
# Ceph EC Performance Journey
# Classic vs. Fast EC

Vikhyat Umrao
Tejas Chaphekar

# Introduction

## Classic EC

- Stripe Size - 4K
- Plug In – Jerasure

## Fast EC

- Stripe Size – 16K
- Plug in – ISA - L
- Flag - allow_ec_optimizations

# Hardware

| Vendor/Model | Role | CPU | RAM( GB) | OSD Disk | BIOS System Profile |
|---|---|---|---|---|---|
| Dell R760 X5D | Ceph OSD nodes | 2 x Intel(R) Xeon(R) Gold 6438N (32c/64T) | 512 | 8, 16 and 24 packs<br><br>Micron NVMe (TLC)<br><br>3.84T | Performance |
| Dell R660 | FIO data loaders (clients) | Intel(R) Xeon(R) Silver 4416+ | 128 | NA | Performance |

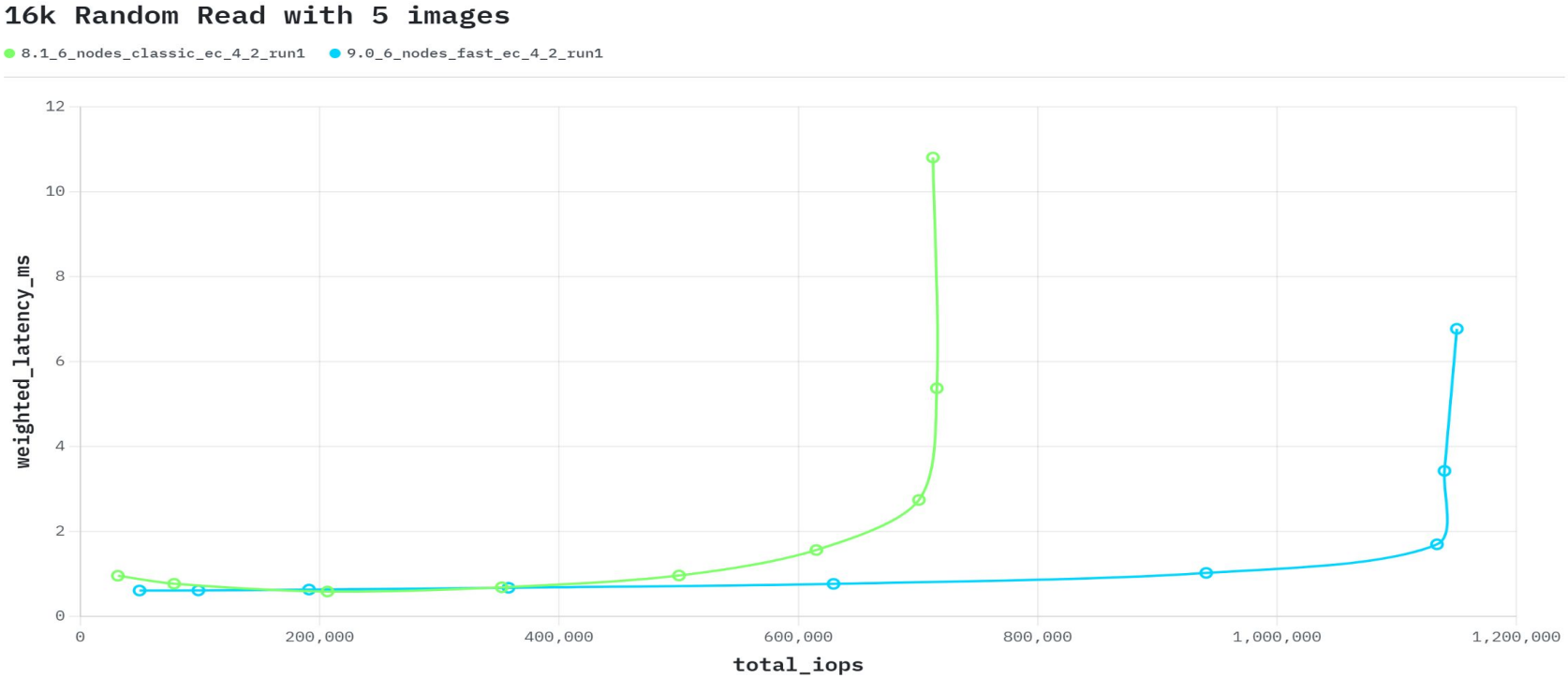
# Cluster Configuration

| Configuration | Value |
|---|---|
| Ceph version Squid | 19.2.1 |
| Ceph version - Tentacle | 20.1.0 |
| OSD Memory Target Autotune | False |
| OSD Memory Target | 16G |
| Autoscaler | Off |
| Balancer | Off |
| Scrubbing | Off |

