



Introduction to Infiniband

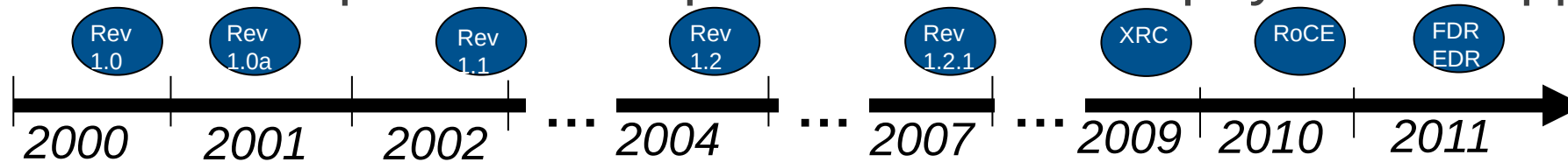
FRNOG 22, April 4th 2014

Yael Shenhav, Sr. Director of EMEA, APAC FAE, Application Engineering

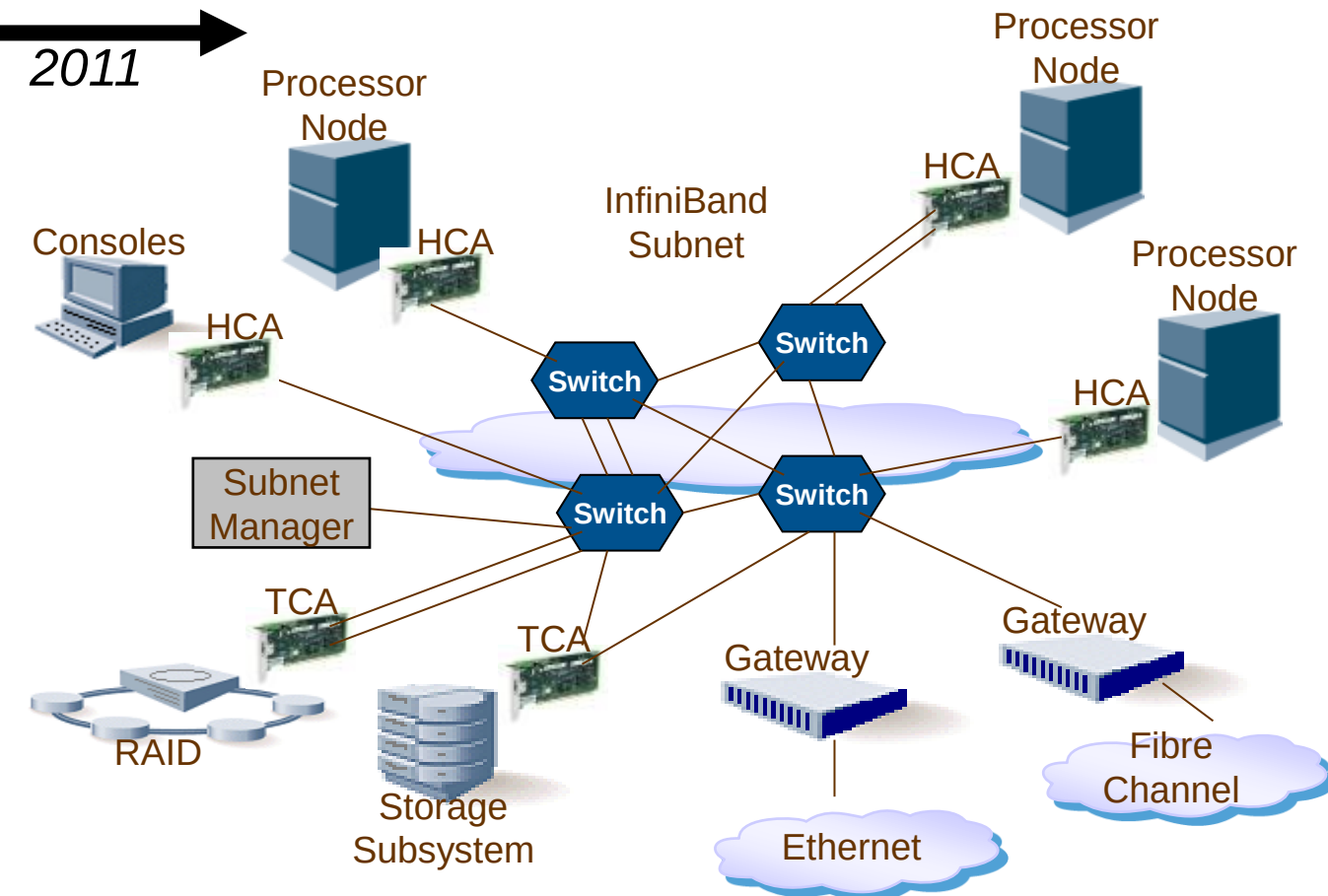
The InfiniBand Architecture



- Industry standard defined by the InfiniBand Trade Association
- Defines System Area Network architecture
 - Comprehensive specification - from physical to applications



- Architecture supports
 - Host Channel Adapters (HCA)
 - Target Channel Adapters (TCA)
 - Switches
 - Routers



■ Performance

- 56Gb/s per link shipping today
- Down to 0.6us application to application latency
- Aggressive roadmap

