



# **BLOCK 3:** **Object Storage & CephFS Deep Dive**

14th October 2026, Zurich



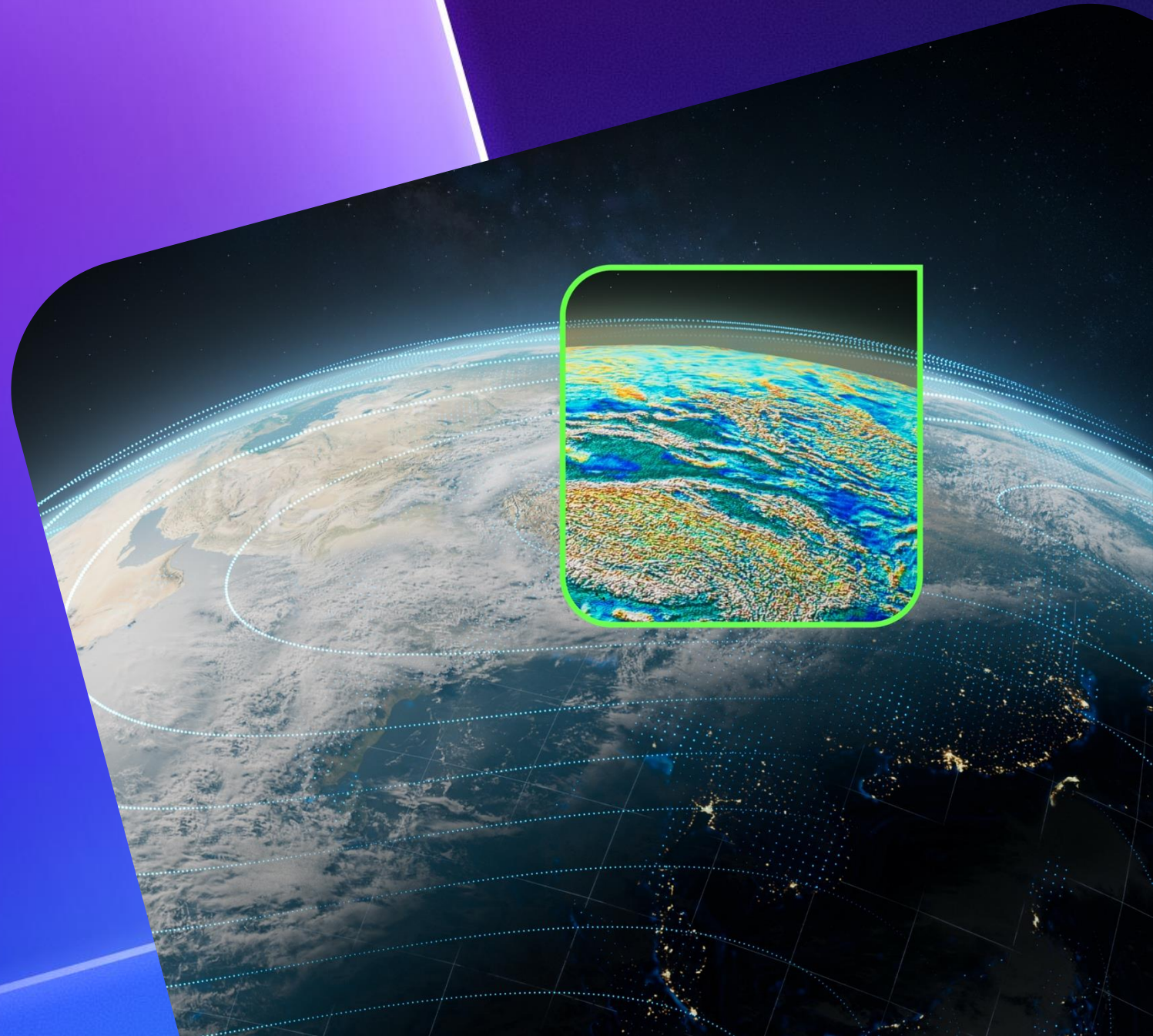
# WELCOME

## Colin Morey

14th October 2026, Zurich



# BENCHMARKING THE JOURNEY TO OBJECT STORAGE





# Table of Contents

- 01 / **Who we are**  
Setting the scope
- 02 / **Our benchmarking process**  
Climbing up the stack
- 03 / **Observations**
- 04 / **Questions**



# WHO WE ARE

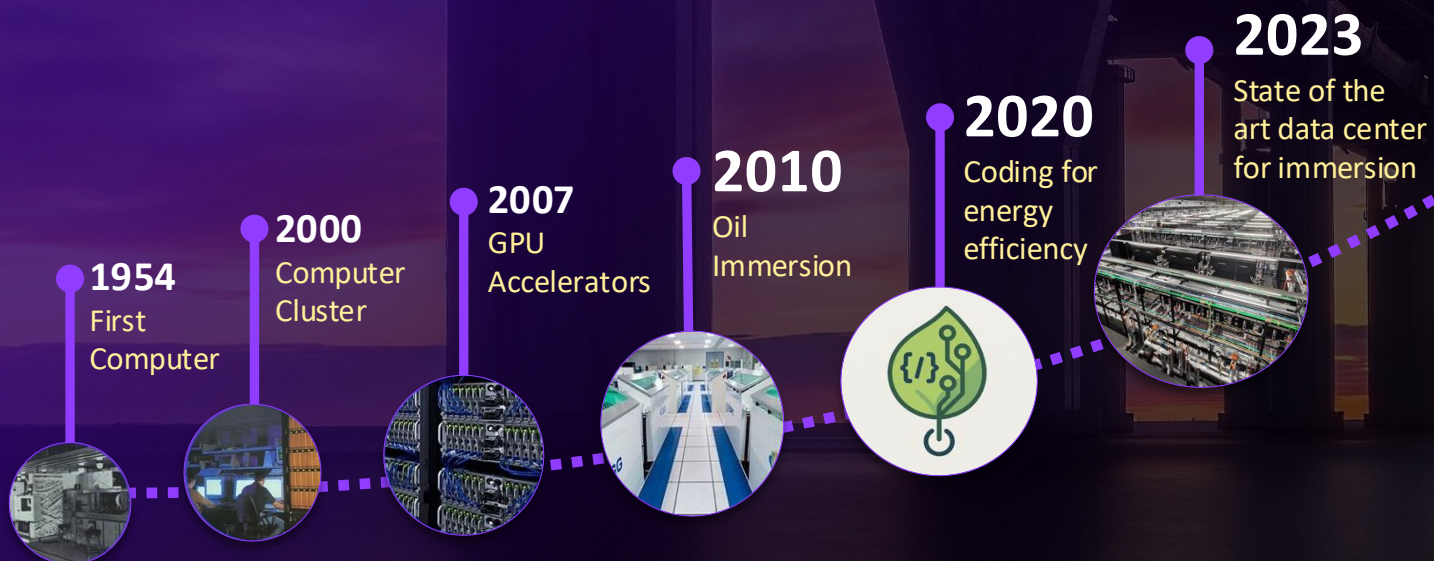


## OUR HPC EXPERIENCE

# HIGH-PERFORMANCE COMPUTING IS CRITICAL TO OUR CORE

Highly optimized HPC is one of the driving factors that enable the success of Viridien's Seismic Imaging business.

