SAMSUNG

Poseidon

NVMe-oF Reference System for EDSFF SSDs

삼성전자 메모리 사업부
소프트웨어 개발팀 김희수
heesoo84.kim@samsung.com

Agenda

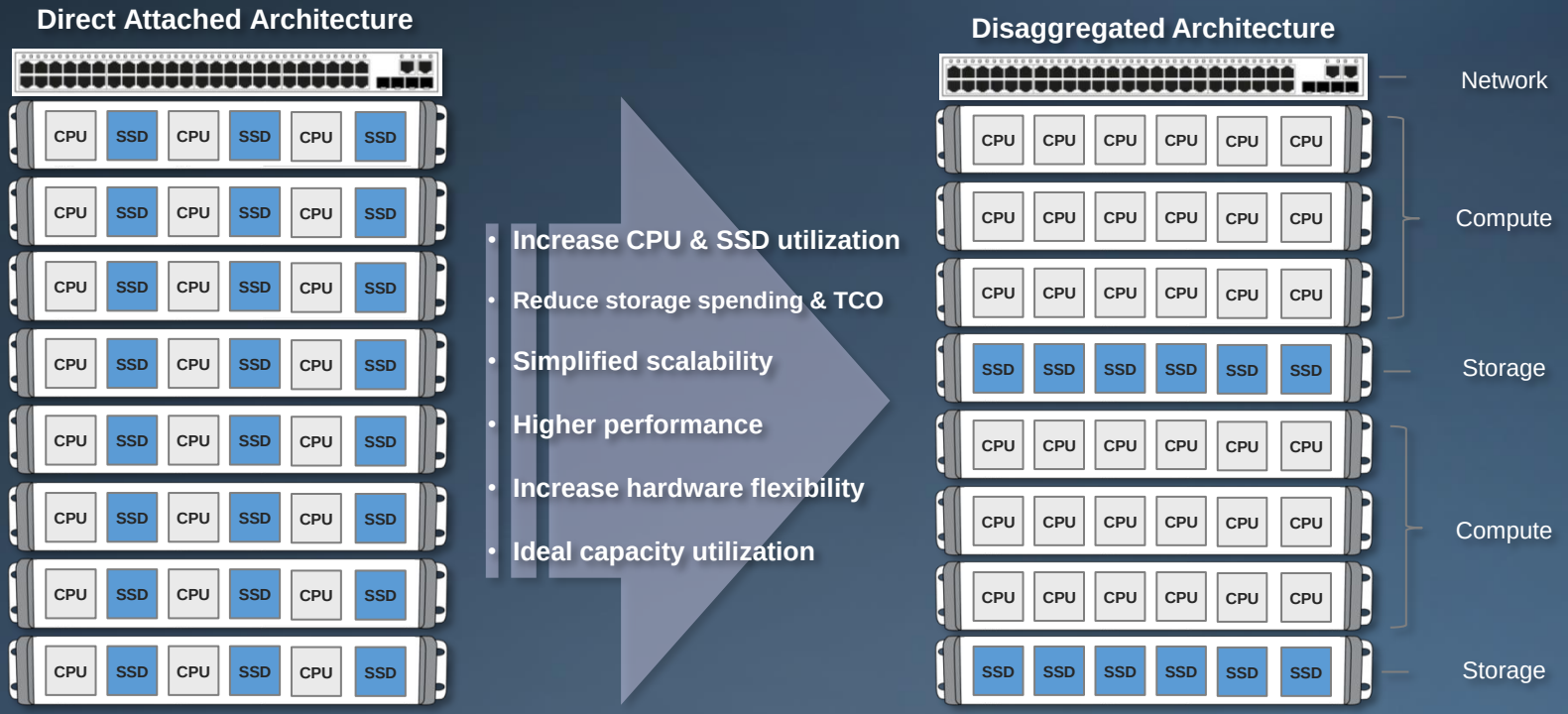
- Disaggregated Architecture for Datacenter Trend
- Slightly Hardware System for NVMe-oF
- Software Solution for NVMe-oF
- Future Works


Datacenter Trend

Disaggregated Architecture

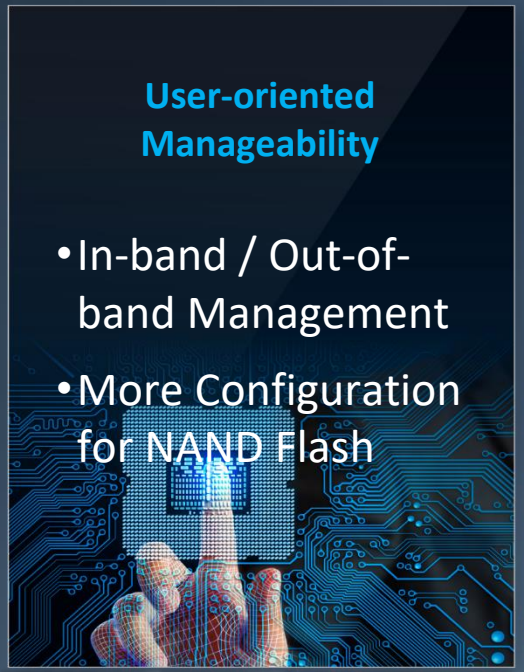
SAMSUNG

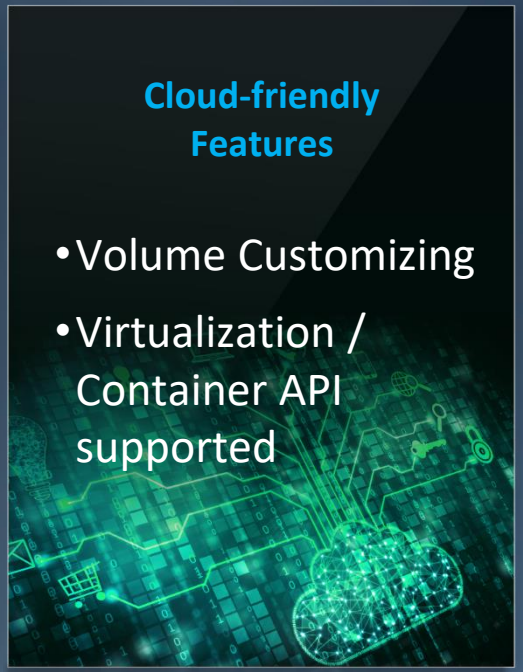
- Storage system is evolving towards the disaggregated architecture



