SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

# Host Management of NVM Express™ Exported NVM Subsystems in PCIe® SSDs

Lee Prewitt, Microsoft
Chaitanya Kulkarni, Nvidia
Mike Allison, Samsung

www.sniadeveloper.org

# Hyperscaler Perspective


## Lee Prewitt

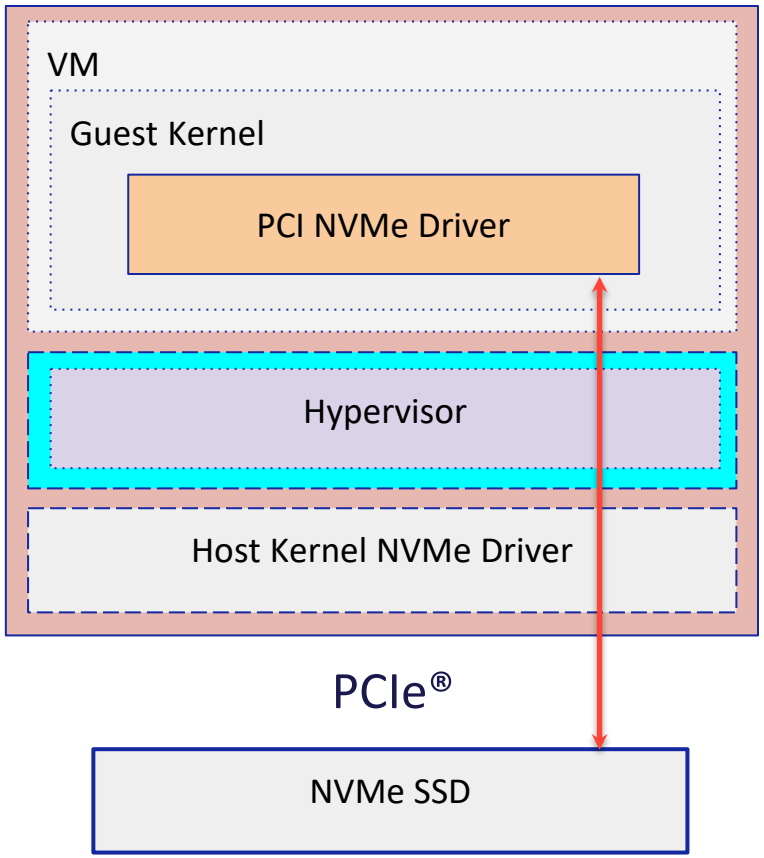
Director Cloud Hardware Storage

Microsoft

# Hyperscaler Perspective

- Challenge: Consistent, Predictable, & Performant VM Guest Experience
  - Live migration should not be observable by the guest VM
  - Live migration should be efficient and performance isolated
  - Hardware iteration should not be observable by the guest VM
  - Includes SSD upgrade, platform change, etc.
- Historical Approach to Achieve these Goals
  - Standards Solve this ChallengHyperscaler-specific logic to make changes invisible to the guest VM
  - For lower latency, vendor specific changes to SSDs were introduced
- Standards Solve this Challenge in a Durable Manner
  - Samsung, Google, Microsoft, and others have led standards activities in NVMe™ and OCP since 2023 to solve this challenge
  - Features are delivered piece by piece to advance the ecosystem
  - Hardware & software are tested at scale in a robust manner




**Robust standards needed to broadly deliver the consistent, predictable, performant VM storage experience**

# Why is Live Migration Needed?

- Customers hate to be interrupted when they are working
  - Need very low interruption rates on imperfect hardware
  - Measured as the number of customer VM events per year
    - Annual Interruption Rate (AIR)
- Live Migration is required to meet AIR goals
- Use cases for Live Migration
  - Actual hardware failure on the node (may not affect all VMs)
  - Predicted hardware failure (allows timely migration before issues occur)
  - Scheduled node maintenance (Hardware or firmware upgrade)
  - Load balancing (allows a mixture of large and small VMs to coexist)

# Why is Para-virtualization Not Enough?

- SSDs are fast
- Para-virtualization is slow
  - Leaves IOPs on the table
- Para-virtualization is expensive
  - Reduces the number of sellable CPU cores on the node

# First Thing is to Get the Hyper-visor out of the IO Path

- TP4165 Tracking LBA Allocation
  - Host migration software only needs to copy the LBAs that are in use
- TP4159 PCIe® Infrastructure for Live Migration
  - Namespace Migration
    - Track changed LBAs between copy passes to minimize work
  - Controller Migration
    - Suspend source controller and migrate its state to the destination controller
- Quality of Service Control
  - Allows for VM resource isolation
  - But more importantly it allows for VM rate limiting during Live Migration

# But This is Not Enough

- Need to get out of trapping the Admin Queue as well
  - Not that CPU intensive, but it does count
  - More importantly, need to minimize the hypervisor surface in the face of confidential computing
- But we still need to make sure that the VM does not see a change in hardware after it's been migrated
- PCIe® Exported NVM Subsystem – How to a consistent lie
  - Abstracts Controller and Namespace Inquiry Data
  - Abstracts Get Log Page (page support as well as returned data)
  - Abstracts Command support

# Are We There Yet?

- Still a couple of pieces needed to complete the picture
- Resource allocation
  - Standardized way attach resources (Namespaces, Queue Pairs, etc.) to portions of a device
  - These sub portions may be PCIe® Physical Functions, SR-IOV Functions or SIOV Assignable Interfaces
- TDISP support for NVMe™
  - Standardize the DEVICE_INTERFACE_REPORT
  - Allows for standardized support for Confidential Compute

