Scalable High IOPS on vSphere ESX and Linux with NVMe/FC

Wenhua Liu, VMware
Jayamohan Kallickal, Broadcom
Broadcom’s Data Center Portfolio

- **Networking**
  - Ethernet Switch ICs
  - Ethernet NIC Adapter
  - Fibre Channel HBA
  - DAS (NVMe/SAS/SATA HBA/MegaRAID)
  - Ethernet NIC
  - Ethernet Switch ICs

- **Server**
  - Server OEMs
  - Fibre Channel Switch OEMs
  - Storage Array
    - (Fibre Channel HBA)
    - (Smart NIC)
    - (SAS/SATA/NVMe/ROC)

- **Disk & Flash Storage**
  - Storage Array
    - (SAS/SATA/NVMe/ROC)
    - (Fibre Channel HBA)
    - (Smart NIC)
**CPU Affinity**

**EQ Per Core**
- Per-CPU WQ/CQ (a “Hardware Queue”)
- Interrupt vector/EQ per CPU
- Interrupt vector/EQ per CPU

**EQ Per Socket**
- One Interrupt Vector/EQ per Socket
FC exchanges
- Adapter has a fixed number
- Needed for SCSI and NVMe
- Exchange assigned to each IO for the duration of the IO
- Partitioning per CPU resulted in few resources per CPU, thus lots of IO “busying”
- Solve by pools per Hardware Queue with resources migrating between Hardware Queues on as-needed basis
Interrupt Handling:
- Disassociate EQ from CQ
  - EQ must be serviced by ISR
  - CQ serviced by Independent Thread

CQ Processing Tenancy
- How much work you do while in the thread
- Large limits put in. If limit reached and work remains, re-schedule

Periodic Queue Pointer Updates to Hardware

Interrupt Rate Management
- Interrupt re-enablement
  - Use architecture-specific re-arming to reduce interrupt rate
  - Interrupt delay largely left “immediate”
- Exception: CPU shared by Interrupt Vectors or HWQs
NVMe Lancer G6 & Prism
1-port & 2-ports IOPs Trend

NVMe SLES 12 SP3 Lancer G6 & Prism IOPs for 12.0.x to 12.4.x
with Prism target

IOps in Millions

12.0. 12.2.1 12.2.2 12.4
Overview of NVMe Device Driver Development in vSphere ESX
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.
NVMe Device Driver
in Current ESXi Release

Storage Stack

NVMe
PCIe Driver

SCSI/NVMe Translation
Controller Configuration
Namespace Discovery
PCIe Specific Things
Scalable Device Driver Model for Future ESXi Release

Storage Stack
- NVMe-SCSI Translation
- NVMe Common Layer

Next Generation Storage Stack
- NVMe-PSA

Kernel Space

NVMe Common Layer
- NVMe RDMA Driver
- 3rd Party FC-NVMe Driver

RDMA Stack
- RoCE Driver
- iWarp Driver
- TCP/IP Stack

User Space

ESXCLI Plugins
- VMKCTL
- ESXCLI Plugins
Features of New Driver Model

- Implements most of common functions defined in NVMe base specification and NVMe-oF specification that are needed for VMware ESXi.
- Common user interface for NVMe device management.
- Transport agnostic driver interface for PCIe based and Fabrics based NVMe driver development.
- Supports auto discovery/connect of NVMe-oF controllers for NVMe/FC.
- Supports persisted connection of NVMe-oF controllers.
- Supports existing SCSI based storage stack and future NVMe native storage stack.
- Much simpler way implementing NVMe transport device driver.
Driver Objects

