NVM Express Software Drivers Update: Microsoft, VMware, UEFI

Lee Prewitt, Microsoft

Uma Parepalli, Stealth Startup

Arvind Jaganath, VMWare

Murali Rajagopal, VMWare

Agenda

- Windows NVMe® Driver Feature Support
- Device Hang Detection & Error Recovery
- Futures

Windows NVMe® Driver Feature Support

- https://docs.microsoft.com/en-us/windows-hardware/drivers/storage/nvme-features-supported-by-stornvme
  - Provides the latest info on command set usage, features and SCSI translation supported by the Windows NVMe driver in various OS releases
- Recent changes
  - Improvements to handling hot-add or remove of namespaces

Device Hang Detection & Error Recovery

- Support defined in Open Compute Project (OCP) Datacenter NVMe® SSD specification and available in Windows Server 2022 and Windows 11
- Specifies detection and recovery mechanisms and ability to identify the type of device hang
- Detection based on either Asynchronous Event Notification (AEN) (preferred) or Controller Fatal Status (CFS) bit
- Feature discovery and recovery based on log page C1h
- Workflow can also be used to handle other device conditions besides device hangs

Device Hang - Feature Discovery

- During device initialization, host reads log page C1h and validates the following conditions are met
  - Bytes 511:496, Log Page GUID, is equal to 5A1983BA3DFD4DABAE3430FE2131D944h
  - Bytes 15:12, Device Capabilities, is non-zero
- Host saves the following fields from log page 0xC1 for use during device hang handling
  - Device Capabilities: provide info on how device alerts host of hangs. Supported values are AEN and CFS. Valid for device to support both mechanisms.
  - Panic Reset Wait Time: amount of time in msec for host to wait for device hang workflow to complete
  - Panic Reset Action: list of all potential resets host can do to recover a device that is hung
- Host sends Asynchronous Event Request to device to wait for a hang condition

Device Hang - Notification

- Two supported mechanisms for device to alert host of a device hang condition:
  1. Controller Fatal Status (CFS) bit in the Controller Status (CSTS) register
  2. Device completes an outstanding Asynchronous Event Request (AER) command with Completion Queue Entry Dword 0 set to the following.
     - Log Page Identifier field set to C1h
     - Asynchronous Event Information field clear to zero
     - Asynchronous Event Type set to 111b (Vendor Specific)
- Device may use one or the other to alert the host for different device hang conditions if both mechanisms supported
- Device saves the following in log page C1h when hang condition is detected:
  - Type of hang provided in Panic ID field
  - Device recovery action
- Device sends debug data via Controller Initiated Telemetry log

Device Hang - Recovery (1 of 3)

- When host detects a device hang (either through AEN or CFS), it waits for Panic Reset Wait Time to allow device to finish its handling
- Host tries one or more resets specified in Panic Reset Action to attempt to bring device back to a state that can service NVMe command(s). Reset should be attempted from least impactful to most impactful
  - NVMe Controller Reset
  - PCIe Function Level Reset
  - PCIe Convention Hot Reset
  - NVM Subsystem Reset
  - PERST# or Power Cycle
- Host determines effectiveness of reset based on ability to complete controller initialization

Device Hang - Recovery (2 of 3)

- After host successfully initializes the controller, it reads the C1h log page to retrieve information about the device hang
  - A non-zero Panic ID field value indicates device is in panic mode
  - The Device Recovery Action field indicates the recommended action
- Host should retrieve the Controller-Initiated Telemetry Log to collect diagnostic data associated with device hang if Panic ID field is non-zero
- Ability to service IO while device is in panic mode depends on the condition encountered
  - If device can't guarantee data integrity, it shall fail IOs with Status Code Type 0x00 (Generic Command Status) and Status Code 0x06 (Internal Error)

Device Hang - Recovery (3 of 3)

- Host will then initiate the recommended Device Recovery Action:
  - No Action Required
  - Format NVM
  - Sanitize
  - Vendor Specific Command
  - Vendor Analysis Required
  - Device Replacement Required

Futures*

- Native NVMe® storage stack
- NVMe over Fabrics initiator support
- NVMe v2.0 specification support

*** Not plan of record**

UEFI NVMe® Drivers

Uma M Parepalli

Sponsored by NVM Express™ organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards

Agenda

- UEFI NVMe® Drivers
  - Introduction
  - Current Status
- Backup Material / Additional Resources
  - NVMe Driver Resources
  - UEFI Conceptual View
  - UEFI Specifications – Current Status

