



June 13-15, 2016

| Marriott San Mateo

| San Mateo, CA

Learn Your Alphabet – SRIOV, NPIV, RoCE, iWARP – to Pump Up Virtual Infrastructure Performance

Dennis Martin
Demartek



Agenda

- ❑ About Demartek
- ❑ I/O Virtualization Concepts
- ❑ RDMA Concepts
- ❑ Examples
- ❑ Demartek Free Resources

Demartek Video



Click to view this one minute video
(available in 720p and 1080p)

Demartek YouTube Channel:

<http://www.youtube.com/user/Demartek/videos>

http://www.demartek.com/Demartek_Video_Library.html

About Demartek

- ❑ Industry Analysis and ISO 17025 accredited test lab
- ❑ Lab includes enterprise servers, networking & storage (DAS, NAS, SAN, 10 / 25 / 40 / 100GbE, 32GFC)
- ❑ 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



The Need For More Bandwidth

► Server and Application Growth

❑ Server Virtualization

- ❑ How many VMs per physical server do you deploy?
- ❑ Compare the number of VMs today vs. one and two years ago

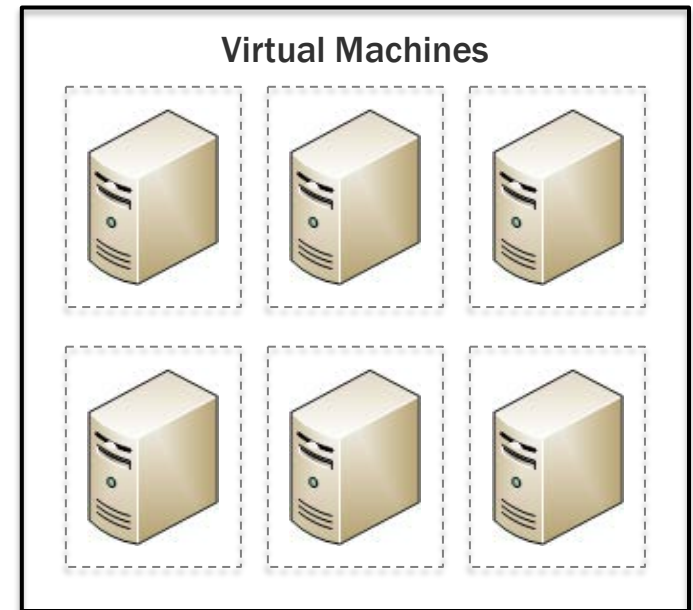
❑ Application Growth

- ❑ Applications processing more data today

❑ Bootstorm test with 90 VMs in one physical server

www.demartek.com/Demartek_Analysis_of_VDI_Storage_Performance_during_Bootstorm.html

