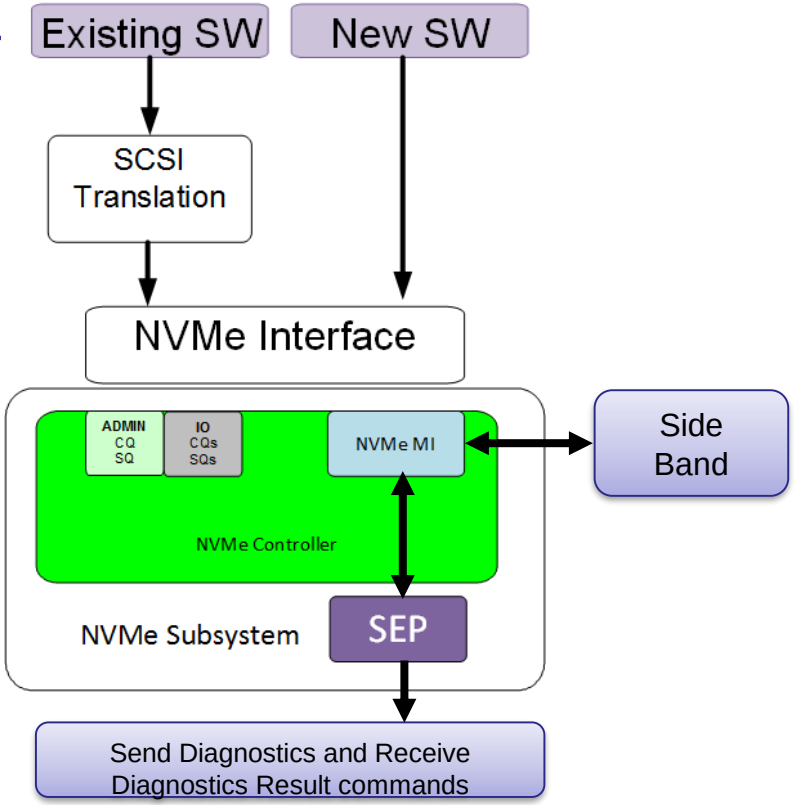
NVMe Enclosure Management and Dual Port Drive

Parag Maharana

SES over NVMe Overview

- Inband SES management via NVMe driver
- Out of band SES management via NVMeMI
- SEP can be exposed as a Namespace or as a separate PCIe function or both
- Existing applications can use SCSI interface to access SES pages based on SCSI to NVMe Translation
- New application can access directly via NVMe admin queue pair