■ Reliable and lossless fabric

- Link level flow control
- Congestion control to prevent HOL blocking

■ Efficient

- Transport Offload
- Kernel bypass
- RDMA and atomic operations

■ QoS

■ Virtualization

■ Application acceleration

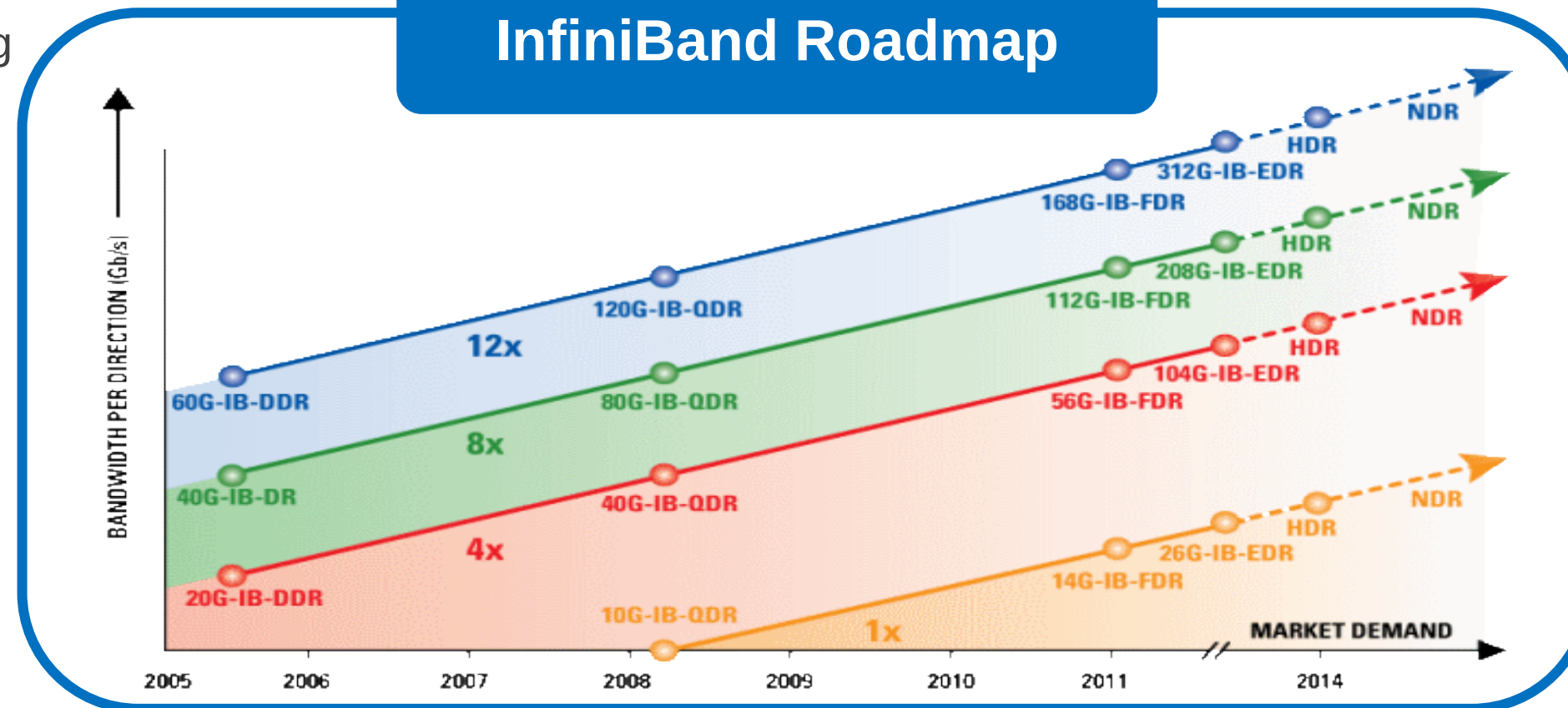
■ Scalable beyond Petascale computing

■ I/O consolidation

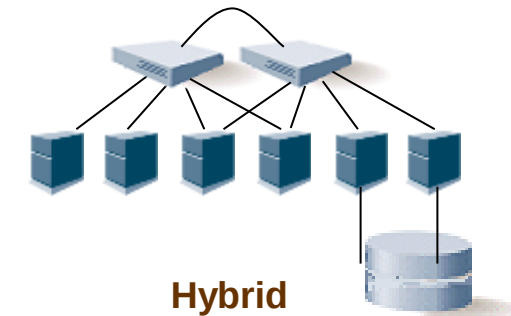
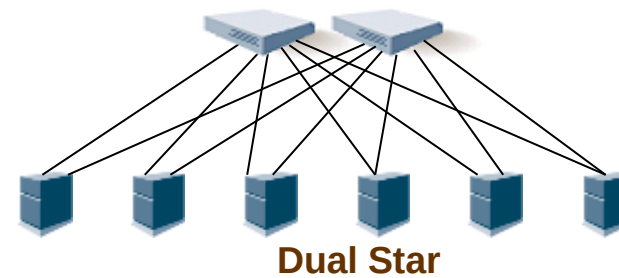
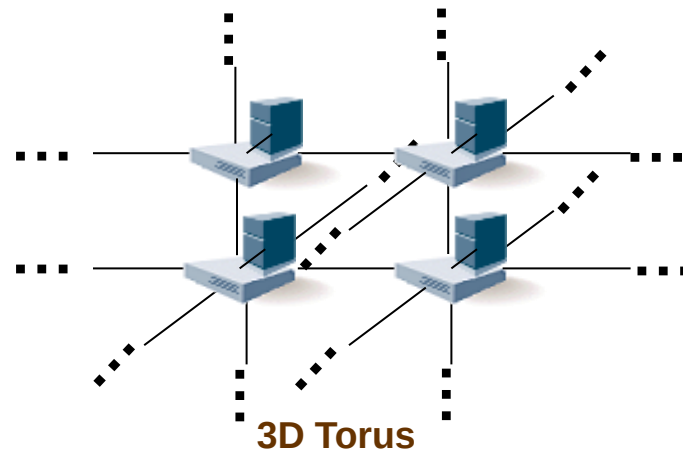
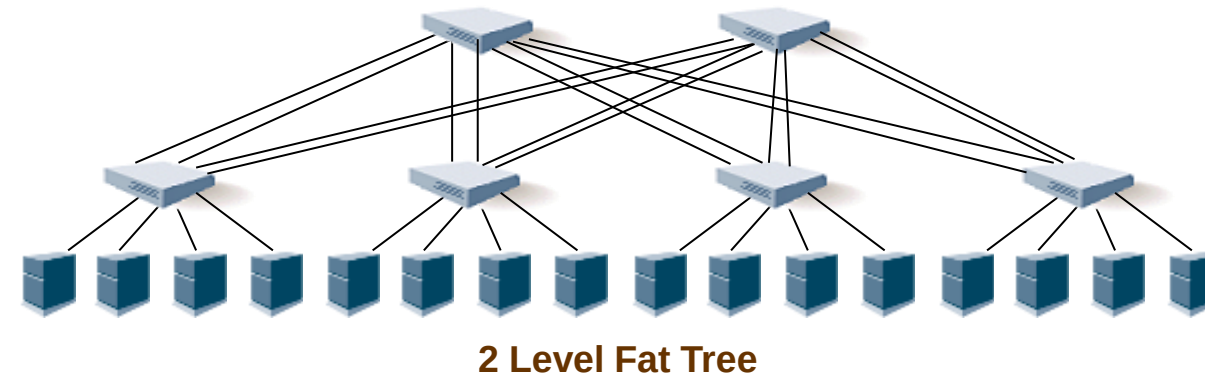
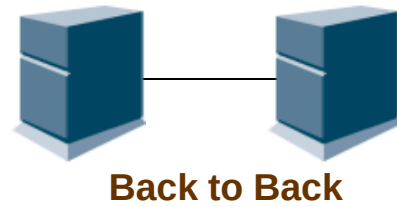
■ Simplified Cluster Management

