NVMe MR-IOV Solution

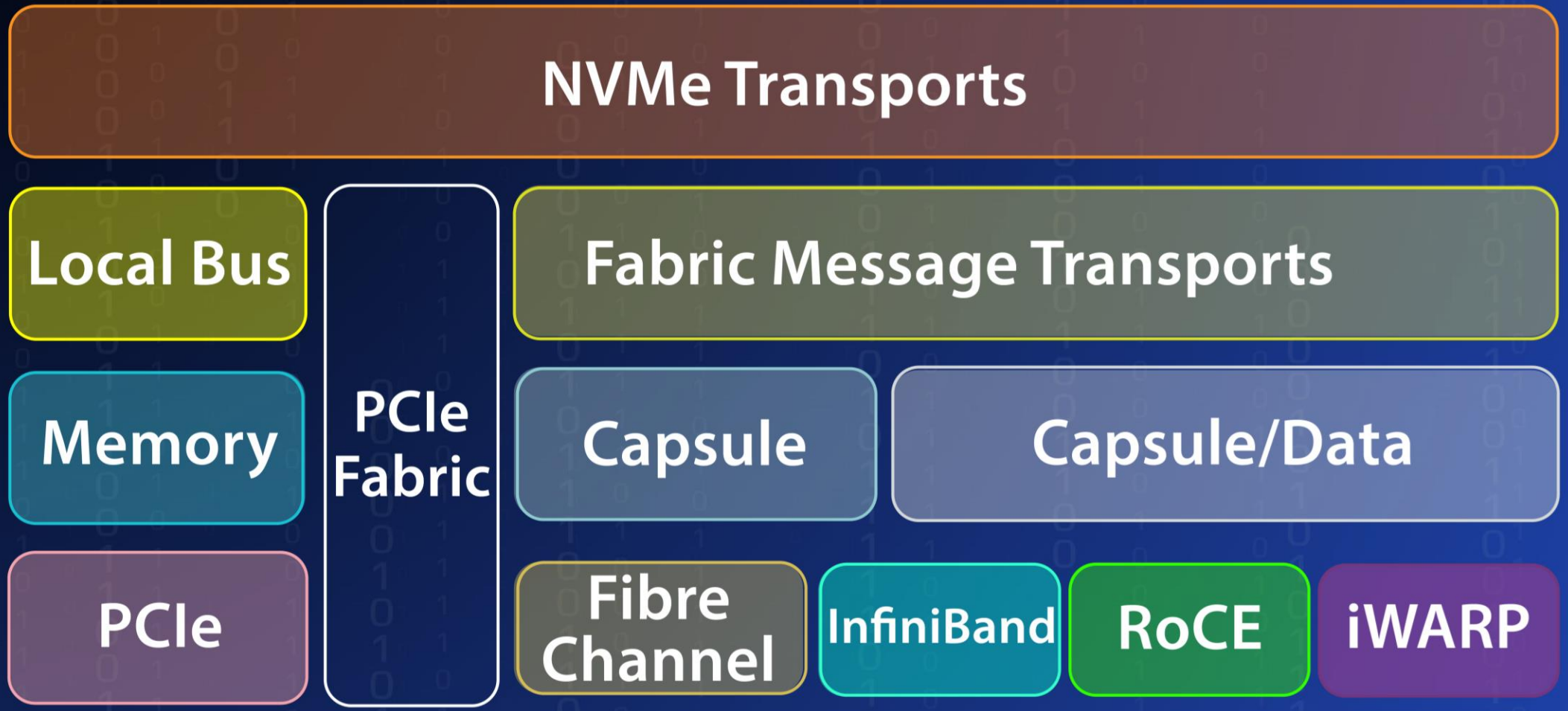
H3 Platform
Brian Pan

Agenda

- NVMe-oF Architecture
- NVMe over PCIe Fabric
- System Architecture
- Sharing NVMe SSD in Fabrics
- Performance Testing Results
- Key Benefits
- Implementation Challenges

