

**E2EVC 2015 BERLIN**

**June 12<sup>th</sup>-14<sup>th</sup>**



# **SMB Direct, The Secret Decoder Ring**



# Didier Van Hoyer

Technical Architect & Technology Strategist



Microsoft MVP in Hyper-V



MEET Member



DELL TechCenter Rockstar



<http://workinghardinit.wordpress.com>



[@workinghardinit](https://twitter.com/workinghardinit)



*Why?*

*Who?*

*When?*

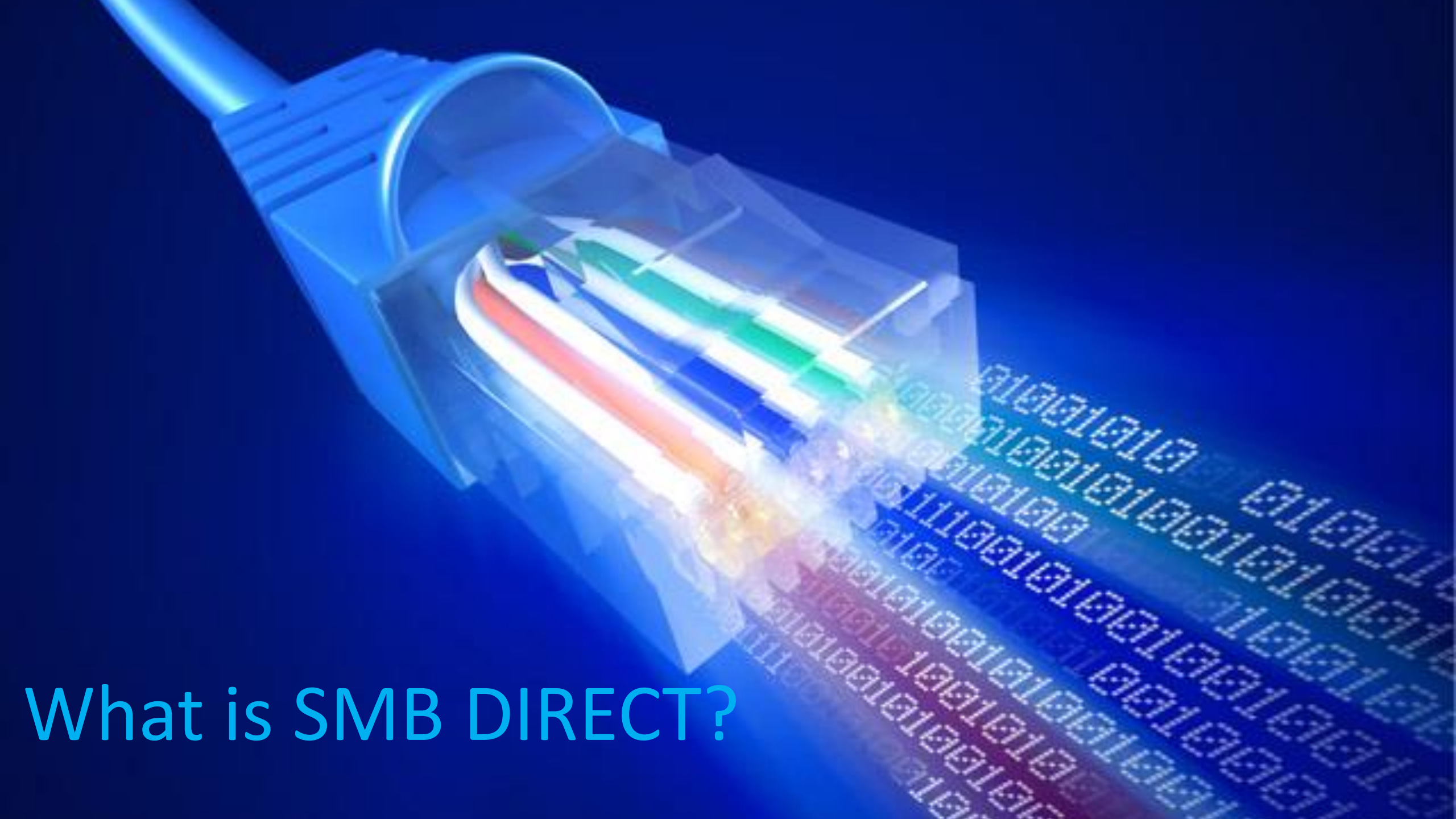
*WHERE?*

*How?*

*What?*







What is SMB DIRECT?

# SMB Direct is SMB over RDMA

## So what is RDMA?

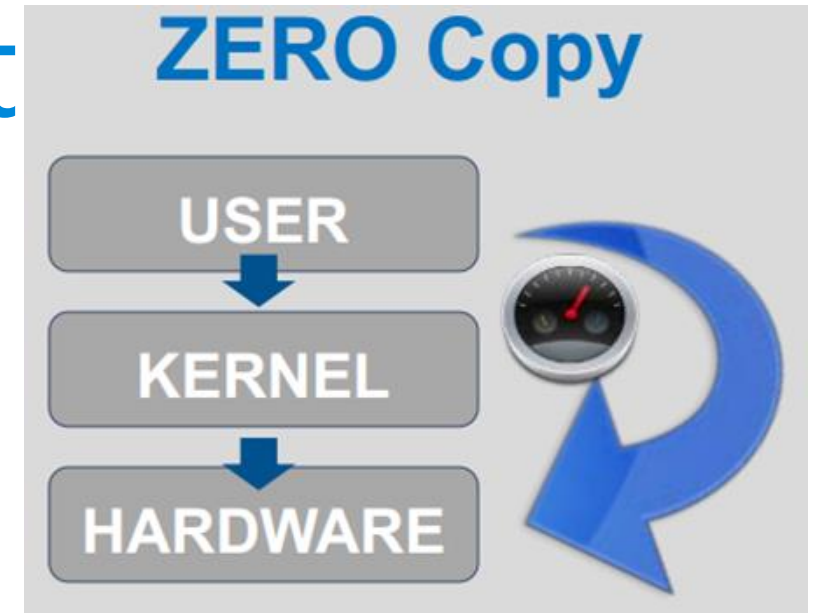
- Direct Memory Access (DMA) allows access (read/write) to the host memory directly, without the intervention of the CPU(s).
- Remote Direct Memory Access (RDMA) extends this ability to remote systems.



