RDMA over Converged Ethernet

Darren J. Harkins – Staff Systems Engineer

May 2017

Mellanox® TECHNOLOGIES
Connect. Accelerate. Outperform.™

RDMA: Remote Direct Memory Access (RDMA)

What is RDMA?

Direct memory access from the memory of one computer to that of another without involving either one's operating system. This permits high-throughput, low-latency networking, omitting the OS and freeing the Processor to other tasks.

✓Higher ***performance*** and lower latency by offloading CPU transport processing.

✓Remote storage at the ***speed*** of direct attached storage (Including 100Gb/s InfiniBand and RoCE*)

- **Enabling Mobility, Scalability & Serviceability**
  - **More User, Scalability & Simplified Management**
  - **Dramatically Lowers CPU Overhead & Reduces Cloud Application Cost**
  - **Highest Throughput (10/40/56/100GbE), SR-IOV & PCIe Gen3/4**

* RDMA Over Converged Ethernet

RoCE

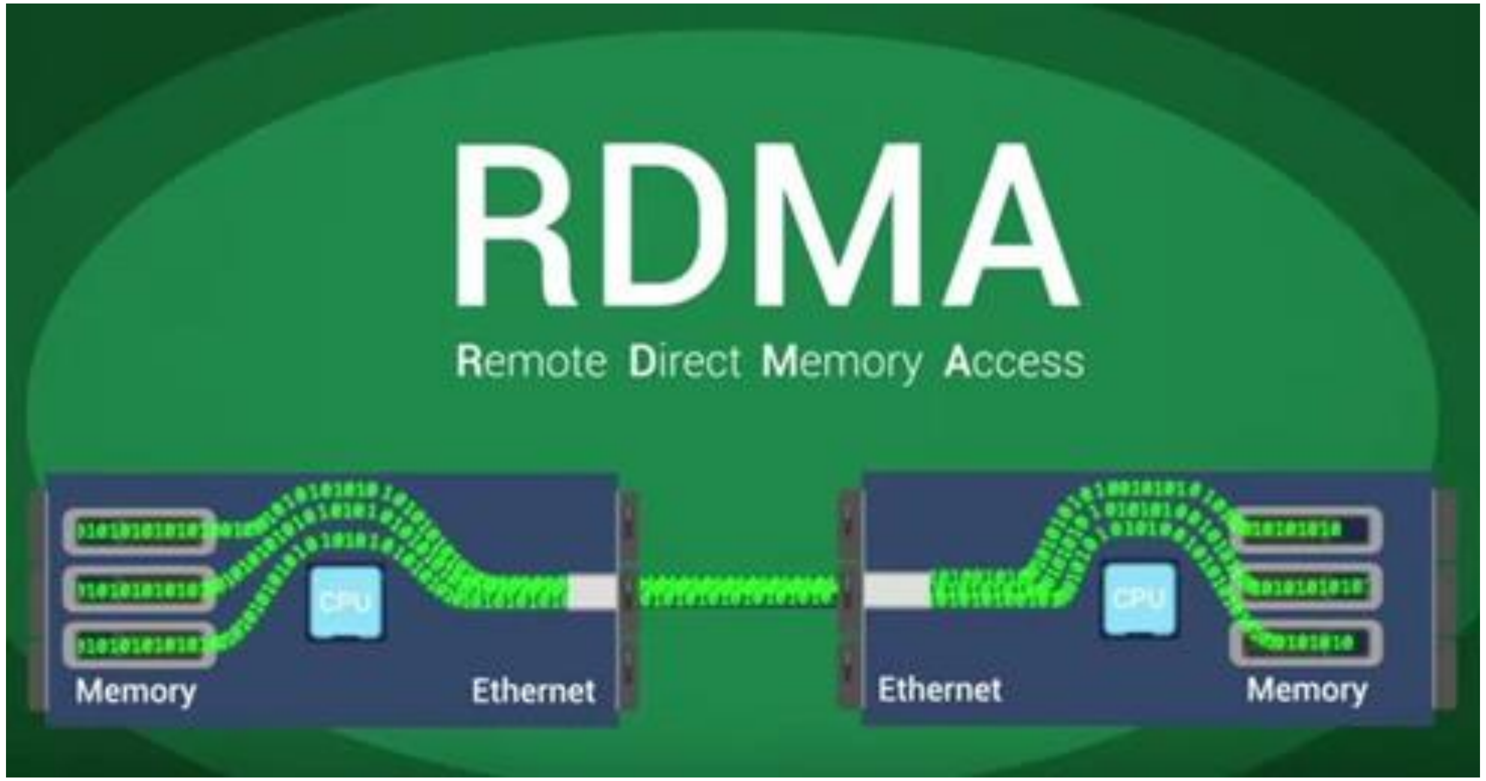
RDMA over Converged Ethernet

RoCE: RDMA over Converged Ethernet

- Well known on InfiniBand
- Works well on a lossless network
- Lower latency than alternative Transport protocols (TCP)
- Significantly lower overhead when offloaded to adaptor

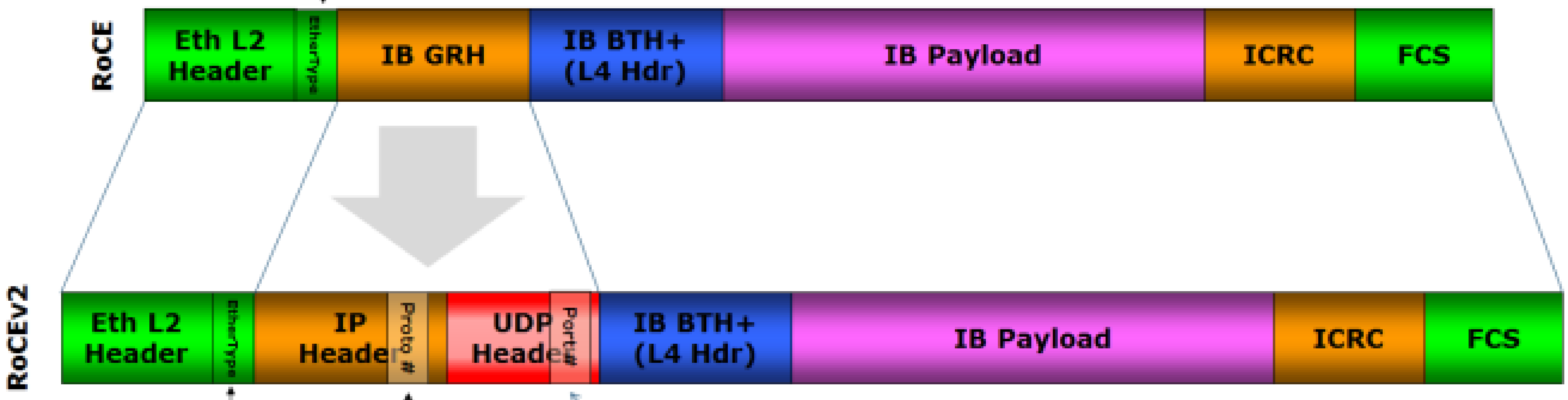
BUT

- Ethernet is not lossless by design
- PFC is required to achieve lossless Ethernet fabric
- PFC (Part of DCB)has a high configuration and management overhead – VLANs, Priorities
- PFC is Layer 2 only



