Shared Storage Using NVMe Over Fabrics

Murali Rajagopal (VMware)
Raj Lalsangi/Madhu Pai (NetApp)
VMware Disclaimer

This presentation may contain product features or functionality that are currently under development.

This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.

Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.

Technical feasibility and market demand will affect final delivery.

Pricing and packaging for any new features/functionality/technology discussed or presented, have not been determined.
Agenda

- NVMe Support in VMWare
- NetApp ONTAP NVMe/f Architecture
Agenda

- NVMe Support in VMWare
- NetApp ONTAP NVMe/f Architecture
vSphere Storage Stack

Offers Rich Set of I/O Services Including NVMe
Storage Stack Architecture

ESXi Storage Stack

ESXi Next Generation Storage Stack

NVMe Transport Device Driver Framework

PCIe Transport Driver

RDMA Transport Driver (RoCEv1, RoCEv2, iWarp)

Fibre Channel Transport Driver

Fibre Channel Transport Abstraction

NVMe-oF Transport Abstraction

NVMe Core Functionality

Stack Interface 1

SCSI NVMe Translation

Stack Interface 2
NVMe Drivers

- NVMe Transport Device Driver Framework
  - PCIe Transport Driver
  - RDMA Transport Driver (RoCEv1, RoCEv2, iWarp)
  - Fibre Channel Transport Driver

- ESXi Storage Stack
- ESXi Next Generation Storage Stack

- Stack Interface 1
  - SCSI NVMe Translation
  - NVMe Core Functionality
  - NVMe-oF Transport Abstraction

- Stack Interface 2

- Driver Interface

- NVMe Core Functionality Stack Interface
  - Stack Interface 1
  - Stack Interface 2
  - Driver Interface

- NVMe Transport Device Driver Framework
  - PCIe Transport Driver
  - RDMA Transport Driver (RoCEv1, RoCEv2, iWarp)
  - Fibre Channel Transport Driver

- ESXi Storage Stack
- ESXi Next Generation Storage Stack
NVMe Driver Support in VMWare

- vSphere 6.5
  - Boot (UEFI)
  - Firmware Update
  - End-to-end protection
  - Deallocate/TRIM/Unmap
  - 4K
  - SMART, Planned hot-remove

- vSphere 6.7
  - Performance enhancements
  - Extended CLI
  - Name space management
  - Async event error handling
  - Enhance diagnostic logs

- Future Direction
  - PCIe Native Hot-plug
  - LED Management
  - NVMe Over Fabric
  - Multiple fabric option
  - Sanitize
  - Telemetry
Core Storage Stack
NVMe Support in Core Storage Stack

Core Storage Stack
- Reduced serialization
- Locality improvements
- vNVMe Adaption layer
- Multiple completion worlds support in NVMe

vSphere 6.5
- Optimized stack - Highly parallel execution for single path local NVMe devices
- Reach target of 90%+ performance of device spec

vSphere 6.7
- Next Generation Storage Stack with ultra-high IOPS
- End-to-end NVMe Stack
- NVMe Multi-pathing, ANA

Future Direction
vNVMe
vNVMe Support in VMWare

- **vSphere 6.5**
  - NVMe 1.0e spec
  - Hot-plug support
  - VM orchestration

- **vSphere 6.7**
  - Performance improvements
  - Async mode support
  - Unmap support

- **Future Direction**
  - NVMe 1.3 spec
  - Parallel execution @backend
  - 4K Support
  - Scatter-gather support
  - Interrupt coalescing
Agenda

- NVMe Support in VMWare
- NetApp ONTAP NVMe/f Architecture
NetApp NVMe/f

• **A SERIES OF FIRSTS IN THE INDUSTRY**
  • 9.4 (May ‘18) – First NVMe/FC solution to the market with bare metal SUSE Linux.
  • 9.5 (Dec ‘18) – Asymmetric Namespace Access (ANA), Redhat Enterprise Linux
  • 9.6 (May ‘19) – Windows support
  • 9.7 (Dec ‘19) – ESX support.

