Flash Memory Summit Session:

Storage System using NVMe over Fabric SSD-Based Ethernet JBOF

Woosuk Chung, Director, Memory Systems R&D

Legal Disclaimer

The information contained in this document is claimed as property of SK hynix. It is provided with the understanding that SK hynix assumes no liability, and the contents are provided under strict confidentiality.

This document is for general guidance on matters of interest only. Accordingly, the information herein should not be used as a substitute for consultation or any other professional advice and services.

SK hynix may have copyrights and intellectual property right. The furnishing of document and information disclosure should be strictly prohibited.

SK hynix has right to make changes to dates, product descriptions, figures, and plans referenced in this document at any time. Therefore the information herein is subject to change without notice.

CONTENTS

Introduction

Performance Evaluation

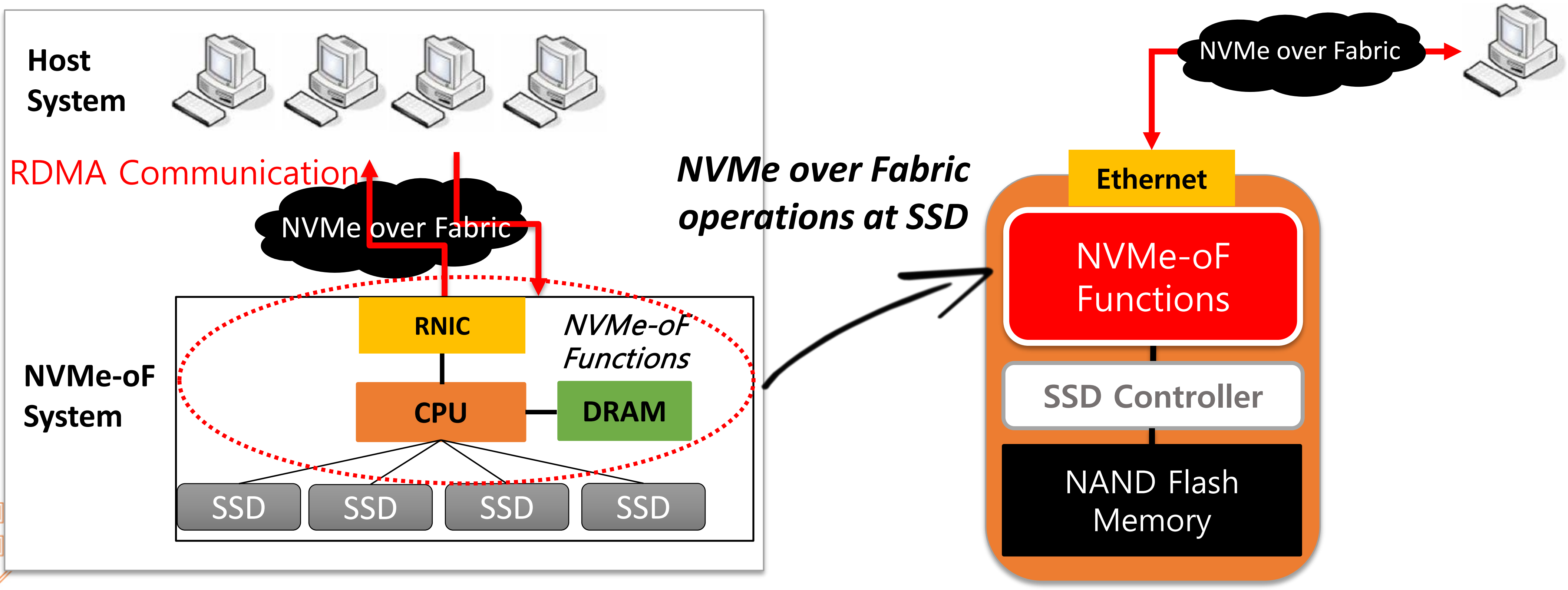
Ceph with NVMe-oF SSD

Summary

Introduction: What is NVMe over Fabric SSD ?