- Centralized route manager
- In-band diagnostics and upgrades

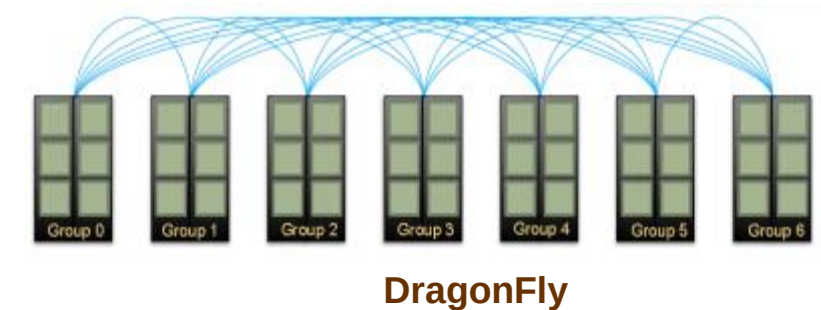
InfiniBand Roadmap



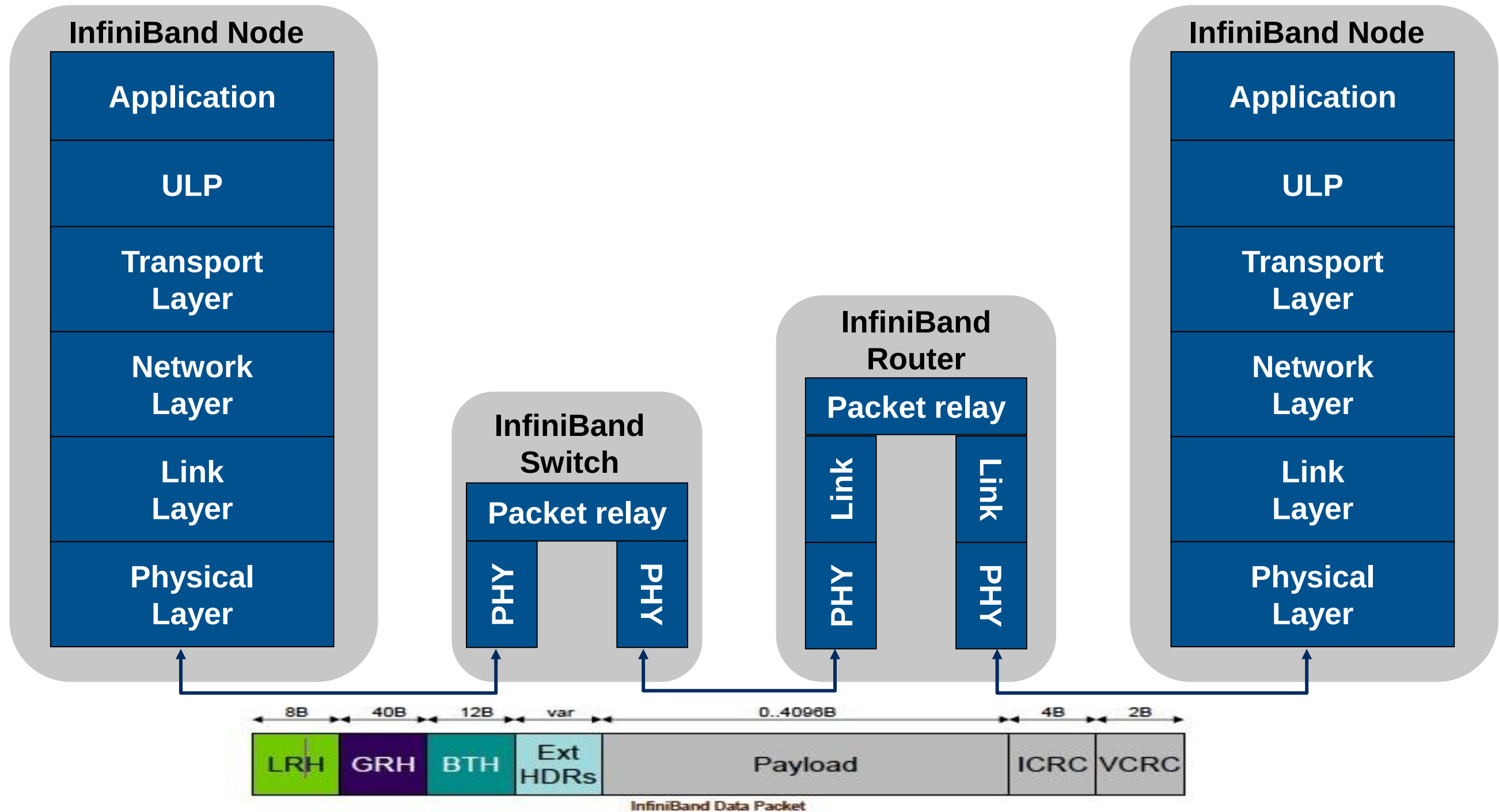
Source: www.infinibandta.org



- Scalability to 1,000s and 10,000s of nodes
- Full/configurable CBB ratio
- Multi-pathing
- Modular switches are based on fat tree architecture



InfiniBand Protocol Layers



■ Media types

- PCB: several inches
- Passive copper: 20m SDR, 10m DDR, 7m QDR, 3m FDR
- Fiber: 300m SDR, 150m DDR, 100/300m QDR&FDR

■ Link encoding

- SDR, DDR, QDR: 8 to 10 bit encoding
- FDR, EDR: 64 to 66 bit encoding

■ Industry standard components

- Copper cables / Connectors
- Optical cables
- Backplane connectors



4X QSFP Copper



4x QSFP Fiber



FR4 PCB

■ Addressing and Switching

- Local Identifier (LID) addressing
- Unicast LID - 48K addresses
- Multicast LID – up to 16K addresses
- Efficient linear lookup
- Cut through switching supported
- Multi-pathing support through LMC

■ Independent Virtual Lanes

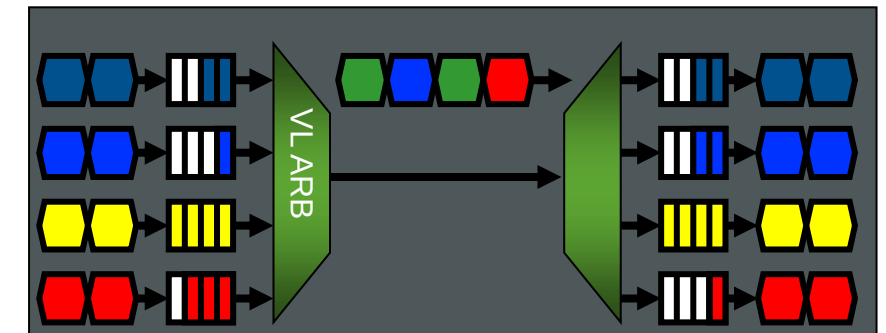
- Flow control (lossless fabric)
- Service level
- VL arbitration for QoS

