NVMe virtualization ideas for Machines on Cloud

Sangeeth Keeriyadath
IBM

k.sangeeth@in.ibm.com

https://www.linkedin.com/in/sangeek

@ksangeek

day2dayunix.sangeek.com

Disclaimer

All of the contents of this presentation are based on my understanding of NVM Express® and my academic interests to explore the possibilities of virtualizing NVMe. The contents in this presentation do not necessarily represent IBM's positions, strategies or opinions.

Topics

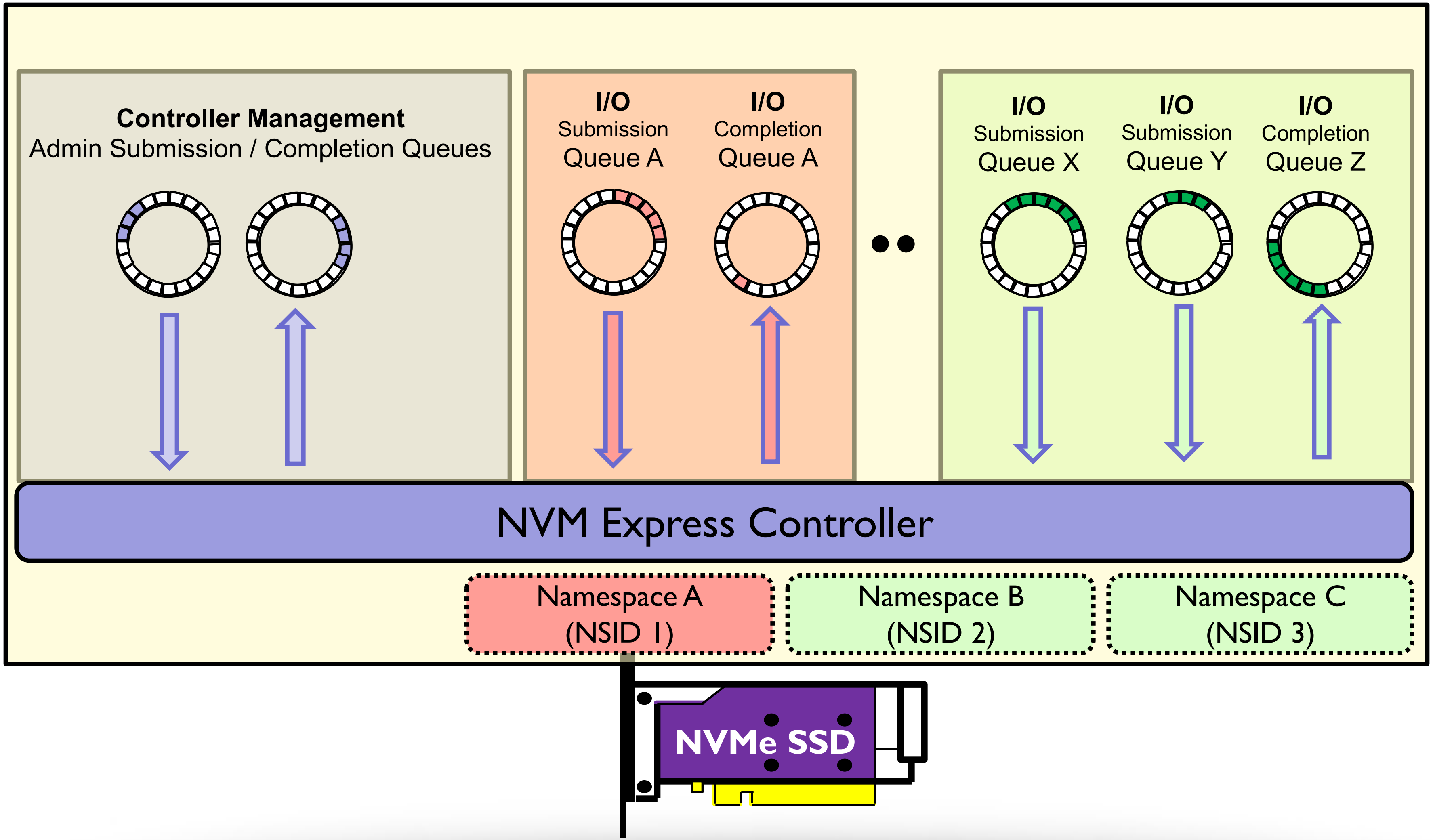
- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

NVMe primer [1] [2]