Vital Virtues of Disaggregated Storage

SAMSUNG



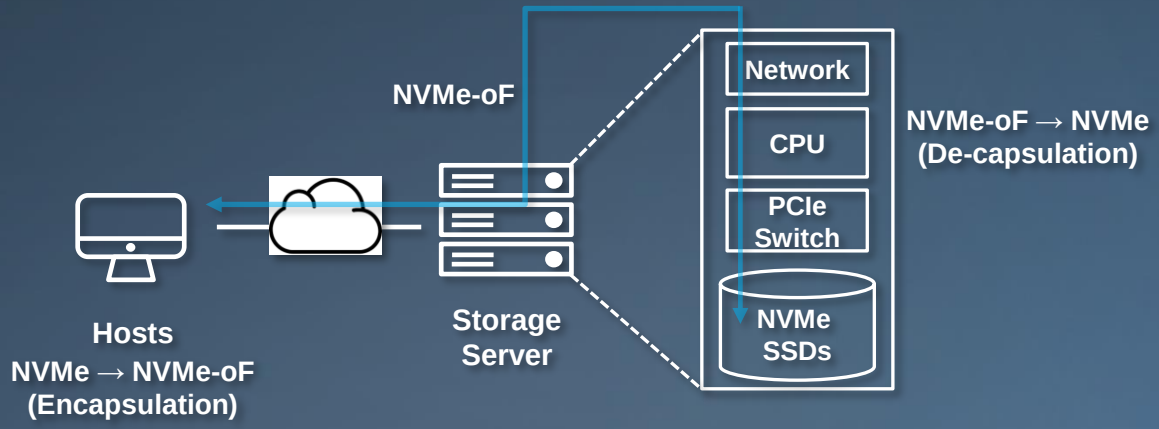


NVMe-oF Interface

- **Can break through the scaling limitation of PCIe-attached NVMe**

※ Up to few hundreds

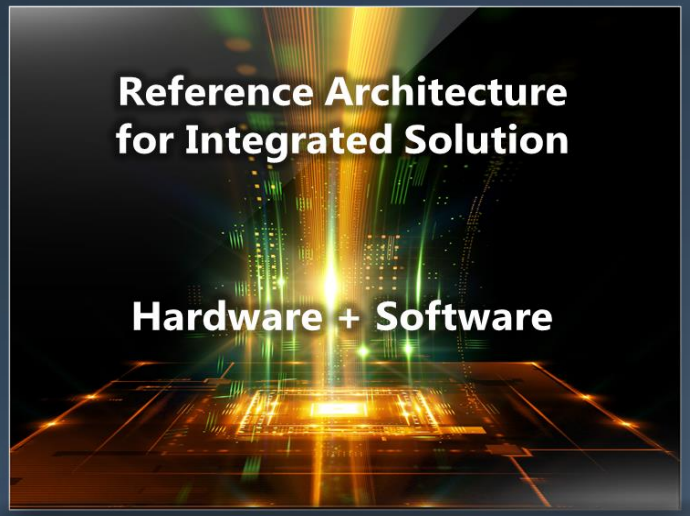
- **Uses a transport protocol over a network to access remote NVMe**
  - End-to-End NVMe semantics across a range of topologies
  - Retains NVMe efficiency and performance over network fabrics

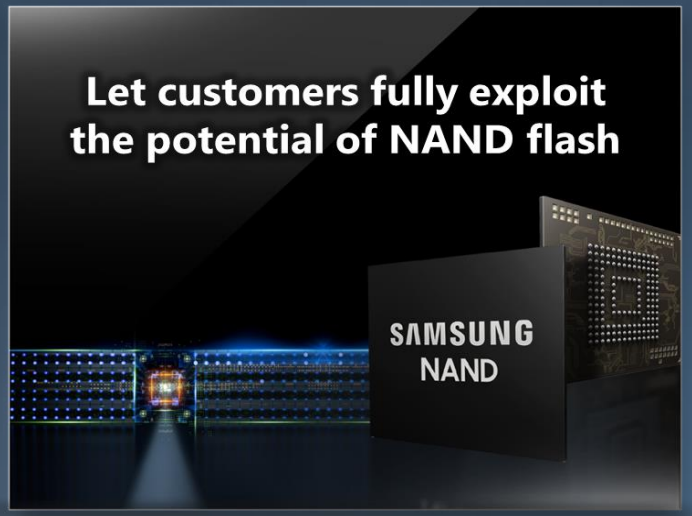


Project Poseidon

SAMSUNG

- Opensourced NVMe-oF Reference System Development Project
- Started from July, 2018, Planning to opensource in March.





Project Poseidon

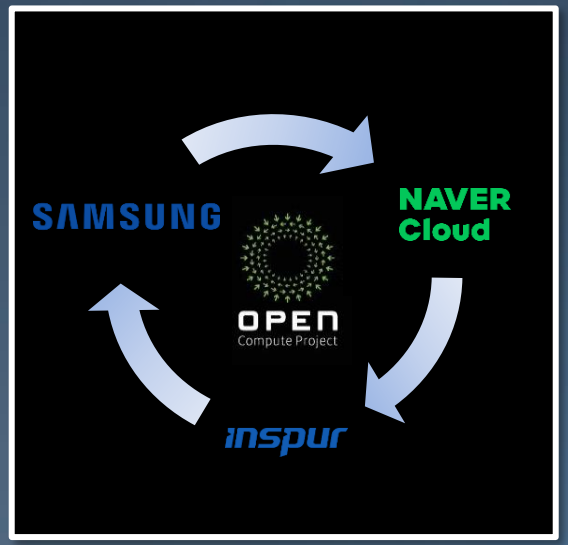
SAMSUNG

- OCP based industrial collaboration b/w
  - Component Vendor ↔ System Vendor ↔ Data Center
- Open-source HW & SW project to expand NVMe eco-system

