



# Scalable and Performant All Flash Ceph Storage with NVMe-oF™

Presented by:

David Tobin - Technical Marketing Engineer  
Platforms Group - Western Digital

# Overview

---



Flash Memory Summit

- Test Environment description
- NVMe-oF Connectivity and integration
- Initial tuning and results
- Enhanced tuning and results
- NVMe-oF recommendations

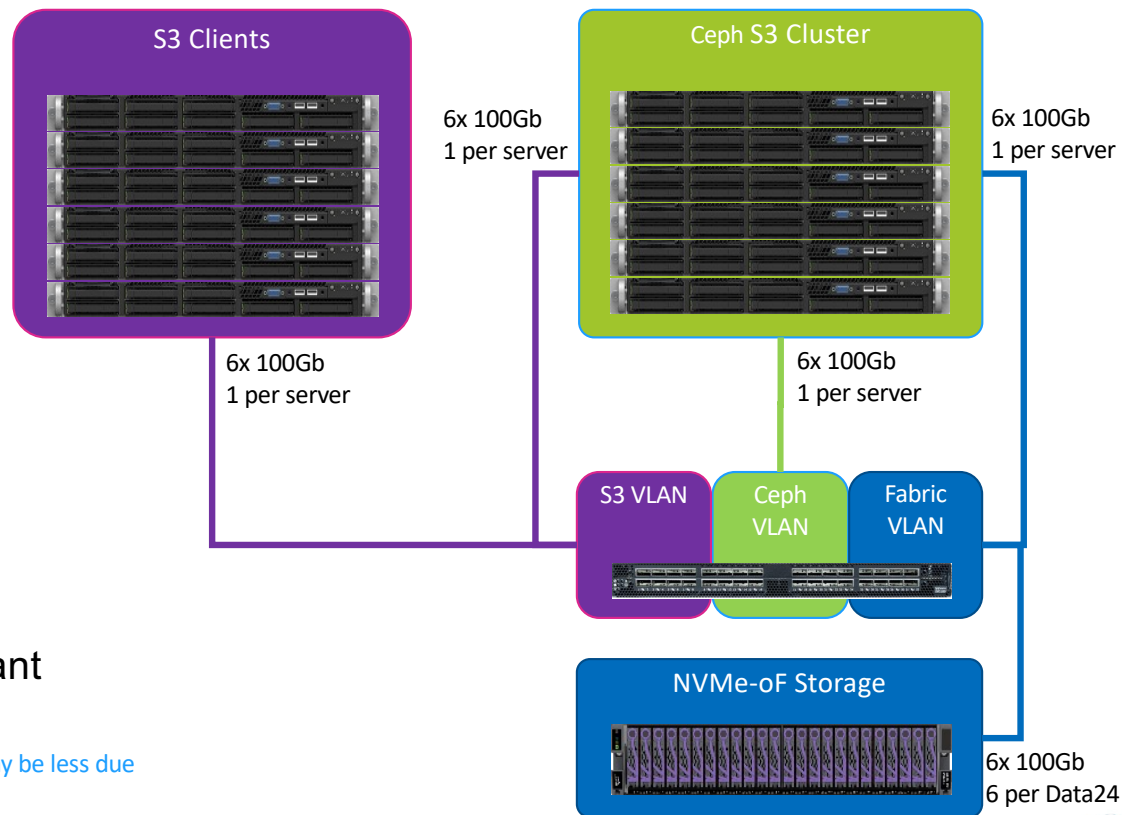
# Test Environment



Flash Memory Summit

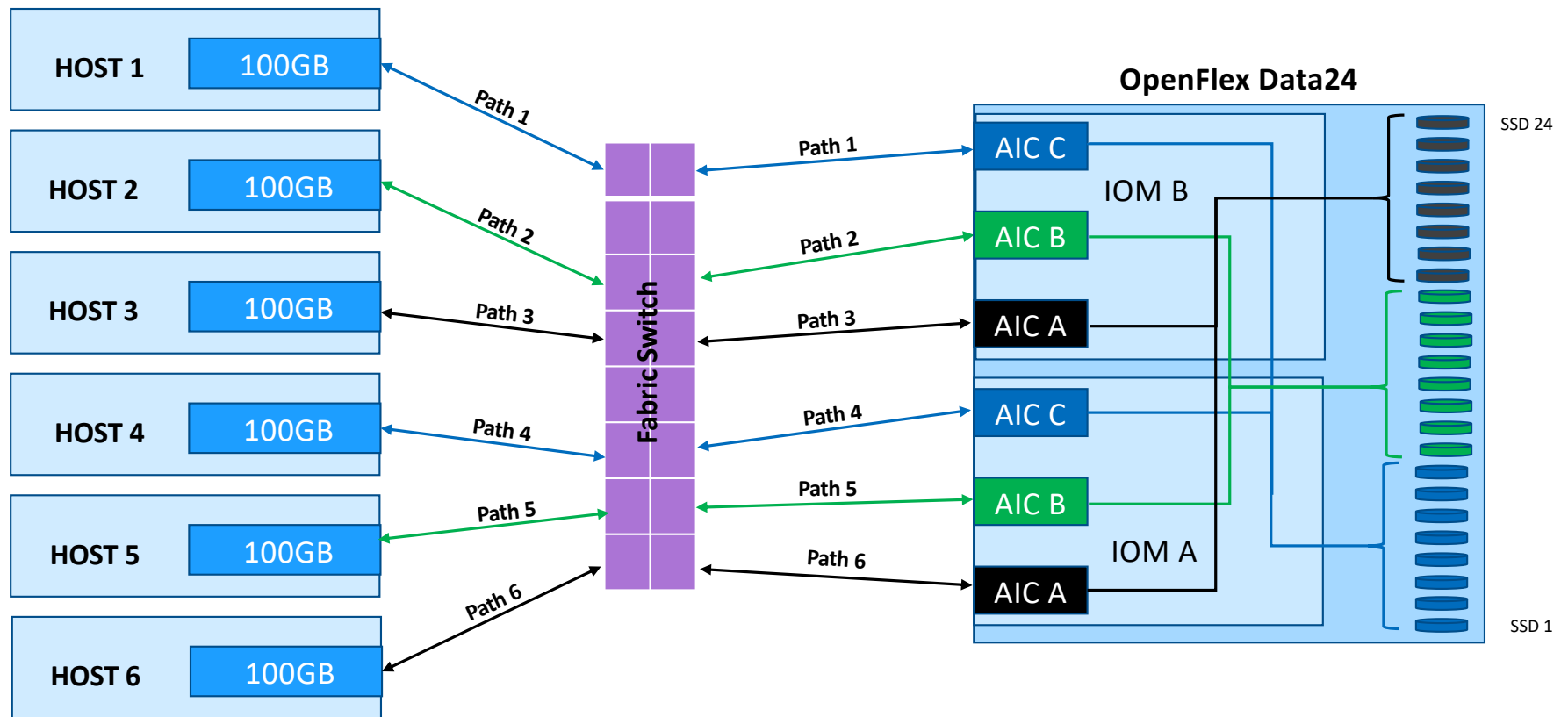
- **6 Clients & 6 Server**
  - 2x Dual port 100 Gb ConnectX® 6
  - 2x Xeon® Gold 6354 (18 core 3GHz)
  - Ubuntu 20.04
  - 512GiB RAM
- **1 Mellanox® SN2700 Switch**
  - 12ports S3 VLAN
  - 12ports Fabric VLAN
  - 6 ports Cluster CLAN
- **1 OpenFlex™ Data24**
  - 24x 3.2TB<sup>1</sup> NVMe™ Devices
  - 6x NVMe-oF Target Interfaces
  - 2x Independent IOMs for fully redundant paths to every NVMe Device

1. One terabyte (TB) is equal to one trillion bytes. Actual user capacity may be less due to operating environment.





# Matterhorn non-HA Switched Topology



# NVMe-oF and RoCE v2 Pre-requisites

---

- **Configure Fabric switch and NICs for Lossless Ethernet**  
Guides available on request
- **Configure IPs on Fabric NIC**  
Use same subnet as NVMe-oF target
- **Verify lossless setting and performance between servers**  
use `ib_write_bw` to measure lossless performance
- **Recommended to upgrade nvme-cli to version 1.15+**
- **Verify Connectivity to Fabric IPs of the OpenFlex Data24**