RoCE : Frame Format

EtherType indicates that packet is RoCE (i.e. next header is IB GRH)



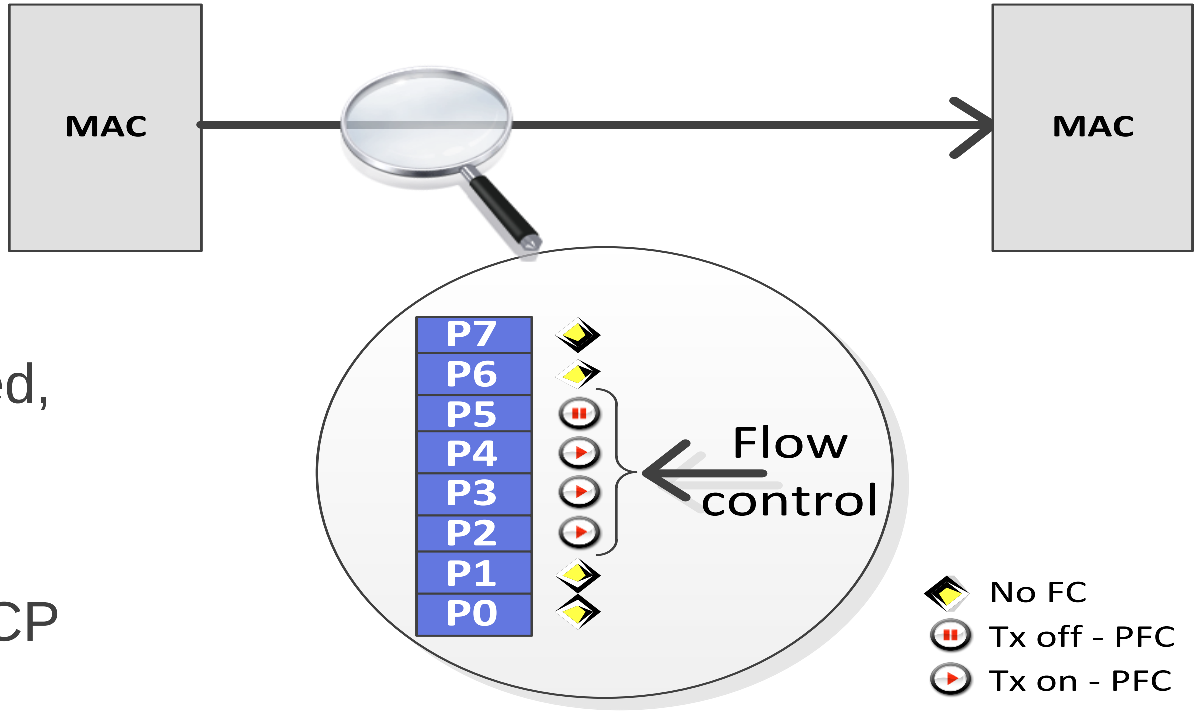
EtherType indicates that packet is IP (i.e. next header is IP)

ip.protocol_number indicates that packet is UDP

UDP dport number Indicates that next header is IB.BTH

Priority Flow Control (PFC)

- By nature Ethernet is a lossy network
- Ethernet provides flow control mechanism which makes it lossless – 2 options:
  - Applied FC over the whole port (Priority Flow Control - 802.3x)
  - Applied FC over specific priority (Priority Flow Control - 802.1Qbb)
- PFC negotiation between switch-host can be done by DCB (Data Center Bridging)
  - Using Data Center Bridging Exchange (DCBX) negotiation
  - End points (switch & host) exchange information about their capabilities
  - If PFC is supported, it will be used
  - If PFC is not supported, Global FC will be used
  - If DCBX is not supported or the PFC capability is not supported, manual configuration is required
- Routers rebuild the layer 2 header
  - Among it the routers rebuild the PCP filed using a DSCP to PCP mapping

