STORAGE DEVELOPER CONFERENCE
SDC 21
BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

A SNIA Event

Challenges and Effects of EDSFF-based NVMe-oF Storage Solution

Duckho Bae, Jungsoo Kim
{duckho.bae, jungsoo0.kim}@Samsung.com
Samsung Electronics

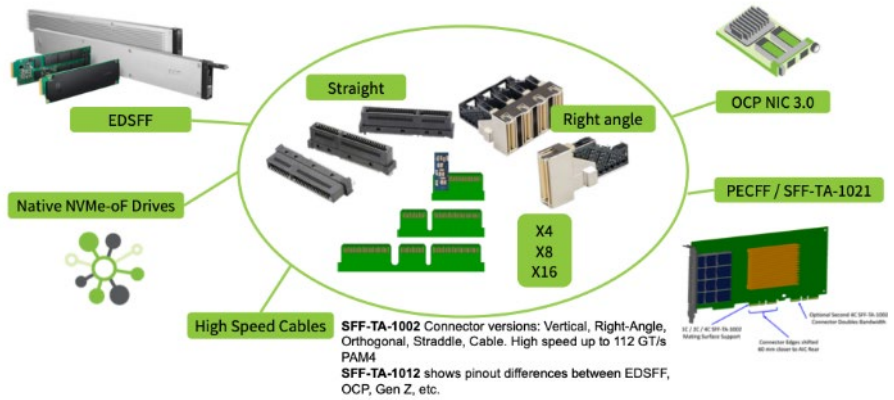
Agenda

- EDSFF
- EDSFF Application
- E1.S Reference Server
- EDSFF in NVMe-oF Solution

EDSFF Form Factor

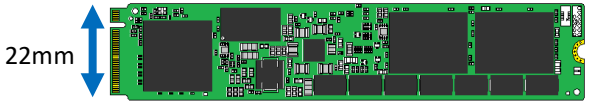
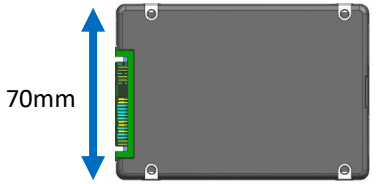
What is EDSFF

- Enterprise and Data Center Standard Form Factor
- Designed to overcome conventional device limitations
  - Improve thermals, power, and scalability
  - High-speed interface
  - Hot-plug support
  - built in LEDs, carrier-less design
- Allow to support new types of devices
- Customizable latch & extension kit design



Source: https://www.snia.org/forums/cmsi/knowledge/formfactors

Conventional SSD Form Factor

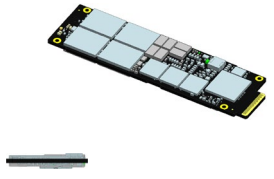
	M.2	U.2
		
Characteristic	• The initial spec. proposal was for client use • Used as a boot drive in server systems • Compact size and high flexibility	• Evolved from 2.5” HDD Form factor • Most general Form Factor • Two different thicknesses: 7mmT, 15mmT • Mostly used for storage device
Limitation	• Limited Performance and Scalability • Front-loading and hot-plug not supported • Vulnerable to warpage because of thin PCB • Low input power (3.3V)	• Power limitation (typical 25W) • Disadvantageous for high speed interface like PCIe Gen4/5

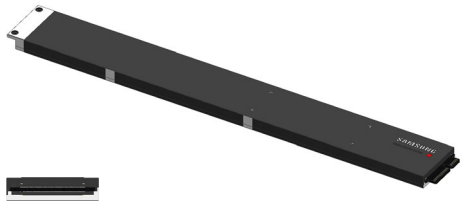
EDSFF E1.X

- 1RU optimized, Offers various thickness

System Density → **Power Capacity**

Capacity ↓

E1.S up to 8 NAND Landing					
Size	31.5 x 111.49 x **5.9mm**	31.5 x 111.49 x **8.01mm**	33.75 x 118.75 x **9.5mm**	33.75 x 118.75 x **15mm**	33.75 x 118.75 x **25mm**
Recommended Power(W)	12W	16W	20W	20W	25W

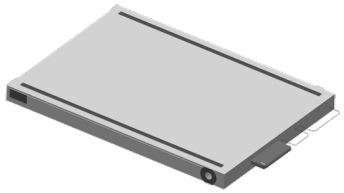
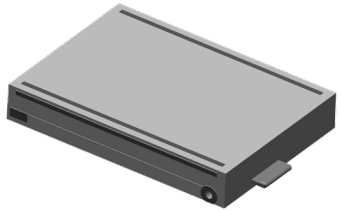
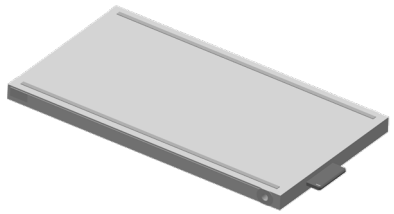
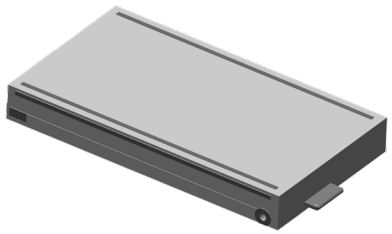
E1.L up to 16 NAND Landing		
Size	38.4 x 318.75 **x 9.5mm**	38.4 x 318.75 x **18mm**
Recommended Power(W)	25W	40W

EDSFF E3.X

- 2RU Optimized, Applicable to various applications

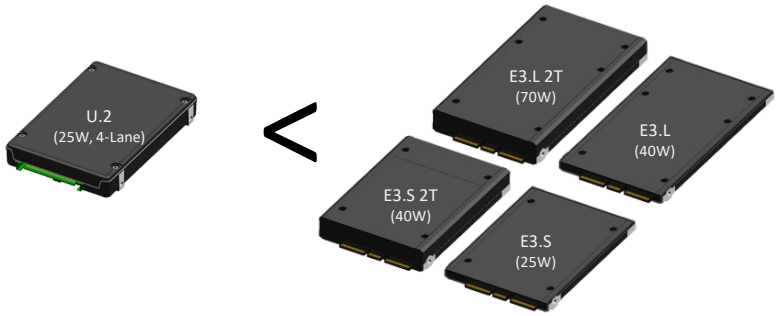
System Density → Power / Capacity

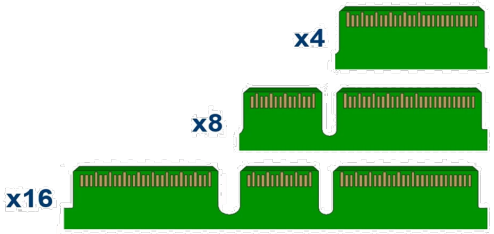
Density ↓ Power / Capacity

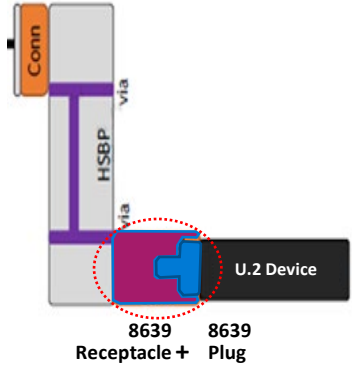
E3.S up to 8 NAND Landing		
Size	76 x 112.75 x **7.5mm**	76 x 112.75 x **16.8mm**
Recommended Power(W)	25W	40W
E3.L up to 16 NAND Landing		
Size	76 x 142.2 **x 7.5mm**	76 x 142.2 x **16.8mm**
Recommended Power(W)	40W	75W

EDSFF – Benefits

- Various power options
  - 25W, 40W, 70W
- Various PCIe interfaces (x4 ~ x16 lanes)
  - x4 ~ x16 lanes
- Better Signal Integrity (SI)
  - Advantageous for high-speed interface (< PCIe Gen5)

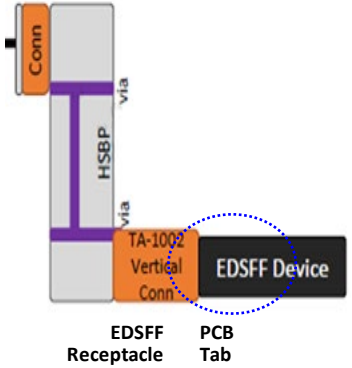






U.2

VS.



