14th ANNUAL WORKSHOP 2018

WINDOWS RDMA (ALMOST) EVERYWHERE

Omar Cardona

Microsoft

[April 2018]

AGENDA

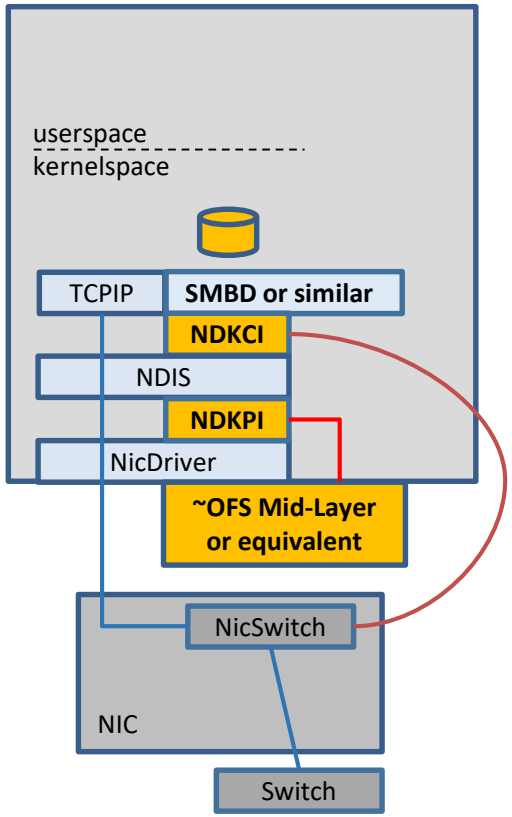
- **How do we use RDMA?**
  - Network Direct
- **Where do we use RDMA?**
  - Client, Server, Workstation, etc.
  - Private Cloud, Public Cloud
- **Internals**
  - Modes, Aggregation, Port Configs
- **Challenges**
  - Silos
  - ECMP/LACP
  - Diagnosis
- **Future Work**
  - Secure Multi-Tenant RDMA VFs

HOW DO WE USE RDMA?

NETWORK DIRECT

Kernel Provider Interface

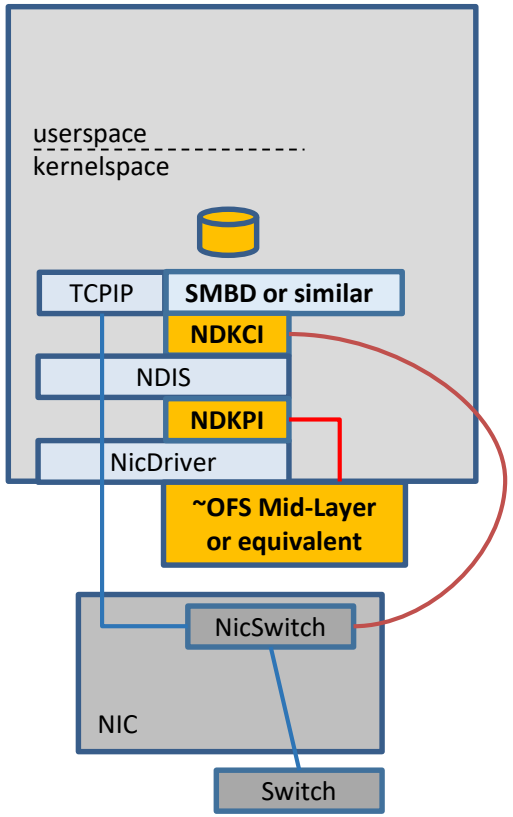
- **ND Kernel Provider Interface**
  - NDKPI
    - Enforceable API definition for Kernel Verbs
  - RC only, !qp_modify, !atomics
  - Agnostic of iWARP, RoCE, IB, etc.



NETWORK DIRECT

Kernel Provider Interface

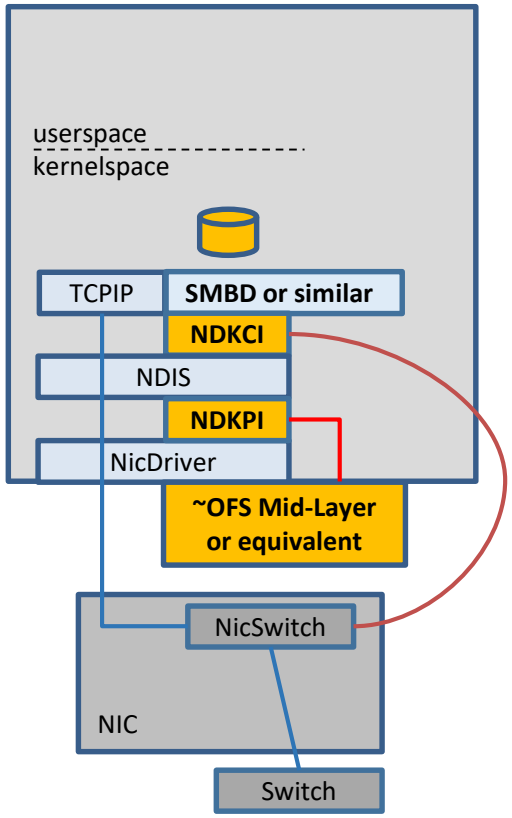
- **ND Kernel Provider Interface**
  - NDKPI
    - Enforceable API definition for Kernel Verbs
  - RC only, !qp_modify, !atomics
  - Agnostic of iWARP, RoCE, IB, etc.
- **Drivers?**
  - Mid-Layer implemented in Drivers
  - Available from all major IHVs



NETWORK DIRECT

Kernel Provider Interface

- **ND Kernel Provider Interface**
  - NDKPI
    - Enforceable API definition for Kernel Verbs
  - RC only, !qp_modify, !atomics
  - Agnostic of iWARP, RoCE, IB, etc.
- **Drivers?**
  - Mid-Layer implemented in Drivers
  - Available from all major IHVs
- **Validation?**
  - HLK - HW Logo Kit
  - PCS – Private Cloud Stress

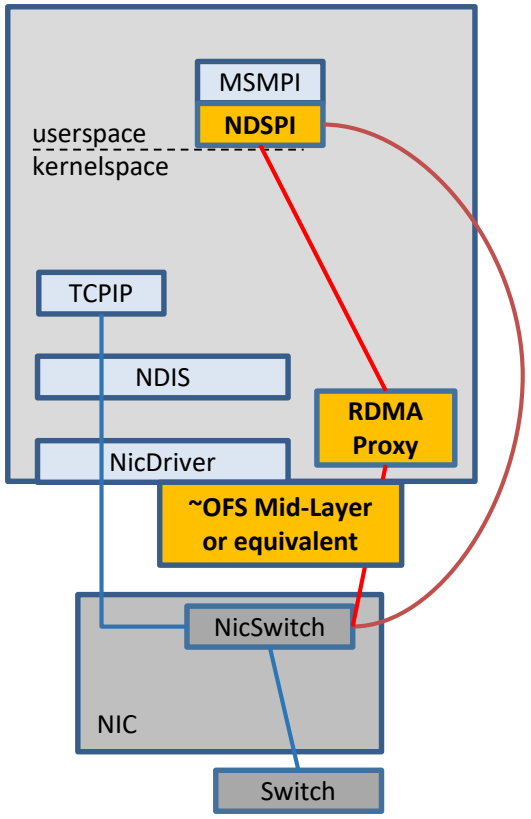


In Windows driver model IHVs are free to implement RDMA support and scope via WinOF, Derivation thereof, or custom software. The HLK and PCS validation vehicles are the mechanism by which functionality and support are tested both in Windows labs and IHV labs.

NETWORK DIRECT

Service Provider Interface

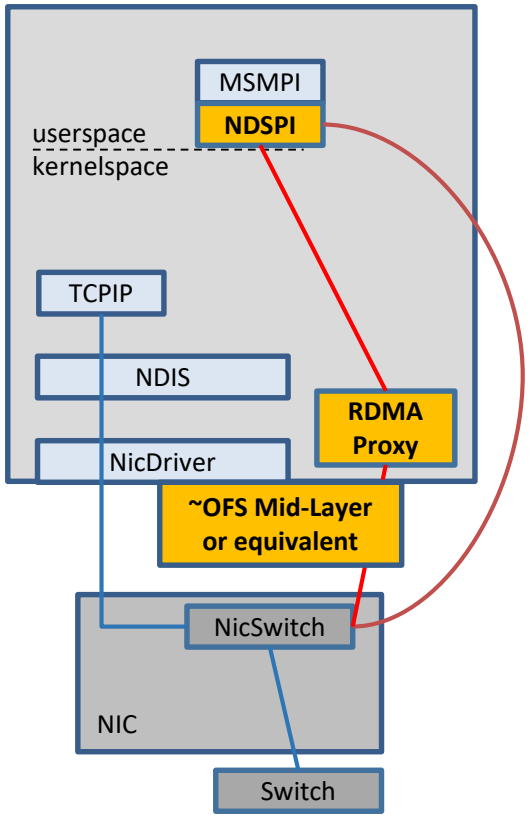
- **ND Service Provider Interface**
  - NDSPI
    - Enforceable API definition for User Verbs
  - IB is primary scenario
  - Agnostic of iWARP, RoCE, or IB



NETWORK DIRECT

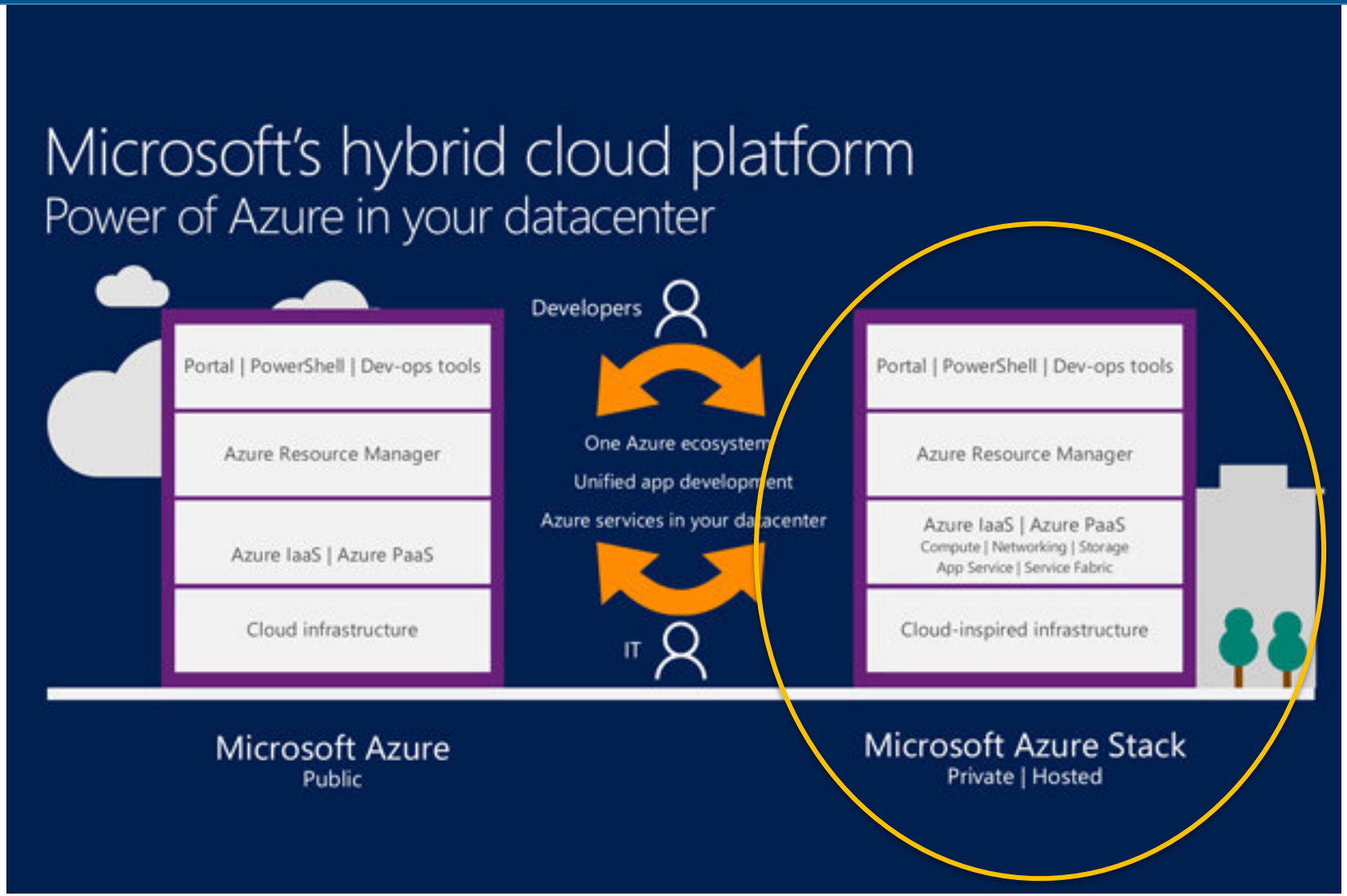
Service Provider Interface

- **ND Service Provider Interface**
  - NDSPI
    - Enforceable API definition for User Verbs
  - IB is primary scenario
  - Agnostic of iWARP, RoCE, or IB
- **Drivers?**
  - Separate windows kernel proxy for control path
  - Mid-Layer implemented in Drivers
  - Separate API from NDKPI
  - HPC focused - Limited IHV availability



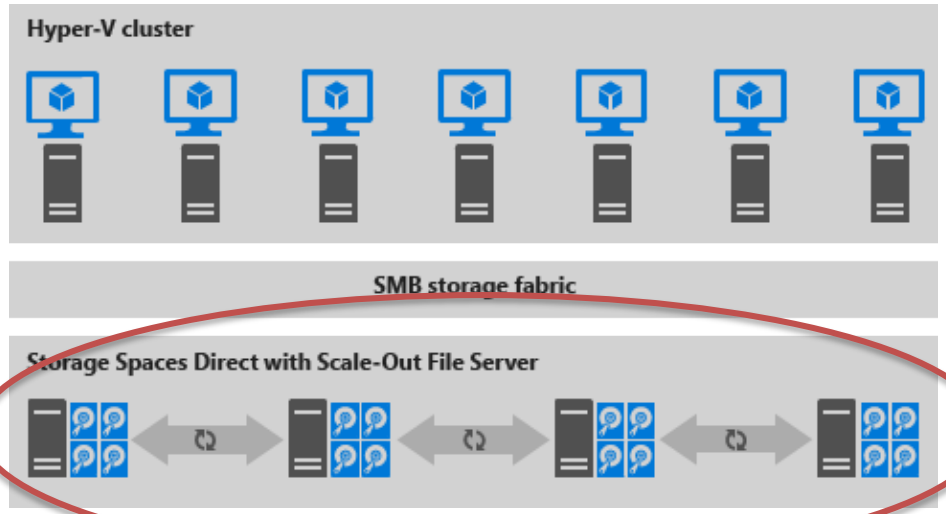
WHERE DO WE USE RDMA?