# VIRIDIEN



**TODAY**

700+ PetaFlops (SP) of compute  
300+ PetaBytes of storage

11 HPC OnPrem Centers,  
including 2 Client centers

HPC Hubs in Houston,  
London and Paris

21K Servers &  
32K GPUs

More than one million HPC  
jobs per day

Strong collaboration with technology  
leaders



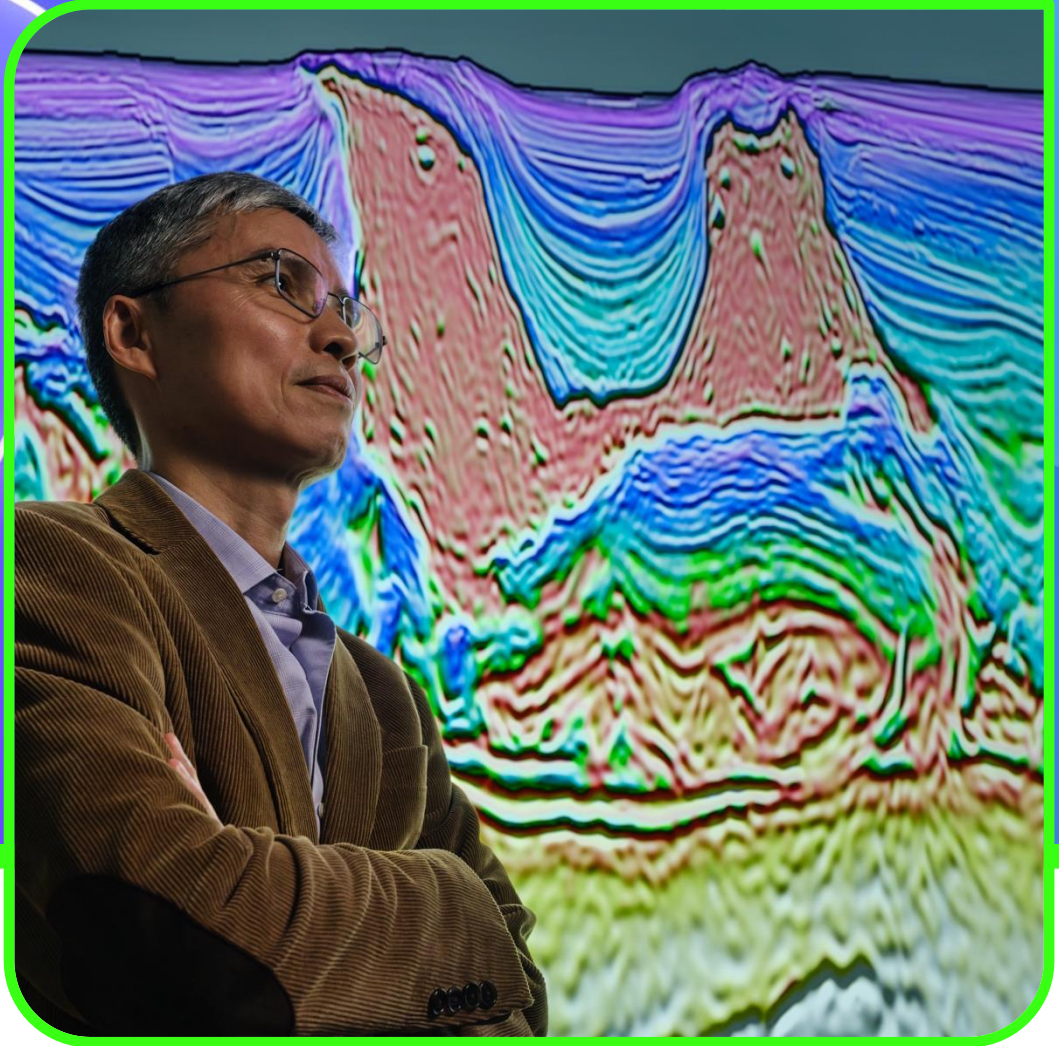
# 300PB+ of online storage

Data, the management,  
and storage of it,  
underpins everything  
Viridien does

- 300PB+ spread across 440 individual storage islands
- Each island consists of 2 or more head nodes and attached raid array
- In-house software manages placement, movement and data lifecycle.
- Goals
  - Consolidate these 440+ storage islands in to single-digit ceph clusters.
  - Increase resilience
  - Decrease operator overhead



# OUR BENCHMARKING PROCESS

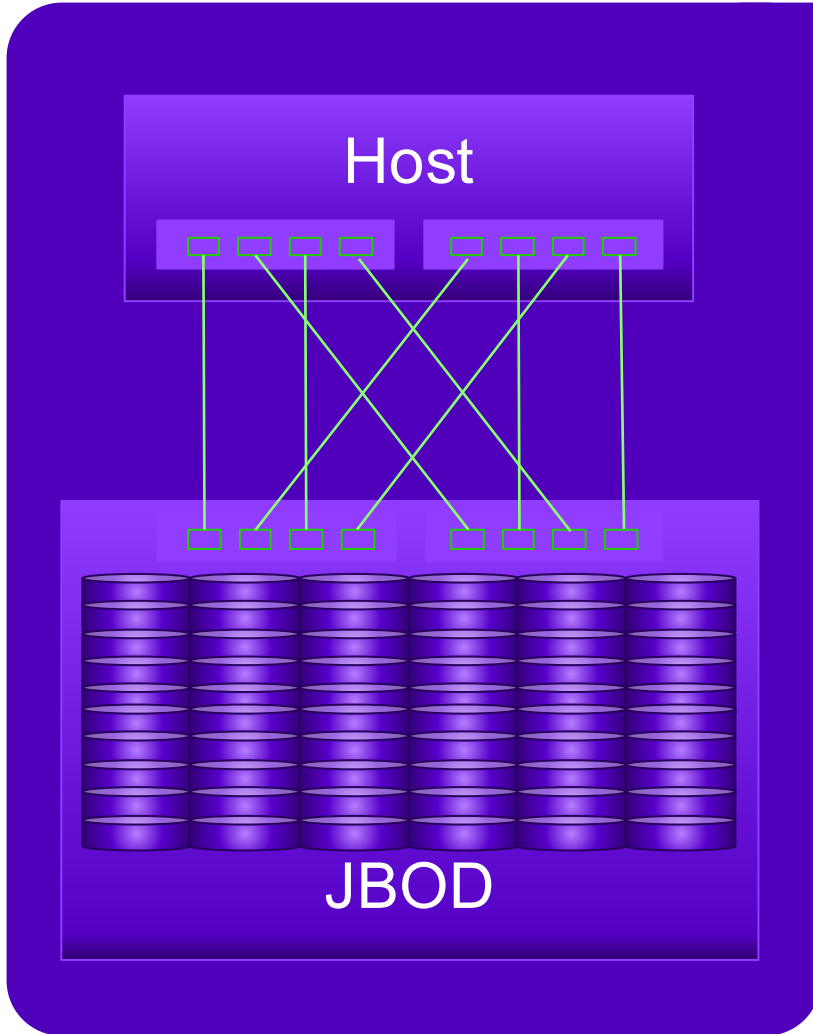


# Our Benchmarking process – walking the stack



- Low level validation
- OSD bench
- RADOS bench
- Client workload

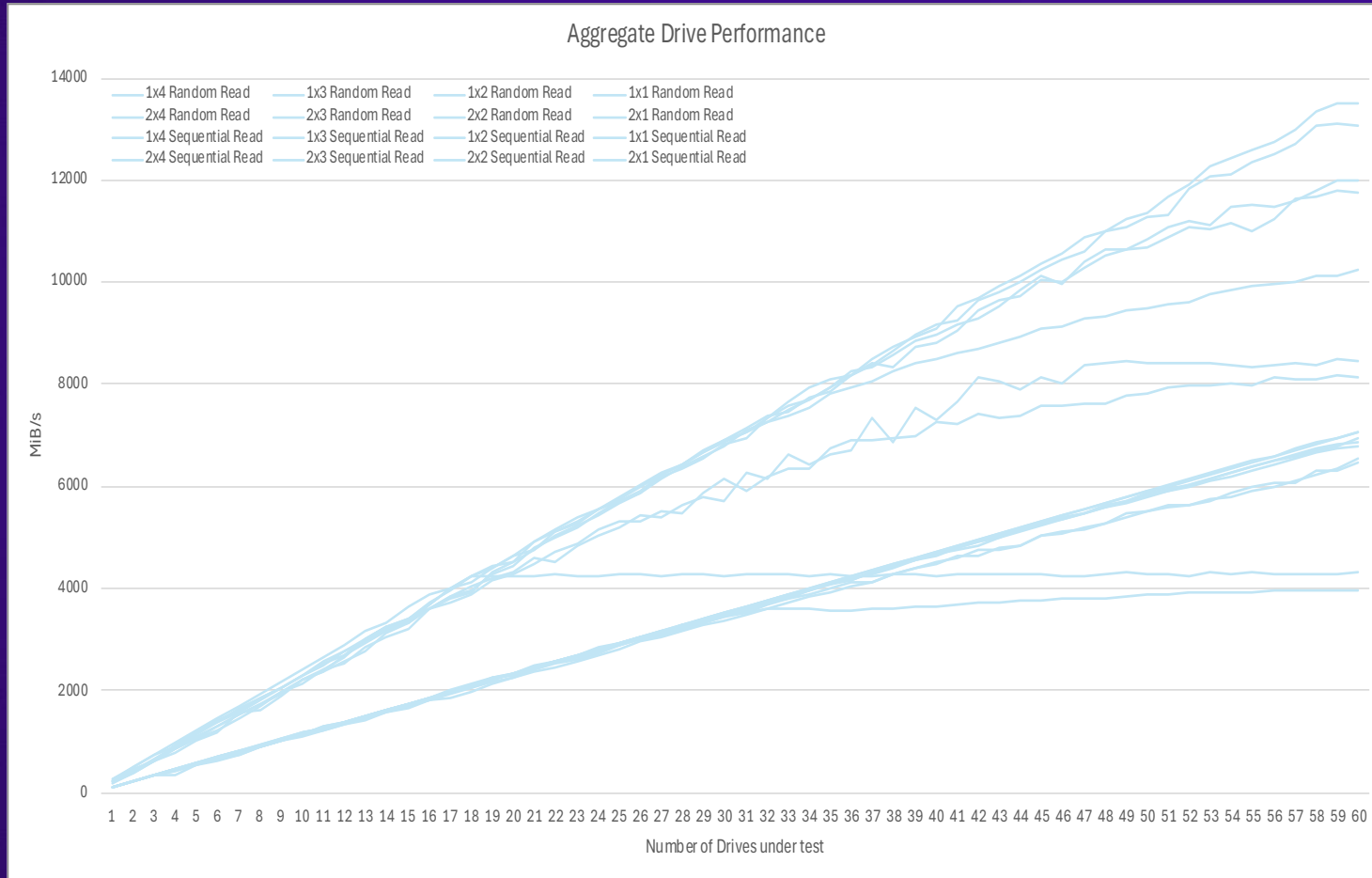
# JBOD system – SAS topology



- 2x Intel(R) Xeon(R) Gold 6430 (32c/64t)
- 256GB
- 2x Broadcom / LSI 9500-16e Tri-Mode HBA
- 1 60-bay JBOD
  - 2x 4 port “controllers”
  - 60x 12TB
- Cross-connected to avoid SPOF, and provide maximum resiliency