# Live Migration flow

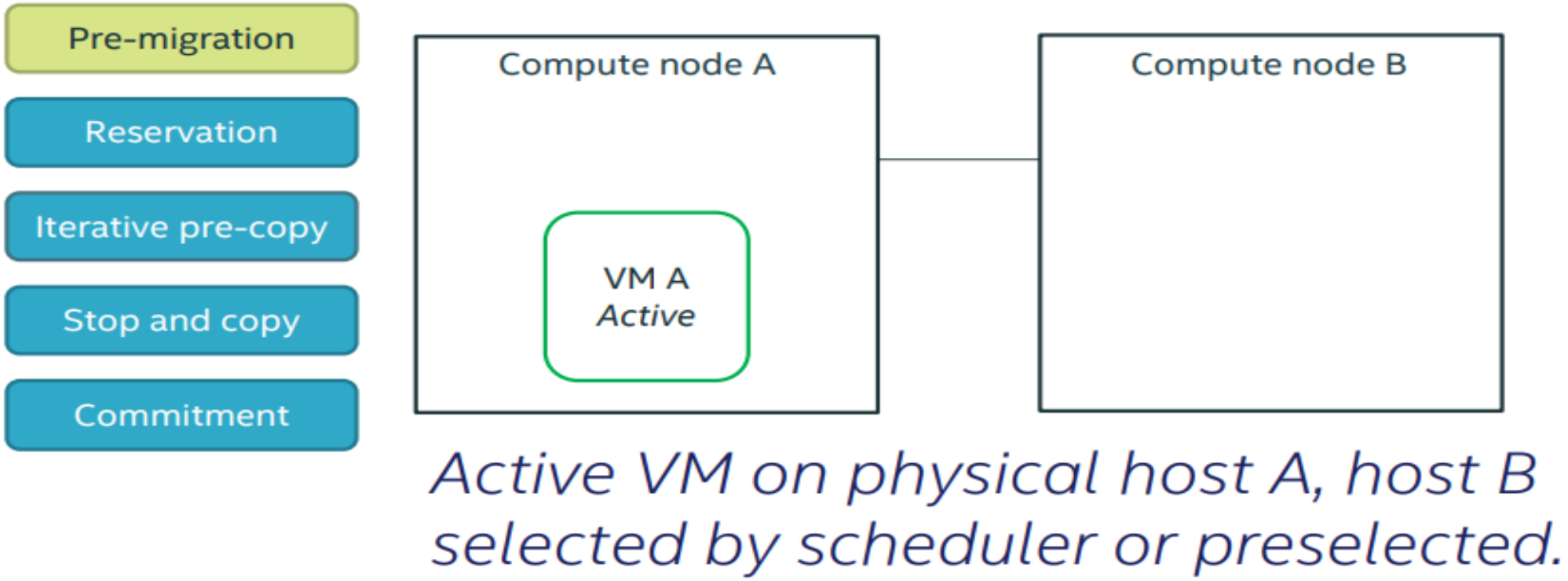
Chaitanya Kulkarni

Director – Systems Architect

NVIDIA

# Pre-migration



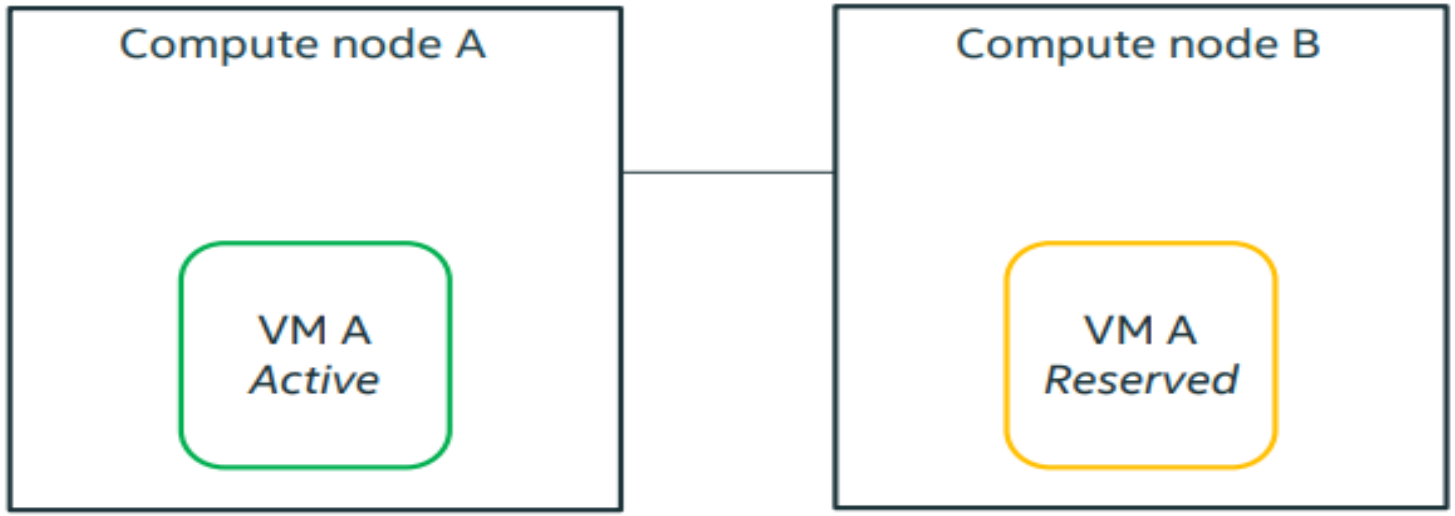

*Active VM on physical host A, host B selected by scheduler or preselected.*

Dive Into VM Live Migration Openstack
Liberty Summit 2015

# Reservation

- Pre-migration
- Reservation
- Iterative pre-copy
- Stop and copy
- Commitment



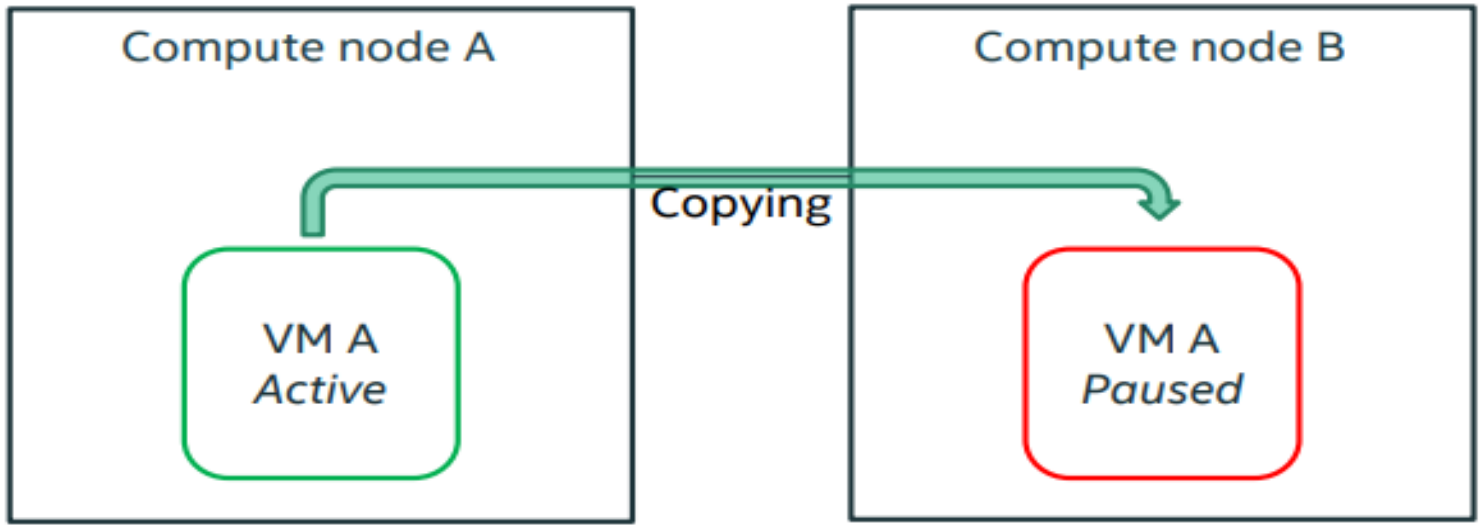

*Confirm availability of resources on host B; reserve a new VM.*

Dive Into VM Live Migration Openstack
Liberty Summit 2015

# Iterative pre-copy

- Pre-migration
- Reservation
- Iterative pre-copy
- Stop and copy
- Commitment



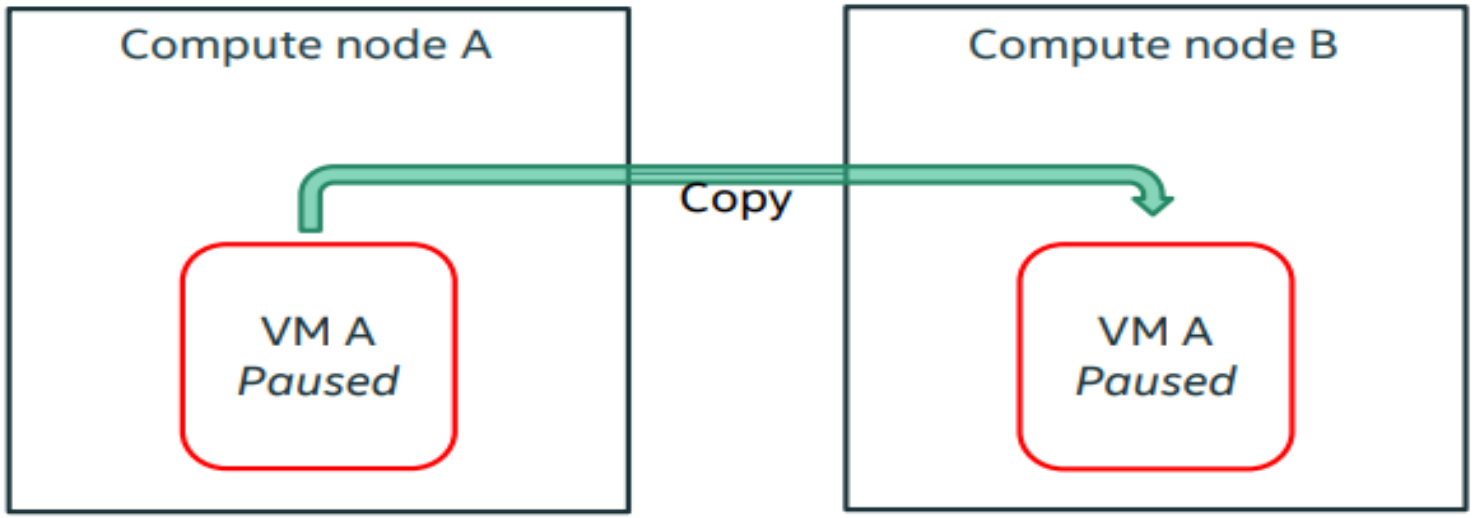

*Memory is transferred from A to B and next dirtied pages are iteratively copied.*

Dive Into VM Live Migration Openstack
Liberty Summit 2015

# Stop and copy

- Pre-migration
- Reservation
- Iterative pre-copy
- **Stop and copy**
- Commitment