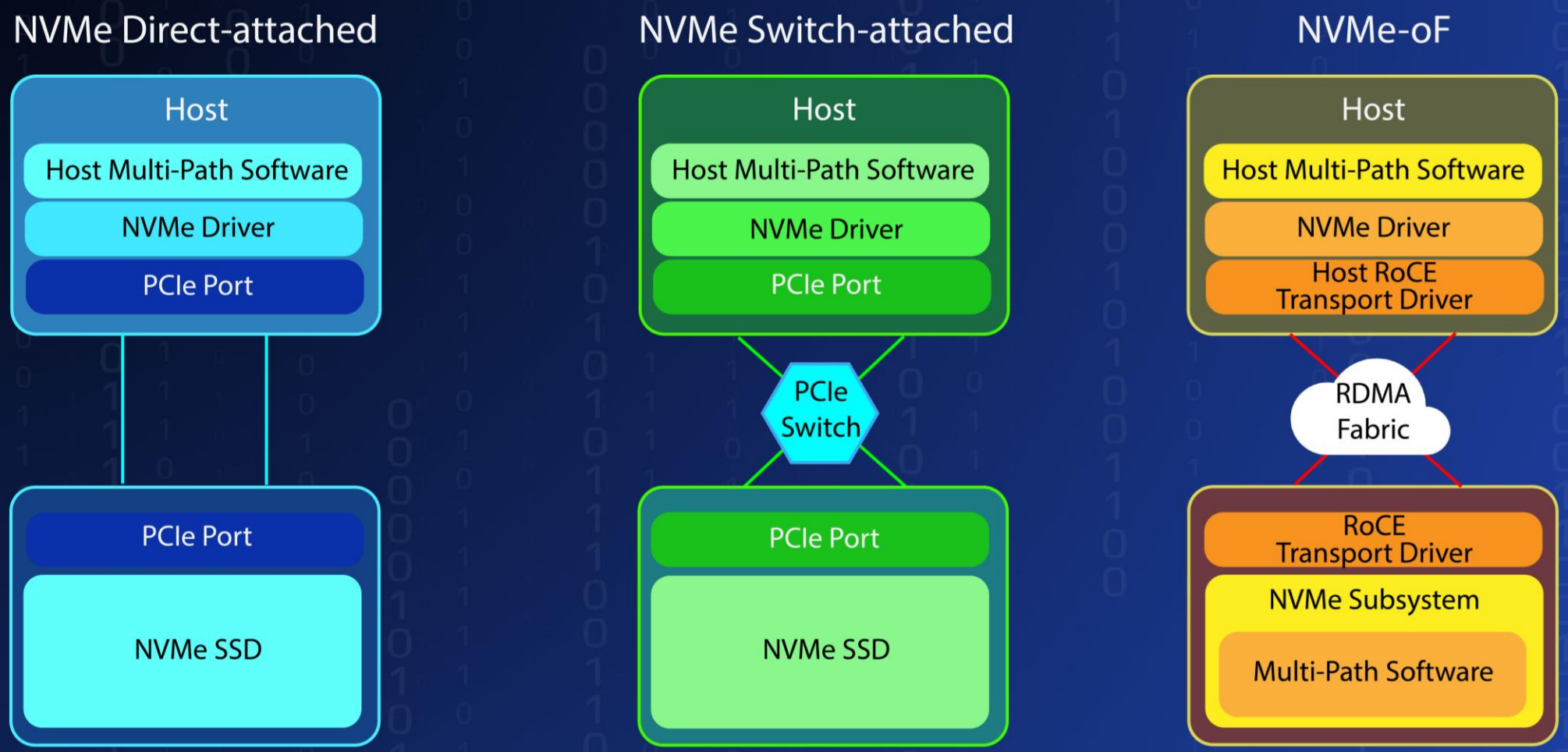
Flash Memory Summit

NVMe Transport

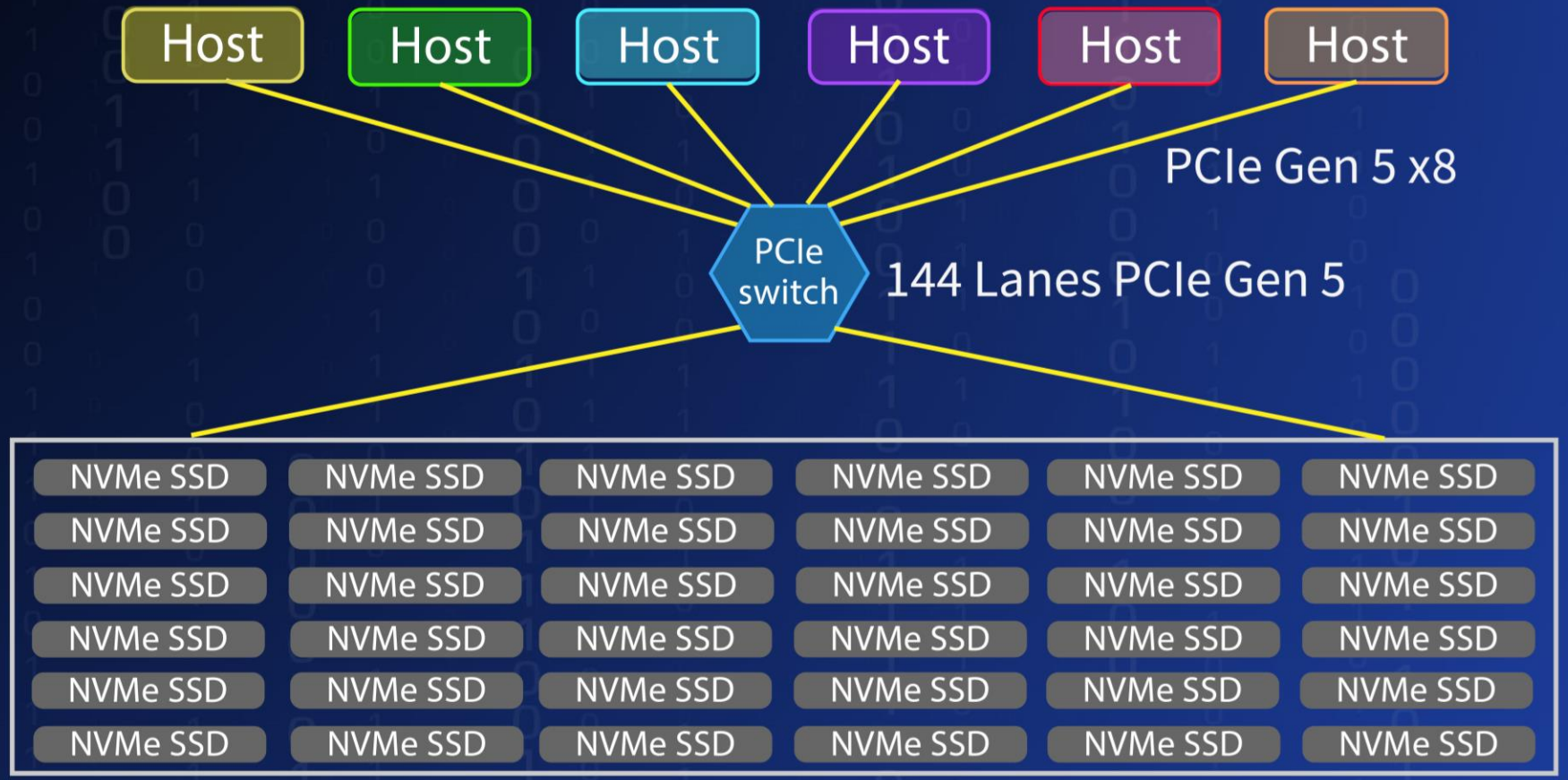