PUBLIC CLOUD - PRIVATE CLOUD



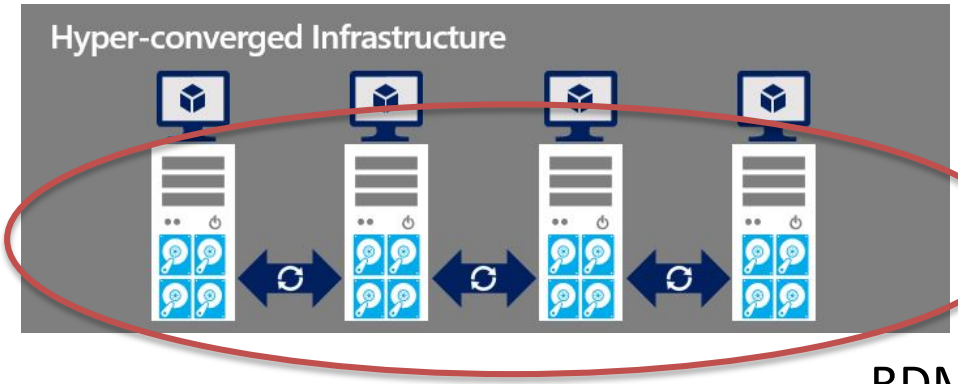
DISAGGREGATE VS HYPER-CONVERGED

RDMA (iWARP or RoCE) – **Dedicated** Fabric



Windows Server 2016

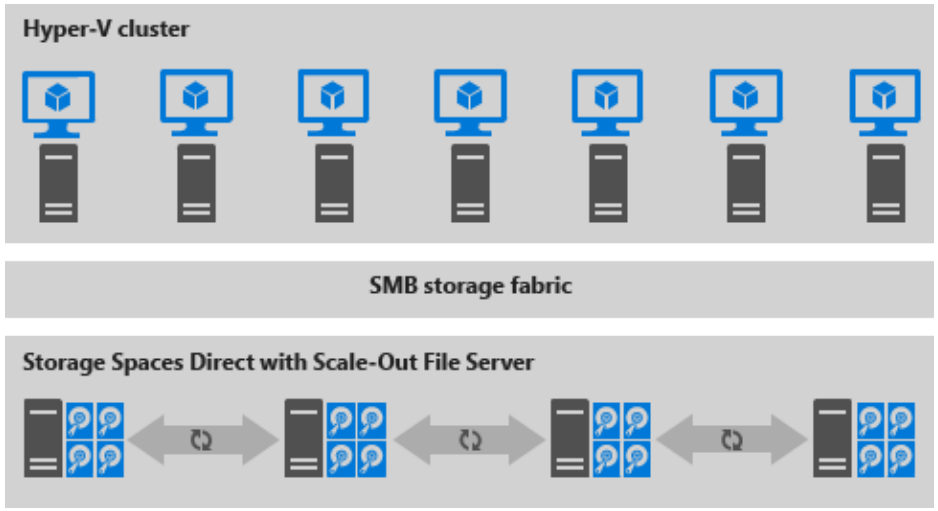
DISAGGREGATE VS HYPER-CONVERGED



Azure Stack

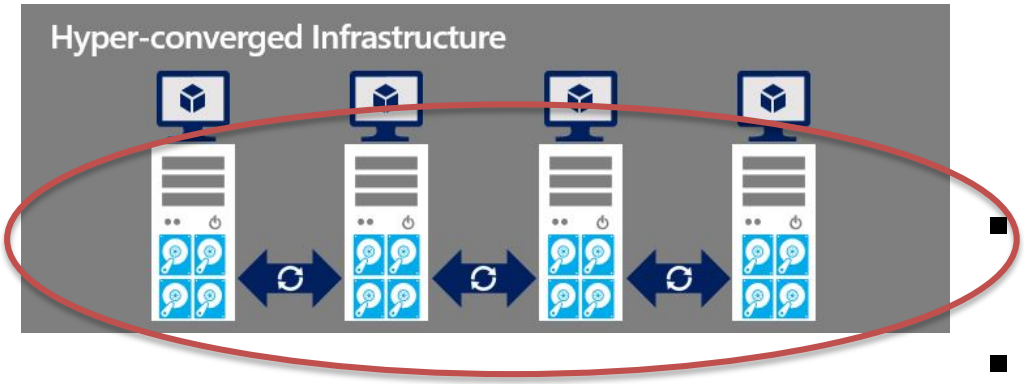
RDMA (iWARP or RoCE) – **Shared** Fabric

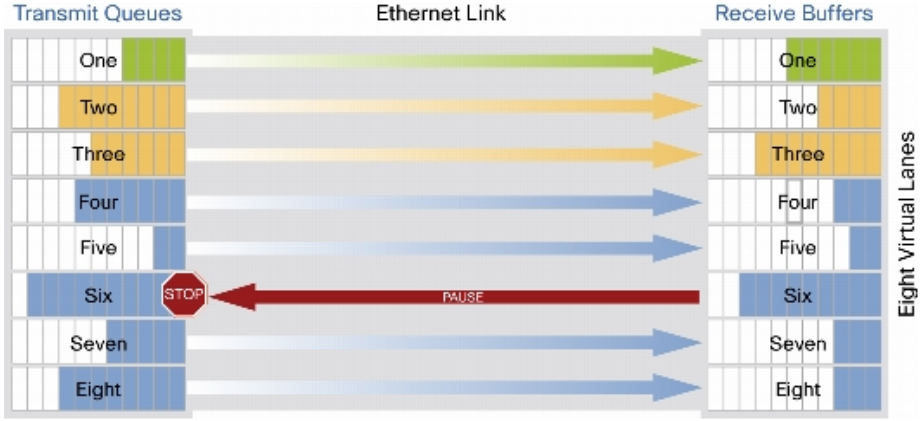
DCB: PFC/ETS



Windows Server 2016

DISAGGREGATE VS HYPER-CONVERGED

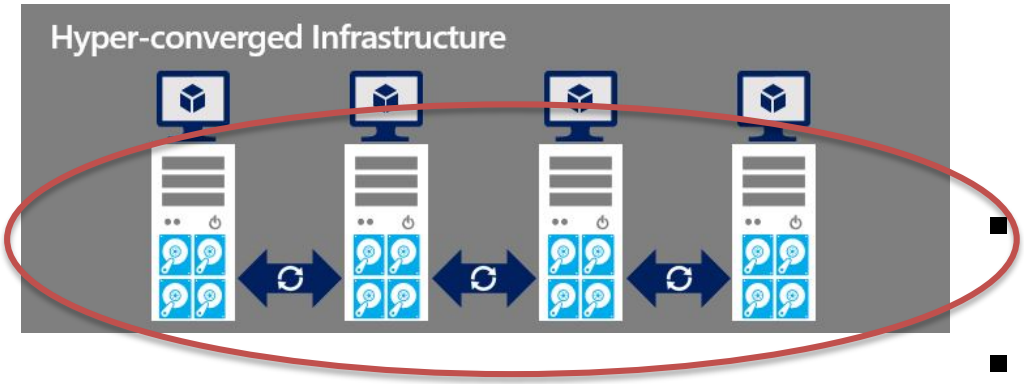


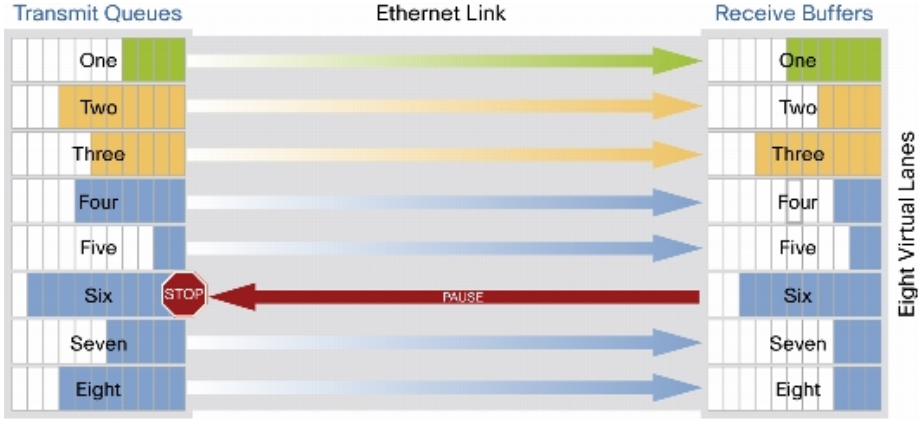


Data Center Bridging (DCB)

- **VLANs**
  - Of course..
- **Priority Flow Control (PFC)**
  - Primarily for RoCE
  - Optional for iWARP
- **Enhanced Transmission Selection (ETS)**
  - TX Reservation Requirements

DISAGGREGATE VS HYPER-CONVERGED

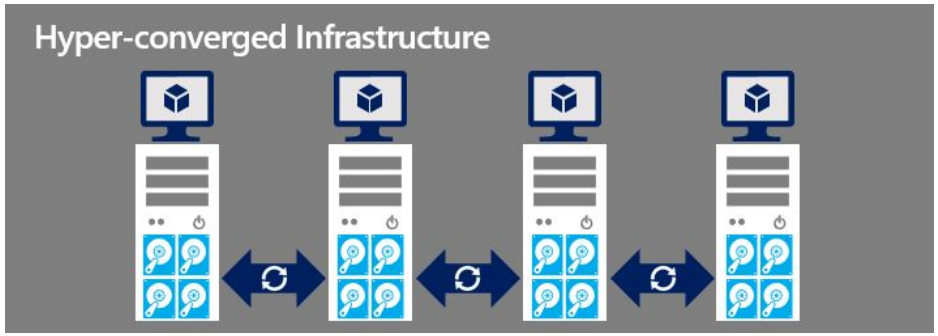


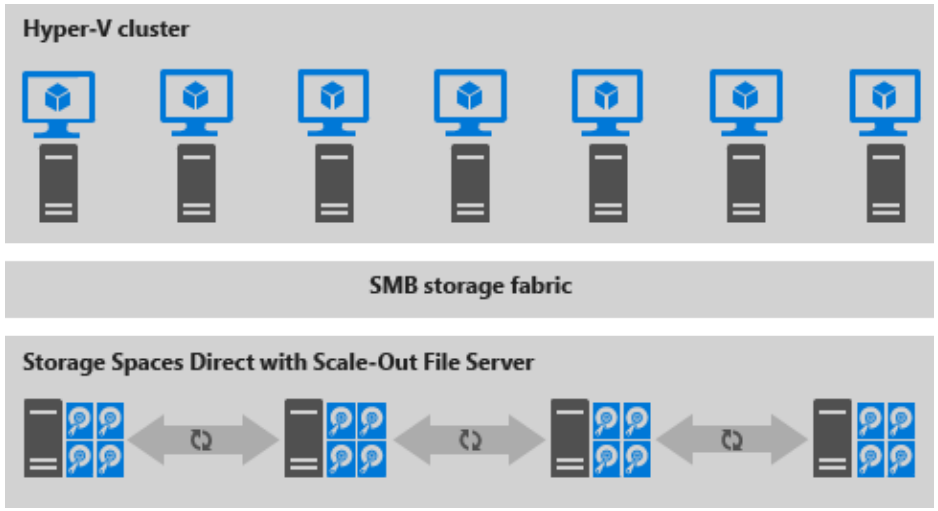


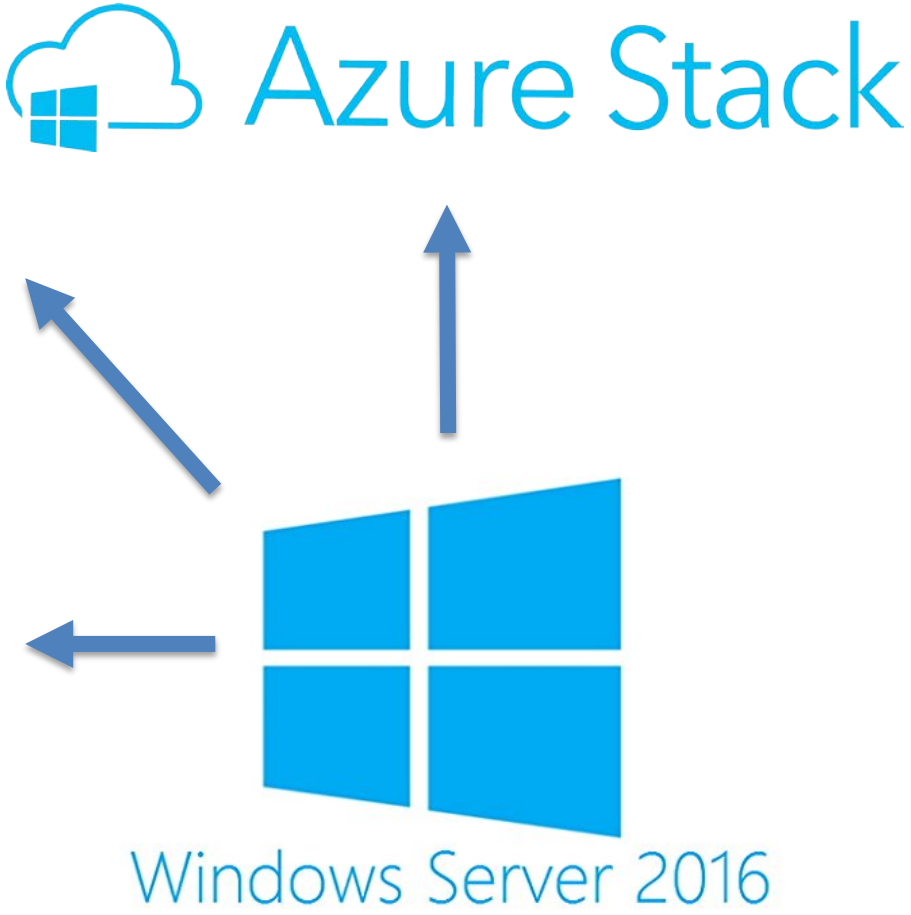
Data Center Bridging (DCB)

- **VLANs**
  - Of course..
- **Priority Flow Control (PFC)**
  - Primarily for RoCE
  - Optional for iWARP
- **Enhanced Transmission Selection (ETS)**
  - TX Reservation Requirements
- **Is DCB required?**
  - RoCE as operational prerequisite
  - iWARP for strict Storage SLAs