NVM Express® (NVMe) is an inherently **parallel** and **high-performing** interface and command set designed for **non-volatile memory based storage**

- The interface provides optimized command submission and completion paths resulting in :
  - Lower latency
  - Increased bandwidth
- Simple command set
- Support for parallel operation (no lock contention) by supporting up to 65,535 I/O Queues with up to 64K outstanding commands per I/O Queue
- Suitable for complementing the benefits of multi-core CPU systems and application parallelism - Command Submission/Completion queue (SQ/CQ) bound to core, process or thread

NVMe view [3]



Controller Management

Admin Submission / Completion Queues

I/O Submission Queue A

I/O Completion Queue A

I/O Submission Queue X

I/O Submission Queue Y

I/O Completion Queue Z

NVM Express Controller

Namespace A (NSID 1)

Namespace B (NSID 2)

Namespace C (NSID 3)

NVMe benefits

- Storage stack reduction (lower CPU utilization, 2 Register writes for command submit/complete)
- Designed considering next generation NVM based devices (65,535 I/O Queues)
- Elastic buffer of big size (64K commands)
- Bind interrupts to CPUs (Support For Multiple MSI-X Interrupts)
- Interconnect choices :
  - NVMe over PCIe
  - NVMe over Fabric (RDMA, Fibre Channel)

Topics

- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

Why virtualize ?

- ❐ Key component of cloud computing
- ❐ Optimizes utilization of physical resources
- ❐ Improves storage utilization
- ❐ Simplified storage management - provides a simple and consistent interface to complex functions
- ❐ Easier high availability(HA) solutions

Learnings : IBM PowerVM success story

IBM® PowerVM® provides the industrial-strength virtualization solution for IBM Power Systems™ servers and blades that run **IBM AIX®**, **IBM i** and **Linux** workloads.
Virtual I/O Server(VIOS) facilitates the sharing of physical I/O resources among guest Virtual Machines(VM). **[4]**

