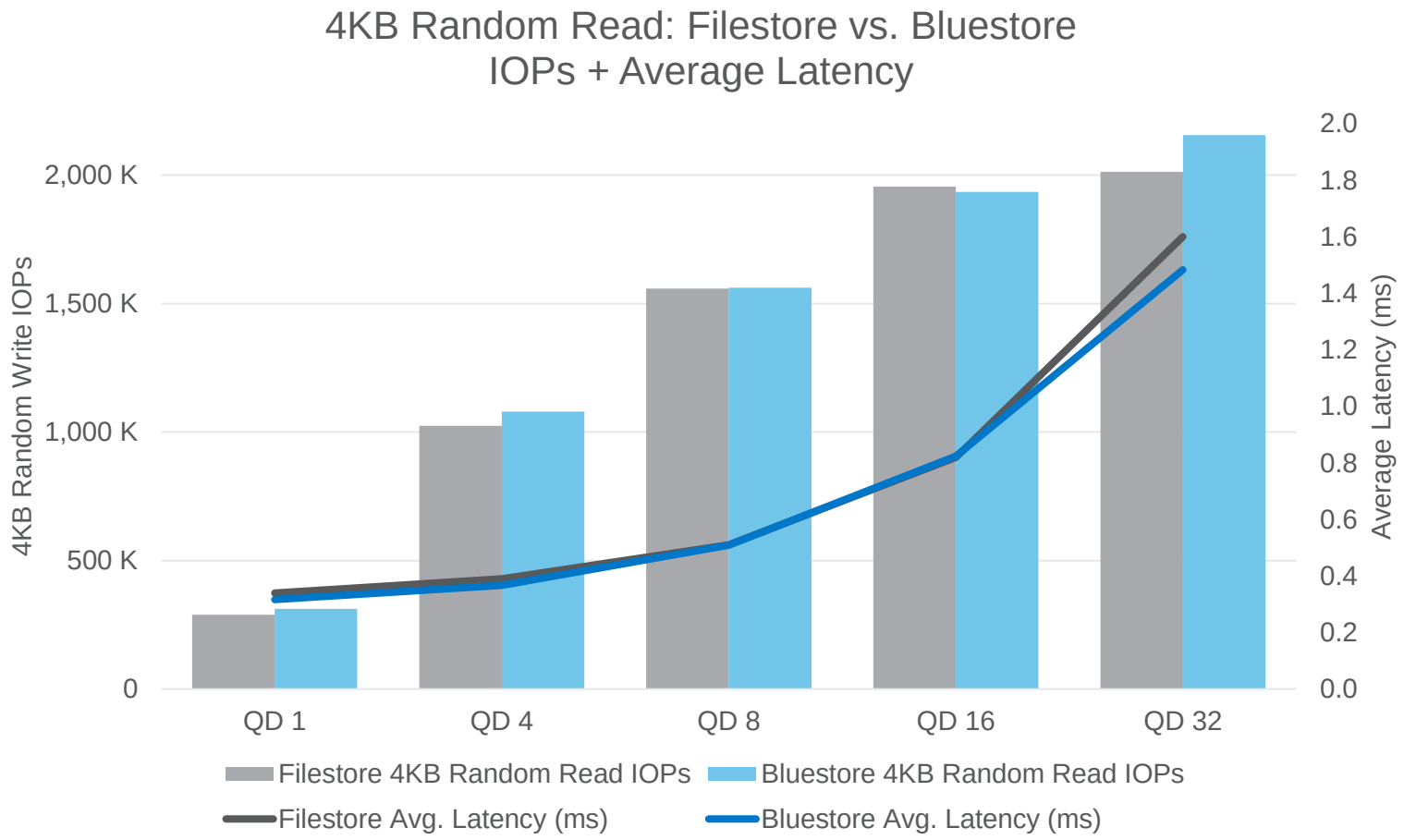
# Bluestore vs. Filestore: 4KB Random Read

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Reads:

- Queue Depth 32
  - Bluestore: 2.15 Million @ 1.5ms Avg. Latency
  - Filestore: 2.0 Million @ 1.6ms Avg. Latency

### Bluestore is slightly more performant than Filestore on 4KB Rand Reads



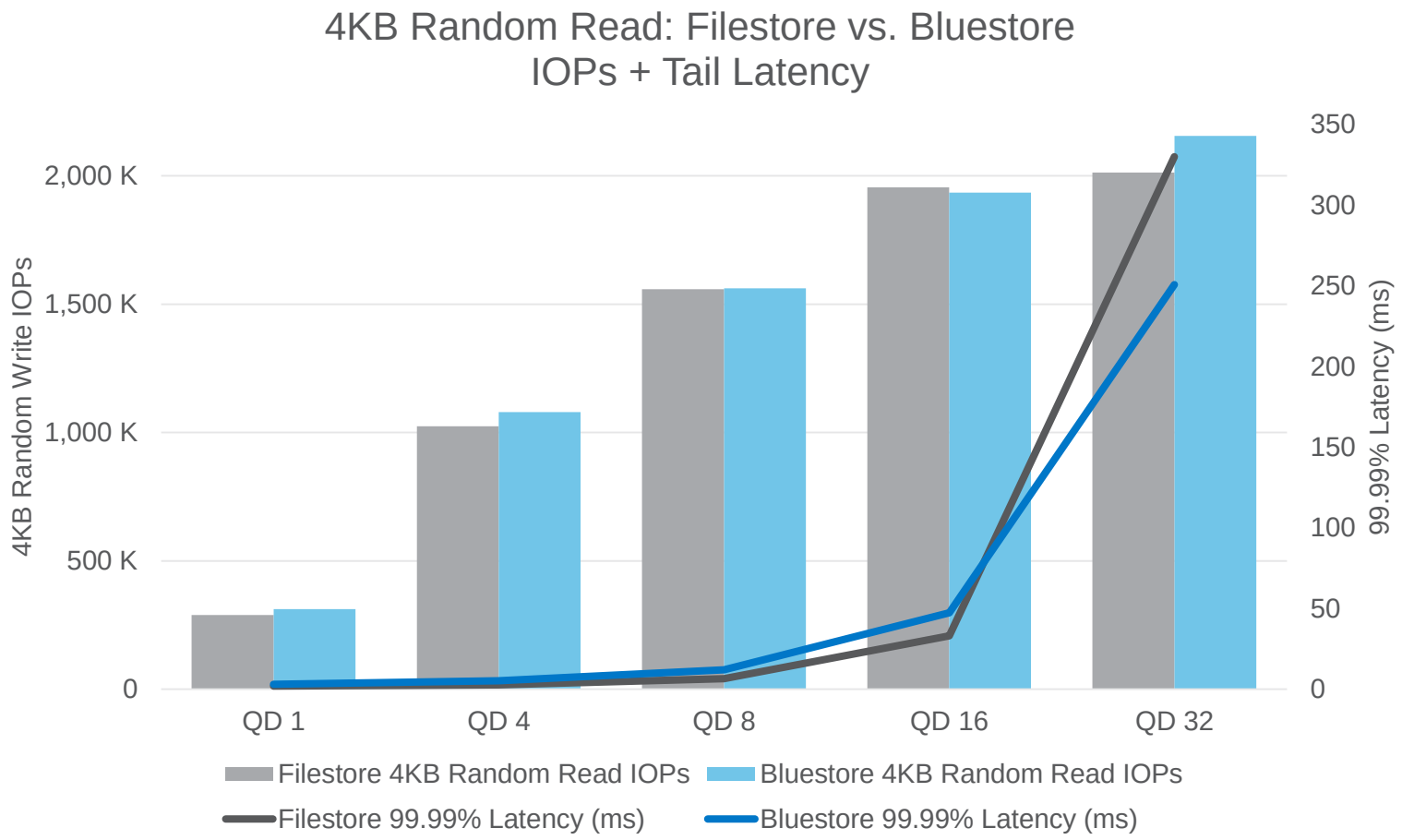
# Bluestore vs. Filestore: 4KB Random Read