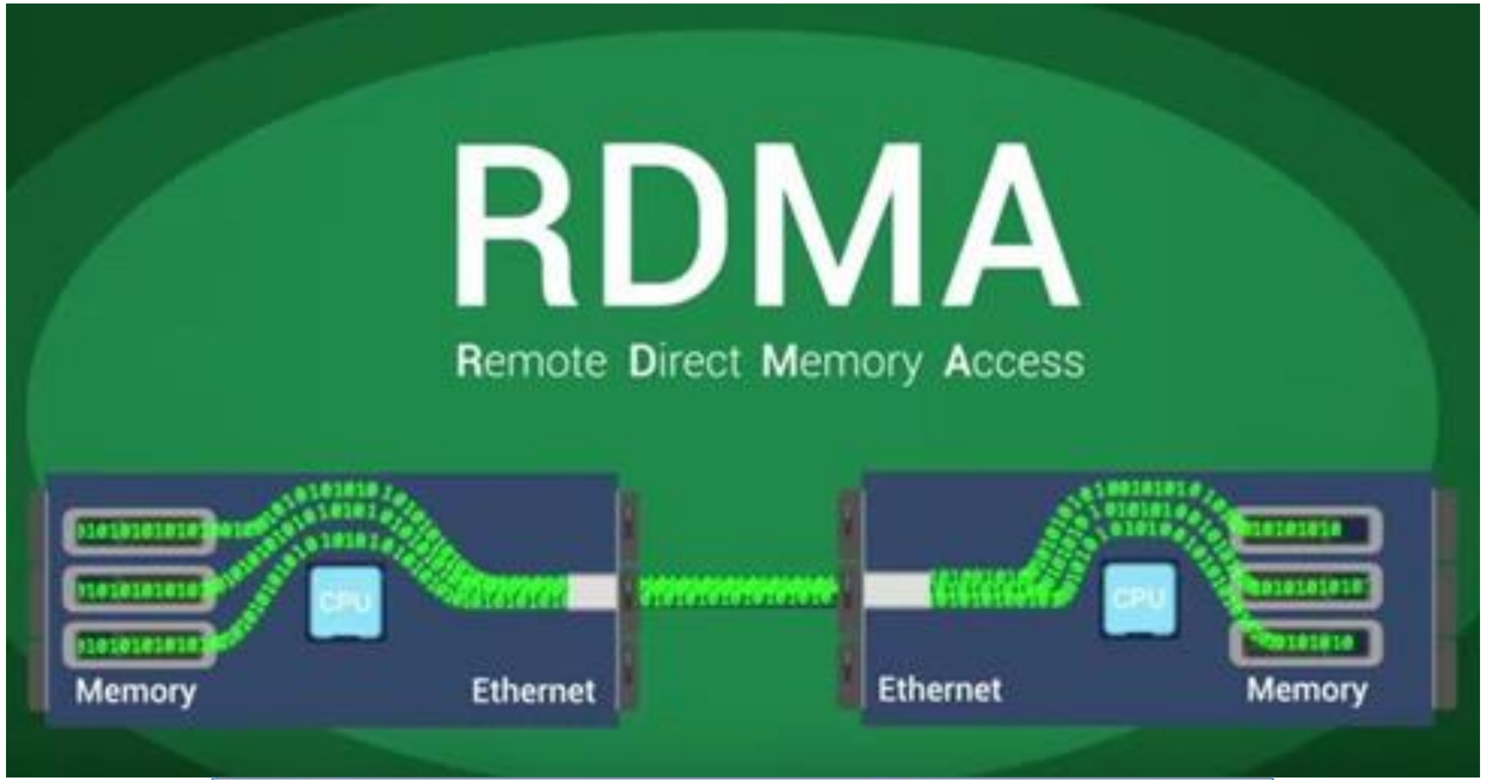


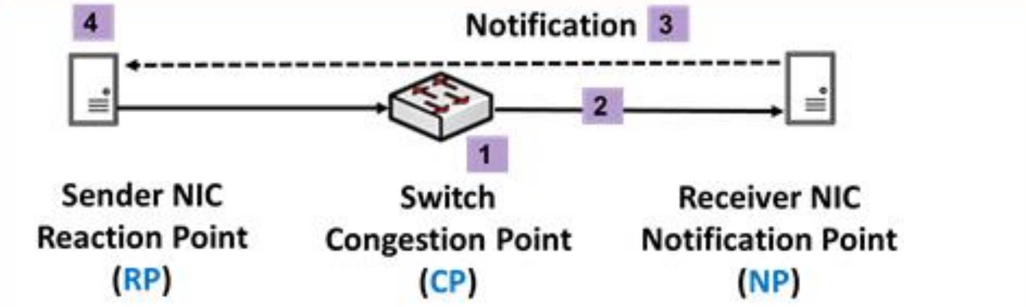
Routable RoCE

RDMA over Converged Ethernet at Layer 3

RoCEv2: Routable RDMA over Converged Ethernet

- Routable RoCE requires a higher level congestion mechanism
  - ECN – Explicit Congestion Notification
- ECN can slow down traffic to prevent congestion
- ECN configuration overhead is lower than PFC, simple and easy





1. Switch detects onset of congestion
2. Switch marks packets' CE (Congestion Experienced) bit
3. Receiver sees CE bit on packets and notifies sender
4. Sender slows transmission rate until congestion risk is gone

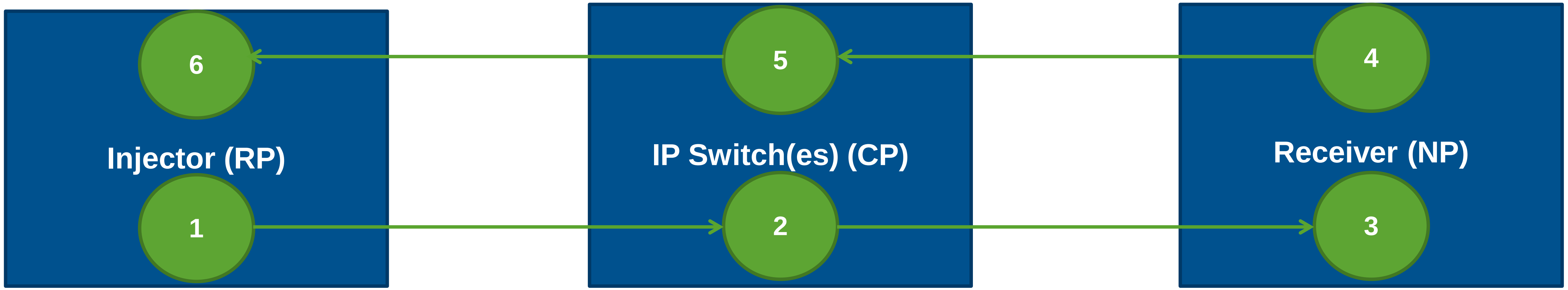
Existing Congestion Control Solutions

- L3/L4 solutions
  - TCP congestion control (Reno, New Reno, Vegas, Cubic)
    - Targets mostly long latency links
    - Buffer hog – fills the buffer to maximum available, relies on drops for signaling
    - Not optimized for data center usage, not optimized for lossless fabric
  - ECN
    - Improves performance in data center scenarios
    - Relies on explicit ACK/NACKs for each transmitted packet
    - Assumes software, TCP/IP latency (~100us)
  - IB Congestion Control
    - Similar to ECN
    - Explicit notifications for congestion marked packets, on special notification packets
- L2 solutions
  - QCN
    - Explicit notification when switch is suffering congestion
    - Host performs rate limit without relying on ACKs
    - L2 only – can’t go through routers

RoCEv2 Congestion Control

- Handles long-lived congestions over lossless fabric
- Per QP rate limitation according to signaling from the fabric
- Uses ECN markings in the IP header for congestion detection
  - No special functionality required
  - Compatible with most modern switches/routers in market
- Reflects the marking to the traffic source using special notification packet
  - Similar to InfiniBand CNP packet
  - Can be on a different, higher priority
- Utilizes a protocol inspired by DCTCP and QCN to control the rate
  - DCTCP provides estimation of the congestion severity in the network
  - QCN decides the transmission speed per QP according to the DCTCP estimation
- Available in Connect-X 3 Pro and above

RoCEv2 Congestion Control – Cont.