# NVMe-oF and RoCE v2 Connection

- Identify NVMe device subnqn

```
nvme discover -trdma -s4420 -a<data24 IP>
```

```
pfedlr650-3 ~ # nvme discover -trdma -s4420 -a192.168.1.11 | grep subnqn
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.1
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.2
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.3
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.4
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.5
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.6
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.7
subnqn: nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.8
pfedlr650-3 ~ #
```

- Connect command

```
nvme connect -trdma -s4420 -a<data24 fabric IP> -n <subnqn of device>
```



# NVMe-oF and RoCE v2 Connection

- Verify connectivity over all Fabric NICs

```
nvme list -v
```

```
[pfedlr650-3 ~ # nvme list -v
NVM Express Subsystems
```

Subsystem	Subsystem-NQN	Controllers
nvme-subsys0	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.1	nvme0, nvme8
nvme-subsys1	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.2	nvme1, nvme9
nvme-subsys2	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.3	nvme10, nvme2
nvme-subsys3	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.4	nvme11, nvme3
nvme-subsys4	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.5	nvme12, nvme4
nvme-subsys5	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.6	nvme13, nvme5
nvme-subsys6	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.7	nvme14, nvme6
nvme-subsys7	nqn.1992-05.com.wdc.openflex-data24-usalp04621qa0003:nvme.8	nvme15, nvme7

# Ceph integration and Configuration

- Ceph can now access NVMe-oF devices as if they were local
- Set objectstore to bluestore  
`osd_objectstore = bluestore`
- Increase SSD Memory and Cache  
`osd_memory_base = 8G`  
`osd_memory_target = 12G`  
`bluestore_cache_size_ssd = 8G`
- Increase threads per OSD  
`osd_op_num_threads_per_shard = 4`
- Set Rados Gateway to Beast  
`rgw_frontends = beast endpoint=192.168.1.122:7480 tcp_nodelay=1`