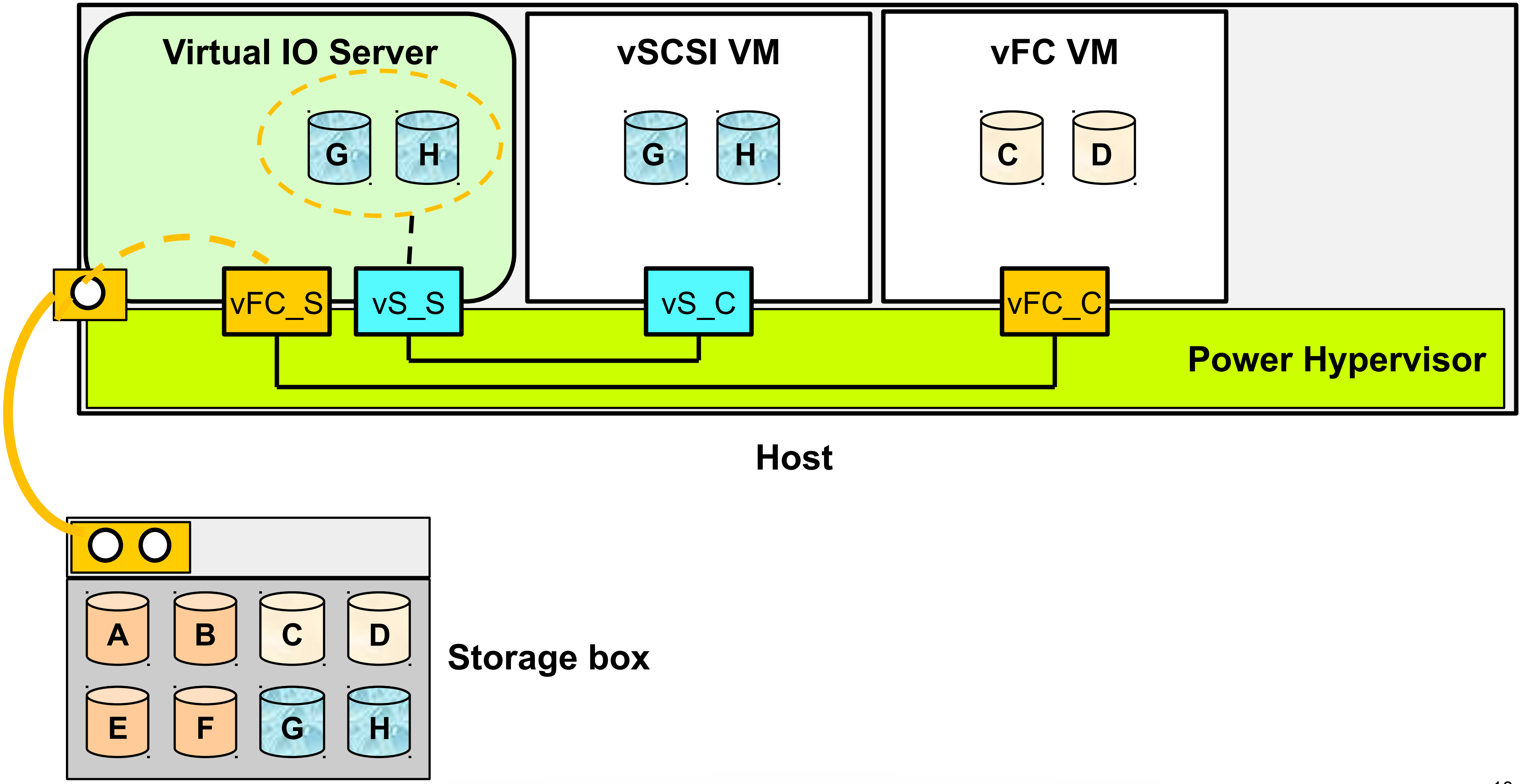
- **NPIV [5]**

N-Port ID Virtualization(NPIV) is a T11 standard technology for virtualization of Fibre Channel networks. Enables connection of multiple VMs to one physical port of a Fibre Channel adapter. VIOS facilitates storage adapter sharing.

- **vSCSI [6]**

Using virtual SCSI, VM can share disk storage, tape or optical devices that are assigned to the VIOS. VIOS does the storage virtualization, performs SCSI emulation and acts as SCSI target.

PowerVM storage virtualization view



Why virtualize NVMe ?

- Limited PCIe slots and need to share the benefits amongst VMs (high VM density)
- Use cases :
  - Server side caching of VM data
  - VM boot disk

NVMe being highly scalable is well suited for various virtualization exploitations

Topics

- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

NVMe virtualization approaches

- ❒ Implementing SCSI to NVMe translation layer on the Hypervisor. (Blind Mode).
- ❒ Pure virtual NVMe stack by distributing I/O queues amongst hosted VMs. (Virtual Mode).
- ❒ SR-IOV based NVMe controllers per virtual functions (Physical mode).

Topics

- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

Blind Mode : SCSI to NVMe translation

- **Motivation**
  - Lot of storage stack ecosystem built around SCSI architecture model, protocol and interfaces
  - Preserve software infrastructure investments
- **Method**
  - Implement a "SCSI to NVMe translator" layer built logically below the operating system SCSI storage stack and above the NVM Express driver

- Detailed in "NVM Express: SCSI Translation Reference" document [7]
  - SCSI primary/block commands to NVMe command mapping
  - Common SCSI Field translations e.g. "PRODUCT IDENTIFICATION" field would be translated to first 16 bytes of the Model Number (MN) field within the "Identify Controller Data"
  - NVMe "Status Code" to SCSI "Status Code, Sense Key, Additional Sense Key"

OS Storage stack : SCSI to NVMe translated

INQUIRY
REPORT LUNS
READ CAPACITY

READ(6), READ(10),
READ(12), READ(16)