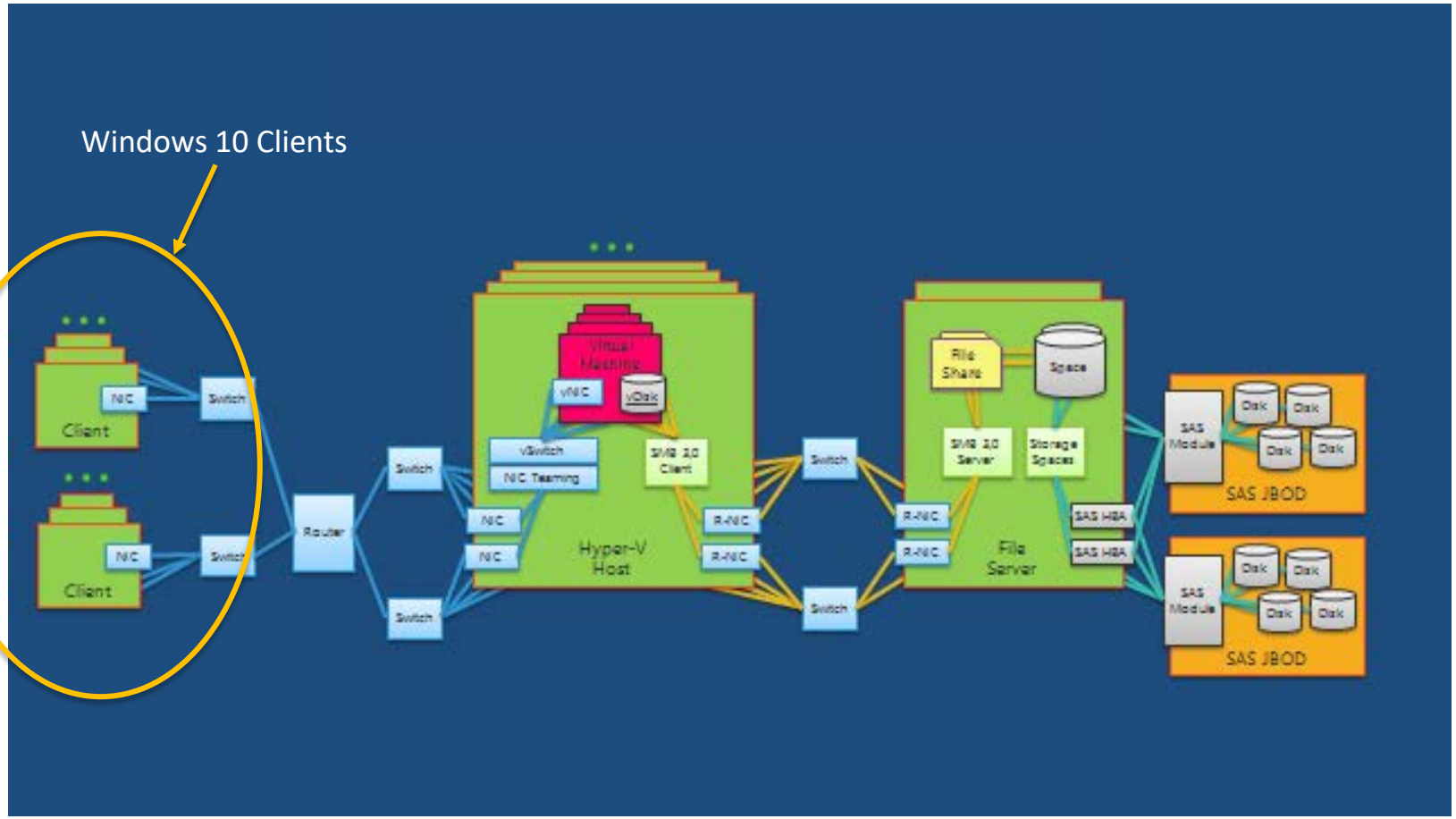
DISAGGREGATE VS HYPER-CONVERGED



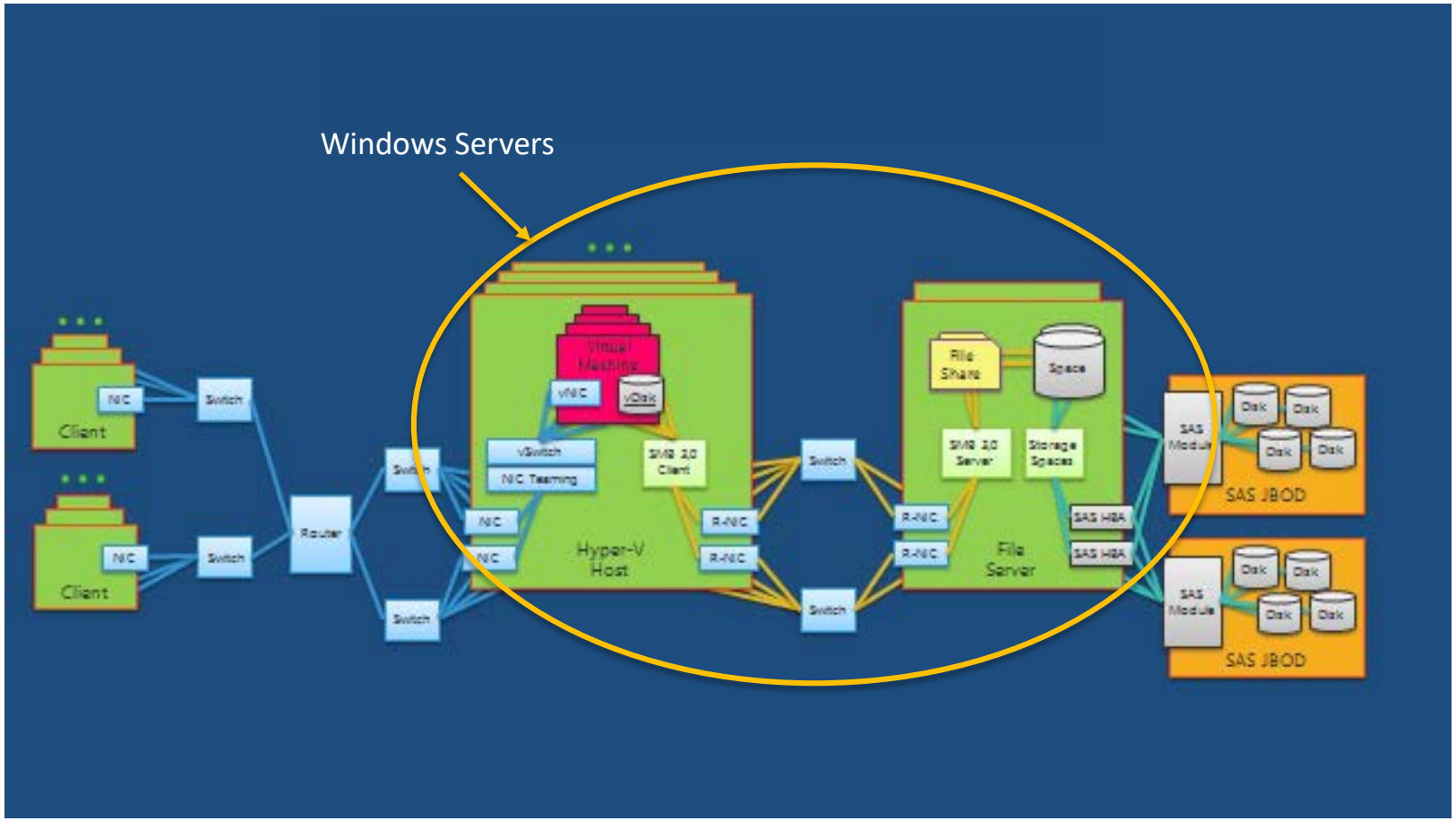




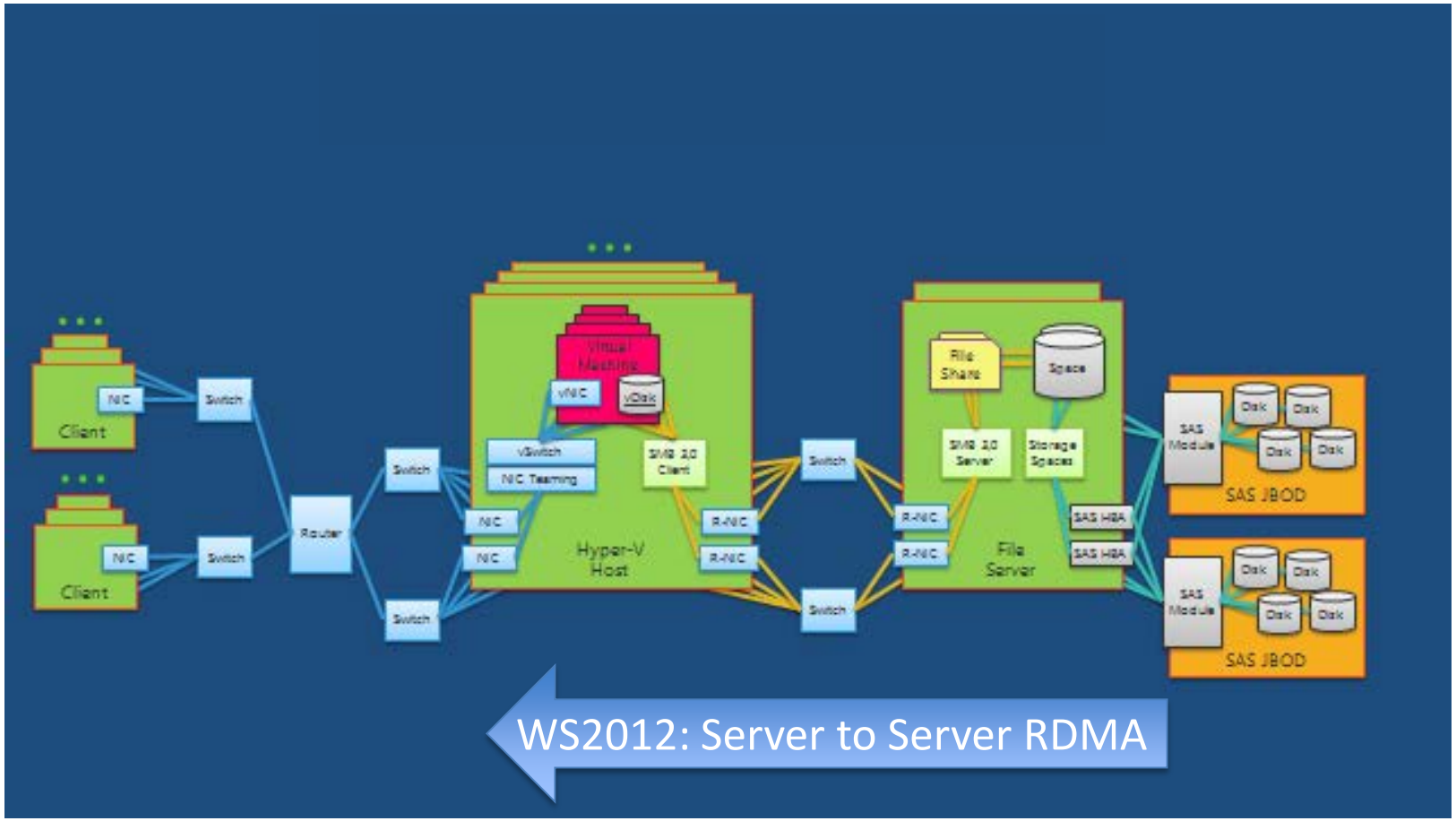
CLIENT RDMA


Windows 10 Clients
Client
NIC
Switch
Client
NIC
Switch
Router
Switch
Switch
NIC
NIC
Virtual Machine
vNIC
vDisk
vSwitch
NIC Teaming
SMB 3.0 Client
Hyper-V Host
R-NIC
R-NIC
Switch
Switch
R-NIC
R-NIC
File Share
Space
SMB 3.0 Server
Storage Spaces
File Server
SAS HBA
SAS HBA
SAS Module
Disk
Disk
Disk
Disk
SAS JBOD
SAS Module
Disk
Disk
Disk
Disk
SAS JBOD

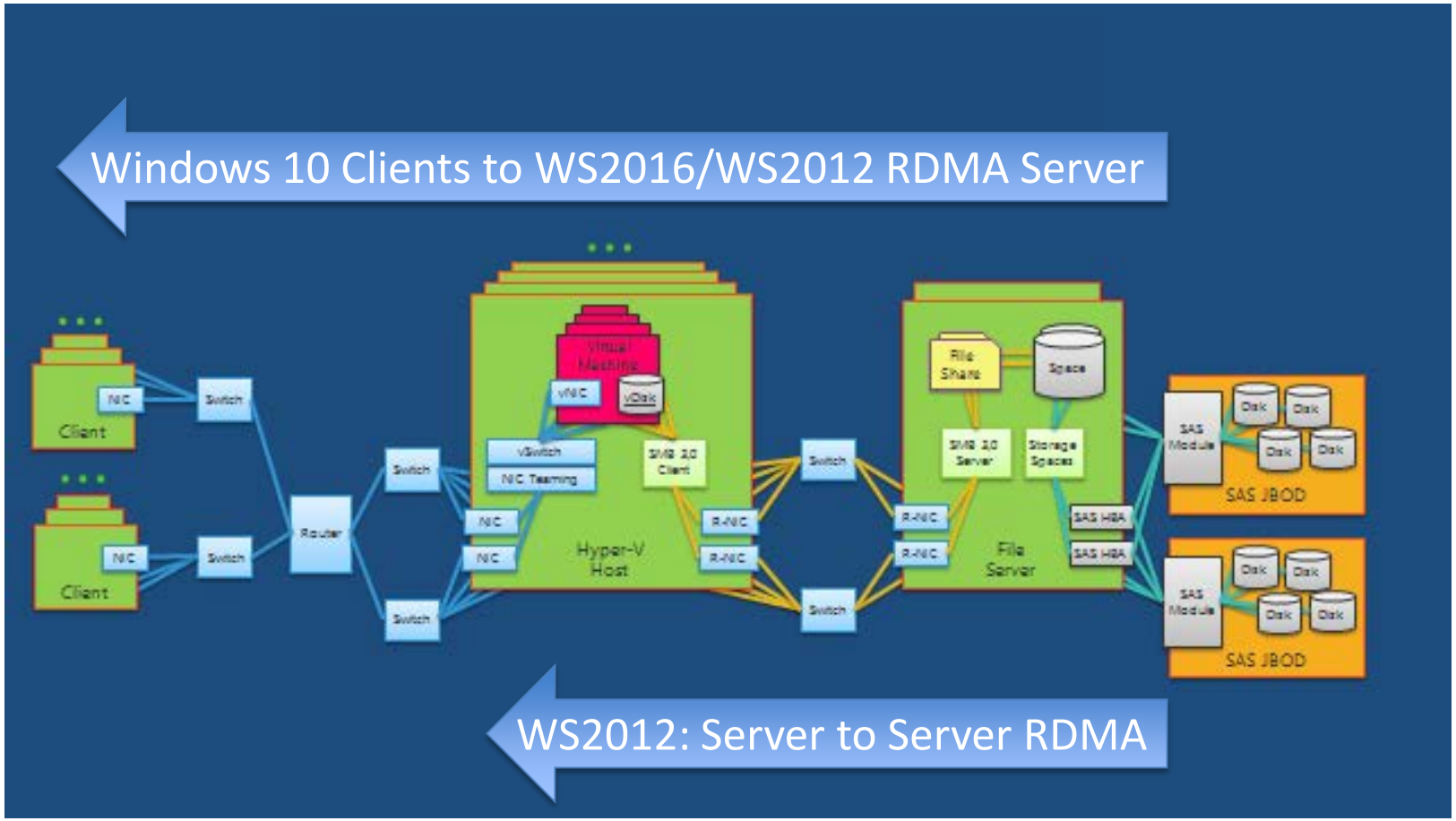
CLIENT RDMA


Windows Servers
Client
NIC
Switch
Client
NIC
Switch
Router
Switch
Switch
Virtual Machine
vNIC
vDisk
vSwitch
NIC Teaming
SMB 3.0 Client
NIC
NIC
Hyper-V Host
R-NIC
R-NIC
Switch
Switch
R-NIC
R-NIC
File Share
Space
SMB 3.0 Server
Storage Spaces
File Server
SAS HBA
SAS HBA
SAS Module
Disk
Disk
Disk
Disk
SAS JBOD
SAS Module
Disk
Disk
Disk
Disk
SAS JBOD