From SNIA website

Architecture Comparisons



NVMe over PCIe Fabric



From SNIA website

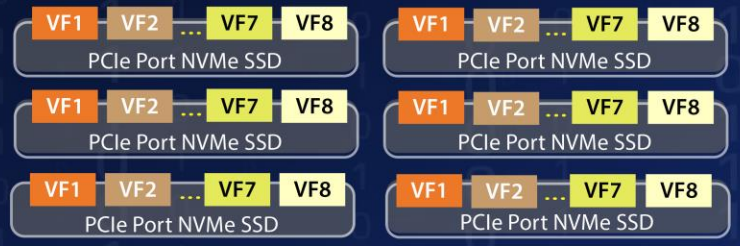
Flash Memory Summit

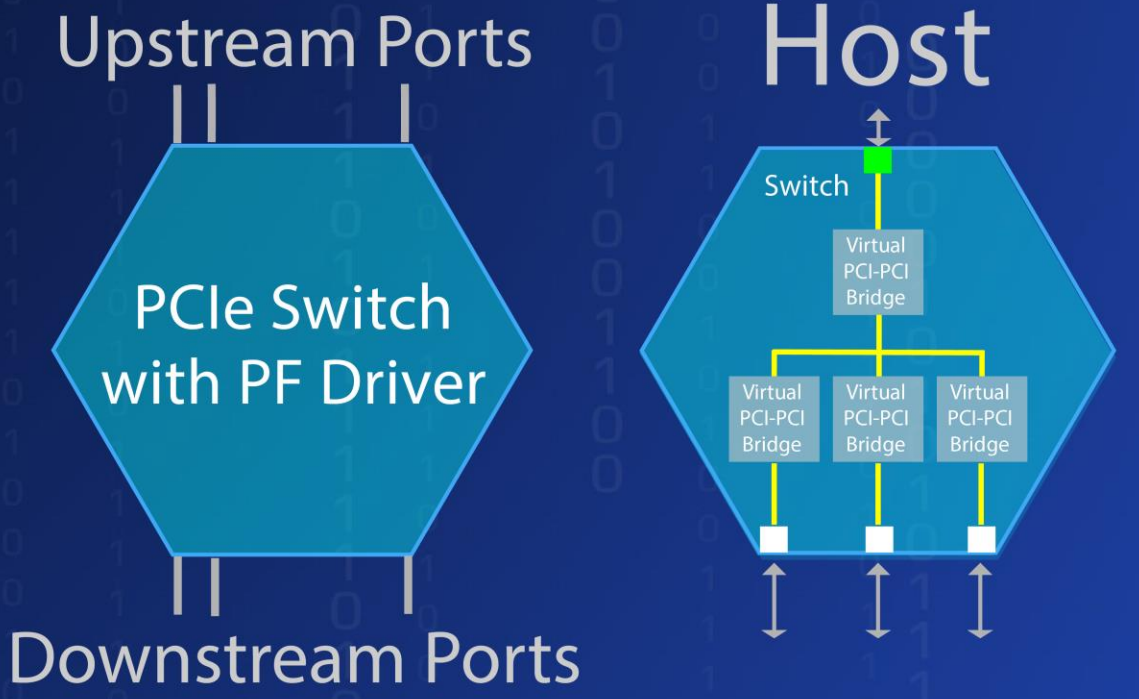
Host Emulation and NVMe Configuration by PCIe Switch