# Benchmarking Process – Low level - JBOD



## Fio 4M tests

Random read, for more realistic drive performance

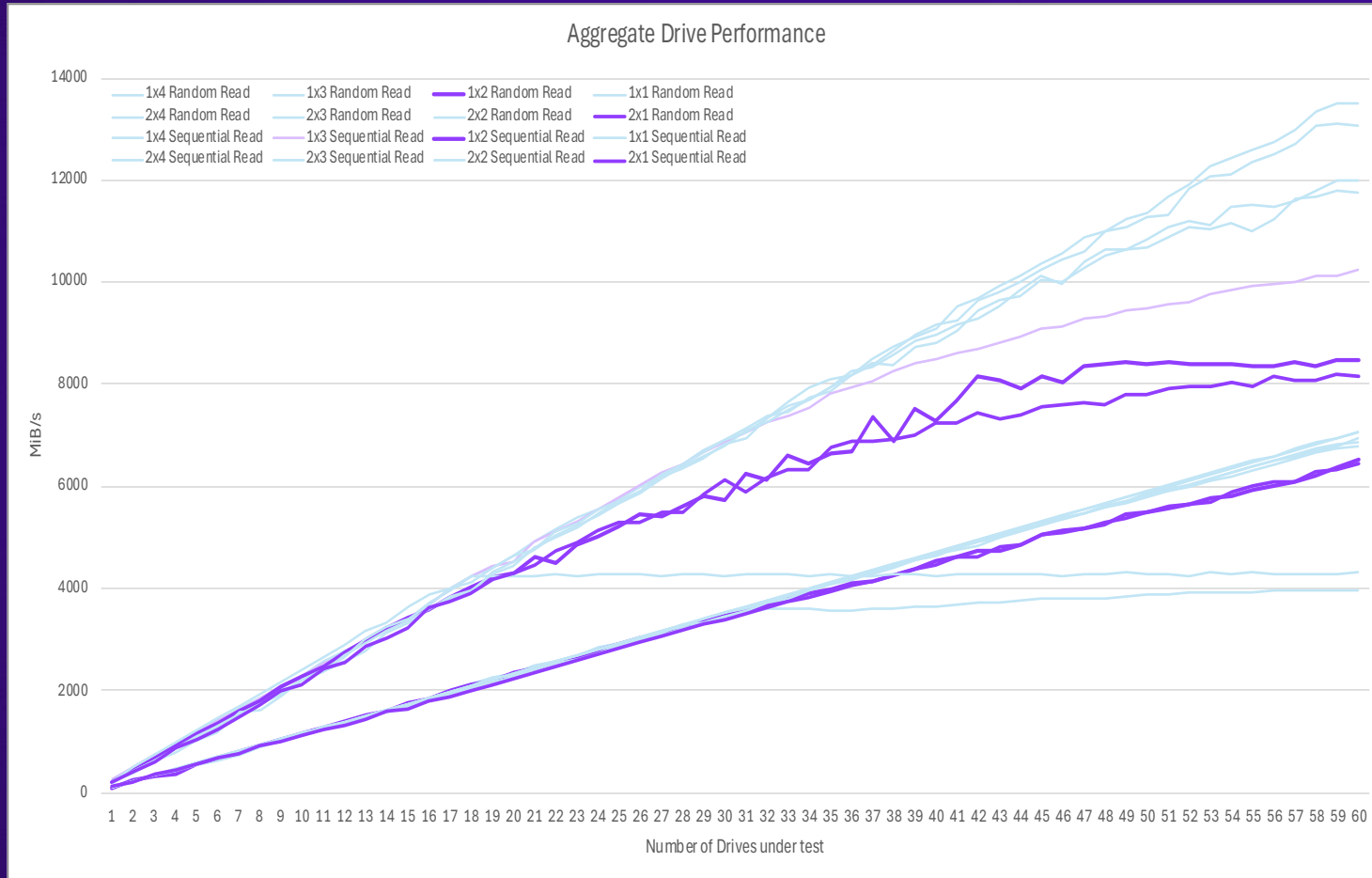
Sequential read, for max throughput, can highlight bottlenecks in SAS chains.



## 1x1 shows clear plateau



# Benchmarking Process – Low level



## Fio 4M tests

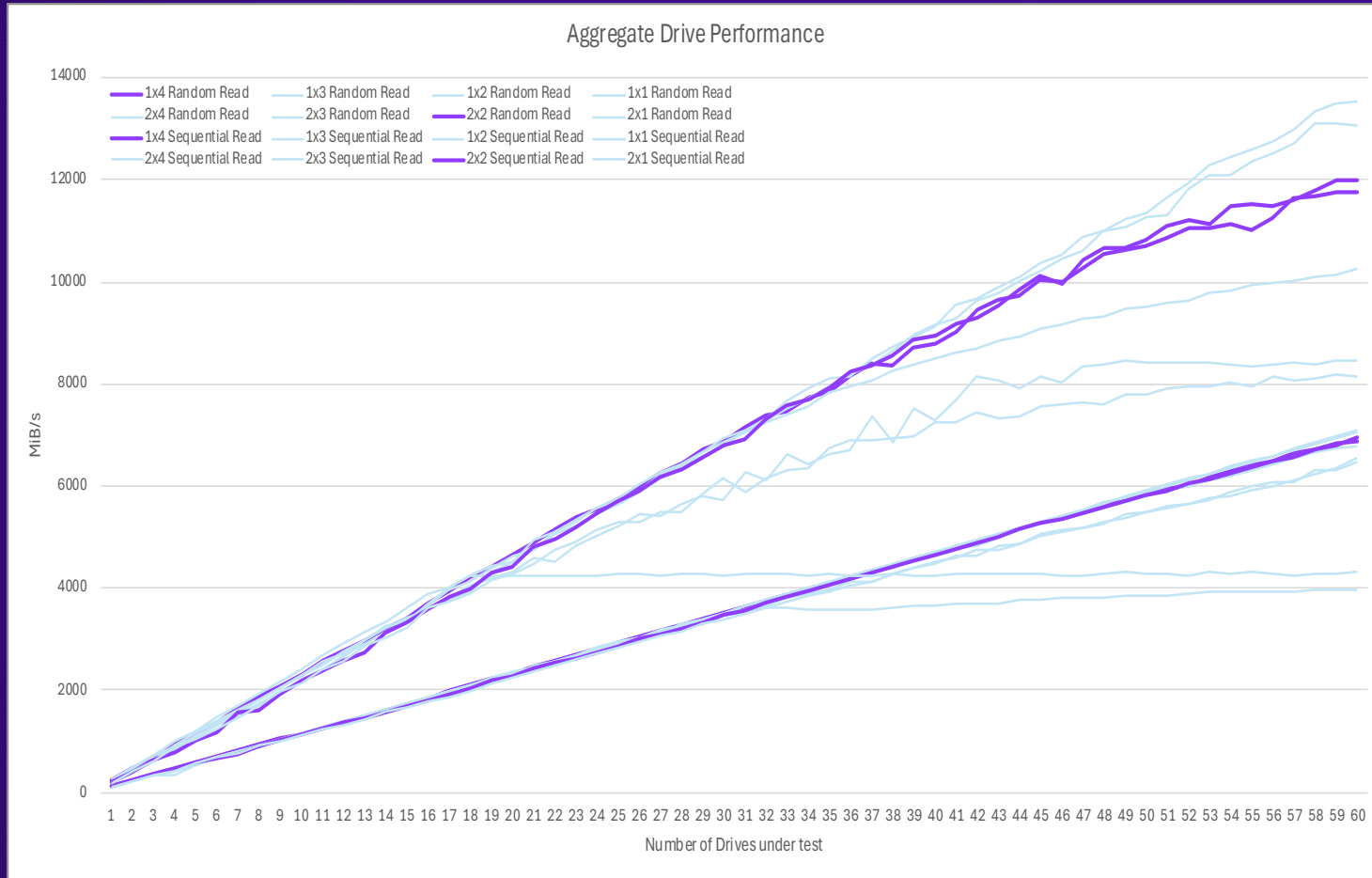
Sequential read, for max throughput, can highlight bottlenecks in SAS chains.