UEFI NVMe® Drivers

- Unified Extensible Firmware Interface (UEFI) is the 64-bit Platform Firmware that replaced the legacy proprietary BIOS
- UEFI NVMe Drivers are part of the Platform Firmware/BIOS (Pre-OS Boot)
- Required for booting OS from NVMe SSDs
- Eliminates the need for proprietary Legacy Option ROM support on NVMe SSDs
- Enables full debug of OS Driver functionality in pre-boot environment
- Any random issues that are hard to detect and trace such as system hangs after several days of testing can be simulated and debugged using UEFI NVMe Debug Drivers

UEFI NVMe® Drivers – Current Status in Last 3 Years

- Industry tested and reliable built-in NVMe Drivers
- Thoroughly tested on ARM OCP platforms in last 3 years at scale
- Plug-and-boot functionality independent of NVMe SSD vendors working successfully at scale
- Standard built-in drivers for all Intel, ARM and AMD Server Platforms including OCP Servers

UEFI NVMe® Drivers – Current Status

- Bug free and seamless handover of boot functionality to all leading Operating Systems
- Fixed the issues related to routing and handling of NVMe Commands thus eliminating legacy SAS/SATA handling of NVMe commands
- New features are being added as needed to the UEFI Specifications

vSphere NVMe® Driver and Stack Support

Sponsored by NVM Express™ organization, the owner of NVMe, NVMe-oF™ and NVMe-MI™ standards

Speakers

Arvind Jaganath

vmware®

Murali Rajagopal

vmware®

Disclaimer

VMware 2022

- 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.
- This information is confidential.

The information in this presentation is for informational purposes only and may not be incorporated into any contract. There is no commitment or obligation to deliver any items presented herein.

Agenda

Motivation and Vision

High-level updates

vSphere Release Features

NVMe®/TCP and iSCSI Performance

Future Roadmap

Motivation and Vision

- Committed to…
  - Bringing latency and performance close to bare-metal
  - Improving Scale – (Namespaces, Paths etc.)
  - Improving Fabric Manageability and Standards Compliance
- Consumability and Resiliency
  - Virtual Volumes (vVols), support for storage migrations, DRS, HA etc., Clustered usage
- Security
  - Authentication and on-the-wire encryption
- Future Proofing - Code refactoring and re-organization
  - Native NVMe stack, Stack fast-paths/optimizations
- Hardware Offloads

High-Level Updates

- NVMe®/TCP
- Standards
  - Fabric Management, Abort
- Performance and Scalability Optimizations
  - Hybrid Polling (SPDK-like) (Direct-Attached)
    - Now as much as ~8M IOPs performance
  - Storage stack analysis and improvements
    - Direct-Attached NVMe and NVMe-oF™
  - Namespaces and Paths Scale
- NVMe Reservations
- Passthrough with NVMe devices
  - Scale and Hot-plug capabilities
  - Resiliency – DPC, LED control
- vVols

- VMware Compatibility Guide - “VMware’s HCL”

https://www.vmware.com/resources/compatibility/search.php?deviceCategory=io&details=1&keyword=nvme&page=3&display_interval=500&sortColumn=Partner&sortOrder=Asc

  - ~1900 listings supporting the latest and greatest HBA/Firmware
  - Supporting a large ecosystem of IO vendors, Array vendors, OEMs
    - NVMe Flash/SSDs (Storage vendors)
    - HBAs (RAID, Tri-mode, RDMA, FC), NIC (supporting NVMe/TCP)
    - OEMs supporting Direct-attached NVMe or NVMe-oF
    - Array vendors supporting (NVMe-oF /RDMA /FC and /TCP)