Physical Server





I/O Virtualization

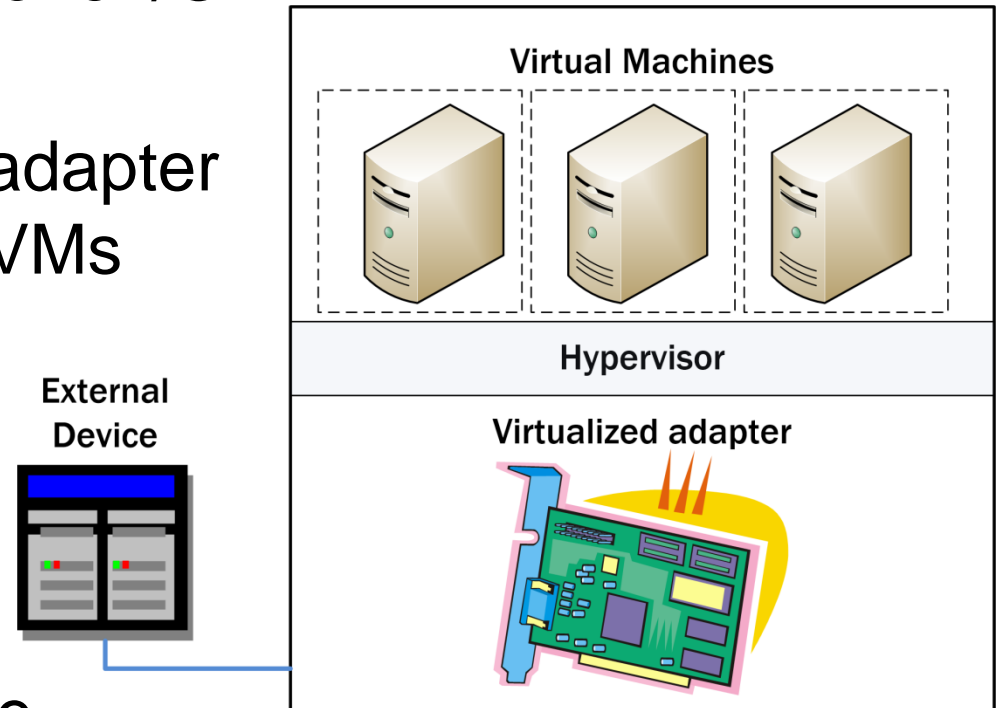


I/O Virtualization

- ❑ 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)

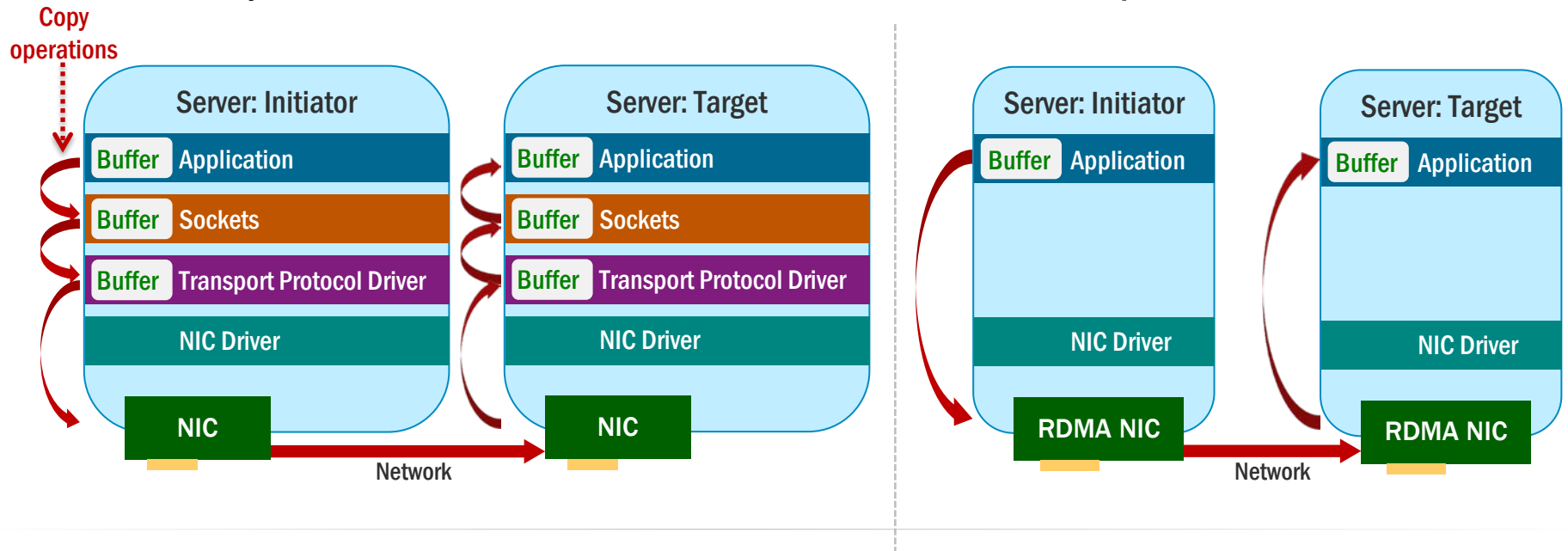


RDMA



Remote Direct Memory Access (RDMA)

- ❑ 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?