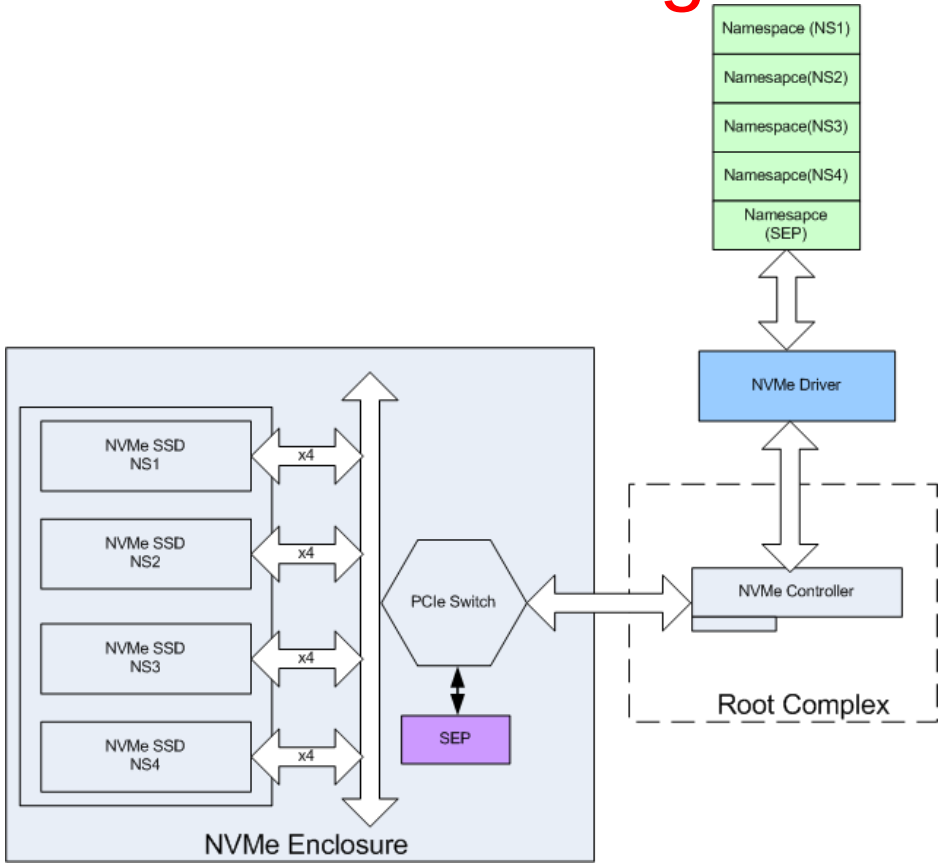


Enclosure Services Management usage model

- Main goal is to use standard SCSI Enclosure Services tools on NVMe. This will leverage investment on existing management tools for SES
- Also generic sg_ses tool in Linux that can be used to manage NVMe enclosure services, can manage NVMe enclosure
- When sg_ses open SEP namespace (e.g. /dev/nvme0ns1) then NVMe driver can translate/convert send and receive diagnostic commands to NVMe admin command that will be send to NVMe HBA or NVMe enclosure
- NVMe Admin commands are issued in-band through PCIe.
- Linux generic SES management tool will work as it is by selecting NVMe controller namespace “sg_ses /dev/nvme0ns1”