EDSFF

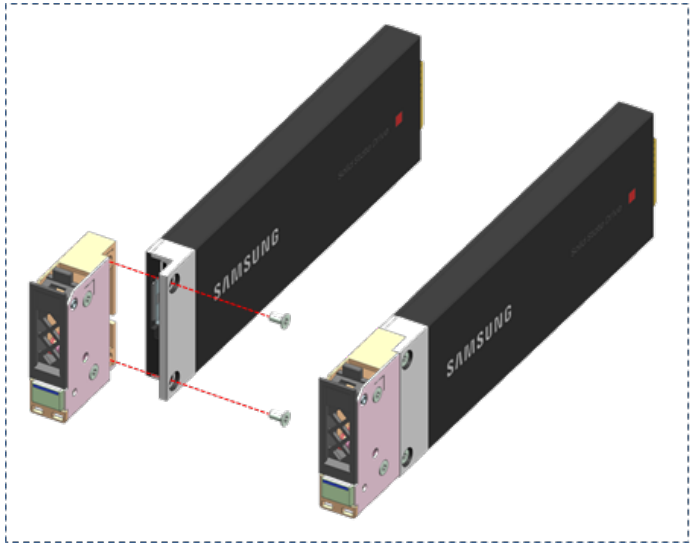
EDSFF Application

E1.S SSD Tool-less Design

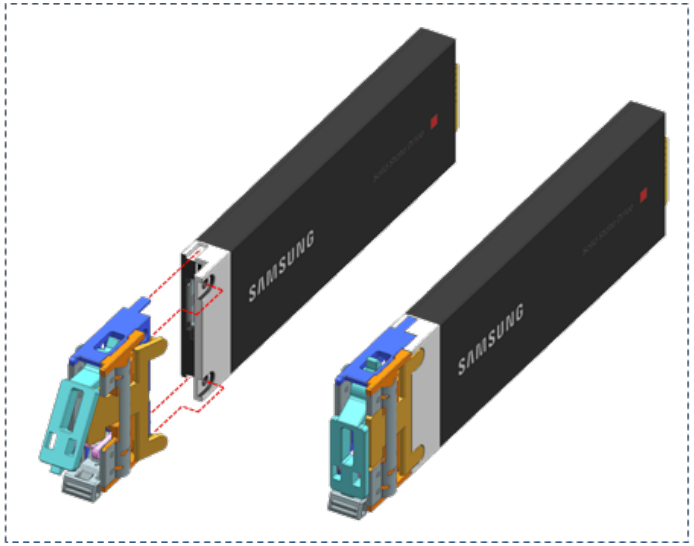
DMC (Device Mgmt. Control)

E1.S Tool-less Design

- DC customers want to improve serviceability in their datacenter by removing the screws
- E1.S + extension kit with screws are the only option in the market, and we developed the innovate new tool-less ext. kit design to satisfy the requirements



Current Design – Screw type



Tool-less Design – Clip type

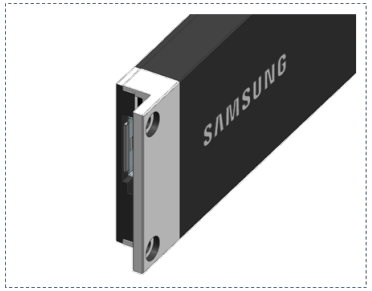
Compatible w/ a screw type extension kit!

E1.S Tool-less Design

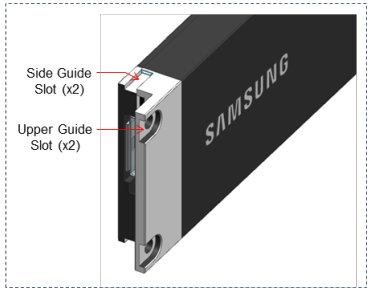
Screw Type **Tool-less Type**

# of drives	Time saved		Cost saving
	per drive	Total	
10M device	36 sec	100,000 h	$2.5M

* Average salary of data center technician is assumed to be $25/h



Standard

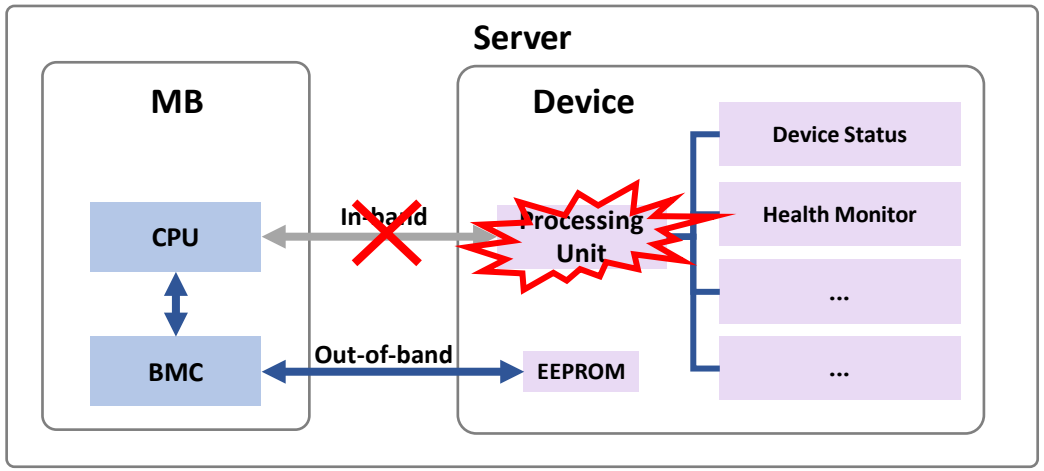


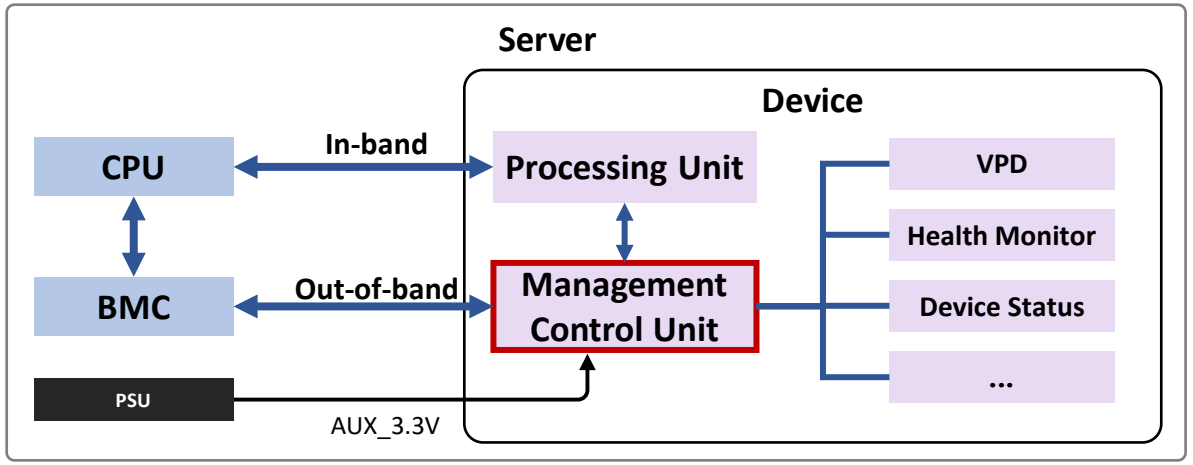
Modified

Want to discuss design upgrades and standardization with the industry!

