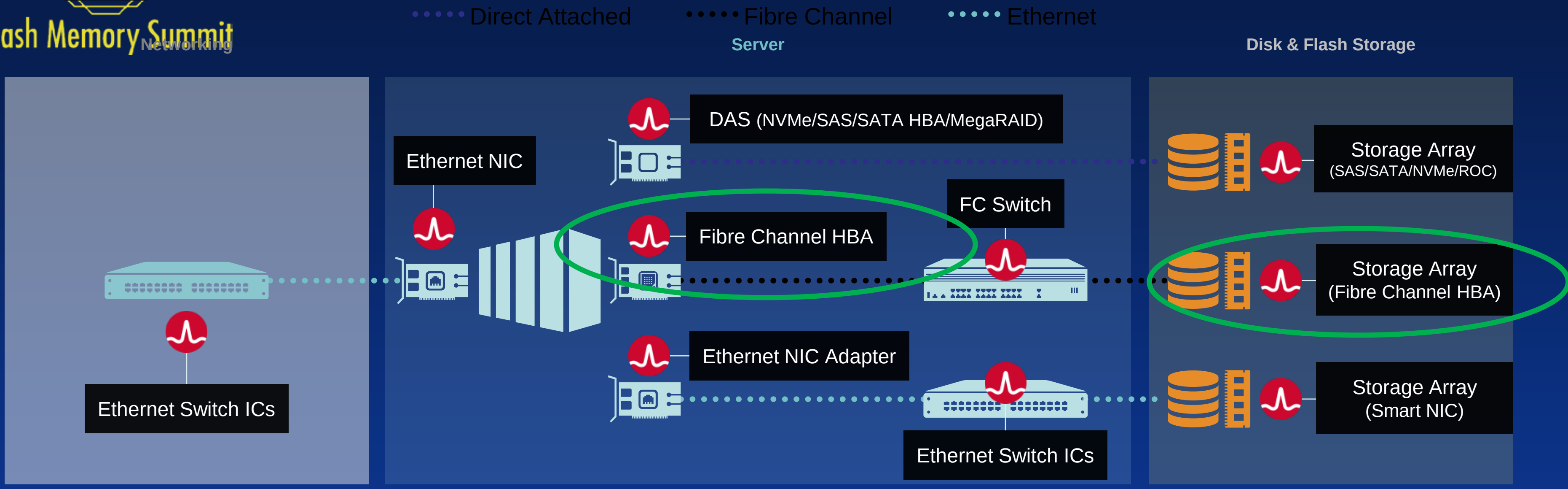
Scalable High IOPS on vSphere ESX and Linux with NVMe/FC

Wenhua Liu, VMware

Jayamohan Kallickal, Broadcom

Broadcom’s Data Center Portfolio



Server OEMs

CISCO DELL Hewlett Packard Enterprise IBM Lenovo

Fibre Channel Switch OEMs

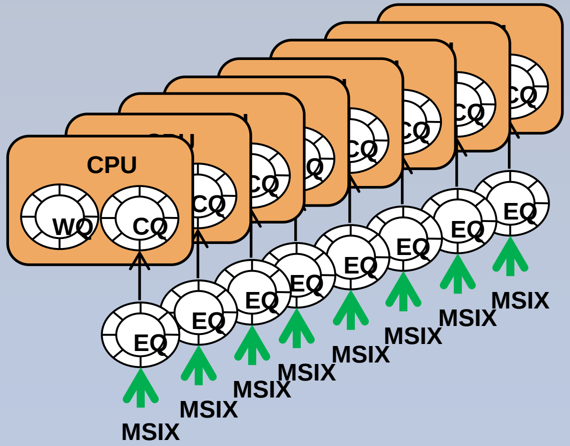
DELL EMC FUJITSU Hewlett Packard Enterprise HITACHI Inspire the Next HUAWEI IBM Lenovo NetApp