Poseidon SW
Storage OS for NVMe-oF

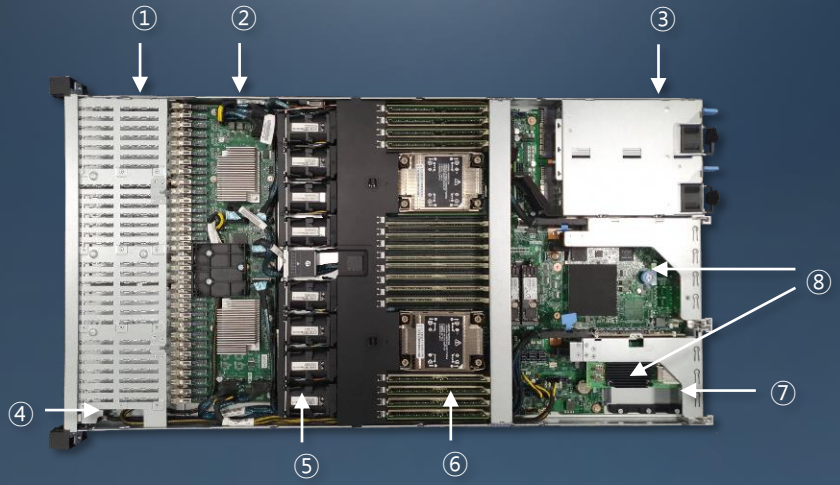
GitHub

Poseidon HW
E1.S/E3.S Ref. System




Hardware System for NVMe-oF

Poseidon V1 Server

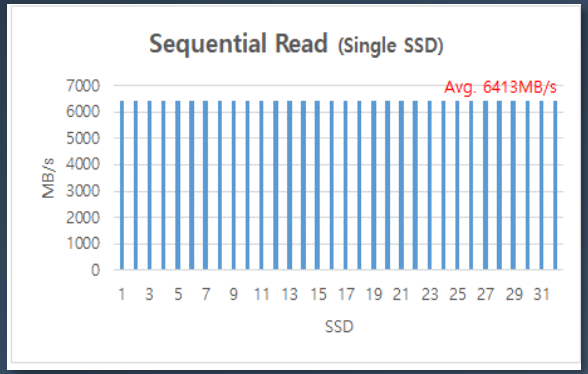


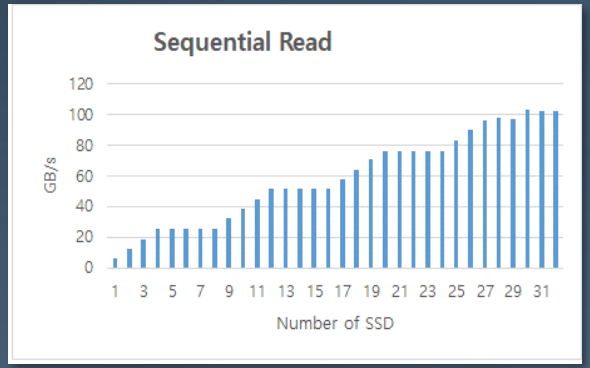
Chassis	EIA standard (19”)
Form Factor	1U
Processor	Next Gen x86 Processor
# of Processors	2
Max # of Memory Slots	32
Memory Speed	3,200 MT/s
Network	RDMA / TCP
Network Speed	Up to 100 GbE * 6 port
PCIe Version	PCIe 4.0
Storage	E1.S SSD * 32ea

① PM9A3 E1.S SSD	32ea
② SSD Backplane	1ea
③ Power Supply Unit (PSU)	2ea
④ Front Panel & IO Module	1ea
⑤ System Fan	8ea
⑥ Motherboard	1ea
⑦ OCP Mezzanine NIC	1ea
⑧ PCIe Slot (FHHL Card)	2ea

IO Performance – Single vs Aggregated

- Each 32 SSDs shows stable IO performance around 6.4GB/s
- Aggregated IO performance achieved > 102GB/s
- Lanes b/w CPU↔PCIe Switch (x16) are saturated with 4 SSDs