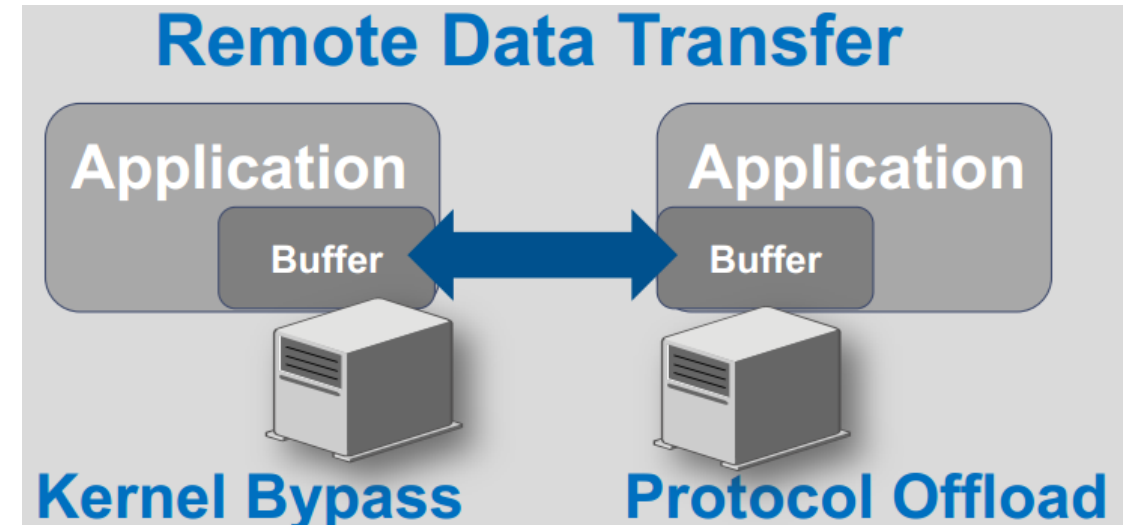
What makes RDMA “great”?

# What makes RDMA "great"

Zero-Copy: applications can transfer data from one host to another without use of the software network stack. Data is being read & written, sent/received directly to memory buffers without being copied over the different network layers.



Kernel Bypass: applications can perform data transfer directly from user space without the need to perform context switches over the kernel.





# What makes RDMA "great"?

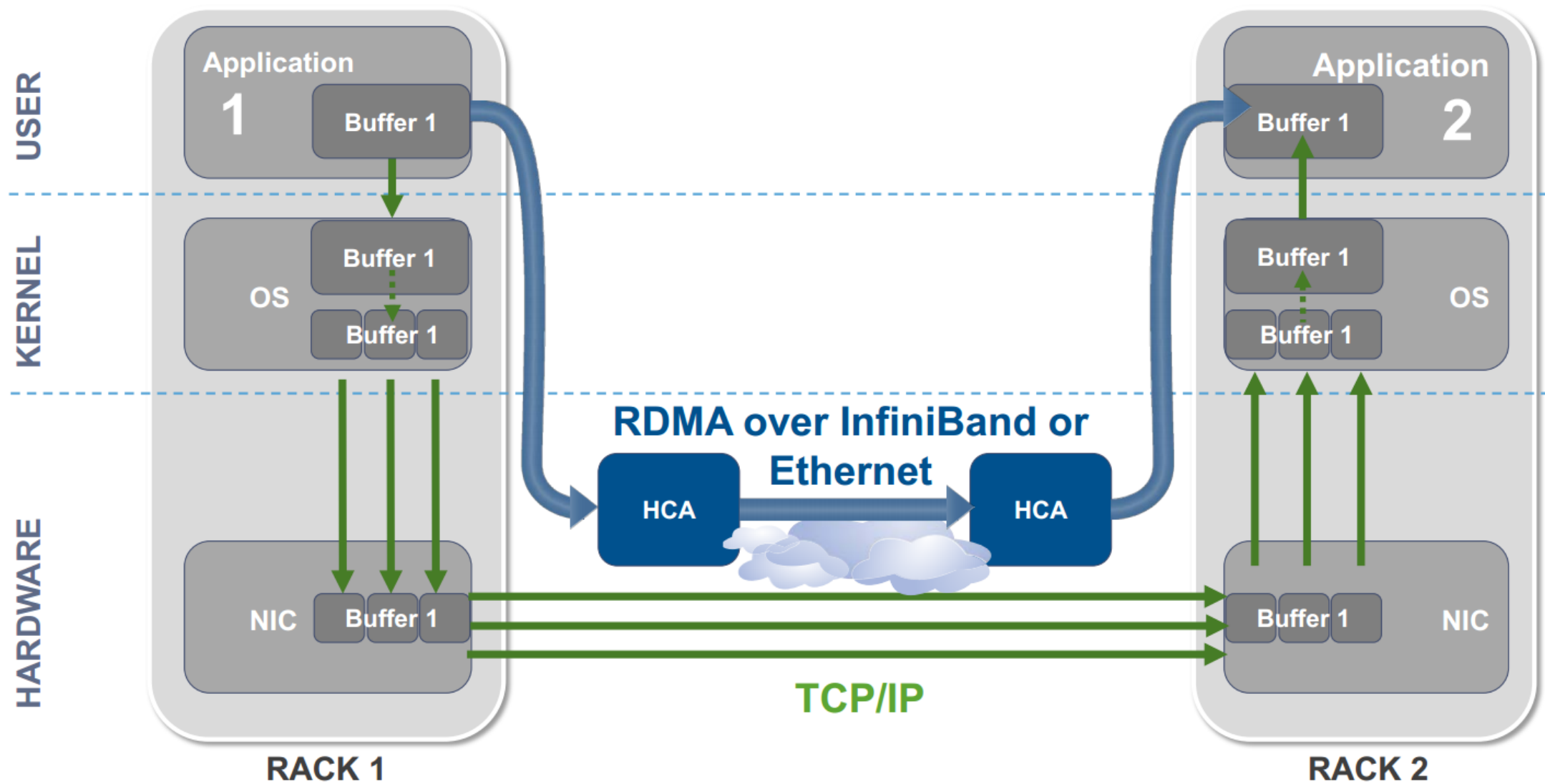
## CPU Bypass / Offload

- Applications can access remote memory without consuming any host CPU cycles in the remote machine. The remote memory machine will be read/written without involving remote processes (meaning CPU cycles).
- The caches in the remote CPU(s) won't be filled with the accessed memory content.