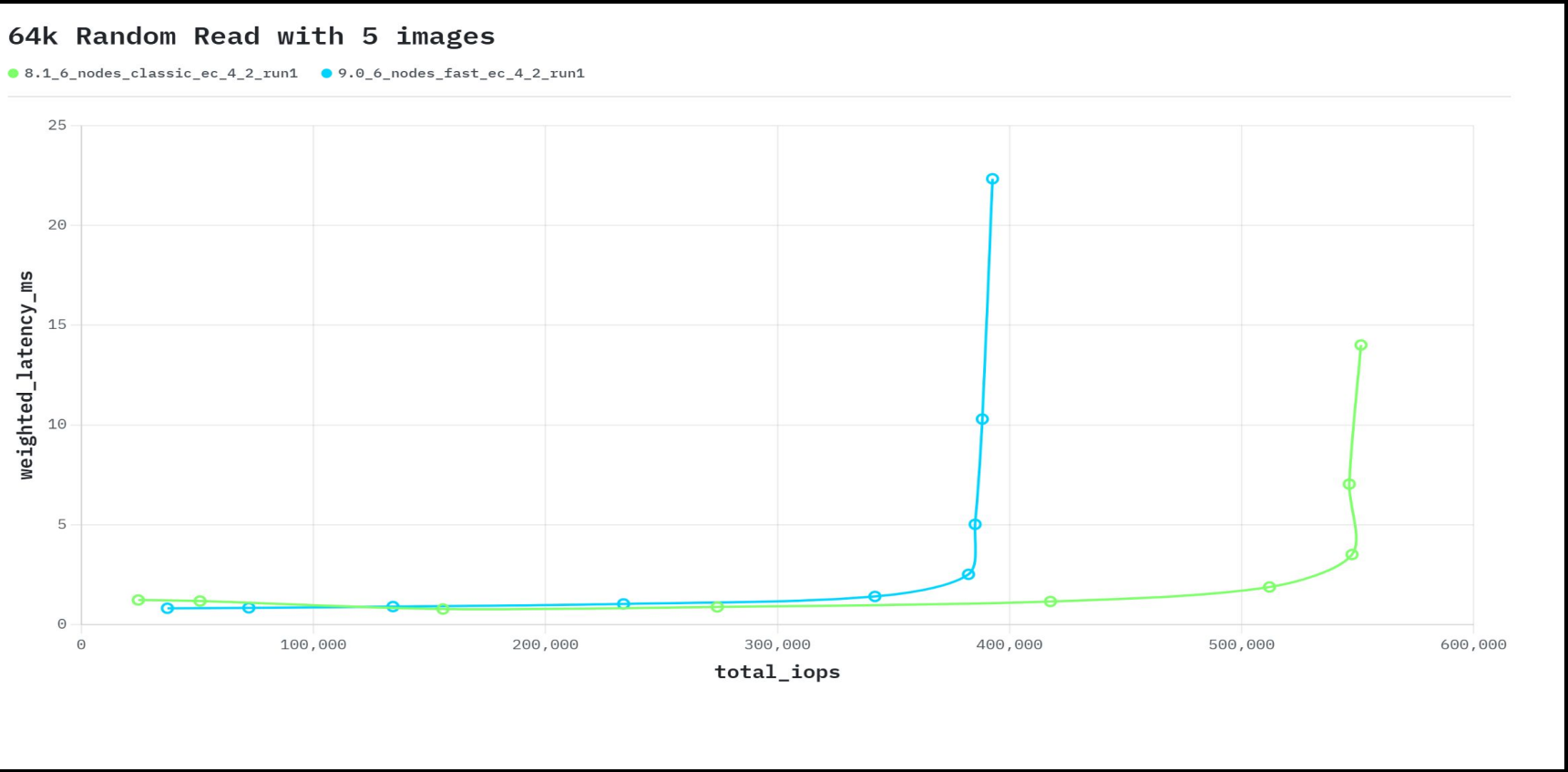
# Test Coverage

| Parameters | Values |
|---|---|
| Images | 30 RBD images with 6 FIO Clients |
| I/O Depths | 1, 2, 4, 8, 16, 32, 64, 128, 256 |
| Pool Types | Replica-3, EC (2+2, EC 4+2, EC 6+2) |
| R/W Ratios | 0R100W, 70R30W, 100R0W |
| Block Sizes | 4KB, 16KB, 32KB, 64KB, 4MB |
| OSD Count | 144 (6 nodes T-shirt size), 288 ( 12 nodes T-shirt size) |

# Performance Latency chart Squid(8.1) vs. Tentacle (9.0) - EC 4+2, 16K Block size, random read, All IO depths



# Performance Latency chart Squid(8.1) vs. Tentacle (9.0) - EC 4+2, 64K Block size, random read, All IO depths



# OSD memory accumulation

# EC Memory Leak Investigation

- Fast Erasure Coding implementation showed progressive memory leak during sustained random read I/O workloads, causing OSD memory exhaustion (20-22GB peak vs 16GB target).
- It was seen that the ECDummyOp objects (~1k) accumulated in OSD memory during normal I/O operations when PGs transition to idle state.
- These objects were not freed until a PG interval change occurred, resulting in memory growth during extended test runs.