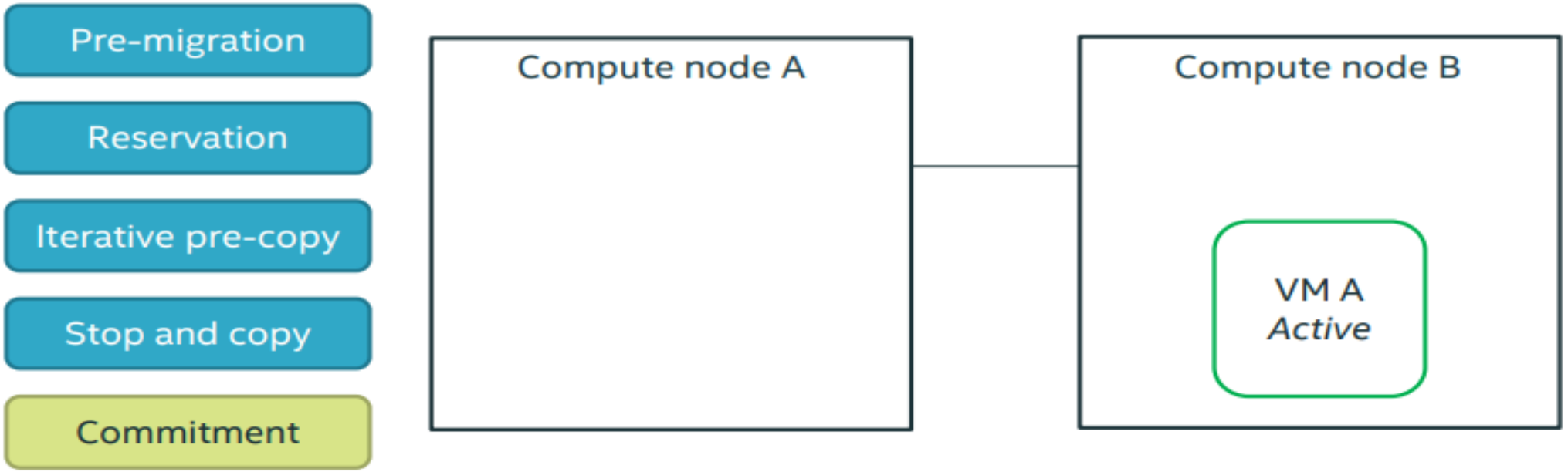



*Suspend VM and copy remaining pages and CPU state.*

Dive Into VM Live Migration Openstack
Liberty Summit 2015

# Commitment




*Host B becomes primary host for VM A.*

Dive Into VM Live Migration Openstack Liberty Summit 2015

# Different Storage Virtualization modes

- VFIO backend
- NVM Express® User Space backend
- Generic Block device backend
- File backend where device is formatted with file system.

# Why Use NVM Express® with VFIO Mode ?

- Virtual machines often make use of direct device access (when configured for the highest possible I/O performance).
- From a device and host perspective, this simply turns the VM into a userspace driver, with the benefits of significantly reduced latency, higher bandwidth.

# Why Use NVM Express® with VFIO Mode ?

- Applications, particularly in the high-performance computing field, also benefit from low-overhead, direct device access from userspace.
- Examples include network adapters (often non-TCP/IP based) and compute accelerators.
- NVM Express Protocol is particularly designed for the high performance where users can get maximum performance out of storage.

# Why Use NVM Express™ with VFIO Mode ?

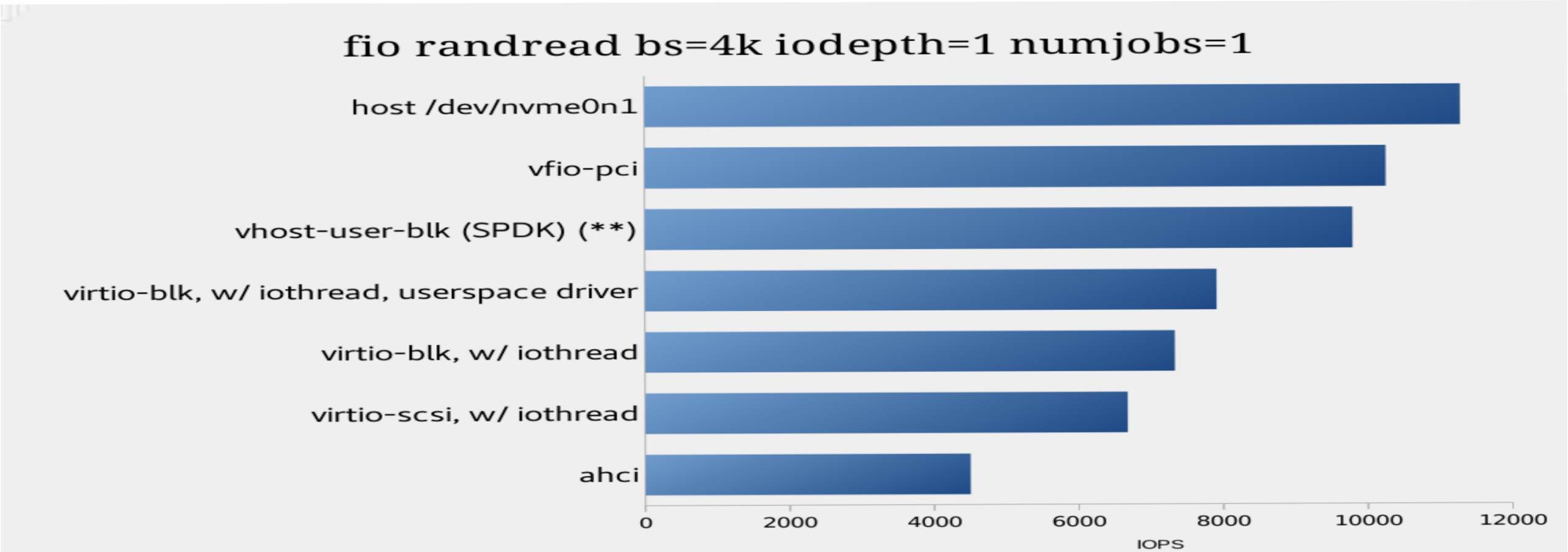
- Prior to VFIO, these drivers had to either go through the full development cycle to become proper upstream driver or to be maintained out of tree.
- OR
- Make use of the UIO framework, which has:
  - no notion of IOMMU protection
  - limited interrupt support
  - requires root privileges to access things like PCI configuration space.

# Why Use NVM Express™ with VFIO Mode ?

- The VFIO driver framework intends to unify these by replacing both the KVM PCI specific device assignment code as well as provide a more secure, more featureful userspace driver environment than UIO.