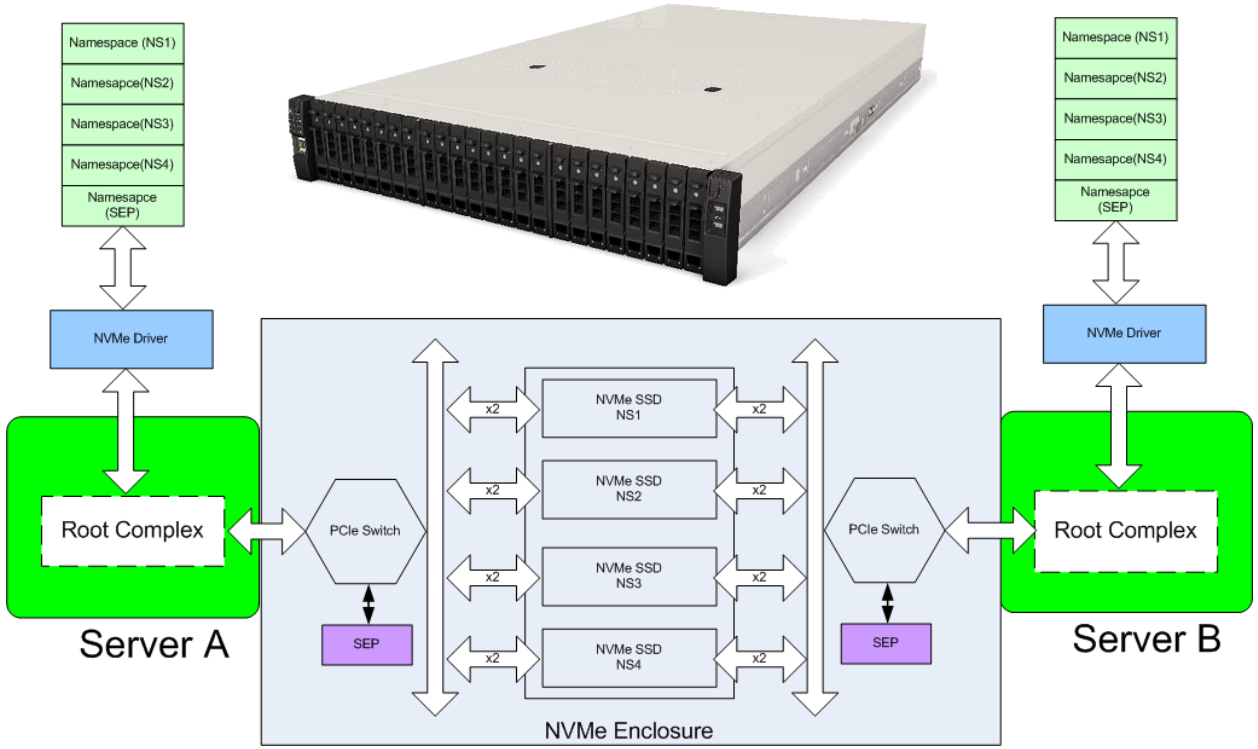
NVMe Enclosure Services Management

- SEP type device may expose as namespace or as a separate PCIe function
- Host SW can detect using Identify data
- Application can send enclosure inventory and status commands through Admin queue
- Admin queue allow tunneling of NVMeMI commands for Send/Receive SES commands
- SEP can be connected to switch through I2C, SGPIO or PCIe.
- Also it is possible to connect NVMe enclosure directly to root complex without going through any NVMe aggregator



SES over NVMe for Multi-Host Topology

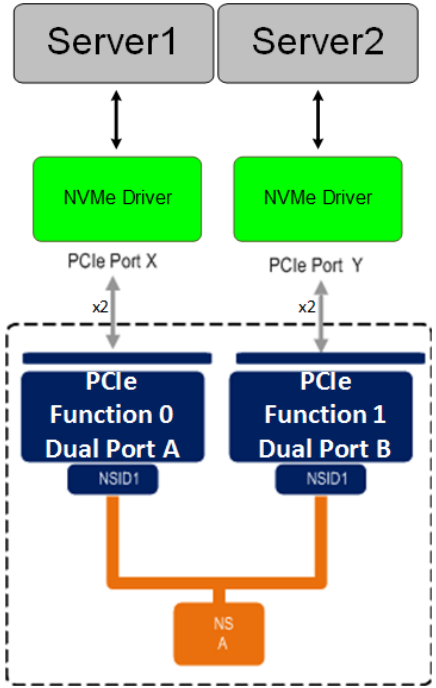
- Multi-Host Topology
- Directly connected to root complex of the servers
- Each Namespace will be expose through a separate PCIe functions
- Each namespace will be handed by separate driver instance



SEAGATE

Dual Ports NVMe Drive

- Single port U.2 NVMe drives are PCIe x4
- Dual Ports U.2 NVMe drives are each PCIe x2
- Dual port drives shared namespace across both ports
- Dual ports can have dedicated namespace per port (mainly for boot)
- Dual port drives required reservation to allow active-active support from two or more servers on shared namespace
- NVMe Subsystem reset will reset entire drive



Dual Ports NVMe SSD

NVMe Persistent Reserve and Release

- Reservation functionalities are compatible with T10 persistent reservations
- Dual port drive can support any one type of Reservation. Typical reservation type is Write Exclusive with Registrant
- Dual port drive need to support 4 PR calls i.e. register, acquire, release and report
- The reservations and registrations can be persisted across all type of resets including NVM Subsystem Resets. This can be enabled via NVMe PTPLS Register
- Reservation can only be cleared by preempt or preempt and abort commands. Also drive can updates PR log pages and AENs
- Dual port drive should handle Host Identifier as well as change of Host Identifier (HOSTID) in case of server replaced or drive swapped

Reservation Type	Reservation Holder		Registrant		Non-Registrant		Reservation Holder Definition
	Read	Write	Read	Write	Read	Write	
Write Exclusive	Y	Y	Y	N	Y	N	One reservation holder
Exclusive Access	Y	Y	N	N	N	N	One reservation holder
Write Exclusive – Registrants only	Y	Y	Y	Y	Y	N	One reservation holder
Exclusive Access – Registrants only	Y	Y	Y	Y	Y	N	One reservation holder
Write Exclusive – All Registrants	Y	Y	Y	Y	Y	N	all Registrants are Reservation Holder
Exclusive Access – All Registrants	Y	Y	Y	Y	N	N	all Registrants are Reservation Holder

NVM I/O Command	Operation/Action
Reservation Register	1. Register a reservation key 2. Unregister a reservation key 3. Replace a reservation key
Reservation Acquire	1. Acquire a reservation on a namespace 2. Preempt reservation held on a namespace 3. Preempt and abort a reservation held on a namespace
Reservation Release	1. Release a reservation held on a namespace 2. Clear a reservation held on a namespace
Reservation Report	1. Retrieve reservation status data structure a. Type of reservation held on the namespace (if any) b. Persist through power loss state c. Reservation status, Host ID, reservation key for each