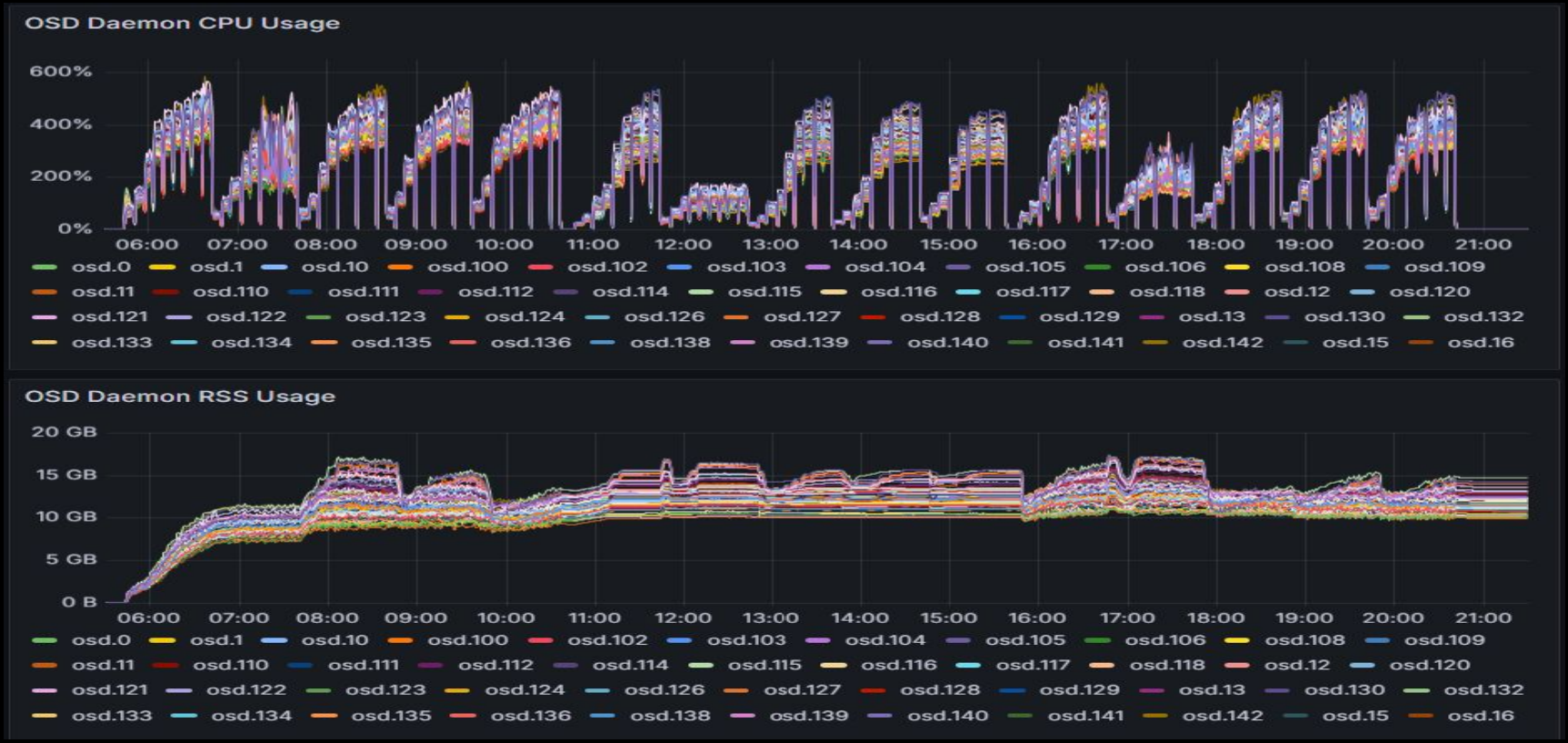
# EC Memory Leak Workaround / Patch

- Implemented  PG repeer operation between test iterations to force PG interval changesand trigger ECDummyOp cleanup.