## Transport protocol acceleration

- Message based transactions (TCP is stream based).
- Scatter/gather entries support (reading multiple memory buffers and sending them as one or reading one and writing it to multiple memory buffers).







Q:Where did RDMA grow up?

# A: Where the benefits were worth the premium price

1. Low latency
  2. High throughput
  3. Reduced CPU footprint
  4. **“lossless” fabrics** (you have to provide this)
- HPC, Financial transactions, medical imaging, storage, backup, cloud computing, ...





Why do I care?

# Why is RDMA important to us?

1. The BIG surge in East-West traffic due to virtualization induces a performance & scalability hit.
2. Storage over file system shares is also highly demanding.

In other words not using RDMA:

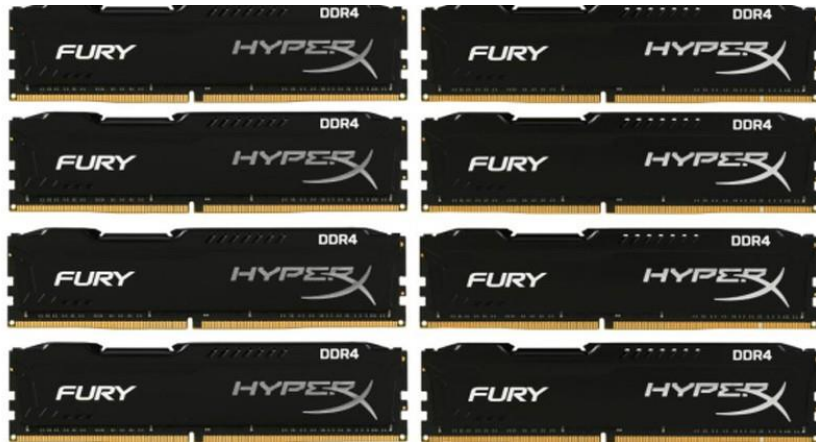
- Slows us down
- Loses us money (can't leverage capabilities of hard & software)

# Modern commodity network hardware





# A modern cloud OS, servers & storage



Didier Van Hoya - WorkingHardInIT

# Flavors of RDMA

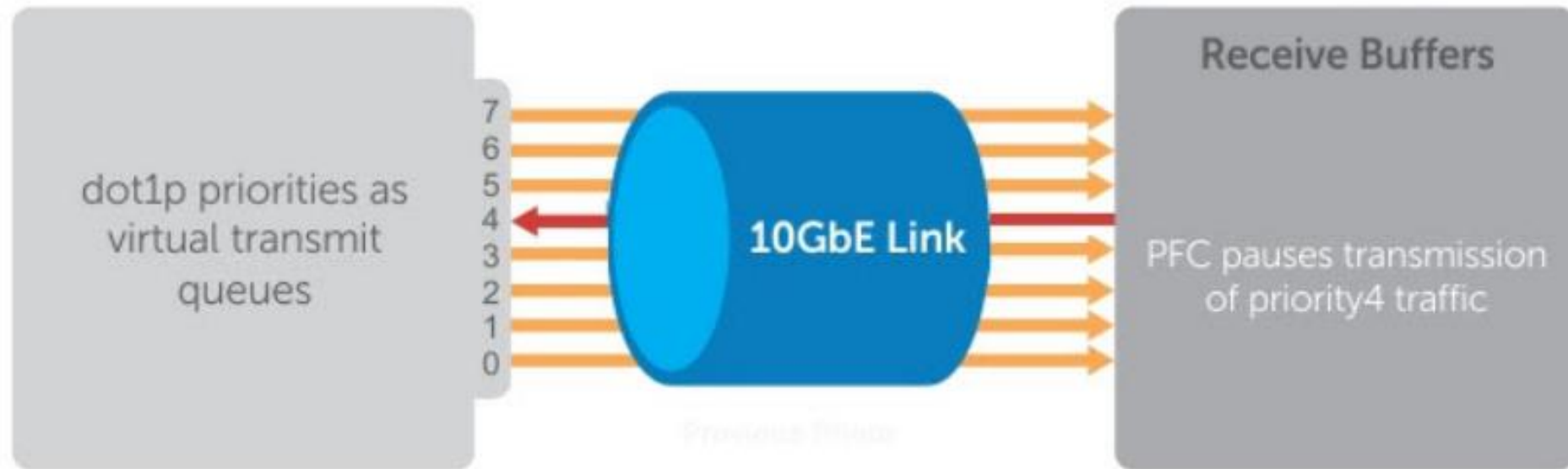
Type (Cards*)	Pros		Cons
<b>Non-RDMA Ethernet</b> (wide variety of NICs)	<ul style="list-style-type: none"> <li>TCP/IP-based protocol</li> <li>Works with any Ethernet switch</li> <li>Wide variety of vendors and models</li> <li>Support for in-box NIC teaming</li> </ul>		<ul style="list-style-type: none"> <li>High CPU Utilization under load</li> <li>High latency</li> </ul>
<b>iWARP</b> (Chelsio T4/T5)	Low CPU Utilization under load Low latency	<ul style="list-style-type: none"> <li>TCP/IP-based protocol</li> <li>Works with any Ethernet switch</li> <li>RDMA traffic routable</li> <li>Offers up to 40Gbps per NIC port today</li> </ul>	<ul style="list-style-type: none"> <li>Requires enabling firewall rules</li> </ul>
<b>RoCE</b> (Mellanox ConnectX-3, Mellanox ConnectX-3Pro, Mellanox ConnectX-4)		<ul style="list-style-type: none"> <li>Ethernet-based protocol</li> <li>Works with Ethernet switches with DCB support</li> <li>Routable RoCEv2 is here</li> <li>Offers up to 100Gbps per NIC port today</li> </ul>	<ul style="list-style-type: none"> <li>RoCEv1 non routable</li> <li>Requires DCB switch with Priority Flow Control (PFC)</li> </ul>
<b>InfiniBand</b> (Mellanox ConnectX-3, Mellanox ConnectX-3Pro, Mellanox ConnectX-4, Mellanox Connect-IB)		<ul style="list-style-type: none"> <li>Switches typically less expensive per port</li> <li>Switches offer high speed Ethernet uplinks</li> <li>Commonly used in HPC environments</li> <li>Offers up to 100Gbps per NIC port today</li> </ul>	<ul style="list-style-type: none"> <li>Not an Ethernet-based protocol</li> <li>RDMA traffic not routable via IP infrastructure</li> <li>Requires InfiniBand switches</li> <li>Requires a subnet manager (typically on the switch)</li> </ul>