Conclusion

- NVMe Dual port drives fits well in NVMe enclosures
- NVMe enclosures can direct connect to server using PCIe
- Also NVMe enclosures can work well with fabrics transport
- NVMe enclosure will provide a fault tolerant infrastructure using redundant path using Dual port drives
- Intelligence NVMe enclosures can function as NVMe appliances or AFA
- SES over NVMe will provide a common/standard mechanism to control all non NVMe devices in enclosure (e.g. Fans, Power supplies, LED, etc)

Thank You! Questions?

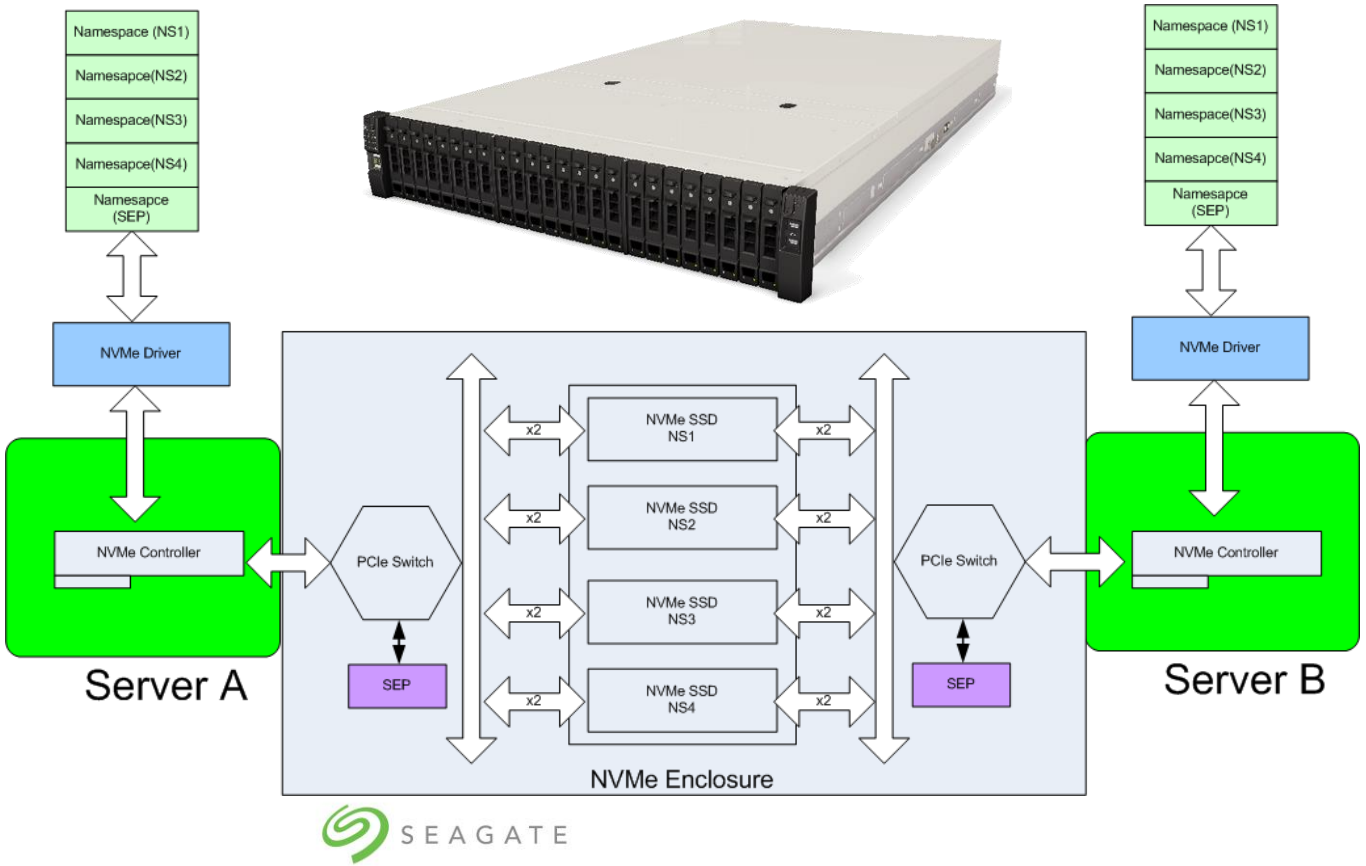
Visit Seagate Booth #505

Learn about Seagate's ever-expanding portfolio of SSDs, Flash solutions and system level products for every segment