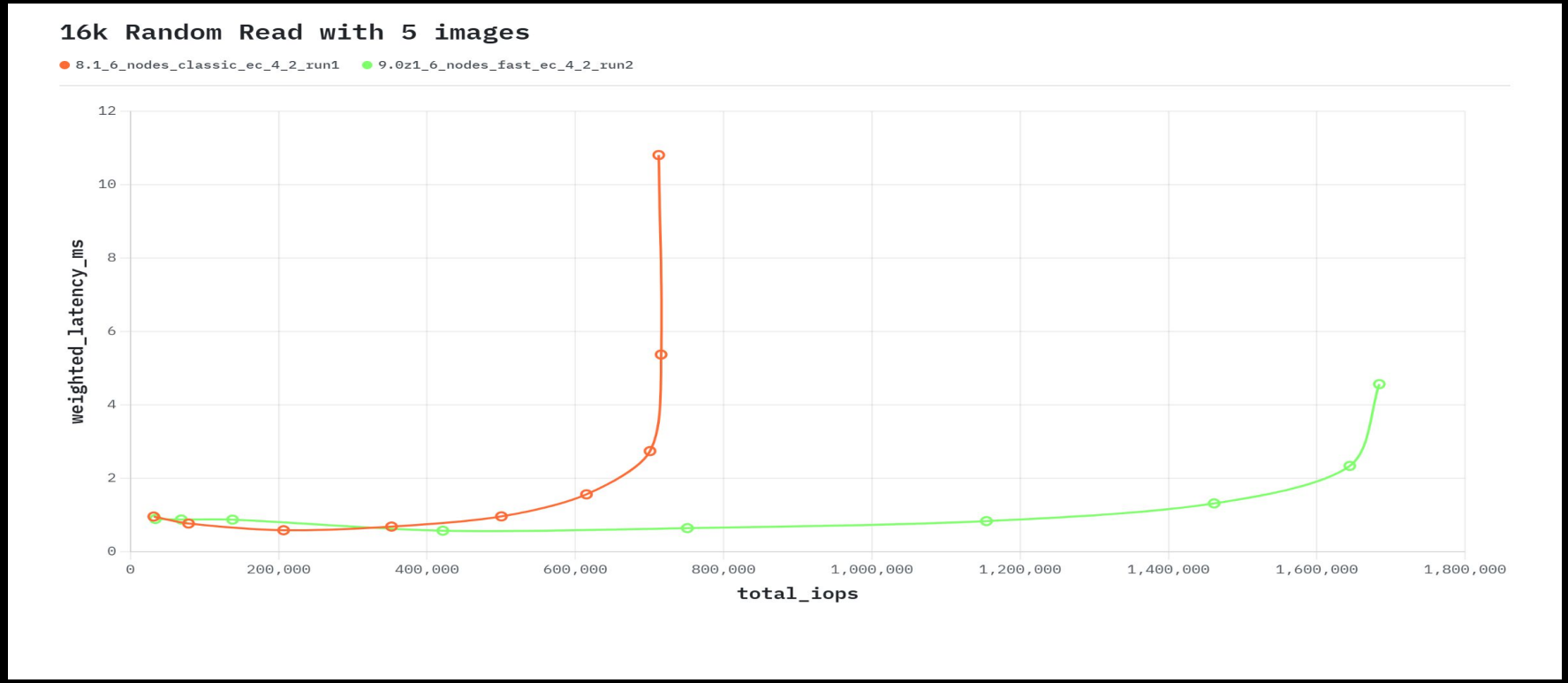
```
ceph pg dump pgs_brief | awk '/^[0-9]+\.[0-9a-f]+/ {print $1}' | xargs -P 32 -n 1 ceph pg repeer
```