# What is DCB?

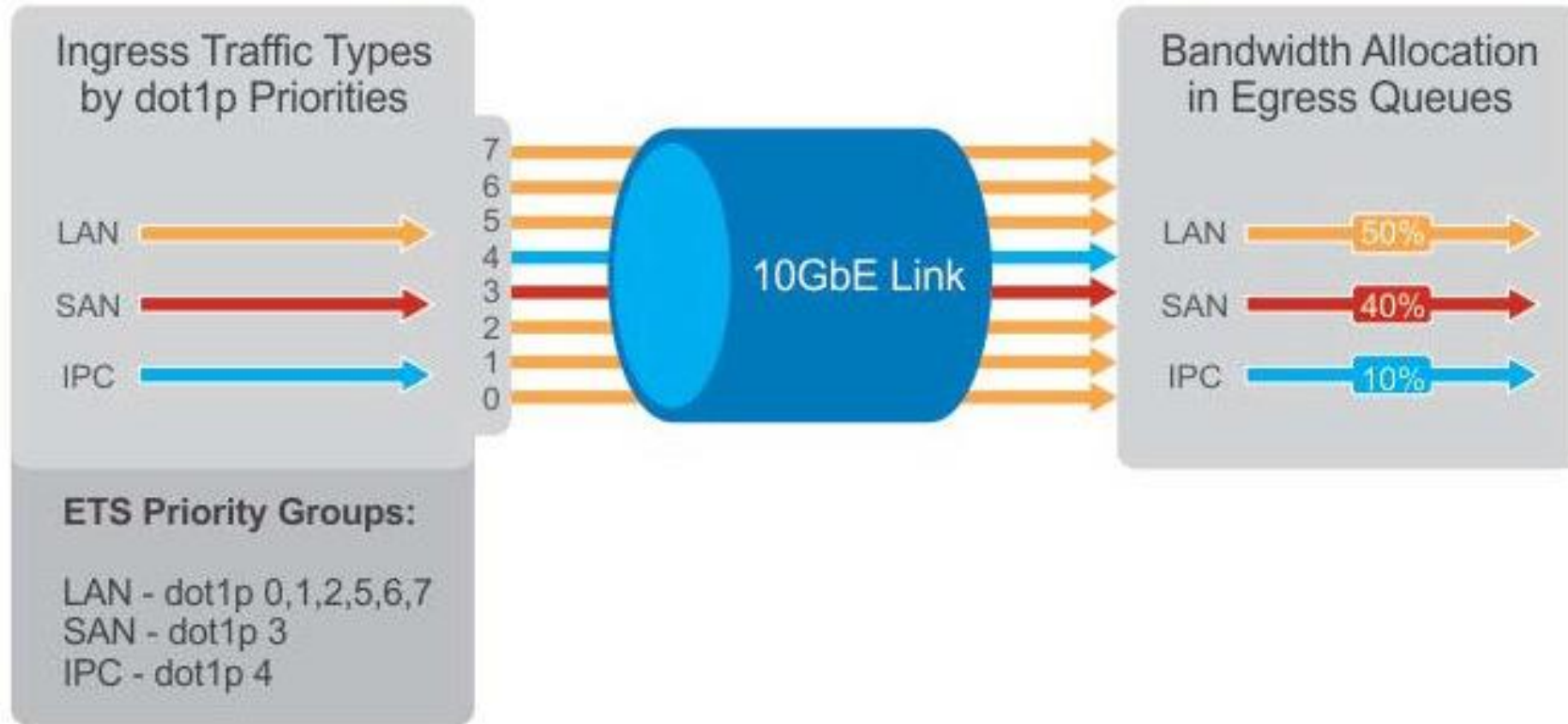
- **Priority-based Flow Control (PFC).** Provides a link-level, flow-control mechanism that can be independently controlled for each priority to ensure zero-loss due to converged-network congestion.
- Quantified Congestion Notification (QCN). Provides end-to-end congestion management for protocols without built-in congestion-control mechanisms. It's also expected to benefit protocols with existing congestion management by providing more timely reactions to network congestion.
- **Enhanced Transmission Selection (ETS).** Provides a common management framework for bandwidth assignment to traffic classes.
- Data Center Bridging Exchange Protocol (DCBx). A discovery and capability exchange protocol used to convey capabilities and configurations of the other three DCB features between neighbors to ensure consistent configuration across the network.



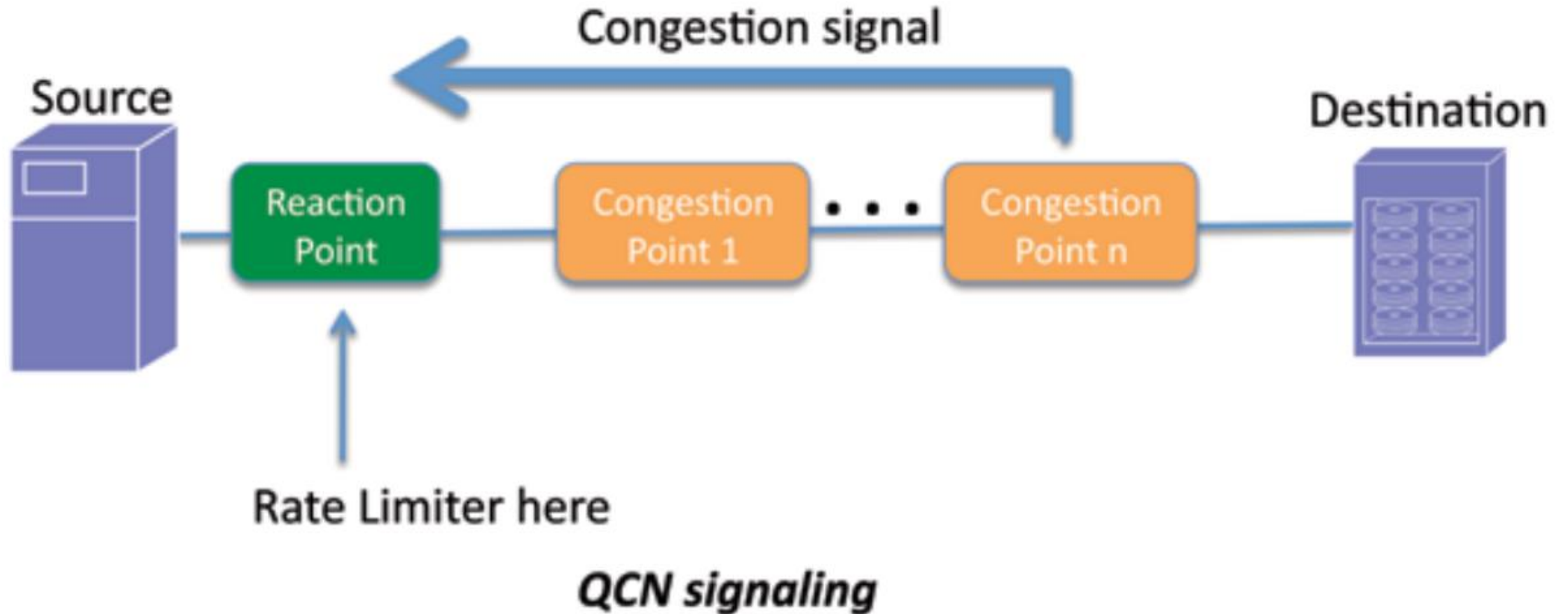
# How does PFC work?