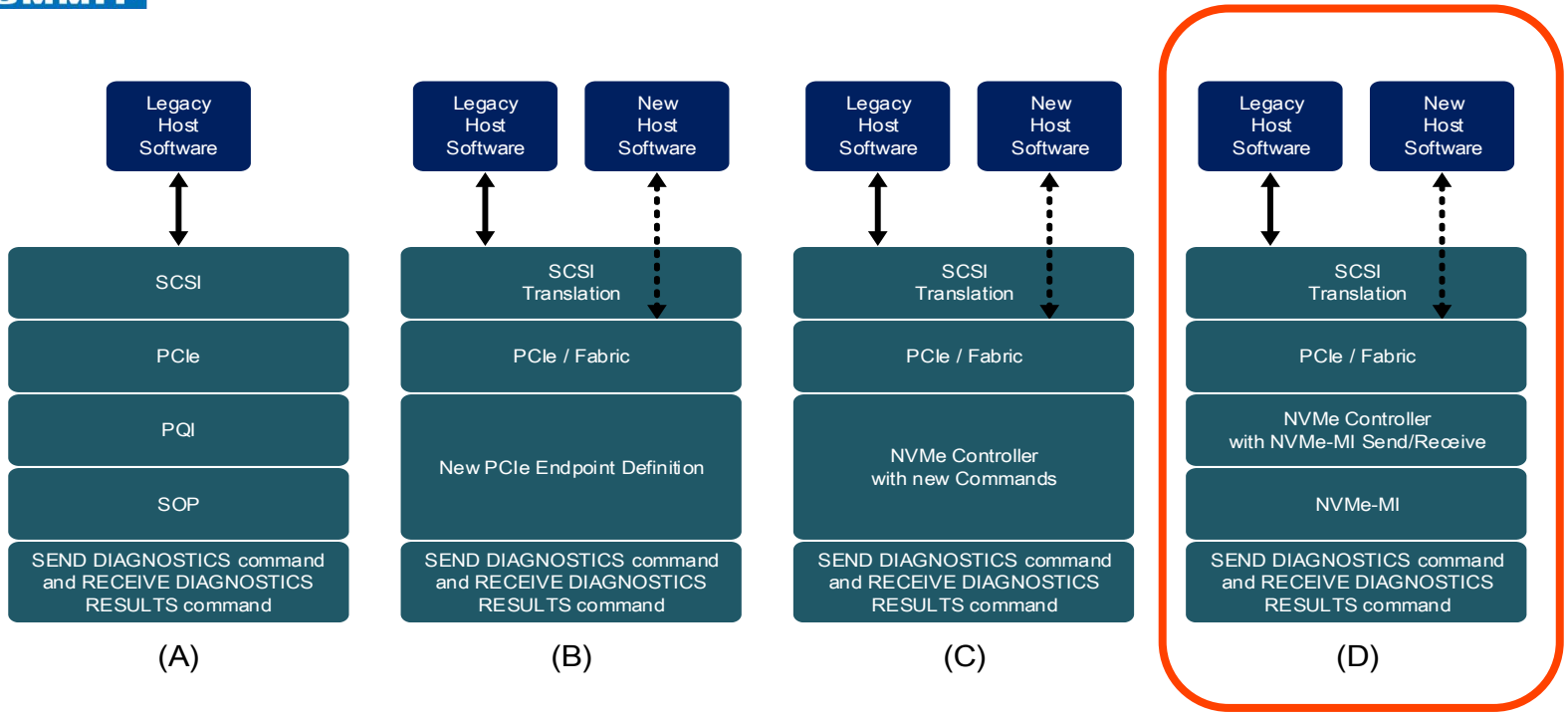
www.seagate.com/flash

SES over NVMe for Multi-Host Topology

- Multi-Host Topology with NVMe aggregator Controller at Host



Possible Options (there may be others)



NVMe workgroup selected option ‘D’ for NVMeMI pass-through model