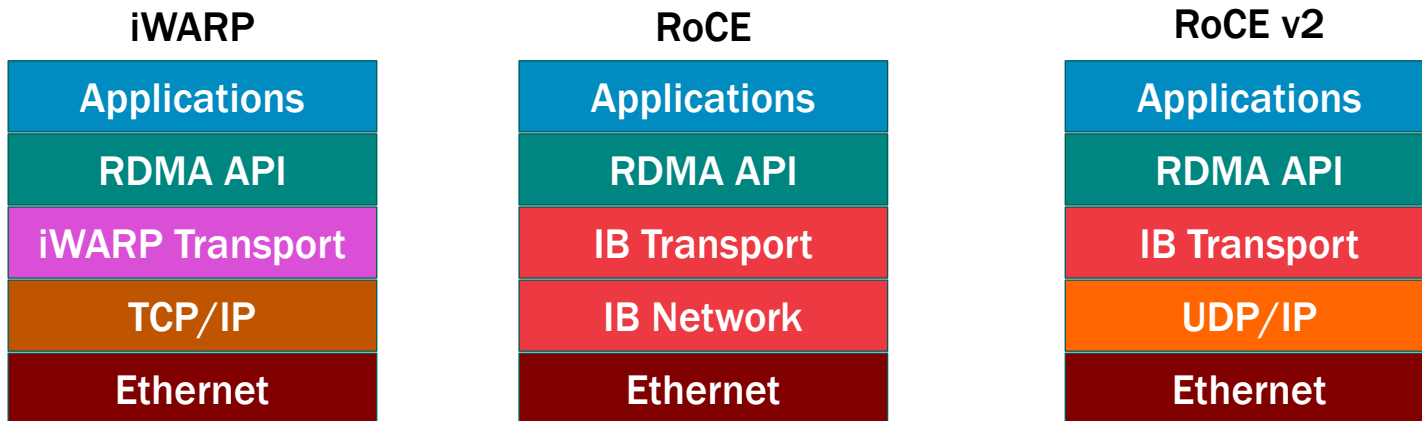
- ❑ InfiniBand (IB) – the default transport protocol
- ❑ Ethernet with RoCE: RDMA over Converged Ethernet
 - ❑ Requires DCB switch (lossless fabric)
- ❑ Ethernet with iWARP: Internet Wide Area RDMA protocol
 - ❑ Runs on top of regular TCP/IP
- ✓ RDMA is available for 10Gb and faster Ethernet technologies