GOOD / NO SENSE

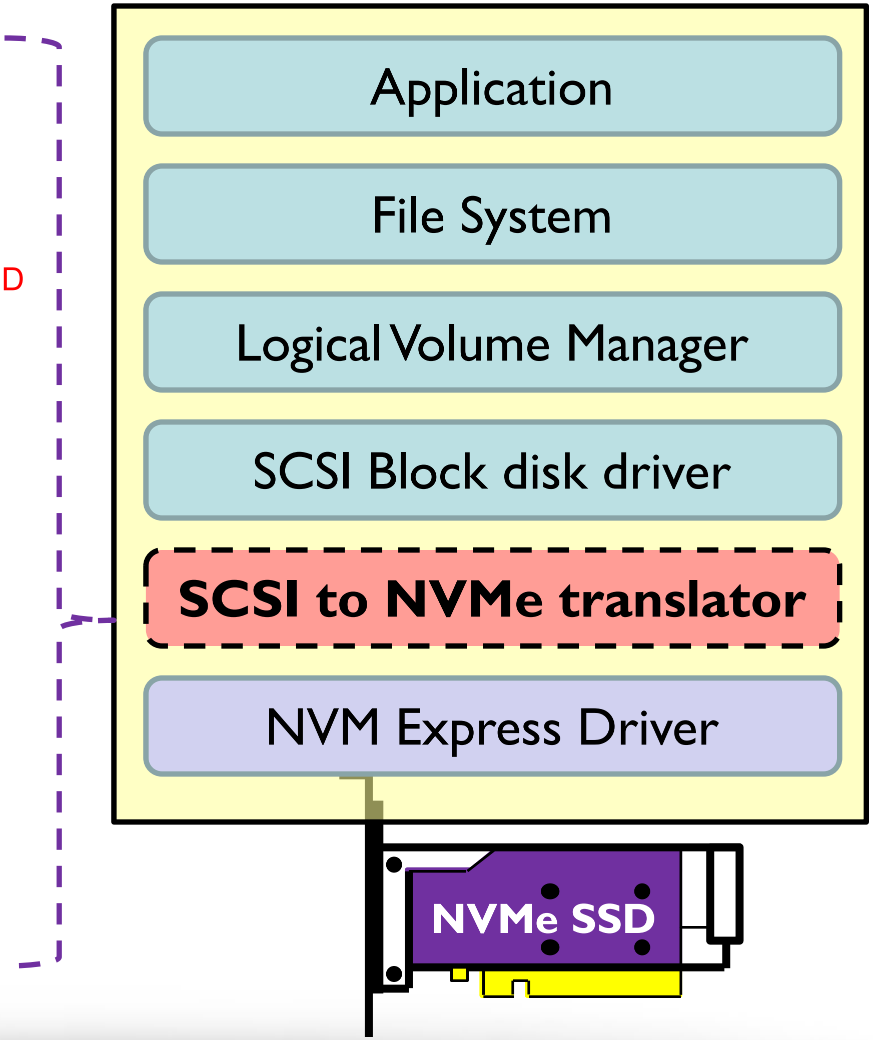
CHECK CONDITION / ILLEGAL REQUEST / INVALID COMMAND OPERATION CODE

Identify command

Read

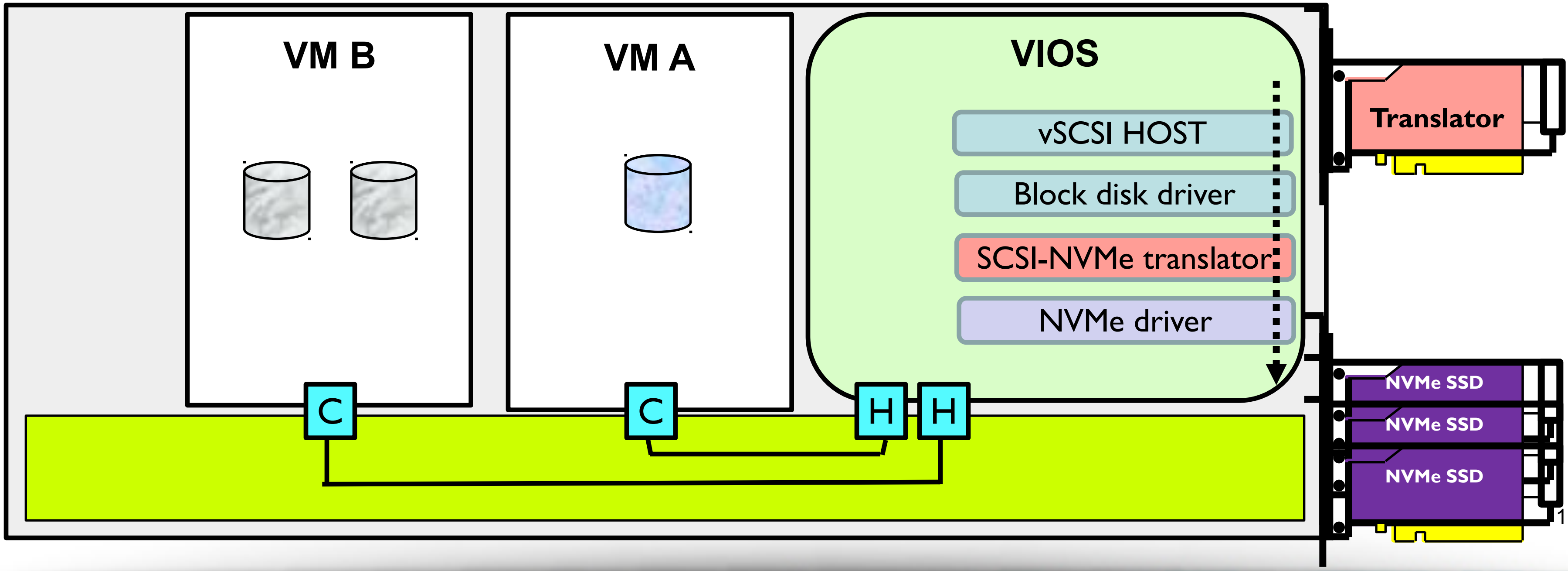
Success Completion

Invalid Command Opcode



Blind Mode View : NVMe unaware VM

- ❐ SCSI to NVMe translation performed on the storage hypervisor
- ❐ Virtual Machine(VM)'s Operating System(OS) storage stack is unmodified