- NVMe Adapter
- NVMe Controller
- Admin/IO Queue
### User Interface

```
[root@localhost:~] esxcli nvme adapter list

<table>
<thead>
<tr>
<th>Adapter</th>
<th>Adapter Qualified Name</th>
<th>Transport Type</th>
<th>Driver</th>
<th>Associated Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmhba32</td>
<td>aqn:nvme_pcie:nqn.2014-08.org.nvmexpress15ad15adVMWare_NVME-0000VMware_Virtual_NVMe_Disk</td>
<td>PCIe</td>
<td>nvme_pcie</td>
<td></td>
</tr>
<tr>
<td>vmhba33</td>
<td>aqn:brcmnvmefc:100000000fa9892f</td>
<td>FC</td>
<td>brcmnvmefc</td>
<td></td>
</tr>
<tr>
<td>vmhba34</td>
<td>aqn:brcmnvmefc:100000000fa98930</td>
<td>FC</td>
<td>brcmnvmefc</td>
<td></td>
</tr>
<tr>
<td>vmhba35</td>
<td>aqn:nvmerdma:24-8a-0-7b-34-32</td>
<td>RDMA</td>
<td>nvmerdma</td>
<td>vmrdma0, vmnic0</td>
</tr>
</tbody>
</table>
```

```
[root@localhost:~] esxcli nvme controller list

<table>
<thead>
<tr>
<th>Name</th>
<th>Controller</th>
<th>Adapter</th>
<th>Transport</th>
<th>Online</th>
</tr>
</thead>
<tbody>
<tr>
<td>nqn.2014-08.org.nvmexpress_15ad_VMware_Virtual_NVMe_Disk__VMWare_NVME-0000</td>
<td>256</td>
<td>vmhba32</td>
<td>PCIe</td>
<td>true</td>
</tr>
<tr>
<td>nqn.2014-08.org.sanblaze:virtualun.prme-hwe-drv-sanblaze-002.0.0#vmhba33#200200110de23a00:200400110de23a00</td>
<td>259</td>
<td>vmhba33</td>
<td>FC</td>
<td>true</td>
</tr>
<tr>
<td>nqn.2014-08.org.sanblaze:virtualun.prme-hwe-drv-sanblaze-002.1.0#vmhba34#200300110de23b00:200500110de23b00</td>
<td>264</td>
<td>vmhba34</td>
<td>FC</td>
<td>true</td>
</tr>
<tr>
<td>nqn.2010-06.com.purestorage:flasharray.4d4bafbf03558e0f#vmhba35#10.20.54.101</td>
<td>266</td>
<td>vmhba35</td>
<td>RDMA</td>
<td>true</td>
</tr>
<tr>
<td>nqn.2010-06.com.purestorage:flasharray.4d4bafbf03558e0f#vmhba35#10.20.54.102</td>
<td>268</td>
<td>vmhba35</td>
<td>RDMA</td>
<td>true</td>
</tr>
</tbody>
</table>
```

```
[root@localhost:~] esxcli nvme namespace list

<table>
<thead>
<tr>
<th>Name</th>
<th>Controller</th>
<th>Namespace ID</th>
<th>Block Size</th>
<th>Capacity in MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>t10.NVMe__VMware_Virtual_NVMe_Disk__VMWare_NVME-0000__00000001</td>
<td>256</td>
<td>1</td>
<td>512</td>
<td>40960</td>
</tr>
<tr>
<td>eui.600110d003a23b000040100000ac7d235</td>
<td>264</td>
<td>1</td>
<td>512</td>
<td>10240</td>
</tr>
<tr>
<td>eui.600110d003a23b000040100000ac7d236</td>
<td>264</td>
<td>2</td>
<td>512</td>
<td>16</td>
</tr>
<tr>
<td>eui.600110d003a23b000040100000ac7d236</td>
<td>264</td>
<td>2</td>
<td>512</td>
<td>2048</td>
</tr>
<tr>
<td>eui.600110d003a23b000040100000ac7d236</td>
<td>264</td>
<td>3</td>
<td>512</td>
<td>8192</td>
</tr>
<tr>
<td>eui.00d80b8cbbc79e4324a9374a00011fc6</td>
<td>266</td>
<td>73670</td>
<td>512</td>
<td>61440</td>
</tr>
<tr>
<td>eui.00d80b8cbbc79e4324a9374a00011fc7</td>
<td>266</td>
<td>73671</td>
<td>512</td>
<td>10240</td>
</tr>
</tbody>
</table>
```