■ Congestion control

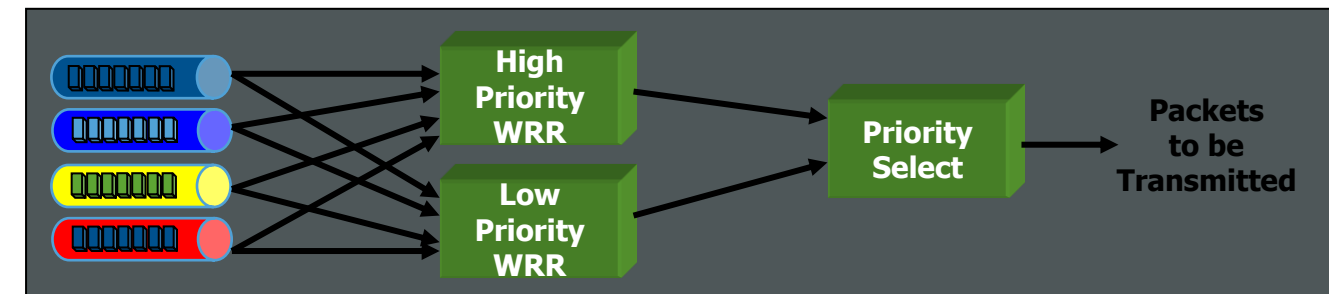
- Forward / Backward Explicit Congestion Notification (FECN/BE CN)

■ Data Integrity

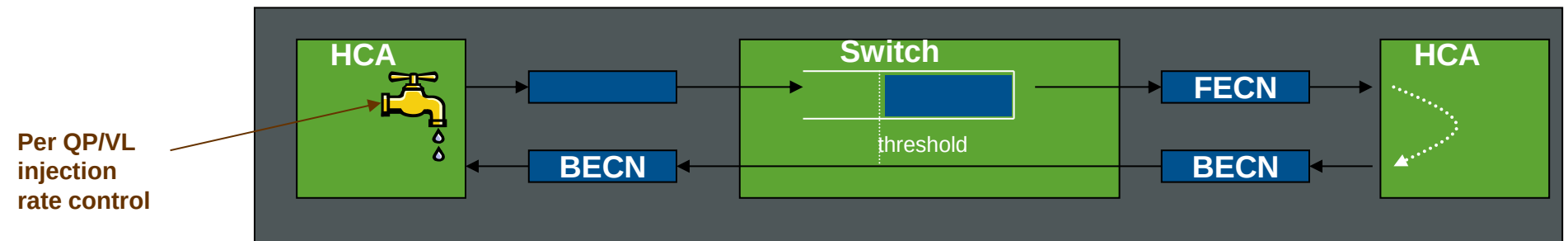
- Invariant CRC
- Variant CRC



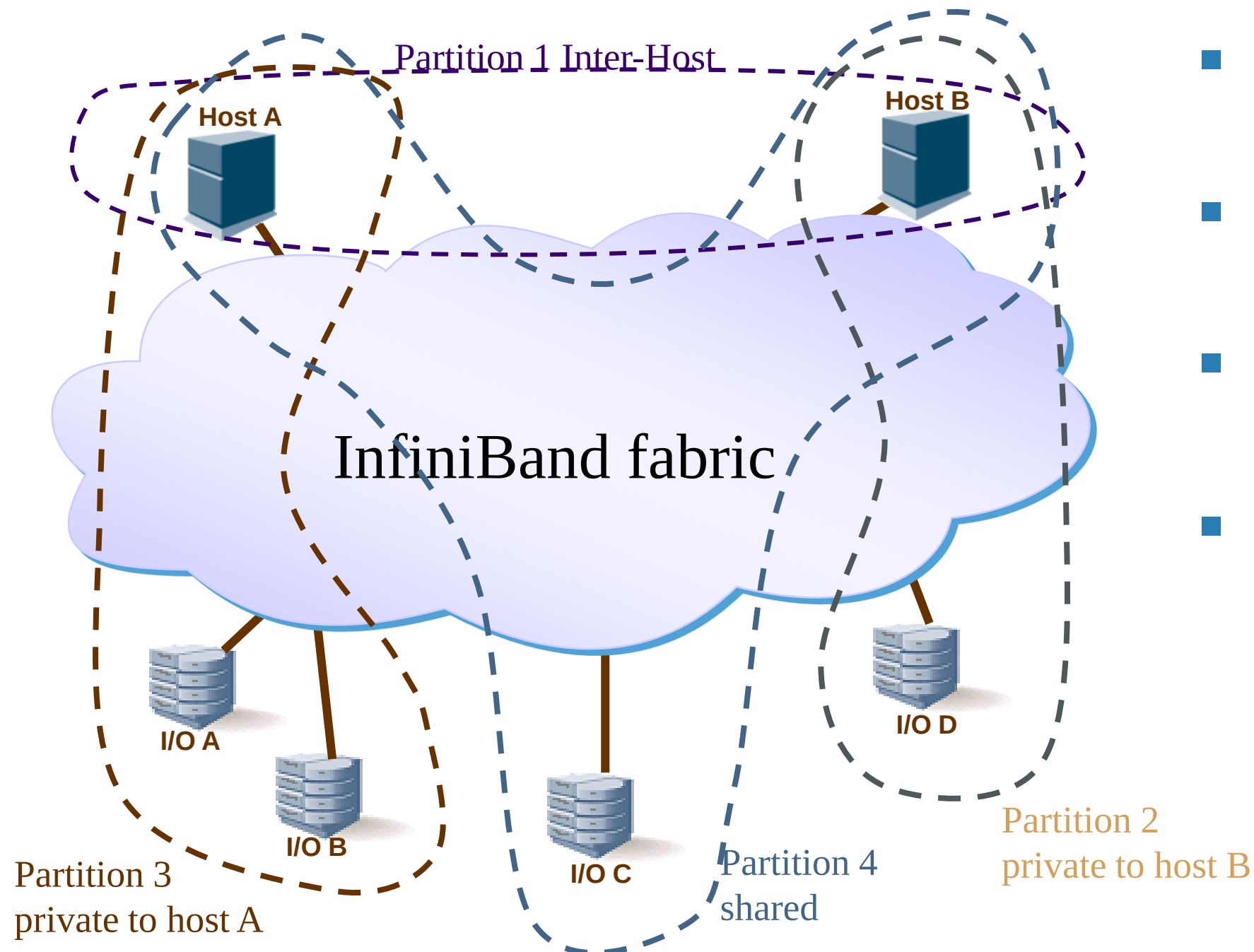
Independent Virtual Lanes (VLs)