1x1 shows clear plateau

Adding additional cable(s) shows clear performance benefit, and raises plateau



# Benchmarking Process – Low level



## Fio 4M tests

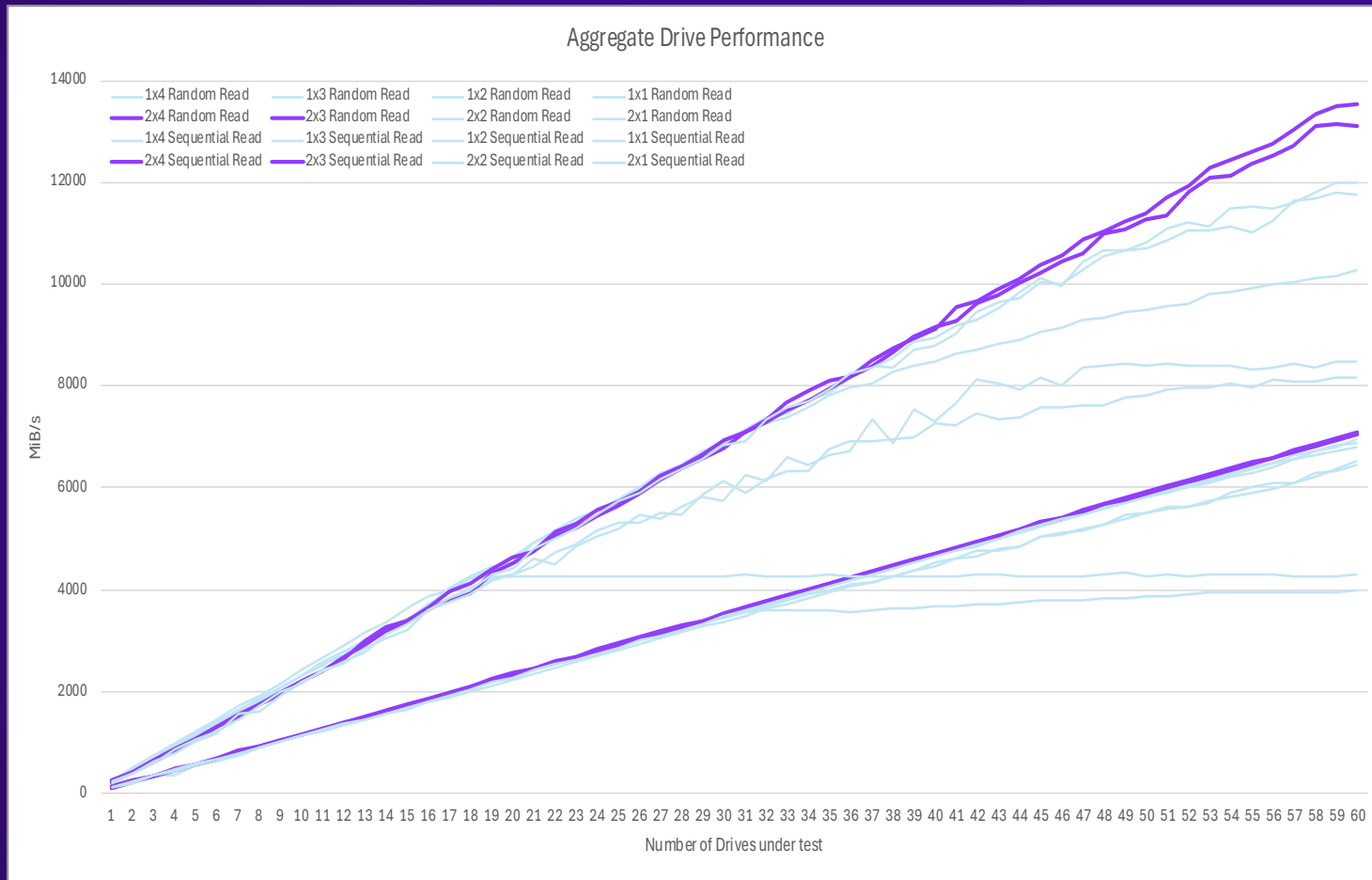
Sequential read, for max throughput, can highlight bottlenecks in SAS chains.

1x1 shows clear plateau

Adding additional cable(s) shows clear performance benefit, and raises plateau

minor difference between 2x2 and 1x4 configuration

# Benchmarking Process – Low level



## Fio 4M tests

Sequential read, for max throughput, can highlight bottlenecks in SAS chains.

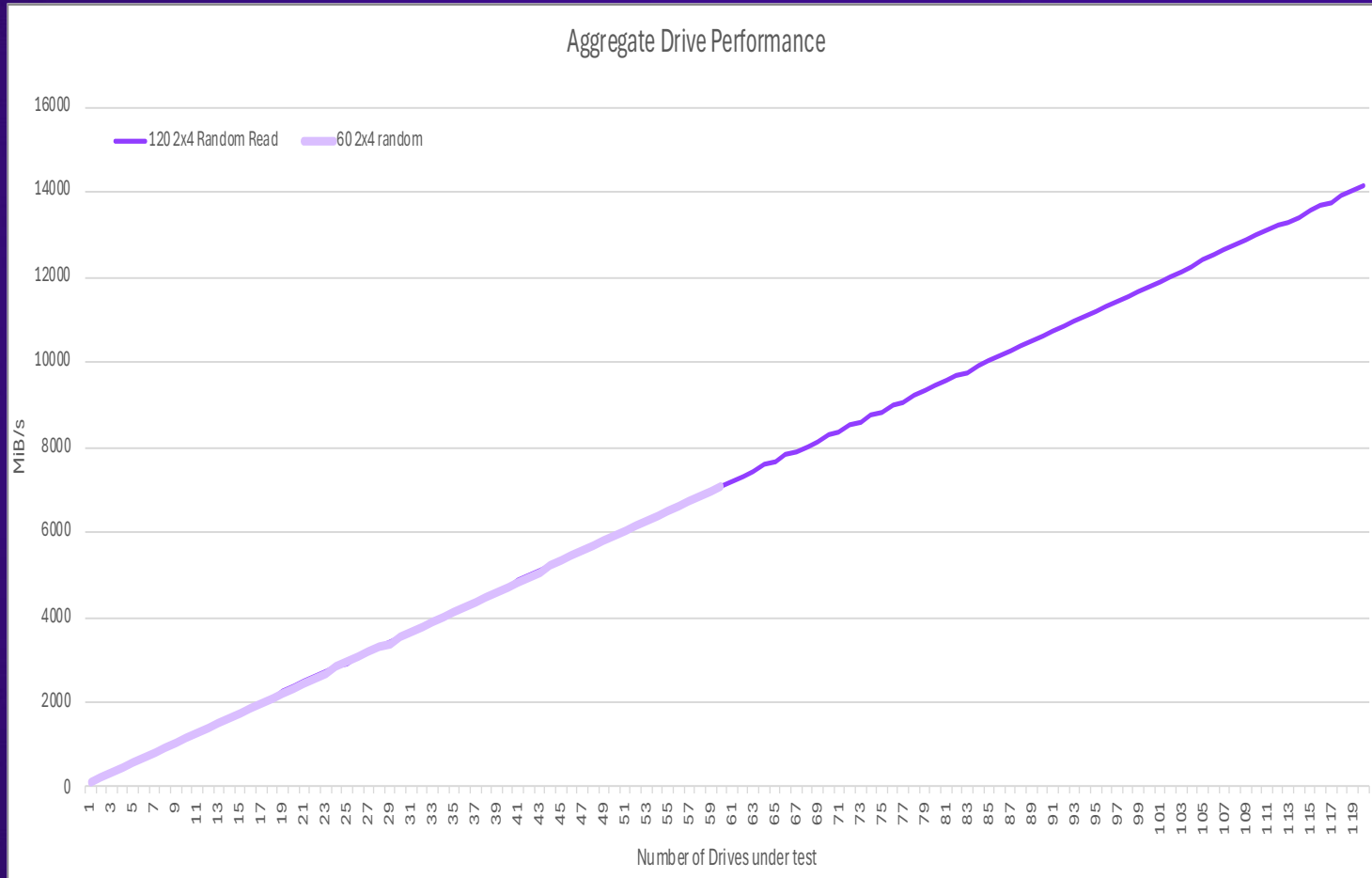
1x1 shows clear plateau

Adding additional cable(s) shows clear performance benefit, and raises plateau

minor difference between 2x2 and 1x4 configuration

2x3 and 2x4 show minimal evidence of plateau

# Benchmarking Process – Low level



## Fio 4M tests

Extending tests to 120 drives shows random tests continue to scale well.