- ❐ Performance consideration : "storage hypervisor" could use a "hardware accelerated SCSI to NVMe translator"



Topics

- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

Virtual Mode : pure NVMe virtual stack

- **Motivation**

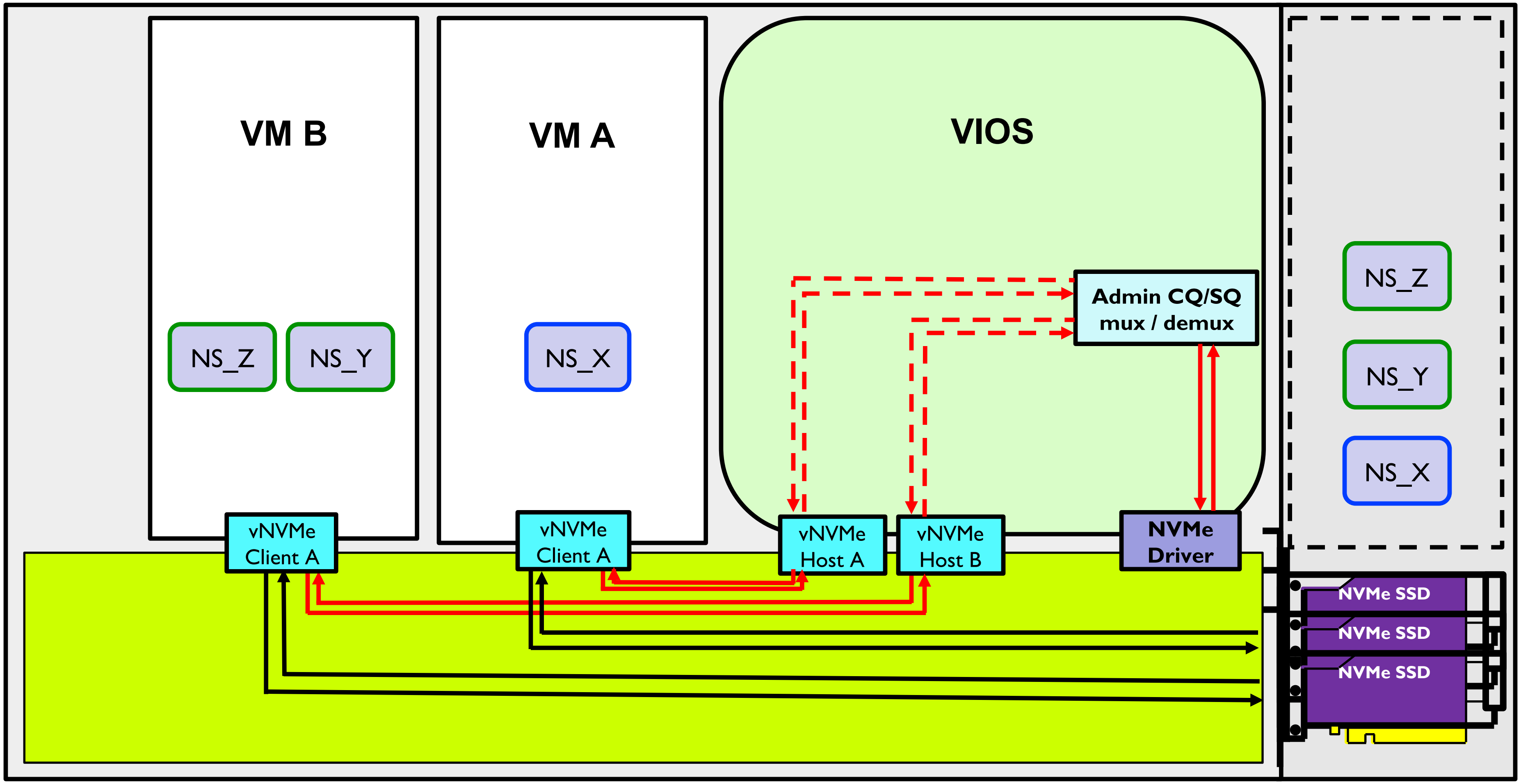
Model suitable for server side storage virtualization solutions and preserve NVMe benefits