vSphere Release Features

Released NVMe® Features in vSphere 7.0u3

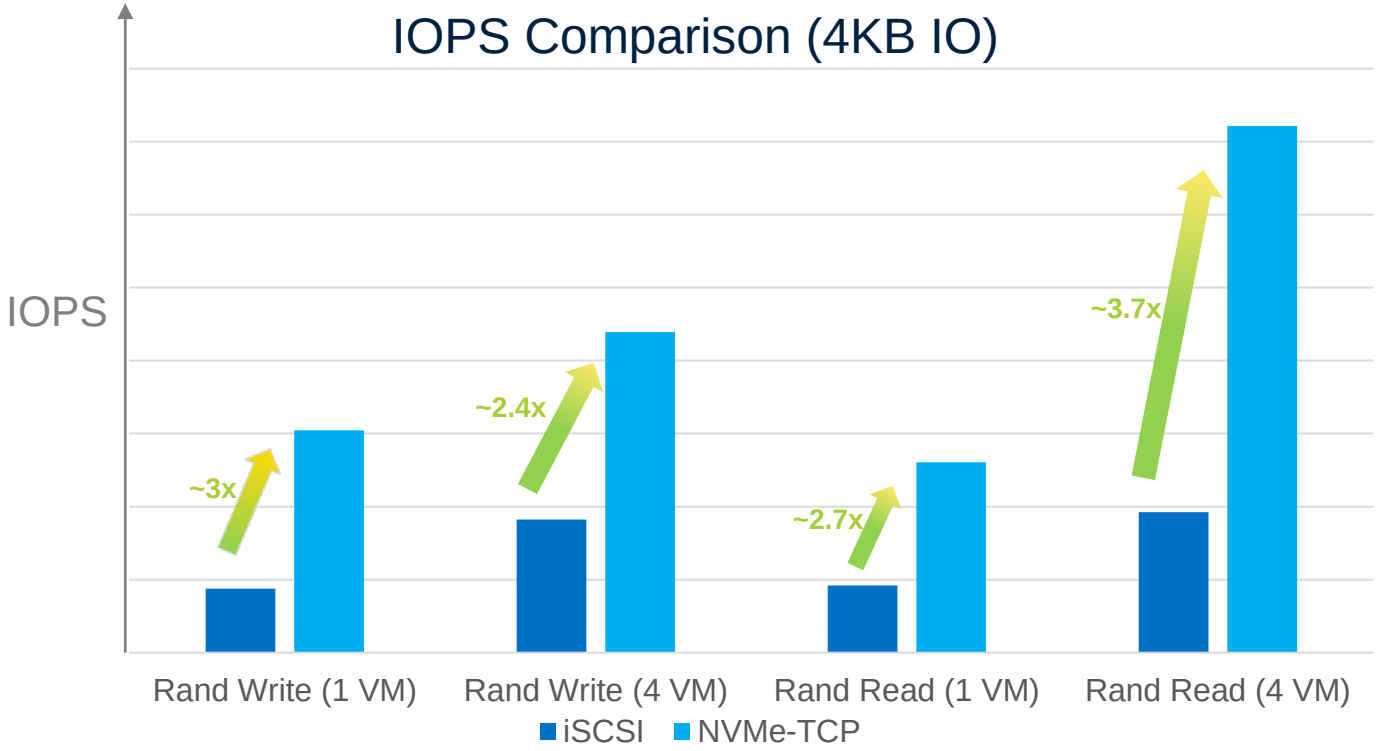
- NVMe-oF™/TCP (Initiator)
  - 2 array vendors certified (more in the process)
- TP 8002 - NVMe-oF Discovery (Partial)
- TP 4097 - Abort Enhancements
- TP 8010 - NVMe-oF Central Discovery Controller (CDC)

NVMe Targeted Features for vSphere 8.0

- TP 8002 - NVMe-oF Discovery (Full)
- TP 8010 - NVMe-oF Central Discovery Controller (CDC)
- NVMe Reservation support for clustered VMDK (WSFC)
- vVols w/NVMe – FC Only
- Support for 256 Namespaces and 2K Paths

NVMe®/TCP Vs iSCSI Performance Comparison

Vendor X Array, 2 node benchmark: FIO



- 1 VM and 4 VM Tests
- Up to ~2.4 x improvement for Random Writes
- Up to ~3.7 x improvement for Random Reads

Future Roadmap

- E2E NVMe® support in our ESXi stack
- Security
  - TP 8006 -In-Band Authentication
  - TP 8011 – NVMe TLS 1.3
  - TP 8019 - Authentication Verification Entity (AVE)
- TP 8012- Boot Over NVMe/TCP
- TP 4034 – Dispersed Namespaces
  - MetroCluster use cases
- TP 8009 - Automated Discovery of IP Discovery Controllers (NVMe-TCP)
- TP 4033 - Advanced Command Retry Enable (ACRE)
- TP 4040 - Max Data Transfer for non-IO Commands (MDTS)
- vVols for other Fabrics
- NVMe-oF™ Offload with SmartNICs
  - CPU core savings and higher performance

Questions?

Architected for Performance