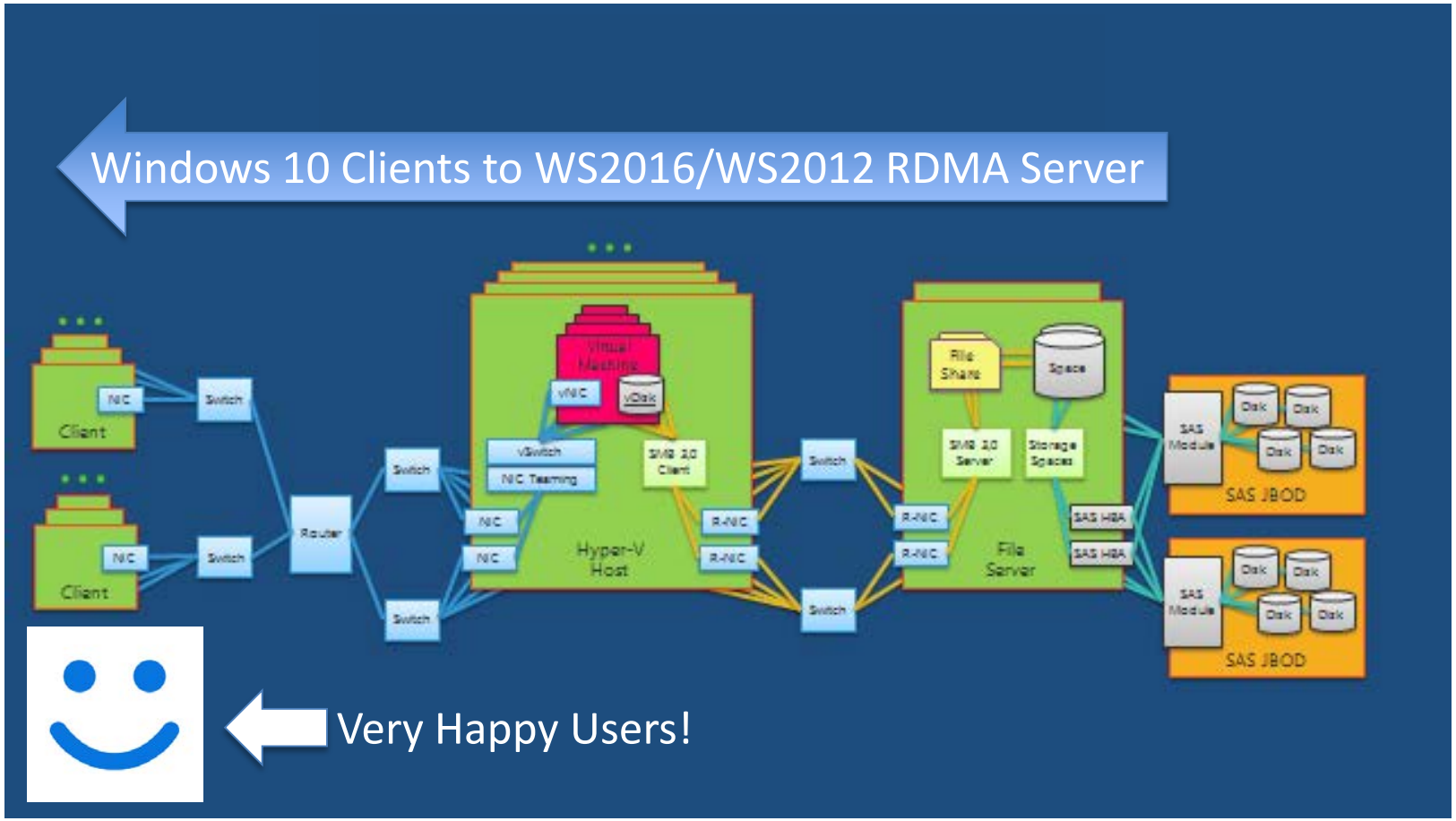
CLIENT RDMA


Client
NIC
Switch
Router
Hyper-V Host
vNIC
vDisk
vSwitch
NIC Teaming
SMB 3.0 Client
R-NIC
File Server
File Share
Space
SMB 3.0 Server
Storage Spaces
SAS HBA
SAS Module
Disk
SAS JBOD
WS2012: Server to Server RDMA

CLIENT RDMA

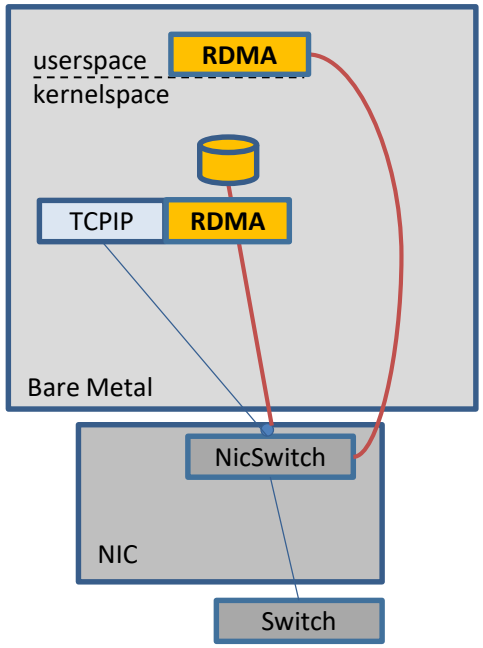

Windows 10 Clients to WS2016/WS2012 RDMA Server
Client
NIC
Switch
Router
Virtual Machine
vNIC
vDisk
vSwitch
NIC Teaming
SMB 3.0 Client
Hyper-V Host
R-NIC
File Share
Space
SMB 3.0 Server
Storage Spaces
File Server
SAS HBA
SAS Module
Disk
SAS JBOD
WS2012: Server to Server RDMA

CLIENT RDMA

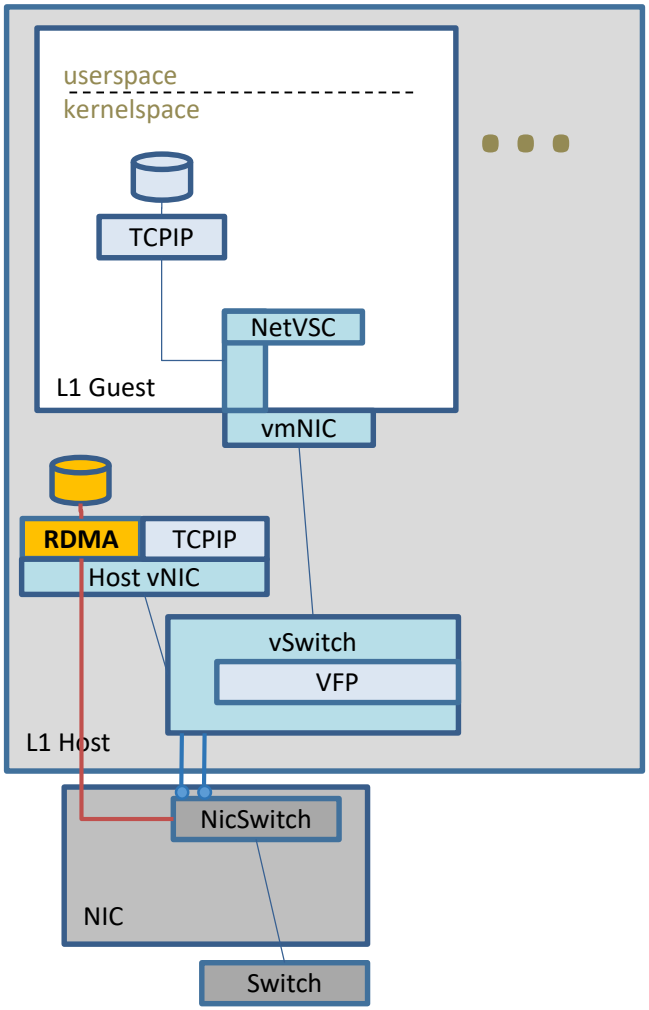

Windows 10 Clients to WS2016/WS2012 RDMA Server
Client
NIC
Switch
Client
NIC
Switch
Router
Switch
Switch
NIC
NIC
Hyper-V Host
Virtual Machine
vNIC
vDisk
vSwitch
NIC Teaming
SMB 3.0 Client
R-NIC
R-NIC
Switch
Switch
R-NIC
R-NIC
File Server
File Share
Space
SMB 3.0 Server
Storage Spaces
SAS HBA
SAS HBA
SAS Module
Disk
Disk
Disk
Disk
SAS JBOD
SAS Module
Disk
Disk
Disk
Disk
SAS JBOD
Very Happy Users!

INTERNALS

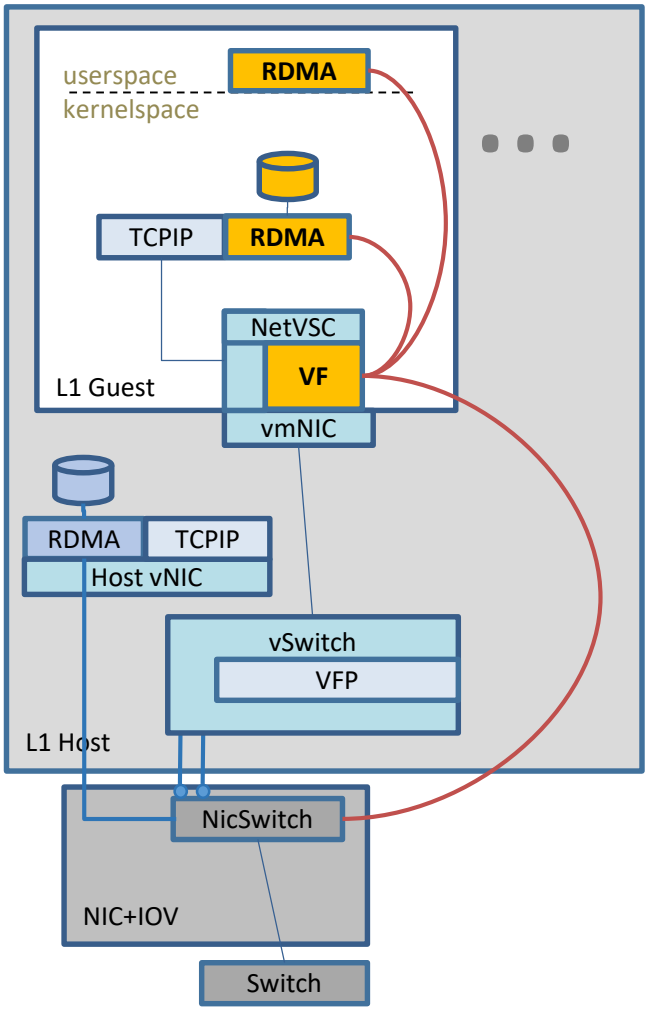
RDMA MODES



Mode 1: Native RDMA



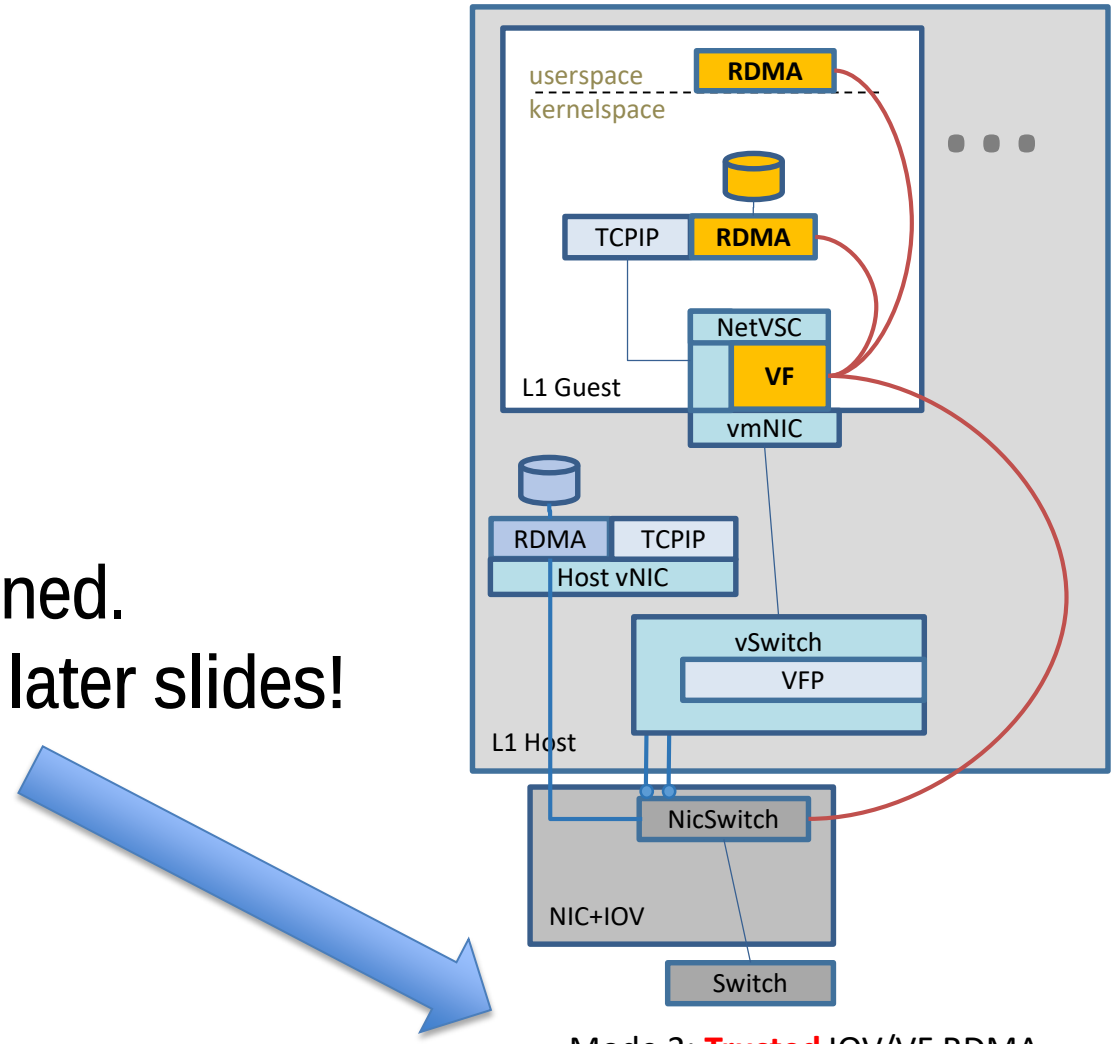
Mode 2: vSwitch RDMA



Mode 3: **Trusted** IOV/VF RDMA

RDMA MODES

Stay Tuned.
More details in later slides!