Create/ Delete namespace

Assign VFs to hosts

Use IO enabler as virtual devices

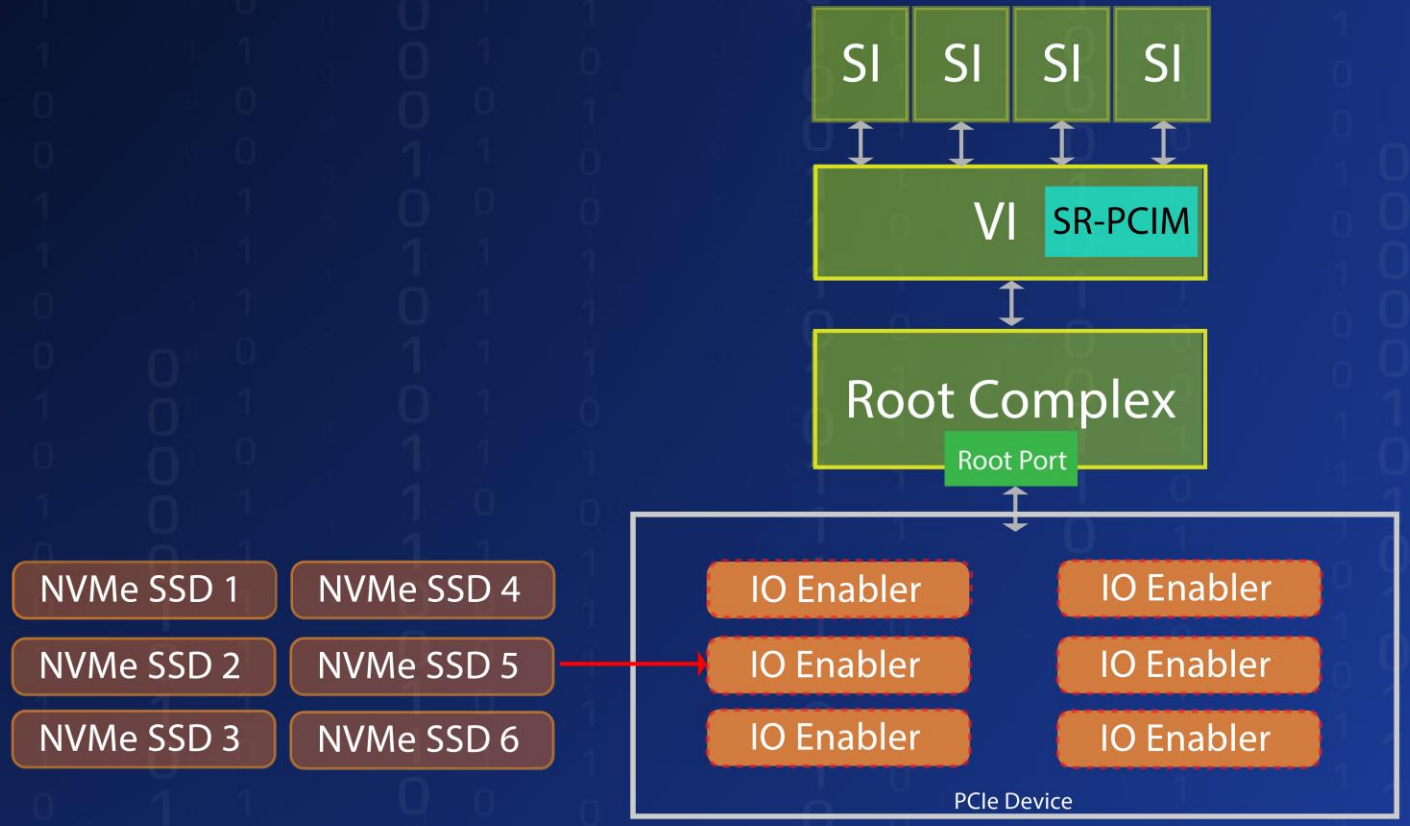




Host View of PCIe Ports

IO enablers are virtual PCIe devices

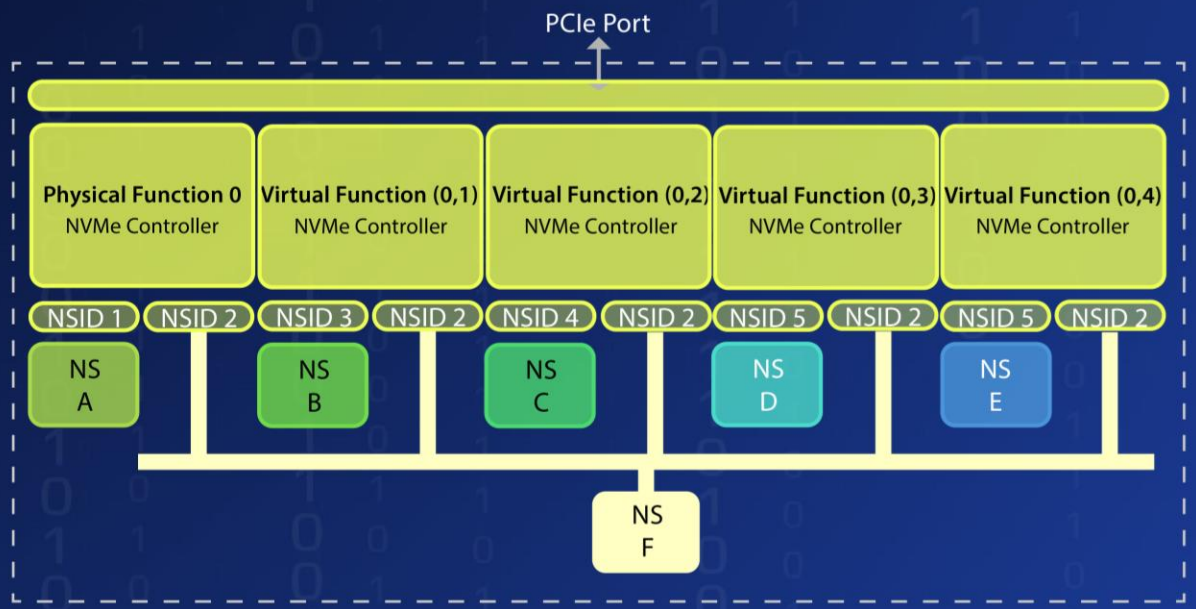
When VFs are assigned to hosts, the IO enabler will be replaced by the VF (NVMe controller)