H/L Weighted Round Robin (WRR) VL Arbitration

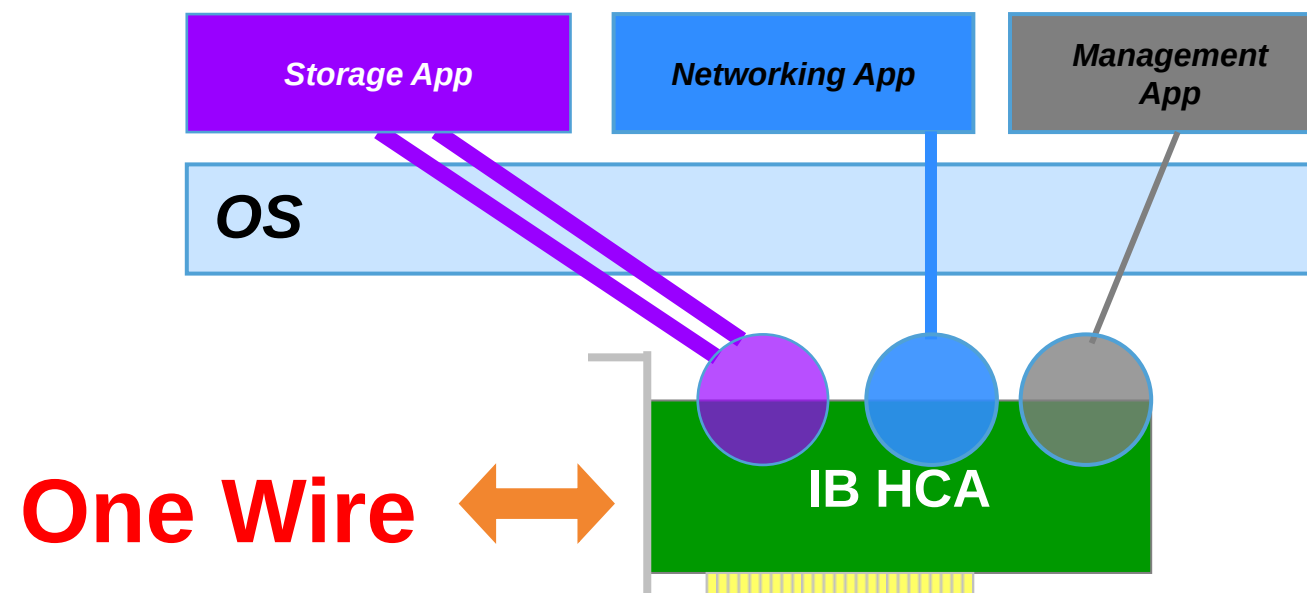


Efficient FECN/BE CN Based Congestion Control



- Logically divide the fabric into isolated domains
- Partial and full membership per partition
- Partition filtering at switches
- Similar to
 - FC Zoning
 - 802.1Q VLANs

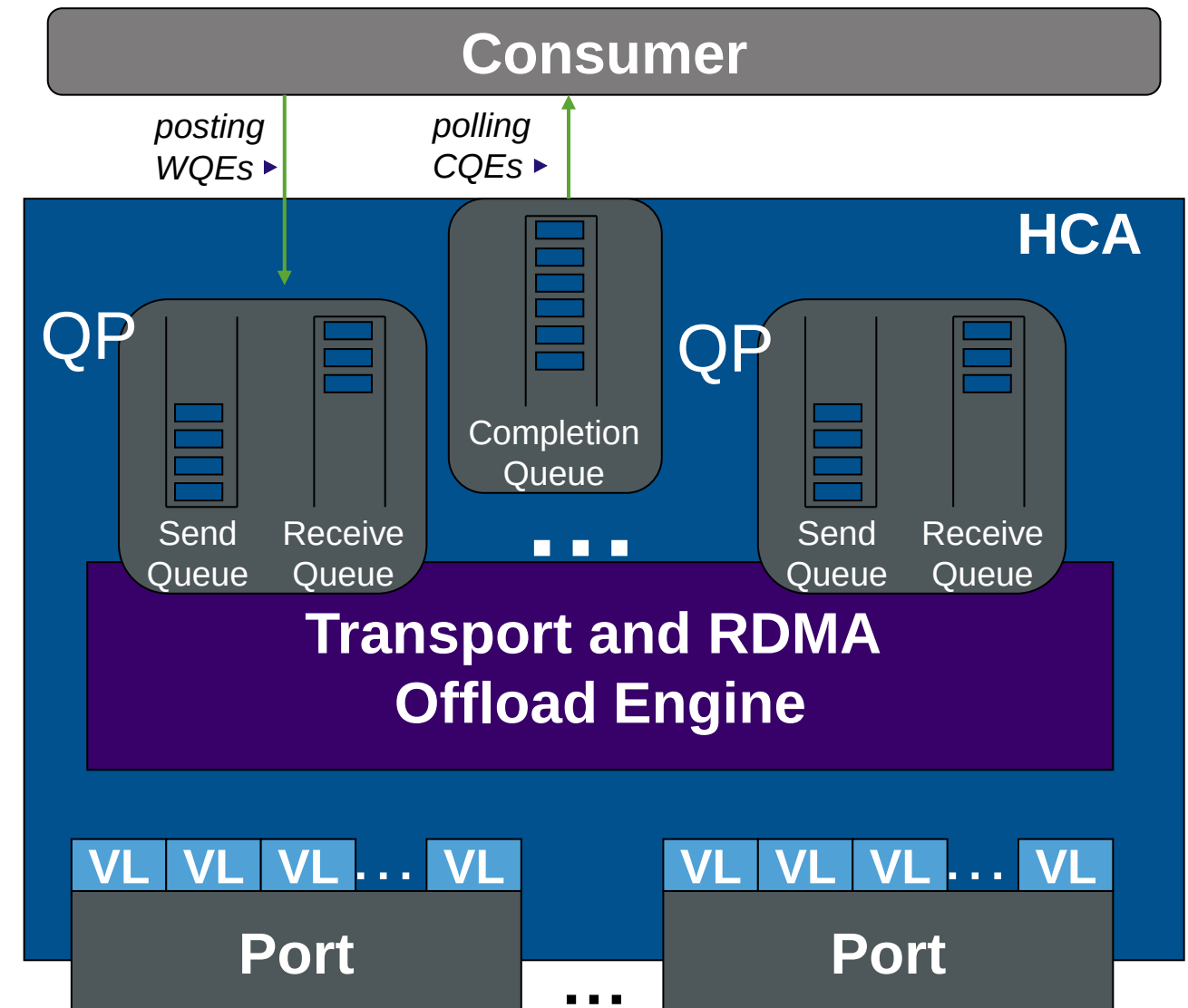
Fabric Consolidation with InfiniBand/RoCE



- High bandwidth pipe for capacity provisioning
- Dedicated I/O channels enable convergence
 - For Networking, Storage, Management
 - Application compatibility
 - QoS - differentiates different traffic types
 - Partitions – logical fabrics, isolation
 - Virtualization with bare metal performance
- Flexibility
 - Soft servers / fabric repurposing