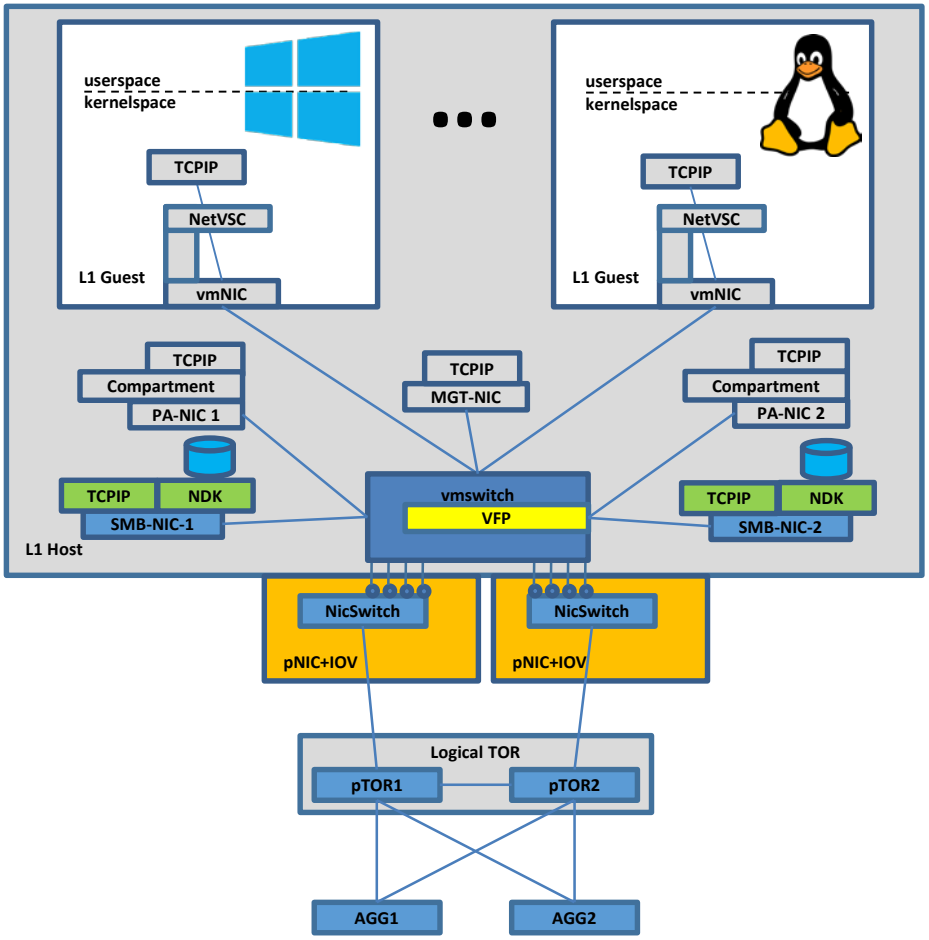


Mode 3: **Trusted** IOV/VF RDMA

SERVER INTERNALS

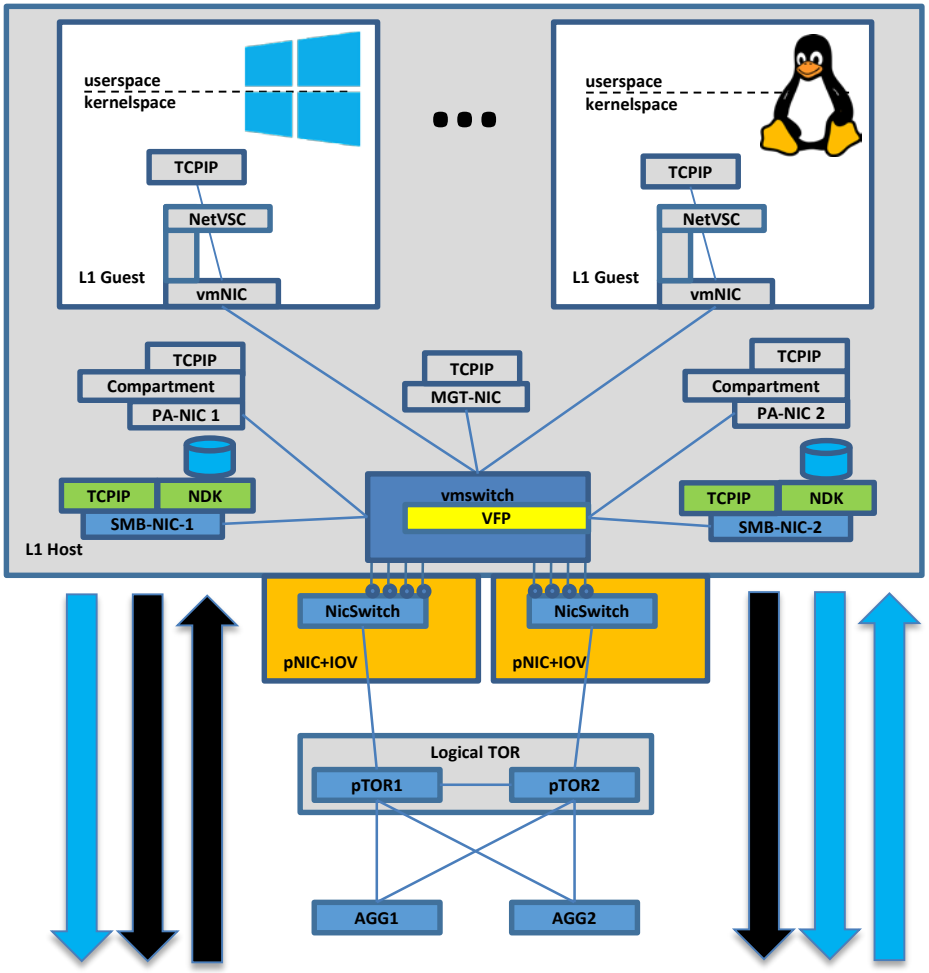
SERVER INTERNALS

- **Deployment**
  - Single NIC Dual Port



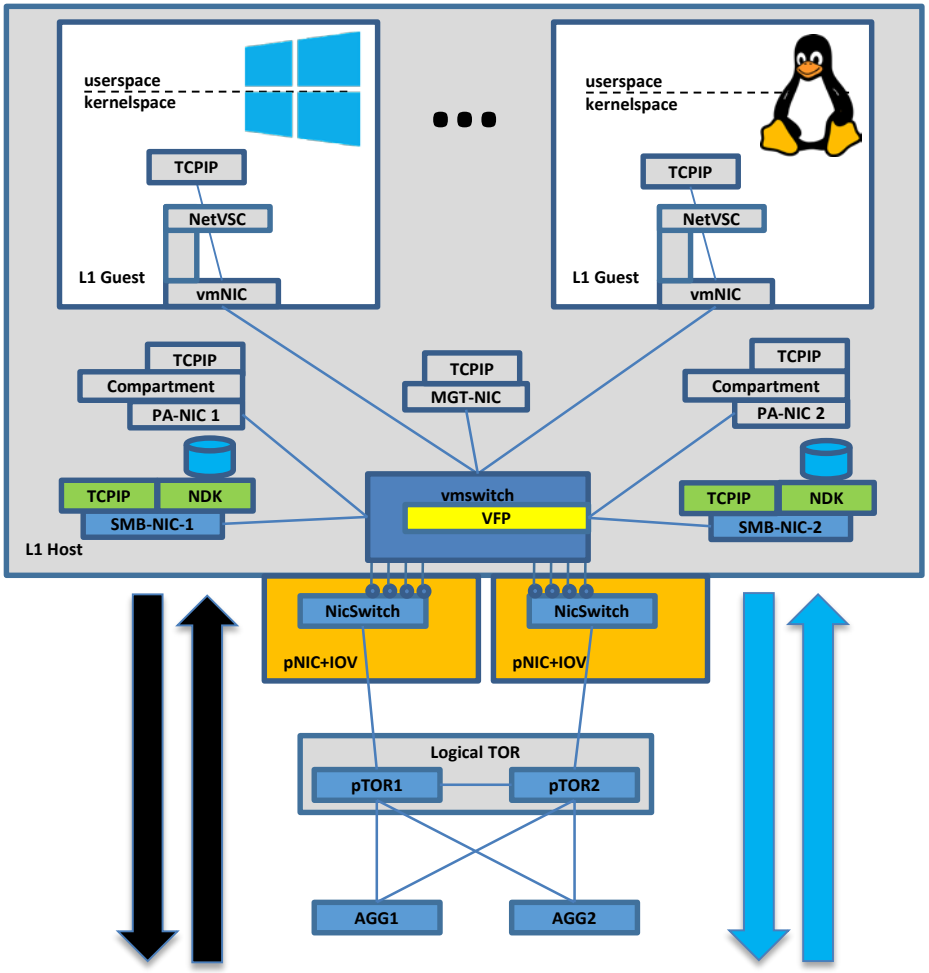
SERVER INTERNALS

- **Deployment**
  - Single NIC Dual Port
  - Switch Independent Teaming/Failover
    - Dynamic Distribution



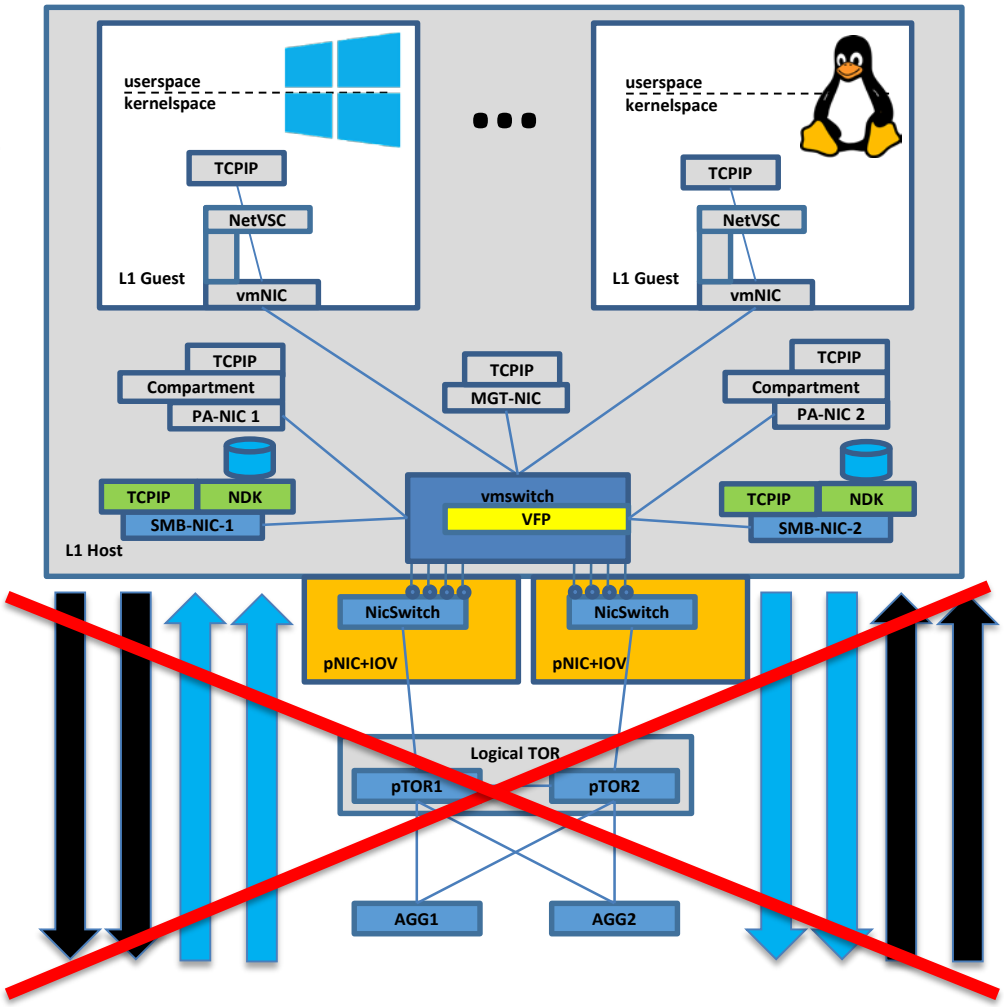
SERVER INTERNALS

- **Deployment**
  - Single NIC Dual Port
  - Switch Independent Teaming/Failover
    - Dynamic Distribution
    - Hyper-V Port Mode



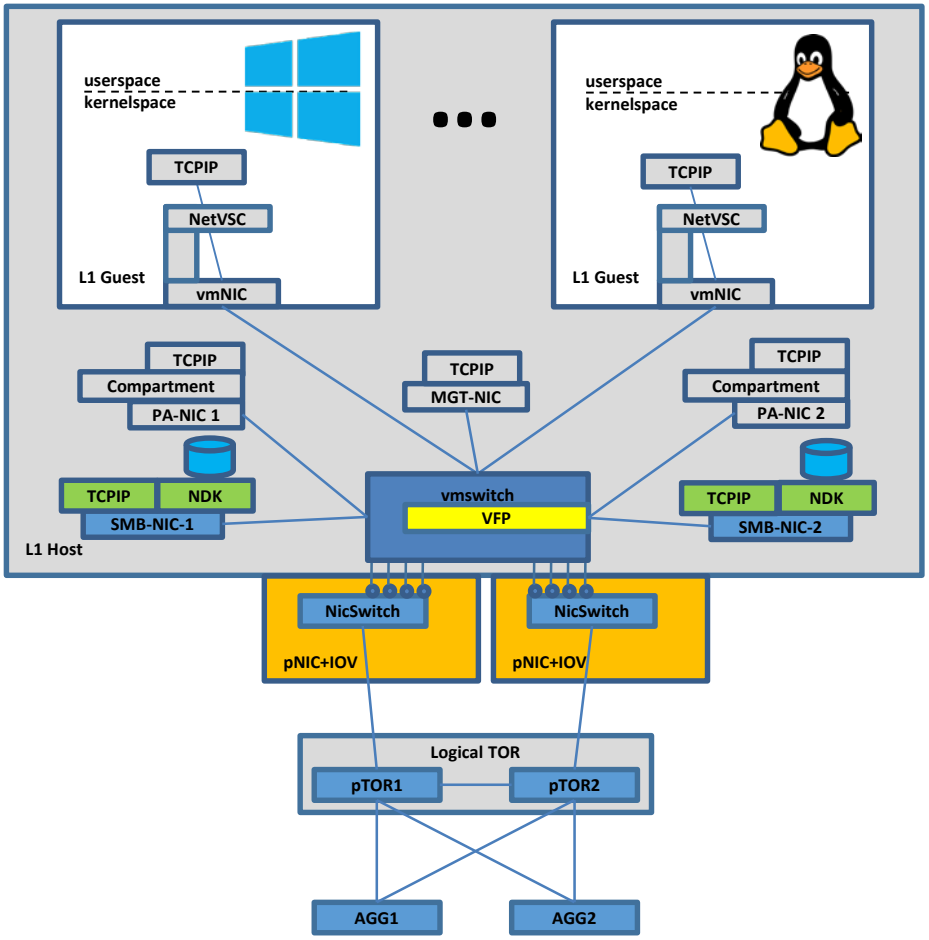
SERVER INTERNALS

- **Deployment**
  - Single NIC Dual Port
  - Switch Independent Teaming/Failover
    - Dynamic Distribution
    - Hyper-V Port Mode
    - !LACP – More later….