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Reads:

- Queue Depth 32
  - Bluestore Tail Latency: 251 ms
  - Filestore Tail Latency: 330 ms

### Bluestore has lower tail latency and higher IOPs at high queue depths

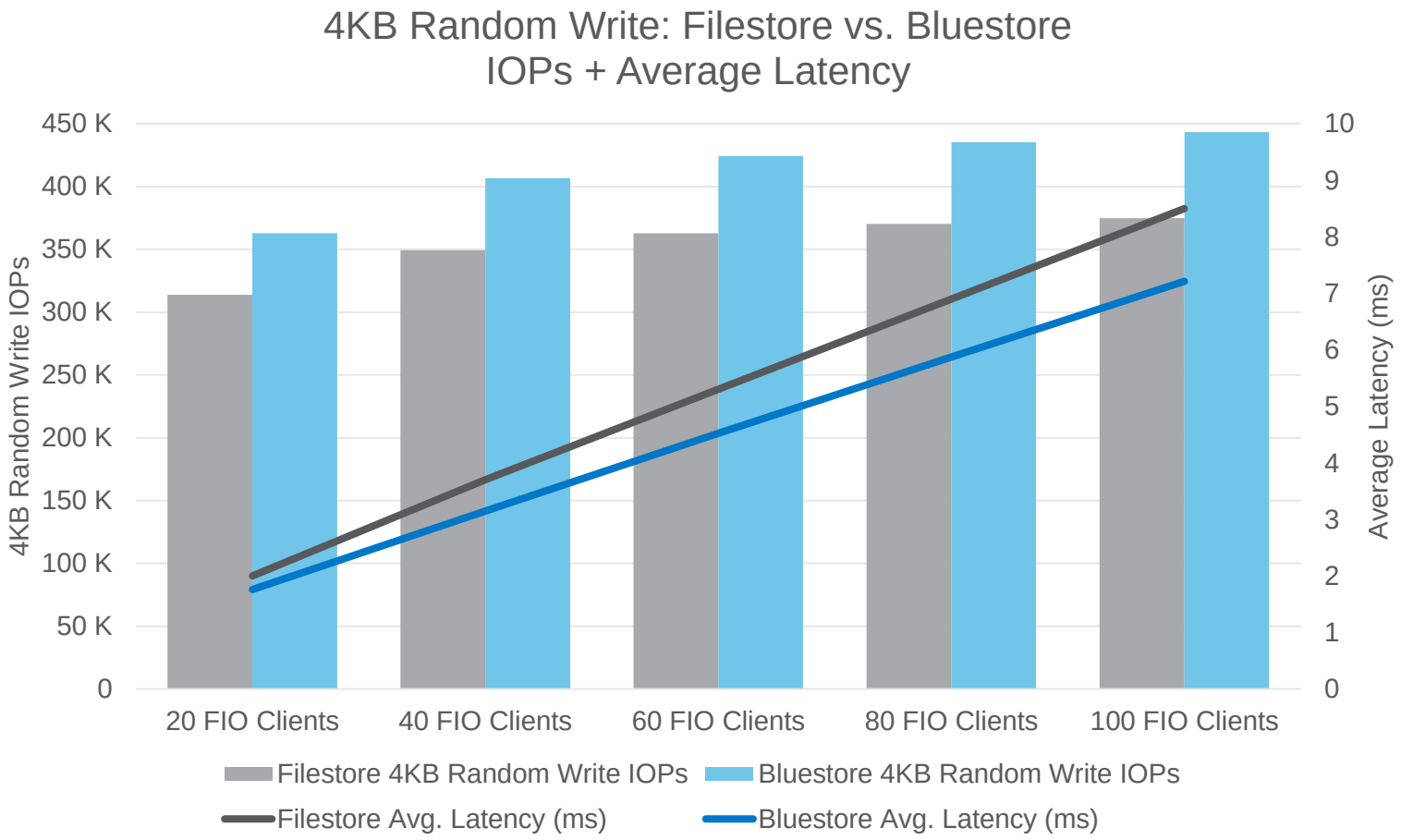


# Bluestore vs. Filestore: 4KB Random Write

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Writes:

- 18% Higher IOPs
- 15% Lower Avg. Latency
- 100 Clients
  - Bluestore: 443k IOPs @ 7ms
  - Filestore: 375k @ 8.5ms