Disk & Flash Storage Arrays OEMs

DELL EMC Hewlett Packard Enterprise IBM NetApp nimble storage PURESTORAGE

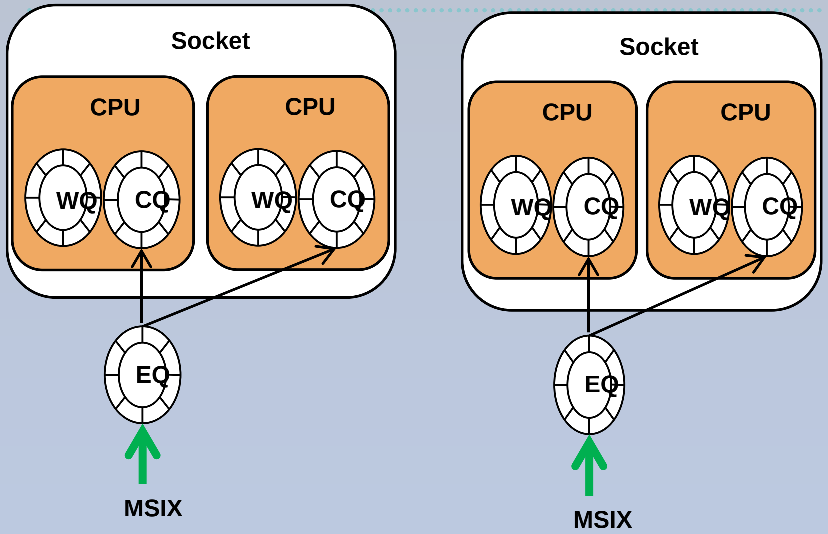
CPU Affinity

EQ Per Core



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

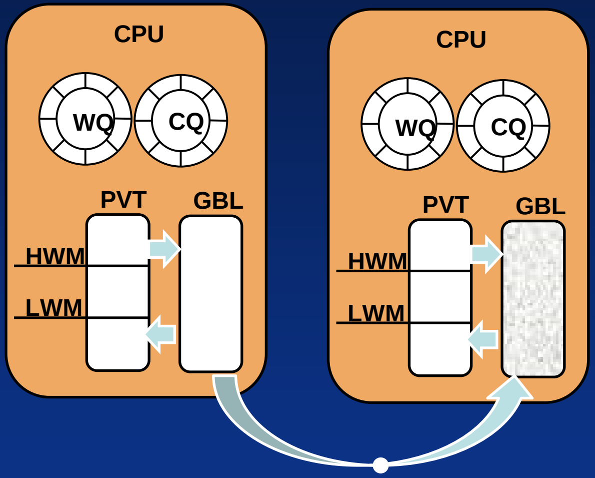
EQ Per Socket



- One Interrupt Vector/EQ per Socket

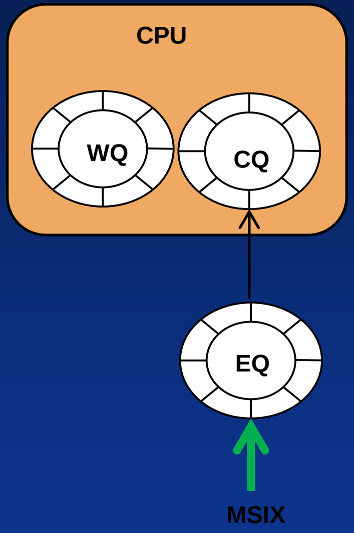
Sharing Adapter Resources

- 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