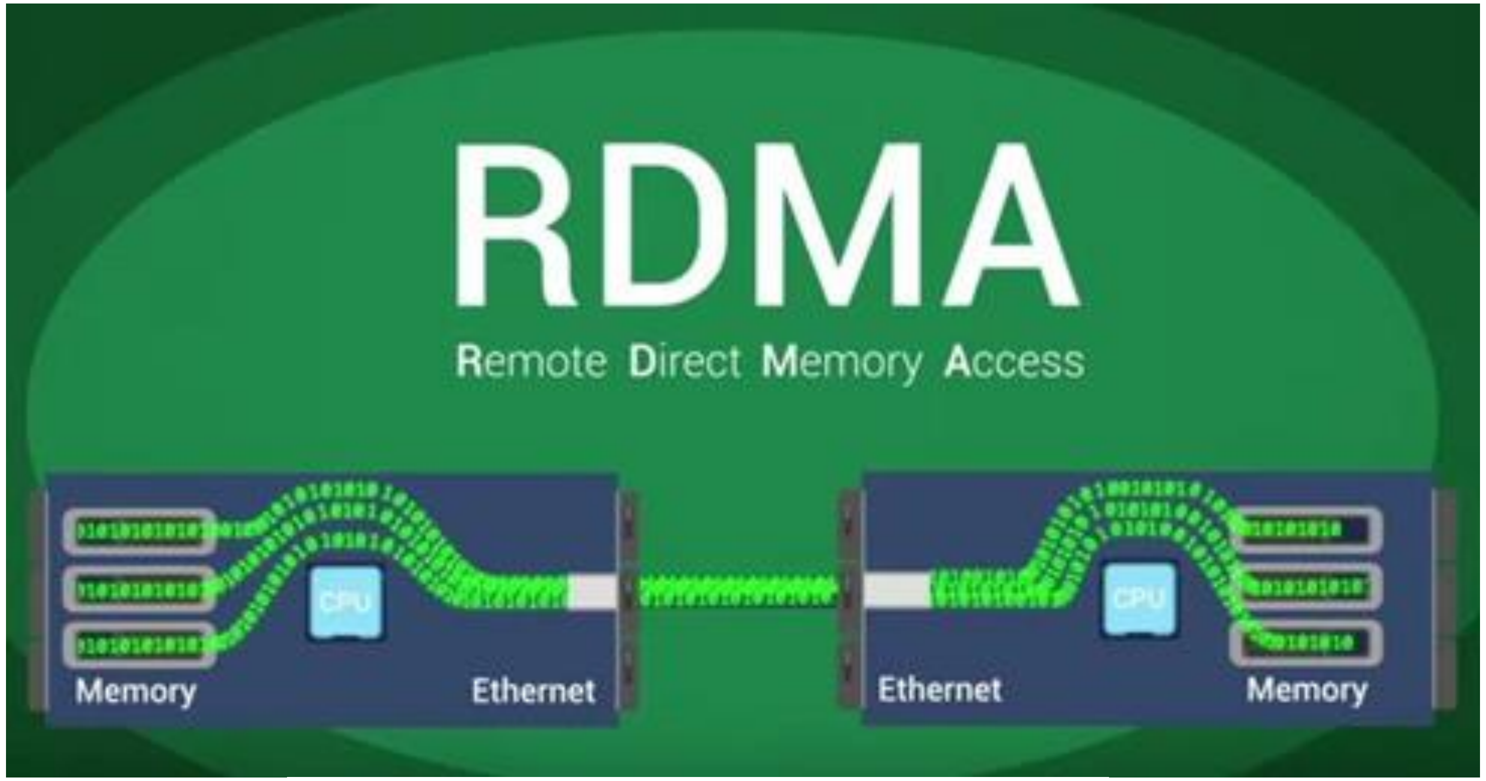
1. Reaction Point (RP) injects ECN-capable packets to network
2. Switch (Congestion Point - CP) marks packets when congestion occurs
3. Notification Point (NP) records the marked packets
4. NP sends periodic information to the RP about the congestion marking observed (“CNP” packet)
5. Switch forwards the notification as a usual packet
6. RP sees the CNP packets, estimates network congestion state and reduces speed
7. RP increases speed when no CNP packets are received for some time