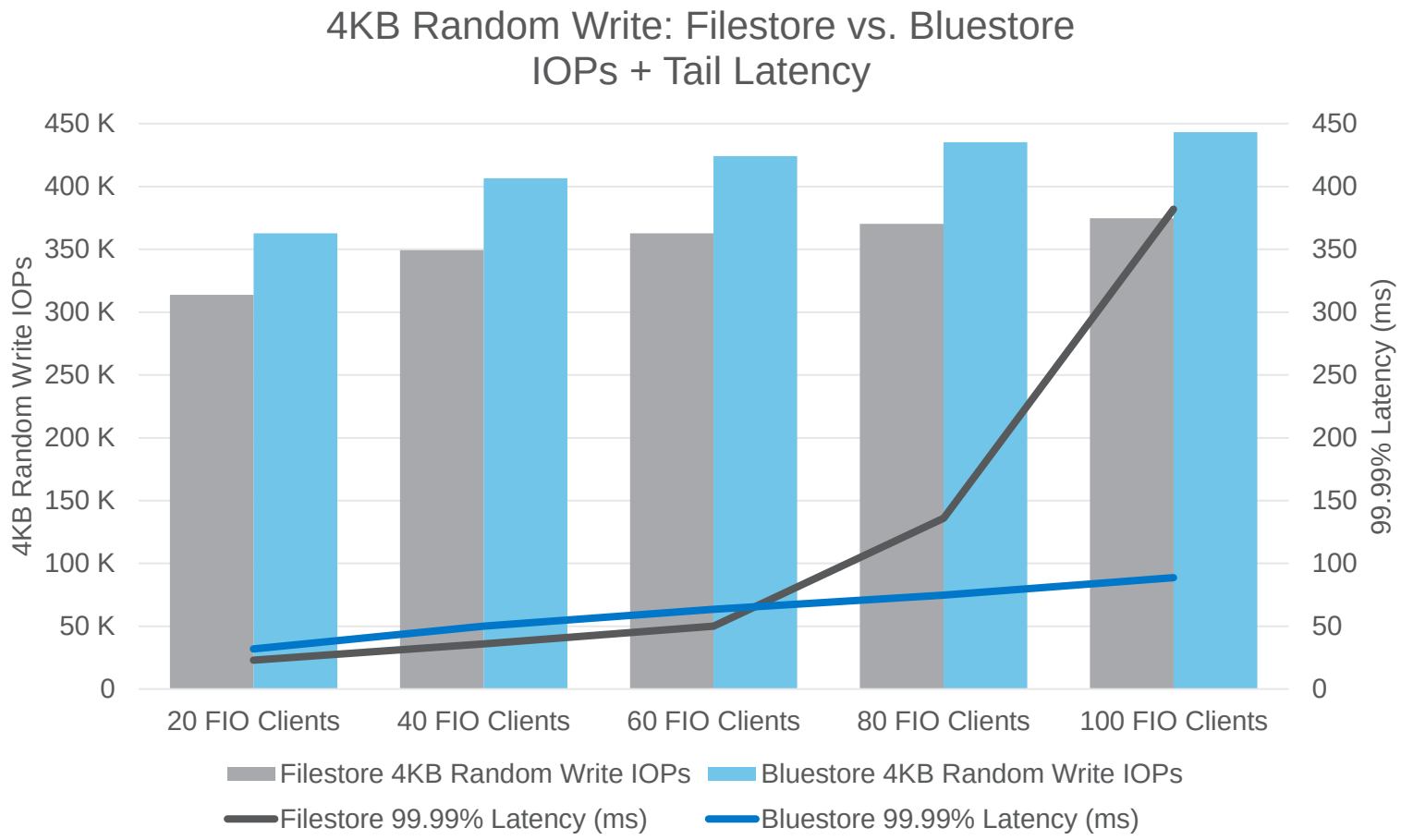


# Bluestore vs. Filestore: 4KB Random Write

## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Writes:

- 18% Higher IOPs
- Up to 70% reduced tail latency
- 100 Clients
  - Bluestore Tail Latency: 89ms
  - Filestore Tail Latency: 382ms



# Bluestore vs. Filestore: 4MB Object Read

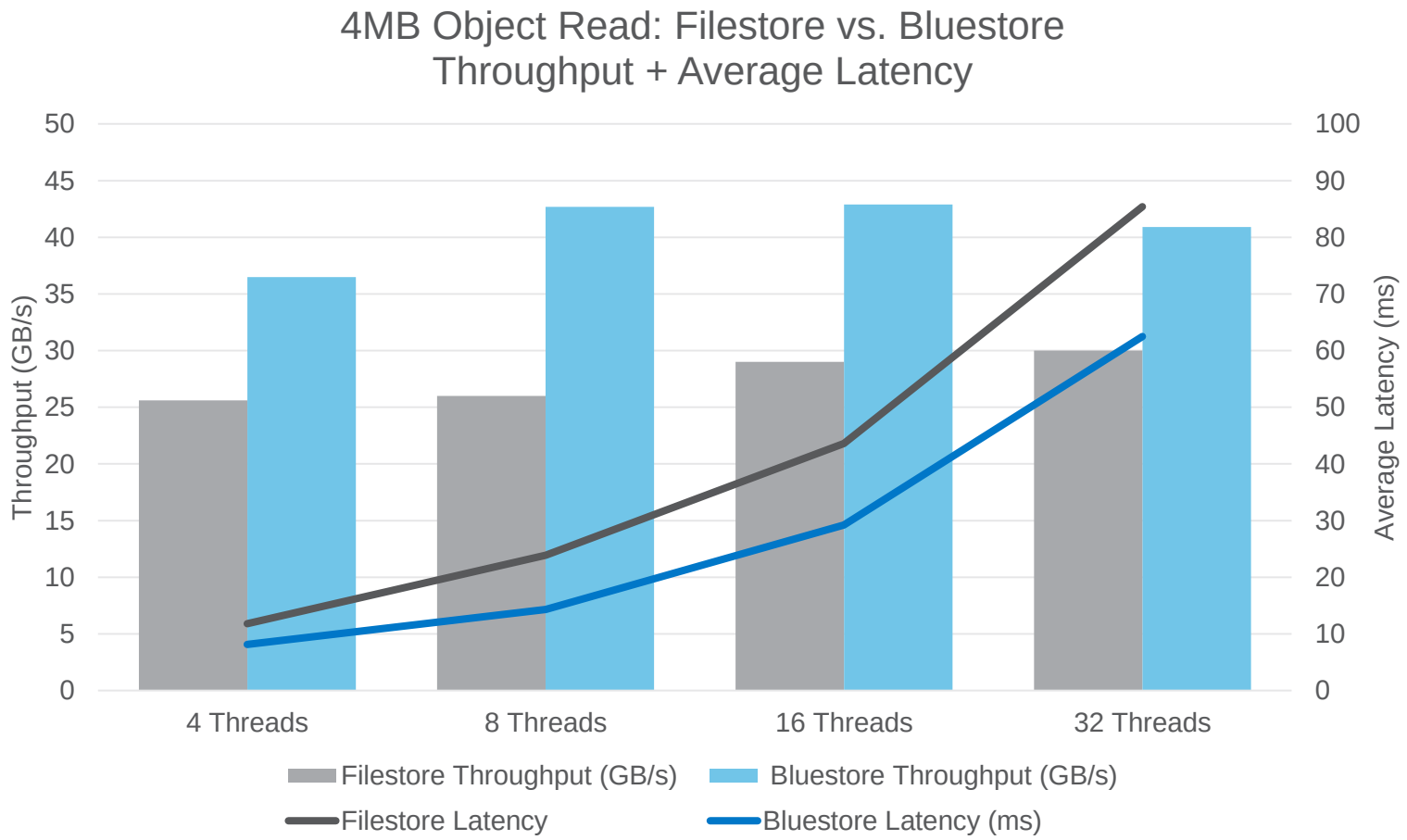
## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Writes:

- 16 Threads:
  - Bluestore: 42.9 GB/s @ 29ms
  - Filestore: 29 GB/s @ 44ms

### Bluestore improves object read

- 48% higher throughput
- 33% lower average latency



# Bluestore vs. Filestore: 4MB Object Write

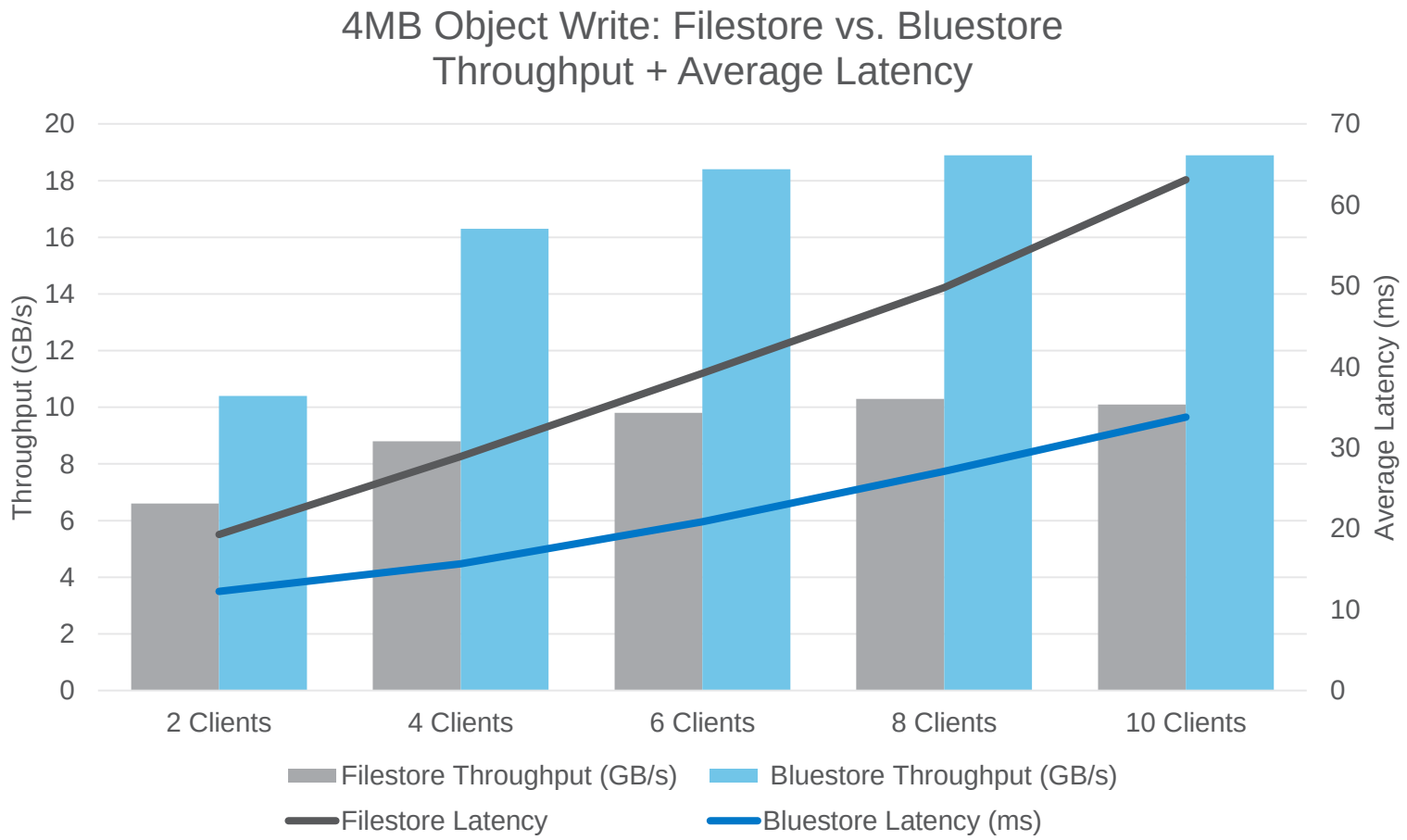
## Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