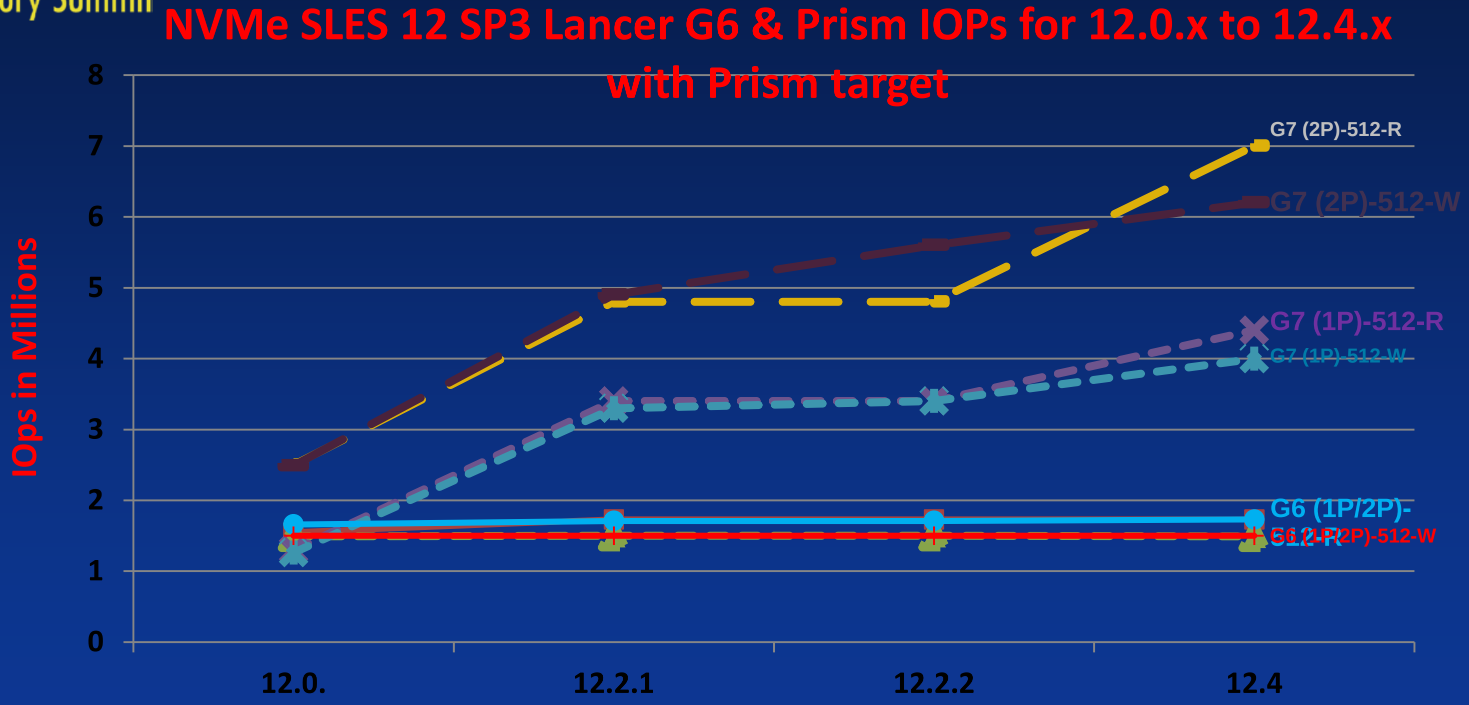


Interrupt Handling

- 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

Flash Memory Summit

Overview of NVMe Device Driver Development in vSphere ESX

Disclaimer

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

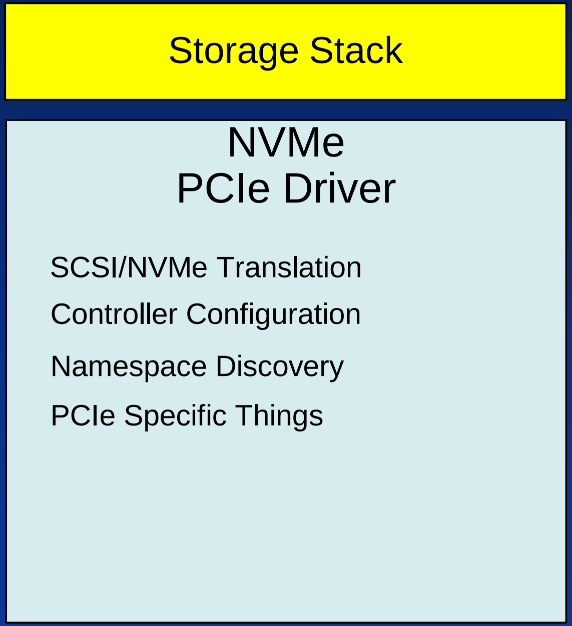
This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.

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.