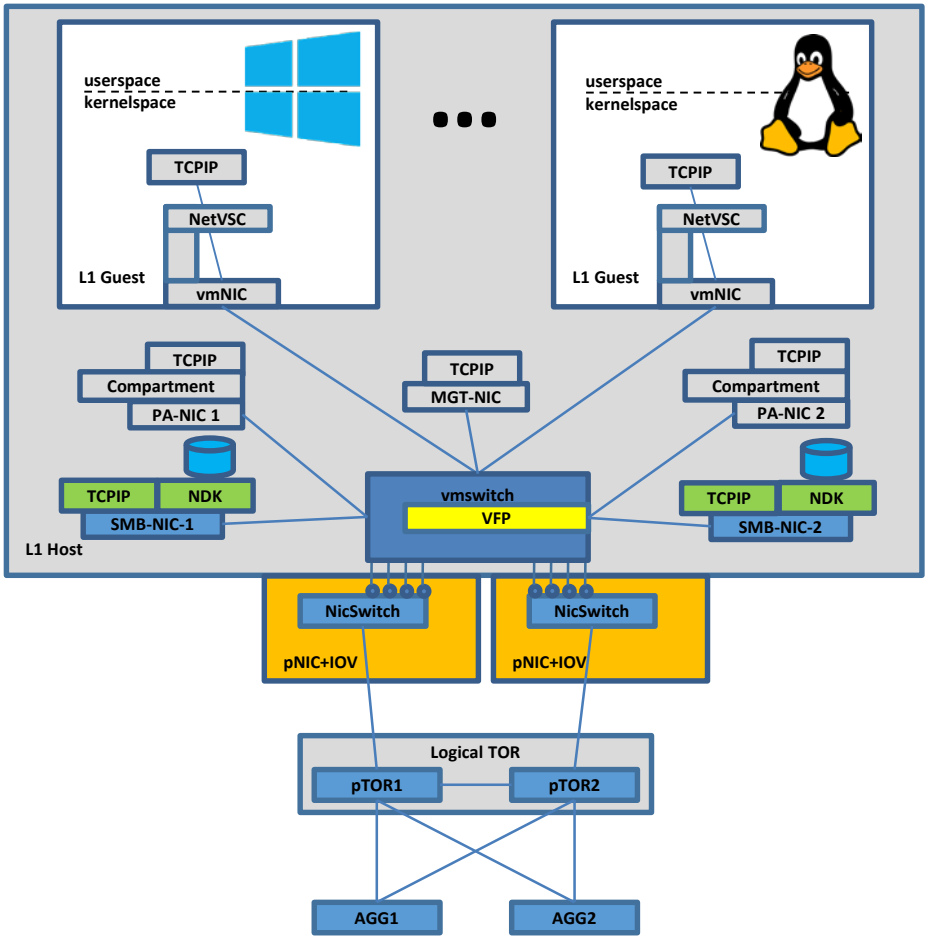
SERVER INTERNALS

- **Deployment**
  - Single NIC Dual Port
  - Switch Independent Teaming/Failover
    - Dynamic Distribution
    - Hyper-V Port Mode
    - !LACP – More later….
- **NicSwitch / VEB**
  - RDMA enabled on PF vPorts



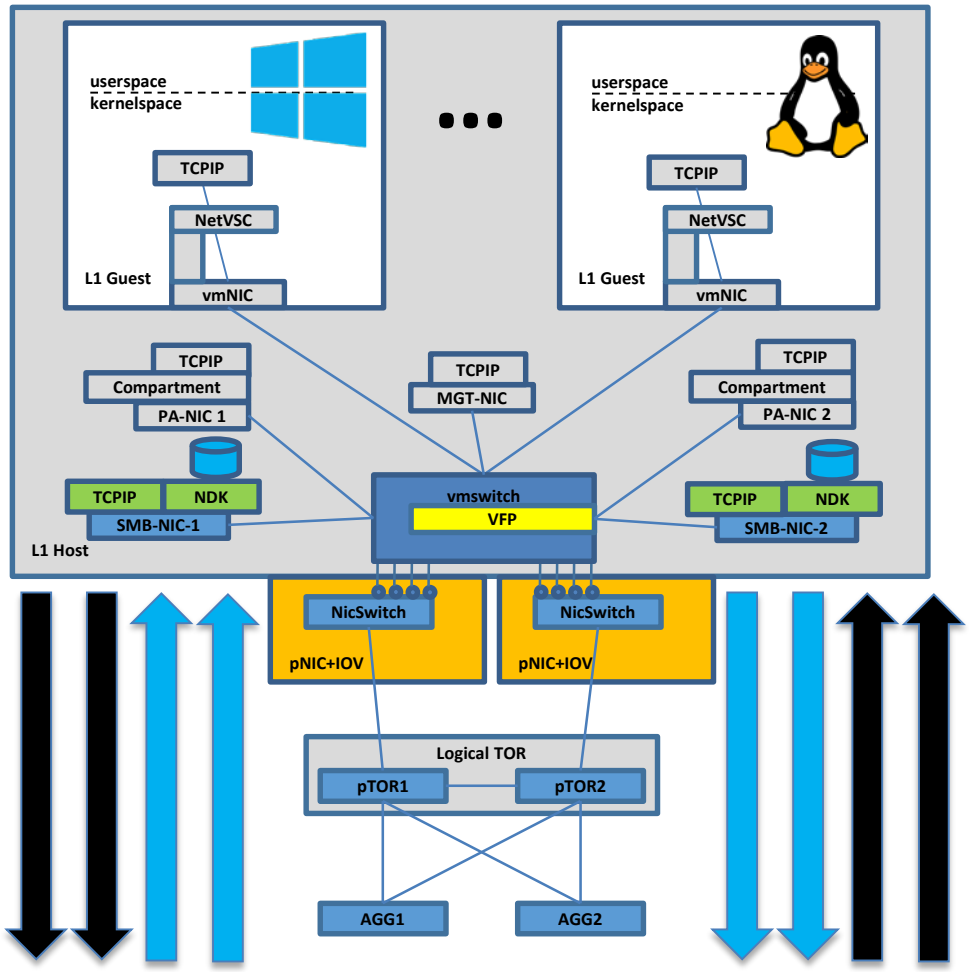
SERVER INTERNALS

- **Deployment**
  - Single NIC Dual Port
  - Switch Independent Teaming/Failover
    - Dynamic Distribution
    - Hyper-V Port Mode
    - !LACP – More later….
- **NicSwitch / VEB**
  - RDMA enabled on PF vPorts
- **Converged Ethernet**
  - DCB: PFC TC3 RDMA
  - DCB: ETS TC3 50% BW Reservation
  - Tenants (VMs) TC0



SERVER INTERNALS

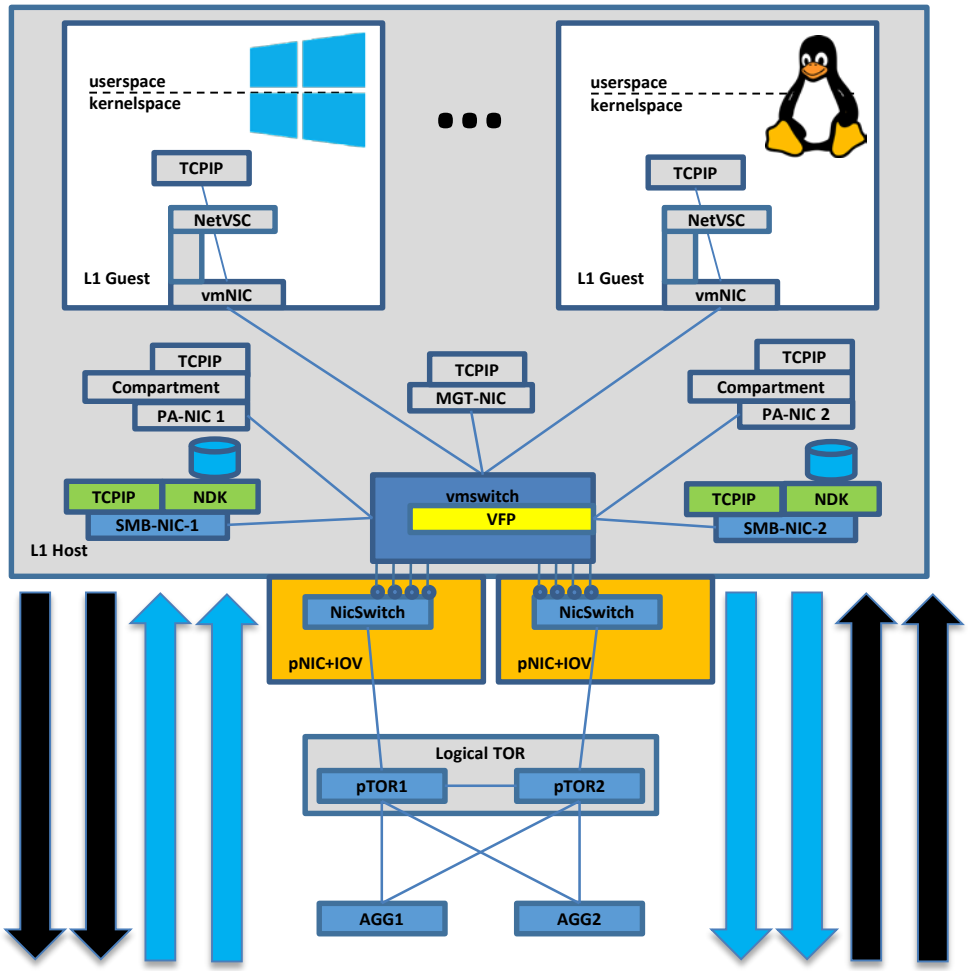
- **Deployment**
  - Single NIC Dual Port
  - Switch Independent Teaming/Failover
    - Dynamic Distribution
    - Hyper-V Port Mode
    - !LACP – More later….
- **NicSwitch / VEB**
  - RDMA enabled on PF vPorts
- **Converged Ethernet**
  - DCB: PFC TC3 RDMA
  - DCB: ETS TC3 50% BW Reservation
  - Tenants (VMs) TC0
- **SMB Multichannel + RDMA**
  - Verb Consumer Teaming!



SERVER INTERNALS

- **Deployment**
  - Single NIC Dual Port
  - Switch Independent Teaming/Failover
    - Dynamic Distribution
    - Hyper-V Port Mode
    - !LACP – More later….
- **NicSwitch / VEB**
  - RDMA enabled on PF vPorts
- **Converged Ethernet**
  - DCB: PFC TC3 RDMA
  - DCB: ETS TC3 50% BW Reservation
  - Tenants (VMs) TC0
- **SMB Multichannel + RDMA**
  - Verb Consumer Teaming!

Very Happy Users!



CHALLENGES

BROWNFIELD SILOS…

Competing Priorities

Is LACP an RDMA deployment blocker? …Yes!

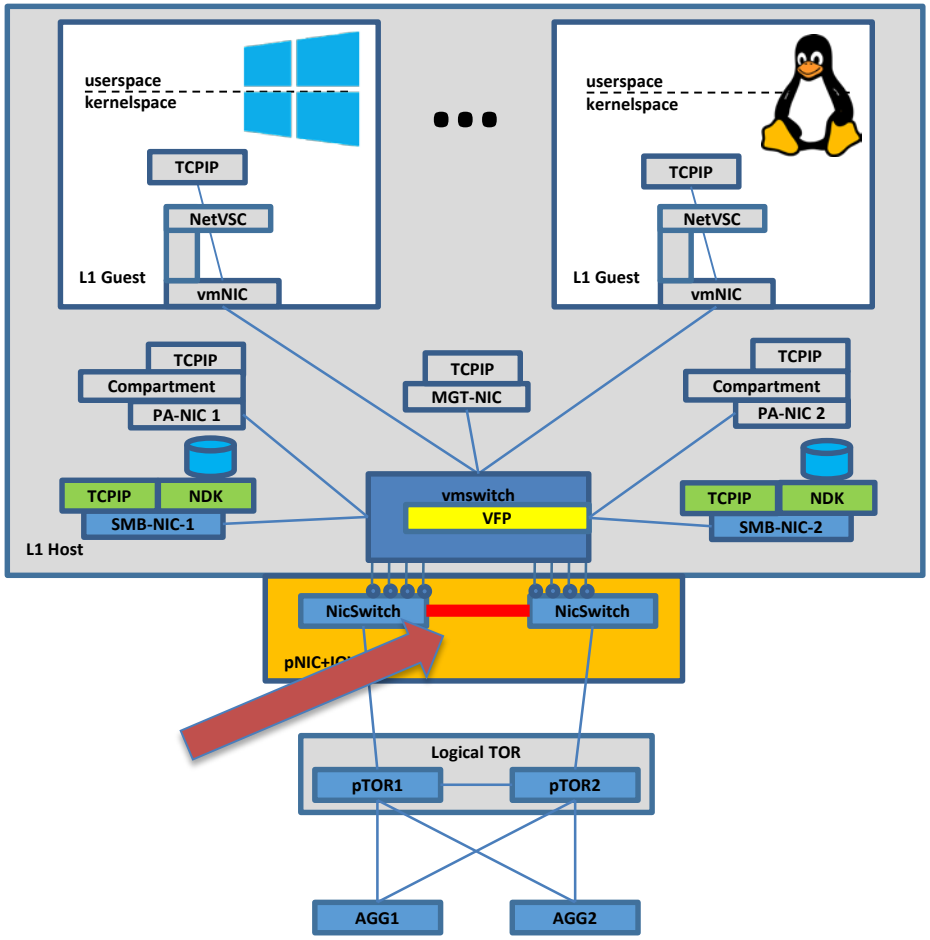
BROWNFIELD SILOS…

Competing Priorities

- **Network Admin**
  - LACP to the Host is de-facto
  - L3 Hops
  - VLAN / COS conversions
  - Appliances
  - Etc. Etc. Etc.
- **Server Admin**
  - Hyper-Converged
  - RDMA everywhere
  - IOV to Tenants / Containers
  - SDN / SDS
  - Etc. Etc. Etc…

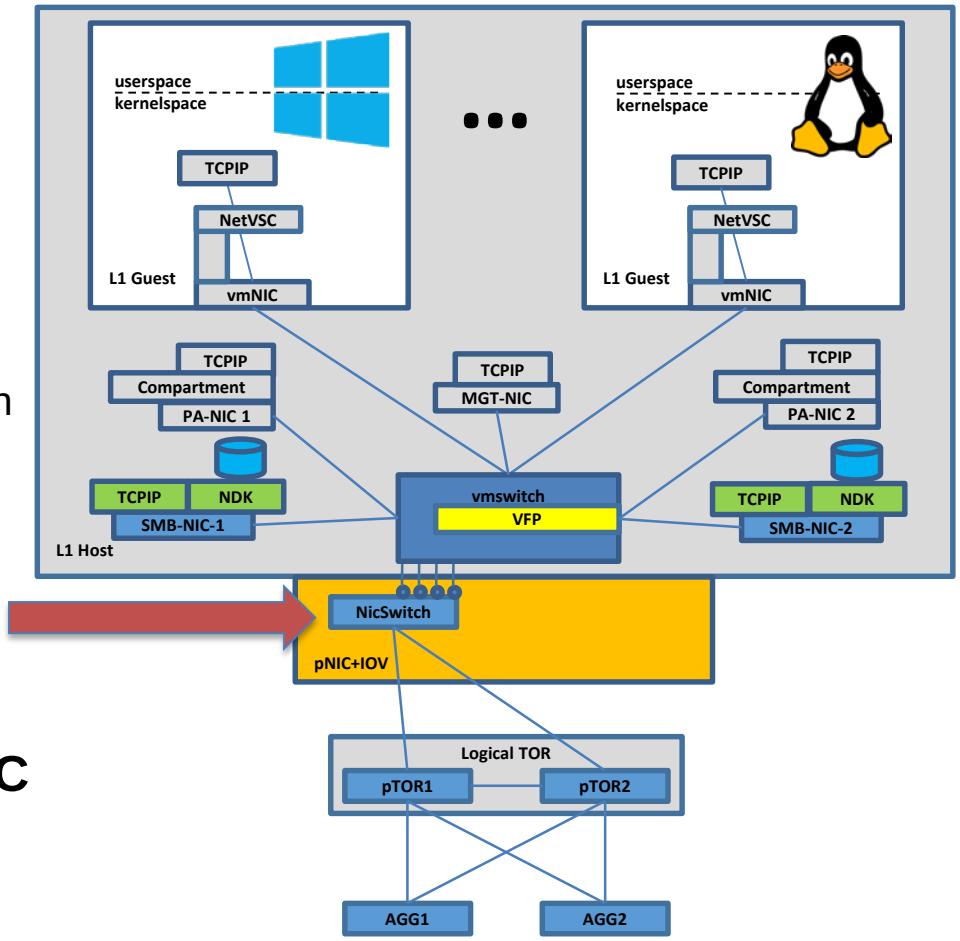
PROPOSAL: ECMP TO THE HOST (1)

- **VPC/Crossbar "Concept"**
  - Offload the appearance of ECMP to the NIC
  - Enables Host NIC Port control
  - Worst case 50% increase in PCI BW
- **RDMA ingress**
  - Always arrives at expected port
  - No LACP distribution problem
- **Capacity planning**
  - PCIe Gen3 x16 = 128Gbps
  - 2x 40G ports w/ 50% oversubscription
  - Is this acceptable overprovisioning?