DMC (Device Mgmt. Control) – E3.S basis

- BMC can access only limited dataset through EEPROM when a device processing unit fails
- Replace the EEPROM with a microcontroller, MCU (Mgmt. Control Unit), and collects most of the device H/W information instead of processing unit
- BMC talks to MCU to collect device status data through OOB (I2C)





EDSFF Reference Server

Poseidon Server

EDSFF Reference System

- E1 and E3 based system can increase the performance and density
- Have more flexibilities than traditional system

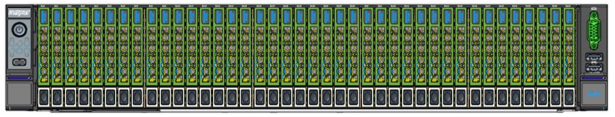
U.2 SSD x 10ea

vs.

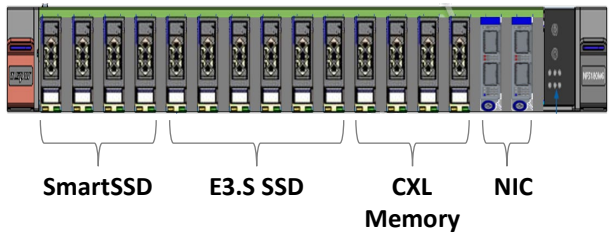
E1.S SSD x 32ea

U.2 SSD x 24ea

vs.



E3.S SSD x 40ea



Poseidon Project

- Open-source HW & SW project for NVMe-oF based shared network storage system
- OCP based industrial collaboration b/w “Component Vendor ↔ ODM ↔ Data Center”