# How does ETS work?



# What about QCN?





# What about QCN?

- Missing in Windows & a lot of switches for now 😊
- Complexity to achieve this across all hops & end to end
- Is the lack of end to end congestion notification an issue?
- End to end might not be needed (hop based in FC – ingress rate limiting - does the job, doesn't work in FCoE due to FCF/Layer 3 MAC rewriting ) ...
- Apparently not (you won't hear anyone put that in writing) as RoCE v2 is now being used in public clouds without problems ...
- Some state the industry silicon is ready but ... the standard/implementation has to come to many switches still...
- Some say it's a worry, it depends on the design & protocol
- Devil in details= layer 2 so not routable ;-)

# What about DCBX?

- Missing in Windows for now.
- It's a convenience issue solved by automation of your DCB configuration (PowerShell).
- But convenience is important and I expect Microsoft to look into this and possibly provide it in future versions of Windows.
- The benefit is that the DCB configuration is learned from the switches.

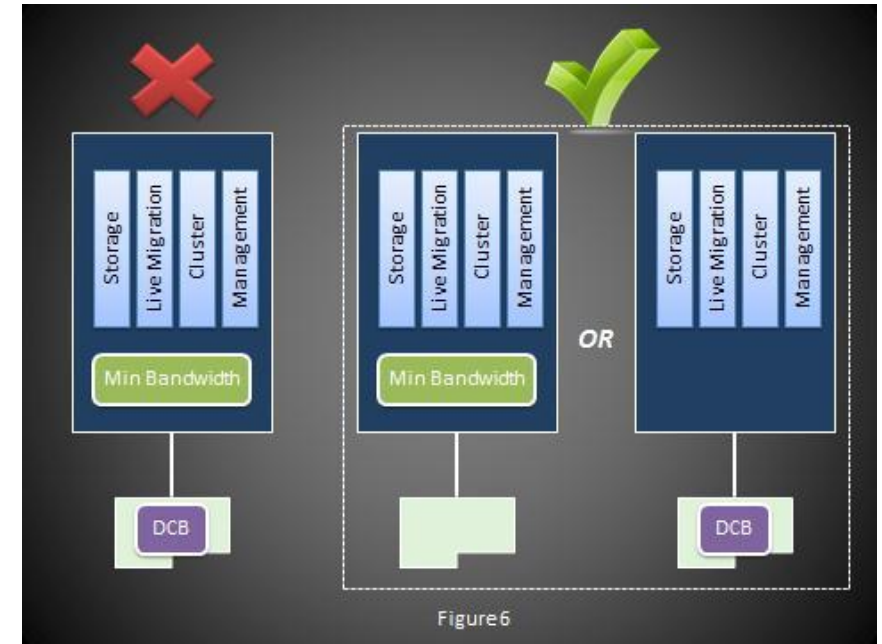
# How do I configure DCB?





# Configure OS, Application & Network

- Don't mix QoS types
- Configure QoS for all payloads
- You need rNICs (RDMA capable NICs)
- Configure your switches end to end (ports, uplinks)
- Configure your host & rNICs
- You must do PFC (this requires **tagged VLANs** on hosts & switches)
- You can do ETS (harder to test & make work in heterogeneous environments)



```

#Install DCB on the hosts
Install-WindowsFeature Data-Center-Bridging
#Mellanox RoCE drivers don't support DCBx, disable it.
Set-NetQosDcbxSetting -willing $False
#Make sure RDMA is enable on the NIC (should be by default)
Enable-NetAdapterRdma -Name RDMA-NIC1
Enable-NetAdapterRdma -Name RDMA-NIC2
#Start with a clean slate
Remove-NetQosTrafficClass -confirm:$False
Remove-NetQosPolicy -confirm:$False

#Tag the RDMA NIC with the VLAN chosen for PFC network
Set-NetAdapterAdvancedProperty -Name "RDMA-NIC-1" -RegistryKeyword "VlanID" -RegistryValue 110
Set-NetAdapterAdvancedProperty -Name "RDMA-NIC-2" -RegistryKeyword "VlanID" -RegistryValue 120

#SMB Direct traffic to port 445 is tagged with priority 4
New-NetQosPolicy "SMBDIRECT" -netDirectPortMatchCondition 445 -PriorityValue8021Action 4
#Anything else goes into the "default" bucket with priority tag 1 :-)
New-NetQosPolicy "DEFAULT" -default -PriorityValue8021Action 1

#Enable PFC (lossless) on the priority of the SMB Direct traffic.
Enable-NetQosFlowControl -Priority 4
#Disable PFC on the other traffic (TCP/IP, we don't need that to be lossless)
Disable-NetQosFlowControl 0,1,2,3,5,6,7

#Enable QoS on the RDMA interface
Enable-NetAdapterQos -InterfaceAlias "RDMA-NIC1"
Enable-NetAdapterQos -InterfaceAlias "RDMA-NIC2"

#Set the minimum bandwidth for SMB Direct traffic to 90% (ETS, optional)
New-NetQosTrafficClass "SMBDirect" -Priority 4 -Bandwidth 90 -Algorithm ETS