- Share NVM storage capacity amongst Virtual Machines(VMs)
- VMs with pure NVMe stack retain the various NVMe performance characteristics :
  - Low latency
  - Parallelism
- Easier on-demand scaling(grow/shrink) implementation
- Suitable for future "NVMe over Fabrics" implementations in data-centers

Virtual Mode : pure NVMe virtual stack

- **Method**
  - “Storage hypervisor(VIOS)” owns the physical NVMe controller and facilitates device sharing amongst guest VMs
  - VIOS governs the admin SQ/CQ of the NVMe device.
  - VIOS manages the namespaces (storage partitions) for use of the VMs.
  - VIOS manages a light-weight “virtual NVMe host adapter” i.e. vNVMeHost
  - VM will see a NVMe device off a “virtual NVMe client driver” i.e. vNVMeClient
  - Hypervisor manages the pairing between the vNVMeClient and vNVMeHost adapters
  - Storage stack on the VM Operating System above the vNVMeClient driver would be agnostic to this virtualization and gets to use NVMe disks.

Virtual Mode View : pure NVMe stack on VM



Topics

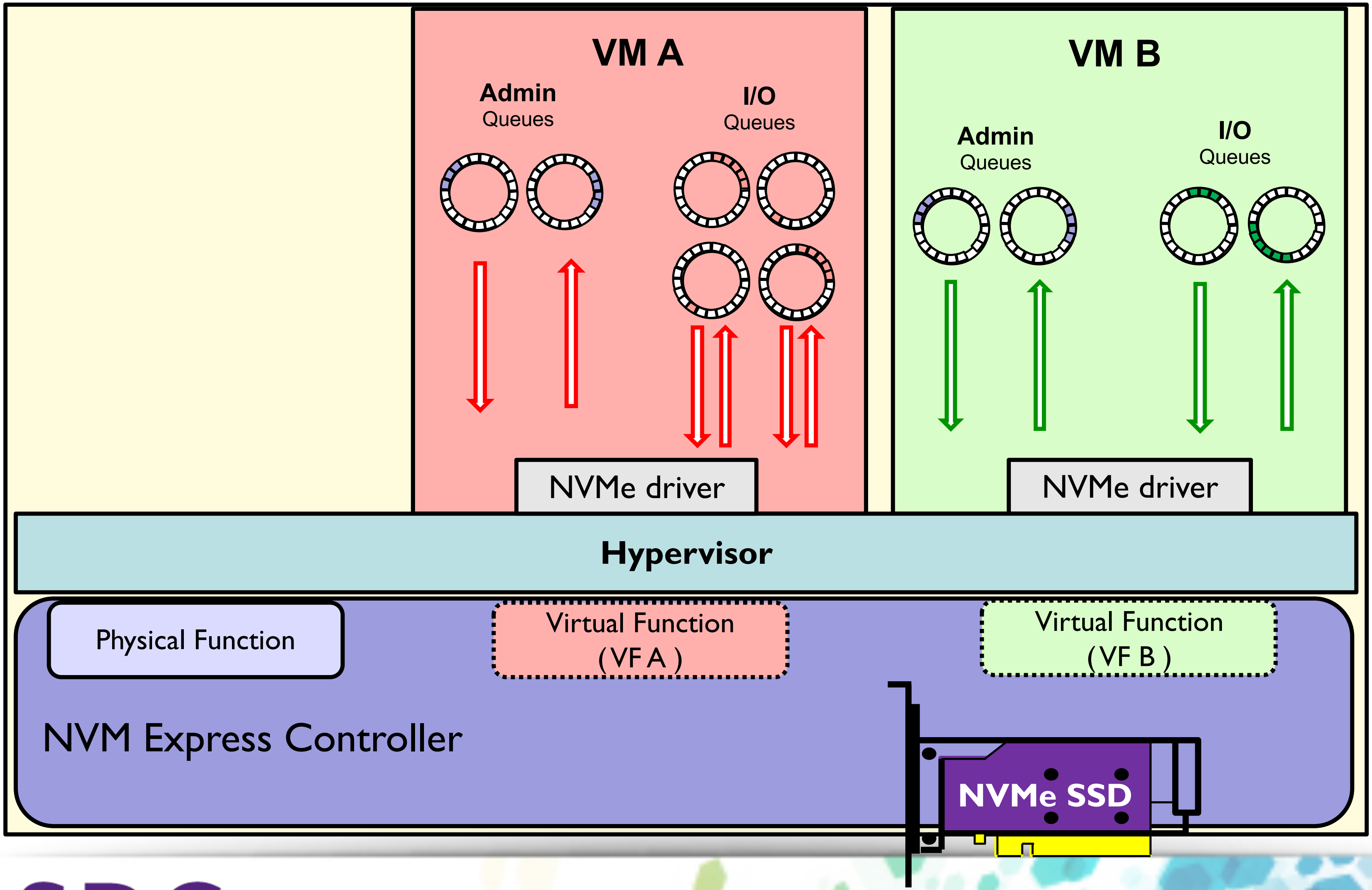
- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