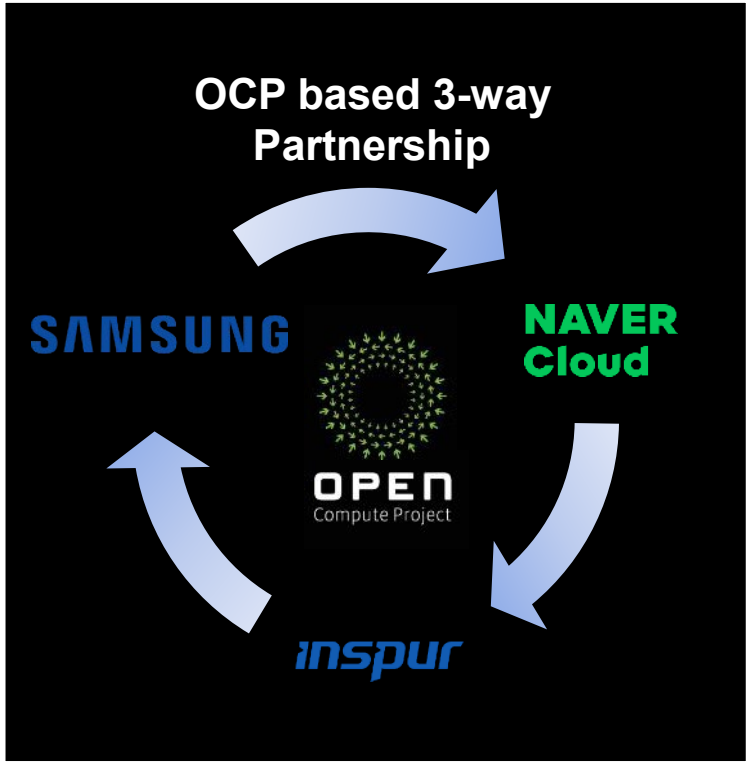
Poseidon SW

Open-source Storage OS for NVMe-oF

+

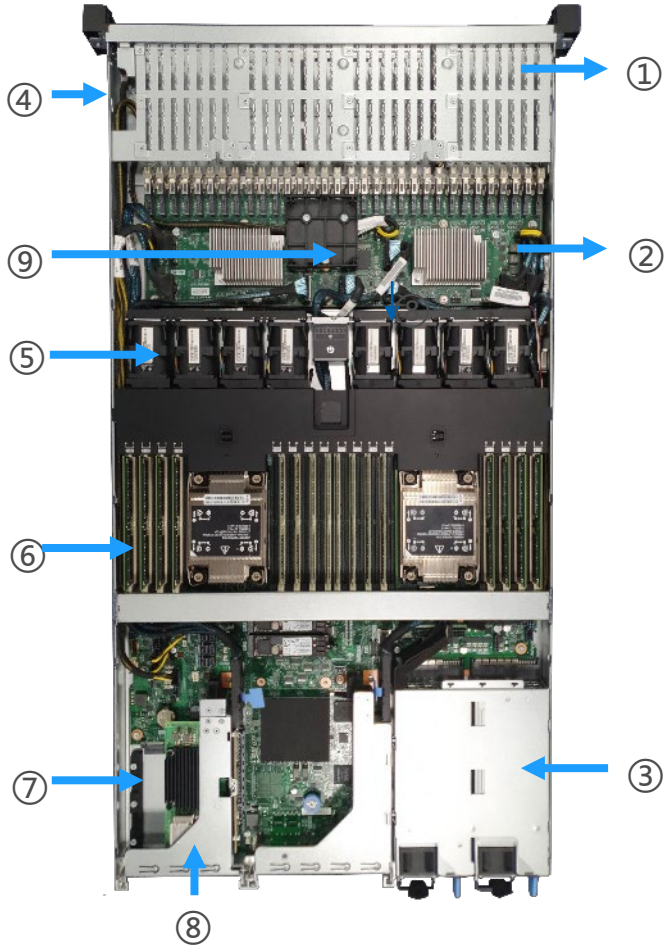
Poseidon HW

PCIe Gen4 E1.S SSD Ref. System



System Design Overview

Front

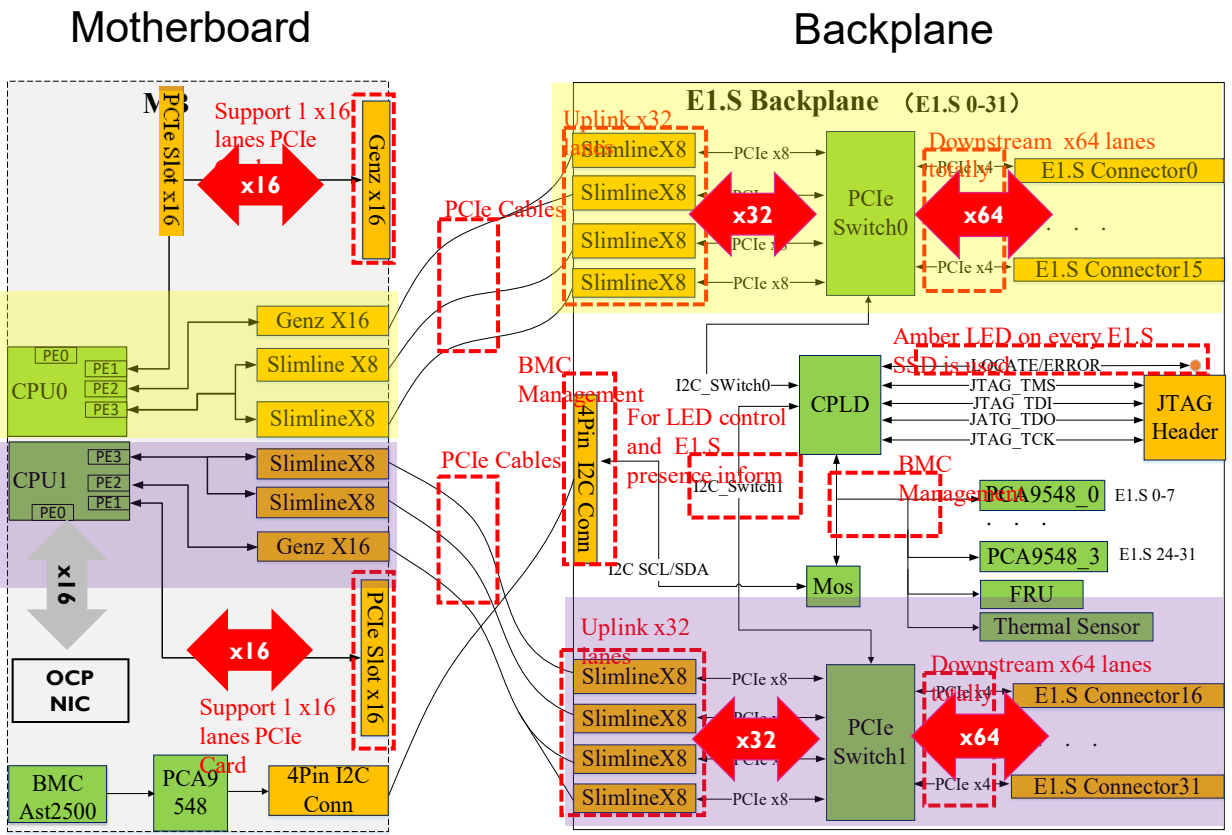


① : E1.S SSD (5.9/8.01/9.5)	32ea
② : 32 E1.S BP	1
③ : PSU	2ea
④ : IO Module	1ea
⑤ : FAN	8ea
⑥ : MB	1ea
⑦ : FHHL Card	2ea
⑧ : OCP NIC V3	1ea
⑨ : NVDIMM Power Module	2ea

Front View

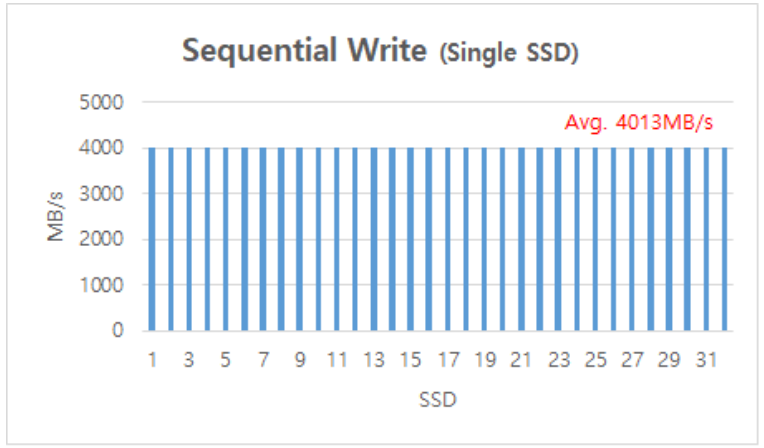
Rear View

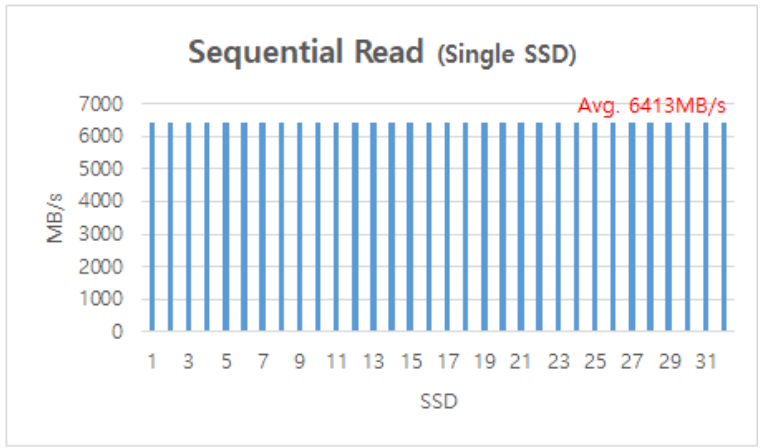
System Diagram

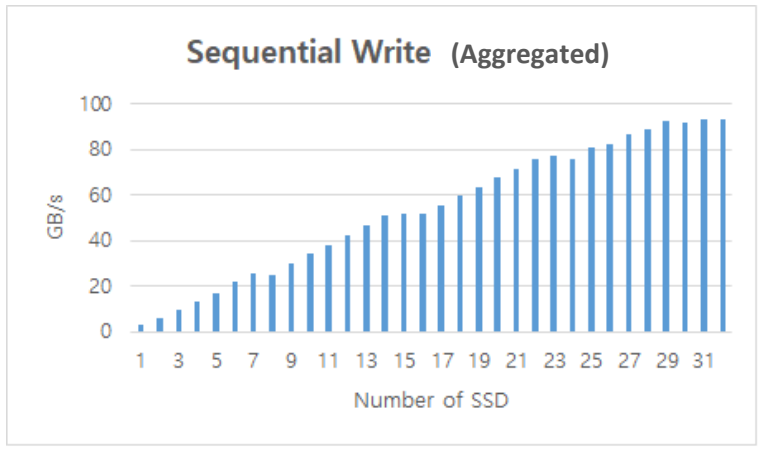


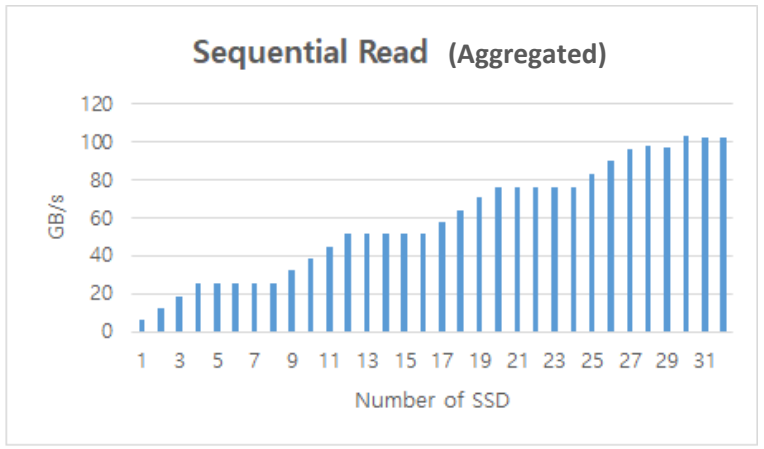
- Symmetrical PCIe topology to minimize socket to socket traffic (UPI)
- Each CPU provides 64 PCIe Gen4 lanes
- Each PCIe Switch provides PCIe Gen4.0 100 lanes (32W typical power)
- NIC (100GbE x 2 Ports) bandwidth limits the total IO bandwidth
- E1.S Connector (Orthogonal type) support PCIe4.0, connected to 32 pcs E1.S