PROPOSAL: ECMP TO THE HOST (2)

- **LACP Offload**
  - Single Physical Port Exposed to Host
  - LACP management and failover in NIC scope
- **No Host visibility for Physical Port management**
  - Not a major problem
  - Report Remaining Link Speed change on member port loss
  - Report Link loss on all member port loss
- **RDMA ingress**
  - From Host PoV, there's only one Port. Everything arrives correctly.
  - LACP distribution problem goes away
- **Is this a viable solution from NIC IHVs?**
  - Does the PCI BW oversubscription problem persist?



DIAGNOSIS

- **Does this sounds familiar?**
  - Customer deploys fabric, RDMA doesn’t work.

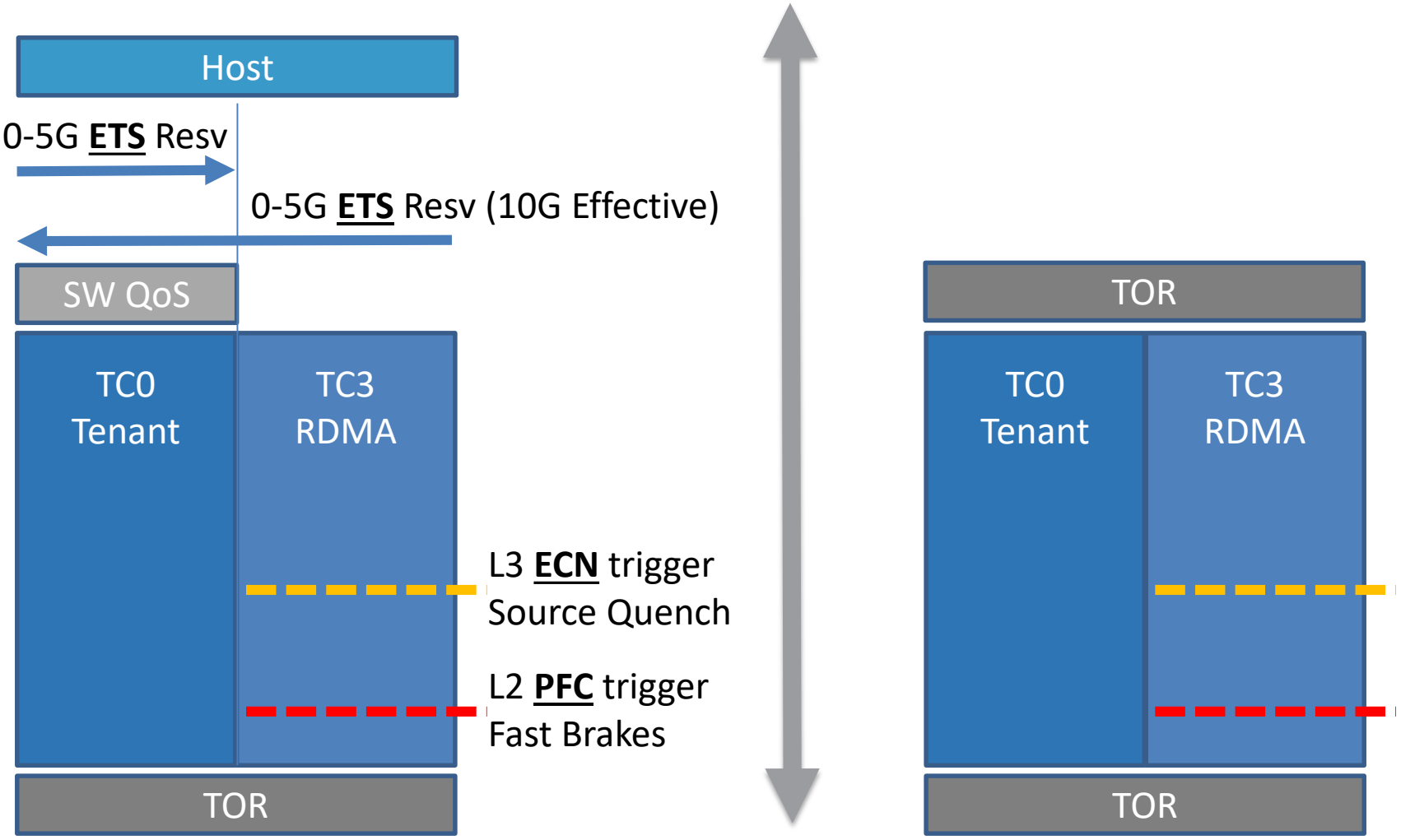
DIAGNOSIS

- **Does this sounds familiar?**
  - Customer deploys fabric, RDMA doesn't work.
- **The investigation begins…**

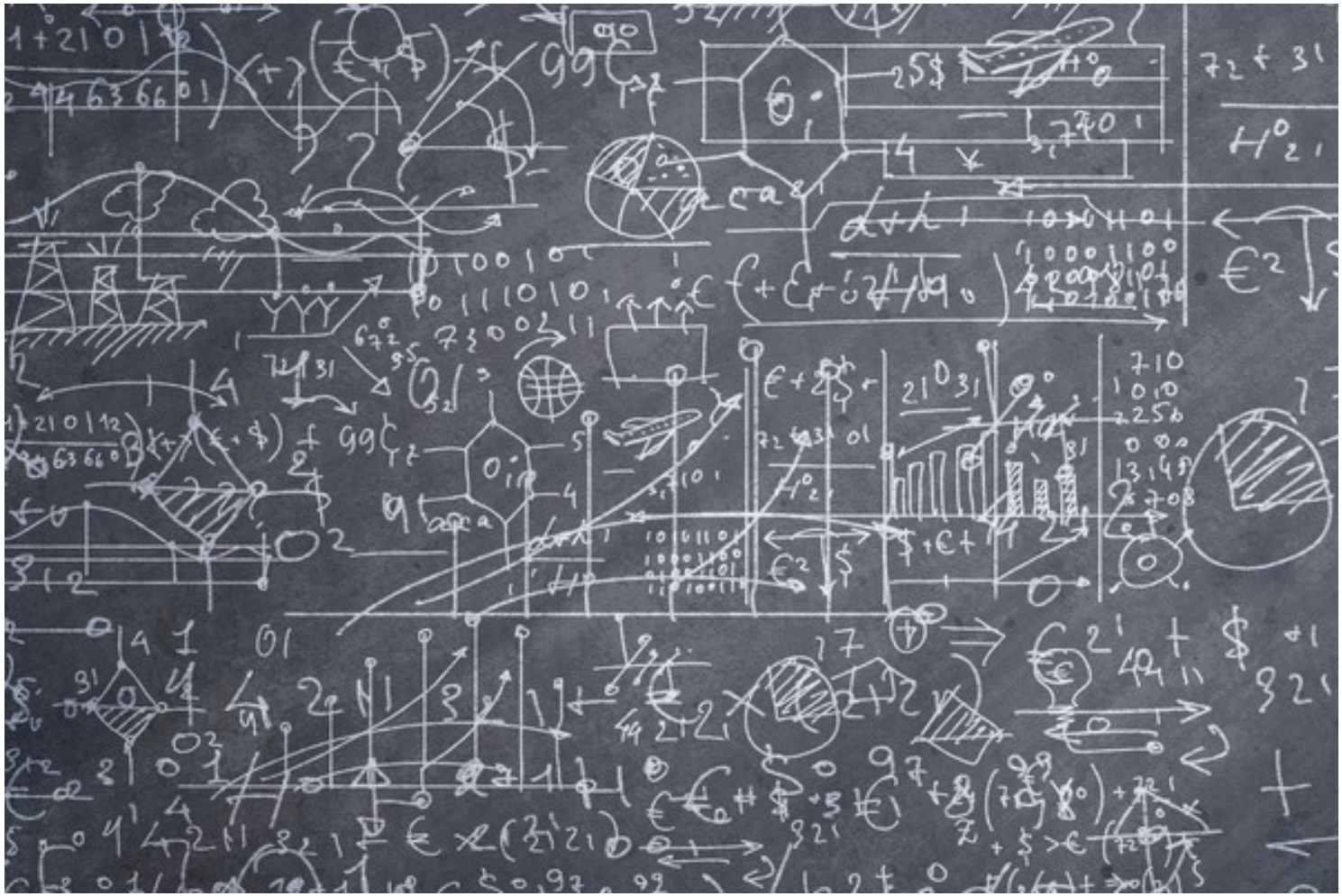
DIAGNOSIS

- **Does this sounds familiar?**
  - Customer deploys fabric, RDMA doesn't work.
- **The investigation begins…**
  - Is it RoCE or iWARP?
  - We're VLANs configured correctly?
  - Was PFC configured correctly on all hops?
  - Was ETS configured correctly?
  - What NICs is the customer using?
  - Which IHV specific counters should be inspected?
  - Which TOR/AGG specific counters should be inspected?
  - Did the TOR/AGG expose per-TC PFC/ETS counters?
  - Does the TOR/AGG support PFC on priority tagged traffic? (ie. VLAN-0)?
  - We're there any L3 Hops?
  - Was 802.1p preserved/propagated across L3 Hops?
  - Was ECN configured properly and relative to PFC?
  - Is there congestion on the fabric?
  - Was incast validated?
  - Is there evidence of cascading HoL?
  - Etc. Etc. Etc…

DIAGNOSIS


Host
0-5G ETS Resv
0-5G ETS Resv (10G Effective)
SW QoS
TC0
Tenant
TC3
RDMA
L3 ECN trigger
Source Quench
L2 PFC trigger
Fast Brakes
TOR
TOR
TC0
Tenant
TC3
RDMA
TOR

DIAGNOSIS



DIAGNOSIS

- **How can we make RDMA turn key?**
  - Dependency on lossless Ethernet is inherently complex for most users…

DIAGNOSIS

- **How can we make RDMA turn key?**
  - Dependency on lossless Ethernet is inherently complex for most users…
- **Simple Actions**
  - Step-by-Step 0-hero deployment guides

DIAGNOSIS

- **How can we make RDMA turn key?**
  - Dependency on lossless Ethernet is inherently complex for most users…
- **Simple Actions**
  - Step-by-Step 0-hero deployment guides
  - TORs → Inconsistent in management and diagnosis

DIAGNOSIS

- **How can we make RDMA turn key?**
  - Dependency on lossless Ethernet is inherently complex for most users…
- **Simple Actions**
  - Step-by-Step 0-hero deployment guides
  - TORs → Inconsistent in management and diagnosis
  - Verb Consumer independent monitoring
    - Host based (ICMP, TCP, RDMA) Heartbeat tools
    - Host logs all RDMA control path / error events

DIAGNOSIS

- **How can we make RDMA turn key?**
  - Dependency on lossless Ethernet is inherently complex for most users…
- **Simple Actions**
  - Step-by-Step 0-hero deployment guides
  - TORs → Inconsistent in management and diagnosis
  - Verb Consumer independent monitoring
    - Host based (ICMP, TCP, RDMA) Heartbeat tools
    - Host logs all RDMA control path / error events
  - Validation tools
    - P2P stress egress unbound UDP + RDMA
      - Expect ETS percentile = (50% RDMA + 50% UDP)
      - Else… ETS misconfiguration on Host or Fabric
    - RDMA connection break/instability → PFC misconfiguration
    - Incast stress N-1
    - System counters correlation with RDMA error event

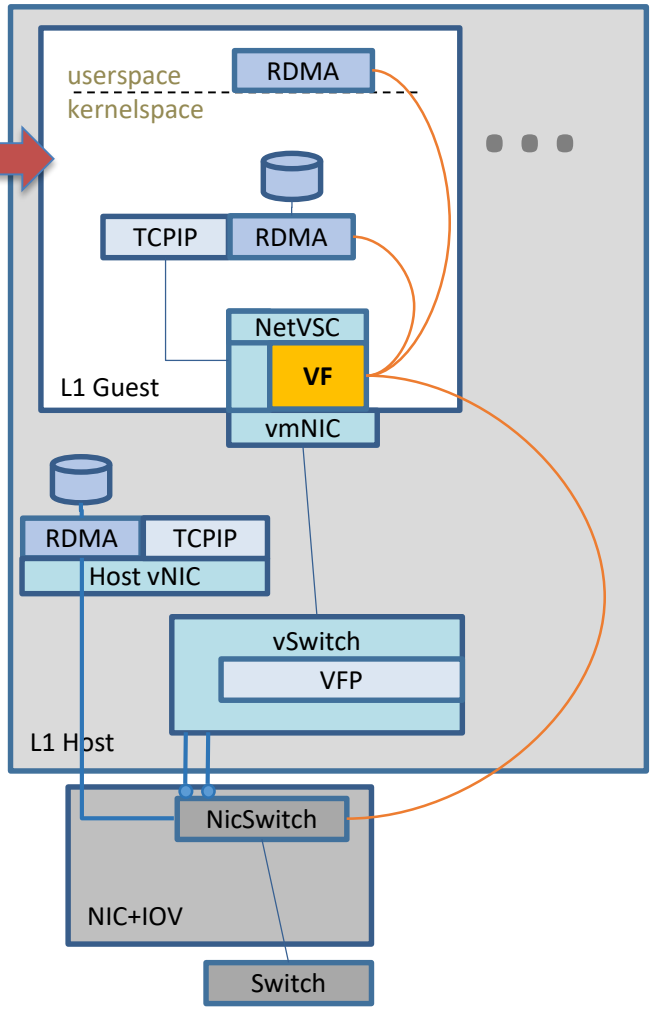
DIAGNOSIS

FUTURE WORK

MULTITENANCY – SECURING RDMA

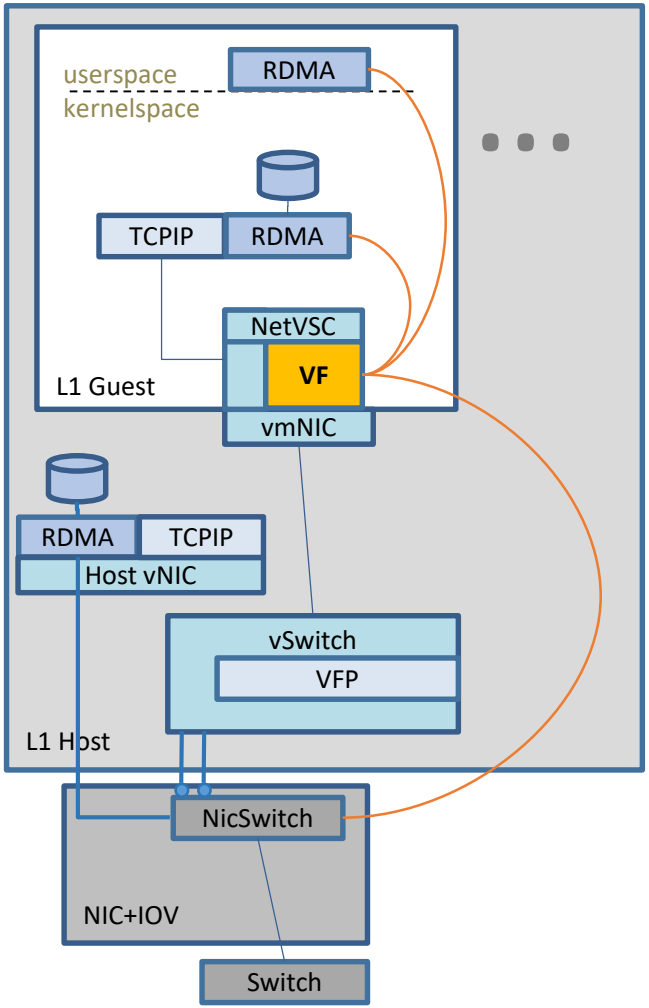
MULTITENANCY – SECURING RDMA

- **Performance and Security conflict**
  - VFs bypass security… Fabric compromised…
  - Acceptable for trusted Guests