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

iWARP and RoCE

- ❑ 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



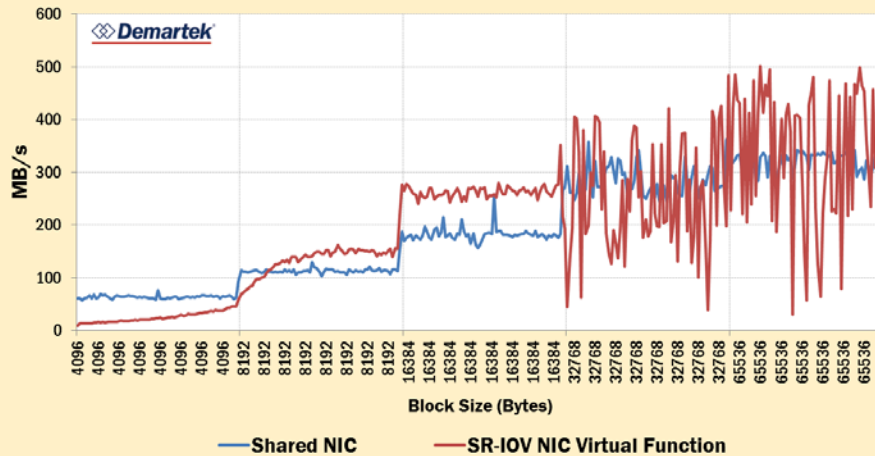


Examples

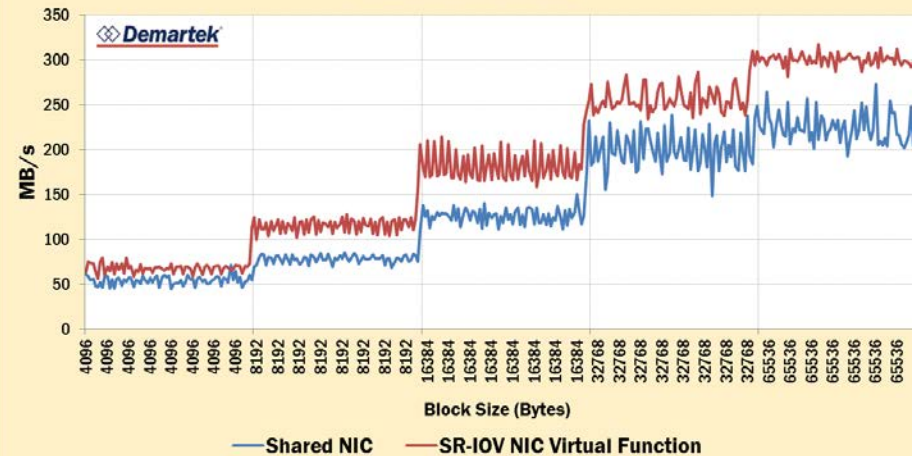


SR-IOV Example – Page 1

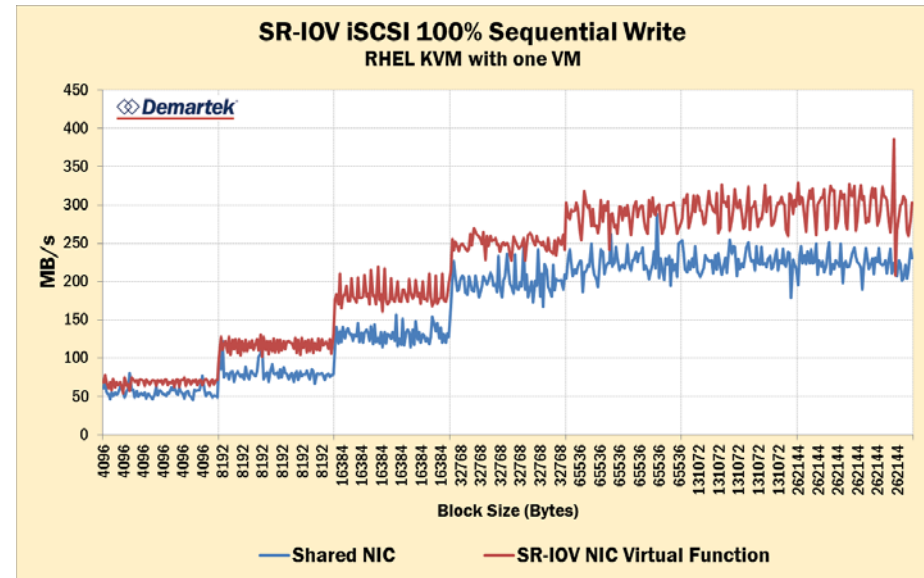
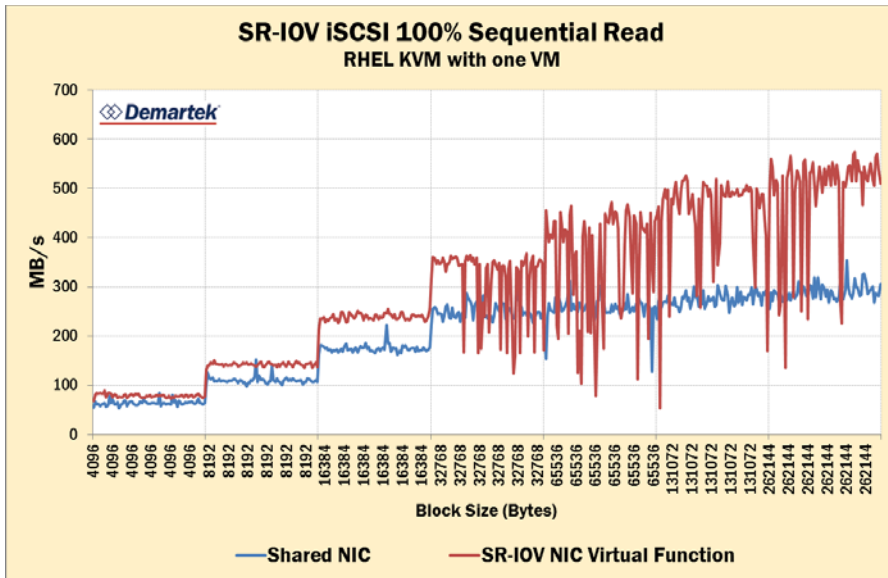
SR-IOV iSCSI 100% Random Read
RHEL KVM with one VM