IO Performance – E1.S Poseidon









*Theoretical B/W limit: 128GB/s

- **Samsung PM9A3 Specification**

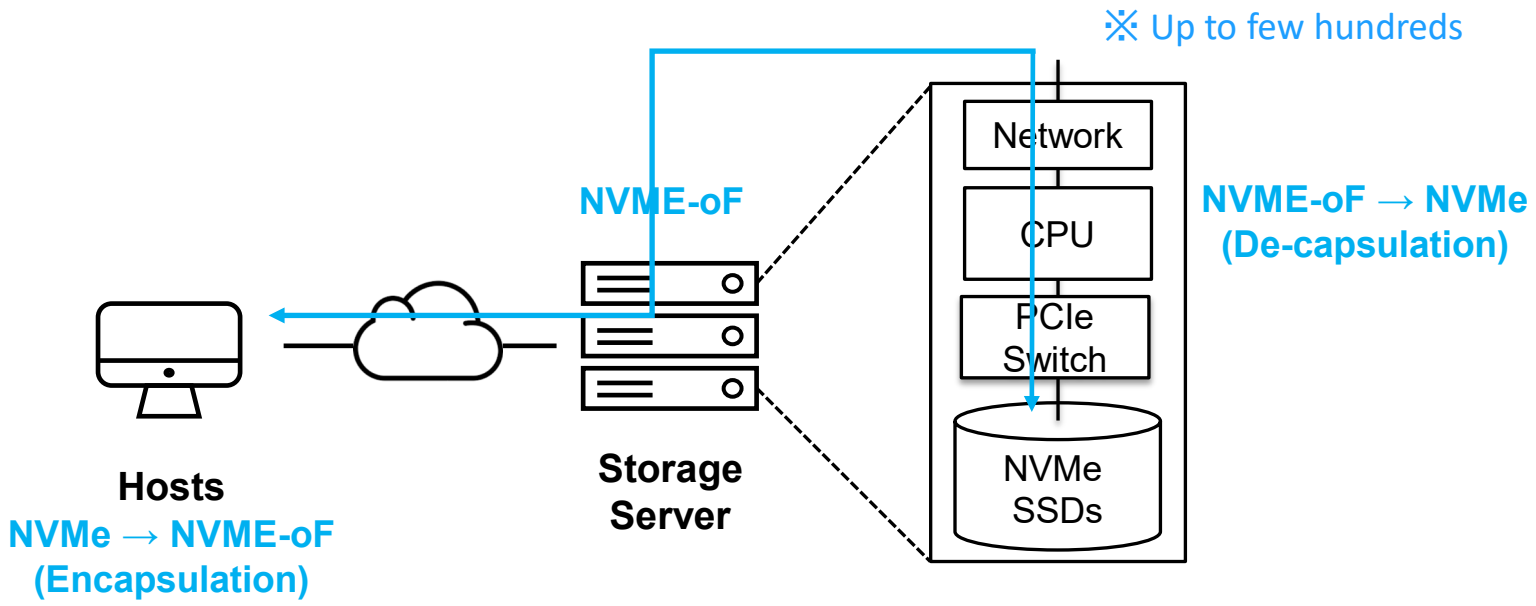
Form factor	E1.S
Capacity	960 GB, 1.92 TB, 3.82 TB, 7.68 TB
Sequential read	Up to 6,500 MB/s
Sequential write	Up to 3,200 MB/s
Random read	Up to 900,000 IOPS
Random write	Up to 150,000 IOPS
Physical Dimensions	31.5 x 111.49 x 5.9 mm
Power consumption	Read: <= 9.7W, Write: <= 11.7W
Host interface	PCIe Gen 4 x4

EDSFF in NVMe-oF Solution

PoseidonOS

NVMe-oF Interface

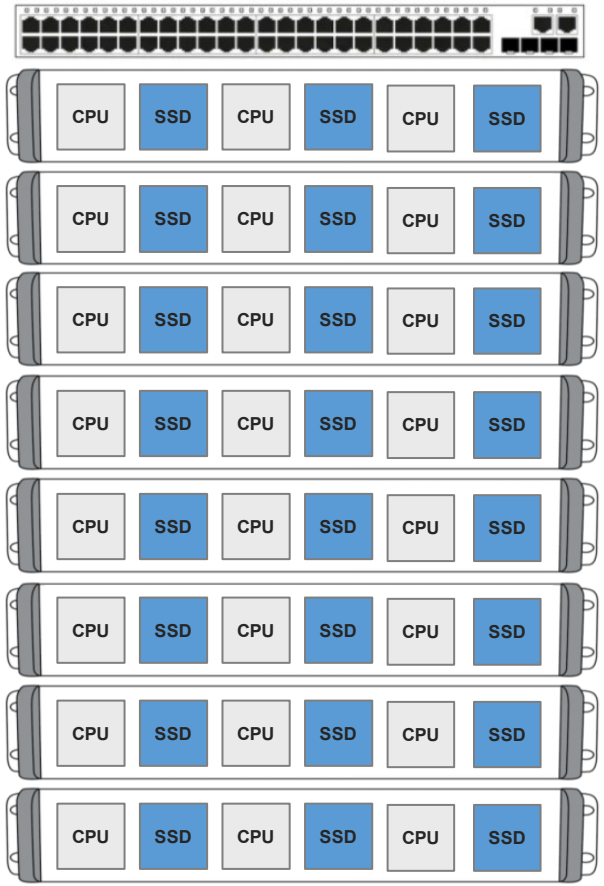
- **Can break through the scaling limitation of PCIe-attached NVMe**
- **Uses a transport protocol over a network to access remote NVMe**
  - End-to-End NVMe semantics across a range of topologies
  - Retains NVMe efficiency and performance over network fabrics



Disaggregated Architecture

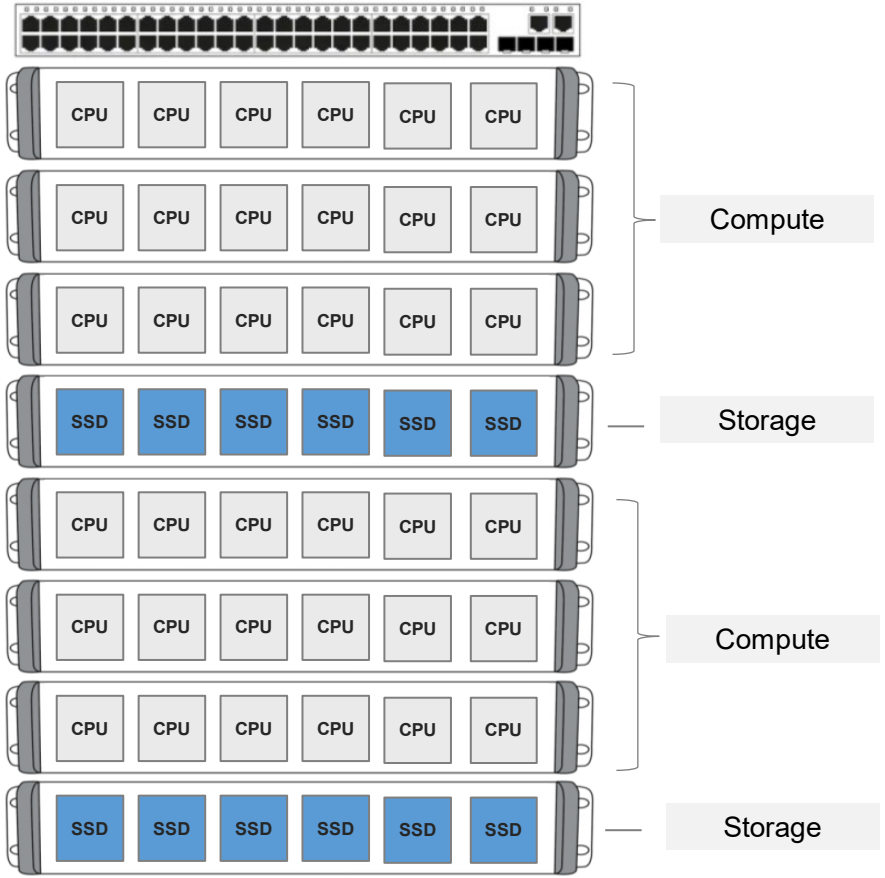
- **NVMe-oF interface brings independent scaling of storage resources**