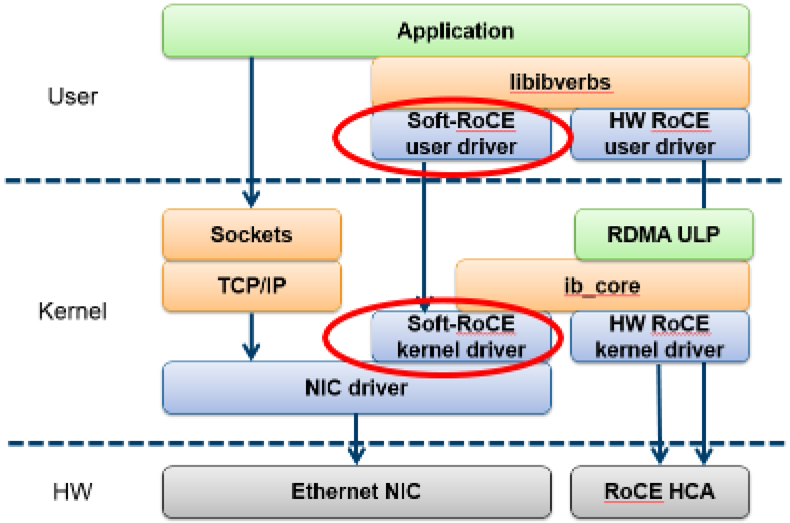
Soft RoCE

RDMA over Converged Ethernet on any NIC

SoftRoCE: RDMA over Converged Ethernet in Software

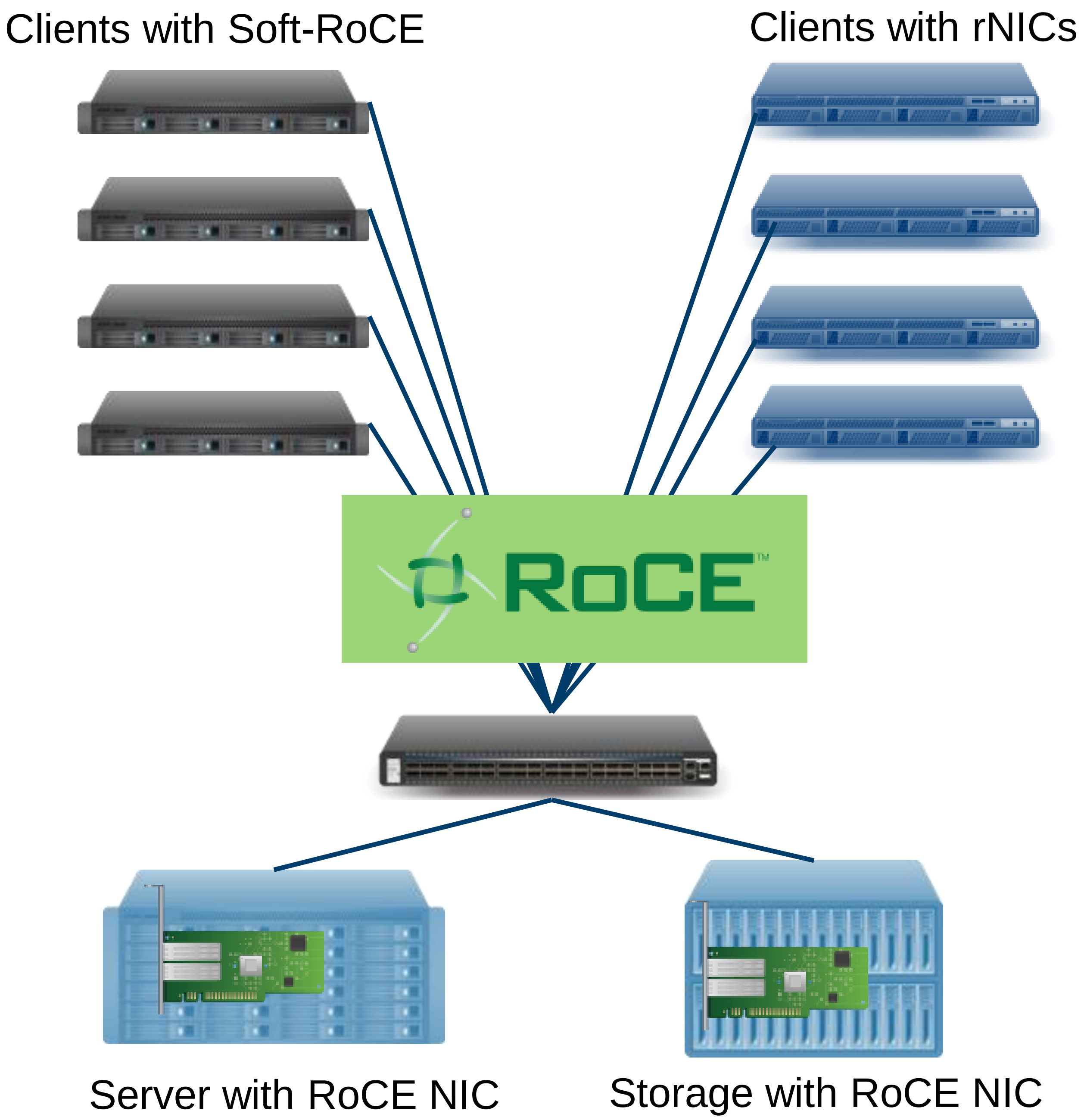
- SoftRoCE – allows non-offloaded adaptors to work with Hardware offloaded adaptors in the same fabric
- Part of MLNX-OFED from 4.0
- Allows integration of RoCE in to test environments





Soft-RoCE Allows Heterogeneous Deployments Anytime

- RoCE Enabled on Any Server, Any NIC
  - RDMA without hardware offload
  - Interoperates with hardware-accelerated RoCE
- Heterogeneous Deployments Can Use RDMA
  - Storage/server with RDMA hardware acceleration benefit from soft-RoCE clients
  - Deploy RoCE while rolling out RoCE adapters
  - Faster, easier prototyping, testing and development
- Top Use Cases
  - Storage array: iSER or NFSoRDMA
  - Clustered file systems: Lustre, GPFS, Gluster
  - Distributed or cloud applications