- Upstream tracker and PR - https://tracker.ceph.com/issues/74433 , https://github.com/ceph/ceph/pull/66961

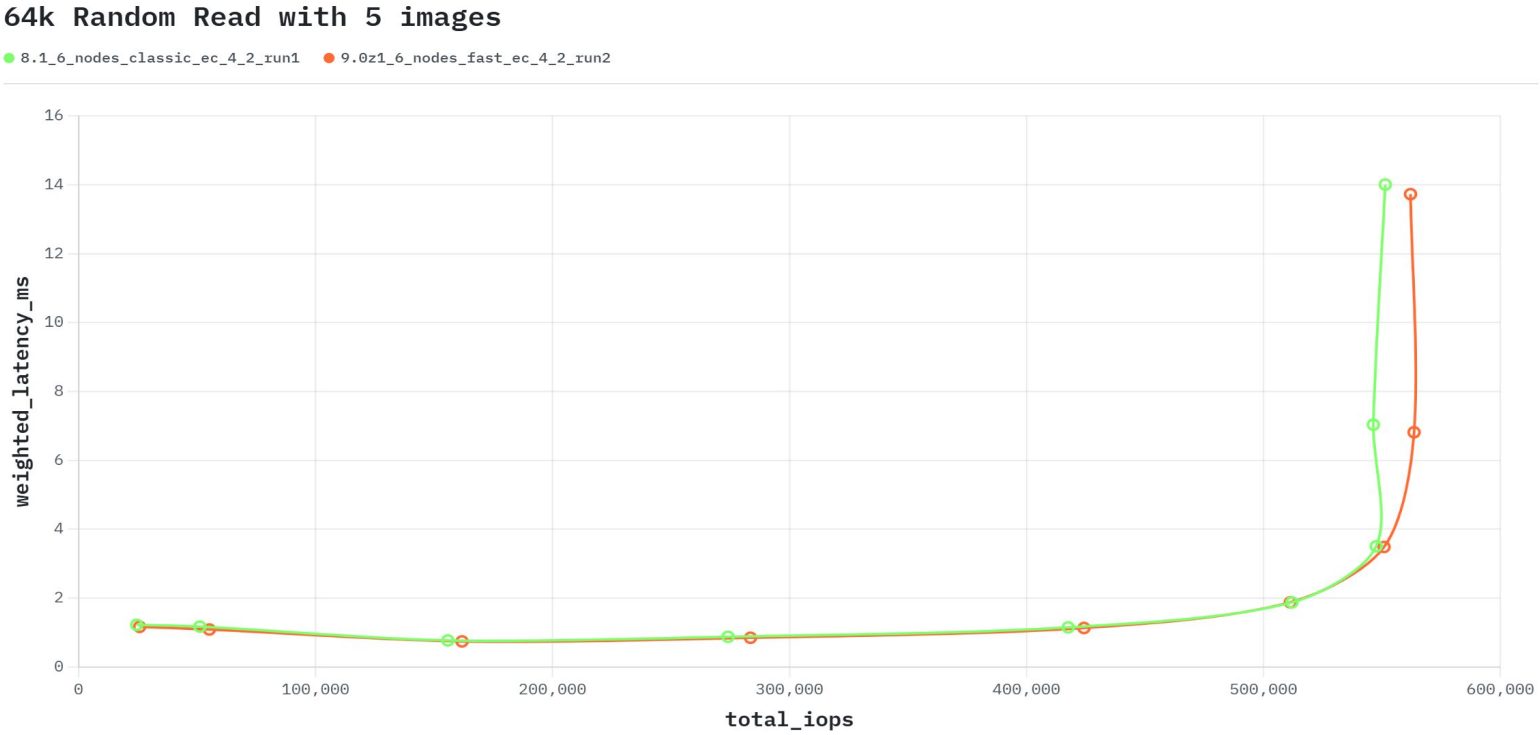
# After the fix OSD memory