• Supported on AFF arrays with 32G FC Target Ports
• No new hardware/fabric!!
• No licensing costs
• Seamless migration from SCSI

Core Design Principles

- Lockless
- Zero Context Switching
- Polling mode
- Small Code path length
- Dynamic threading
- Zero copy
Innovation Velocity

- Reuse wherever possible
  - Leverage SAN management stack
  - Leverage existing FCP driver
  - Build new flows into existing components
  - Open source components

- Partner wherever possible
  - Broadcom for FC transport layer

9 Months to deliver a new Blocks protocol in ONTAP!!
Open Source Leverage

- **SPDK**
  - Storage Performance Development Kit
  - FreeBSD licensed software
    - Intel powered
    - [http://www.spdk.io/](http://www.spdk.io/)
  - Implements the NVMeoF Protocol stack
  - Implements Core Design Principles
    - Poll mode driver
    - Lean and optimized code path length
    - Modular with multiple libraries
    - Static number of threads
    - RDMA transport
    - Built for user space OFED verbs
    - Used DPDK – another Intel framework
Top level Architecture

- **SCSI Protocol Stack**
- **FC Transport/Driver**
- **Netapp Mgmt/Control Plane**
- **NVMe/f Protocol Stack**
- **NVMe/f Target Component**

**FC Physical Port**
NVMe/f Target Component

NVMe/f Target Component

NVMe/f Protocol Stack

FN Transport

BDEV

SPDK – NVMe/f Layer

EAL

BDAL/WAFL Interface

NVMe/f "App"

Event Reactor framework

NetApp

NetApp /Brcm

Open Source

In-progress
Protocol Objects

- Host
- Fabric
- Subsystem
- Namespace
- Session
- Controller
- Queues
  - SQ
  - CQ
ONTAP Objects

- Vserver
- Subsystems
- Namespaces
  - Create
  - Map
- NVMe/FC Lifs
- Hosts
- Controller
NVMe/FC Performance

- FC SCSI (8.3.1)
- FC SCSI (9.3)
- NVMe/FC (9.4-9.7)

- Poll mode drivers
- Optimize code path, lockless design
- Minimize context switches
- Dynamic thread pool
- IOPS

Latency
Industry Leading Performance

NVMe / FC IOPS:
- Towards 1 million / controller @.5ms
- ~100us latency up to 700K IOPS

Note: Single Node A800 with 10-node SLES 12.3 Oracle RAC 75% read/ 25% write, FCP Vs. NVMe/FC tests
ONTAP NVMe/f Next Steps – Disclaimer

- Dates are not included
- Plans and Priorities are always subject to change
ONTAP NVMe/f Next Steps

Plans and priorities are always subject to change

- Protocol Currency
  - Support for relevant TPs

- Support for NVMe Reservations
- NVMe/RoCE support
- NVMe/TCP support

- Support for TP-4078
  - Allows the NVMe controller to restrict or limit the number of attached namespaces.

- Support for TP-4027
  - Support Identify CNS 17h
  - Support for Namespace Attach/Detach
    - Support for Identify CNS 10h (Allocated Namespace ID list)

- Support for TP-8001
  - Graceful Disconnect

- Support for TP-8005
  - Fabric SQ Flow Control
  - .......
Future Directions

- VMware will continue to support relevant NVMe specifications in future vSphere releases
- VMware is also involved in SNIA’s Swordfish specifications on NVMe-oF Management
NVMe Ecosystem @VMware

• Available as part of base ESXi image from vSphere 6.0 onwards

  - Faster innovation with async release of VMware NVMe driver

• VMware led vSphere NVMe Open Source Driver project to encourage ecosystem to innovate

  - https://github.com/vmware/nvme

• VMware NVMe Driver Ecosystem


  - Close to 300 third party NVMe devices certified on VMware NVMe driver