```

# Disable 802.3x flow control (global pause)

```
FTOS#configure
```

```
FTOS(conf)#interface range tengigabitethernet 0/0 -47
```

```
FTOS(conf-if-range-te-0/0-47)#no flowcontrol rx on tx on
```

```
FTOS(conf-if-range-te-0/0-47)#exit
```

```
FTOS(conf)#interface range fortyGigE 0/48 , fortyGigE 0/52
```

```
FTOS(conf-if-range-fo-0/48-52)#no flowcontrol rx on tx off
```

```
FTOS(conf-if-range-fo-0/48-52)#exit
```



# Enable DCB & Configure VLANs

```
FTOS(conf)#service-class dynamic dot1p
FTOS(conf)#dcb enable
FTOS(conf)#exit
FTOS#copy running-config startup-config
FTOS#reload
```

```
FTOS#configure
FTOS(conf)#interface vlan 110
FTOS (conf-if-vl-vlan-id*)#tagged tengigabitethernet 0/0-47
FTOS (conf-if-vl-vlan-id*)#tagged port-channel 3
FTOS (conf-if-vl-vlan-id*)#exit
```

# Create & configure DCB Map Policy

```
FTOS(conf)#dcb-map SMBDIRECT
```

```
FTOS(conf-dcbmap-profile-name*)#priority-group 0  
bandwidth 90 pfc on
```

```
FTOS(conf-dcbmap-profile-name*)#priority-group 1  
bandwidth 10 pfc off
```

```
FTOS(conf-dcbmap-profile-name*)#priority-pgid 1 1 1 1  
0 1 1 1
```

```
FTOS(conf-dcb-profile-name*)#exit
```

PFC => IEEE 802.1Qb

# Apply DCB map to the switch ports & uplinks

```
FTOS(conf)#interface range ten 0/0 – 47
FTOS(conf-if-range-te-0/0-47)#dcb-map SMBDIRECT
FTOS(conf-if-range-te-0/0-47)#exit
FTOS(conf)#interface range fortyGigE 0/48 , fortyGigE 0/52
FTOS(conf-if-range-fo-0/48,fo-0/52)#dcb-map SMBDIRECT
FTOS(conf-if-range-fo-0/48,fo-0/52)#exit
FTOS(conf)#exit
FTOS#copy running-config startup-config
```



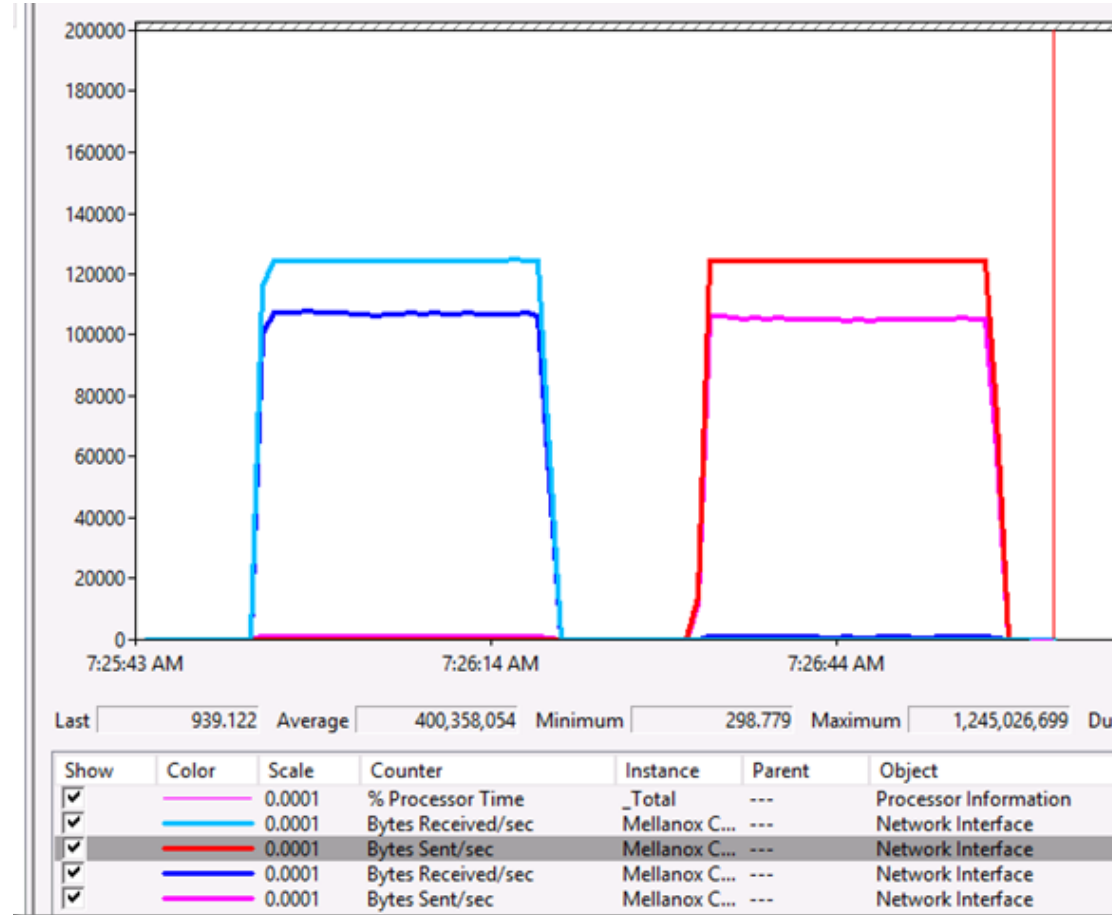
# Demo Time

# Jumbo Frames?

- Infiniband?
  - Exists, MTU Size up to 4K.
  - Impact on SMB Direct?
- RoCE?
  - Yes, familiar 1500 MTU size up to 9K (or more depending on your NIC/Switches).
  - Impact on SMB Direct?
  - Also useful if you fall back to non RDMA (TCP/IP)
- iWarp?
  - Yes, familiar 1500 MTU size up to 9K (or more depending on your NIC/Switches)
  - Impact on SMB Direct?
  - Also same as above, handy during fail back to non RDMA (TCP/IP)!