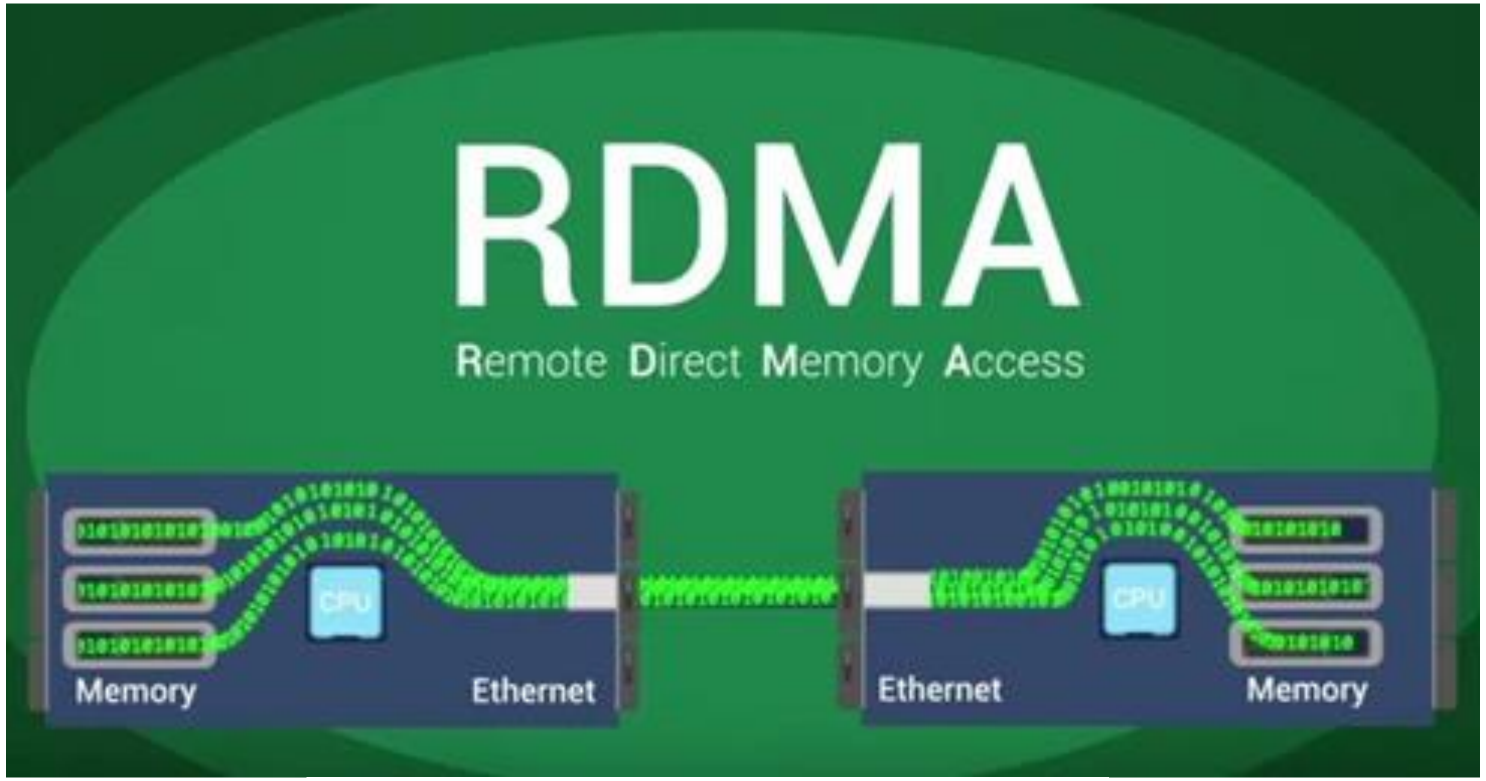


Resilient RoCE

RDMA over Converged Ethernet with re-trans and re-order

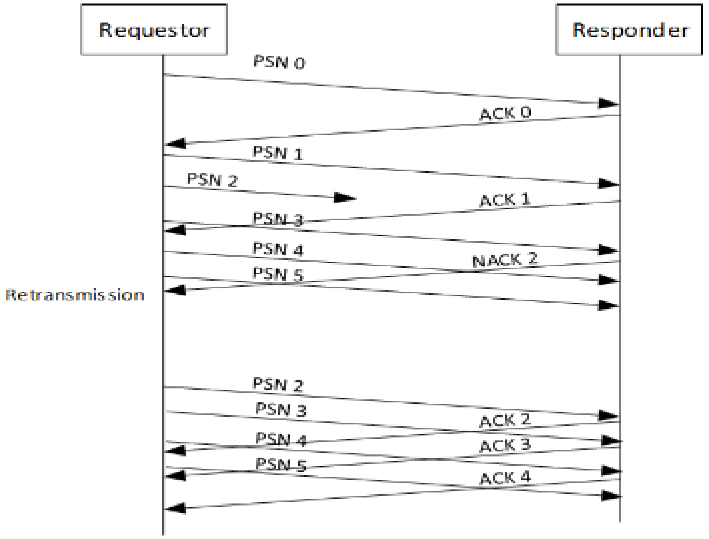
RoCE: Resilient RDMA over Converged Ethernet

- Resilient RoCE can cope with packet loss and Out of Order packets
- ECN is suggested but not required
- Out of Order packets are held in buffer to fill the gaps. Re-ordered packets are then written to memory
- Missing packets are requested from the sender



SO

- No loss – everything is fast
- Some loss – slows down, but stays in working order
- Still significantly better than TCP/IP



SR-IOV and RoCE

RoCE in Virtual Machines

RoCE Is an Open Standard

- IBTA and IETF
  - Steering Committee: Cray, Emulex, HP, IBM, Intel, Mellanox, Microsoft, Oracle,
  - RoCE specification first released in 2010
  - Most widely deployed Ethernet RDMA standard

- Multi-Vendor Support
  - RoCE NICs Today: Mellanox & Emulex
  - Other NIC vendors plan to support
  - Soft-RoCE on any Ethernet adapter (with PFC capability)
  - Almost any data center switch

INFINIBAND™ TRADE ASSOCIATION

IETF®

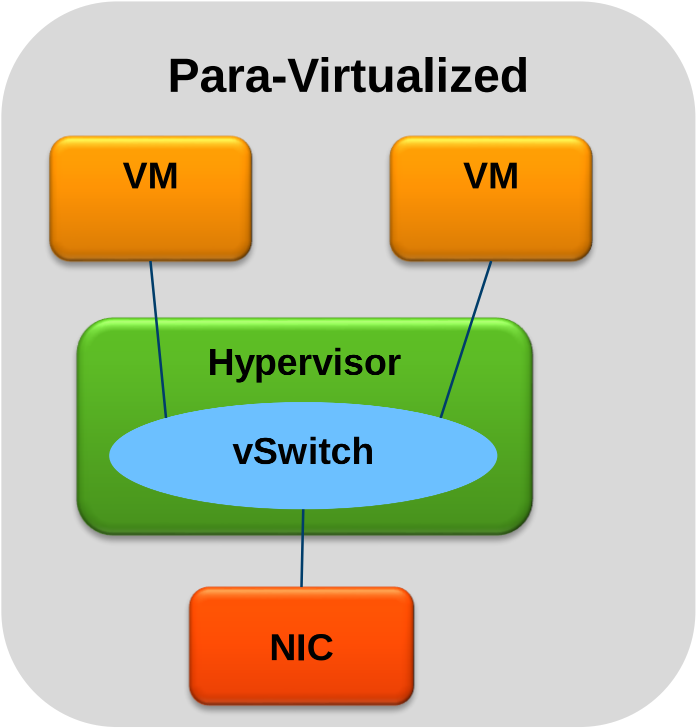
RoCE™

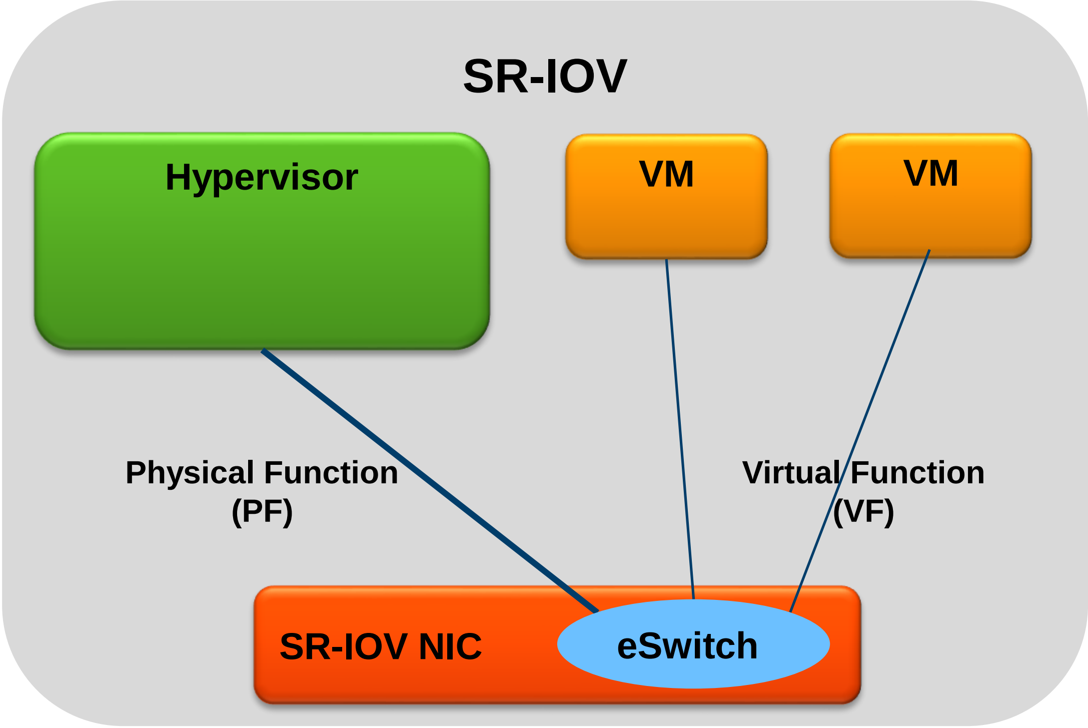
RoCE - Highlights

- RDMA Verbs API
  - Transparent to Applications/ULPs
- Ethernet Management Practices
- Purpose Built IB-RDMA Transport Protocol
  - Connected Services (RDMA and Send/Recv)
  - Datagram Services
  - Atomic Operations
  - User Level Multicast
- User Level IO Access / Kernel Bypass / Zero Copy
- RoCE De-multiplexing (Converged NICs)
  - Based on Ethertype – RoCEv1
  - Based on UDP d.port – RoCEv2