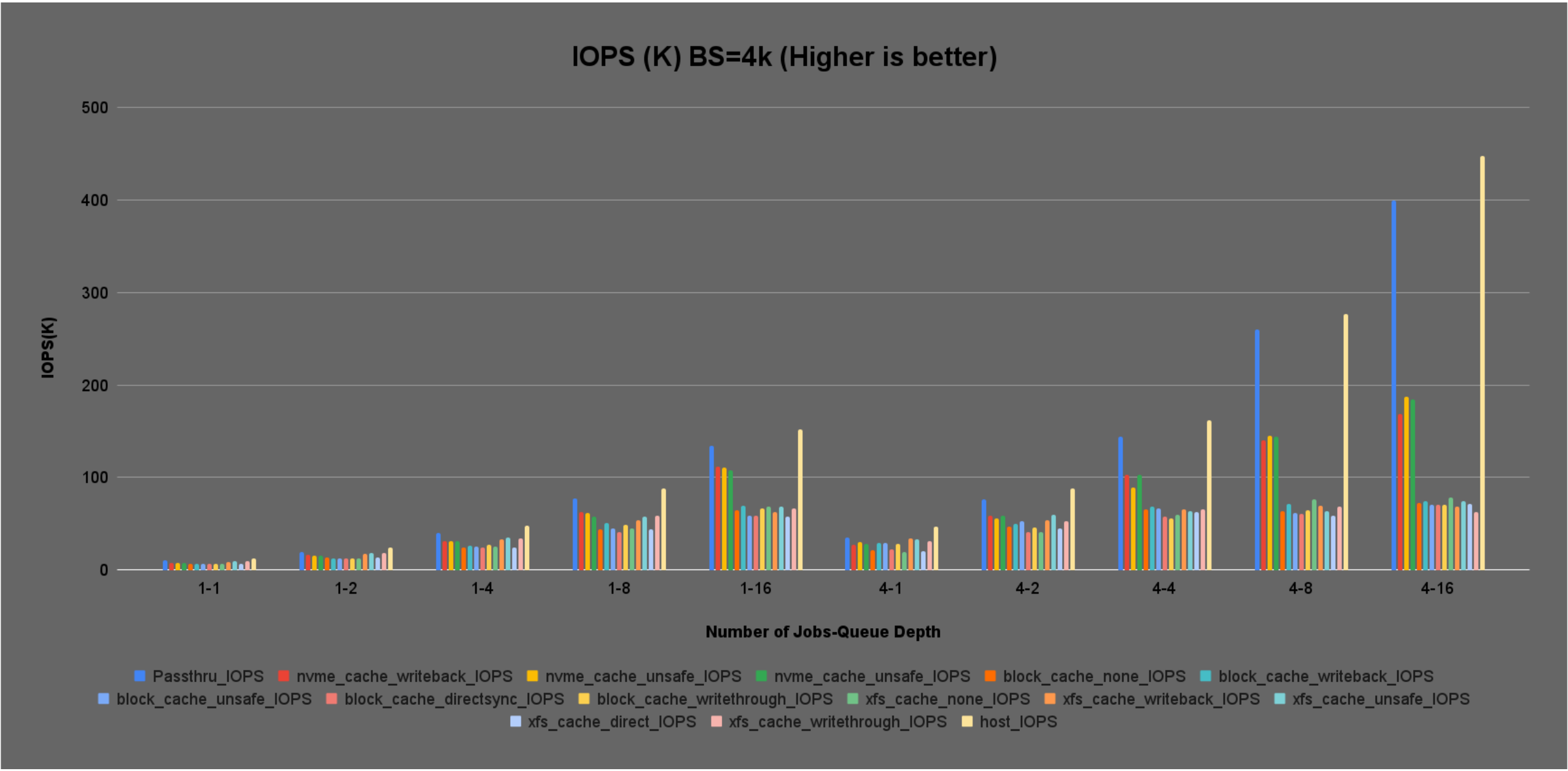

Backend: NVMe, Intel® SSD DC P3700 Series 400G
Host: Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz, Fedora 28
Guest: Q35, 1 vCPU, Fedora 28
QEMU: 8e36d27c5a
(**): SPDK poll mode driver threads take 100% host CPU cores, dedicatedly

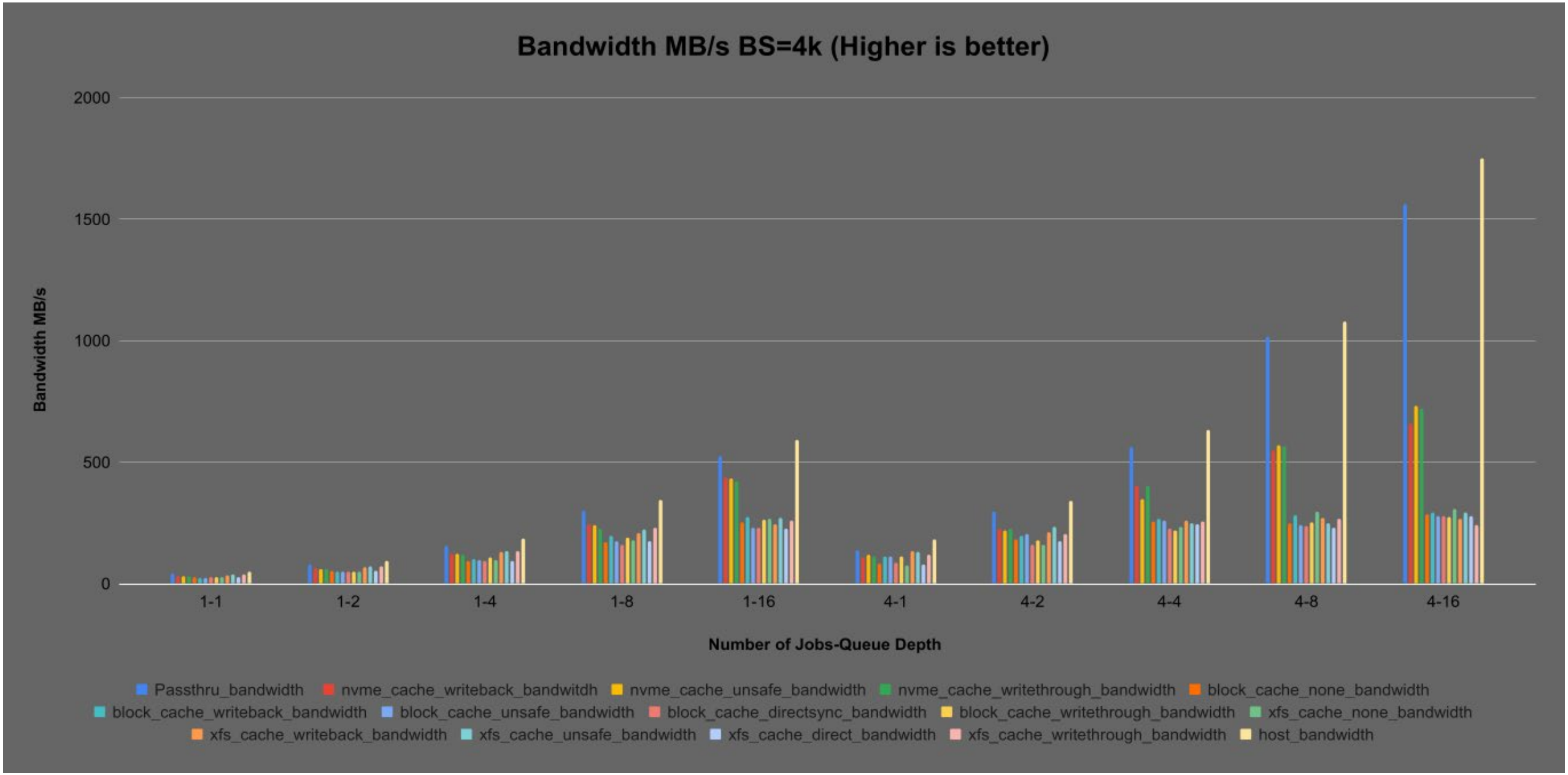
# Performance Matrix

- IOPS (K)/Bandwidth (MB/s)/Latency
- CPU Guest User/ System
- CPU Host User/System
- IOPS Per Core/Bandwidth Per Core
- Block Size 4k, jobs 1 and 4
- Queue Depth 1/2/4/8/16
- Backend Categories:-
  - Pass-through (VFIO)
  - QEMU Userspace NVMe™ driver NVMe controller (3 Modes)
  - QEMU virio-blk on NVMe controller (5 Modes)
  - File created on XFS formatted on NVMe controller (5 Modes)