*Theoretical B/W limit: 128GB/s

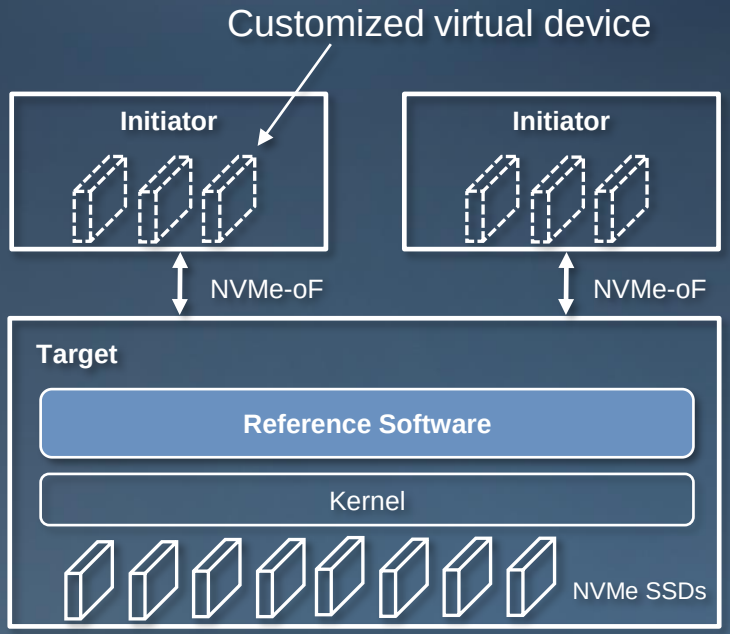
- **Samsung PM9A3 Specification**

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


Software Solution for NVMe-oF

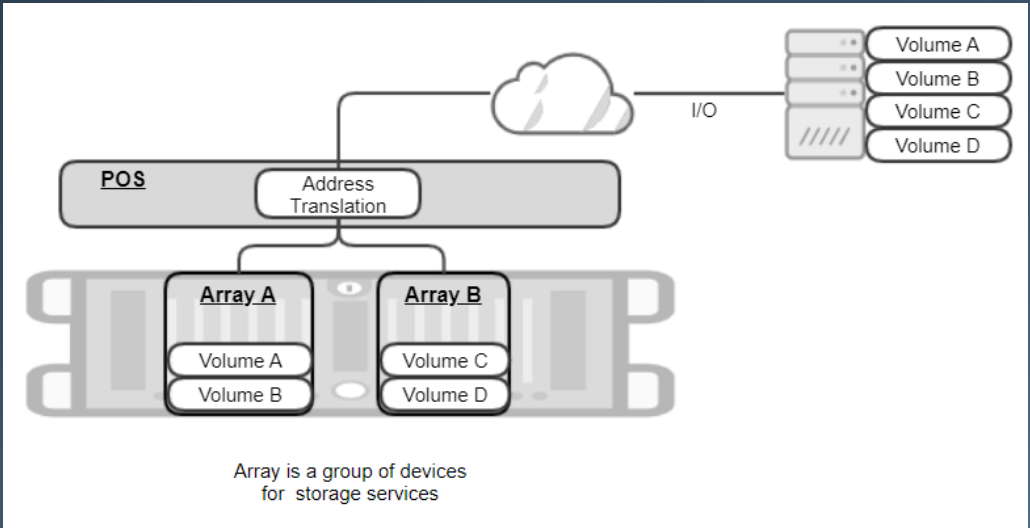
PoseidonOS Concept

- 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, …



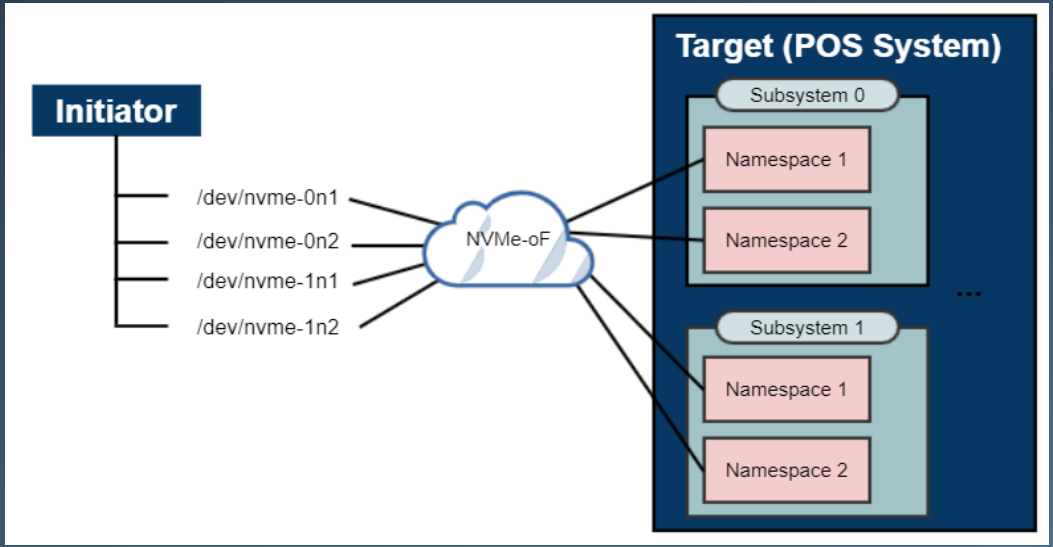
PoseidonOS Array

- User-defined collection of Physical NVMe SSDs
- Can contain 1 or more POS volumes



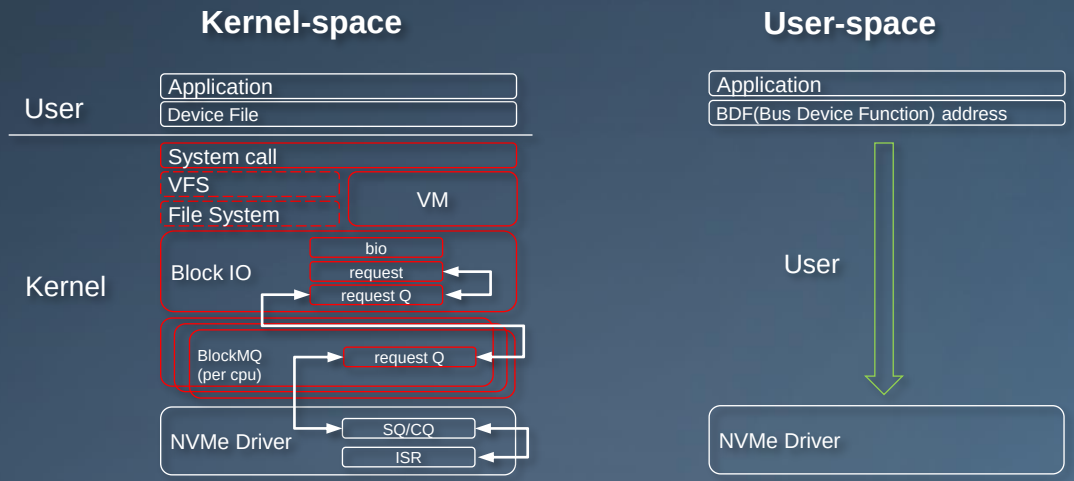
NVM Subsystem Perspective

- Each logical volume (i.e., namespace) is mapped to namespace
- Max 256 volumes are supported