### 4KB Random Writes:

- 6 Clients:
  - Bluestore: 18.4 GB/s @ 21ms
  - Filestore: 9.8 GB/s @ 39ms

### Bluestore improves object Write

- 2x higher throughput
- Half average latency

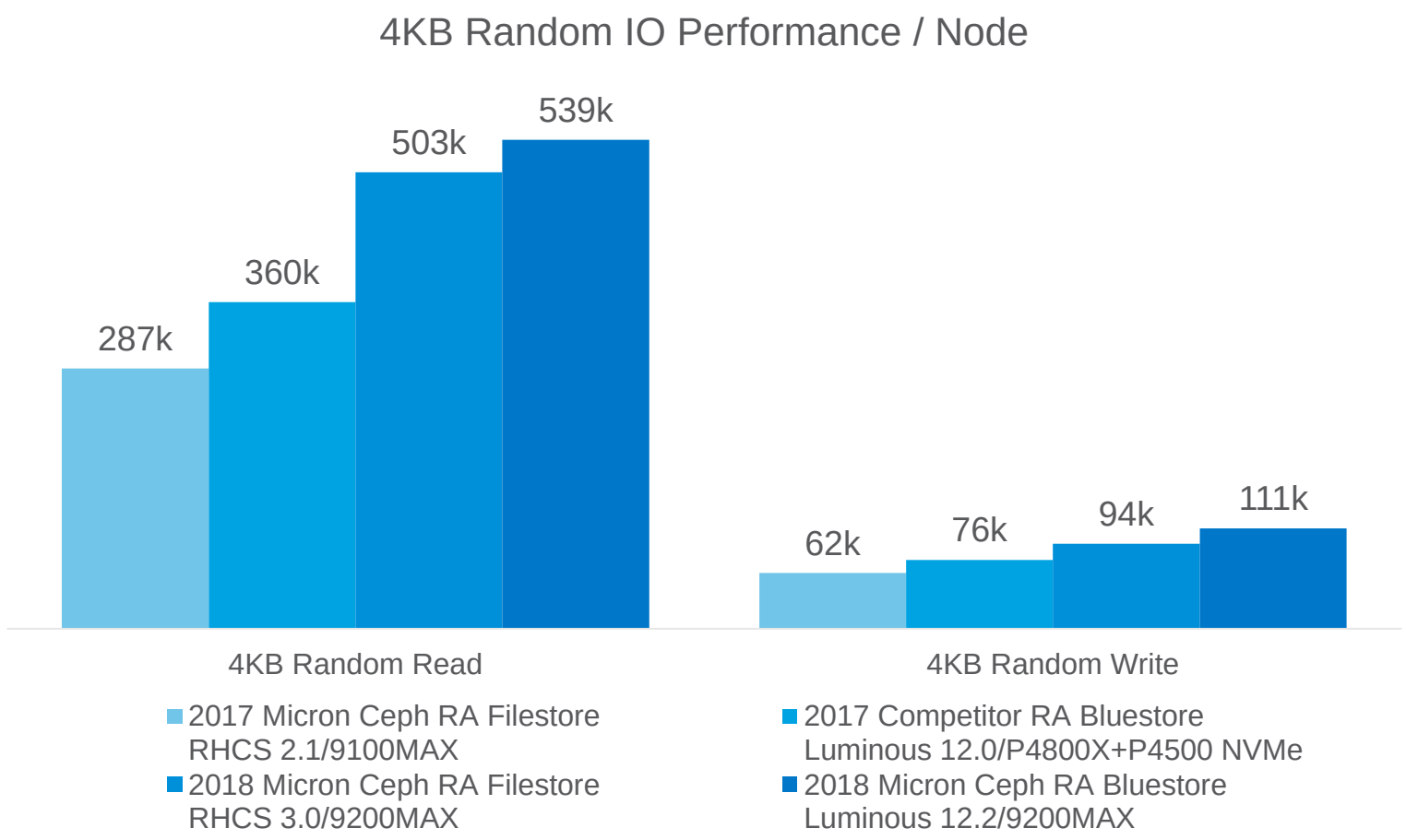


# Performance Comparison: 4KB Random Block

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

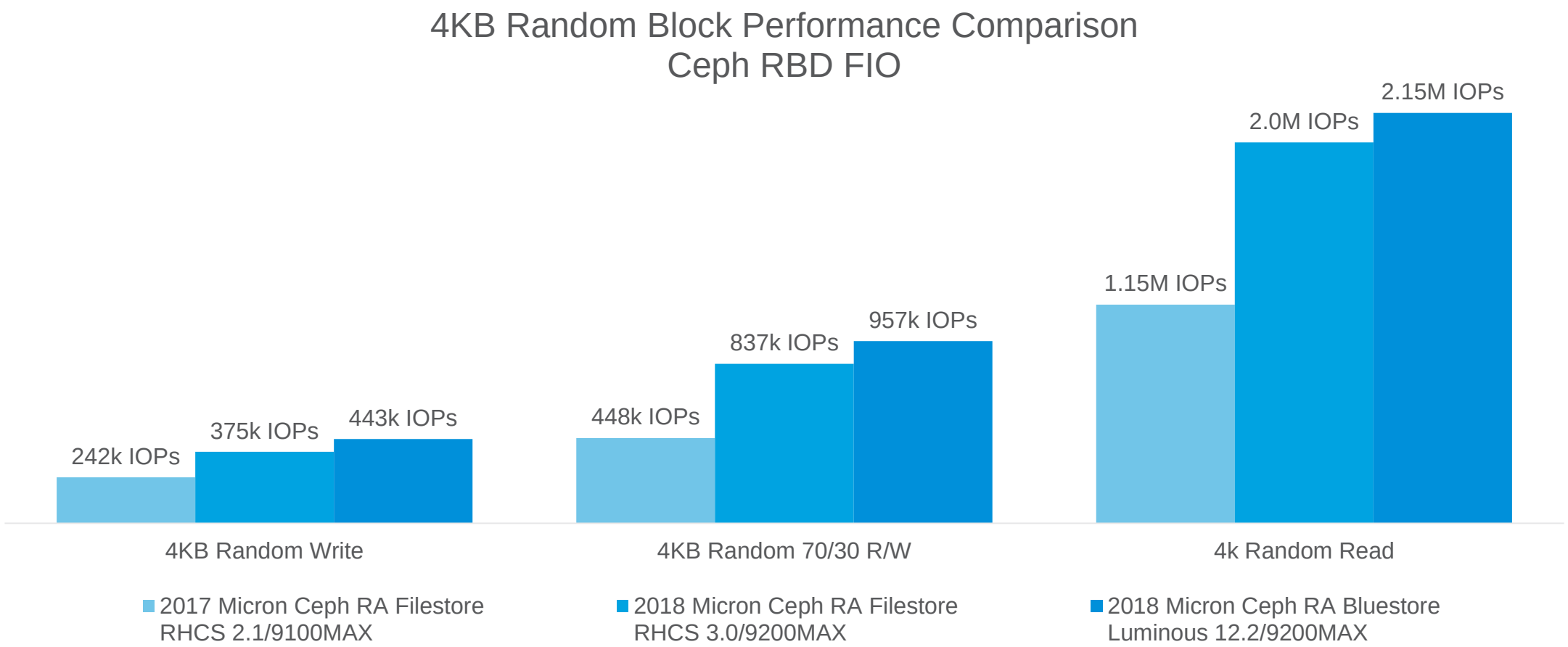
### 4KB Random IOPs

- 4KB Random Reads
  - 1.75X Micron 2017 RA tIOPs Micron 2017 RA to 2018 RA
  - 1.9X  IOPs Micron 2017 RA to 2018 RA + Bluestore
- 4KB Random Writes
  - 1.5X IOPs Micron 2017 RA to 2018 RA
  - 1.8X IOPs o 2018 RA + Bluestore