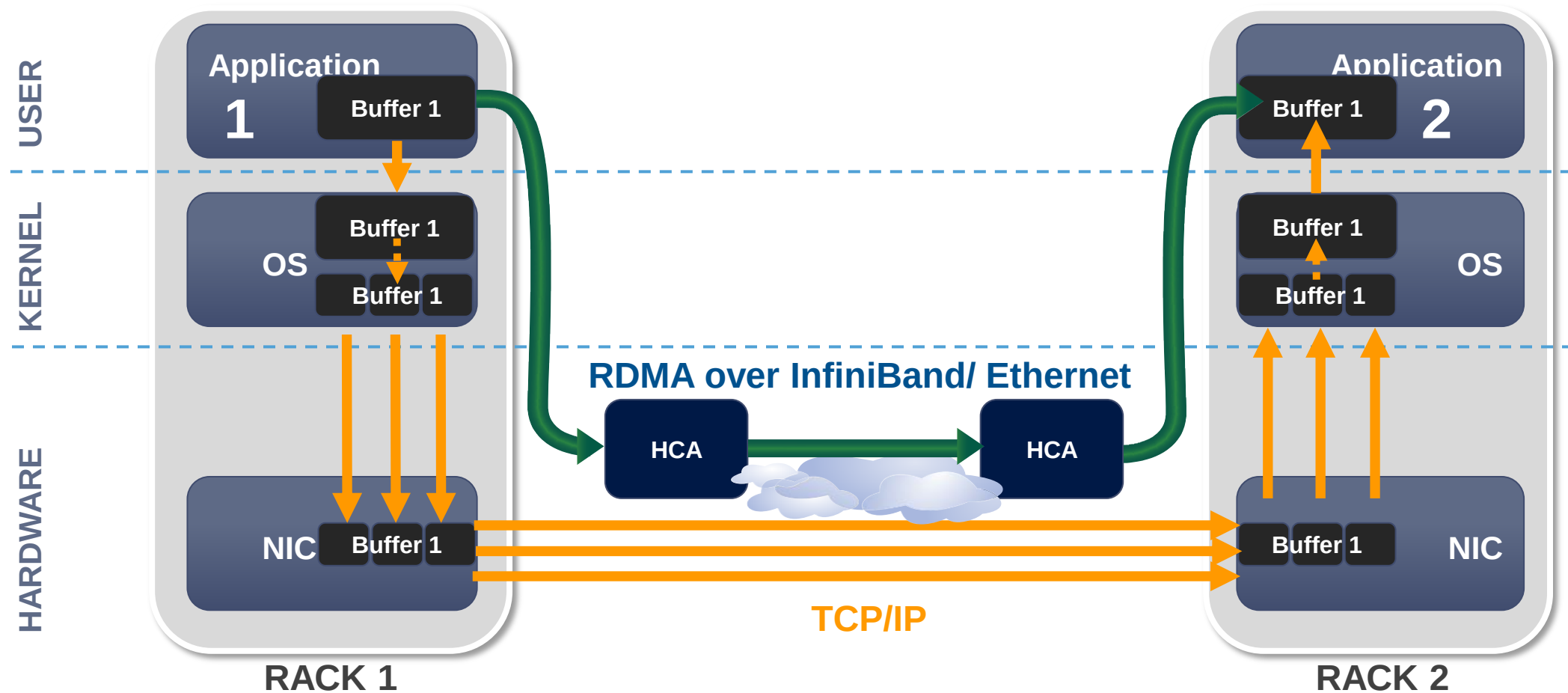
Transport – Host Channel Adapter (HCA) Model

- Asynchronous interface - Verbs
 - Consumer posts work requests
 - HCA processes
 - Consumer polls completions
- Transport executed by HCA
- I/O channel exposed to the application
- Polling and interrupt models supported