Userspace Advantage

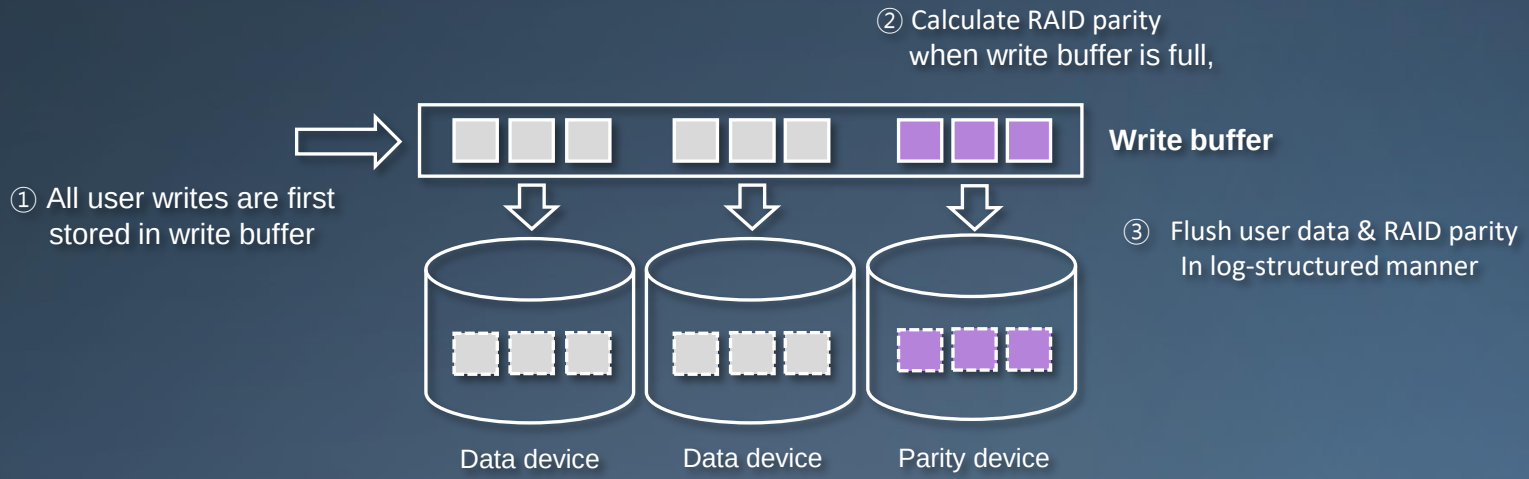
- **User-space NVMe-oF / NVMe IO**
  - Avoids overheads of system calls and data copies
  - Spends more CPU cycles for storage services
  - Enables better latency and IOPS