# Performance Comparison: 4KB Random Block

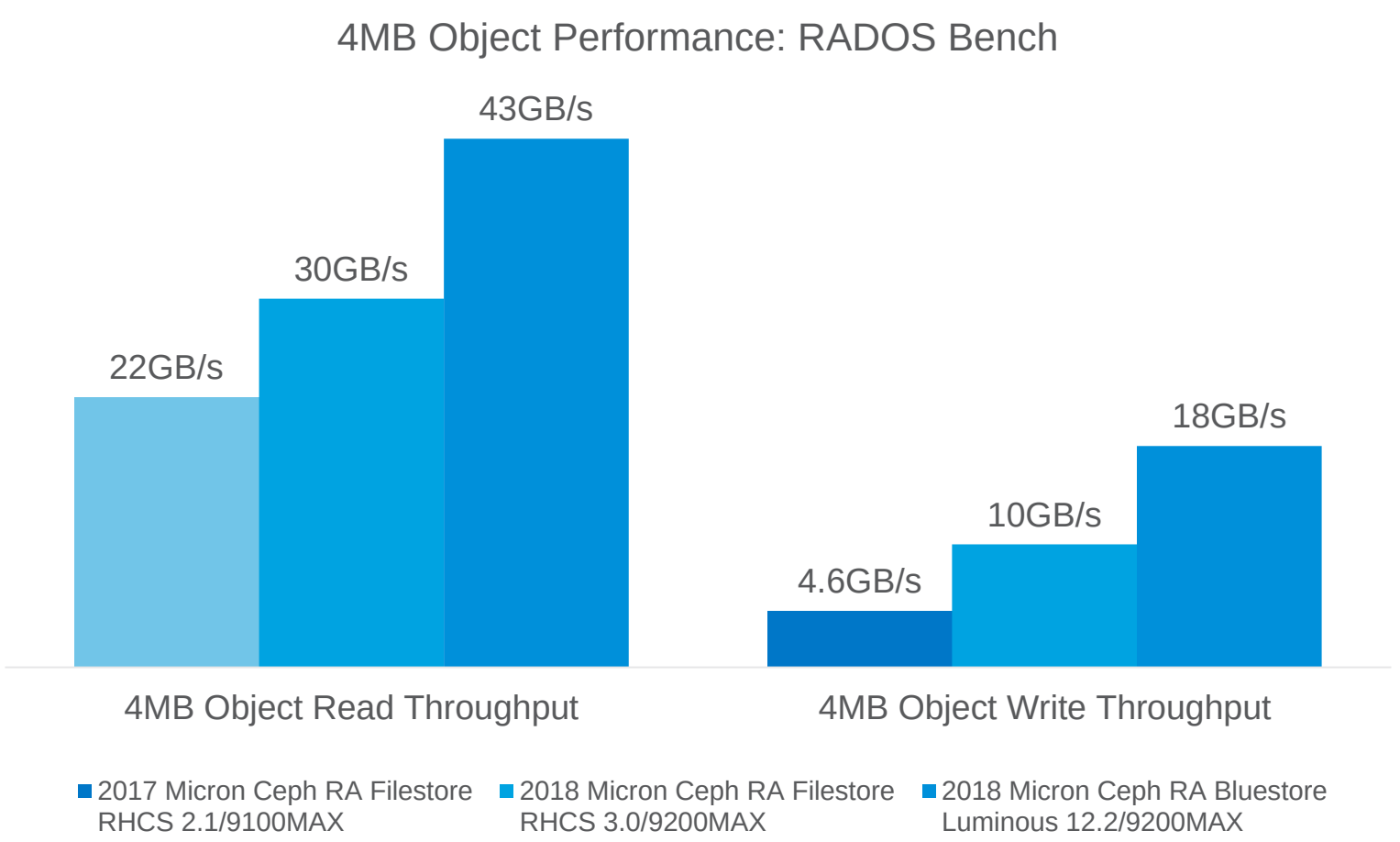
## 2017 Micron RA vs. 2018 Micron RA + Bluestore



# Performance Comparison: 4MB Object
## 2017 Micron RA vs. 2018 Micron RA + Bluestore

### 4MB Object Throughput

- Read Throughput
  - 1.4X Throughput 2017 to 2018
  - 1.95X Throughput 2017 to 2018+Bluestore
- Write Throughput
  - 2.2x Throughput 2017 to 2018
  - 3.9x Throughput 2017 to 2018+Bluestore




Thanks All
These slides will be available on the OpenStack conference website
Reference architecture available now!

# Micron Ceph Collateral

Micron + Red Hat + Supermicro ALL-NVMe Ceph RA

Micron NVMe Reference Architecture:

https://www.micron.com/~/media/documents/products/technical-marketing-brief/micron_9200_ceph_3,-d-,0_reference_architecture.pdf

Protip: Just Google “Micron 9200 Ceph”

# Be Revolutionary. Be SOLID.


Micron®

# Micron Ceph Appendix

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

```
[global]
auth client required = none
auth cluster required = none
auth service required = none
auth supported = none
mon host = xxx
ms_type = async
rbd readahead disable after bytes = 0
rbd readahead max bytes = 4194304
mon compact on trim = False
mon_allow_pool_delete = true
osd pg bits = 8
osd pgp bits = 8
osd_pool_default_size = 2
mon pg warn max object skew = 100000
perf = True
mutex_perf_counter = True
throttler_perf_counter = False
rbd cache = false
mon_max_pg_per_osd = 800

[mon]
mon_osd_max_split_count = 10000

[osd]
osd journal size = 20480
osd mount options xfs =
noatime,largeio,inode64,swalloc
osd mkfs options xfs = -f -i size=2048
osd_op_threads = 32
filestore_queue_max_ops = 5000
filestore_queue_committing_max_ops = 5000
journal_max_write_entries = 1000
journal_queue_max_ops = 3000
objecter_inflight_ops = 102400
filestore_wbthrottle_enable = False
osd mkfs type = xfs
filestore_max_sync_interval = 10
osd_client_message_size_cap = 0
osd_client_message_cap = 0
osd_enable_op_tracker = False
filestore_fd_cache_size = 64
filestore_fd_cache_shards = 32
filestore_op_threads = 6
filestore_queue_max_bytes=1048576000
filestore_queue_committing_max_bytes=1048576
000
journal_max_write_bytes=1048576000
journal_queue_max_bytes=1048576000
ms_dispatch_throttle_bytes=1048576000
objecter_infilght_op_bytes=1048576000
```