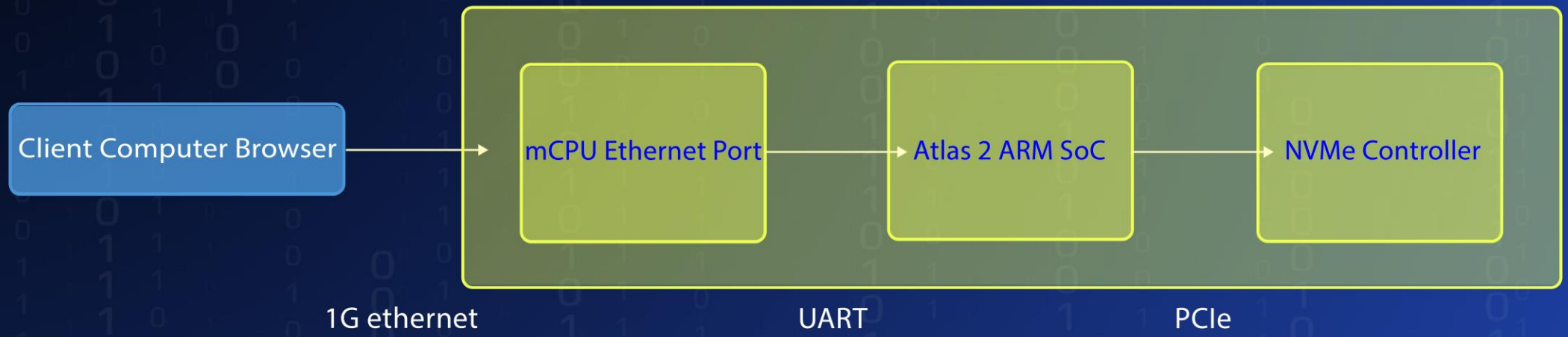
NVMe SSD Configuration

The PF driver on the PCIe switch will

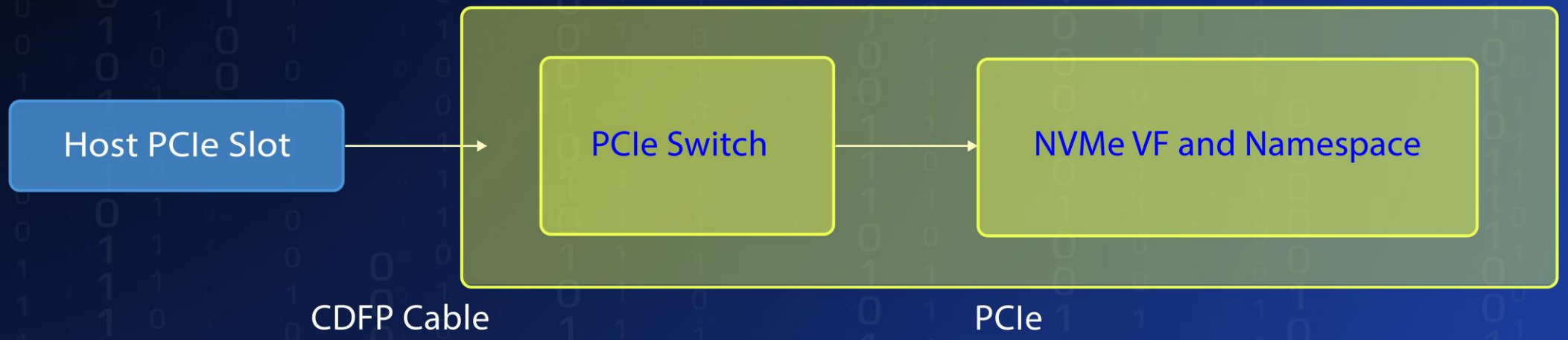
1. Create/ delete the namespace.
2. Set up the namespace as private or shared namespace.
3. Enable the virtual functions of the NVMe SSD.
4. Assign the VFs to connected hosts