Physical Mode : SR-IOV [8]

- **Motivation**
  - Single Root I/O Virtualization (SR-IOV) is a specification that allows a PCIe device to appear as multiple separate physical PCIe devices
  - Inherent QoS capabilities and configuration

- **Method**
  - Partition adapter capability logically divided into multiple separate PCI function called “virtual functions”
  - Each “virtual function” could be individually assigned to a Virtual Machine(VM)

NVMe SR-IOV view



Topics

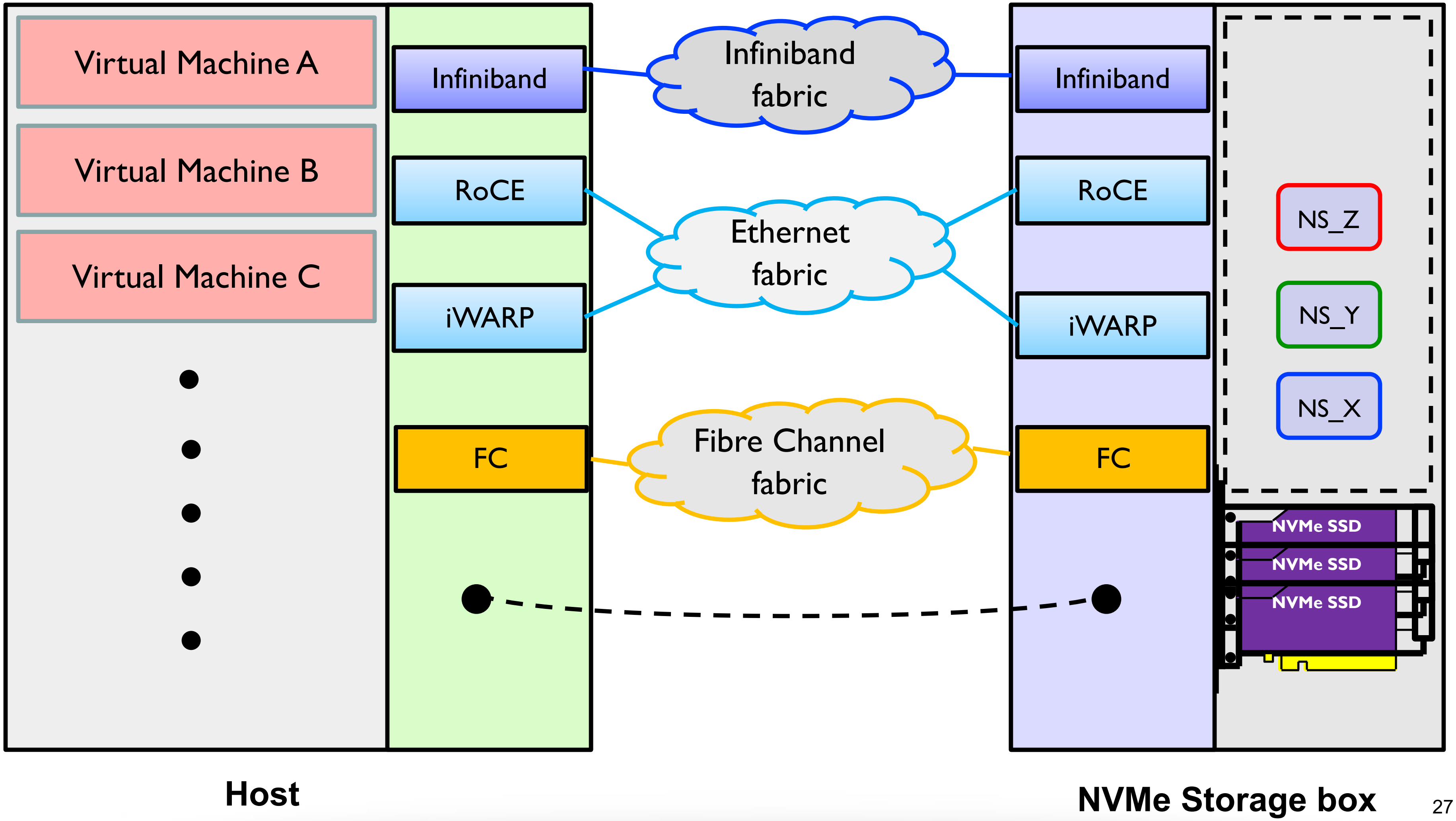
- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

