Storage Protocol Offload for Virtualized Environments

Session 301-F

Dennis Martin, President

Demartek®

Santa Clara, CA
August 2016

Agenda

- About Demartek
- Offloads
- I/O Virtualization Concepts
- RDMA Concepts
- Overlay Networks and Tunneling

About Demartek

- Industry Analysis and ISO 17025 accredited test lab
- Lab includes enterprise servers, networking & storage (DAS, NAS, SAN, 10/25/40/100 GbE, 16/32 GFC)
- We prefer to run real-world applications to test servers and storage solutions (databases, Hadoop, etc.)
- Demartek is an EPA-recognized test lab for *ENERGY STAR Data Center Storage* testing
- Website: www.demartek.com/TestLab

SNIA Emerald™ RECOGNIZED TESTER

Flash Storage Brings Expectations

- Flash storage changes the dynamic in enterprise data centers and often moves the bottleneck
- Networks must keep pace, including network adapters
- There are several technologies designed to improve performance or reduce latency available today

Offloads

- A number of functions can be offloaded onto adapters
  - “hardware offloads”
  - This improves (lowers) host CPU utilization
  - This can improve IOPS or FPS, throughput and/or latency
- Functions include:
  - Various TCP/IP functions: checksums, large send, etc.
  - iSCSI & FCoE – turns a “network adapter” into a “storage controller”
- Other examples:
  - RAID controllers, Fibre Channel adapters, Graphics cards (GPUs)

NIC Port Partitioning

- Creation of multiple PCIe functions for each adapter port
  - Known by various names: “NPAR”, “Universal Multi-Channel”, etc.
- These partitions appear to the O.S. or hypervisor as separate physical adapters, each with its own MAC address
- Bandwidth can be allocated and managed among the partitions
- 10GbE adapters: typically up to 4 partitions per port
  - Faster speeds may support more than 4 partitions per port

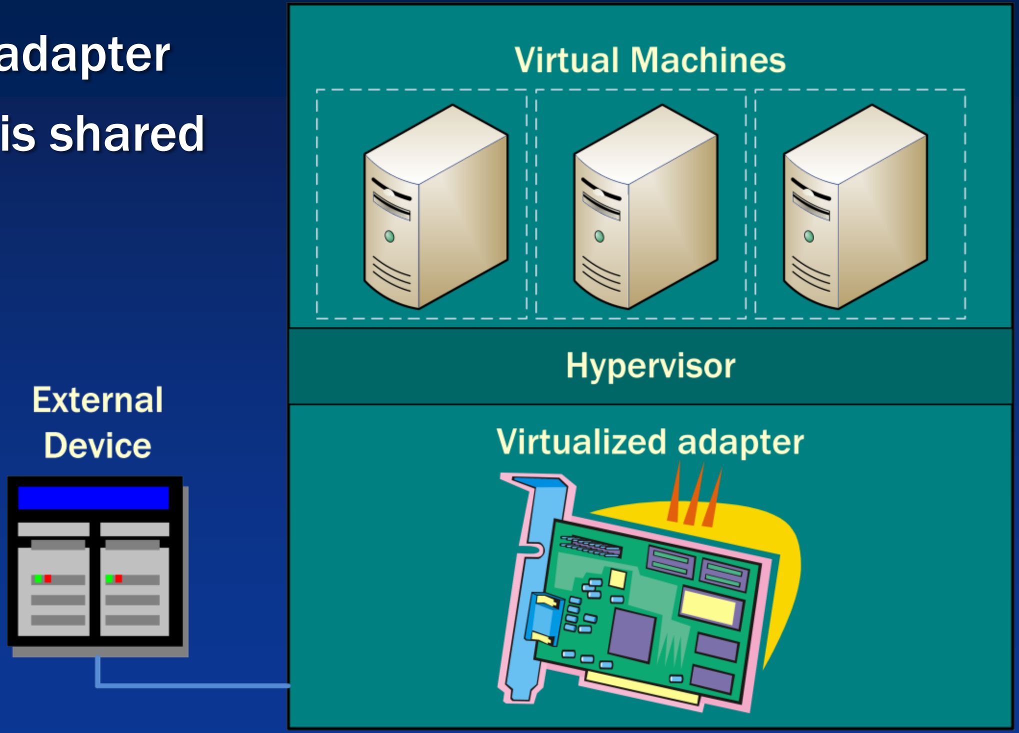
I/O Virtualization

I/O Virtualization

- Virtualizing the I/O path between a server and an external device
- Can apply to anything that uses an adapter in a server, such as:
  - Ethernet Network Interface Cards (NICs)
  - Disk Controllers (including RAID controllers)
  - Fibre Channel Host Bus Adapters (HBAs)
  - Graphics/Video cards or co-processors
  - SSDs mounted on internal cards