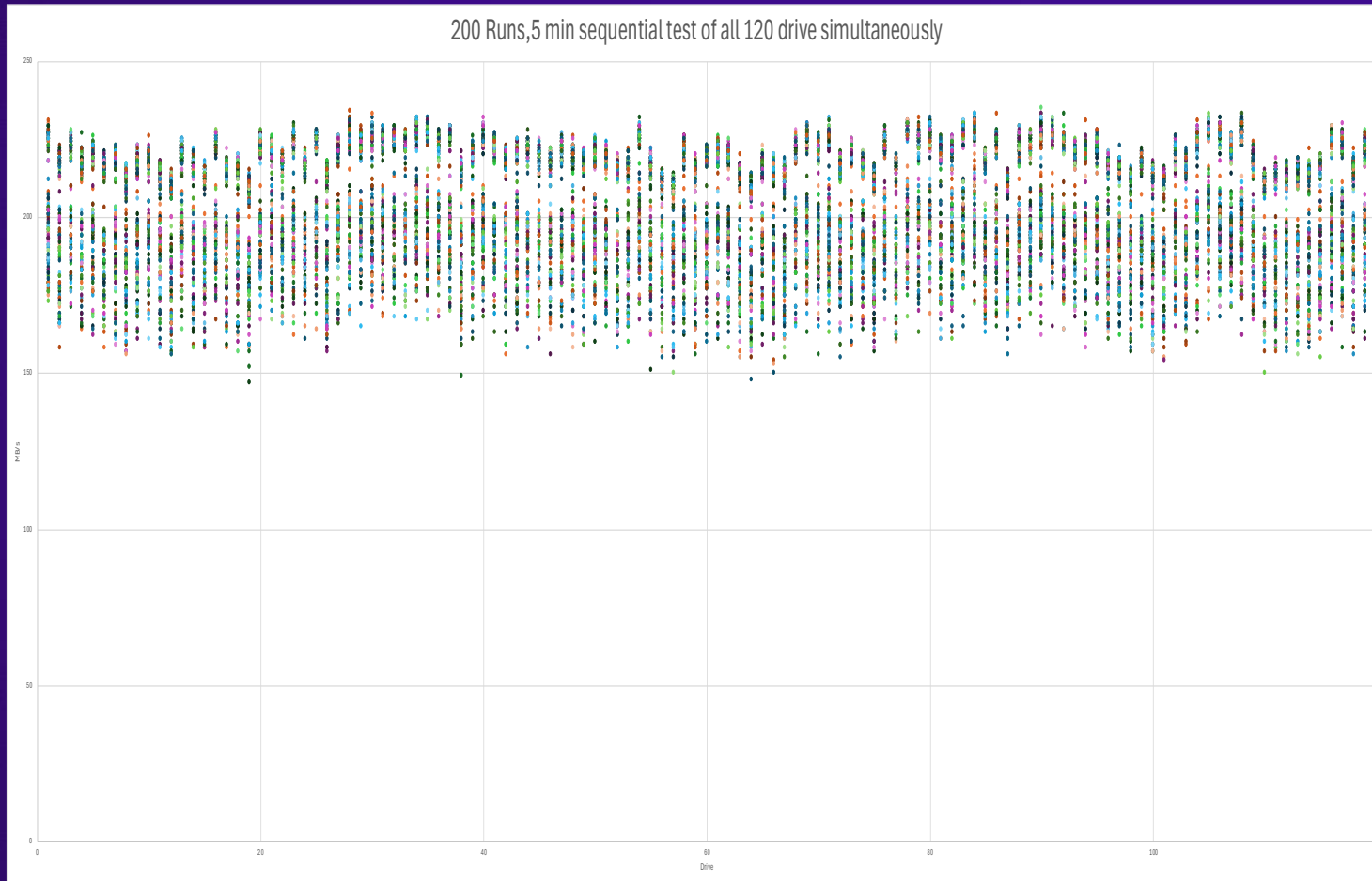
# Benchmarking Process – Low level



## Fio 4M tests

However, a new problem emerged, and we stepped back to smokescreen each drive, some drives had outlying behaviors,

# Benchmarking Process – Low level



## Fio 4M tests

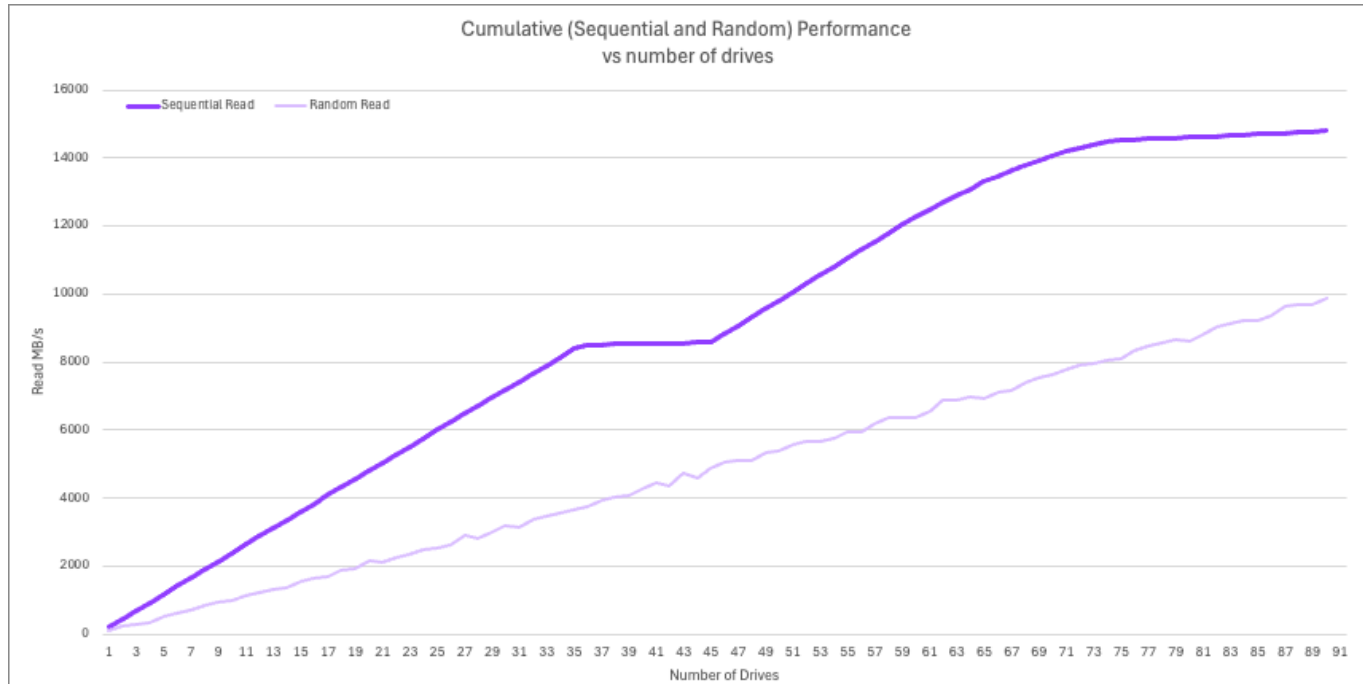
However, a new problem emerged, and we stepped back to smokescreen each drive, some drives had outlying behaviors,

Swapping suspect drives led us to a new pattern.

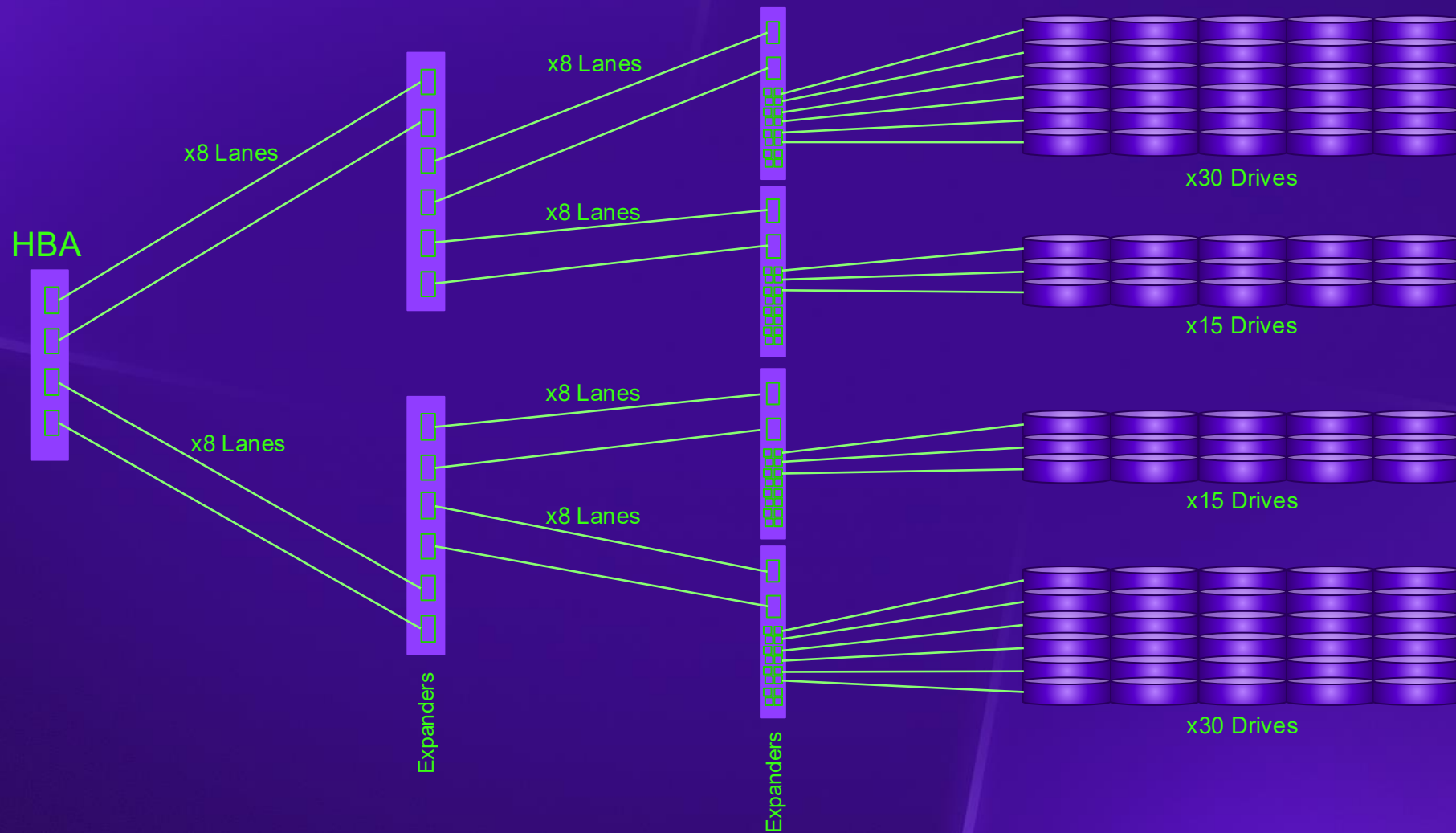


# Benchmarking Process – Low level - HCI

- Intel(R) Xeon(R) 6731E (96 E-core)
- 512GB system memory
- 90x 12TB data
- 2x SATA SSD (boot)
- 1x Broadcom RAID card
- 1x Broadcom SAS38xx based 16-port internal card