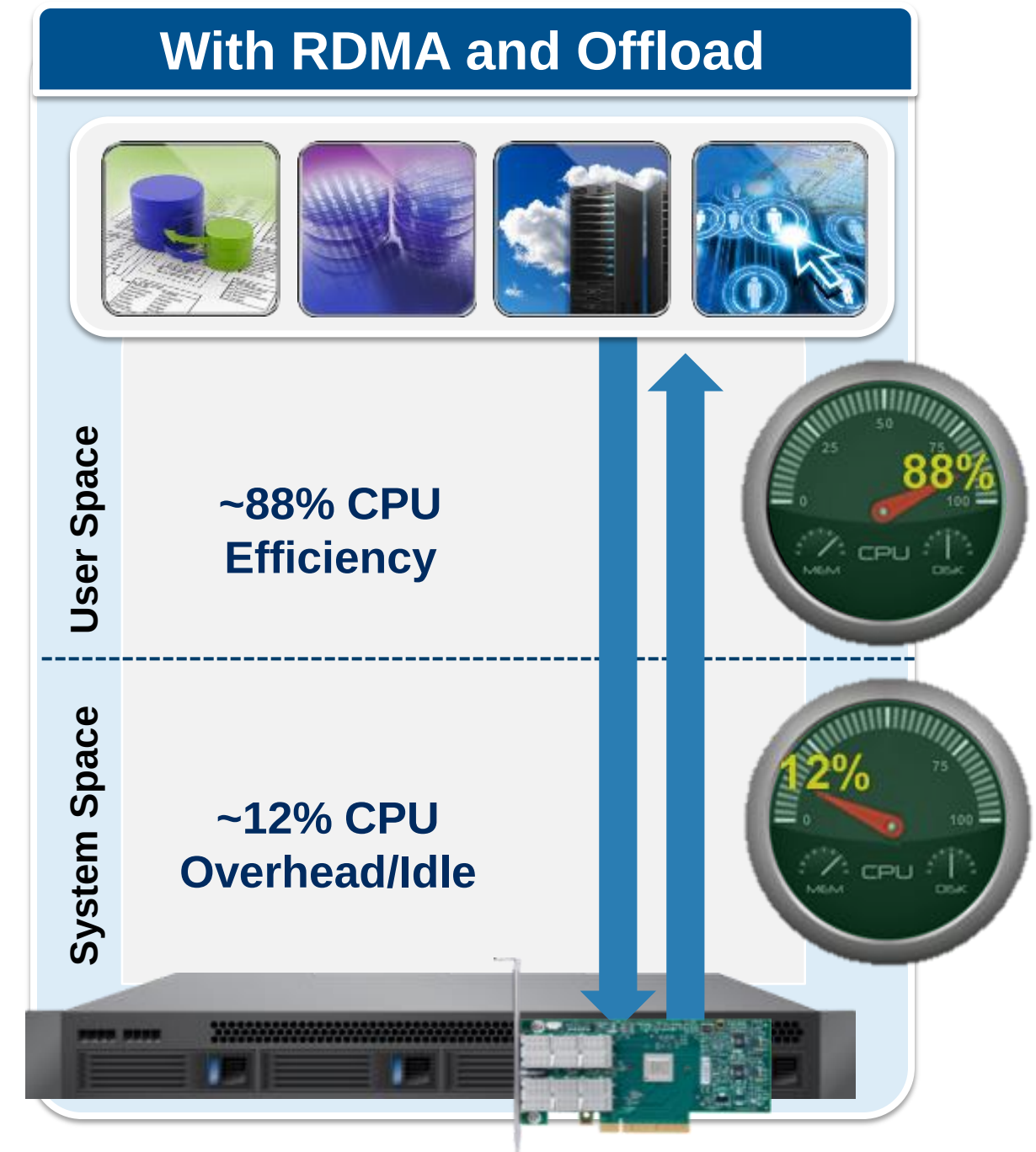
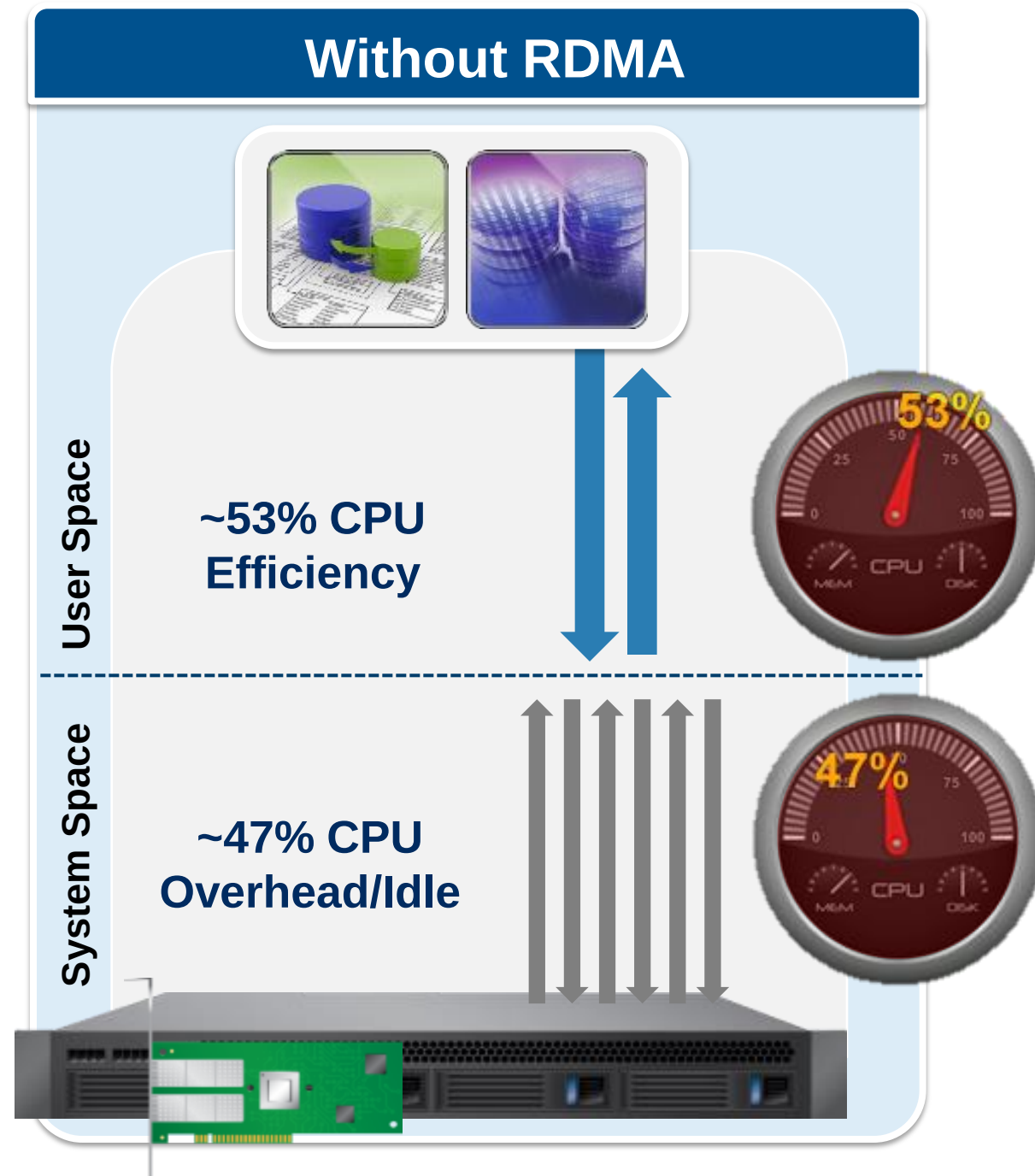


What is RDMA?

- The 3 Remote Direct Memory Access Protocol (RDMA) goodies
 - Transport Offload
 - Kernel Bypass
 - RDMA and Atomic Operations



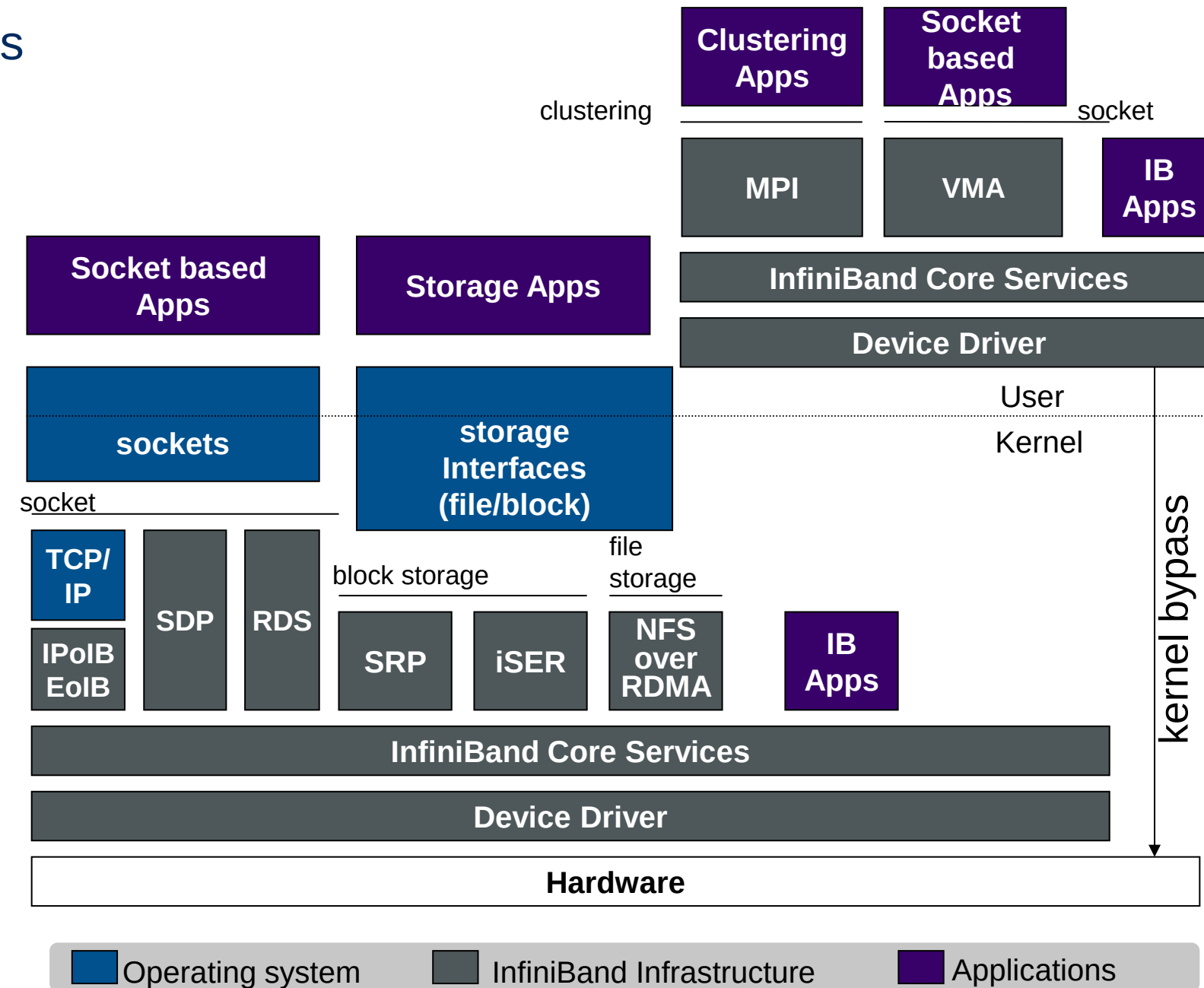
I/O Offload Frees Up CPU for Application Processing