<https://workinghardinit.wordpress.com/2013/11/25/live-migration-can-benefit-from-jumbo-frames/>

# Jumbo Frames?



<https://workinghardinit.wordpress.com/2013/11/25/live-migration-can-benefit-from-jumbo-frames/>



# Configure SMB Direct On The Windows Host

- Infiniband?
  - <https://technet.microsoft.com/en-us/library/dn583823.aspx>
- RoCE?
  - <https://technet.microsoft.com/en-us/library/dn583822.aspx>
- iWarp?
  - <https://technet.microsoft.com/en-us/library/dn583825.aspx>

# Prevent live migration starving storage traffic

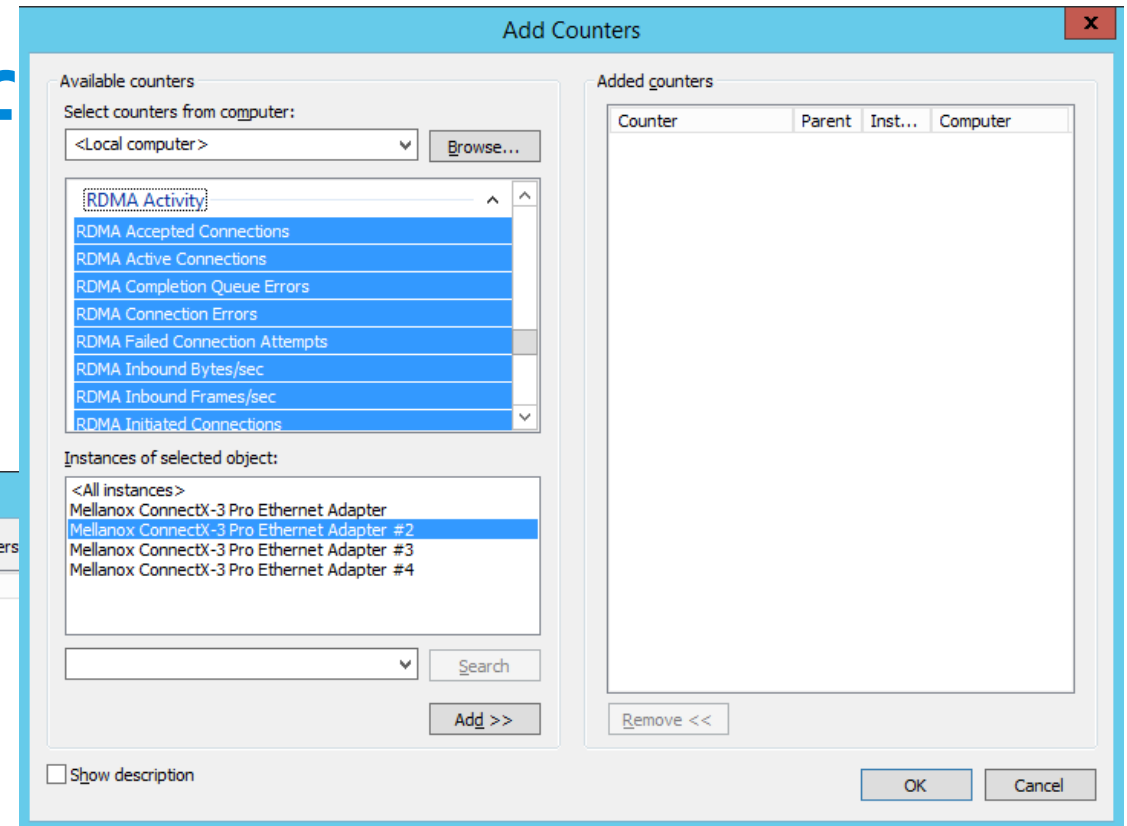
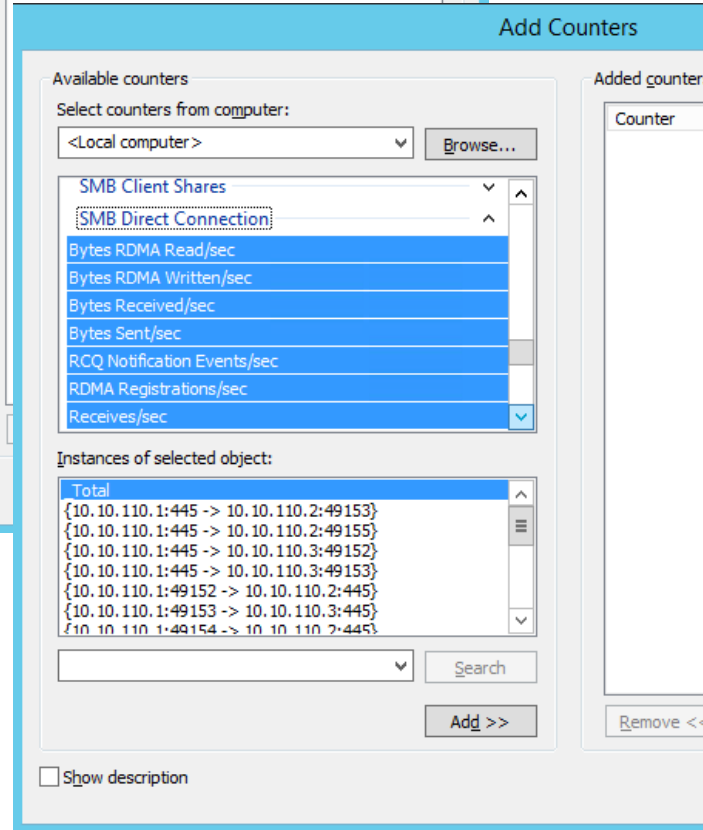
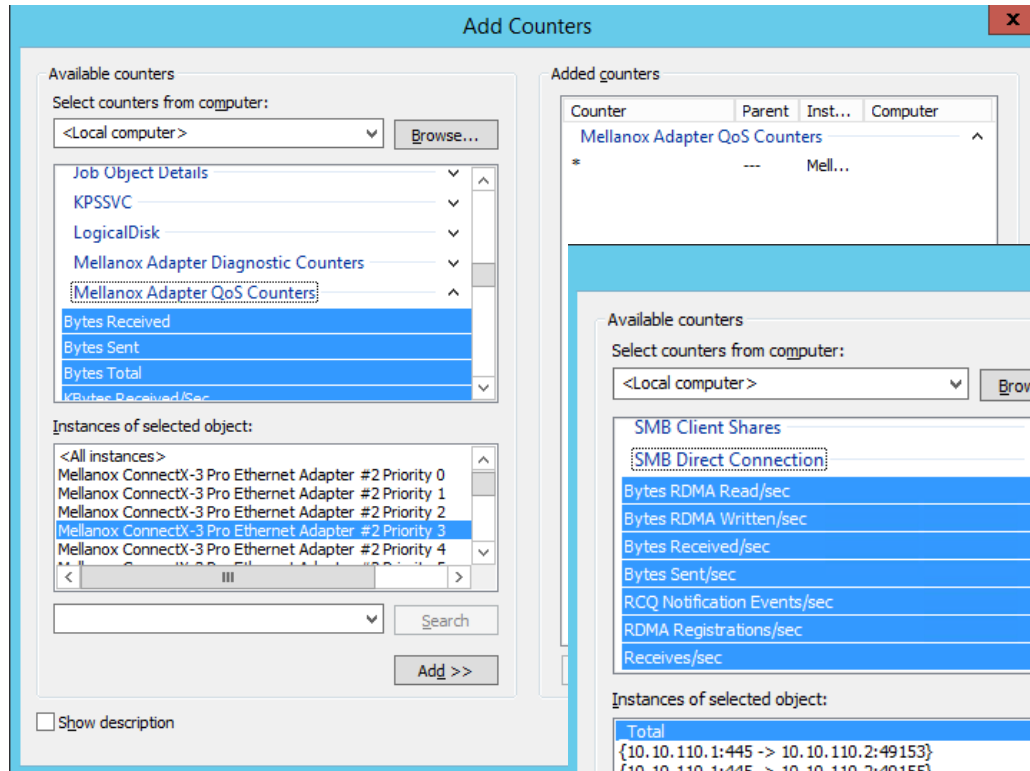
Modules:	SmbShare
Name:	SMBBand
<div>Get-SmbBandWidthLimit Remove-SmbBandwidthLimit Set-SmbBandwidthLimit</div>	