NVMe Device Driver in Current ESXi Release



Storage Stack

NVMe
PCIe Driver

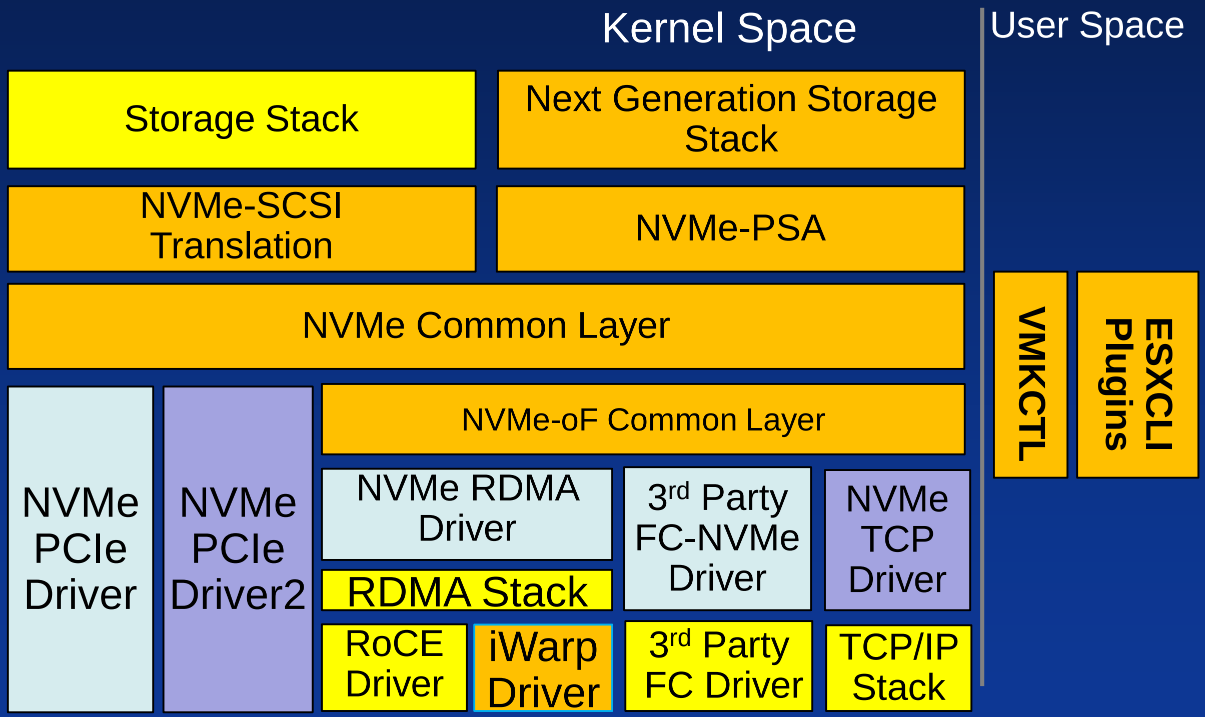
SCSI/NVMe Translation

Controller Configuration

Namespace Discovery

PCIe Specific Things

Scalable Device Driver Model for Future ESXi Release



Kernel Space

User Space

Storage Stack

Next Generation Storage Stack

NVMe-SCSI Translation

NVMe-PSA

NVMe Common Layer

NVMe PCIe Driver

NVMe PCIe Driver2

NVMe-oF Common Layer

NVMe RDMA Driver

RDMA Stack

RoCE Driver

iWarp Driver

3rd Party FC-NVMe Driver

3rd Party FC Driver

NVMe TCP Driver

TCP/IP Stack

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
Adapter  Adapter Qualified Name                                                                    Transport Type  Driver      Associated Devices
-------  ----------------------------------------------------------------------------------------  --------------  ----------  ------------------
vmhba32  aqn:nvme_pcie:nqn.2014-08.org.nvmexpress15ad15adVMWare_NVME-0000VMware_Virtual_NVMe_Disk  PCIe            nvme_pcie
vmhba33  aqn:brcmnvmefc:10000090fa94892f                                                           FC              brcmnvmefc
vmhba34  aqn:brcmnvmefc:10000090fa948930                                                           FC              brcmnvmefc
vmhba35  aqn:nvmerdma:24-8a-07-b4-34-32                                                            RDMA            nvmerdma    vmrdma0, vmnic0

[root@localhost:~] esxcli nvme controller list
Name                                                                                                       Controller  Adapter  Transport  Online
                                                                                                           Number               Type
------------------------------------------------------------------------------------------                 ----------  -------  ---------  ------
nqn.2014-08.org.nvmexpress_15ad_VMware_Virtual_NVMe_Disk________________VMWare_NVME-0000                          256  vmhba32  PCIe         true
nqn.2014-08.org.sanblaze:virtualun.prme-hwe-drv-sanblaze-002.0.0#vmhba33#200200110de23a00:200400110de23a00        259  vmhba33  FC           true
nqn.2014-08.org.sanblaze:virtualun.prme-hwe-drv-sanblaze-002.1.0#vmhba34#200300110de23b00:200500110de23b00        264  vmhba34  FC           true
nqn.2010-06.com.purestorage:flasharray.4d4bafbf03558e0f#vmhba35#10.20.54.101                                      266  vmhba35  RDMA         true
nqn.2010-06.com.purestorage:flasharray.4d4bafbf03558e0f#vmhba35#10.20.54.102                                      268  vmhba35  RDMA         true

[root@localhost:~] esxcli nvme namespace list
Name                                                                              Controller Number  Namespace ID  Block Size  Capacity in MB
--------------------------------------------------------------------------------  -----------------  ------------  ----------  --------------
t10.NVMe____VMware_Virtual_NVMe_Disk________________VMWare_NVME-0000____00000001                256             1         512           40960
eui.600110d003e23b0004010000ac07d235                                                            264             1         512           10240
eui.600110d003e23b0004010000ac07d236                                                            264             2         512              16
eui.600110d002e23a0003000000c5728fa4                                                            259             1         512            8192
eui.600110d002e23a0003000000c5728fa5                                                            259             2         512            2048
eui.600110d002e23a0003000000c5728fa6                                                            259             3         512            8192
eui.00d80b8cbcc79e4324a9374a00011fc6                                                            266         73670         512           61440
eui.00d80b8cbcc79e4324a9374a00011fc7                                                            266         73671         512           10240
eui.00d80b8cbcc79e4324a9374a00011fc6                                                            268         73670         512           61440
eui.00d80b8cbcc79e4324a9374a00011fc7                                                            268         73671         512           10240
```