Reminder: Single Root I/O Virtualization (SR-IOV)

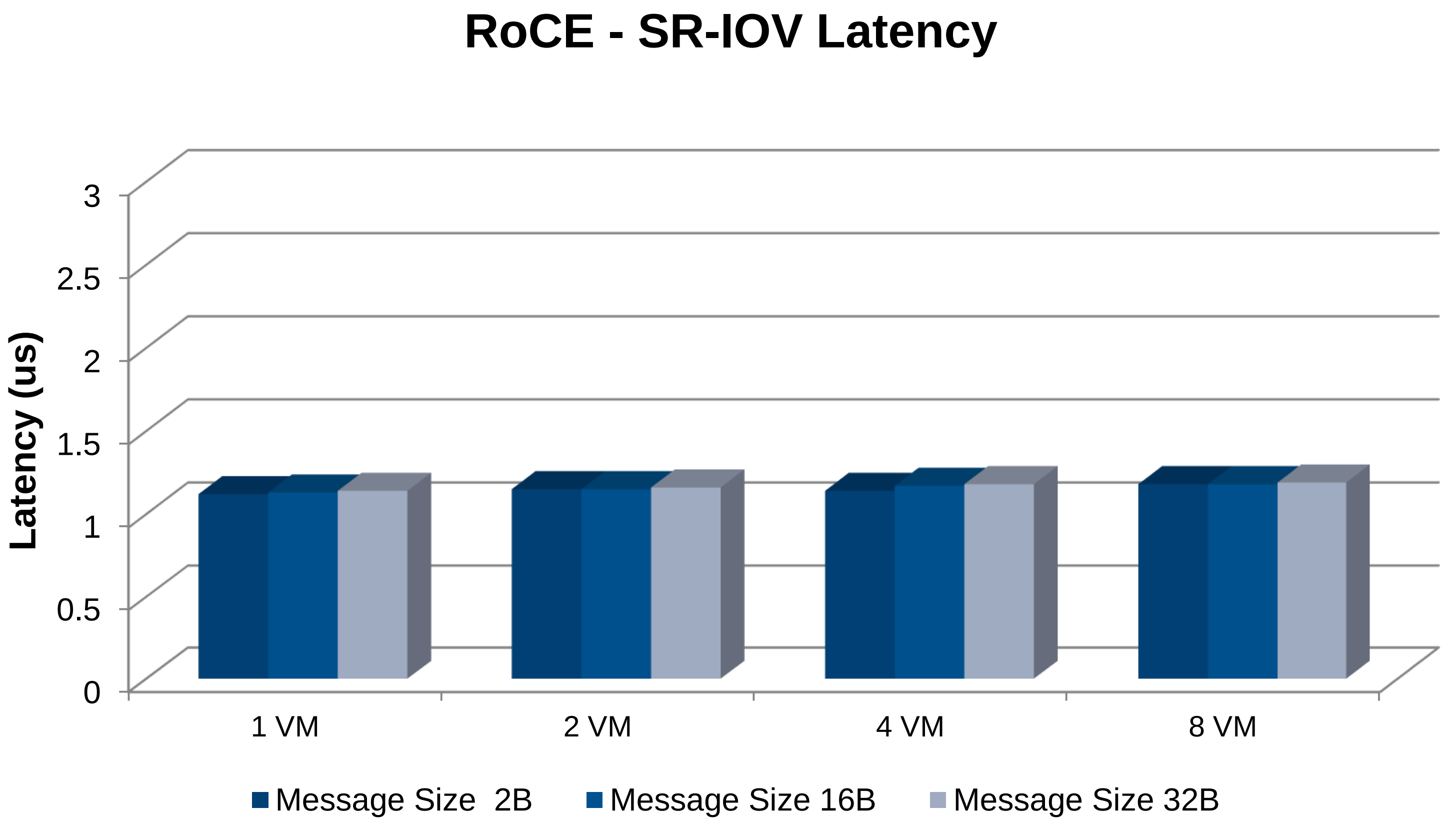
- PCIe device presents multiple instances to the OS/Hypervisor
- Enables Application Direct Access
  - Bare metal performance for VM
  - Reduces CPU overhead
- Enable RDMA to the VM
  - Low latency applications benefit from the Virtual infrastructure