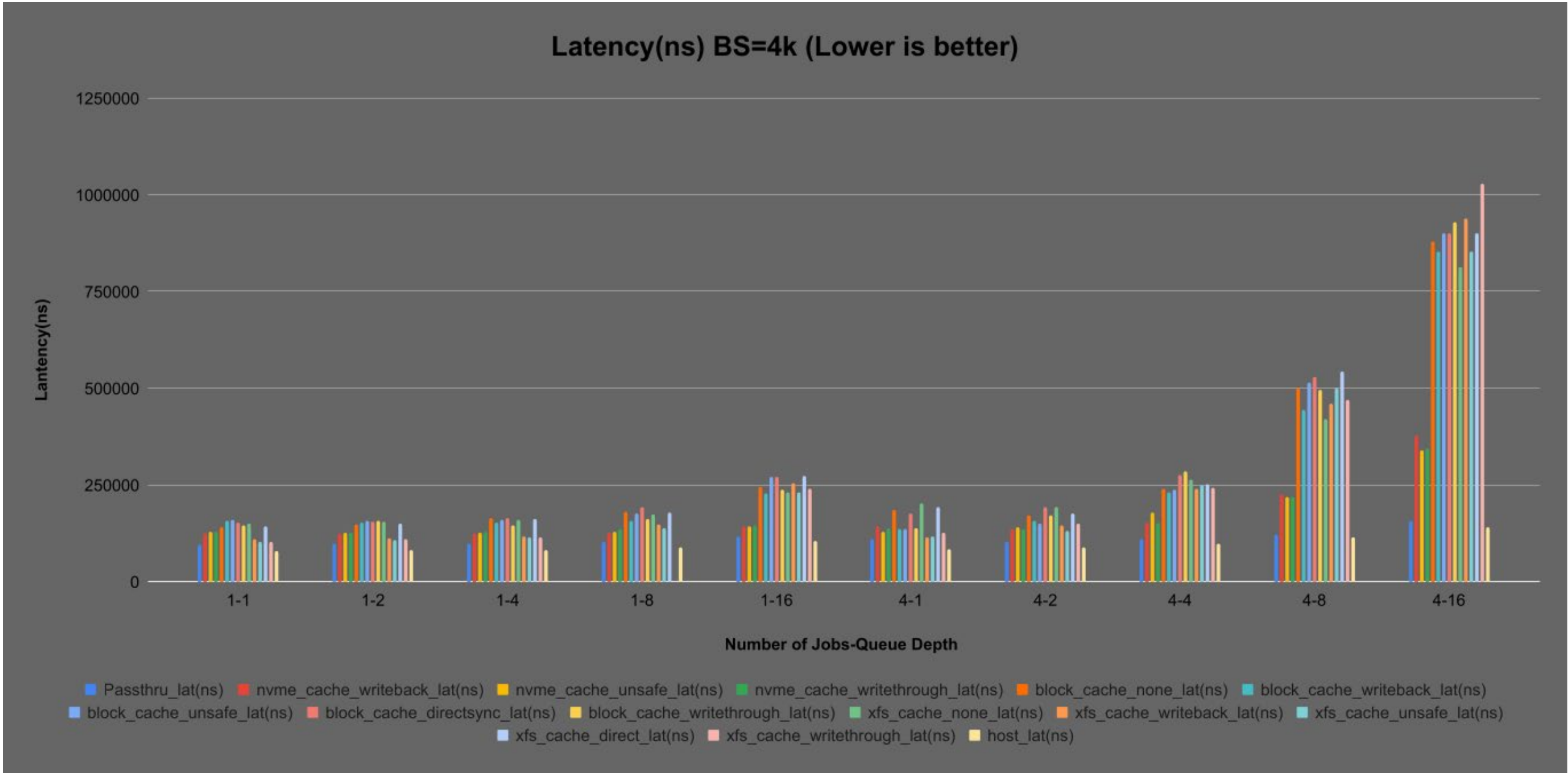
 



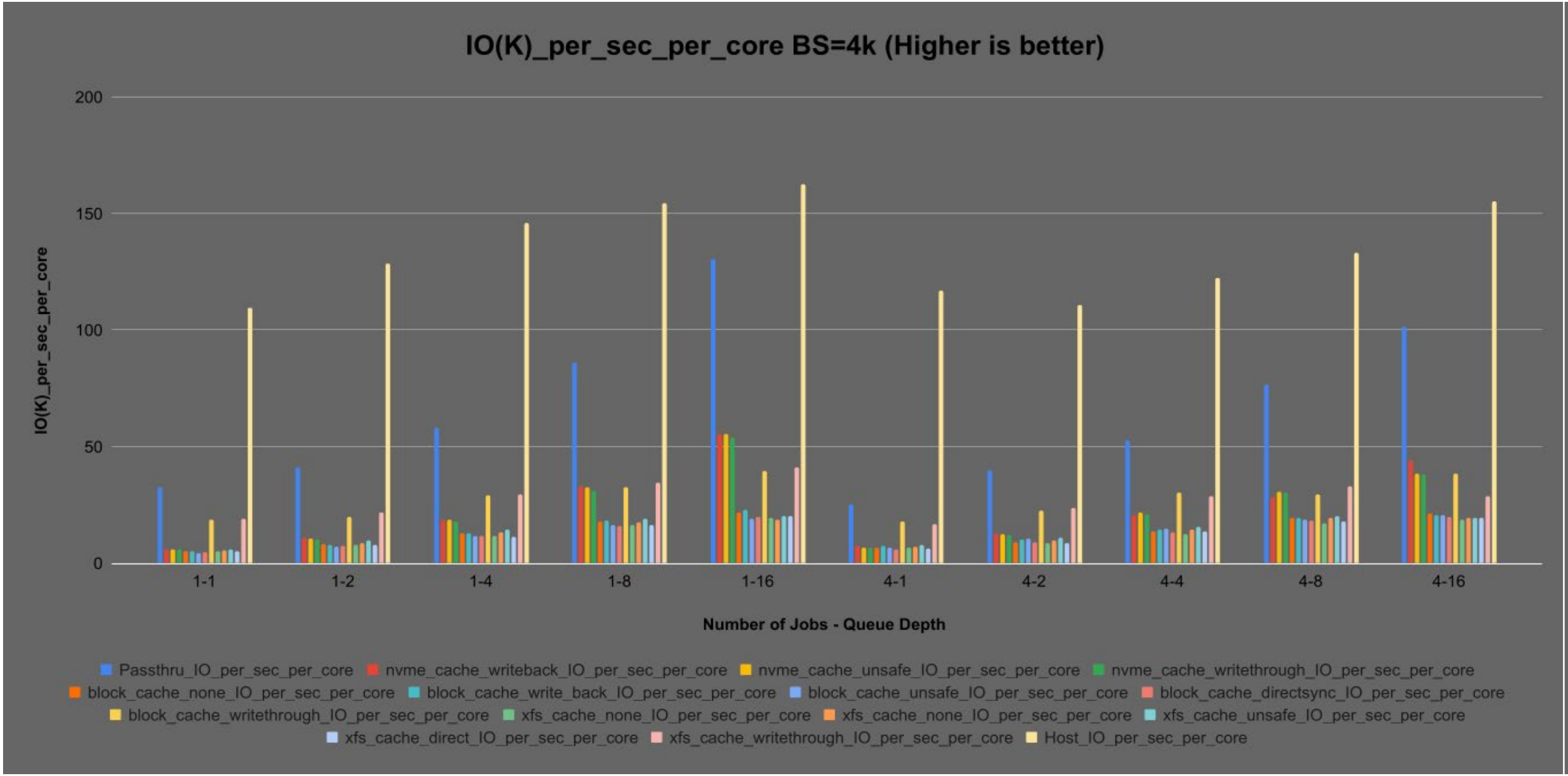

 