MULTITENANCY – SECURING RDMA

- **Performance and Security conflict**
  - VFs bypass security… Fabric compromised…
  - Acceptable for trusted Guests
- **How can we secure tenants?**

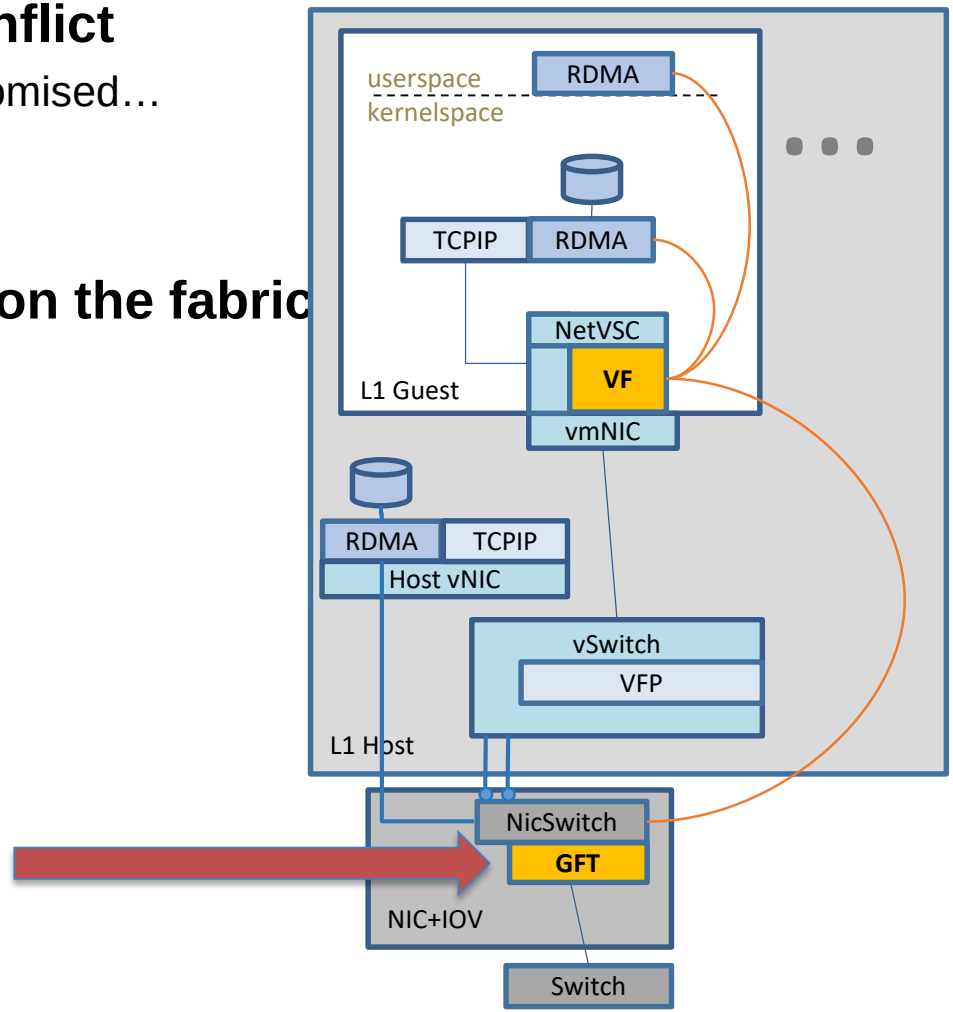


MULTITENANCY – SECURING RDMA

- **Performance and Security conflict**
  - VFs bypass security… Fabric compromised…
  - Acceptable for trusted Guests
- **How can we secure tenants?**

1. Control <u>what</u> tenant places on the fabric

- GFT – Generic Flow Tables
  - Parse, Push/Pop, Transpose…

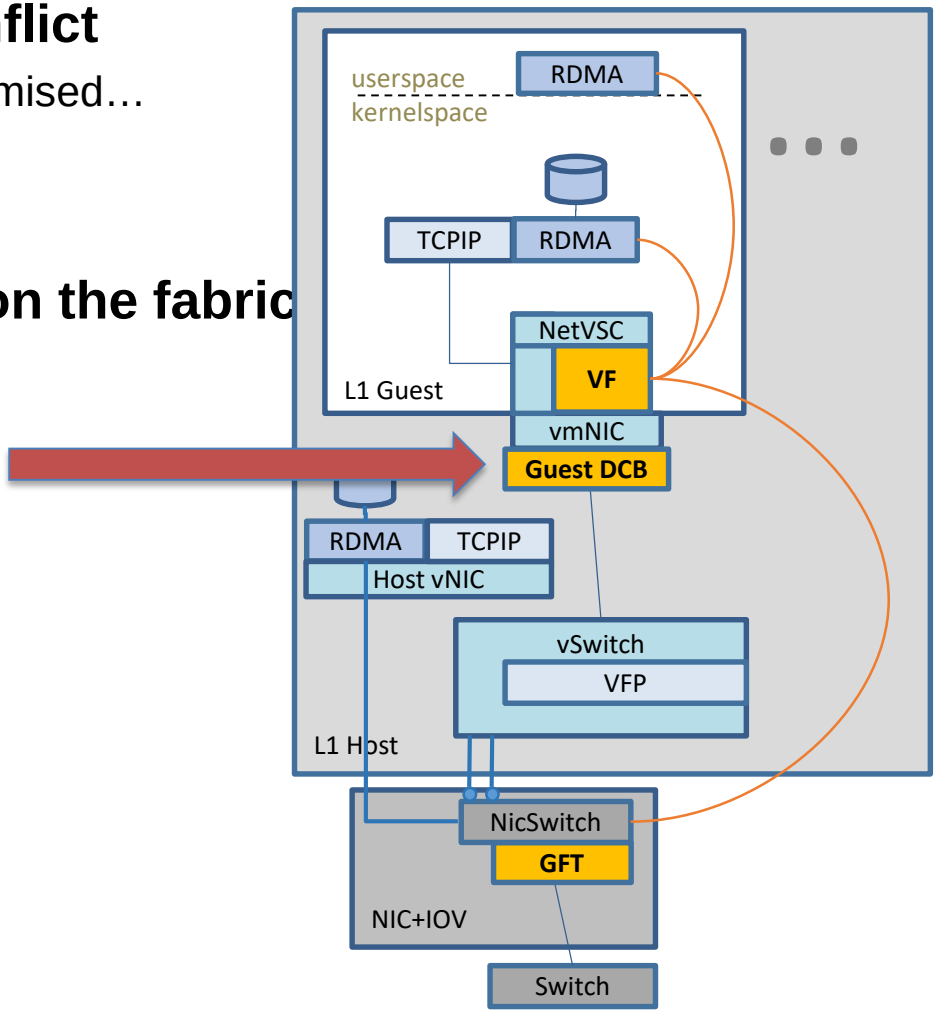


MULTITENANCY – SECURING RDMA

- **Performance and Security conflict**
  - VFs bypass security… Fabric compromised…
  - Acceptable for trusted Guests
- **How can we secure tenants?**

1. Control <u>what</u> tenant places on the fabric

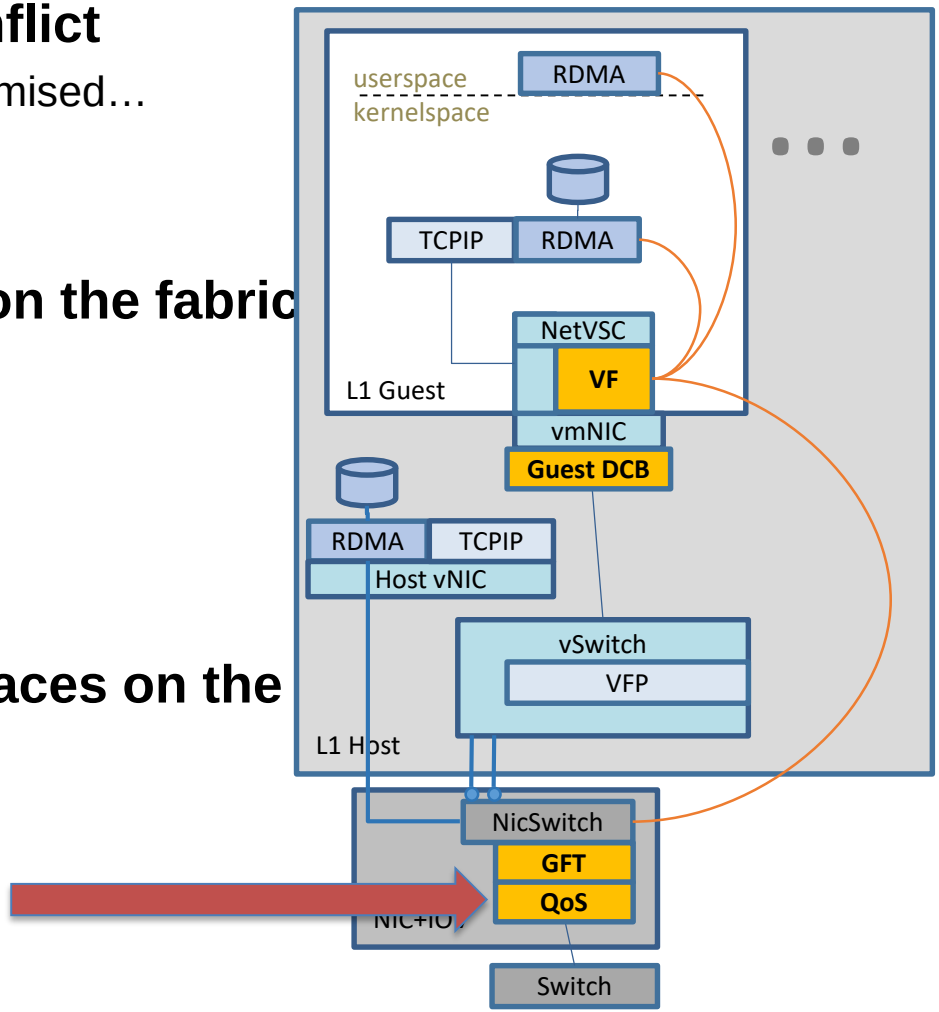
- GFT – Generic Flow Tables
  - Parse, Push/Pop, Transpose…
- Tenant DCB
  - VF level conversion
  - Automatic DCB correction



MULTITENANCY – SECURING RDMA

- **Performance and Security conflict**
  - VFs bypass security… Fabric compromised…
  - Acceptable for trusted Guests
- **How can we secure tenants?**

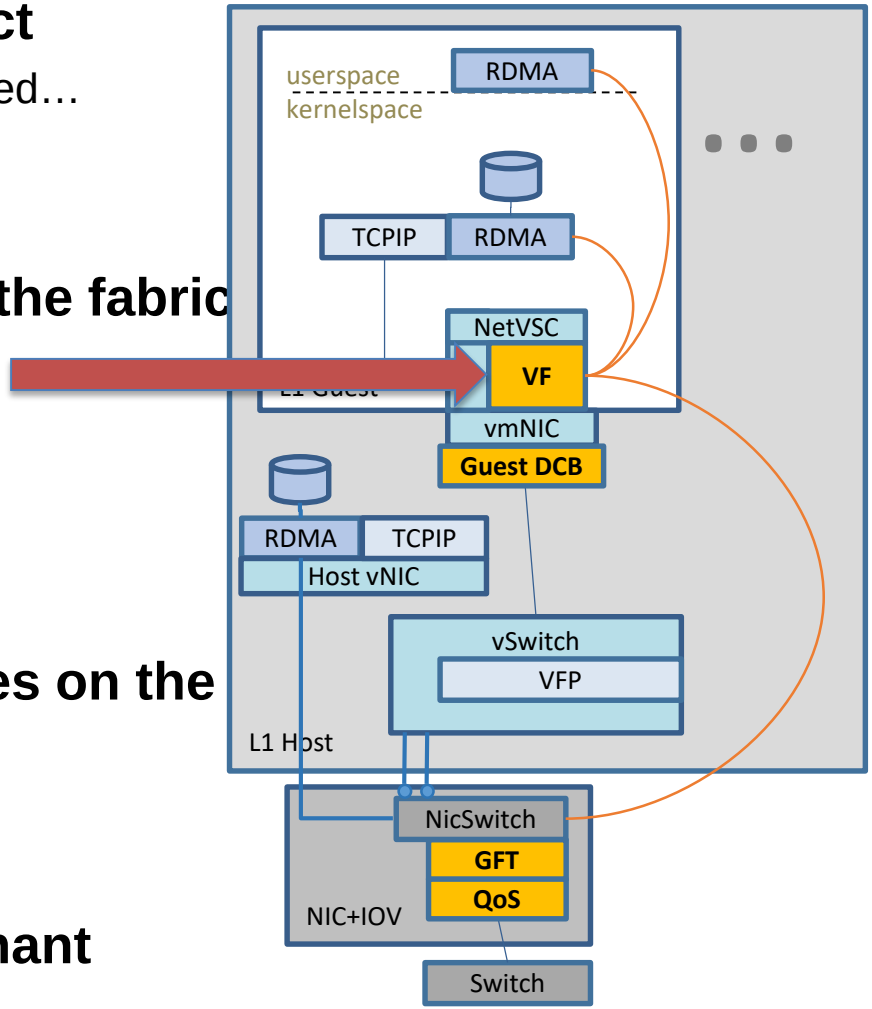
1. **Control <u>what</u> tenant places on the fabric**
   - GFT – Generic Flow Tables
     - Parse, Push/Pop, Transpose…
   - Tenant DCB
     - VF level conversion
     - Automatic DCB correction
2. **Control <u>how much</u> tenant places on the fabric**
   - Per-TC HW QoS
   - Send: Caps/Reservations. Recv: Caps



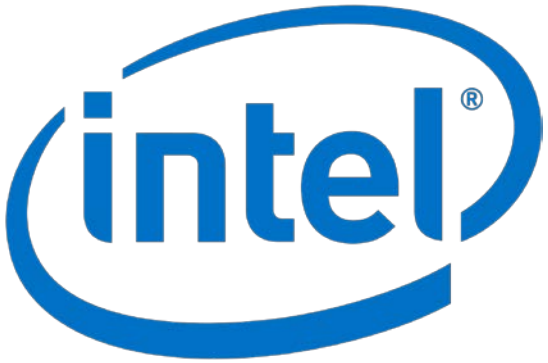
MULTITENANCY – SECURING RDMA

- **Performance and Security conflict**
  - VFs bypass security… Fabric compromised…
  - Acceptable for trusted Guests
- **How can we secure tenants?**

1. **Control <u>what</u> tenant places on the fabric**
   - GFT – Generic Flow Tables
     - Parse, Push/Pop, Transpose…
   - Tenant DCB
     - VF level conversion
     - Automatic DCB correction
2. **Control <u>how much</u> tenant places on the fabric**
   - Per-TC HW QoS
   - Send: Caps/Reservations. Recv: Caps
3. **Control <u>what</u> HW resources tenant consumes**
   - Limit QP, CQ, PD, MR, etc.



RDMA HW PARTNERS



14th ANNUAL WORKSHOP 2018

THANK YOU

Omar Cardona

Microsoft

14th ANNUAL WORKSHOP 2018

We're Hiring!
Help us build the future of Multi-tenant Secure RDMA

ocardona@microsoft.com