# Read Performance



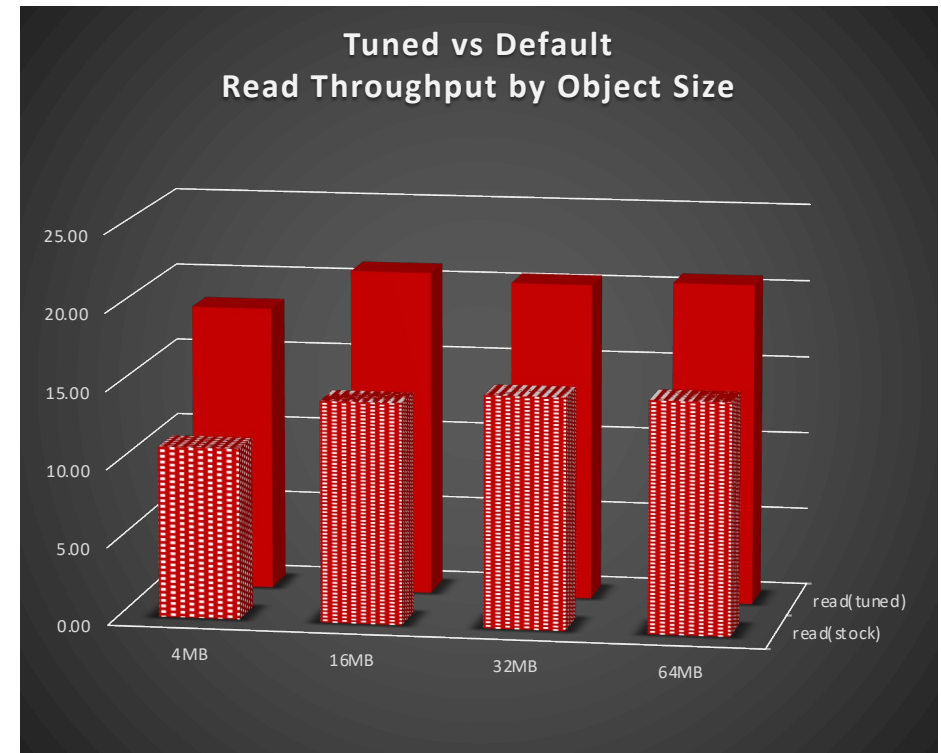
Flash Memory Summit

- **Default Configuration Results**

- 16 OSDs per Node
- 26-33% CPU Utilization per Node
- 12-16 CPU Cores active
- 18% Memory Used
- Average 30% slower

- **Tuned Configuration Results**

- 50-75% CPU Utilization per Node
- 28-32 CPU Cores Active
- 45% Memory Used



# Results



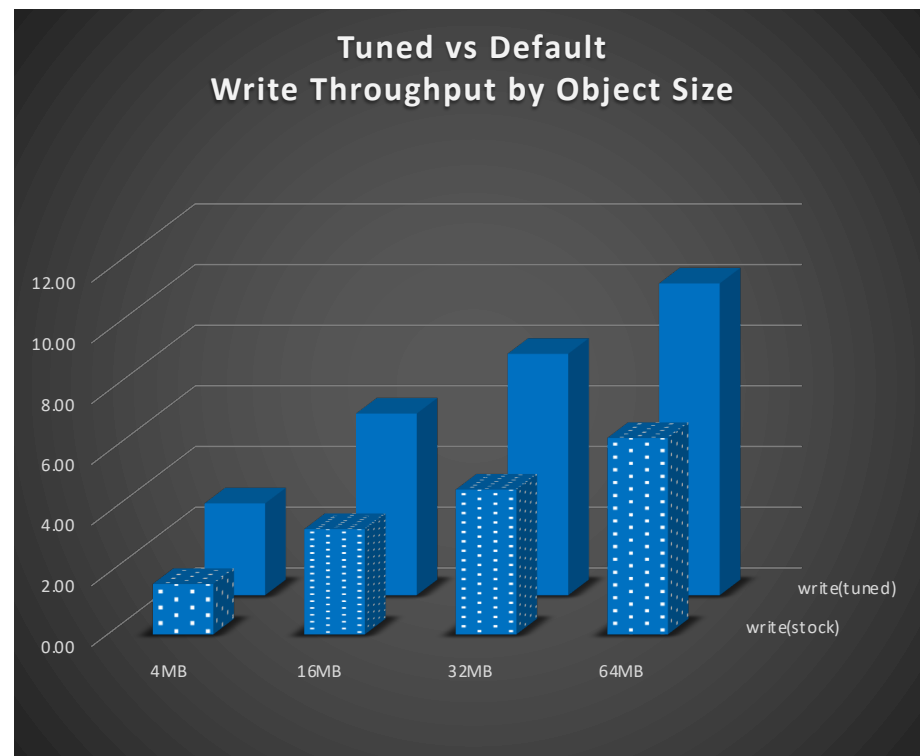
Flash Memory Summit

- **Default Configuration Results**

- 16 OSDs per Node
- 48-63% CPU Utilization per Node
- 18-24 CPU Cores active
- 25% Memory Used
- Average 45% slower

- **Tuned Configuration Results**

- 80-100% CPU Utilization per Node
- 36 CPU Cores Active
- 67% Memory Used



## What's next

---

- Current focus is on customer deployments of this solution
  - This will be a fully redundant network setup
  - 24 OSDs per node (numbers used in my testing based on this value)
  - NVMe native load balancing will also be used
- Next bottle neck is believed to be the RadosGW
  - Likely requires systems with more CPU Cores is required
- More settings to be evaluated

# Questions?

Western Digital and OpenFlex are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. Ceph is a trademark or registered trademark of Red Hat, Inc. or its subsidiaries in the United States and other countries. Intel and Xeon are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries. Mellanox and ConnectX are registered trademarks of Mellanox Technologies, Ltd. The NVMe and NVMe-oF word marks are trademarks of NVM Express, Inc.