- **Moving NVMe over Fabric functions from storage system to SSD**

Conventional NVMe over Fabric System **NVMe over Fabric SSD**

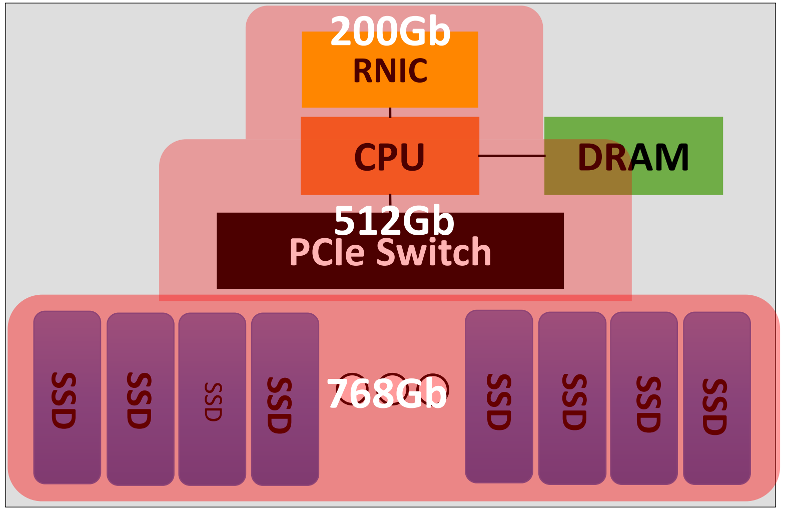


Introduction: Benefits

- High performance scaling
- Low power consumption, reduced cost

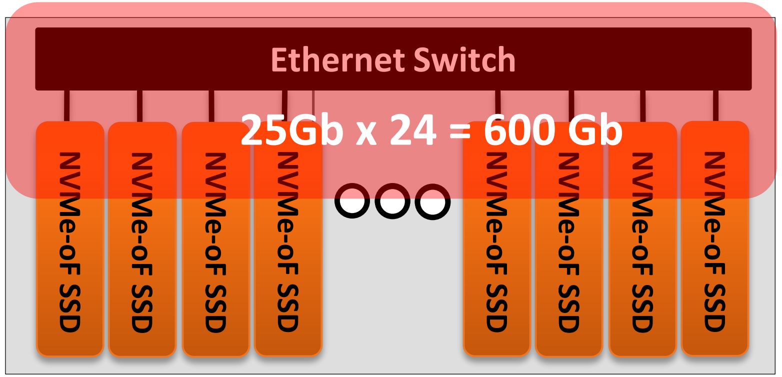
Conventional NVMe-oF JBOF

Dedicated CPU & RNIC
Additional Power Consumption



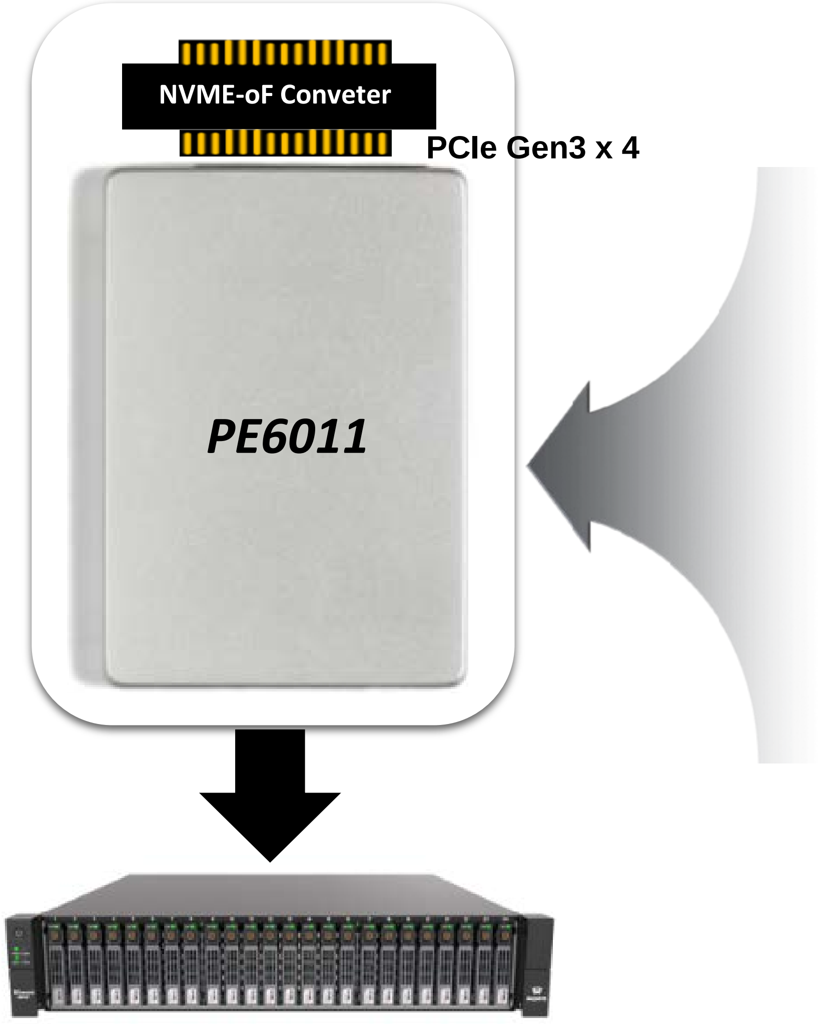
NVMe-oF SSD Based JBOF

No CPU&DRAM



Introduction: SK hynix's NVMe-oF SSD Prototype

NVMe-oF SSD Prototype



Aupera
Making Video Alive

SK hynix

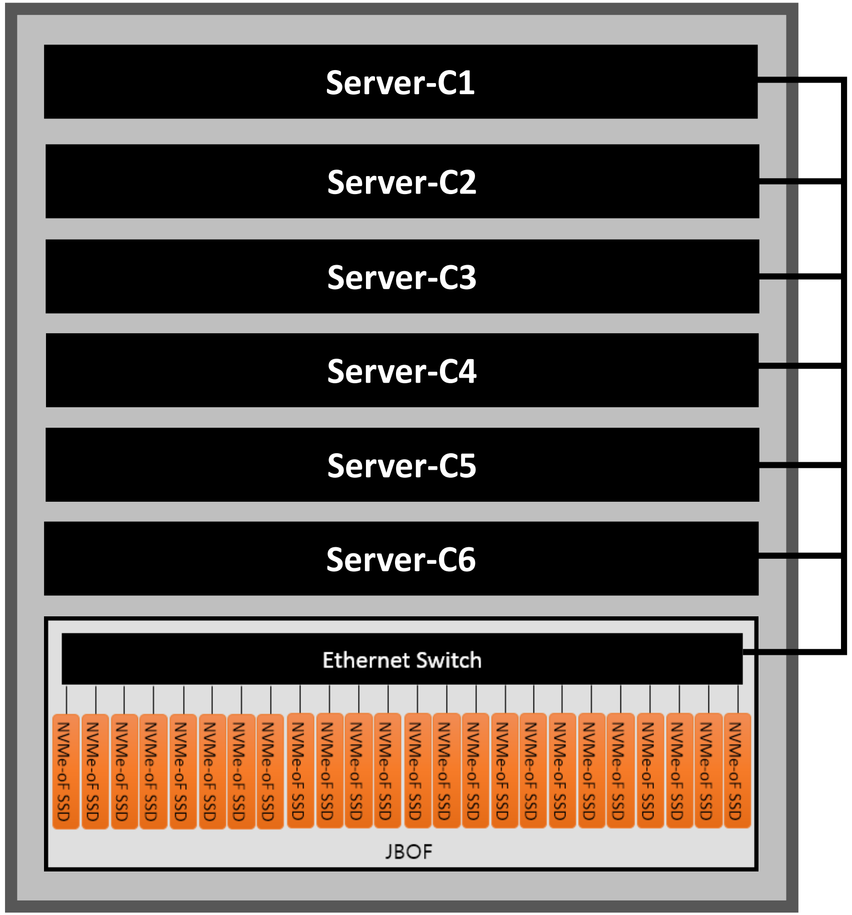
Item	PE6011
Interface	PCIe Gen3 x 4 / NVMe 1.3
NAND	SK hynix 3D V4 TLC
Form Factor	U.2 7 mm
Capacity	3840 GB

MARVELL

Item	88SN2400
Network	Dual 25GE RDMA
RDMA Protocol	RoCEv2
PCIe Interface	PCIe Gen3