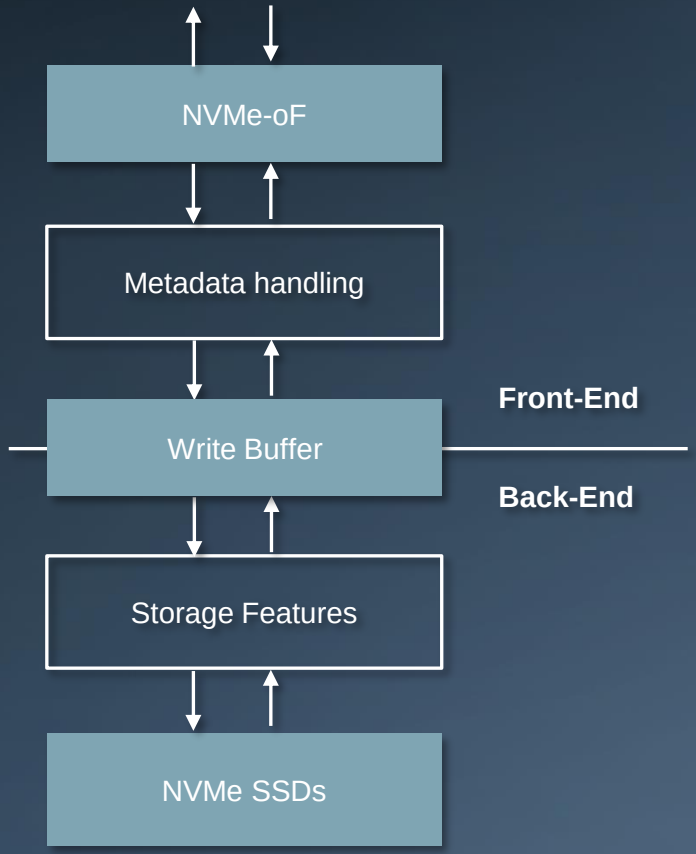
Log-structured RAID

SAMSUNG

- **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**

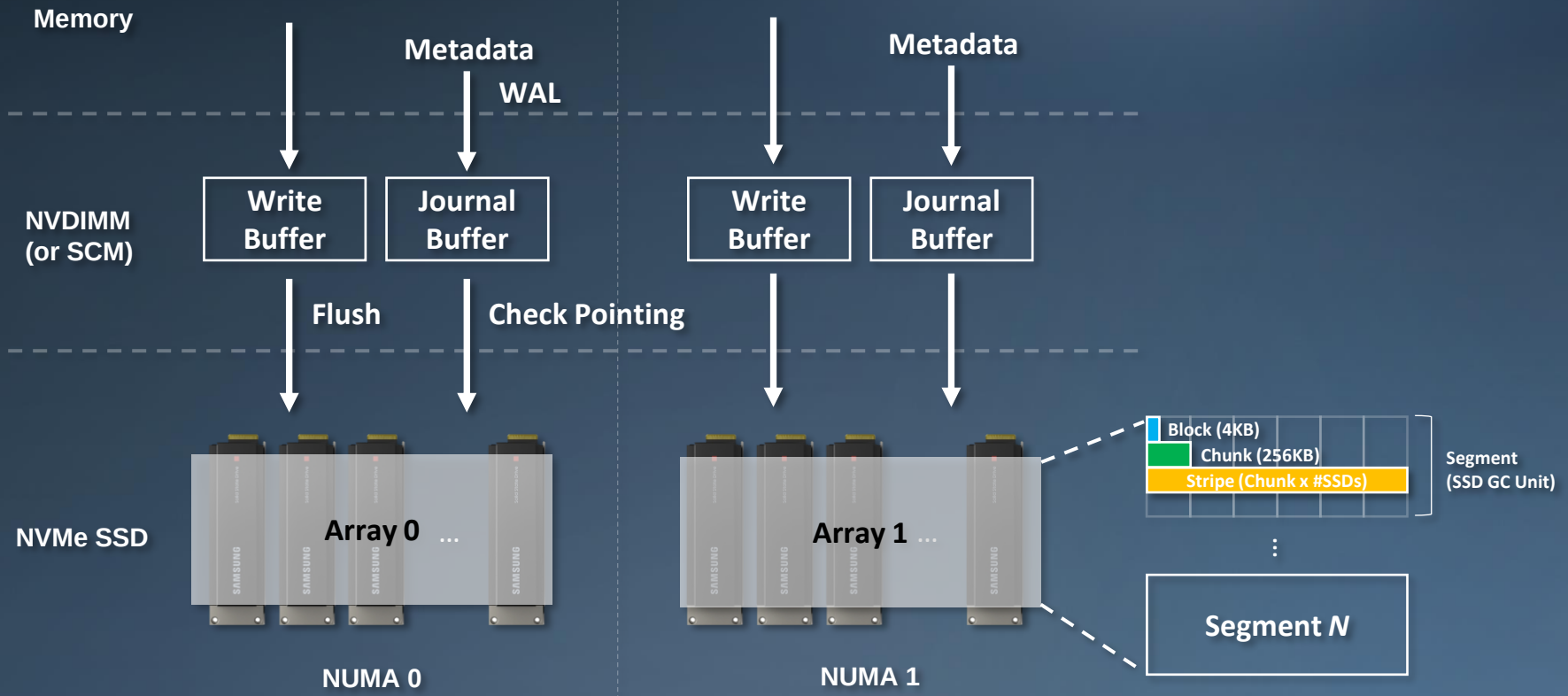


IO Stack



Front-End	Back-End
Latency-oriented	Throughput-oriented
User IO, Network only	RAID, Flush, GC, …
Dedicated thread (No context switch)	Thread pool
Direct device interaction	Indirect device interaction

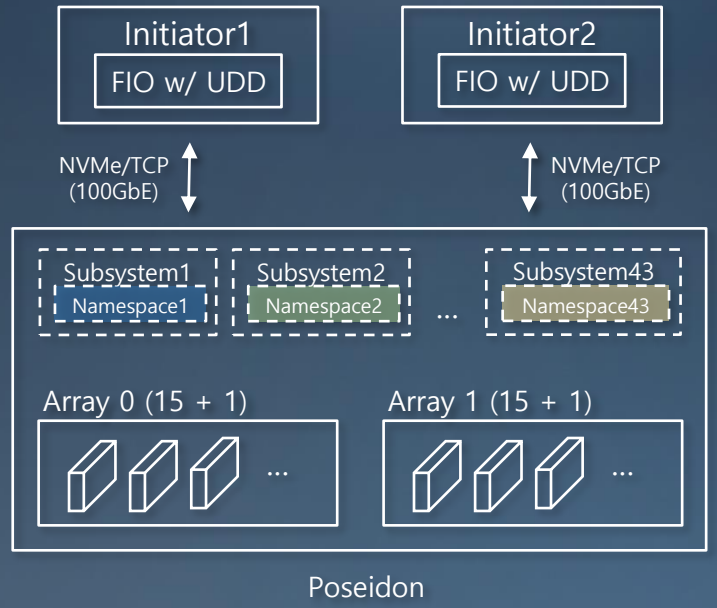
Storage Hierarchy in PoseidonOS



Experiment Environments

SAMSUNG

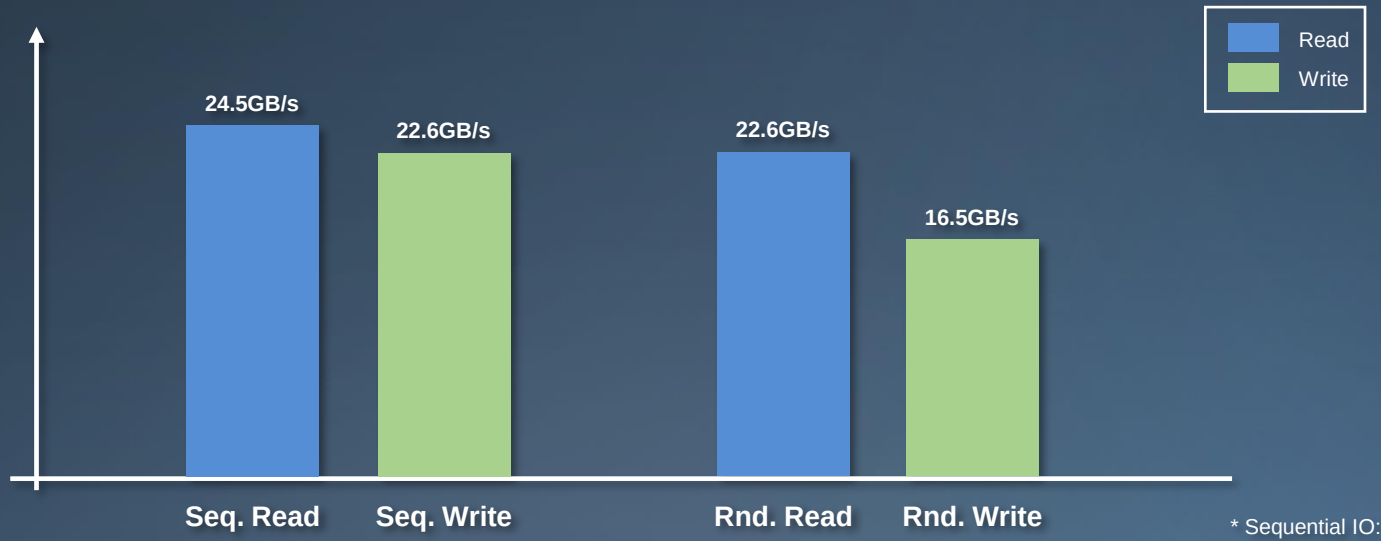
- **PCIe Gen4 SSD * 32**
- **200GbE Network Connection**
  - NVMe/TCP
- **2 Arrays / 43 Volumes**
- **RAID 5 (15 + 1)**
- **Using uDriver in initiator-side**