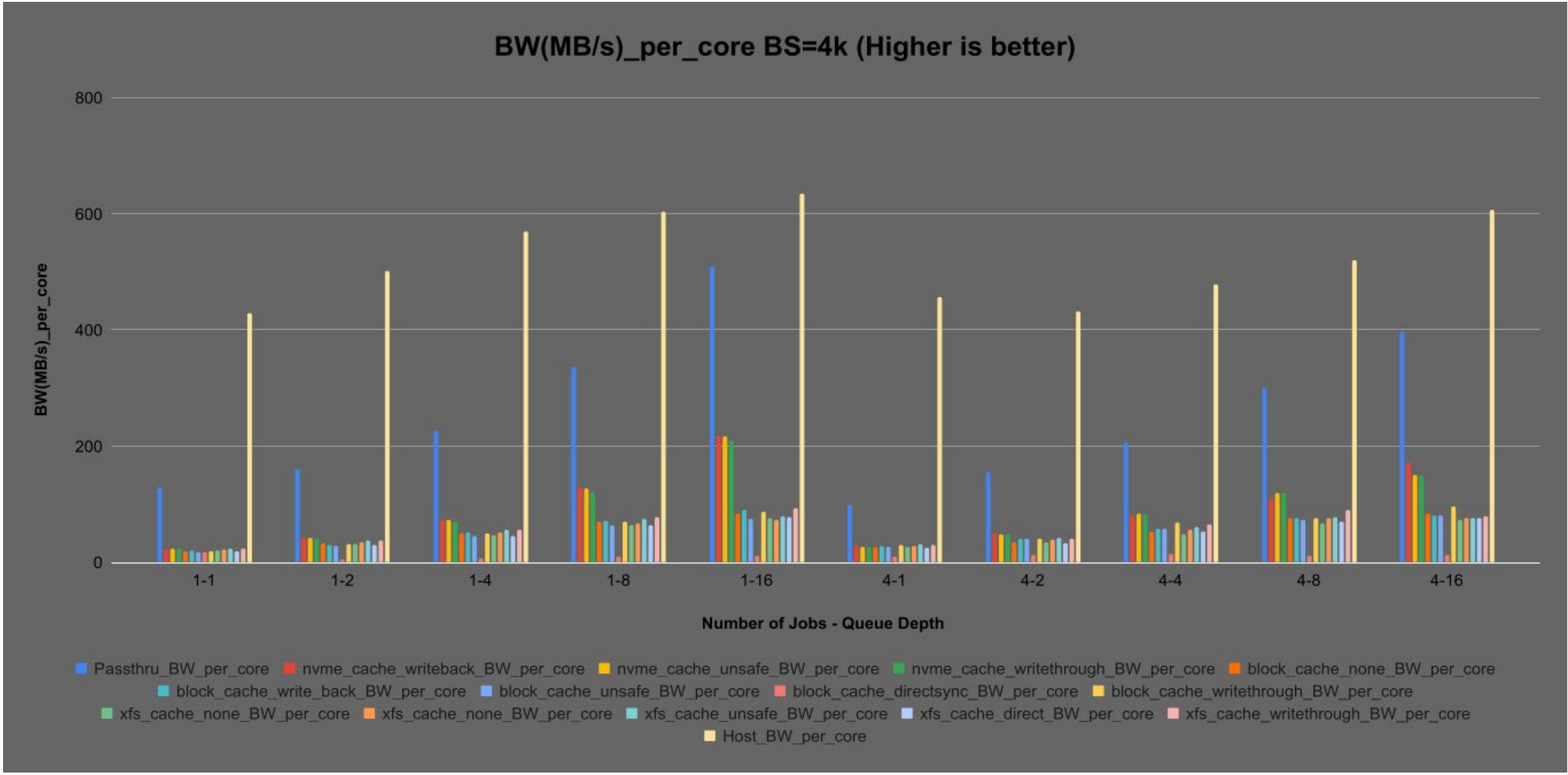

 







# VFIO NVM Express™ Live Migration FSM

- Supporting Live Migration includes creating vfio-nvme implementation that will support VFIO live migration Finite State Machine (FSM).
- This also includes support from the NVM Express protocol that will allow us to execute the subsequent command that are sent from the VFIO FSM.

# NVMe™ PCIe® Exported NVM Subsystems


## Mike Allison

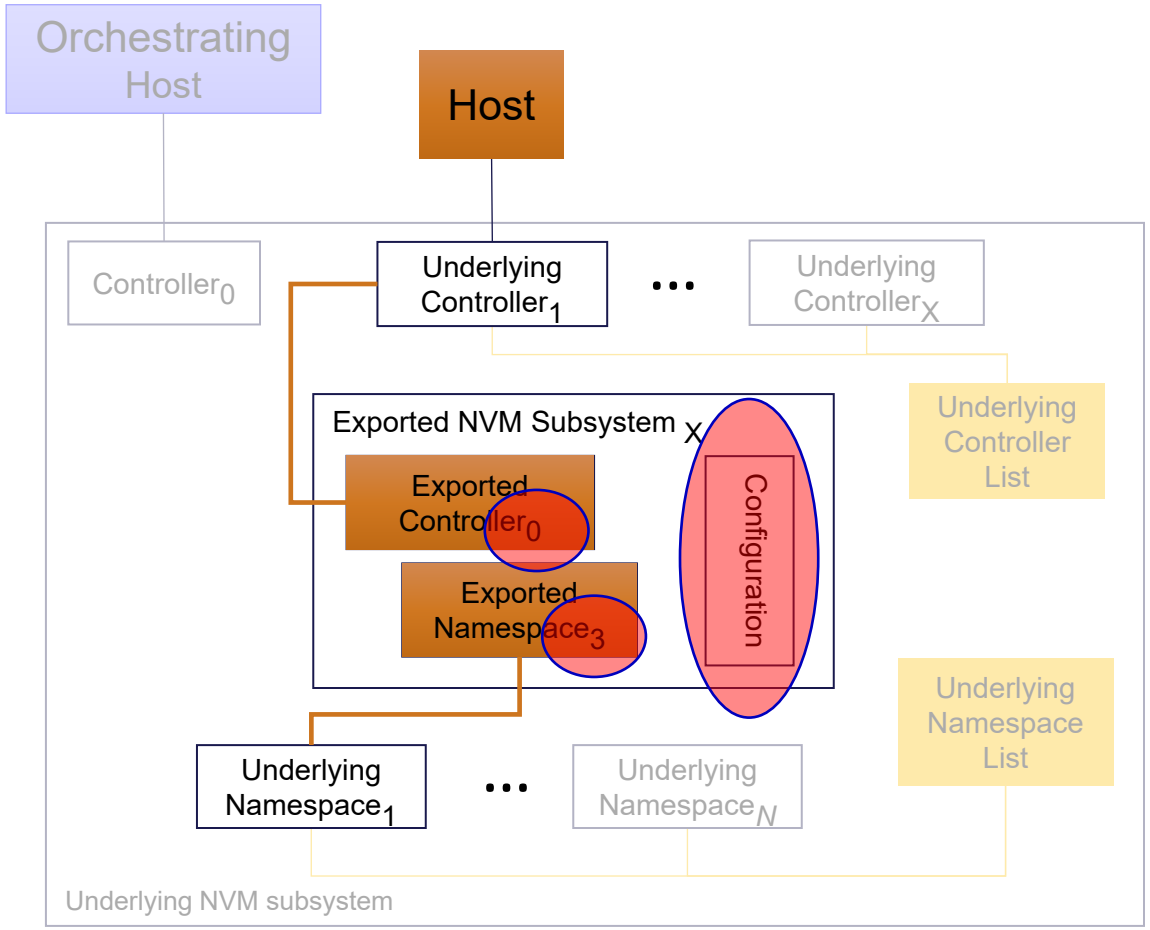
Sr. Director NAND Product Planning - Standards

Samsung

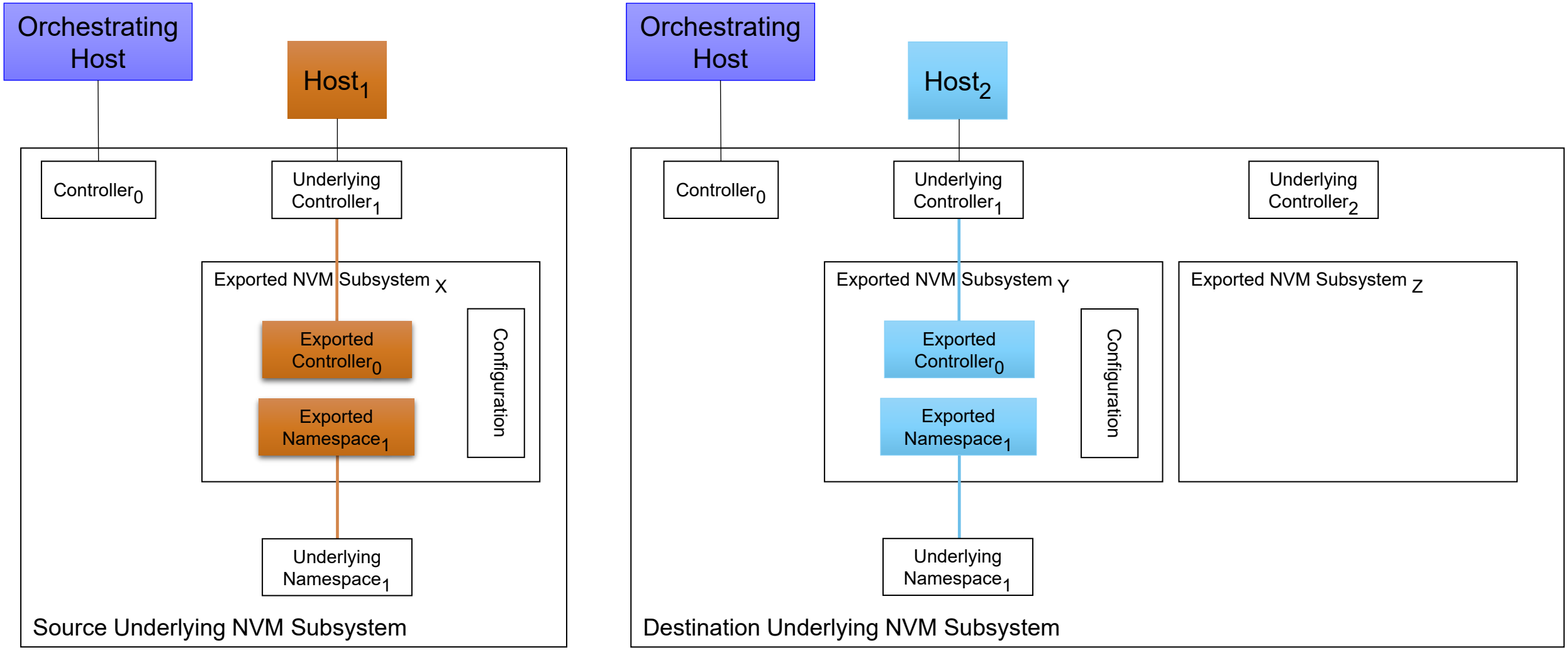
# NVMe™ Exported NVM Subsystems for PCIe®