Performance Evaluation: Configuration & Environment

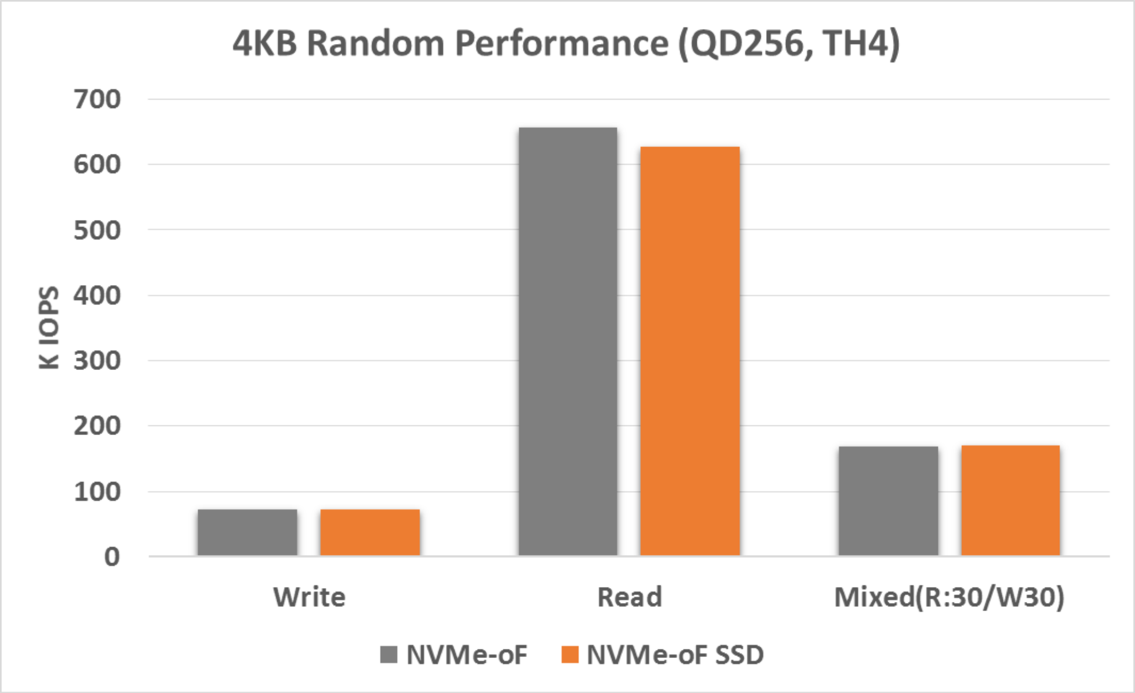
- 24 x NVMe-oF SSD, total capacity 92TB
- 6 storage servers

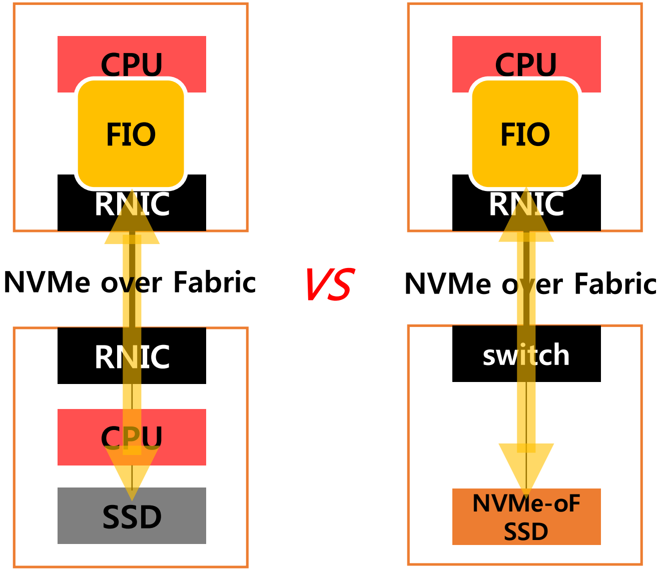


Items		Description
Servers	Hardware	- Xeon Gold 6136 CPU @3.00GHz (2 Sockets – 24 Threads per socket) - 192 GB Memory
	Software	Ubuntu 18.04.2 LTS (GNU/Linux 4.15.0-47-generic x86_64)
RDMA Network Card		Mellanox ConnectX-5 (MCX516A-CCAT)
Benchmark		FIO 3.13

Performance Evaluation: I/O Performance Result

- Single device performance is almost identical between conventional NVMe-oF and NVMe-oF SSD