I/O Virtualization General Diagram

- Multiple VMs sharing one I/O adapter
- Bandwidth of the I/O adapter is shared among the VMs
- Virtual adapters created and managed by adapter (not hypervisor)
- Improved performance for VMs and their apps.



Benefits of I/O Virtualization

- Increases utilization of adapters
- Expensive adapters can be shared rather than dedicated to a single server/O.S.
- Decreases power consumption and cooling needs in some cases
- Reduced rack space servers can be deployed in some cases
- O.S. and hypervisor device management tasks can be offloaded to the adapter, increasing overall performance

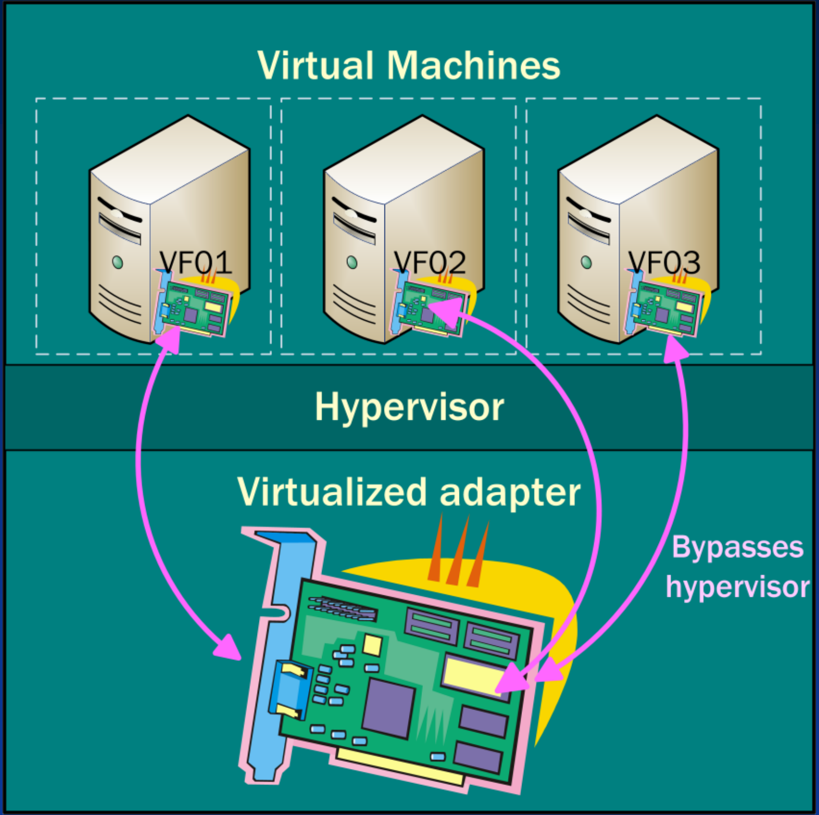
I/O Virtualization Today

- **SR-IOV** (Ethernet)
  - Single Root I/O Virtualization (PCIe bus specification)
  - Enables multiple guest operating systems to simultaneously access an I/O device or adapter without having to trap to the hypervisor on the main data path
  - Works with I/O virtualization functions of host processor

- **NPIV** (Fibre Channel)
  - N_Port ID Virtualization
  - Enables multiple guest operating systems to simultaneously share a single Fibre Channel port id (similar concept to SR-IOV)

Virtual Functions (VF)