- Expanding the NVMe-oF™ Exported NVM Subsystem concept to PCIe
- Adding an Exported Controller
- Defining Exported NVM Subsystem Template that specifies:
  - Static Supported capabilities
  - Configuration data format
    - Selectable Supported capabilities
    - Unique identifiers
    - Etc.
  - Migration data format



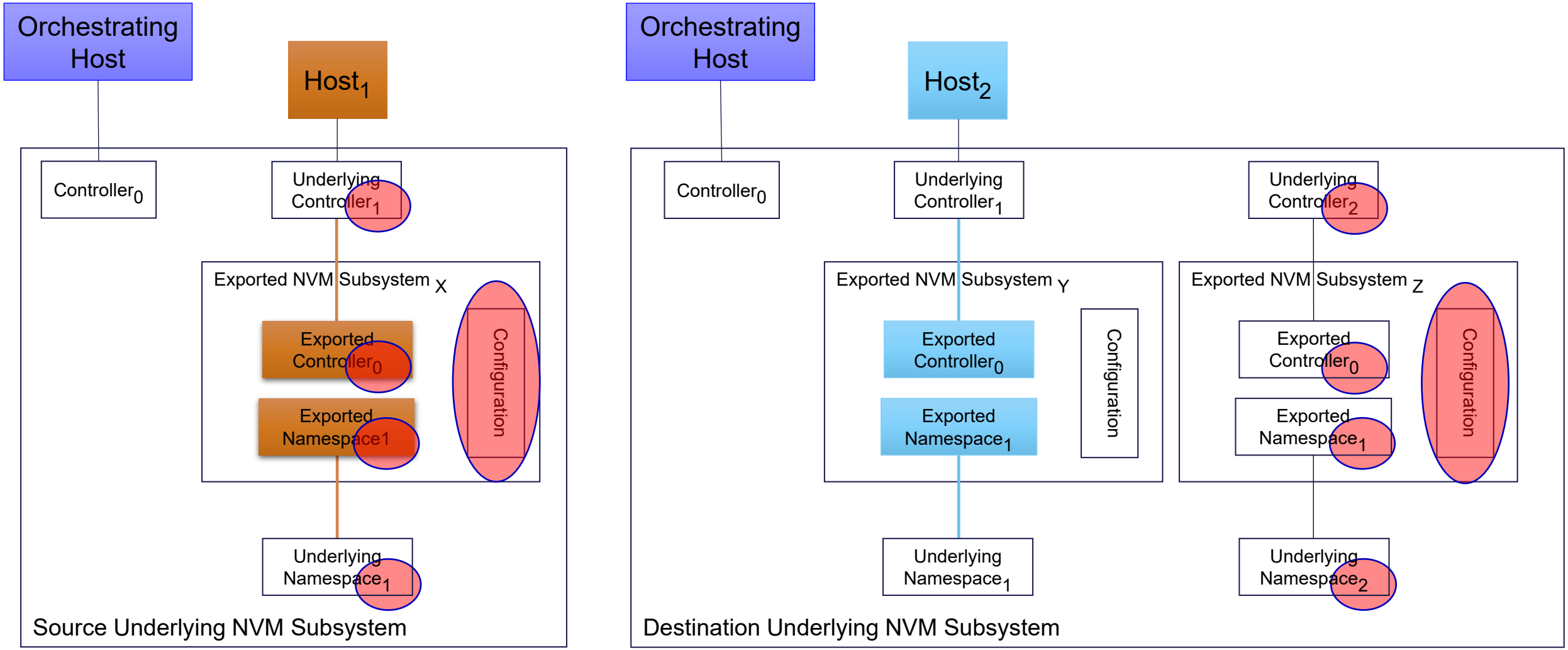

 

# NVMe™ Exported NVM Subsystems Migration



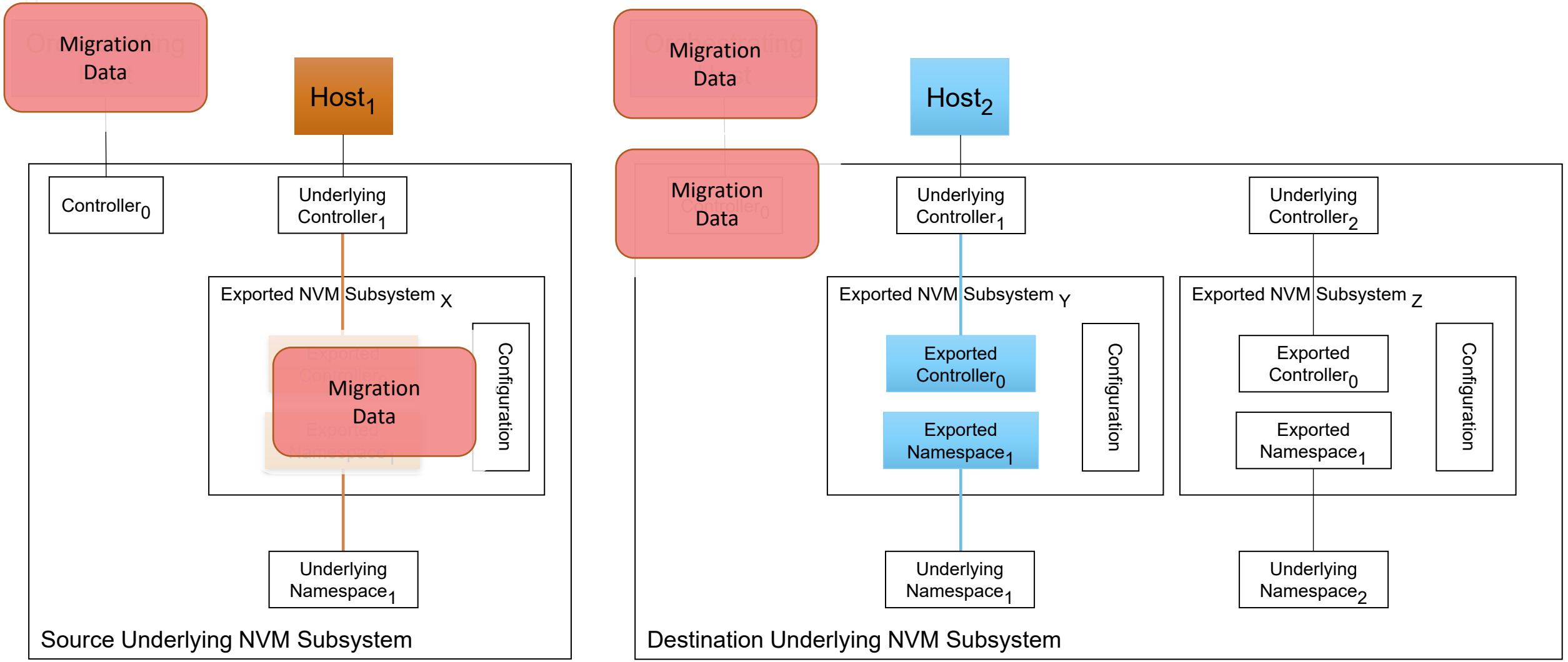

 