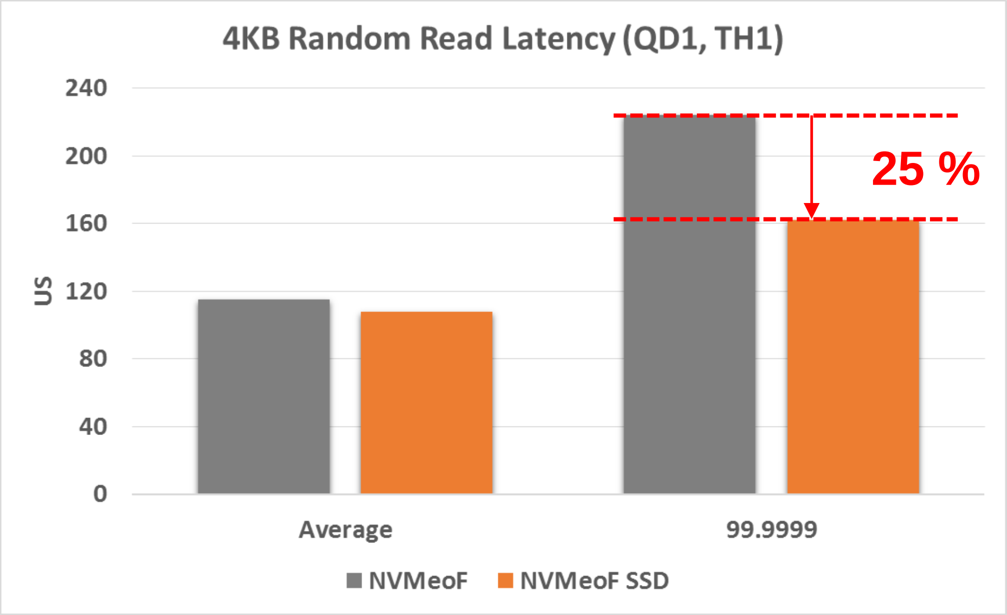


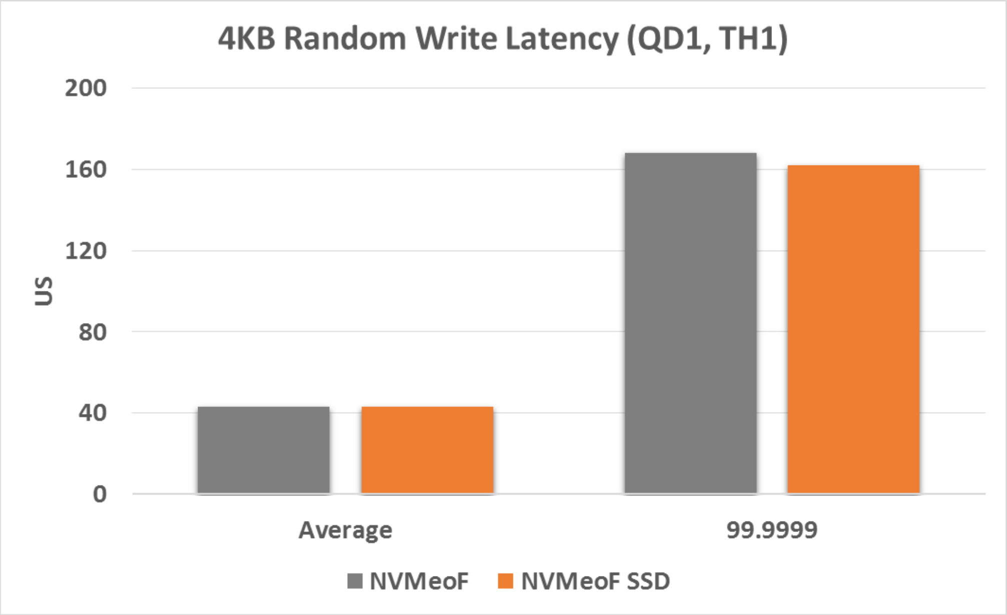
< Conventional NVMe-oF >

< NVMe-oF SSD>

Performance Evaluation: Quality of Service

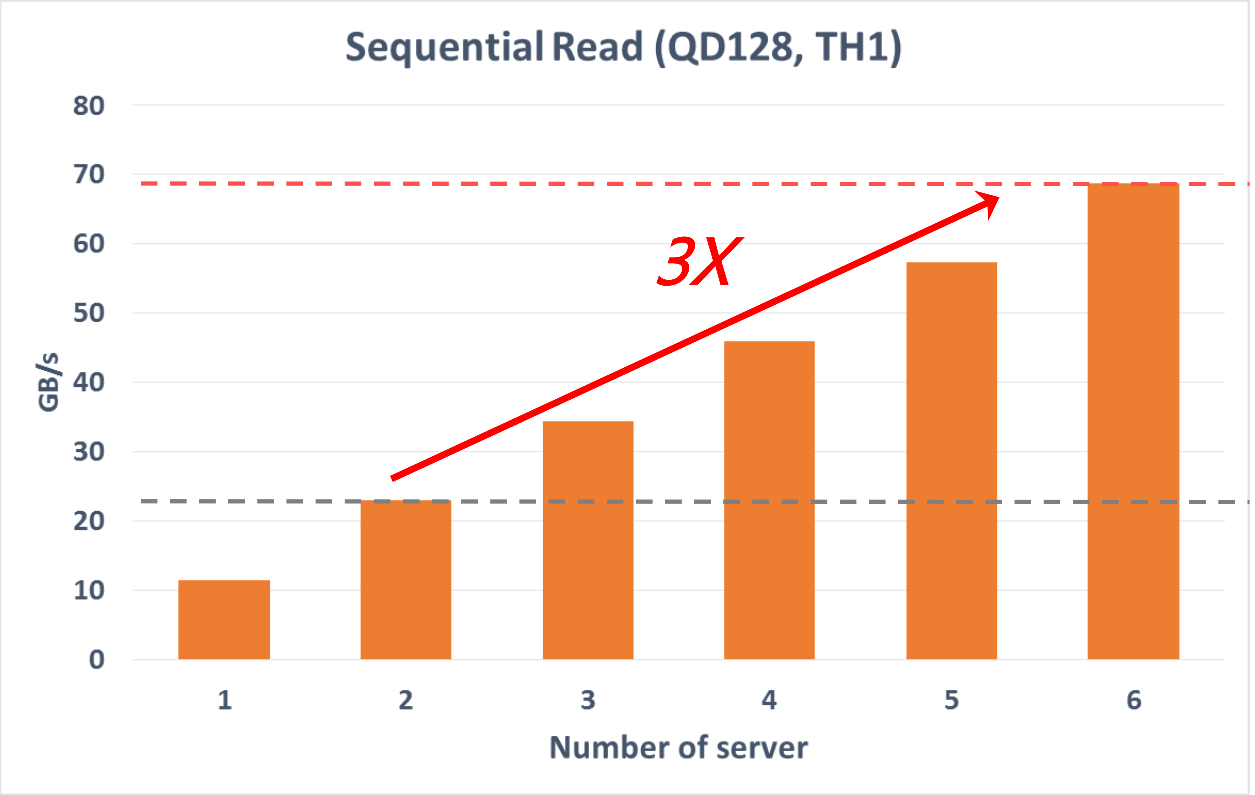
- **Enhanced Read QoS in NVMe-oF SSD**
  - **Offloading NVMe-oF functions to SSD reduces the latency**