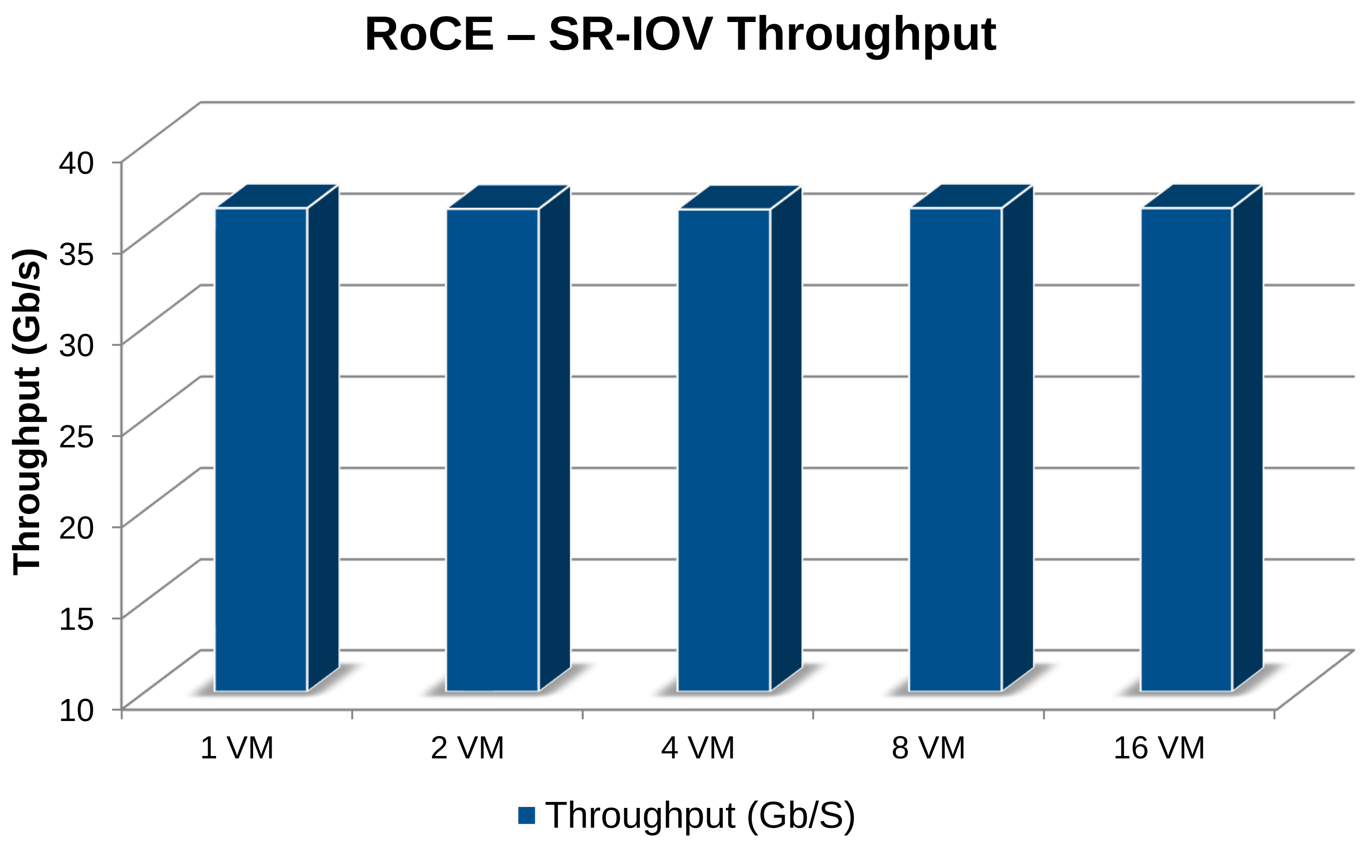




SR-IOV Boosts Ethernet Performance

- SR-IOV Accelerates RoCE
  - Enables native RoCE performance in virtualized environments

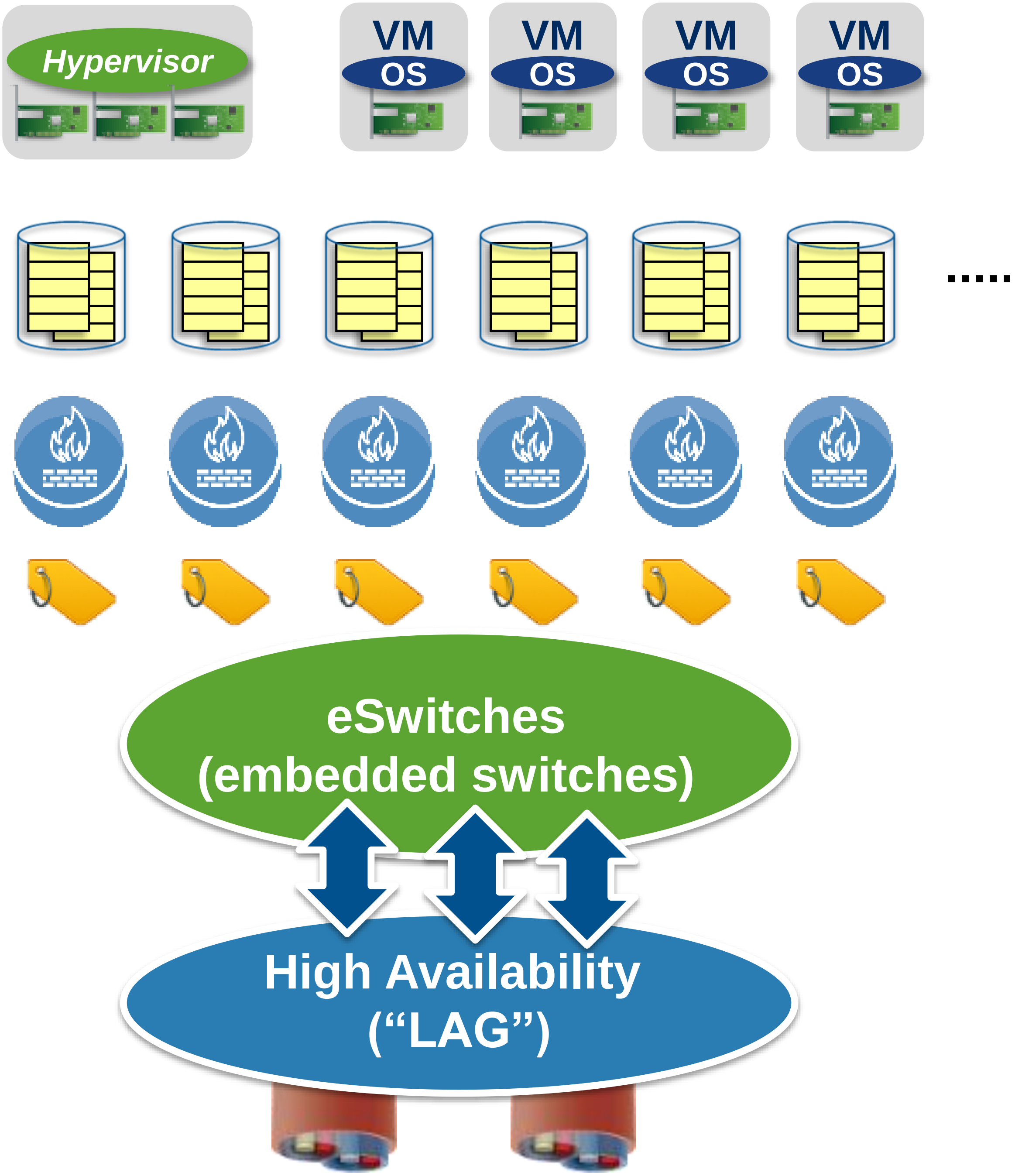




No Performance Compromise in Virtualized Environment

Advanced Virtualization & eSwitch Capabilities

- Scale up Virtualization:
  - High scale SR-IOV with 127 Virtual Functions (VFs)
  - 512 schedule queues
- Advanced virtualization solutions dictate Hypervisors bypass to enable optimal performance (SR-IOV)
- Hypervisor Bypass requires embedded-switch
  - VM Switching & QoS
  - Congestion Control
  - Security filters (ACLs, anti-spoofing)
  - L2 Tunneling
  - High Availability (HW based LAG)
- Benefits
  - Maximize performance in virtualized environment bypassing OS hypervisors



Question Time

Darren Harkins

darren@mellanox.com

+44 (0) 7944 786 208

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™