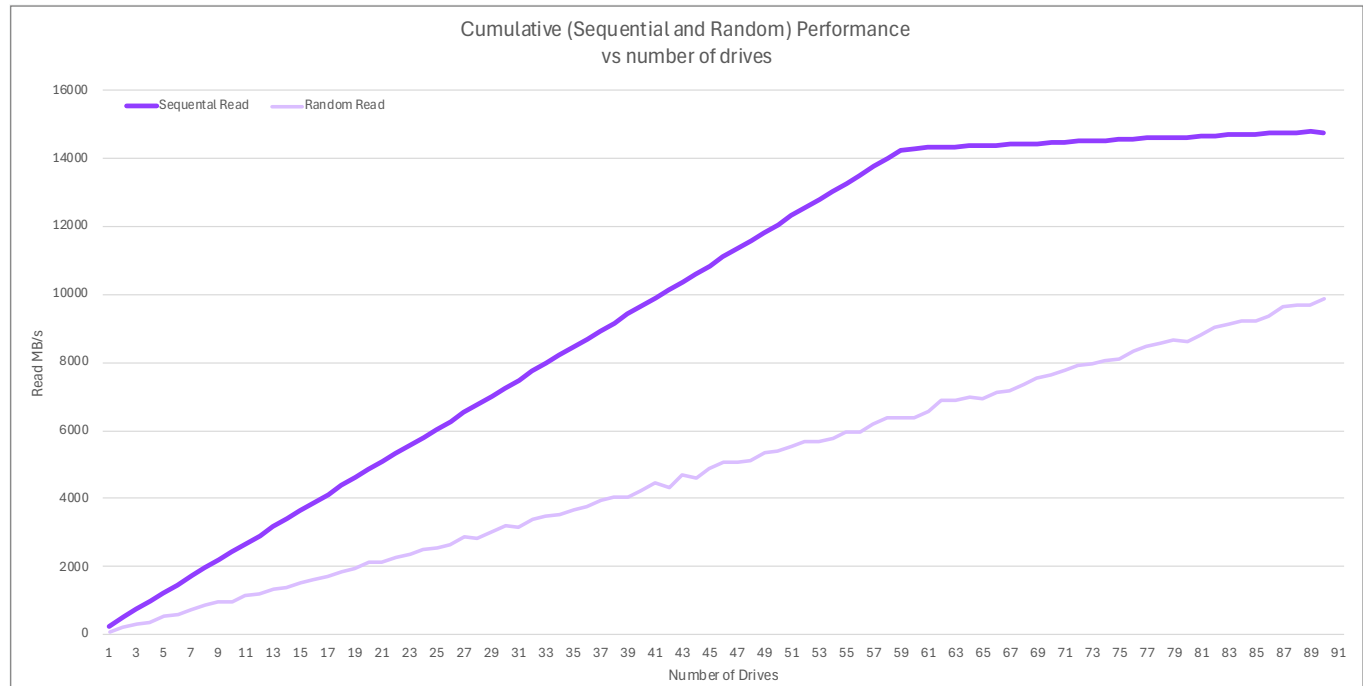
# Benchmarking Process – Low level - HCI





# Benchmarking Process – Low level - HCI

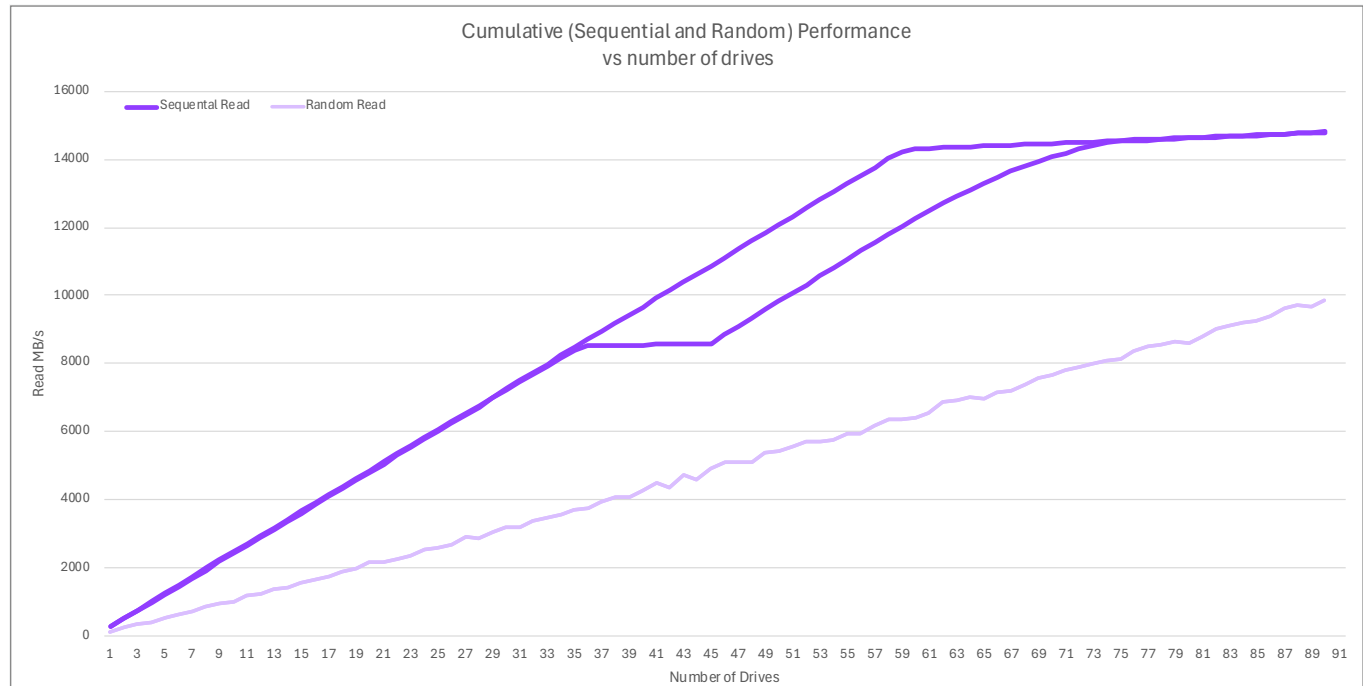
- Expanders can be a bottleneck too
- Reorder access, move plateau





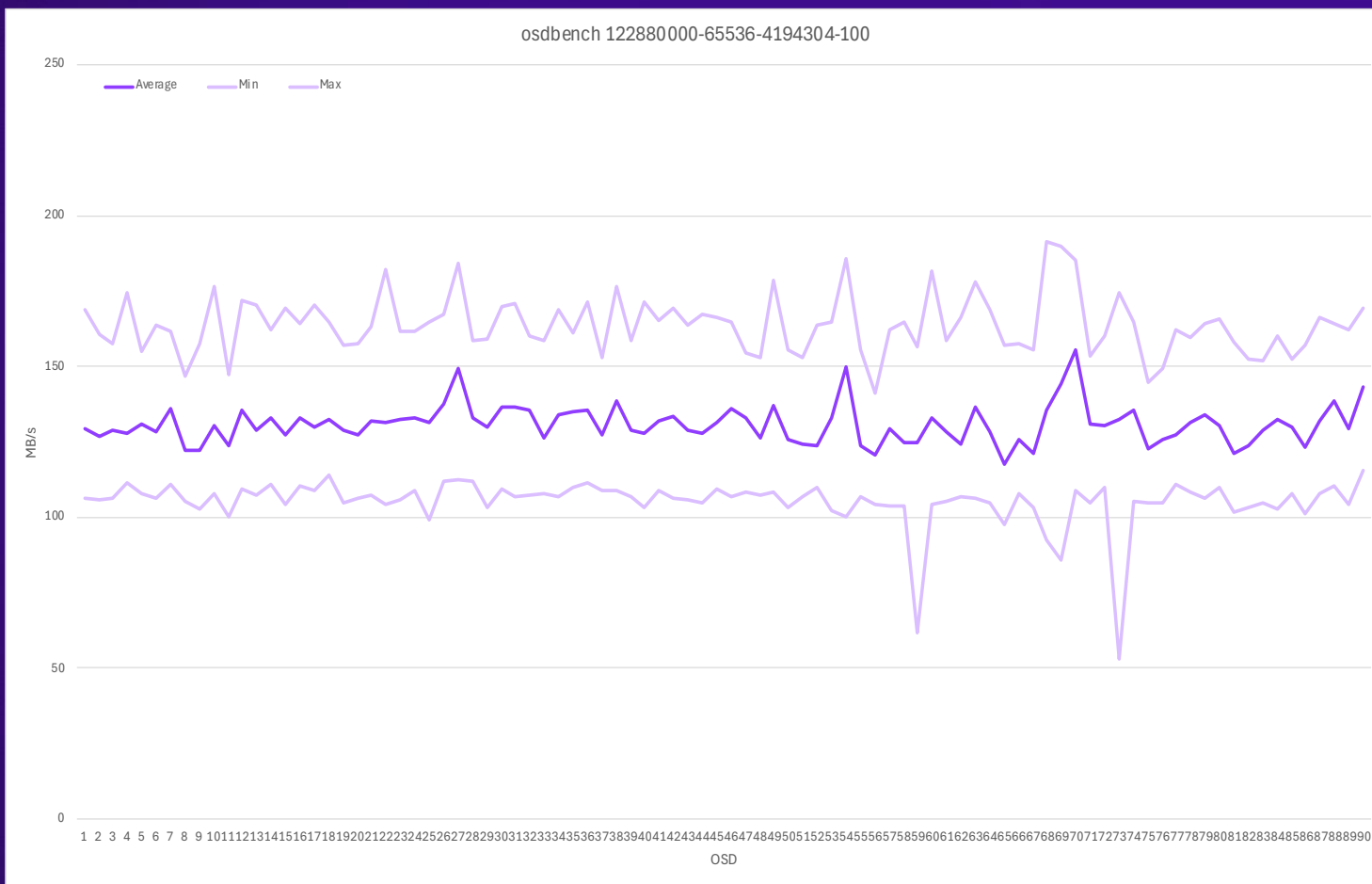
# Benchmarking Process – Low level - HCI