SR-IOV iSCSI 100% Random Write
RHEL KVM with one VM



SR-IOV Example – Page 2



NPIV Example #1

View from Brocade FC switch Name Server with NPIV ports active

NPIV Column



FDMI Host Name	WWN Compa...	NPIV(or)Virtu...	Host vs. Target	Mem
	Qlogic Corpo...	Physical	Target	
	Qlogic Corpo...	Physical	Target	
	Qlogic Corpo...	Physical	Target	
	Qlogic Corpo...	Physical	Target	
DMRTK-SRVR-J	NetXen, Inc.	Physical	Initiator	
DMRTK-SRVR-J	NetXen, Inc.	Physical	Initiator	
		NPIV	Unknown(initiator/target)	
		NPIV	Unknown(initiator/target)	
		NPIV	Unknown(initiator/target)	
		NPIV	Unknown(initiator/target)	
	Emulex Corp...	Physical	Initiator	
		NPIV	Unknown(initiator/target)	
		NPIV	Unknown(initiator/target)	
		NPIV	Unknown(initiator/target)	
		NPIV	Unknown(initiator/target)	
	Emulex Corp...	Physical	Initiator	

brocade6510-1 - Name Server

☒ Auto Refresh Auto-Refresh Interval: 15 seconds Number of Devices: 17

Domain	User P...	Port ID	Port T...	Device Port WWN	Device Node WWN	Device Name	Capability	FDMI Host Name	WWN Compa...	NPIV(or)V...	Host vs. Target	Mem
1(0x1)	0	0X010000	N	21:00:00:24:ff:38:53:fe	20:00:00:24:ff:38:53:fe	QLE2562 FW-v4 04 04 DVR-v8 02 01-k4-tgt	NS		Qlogic Corpo...	Physical	Target	
1(0x1)	1	0X010100	N	21:00:00:24:ff:38:54:2c	20:00:00:24:ff:38:54:2c	QLE2562 FW-v4 04 04 DVR-v8 02 01-k4-tgt	NS		Qlogic Corpo...	Physical	Target	
1(0x1)	2	0X010200	N	21:00:00:24:ff:38:54:10	20:00:00:24:ff:38:54:10	QLE2562 FW-v4 04 04 DVR-v8 02 01-k4-tgt	NS		Qlogic Corpo...	Physical	Target	
1(0x1)	3	0X010300	N	21:00:00:24:ff:3a:ff:b4	20:00:00:24:ff:3a:ff:b4	QLE2562 FW-v4 04 04 DVR-v8 02 01-k4-tgt	NS		Qlogic Corpo...	Physical	Target	
1(0x1)	4	0X010400	N	21:00:00:0e:fe:08:c8:b1	20:00:00:0e:fe:08:c8:b1	QLE6362 FW-v6 04 00 DVR-v9 1.10.28	NS	DMRTK-SRVR-J	NetXen, Inc.	Physical	Initiator	
1(0x1)	5	0X010500	N	21:00:00:0e:fe:08:c8:b0	20:00:00:0e:fe:08:c8:b0	QLE6362 FW-v6 04 00 DVR-v9 1.10.28	NS	DMRTK-SRVR-J	NetXen, Inc.	Physical	Initiator	
1(0x1)	6	0X010604	N	c0:03:ff:1d:db:5c:00:12	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	6	0X010603	N	c0:03:ff:1d:db:5c:00:10	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	6	0X010605	N	c0:03:ff:1d:db:5c:00:16	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	6	0X010601	N	c0:03:ff:1d:db:5c:00:08	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	6	0X010600	N	10:00:00:00:c9:f8:04:32	20:00:00:00:c9:f8:04:32	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS		Emulex Corp...	Physical	Initiator	
1(0x1)	6	0X010602	N	c0:03:ff:1d:db:5c:00:0c	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	7	0X010702	N	c0:03:ff:1d:db:5c:00:0e	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	7	0X010703	N	c0:03:ff:1d:db:5c:00:14	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	7	0X010704	N	c0:03:ff:1d:db:5c:00:18	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	7	0X010701	N	c0:03:ff:1d:db:5c:00:0a	c0:03:ff:00:00:ff:ff:00	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS			NPIV	Unknown(initiator/target)	
1(0x1)	7	0X010700	N	10:00:00:00:c9:f8:04:33	20:00:00:00:c9:f8:04:33	Emulex 81Y1864 FV1.0.11.108 DV2.72.012.001 ...	NS		Emulex Corp...	Physical	Initiator	