***Set-SmbBandwidthLimit -Category LiveMigration -BytesPerSecond 10GB***

***Set-SmbBandwidthLimit -Category VirtualMachine -BytesPerSecond 10GB***

<https://workinghardinit.wordpress.com/2013/09/03/preventing-live-migration-over-smb-starving-csv-traffic-in-windows-server-2012-r2-with-set-smbbandwidthlimit/>

# PerfMon is your f



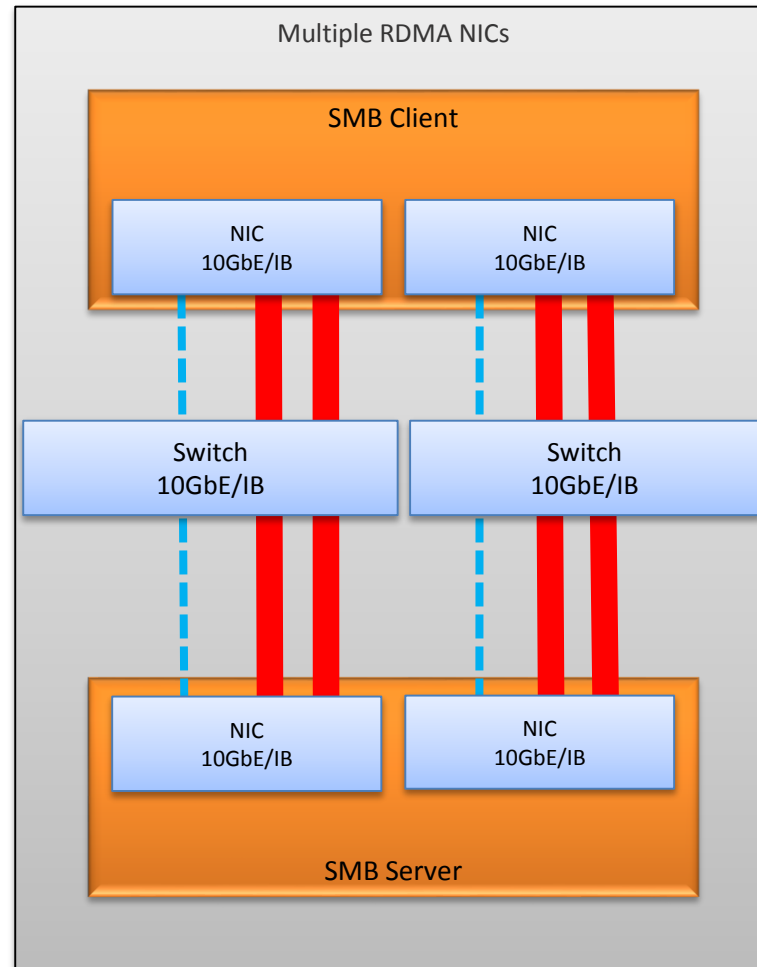
# SMB Multichannel & Direct “Break” sometimes

- Enabling/Disabling RDMA
- Enabling/Disabling Multichannel
- Enabling/Disabling the NICs
- In rare case a firmware upgrade to the switches can require a reboot of the host(s)

**Keep the above in mind when testing & experimenting or trouble shouting. Always start from a clean, known state.**



# SMB Direct requires SMB Multichannel



# BIOS Power Optimizations still apply!

- Disable C states in BIOS / UEFI
- Disable C1E states in BIOS / UEFI
- Disable PCIe Link Power Management in BIOS, basically set all power optimizations to max performance
- Optimize memory for speed
- The faster the cards & bus speeds the better ...

<https://workinghardinit.wordpress.com/2013/06/10/still-need-to-optimizing-power-settings-on-dell-12th-generation-servers-for-lightning-fast-hyper-v-live-migrations/>

**Cat.7**



# The Agony of Choice ...

- Infiniband will be around for a long time (just like FC)
  - What will they'll do to offset Ethernet speed growth?
    - strategy seems RoCE but iWarp is giving them a fight for their money.
- Ethernet is likely to grow fast in
  - New deployments (no infiniband in place)
  - > 10Gbps deployments only if price/Gbps drops low enough compared to Infiniband, that's where Mellanox is outperforming Chelsio (cheap swithes & cards),
- What Flavor Should I use?
  - RoCE has shown great potential as the official heir to infiniband but must address
    - concerns around real life loss less routability
    - complexity of DCB (PFC/ETS) in a heterogeneous world
  - iWarp has the advantage of TCP/IP routability & ease of deployment but must
    - address concerns around need of DCB configurations in larger / high performance deployment & the overhead of relying on the TCP/IP stack
  - Remember ... there are successful SMB 3 / SOFS / Storage Spaces deployments out there without SMB Direct ... (it all depends) but build for the future ...