- Expanders can be a bottleneck too
- Reorder access, move plateau





# Benchmarking Process – OSD Bench - HCI



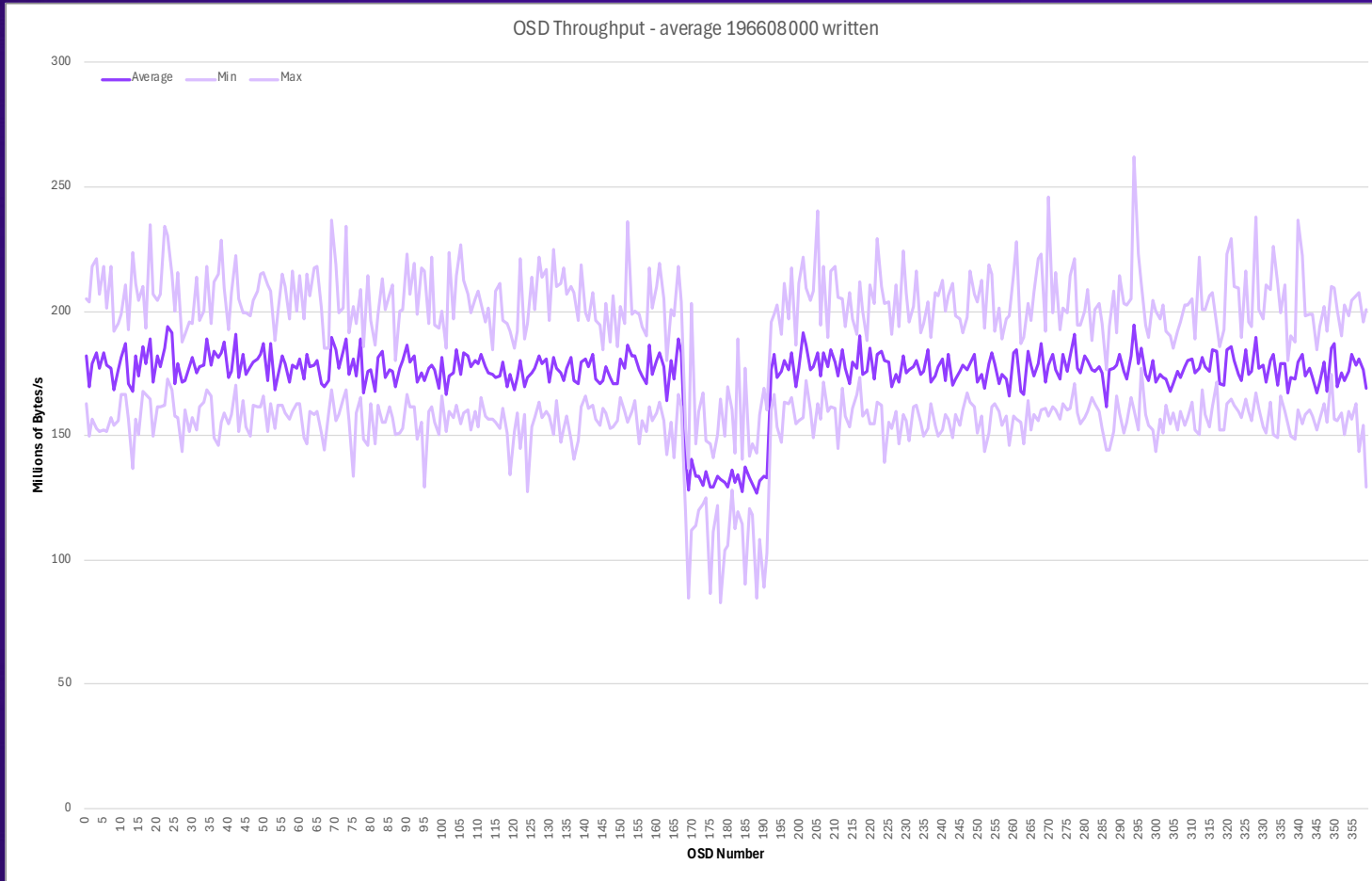
## OSD bench

HCI host results

Sum of average= 11,722MB/s



# Benchmarking Process – OSD Bench - Cluster



## OSD bench – test cluster

16 hosts, 24 drives each

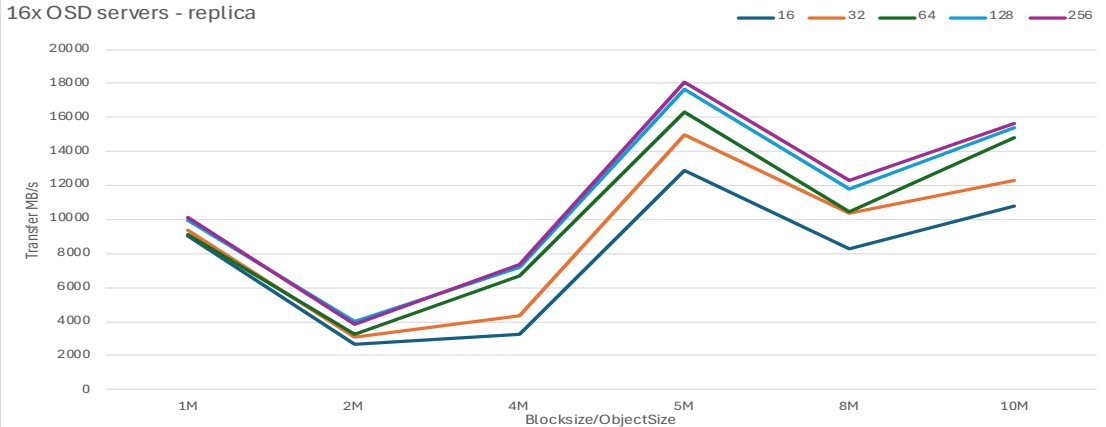
1 node of 24 drives was  
underperforming

Drives and cables reseated.