# NVMe™ Exported NVM Subsystems Migration



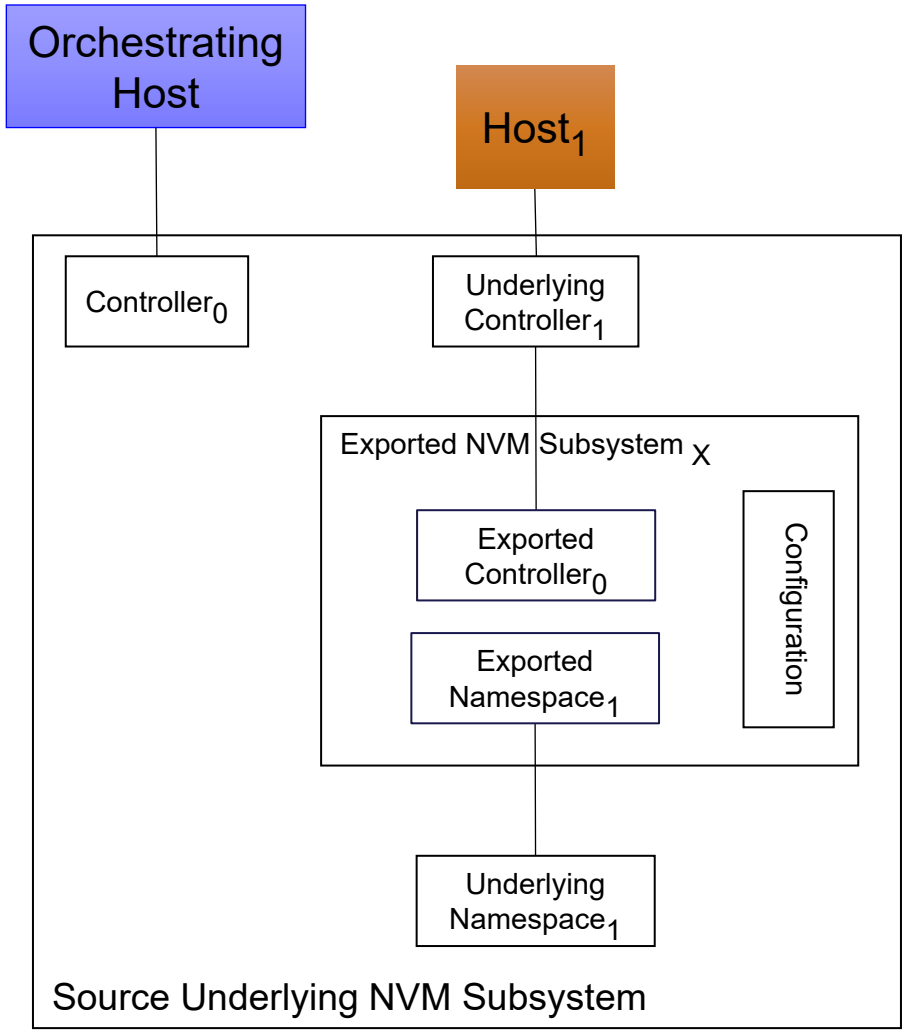

 

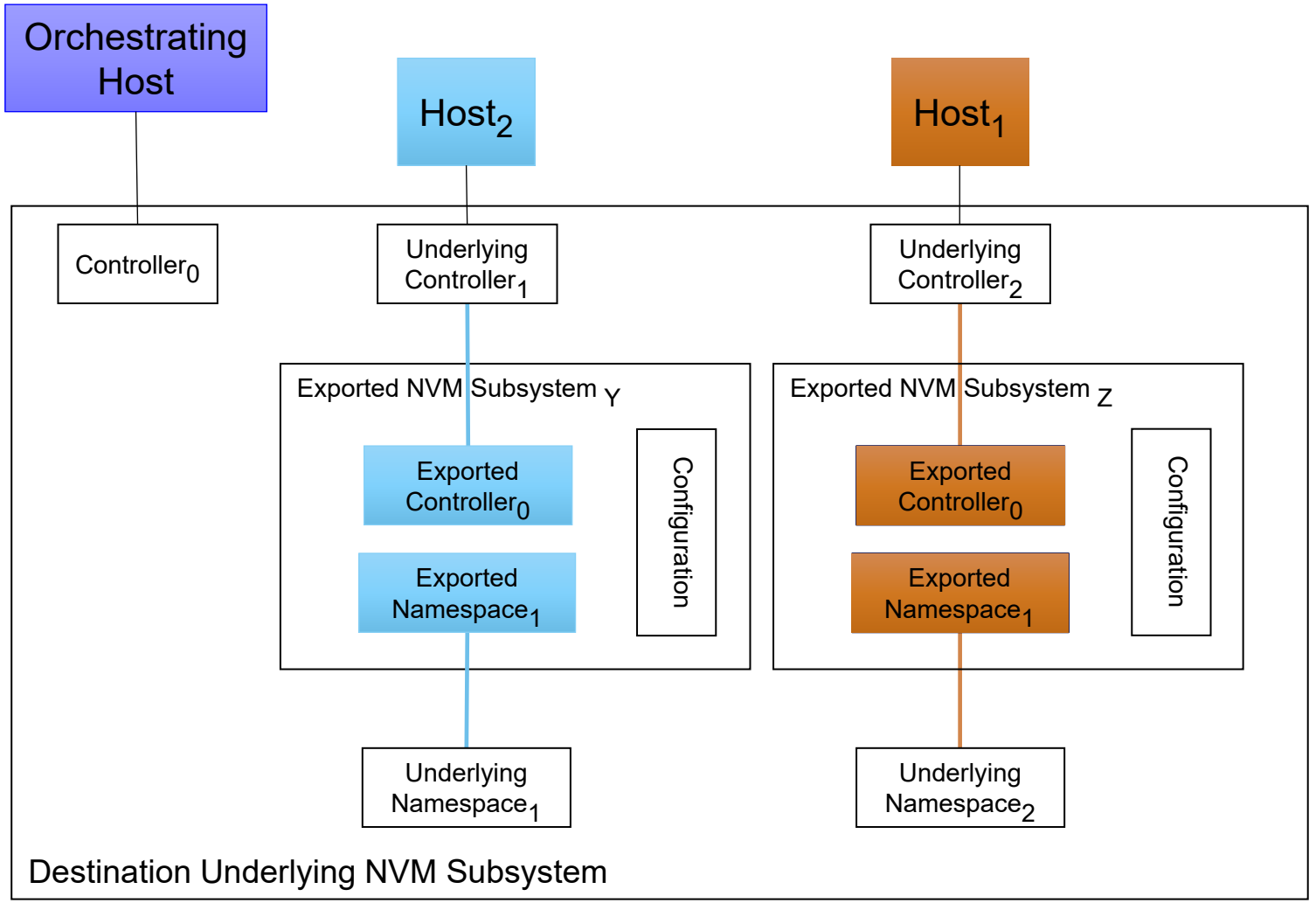
# NVMe™ Exported NVM Subsystems Migration



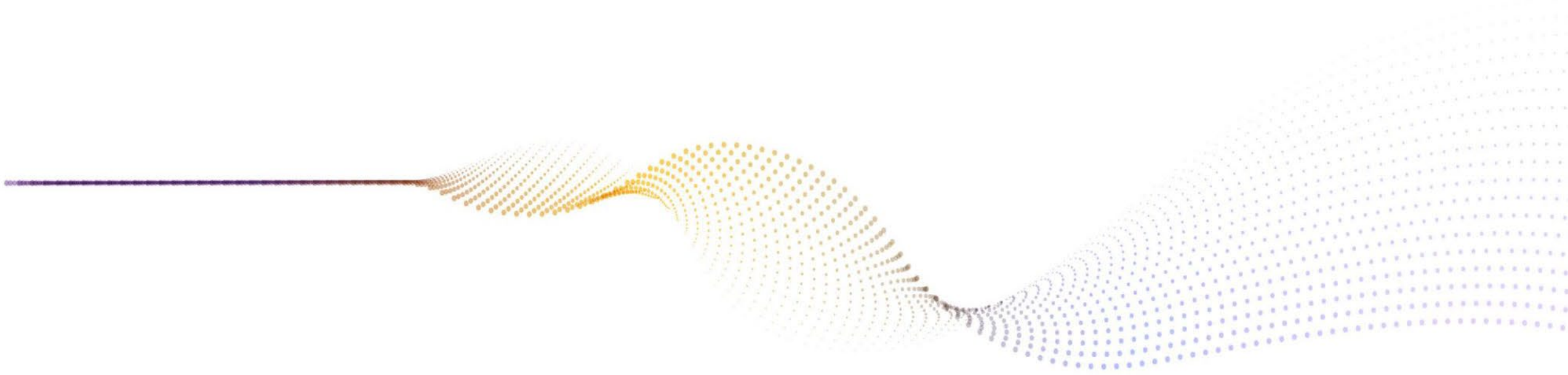

 

# NVMe™ Exported NVM Subsystems Migration





# Thank you for attending!

Please remember to rate this session. You get access the presentations at
http://sniadeveloper.org/conference