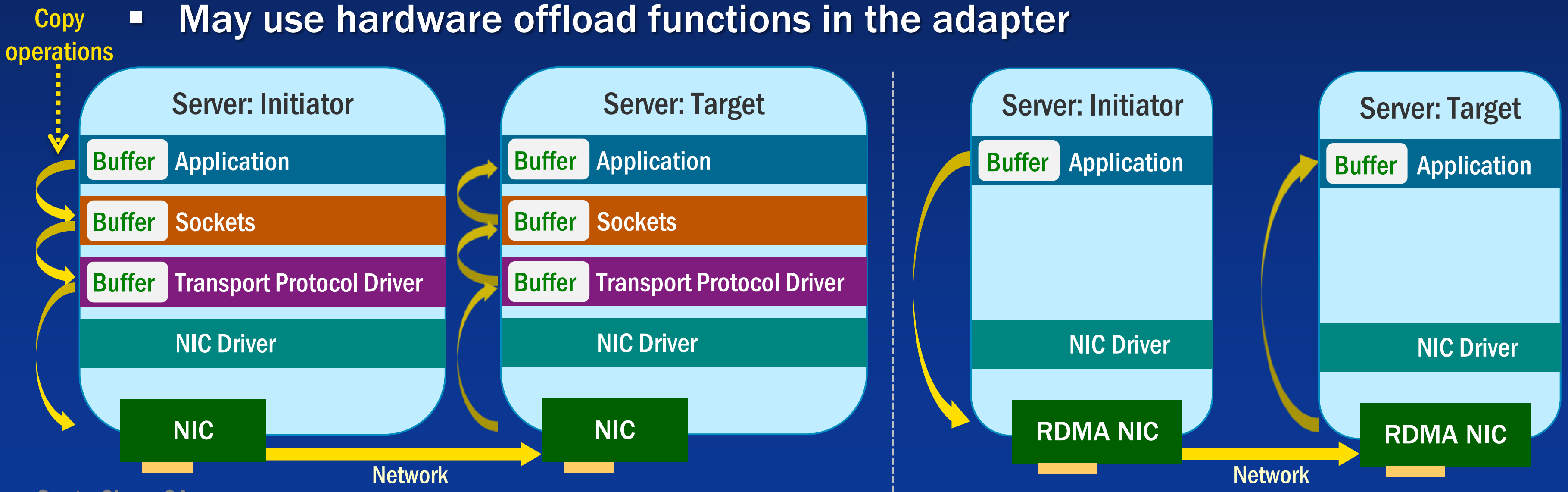
- For SR-IOV and NPIV, virtual functions are created that can be allocated to virtual machines
  - Ethernet NICs: VFs get unique MAC addresses
  - Fibre Channel: VFs get unique WWN
- Hypervisor does not see the VFs
- Adapter manages the VFs



RDMA

RDMA – Remote Direct Memory Access

- Enables more direct movement of data in/out of server
  - RDMA bypasses system software network traffic stack components
  - Bypasses multiple buffer copies, reduces CPU utilization, reduces latency
  - May use hardware offload functions in the adapter



What Networks Can Use RDMA?

- InfiniBand (IB) – the default transport protocol
- Ethernet with RoCE: RDMA over Converged Ethernet
  - Requires DCB switch (lossless fabric)
- Ethernet with iWARP: Internet Wide Area RDMA protocol
  - Runs on top of regular TCP/IP

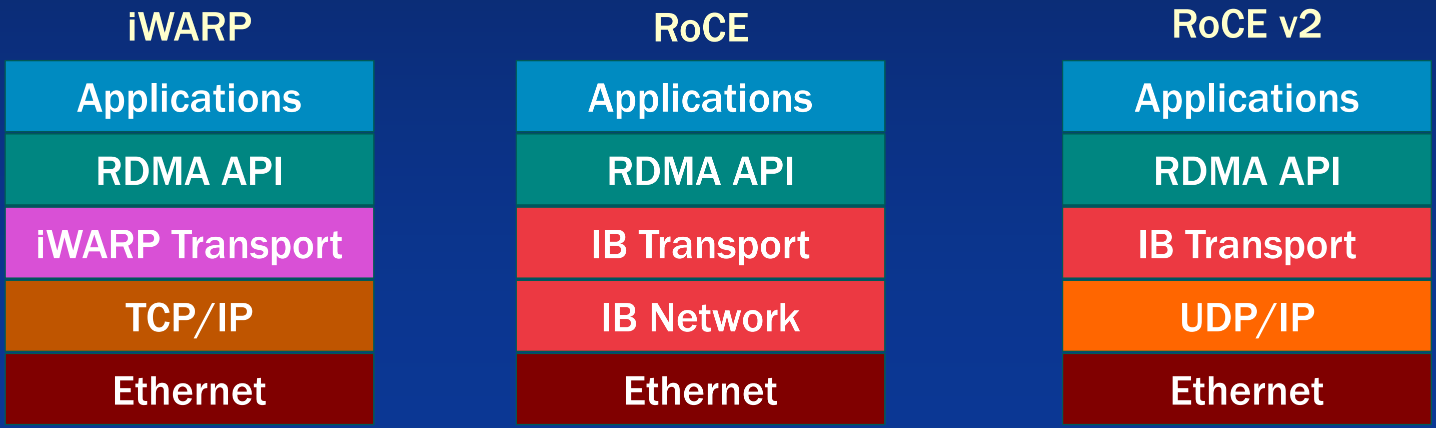
✓ RDMA is available for 10Gb and faster Ethernet technologies