Direct Attached Architecture



- **Increase CPU & SSD utilization**
- **Reduce storage spending & TCO**
- **Simplified scalability**
- **Higher performance**
- **Increase hardware flexibility**
- **Ideal capacity utilization**

Disaggregated Architecture

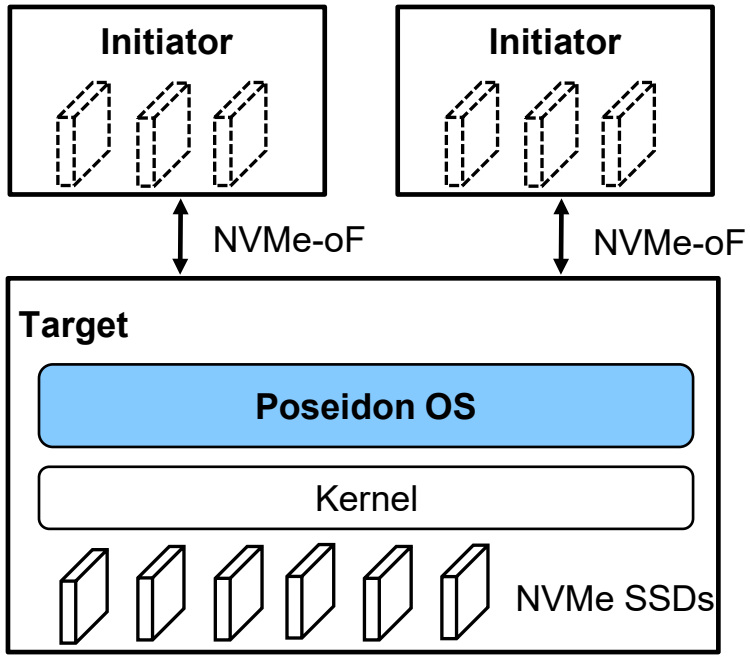


Vital Virtues for NVMe-oF Solution

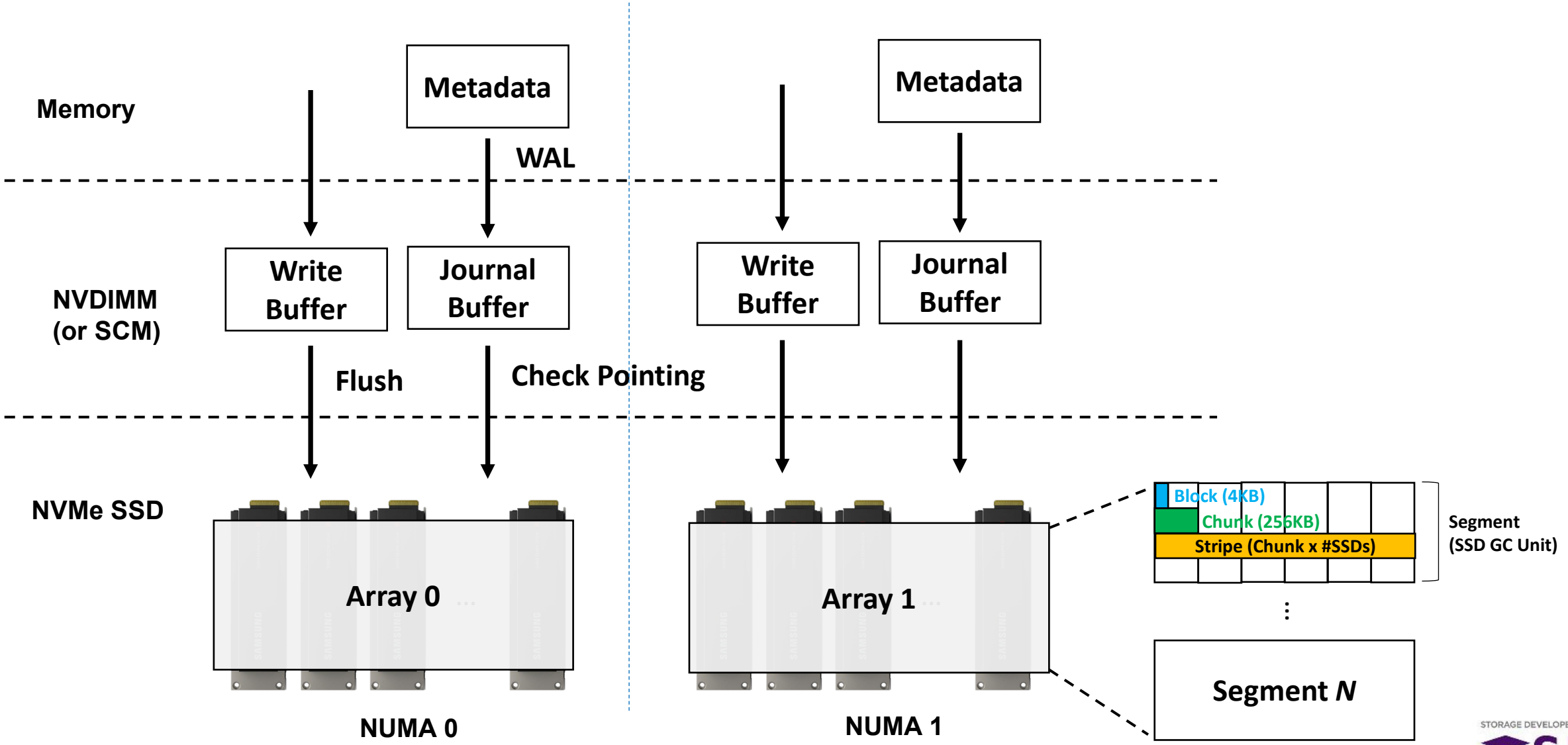
1. **Ultimate and stable performance for resource sharing**
2. **High availability for Peta-byte scale capacity**
3. **Efficient metadata management for storage features at Peta-byte scale**
4. **Optimized CPU, memory, network resources utilization**

PoseidonOS

- **User-space storage OS for NVMe-oF**
- **Provide PCIe Gen4 performance via network**
  - Up to 200GbE
- **Support valuable storage features**
  - NUMA-Aware, Volume Mgmt, Perf Throttling, SW RAID, …
- **Easily integrate with upper orchestration layers**
  - RESTful, CSI, …