Management Path through Ethernet



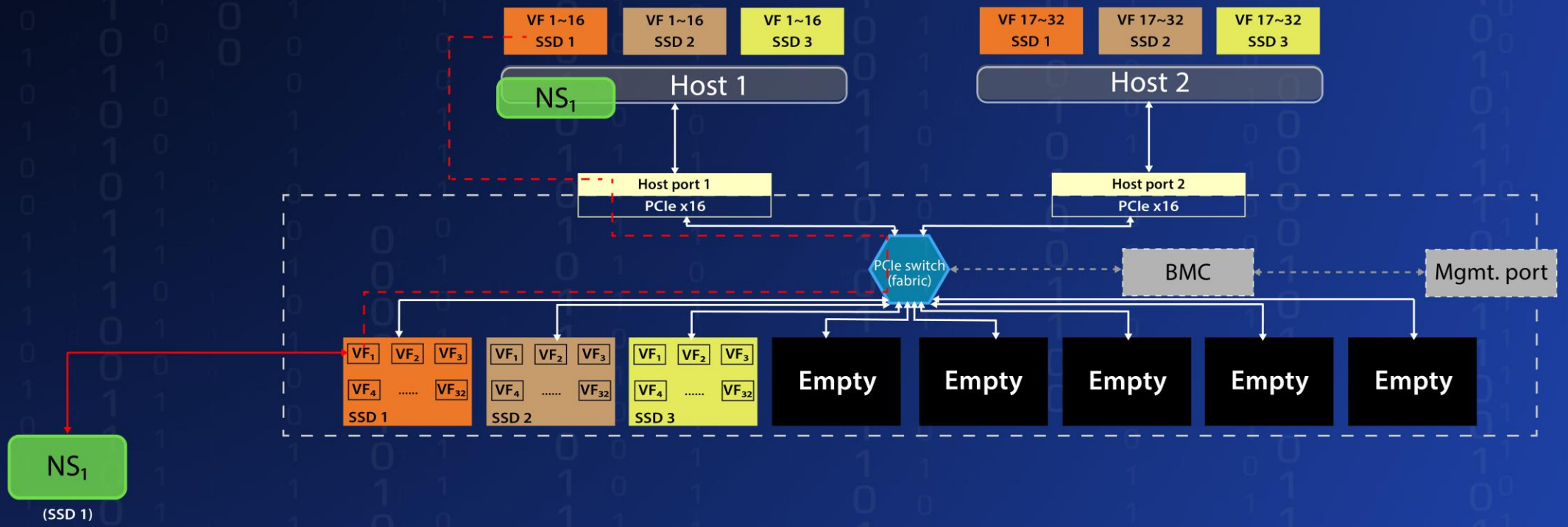
Data Path Through PCIe



System Architecture of NVMe MR-IOV

VF assignment

Namespace attached to VF



Key Benefits of MR-IOV Solution

NVMe over PCIe fabric allows NVMe-based storage to be shared among multiple servers (server cluster).

Benefits of NVMe MR-IOV Include

- Extremely Low Latency (Sub Microsecond)
- Highest Bandwidth (PCIe Gen 5 x4)
- Flexibility of NVMe Sharing in a Cluster (Namespace Sharing)
- Higher NVMe Utilization

4K RR Performance @4K, QD=256

Flash Memory Summit

4K Random Read (3 Hosts are accessing Kioxia CM7 concurrently)

	Host_1 (x8) AMD 7302 16-Core RHEL 8.5	Host_2 (x16) AMD 7452 32-Core Ubuntu 20.04.3 LTS	Host_3 (x8) AMD 7302 16-Core Ubuntu 20.04.1 LTS	Remark
1x SSD Kioxia CM7	810K	810K	800K	Spec: 2,450K
2x SSDs Kioxia CM7	1,530K	1,530K	1,580K	Spec: 4,900K

Read/ Write Latency from VM

4K Random Read/Write from/to Kioxia CM7 SSD by 3 VMs in one host

	VM_1 (8 vCPUs)	VM_2 (8 vCPUs)	VM_3 (8 vCPUs)
4K Random Read	avg=65.06	avg=65.37	avg=65.13
4K Random Write	avg=10.6	avg=10.3	avg=10.1

The specification of read latency is 70 µs, and write latency 10 µs.

Flash Memory Summit

Implementation Challenges

- **Host bus number limitation**
  - All the VFs will be enumerated as a single device.
  - The bus number of the PCIe root port is not enough.
- **NVMe SSD with SR-IOV capability**
  - The NVMe SSD should be with the standard NVMe SR-IOV capability.
  - The enablement of SSD SR-IOV is based on the NVMe SSD implementation.
- **PCIe switch enumeration capability**
  - The PCIe switch does not have enough memory to enumerate all the VFs of the NVMe SSDs.

Flash Memory Summit