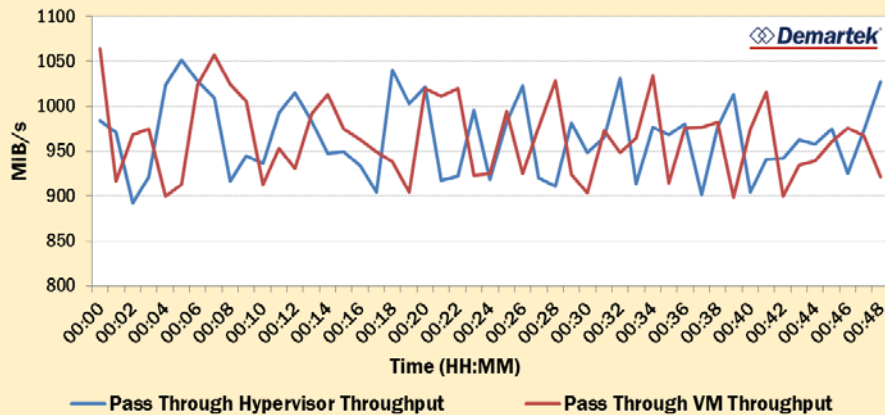
Refreshed : 3:30:35 PM

Free Professional Management Tool brocade6510-1 FID 128 User: admin Role: admin

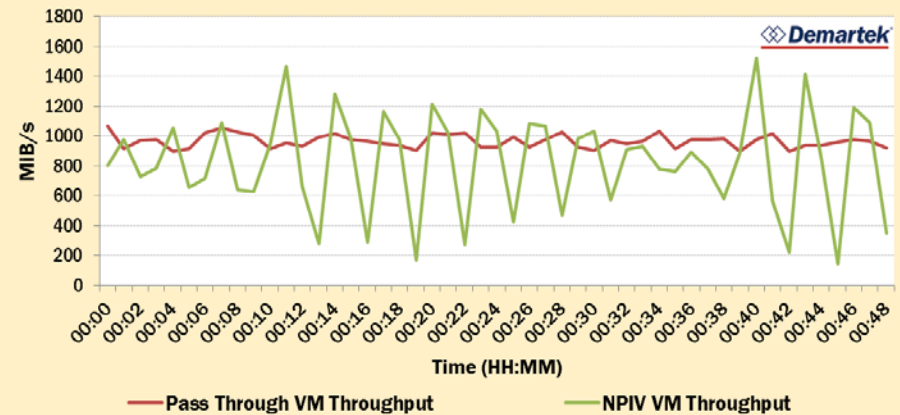
NPIV Example #2

► 16GFC Hyper-V Test Comparing “Pass Through” vs NPIV

Pass-Through Throughput at VM and Hypervisor



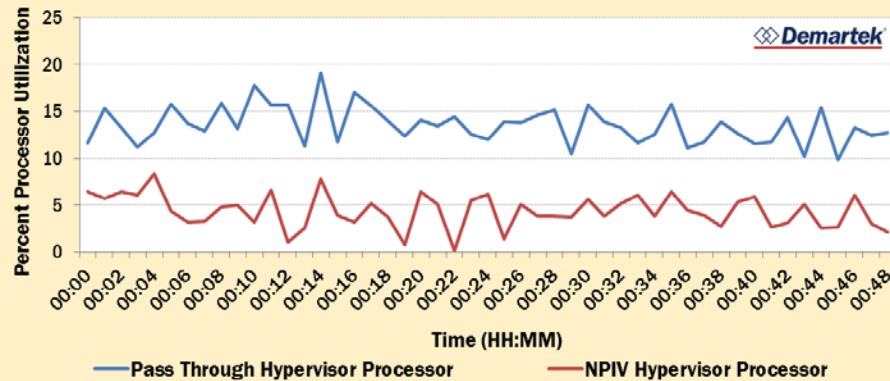
Pass-Through vs NPIV Throughput at VM



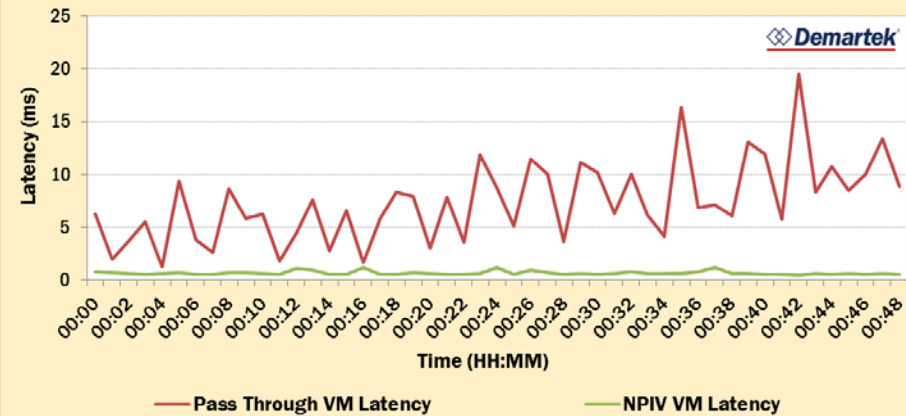
NPIV Example #2

► 16GFC Hyper-V Test Comparing “Pass Through” vs NPIV

**Hypervisor Processor Utilization
Pass Through vs NPIV**



Pass-Through vs NPIV Latency at VM



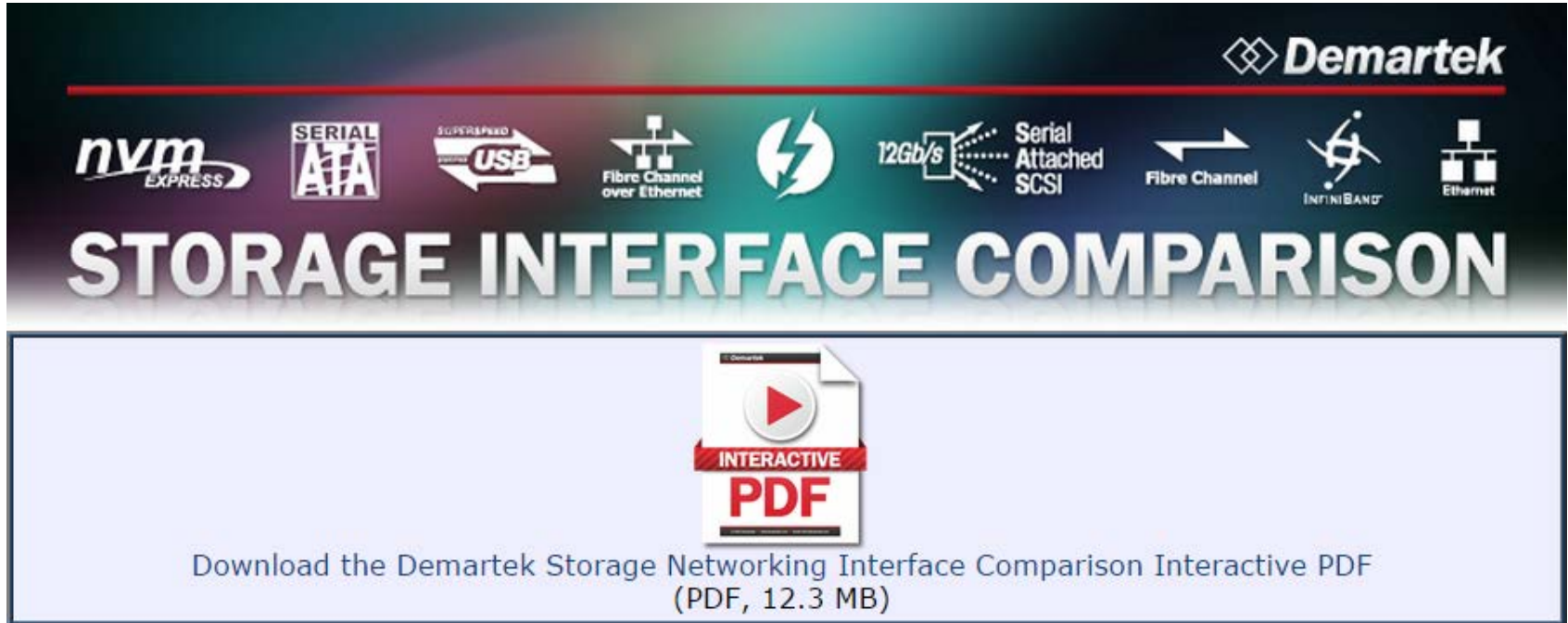
Demartek Testing of RDMA Technologies

- ❑ We are currently testing the performance of some of these RDMA technologies
 - ❑ RoCE
 - ❑ iSER