NVMe over Fabrics

- Extends NVMe (beyond PCIe) onto data center fabrics such as Ethernet, Fibre Channel and InfiniBand
- Distance connectivity to storage systems with NVMe devices
- Scaling out of the NVMe devices in large solutions
- Eliminate unnecessary protocol translation

- Two types of fabric transports for NVMe are currently under development:
  - NVMe over Fabrics using RDMA (InfiniBand, RoCE and iWARP)
    - RDMA verbs
  - NVMe over Fabrics using Fibre Channel (FC-NVMe) **[10]**
    - Standard defining mapping of NVMe commands over FC
    - Backward compatible with Fibre Channel Protocol (FCP) transporting SCSI

NVMe over fabrics view



NVMe over Fabrics virtualization

❒ NVMe over Fabrics using FC (NPIV) [9]

❒ NVMe over Fabrics using RDMA

Other deliberations

- SSD endurance – Drive Writes Per Day (DWPD)
- Virtual Machine migration
- Interrupt driven model or polling
- Virtualization Support (VM_IDs in Frame)

Topics

- NVMe primer
- Past storage virtualization learnings
- Virtualization considerations
- NVMe virtualization approaches
  - Blind Mode : SCSI to NVMe translation
  - Virtual Mode : pure NVMe virtual stack
  - Physical Mode : SR-IOV
- NVMe over Fabrics virtualization
- Other deliberations

References

- [1] NVM Express™ Infrastructure - Exploring Data Center PCIe® Topologies : http://www.nvmexpress.org/wp-content/uploads/NVMe_Infrastructure_final1.pdf
- [2] A Comparison of NVMe and AHCI “Don H Walker” : https://sata-io.org/system/files/member-downloads/NVMe%20and%20AHCI_%20_long_.pdf
- [3] NVM Express Overview http://www.nvmexpress.org/nvm-express-overview/
- [4] IBM PowerVM Virtual I/O Server overview : http://www.ibm.com/support/knowledgecenter/POWER8/p8hb1/p8hb1_vios_virtualioserveroverview.htm
- [5] IBM PowerVM “Virtual Fibre Channel” : http://www.ibm.com/support/knowledgecenter/POWER8/p8hb1/p8hat_vfc.htm
- [6] IBM PowerVM “Virtual SCSI” : http://www.ibm.com/support/knowledgecenter/POWER8/p8hb1/p8hb1_vios_concepts_stor.htm
- [7] NVM Express: SCSI Translation Reference : http://www.nvmexpress.org/wp-content/uploads/NVM-Express-SCSI-Translation-Reference-1_1-Gold.pdf
- [8] I/O Virtualization in Enterprise SSDs : http://www.snia.org/sites/default/files/SDC15_presentations/virt/ZhiminDing_IO_Virtualization_eSSD.pdf
- [9] FC-NVMe NVMe over Fabrics “QLOGIC” : http://www.qlogic.com/Resources/Documents/WhitePapers/Adapters/WP_FC-NVMe.pdf
- [10] Networking Flash Storage? Fibre Channel Was Always The Answer! : http://www.snia.org/sites/default/files/DSI/2016/presentations/stor_networks/MarkJonesNetworking_Flash_Fibre_Channel_Jun16-nc.pdf

Credits [IBM Systems]

- Hemanta Dutta, AIX Storage & IO SW
- Mallesh Lepakshaiah, Virtual IO Server
- Ninad Palsule, AIX Virtual Server Storage
- Sanket Rathi, AIX Storage Device Drivers
- Sudhir Maddali, AIX Storage Device Drivers
- Venkata Anumula, AIX Storage Device Drivers

Thank You