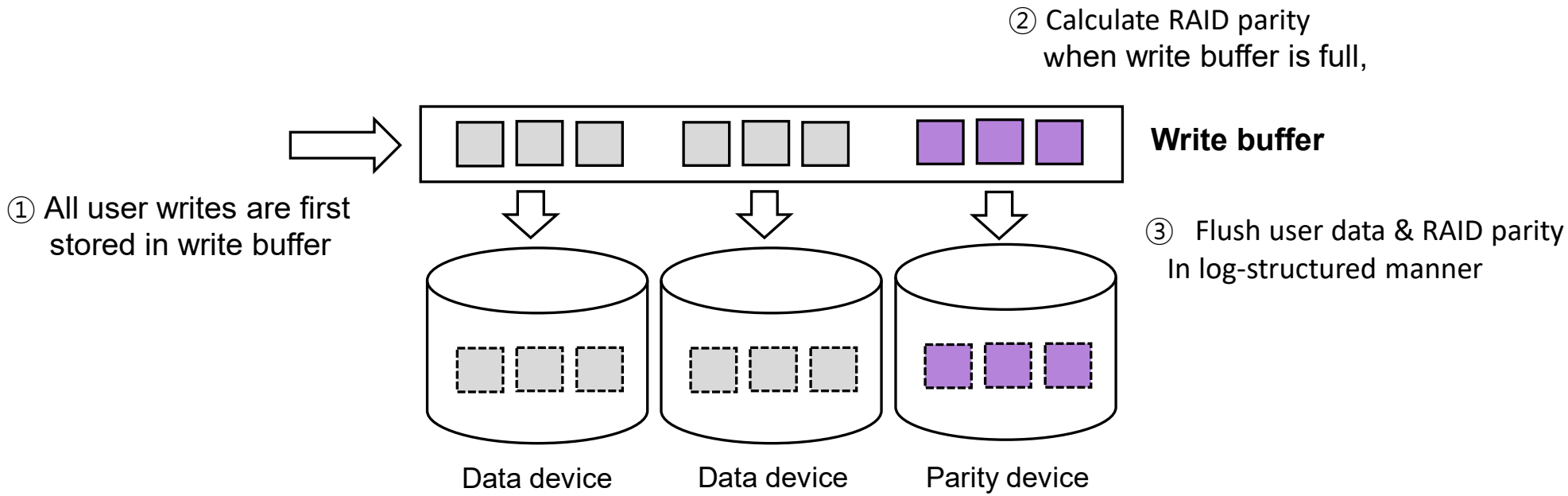


Storage Hierarchy in PoseidonOS



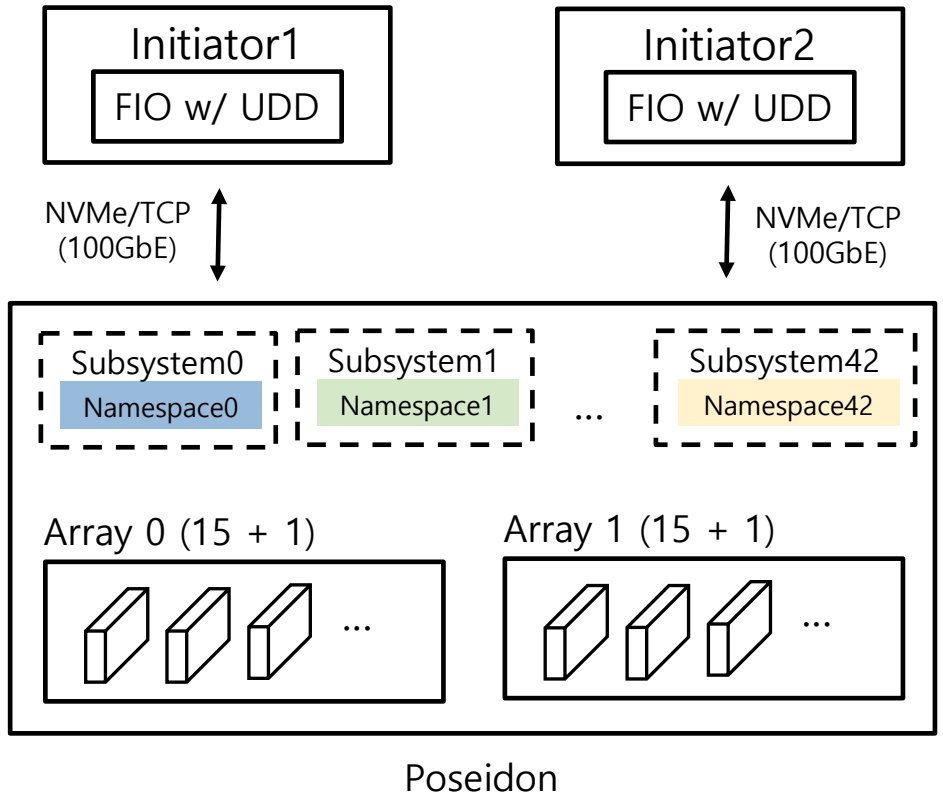
Log-structured RAID

- **Necessary to support software RAID/EC for NVMe drives**
- **Log-structured RAID approach follows naturally since user data is stored to SSDs in log-structured manner**
  - Can reduce WAF and QoS impact for user IO



Experiment Environments

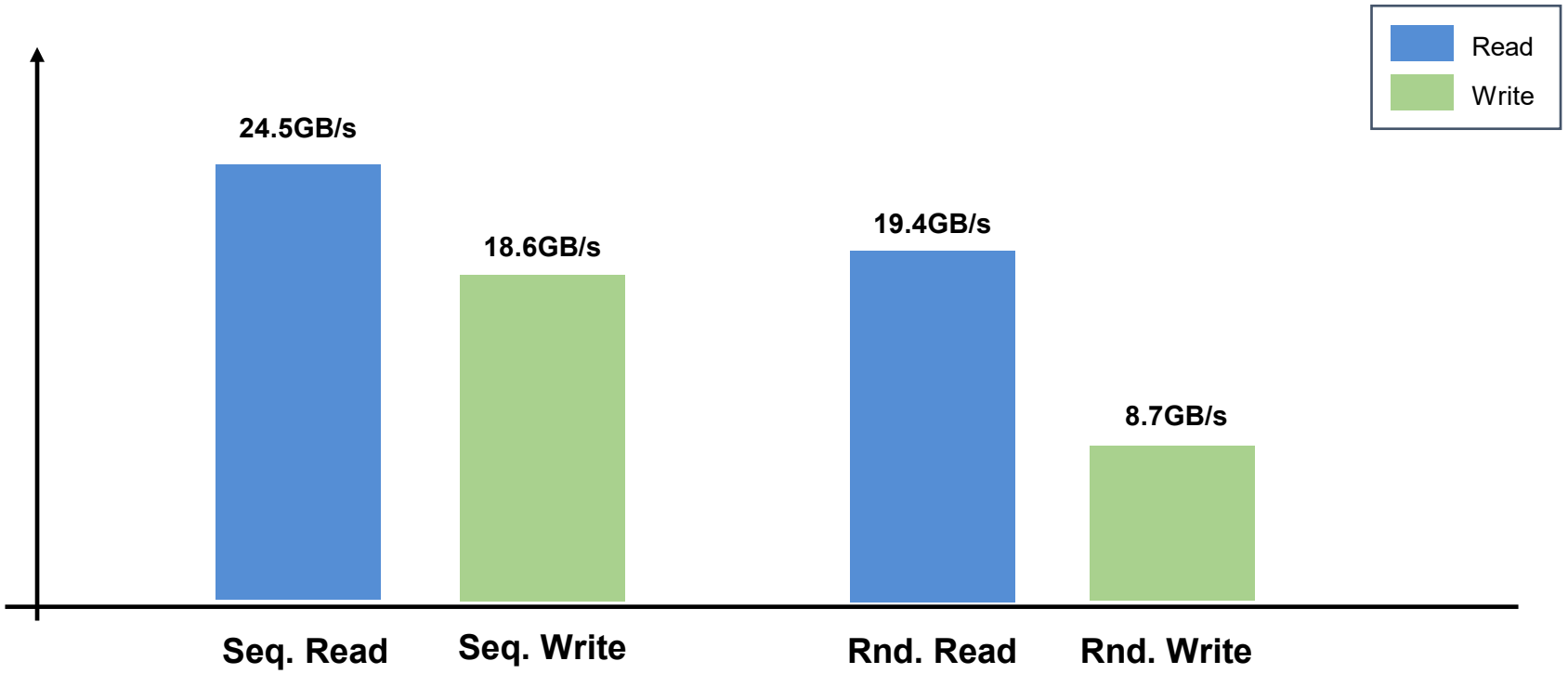
- **PCIe Gen4 SSD * 32**
- **200GbE Network Connection**
  - **NVMe/TCP**
- **2 Arrays / 43 Volumes**
- **RAID 5 (15 + 1)**
- **Using *u*Driver in initiator-side**