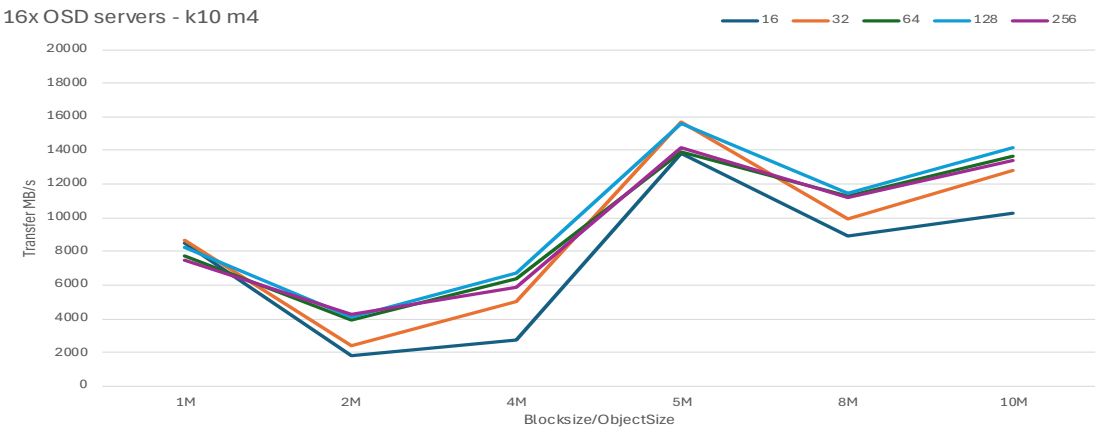


# Benchmarking Process – Rados Bench

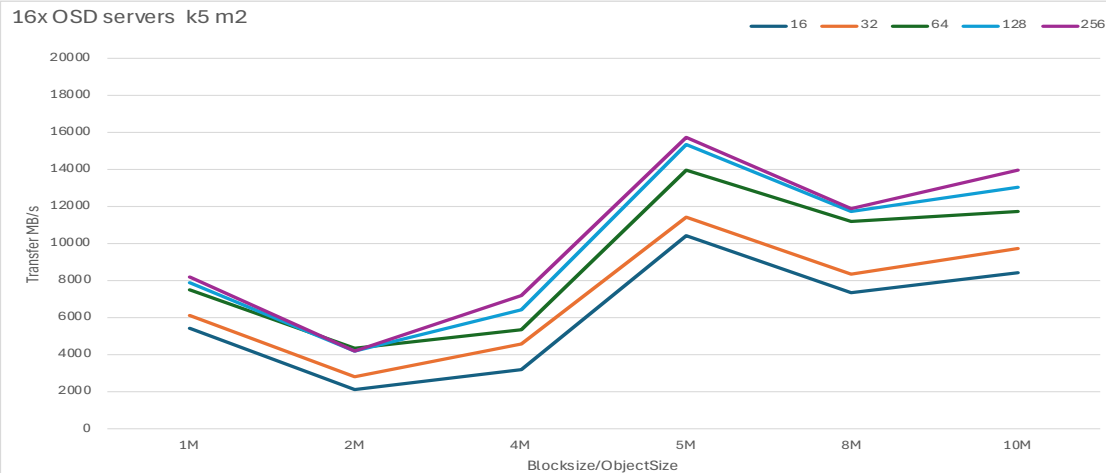
16x OSD servers - replica



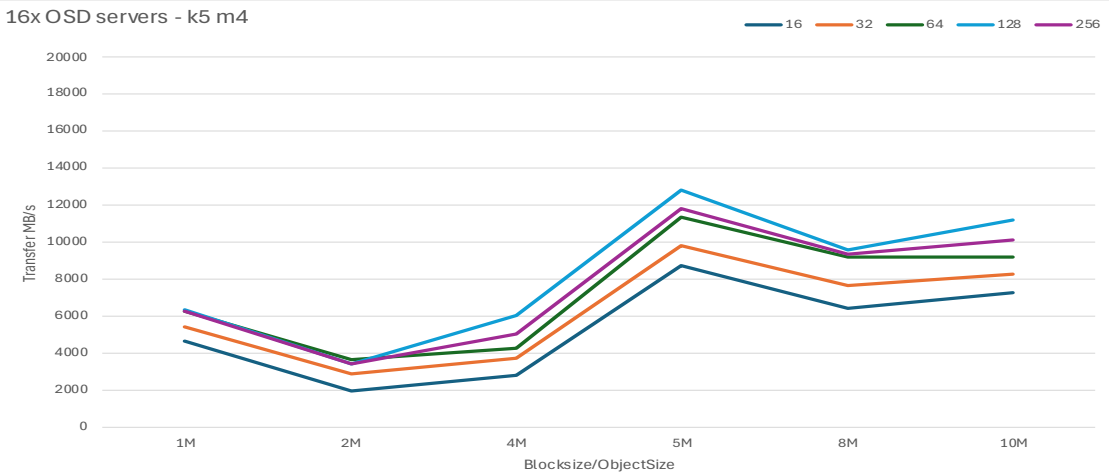
16x OSD servers - k10 m4



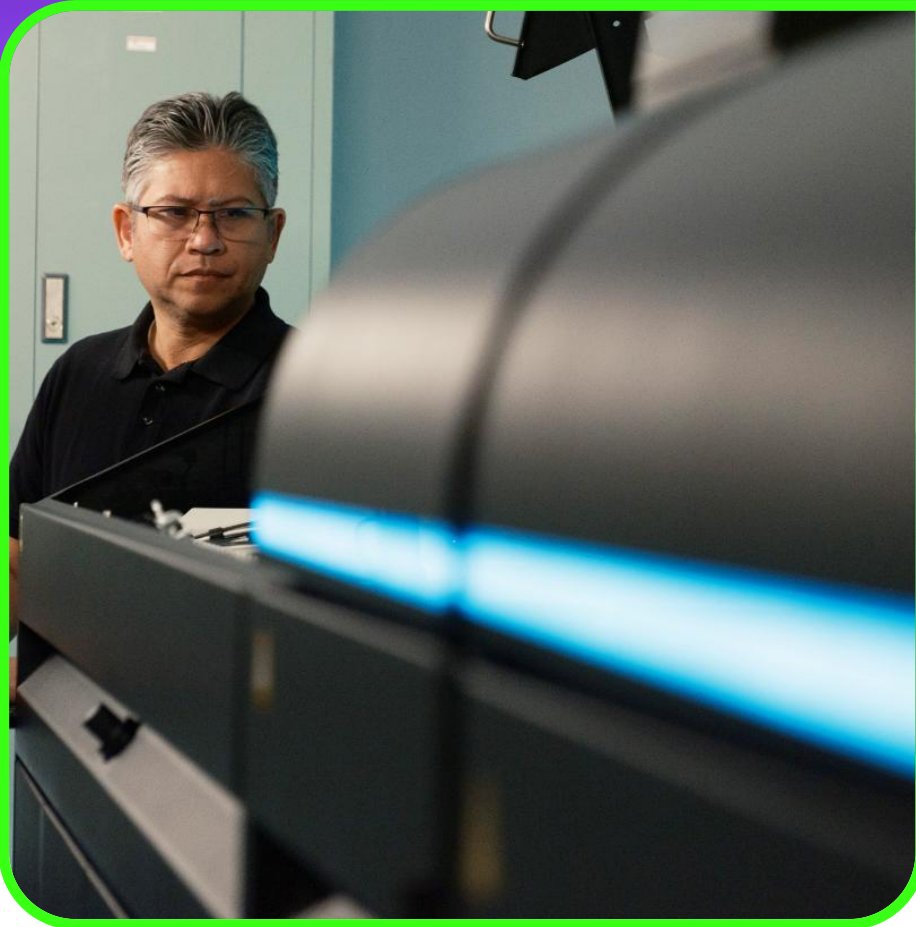
16x OSD servers k5 m2



16x OSD servers - k5 m4



Each OSD server = 24 OSD



# OBSERVATIONS



# The Metadata problem

## BlockDB/WAL/Data

Recommendation: separate BlockDB, WAL and Data devices

Problem: 3x number of devices in single node, 24-bay storage node becomes 8 OSD

Solution: locate multiple BlockDB/WAL on faster devices

- but what is the ratio?

# The metadata problem



## BlockDB/WAL/Data

Recommendation:  
separate BlockDB, WAL  
and Data devices

Problem: 3x number of  
devices in single node, 24-  
bay storage node becomes  
8 Data

Solution: Colocate BlockDB  
and WAL, share 1 NVMe  
with multiple Data drives



# JBOD – Lots of devices – Raw device scanning

## The problem

- *Squid* onwards was not starting OSDs on a highly connected (1x4 multipath) JBOD system
- Systems with multiple paths can take a long time for OSD to startup
- 2 very similar bugs, (<https://tracker.ceph.com/issues/73107> and <https://tracker.ceph.com/issues/74804>)
- Mikael Öhman posted <https://www.mail-archive.com/ceph-users@ceph.io/msg31257.html>
- Followed his example and rebuilt a new container with the call to *raw activate* removed,
  - `sed -i '53,59s/^/ #/' /usr/lib/python3.12/site-packages/ceph_volume/activate/main.py`
- Further investigations show internal communication/output capture isn't working correctly

# Erasure coding - HCI



## Performance

Increases as number of drives in pool increases

Increases as number of active drives increases

Up to ~60drives

# Erasure coding - HCI

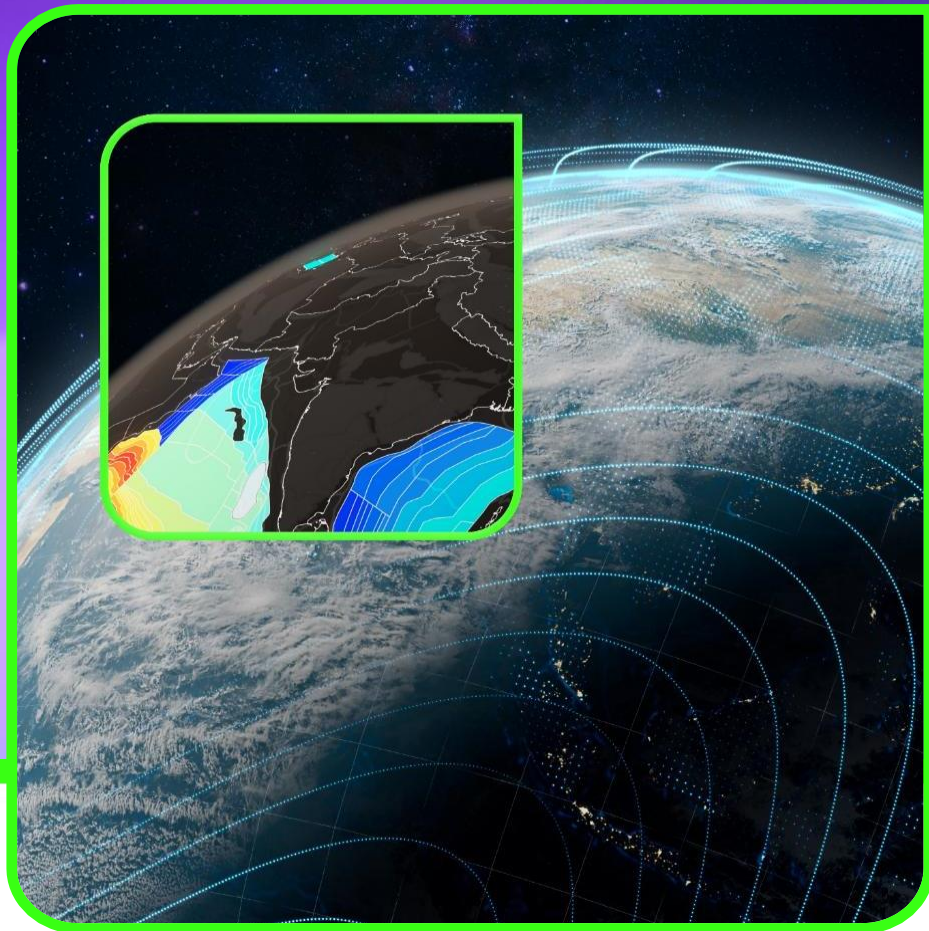


## Performance

Increases as number of drives in pool increases

Increases as number of active drives increases

This holds for Write and Random workloads but in this system sequential performance suffers at higher counts



# CLIENT LOADS



# THANK YOU

Presentation for:  
Ceph Day Zürich  
October 14<sup>th</sup>, 2026

Contact:  
Colin Morey  
HPC Architect  
[Colin.Morey@viridiengroup.com](mailto:Colin.Morey@viridiengroup.com)  
[www.viridiengroup.com/](http://www.viridiengroup.com/)