* 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

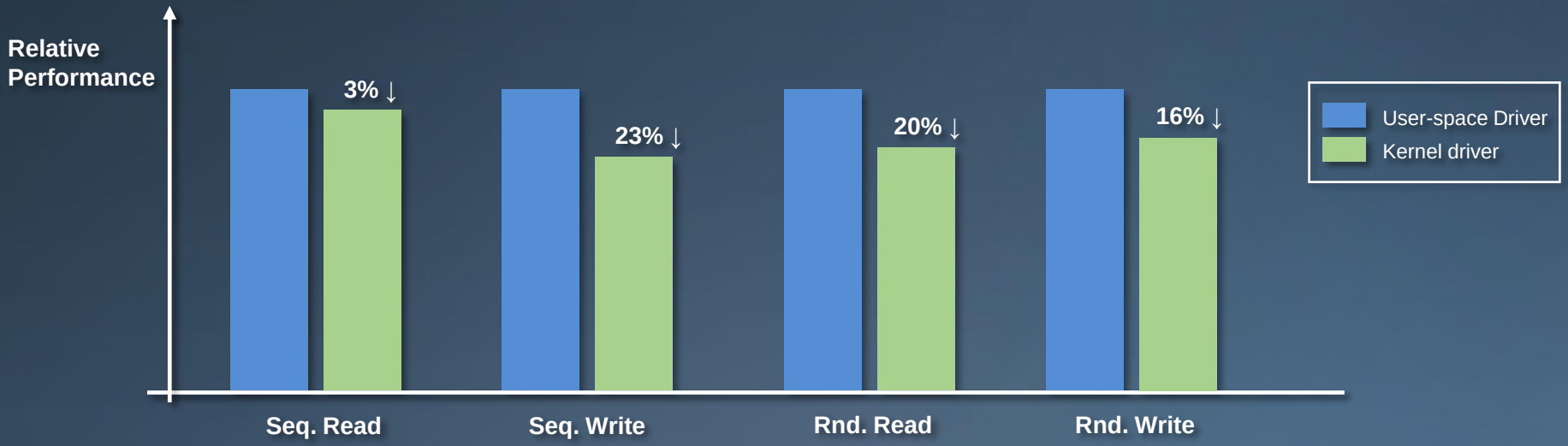


* Sequential IO: 128KB, QD4
Random IO: 4KB, QD128

Performance - Initiator SW Stack

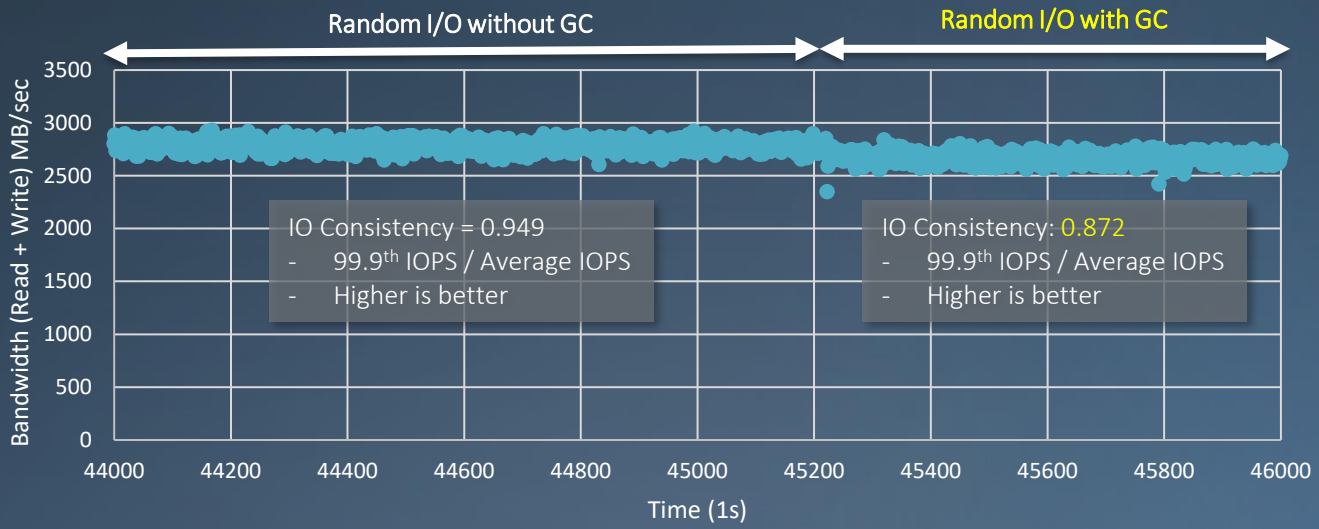
SAMSUNG

- 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**



Future Works

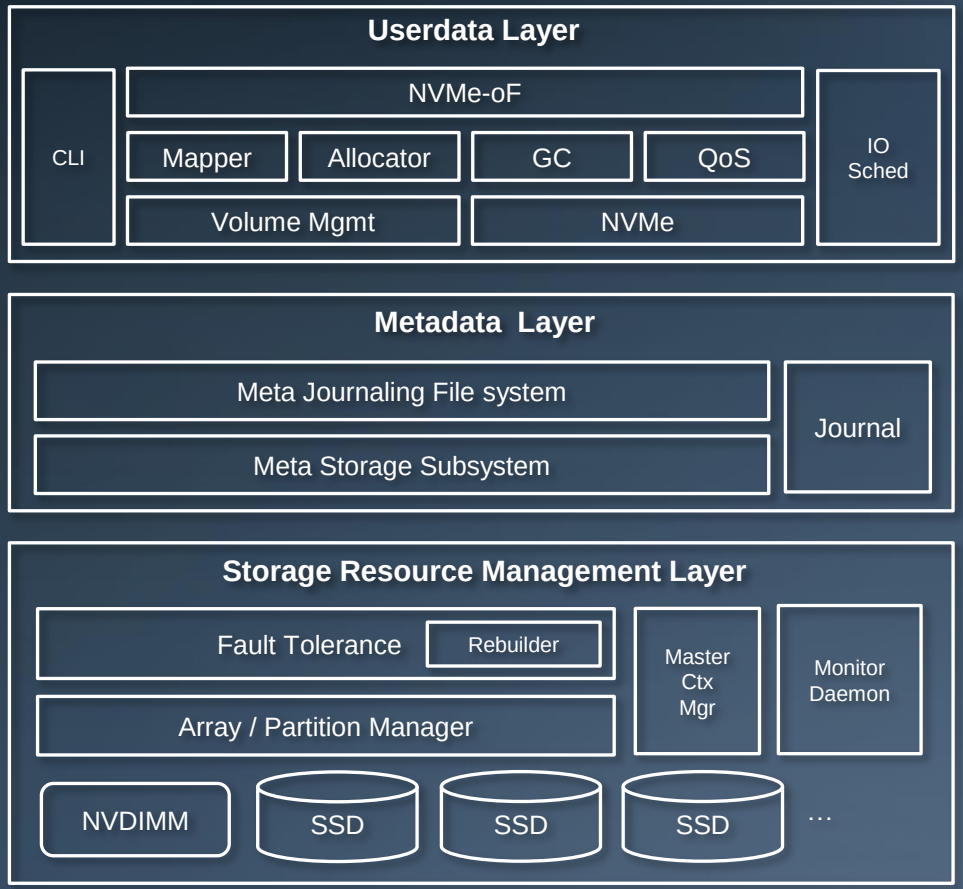
SAMSUNG

- Datacenter adoption and provide for a real cloud service
- 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

Internals



- User I/O handling
- Provides NVMe-oF connectivity