* Intel Xeon CPU (3Ghz, 48 Cores) * 2ea, DDR4-3200 32GB * 32ea, PM9A3 4TB * 32ea, MLNX CX-5 * 2
Ubuntu 5.3.0-24-generic, poseidonos-0.9.10

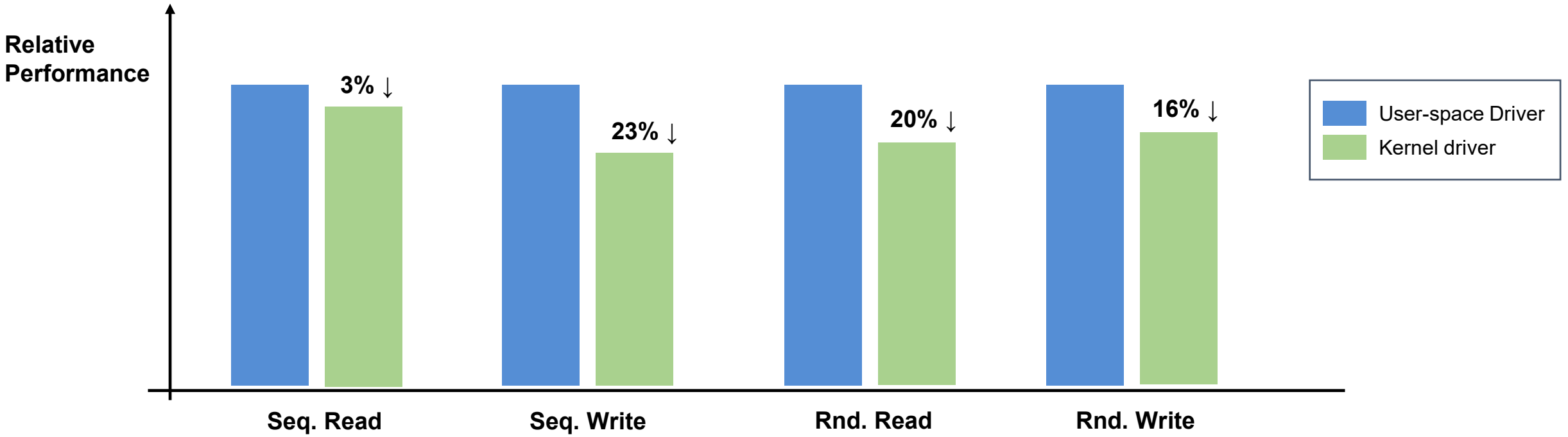
Performance Numbers

- **Achieved up to 200GbE Performance via NVMe/TCP**
- **Random Write has room for improvement**



Performance - Initiator SW Stack

- **User-space driver shows superior performance**
  - Except for Seq. read, kernel driver has up to 23% performance drop

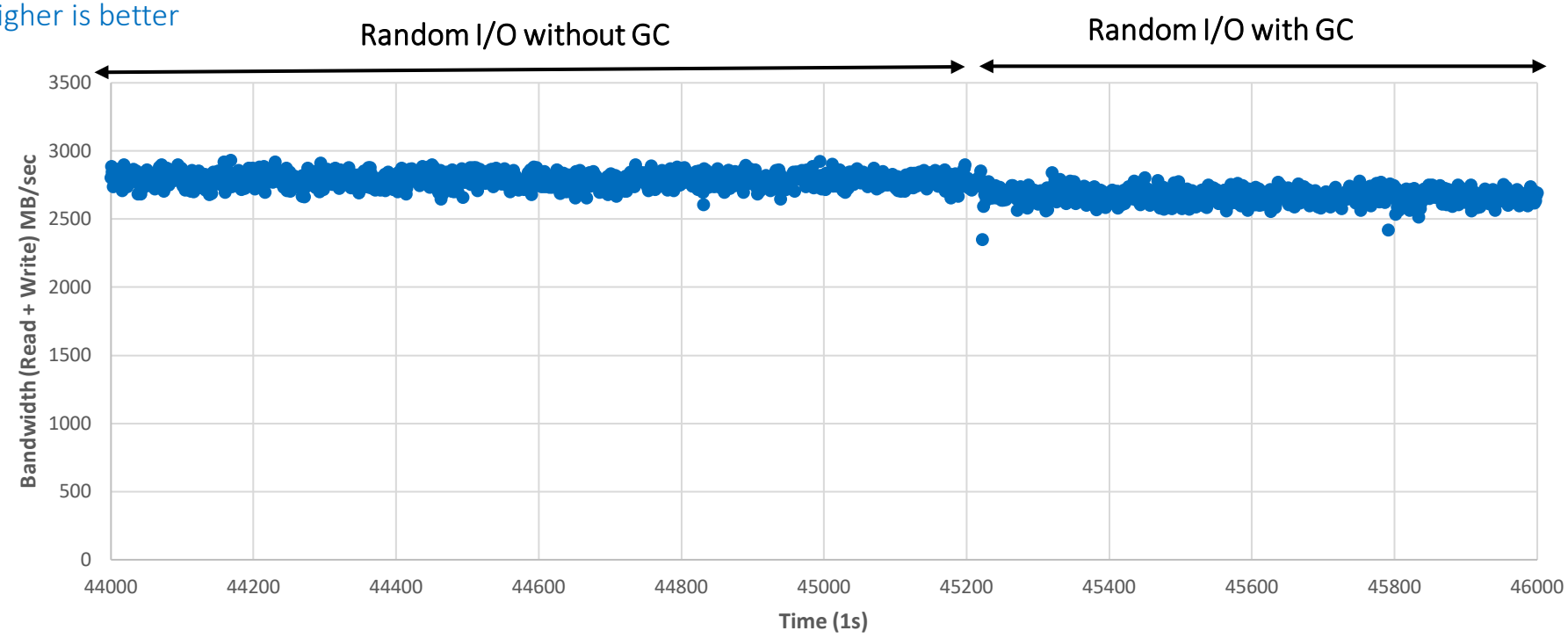


IO Consistency

- **Provide stable QoS in case of mixed IO (7:3)**
- **Internal IO drops IO consistency slightly**

IO Consistency = 0.949
- 99.9th IOPS / Average IOPS
- Higher is better

IO Consistency: 0.872
- 99.9th IOPS / Average IOPS
- Higher is better



Future Work

- Support innovative devices (ex. ZNS, QLC)
- Support more features
- Provide developers toolkits
- Enable PCIe Gen5 performance

- **Available at Github**
  - https://github.com/poseidonos/poseidonos

Thank you for Watching Our Presentation

Please take a moment to rate this session.

Your feedback is important to us.