# 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 = xxxxxxx
 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
```

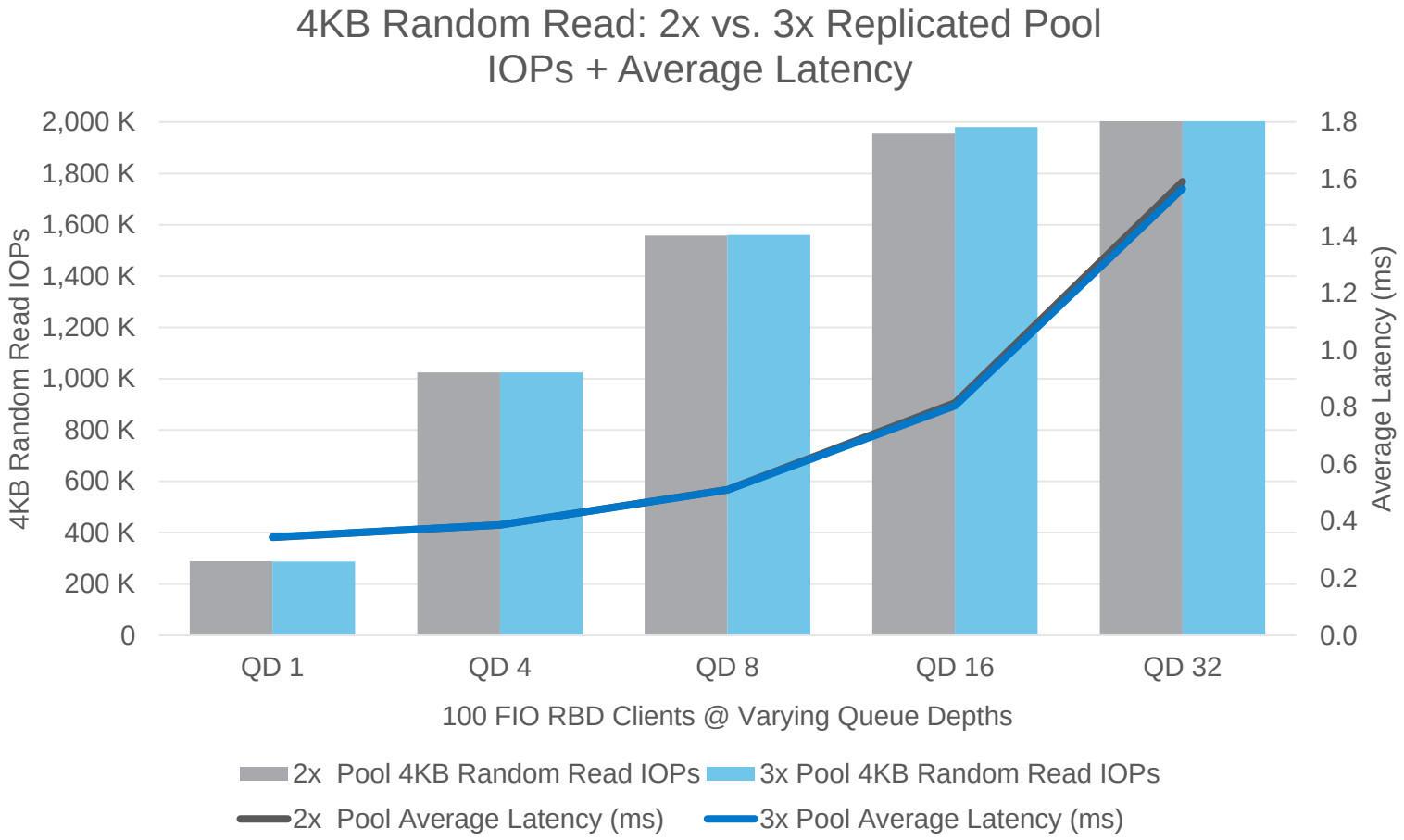
# But What About 3x Replicated Pools?

## 2x vs. 3x Replicated pools: 4KB Random Read

4KB Random Reads are the same on 2x and 3x replicated pools

Queue Depth 32:

- 2x: 2.01M IOPs @ 1.59ms
- 3x: 2.05M IOPs @ 1.57ms



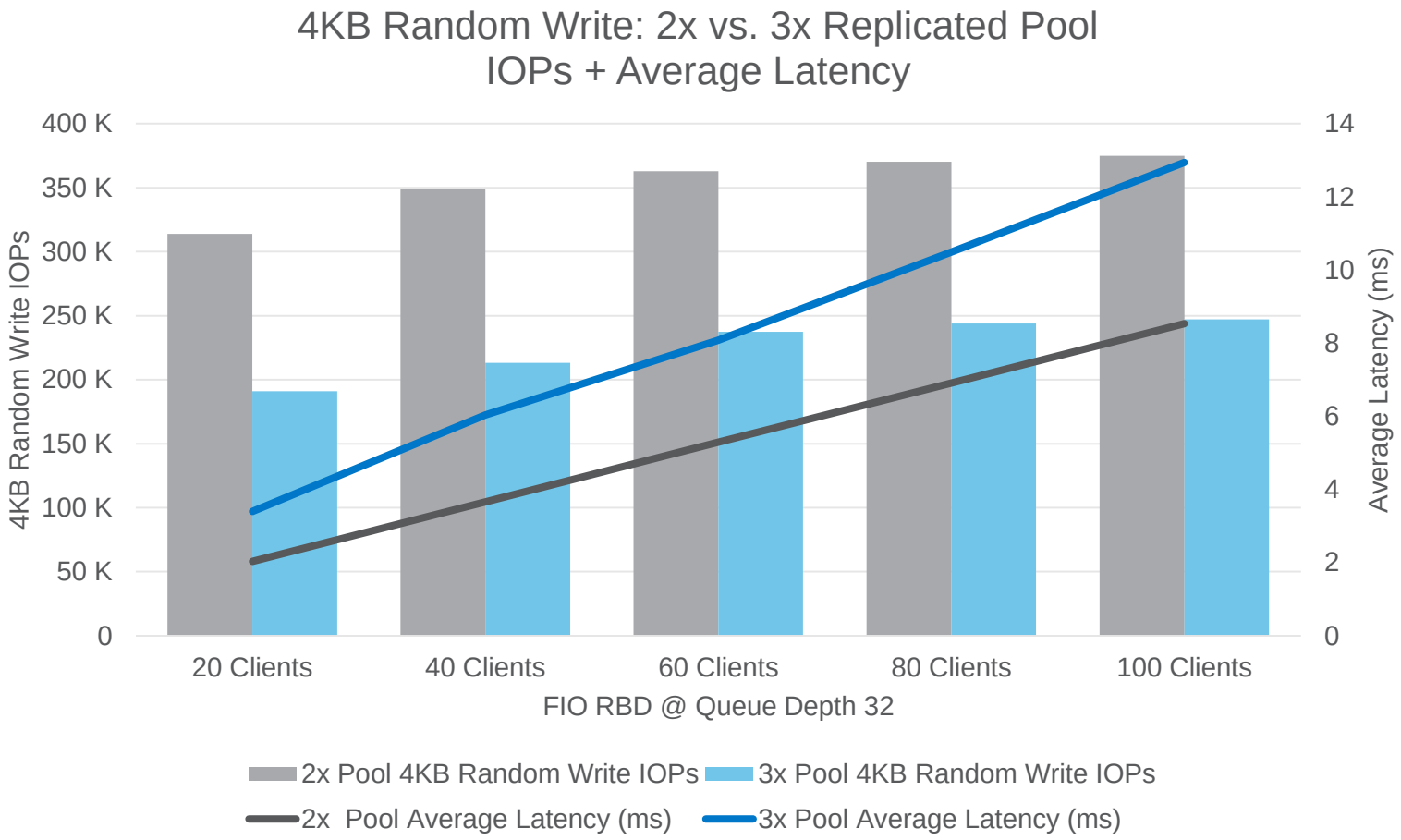
# But What About 3x Replicated Pools?

## 2x vs. 3x Replicated pools: 4KB Random Write

4KB Random Writes are exactly what you'd expect: 33% reduced IOPs