Performance Evaluation: NVMe-oF SSD JBOF

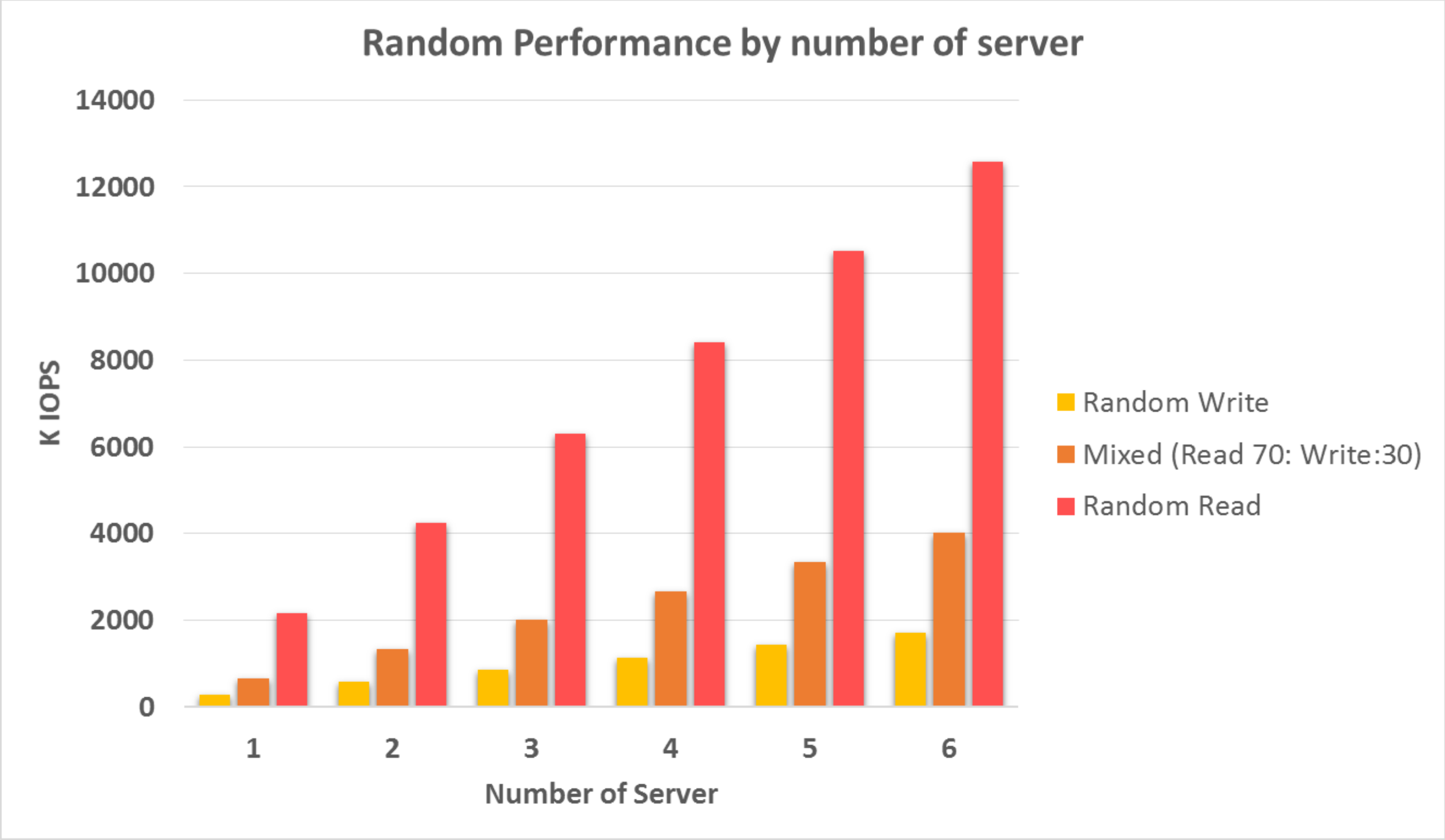
- **Providing 3x higher scalable performance**
  - **Single NVMe-oF SSD JBOF can scale performance for up to 6 servers**
  - **Single conventional NVMe-oF JBOF cannot scale performance beyond 2 servers**
  - **Max. performance can increase by 3x using single NVMe-oF SSD JBOF**



Maximum B/W of Conventional NVMe-oF JBOF

Performance Evaluation: NVMe-oF SSD JBOF

- **High performance scalability in random I/O**
  - **Performance increased in proportion to the number of server**

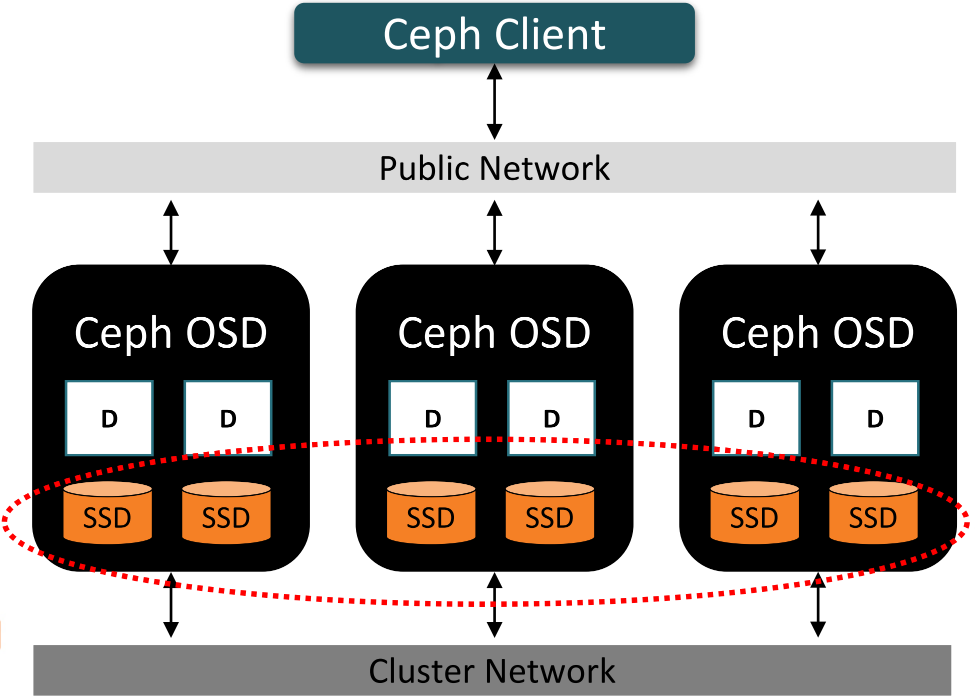


Ceph with NVMe-oF SSD

- More flexible scale-out storage system with NVMe-oF
  - Disaggregating storage from Ceph server using NVMe-oF JBOF