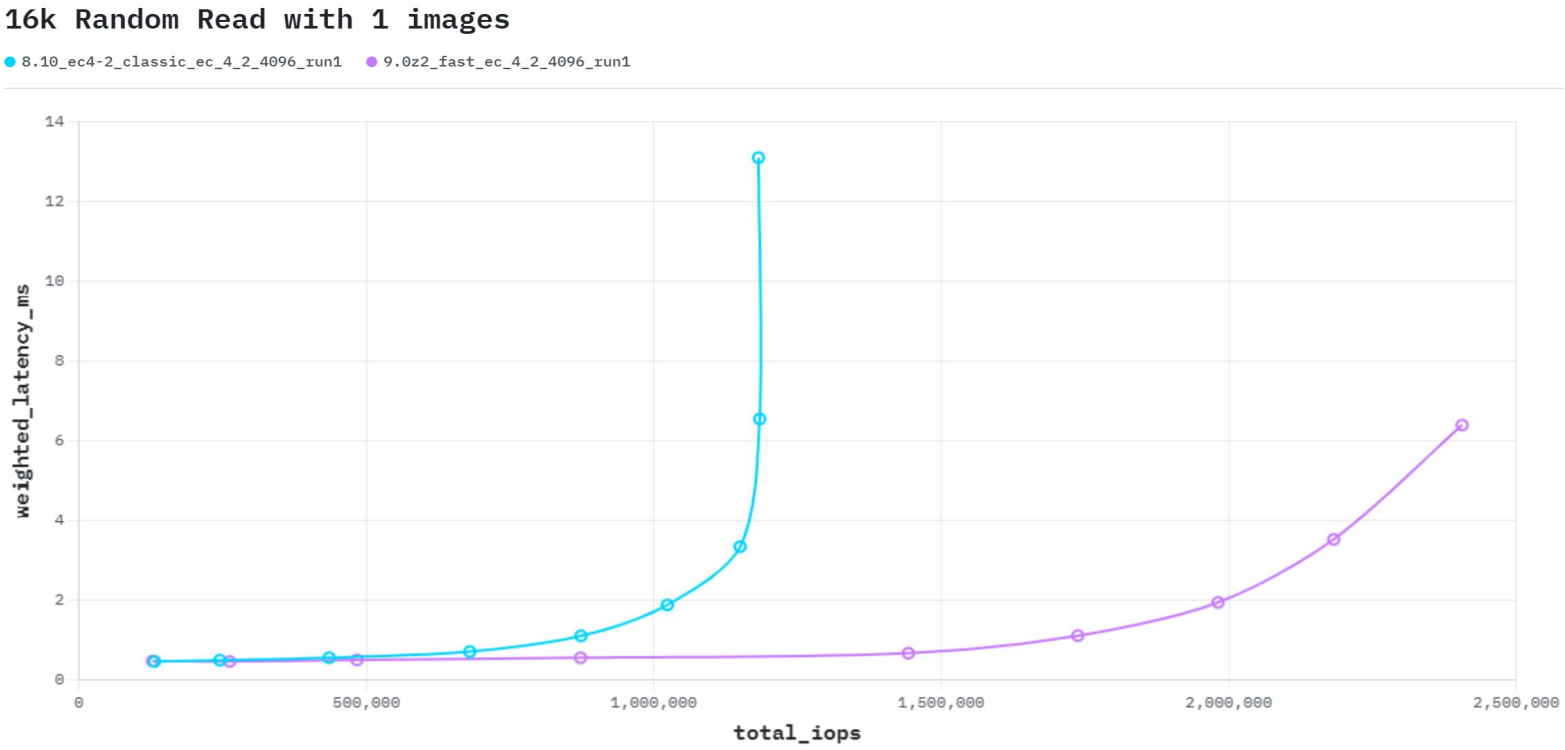
# Performance Latency chart Squid(8.1) vs. Tentacle (9.0.z1) - EC 4+2, 16K Block size, random read, All IO depths – After the fix