- Logical Device mgmt - Volume
- Performance Optimization
- User IO QoS

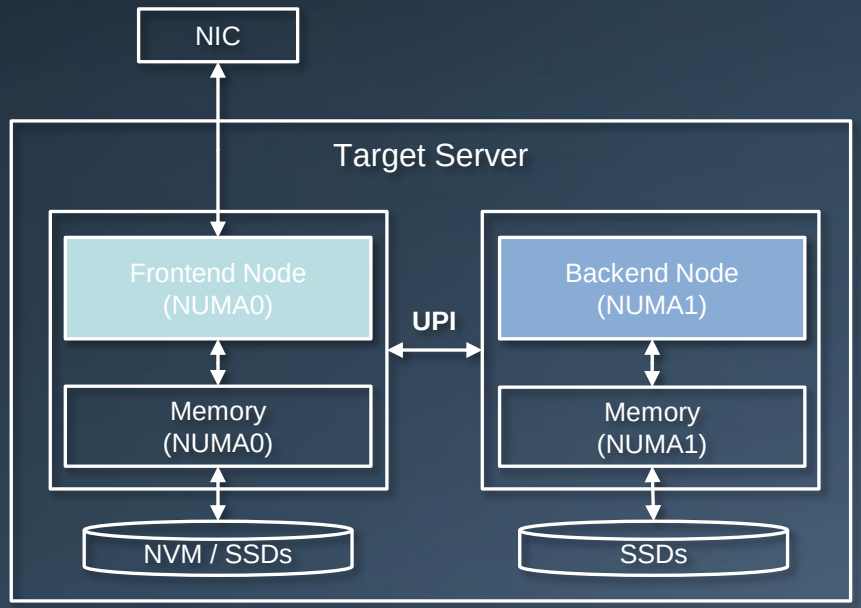
- Guarantee ACID of metadata
- Metadata I/O handling
- Journaling and Restore

- Provides Fault Tolerance Feature - RAID / EC
- Partition mgmt - System / User / Meta Area
- Physical device mgmt - SSD Array, NVDIMM
- SSD device monitoring

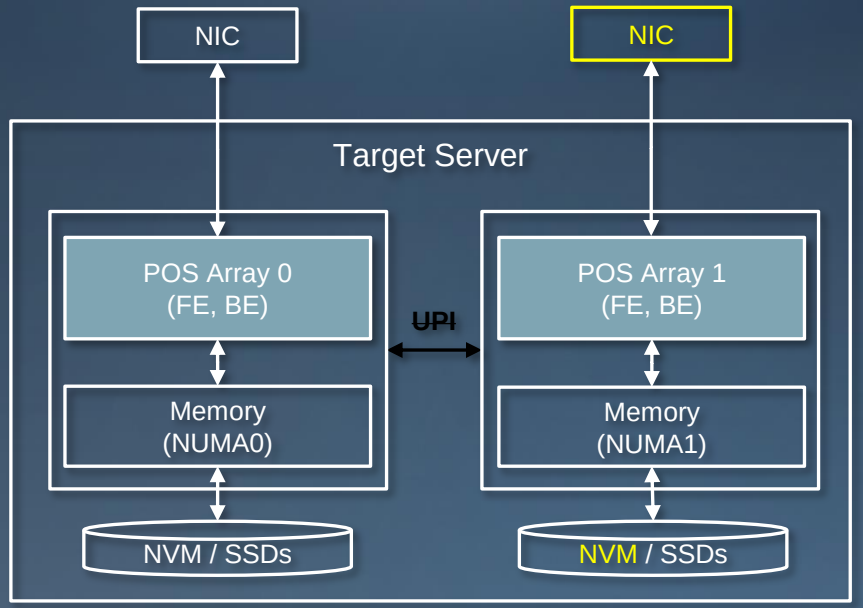
Use cases

SAMSUNG

Single Array



Two Array (NUMA-aware)



- NUMA0: User IO, Network,
- NUMA1: Flush, RAID, GC, …