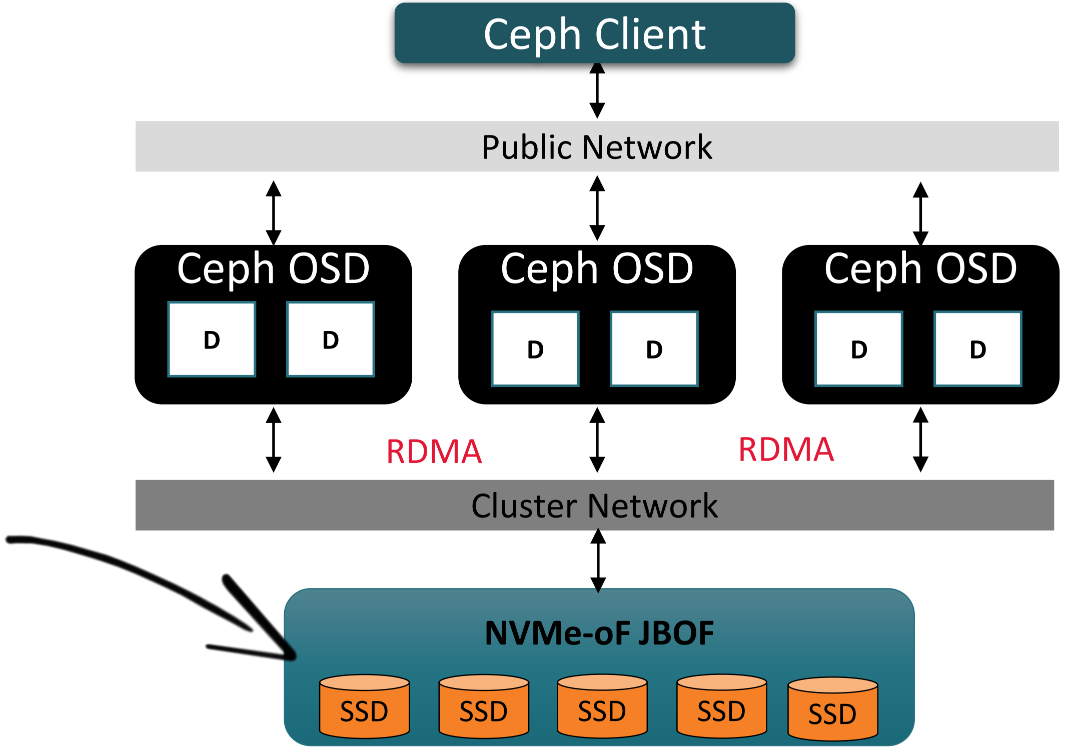
Conventional Ceph Storage Cluster

- Configured with local NVMe SSD



NVMe-oF JBOF as Backend of Ceph OSD

- Configured with remote NVMe SSD via NVMe-oF

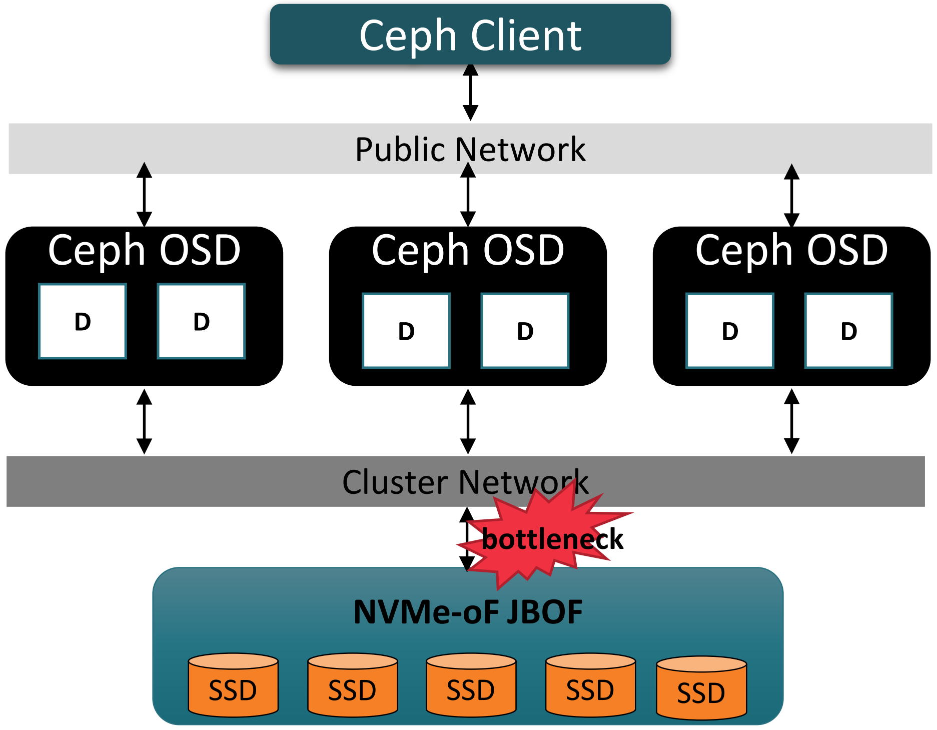


Ceph with NVMe-oF SSD

- **NVMe-oF SSD JBOF allows cost effectiveness & high scalability**

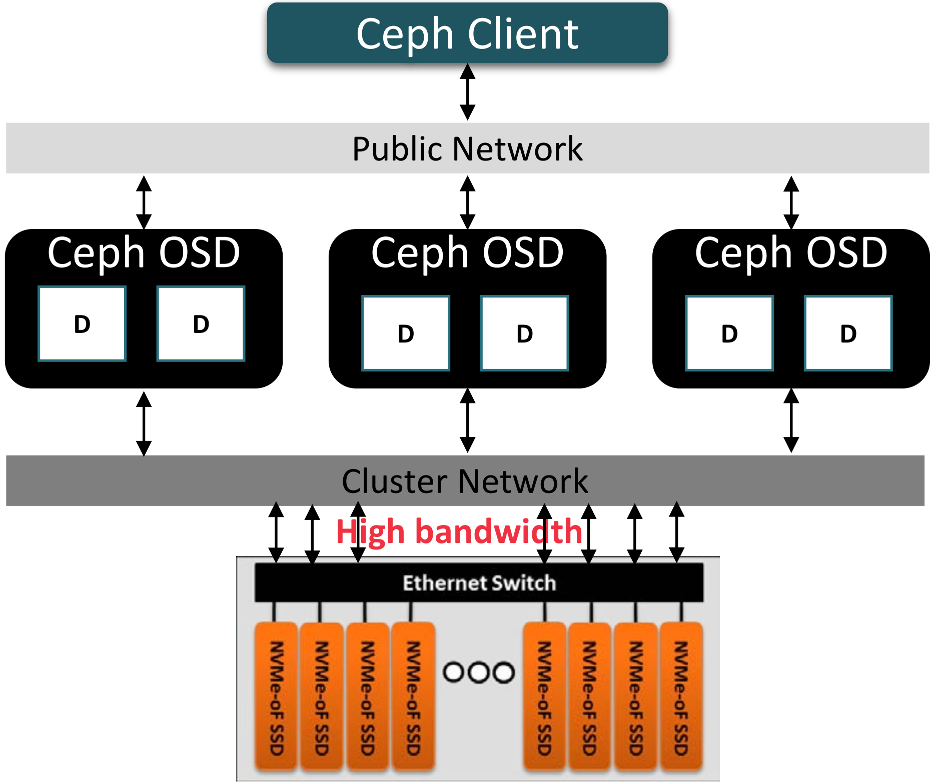
NVMe-oF JBOF

- **Limited bandwidth between Ceph OSD and NVMe-oF JBOF**



NVMe-oF SSD JBOF

- **Eliminate network bandwidth bottleneck**
- **Low cost & high scalable performance**