100 Clients:

- 2x: 375k IOPs @ 8.5ms
- 3x: 247k IOPs @ 12.9ms



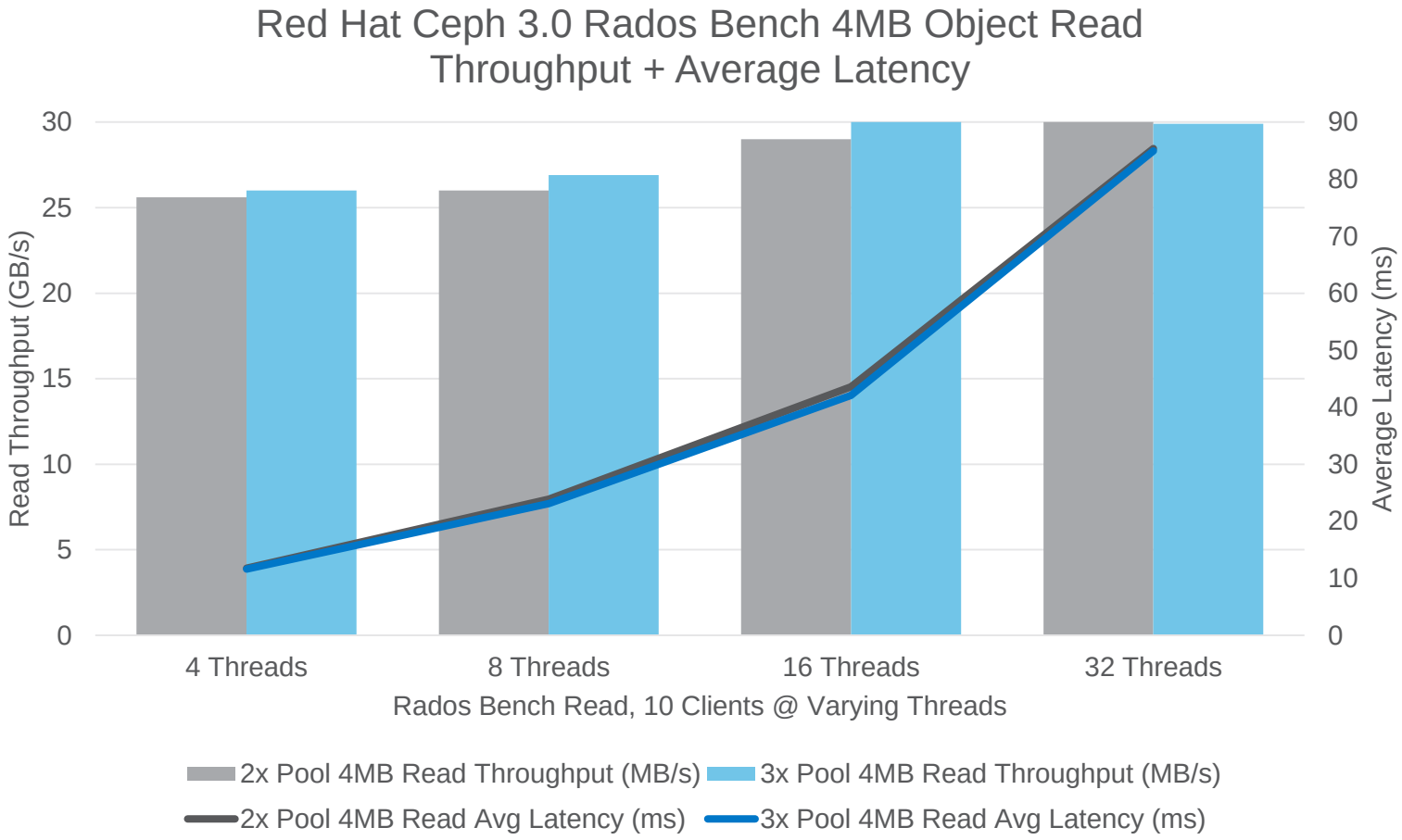
# But What About 3x Replicated Pools?

## 2x vs. 3x Replicated pools: 4MB Object Read

4MB Object reads are nearly the same

32 Threads:

- 2x: 30.0 GB/s @ 85.4ms
- 3x: 29.9 GB/s @ 85.0ms



# But What About 3x Replicated Pools?

## 2x vs. 3x Replicated pools: 4MB Object Writes

4MB Object Writes are ~33% lower on a 3x pool

8 Clients:

- 2x: 10.3 GB/s @ 49.8ms
- 3x: 6.7 GB/s @ 76.7ms