RDMA Applications

- **iSER**: iSCSI Extensions for RDMA (Ethernet)
- **SRP**: SCSI RDMA Protocol (IB)
- **SMB Direct**: Windows Server feature for file servers that takes advantage of RDMA-capable network adapters (Ethernet or IB)
- **NFS over RDMA**: Linux RDMA transport for NFS (Ethernet or IB)
- **NVMe over Fabrics**: RDMA-enabled networks are ideal for this (although not the only way)
- RDMA-enabled distributed filesystems
- RDMA-enabled scale-out distributed SAN or caching

RoCE and iWARP

- iWARP and RoCE adapters cannot communicate via RDMA to each other
  - iWARP adapters speak RDMA only with other iWARP adapters
  - RoCE adapters speak RDMA only with other RoCE adapters



iWARP	RoCE	RoCE v2
Applications	Applications	Applications
RDMA API	RDMA API	RDMA API
iWARP Transport	IB Transport	IB Transport
TCP/IP	IB Network	UDP/IP
Ethernet	Ethernet	Ethernet

Overlay Networks & Tunneling

Overlay Networks and Tunneling

- In large-scale environments we may desire multiple virtual networks on the same physical network
  - Multi-tenant environments: isolate clients from each other
- Accomplished by “tunneling” or “encapsulating” the virtual network traffic within physical Ethernet packets
  - Potentially millions of secure, private networks running over a physical network
  - Extends virtual networks from the datacenter into the cloud
- Requires adapter modifications

VXLAN, STT, NVGRE & GRE

- These protocols modify the Ethernet packet structure to provide a new virtual network identifier
  - Not the same as VLAN tagging
  - Requires support by the adapter (another offload function)
  - Some older adapters can’t support this, affects their offload functions
- VMware: VXLAN, STT (stateless tunneling protocol)
- Microsoft Windows: NVGRE
- Linux: GRE

GENEVE

- Generic Network Virtualization Encapsulation (GENEVE) is a way to combine the other tunneling protocols into one protocol
- Co-authored by Intel, Microsoft, Red Hat and VMware
- Currently in draft form at the IETF
    - https://datatracker.ietf.org/doc/draft-ietf-nvo3-geneve/

Demartek Presentations

- These presentations will be posted to:
  www.demartek.com/flashmem

  - 102-C "How Flash-Based Storage Performs on Real Applications"
  - 301-F "Storage Protocol Offload for Virtualized Environments"

  - Storage Valley Supper Club (Thursday night, August 11):
    "NVMe over Fabrics is Headed Our Way"

Demartek Free Resources

- Demartek SSD Zone – www.demartek.com/SSD
- Demartek iSCSI Zone – www.demartek.com/iSCSI
- Demartek FC Zone – www.demartek.com/FC
- Demartek SSD Deployment Guide
  www.demartek.com/Demartek_SSD_Deployment_Guide.html
- Demartek commentary: “Horses, Buggies and SSDs”
  www.demartek.com/Demartek_Horses_Buggies_SSDs_Commentary.html
- Demartek Video Library - http://www.demartek.com/Demartek_Video_Library.html

Performance reports, Deployment Guides and commentary available for free download.

Thank You!

Demartek®

Demartek public projects and materials are announced on a variety of social media outlets. Follow us on any of the above.

SUBSCRIBE TO THE DEMARTEK NEWSLETTER

Sign-up for the Demartek monthly newsletter, *Demartek Lab Notes*.
www.demartek.com/newsletter