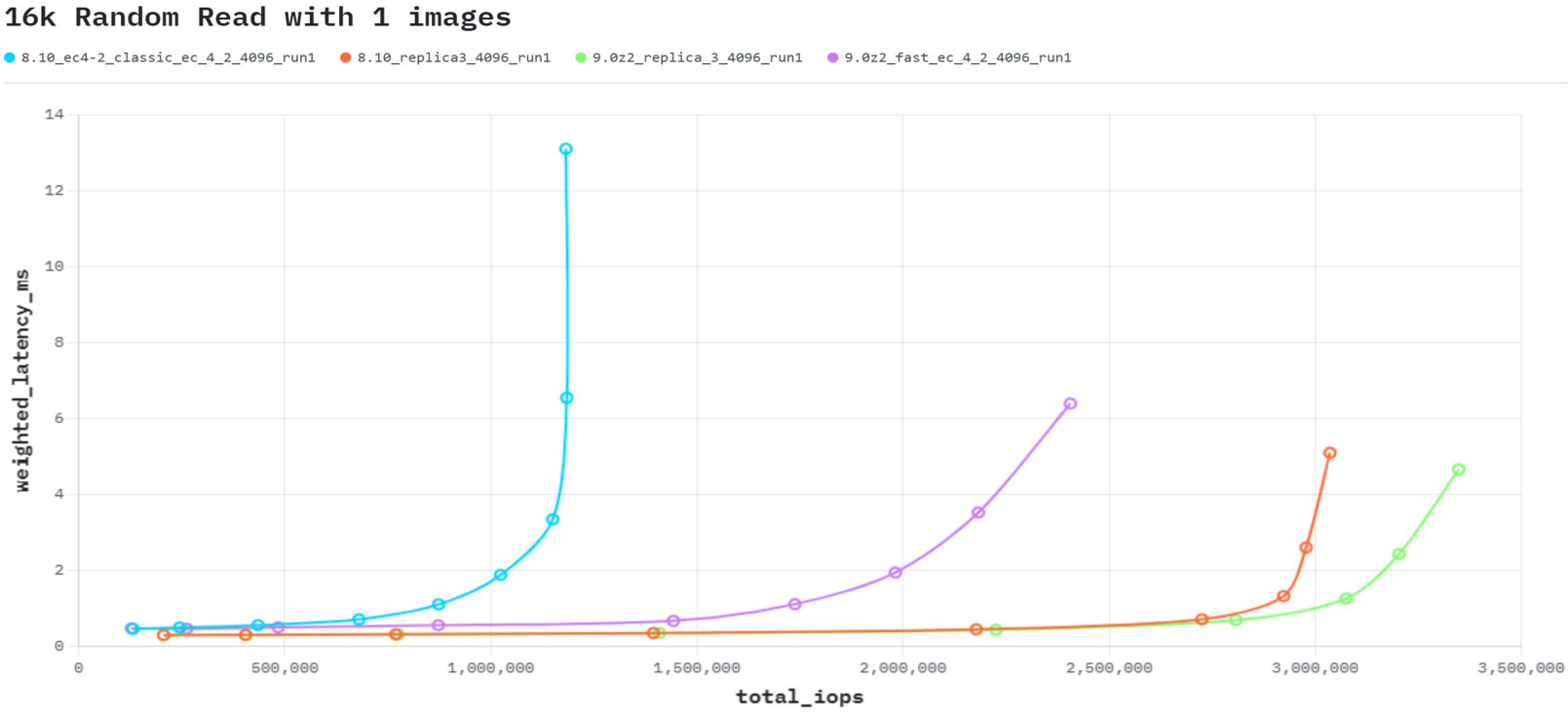
# Performance Latency chart Squid(8.1) vs. Tentacle (9.0.z1) - EC 4+2, 64K Block size, random read, All IO depths – After the fix



# 12 nodes T-shirt size comparison Squid(8.1) vs. Tentacle (9.0.z2), EC 4+2, 16K Block size, random read, All IO depths – After the fix



# 12 nodes T-shirt size comparison Squid(8.1) vs. Tentacle (9.0.z2), EC 4+2 and Replica 3, 16K Block size, random read, All IO depths – After the fix



# Thank you

vikhyat@ibm.com