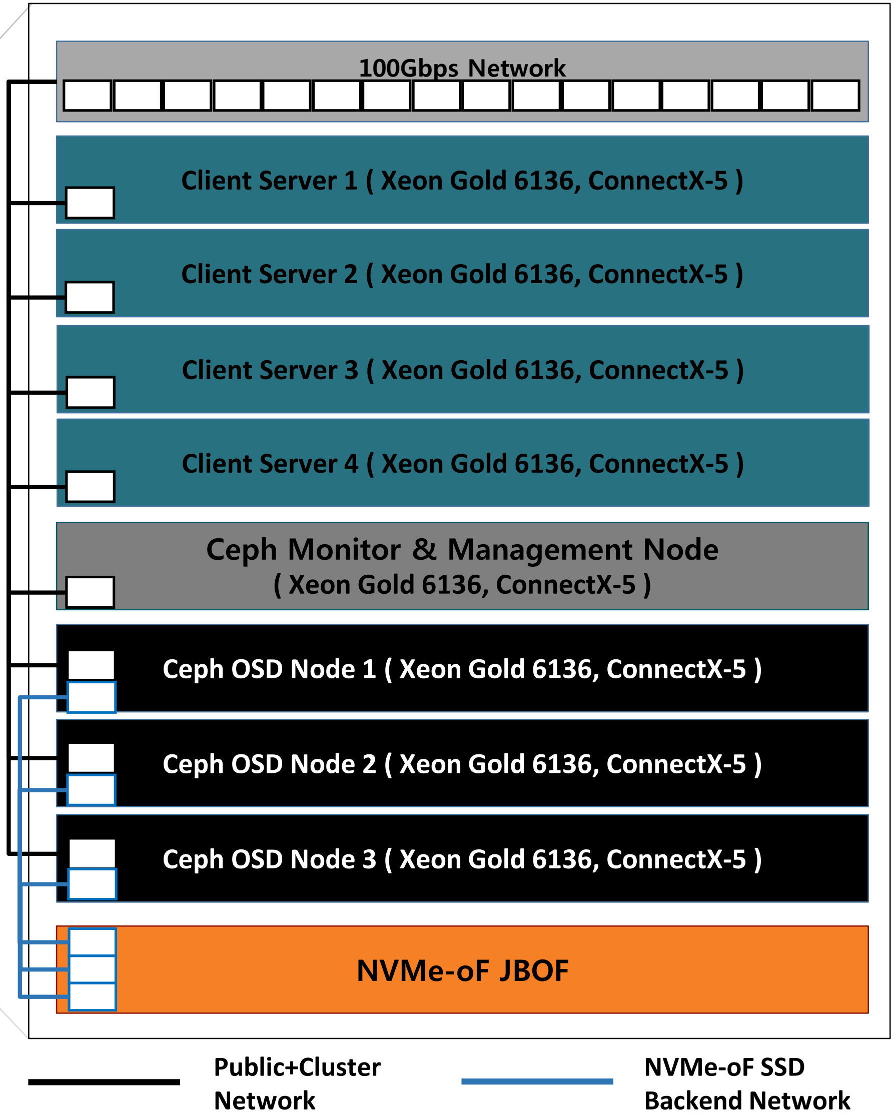
Ceph with NVMe-oF SSD: PoC Cluster Configuration

System Configurations

- 4 x Client nodes
- 3 x OSD nodes
- 1 x NVMe-oF JBOF
- 100Gpbs Network Switch

Ceph Configurations

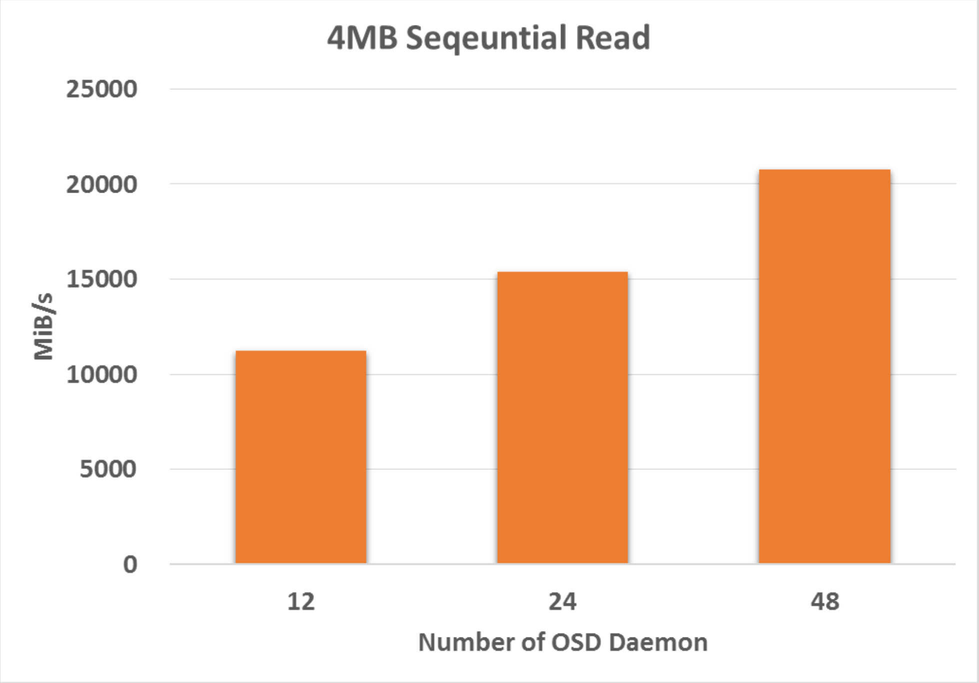
- 1TB RBD Volume on Client nodes
- Replica : 3
- 4 OSDs per NVMe-oF SSD
- PG : 2048


100Gbps Network
Client Server 1 (Xeon Gold 6136, ConnectX-5)
Client Server 2 (Xeon Gold 6136, ConnectX-5)
Client Server 3 (Xeon Gold 6136, ConnectX-5)
Client Server 4 (Xeon Gold 6136, ConnectX-5)
Ceph Monitor & Management Node
(Xeon Gold 6136, ConnectX-5)
Ceph OSD Node 1 (Xeon Gold 6136, ConnectX-5)
Ceph OSD Node 2 (Xeon Gold 6136, ConnectX-5)
Ceph OSD Node 3 (Xeon Gold 6136, ConnectX-5)
NVMe-oF JBOF
Public+Cluster Network
NVMe-oF SSD Backend Network