- ❑ Demartek is developing the ***RoCE Deployment Guide*** that will be published this summer
 - ❑ Will include technology from several vendors
 - ❑ Will include 10GbE, 25GbE, 40GbE and 100GbE

Future Possibilities

- ❑ Look for more solutions that support RDMA for:
 - ❑ File servers: SMB Direct (Windows) and NFS/RDMA (Linux)
 - ❑ Block storage: iSER (iSCSI) and NVMe over Fabrics

Storage Interface Comparison



The banner features the Demartek logo in the top right corner. Below it, a row of storage interface logos is displayed: nvm EXPRESS, SERIAL ATA, SUPER SPEED USB, Fibre Channel over Ethernet, 12Gb/s Serial Attached SCSI, Fibre Channel, INFINIBAND, and Ethernet. The central text reads "STORAGE INTERFACE COMPARISON" in large, bold, white letters. Below this, there is a button with a play icon and the text "INTERACTIVE PDF". At the bottom, a text box contains the link "Download the Demartek Storage Networking Interface Comparison Interactive PDF (PDF, 12.3 MB)".

STORAGE INTERFACE COMPARISON

Download the Demartek Storage Networking Interface Comparison Interactive PDF
(PDF, 12.3 MB)

- ❑ HTML and downloadable interactive PDF version available
- ❑ Search engine: “storage interface comparison”
- ❑ www.demartek.com/Demartek_Interface_Comparison.html

24

Demartek Free Resources

- ❑ Demartek SSD Zone
www.demartek.com/SSD
- ❑ Demartek iSCSI Zone
www.demartek.com/iSCSI
- ❑ Demartek Fibre Channel 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 public projects and materials are announced on a variety of social media outlets. Follow us on any of the above.



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