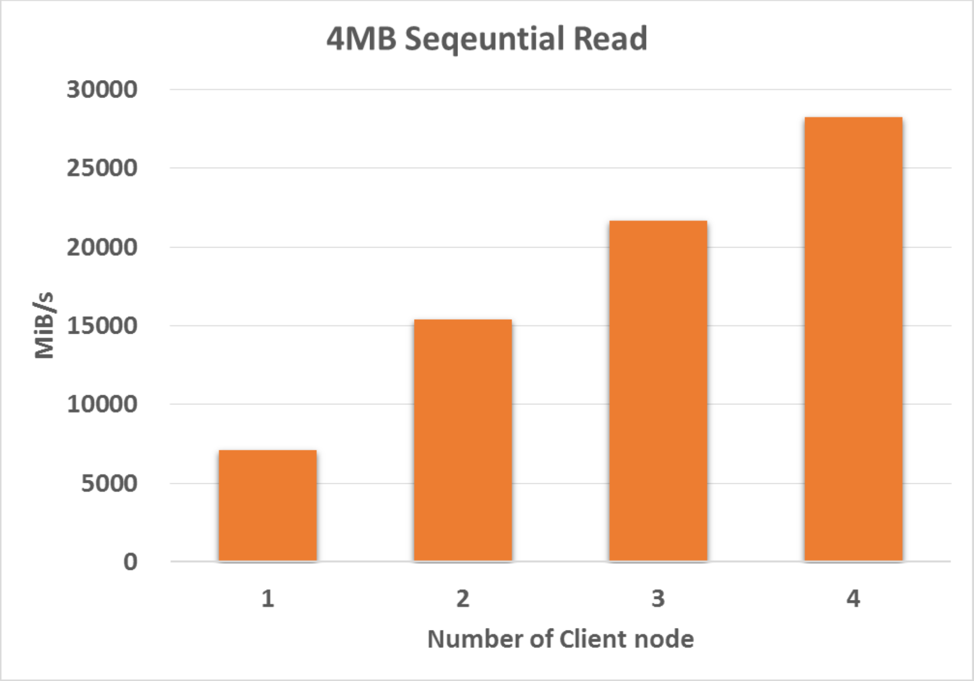
Ceph with NVMe-oF SSD: Performance Evaluation

- **Providing scalable performance for ODSs & Client Servers**
  - **Performance scale by increasing #s of OSD, client servers**

Performance by Number of OSD Daemon



Performance by Number of Client Node



Summary

1. NVMe over Fabric SSD

- Offloading host NVMe over Fabric functions into SSD
- Competitive single device performance in storage system
- High performance scalability validated

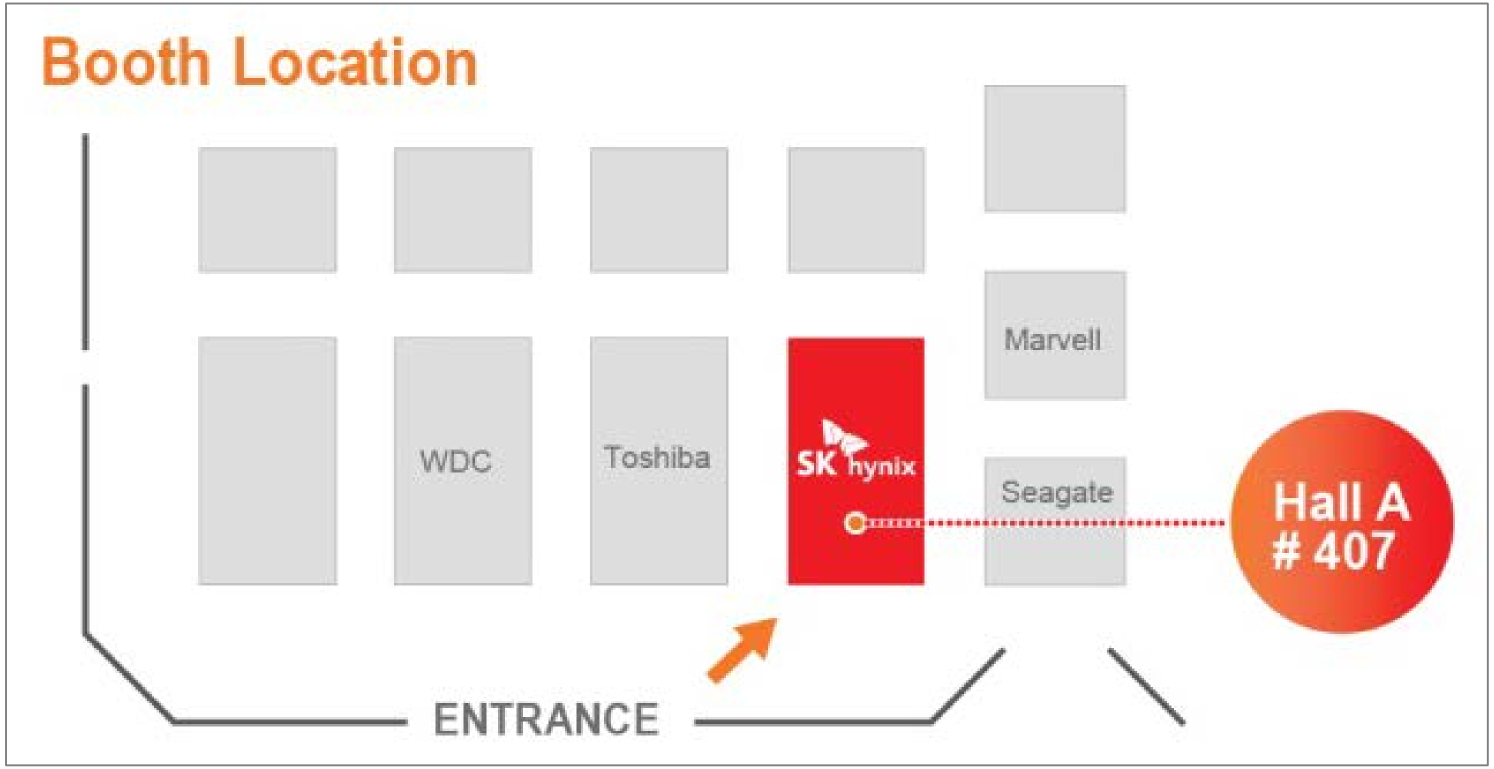
2. Ceph with NVMe-oF SSD

- Cost effective & high scalable solution for Ceph cluster
- Providing scalable performance for Ceph ODSs & Ceph Client Servers

3. Next Works

- Continue to evaluate Ceph storage cluster with NVMe-oF SSD
- Performance optimization for NVMe-oF SSD based Ceph storage

Learn more about SK hynix



Visit SK hynix
@ booth #407

Experience SK hynix products and demos & get a free giveaway!

Thank you

Growing together

for better tomorrow