Upper Layer Protocols



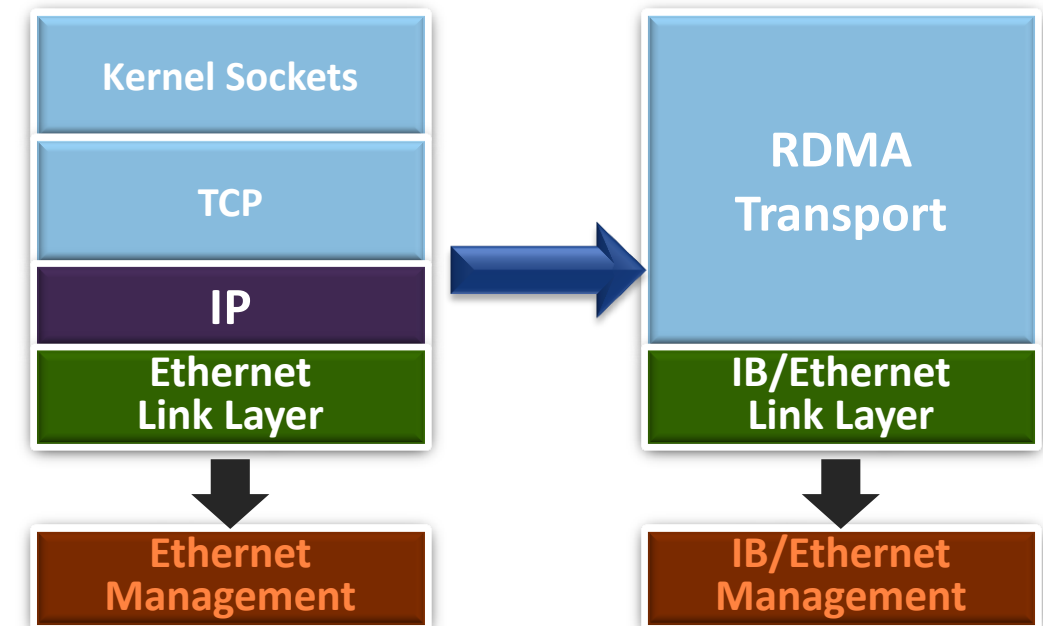
- ULPs connect InfiniBand to common interfaces
- Supported on mainstream operating systems
- Clustering
 - MPI (Message Passing Interface)
 - RDS (Reliable Datagram Socket)
- Network
 - IPoIB/EoIB (IP/Eth over InfiniBand)
 - SDP (Socket Direct Protocol)
 - VMA – usermode socket accelerator
- Storage
 - SRP (SCSI RDMA Protocol)
 - iSER (iSCSI Extensions for RDMA)
 - NFS over RDMA



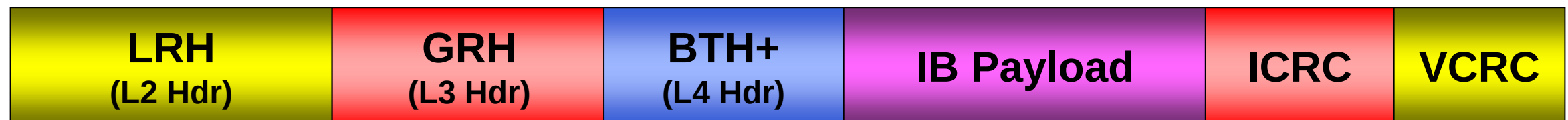
RoCE – RDMA over Converged Ethernet



- InfiniBand transport over Ethernet
- API Compatible
- Efficient, light-weight transport, layered directly over Ethernet
 - FCoE equivalent for high-performance IPC traffic
- Takes advantage of DCB Ethernet
 - PFC, ETS, and QCN



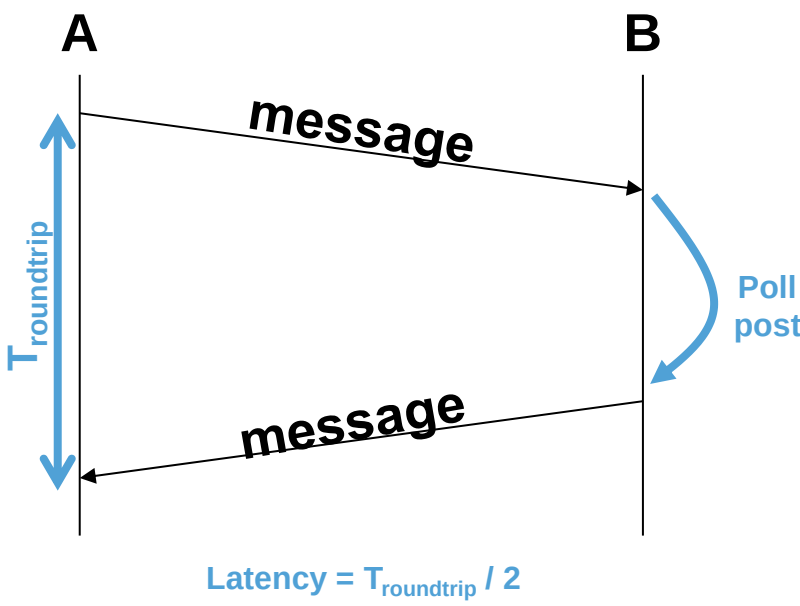
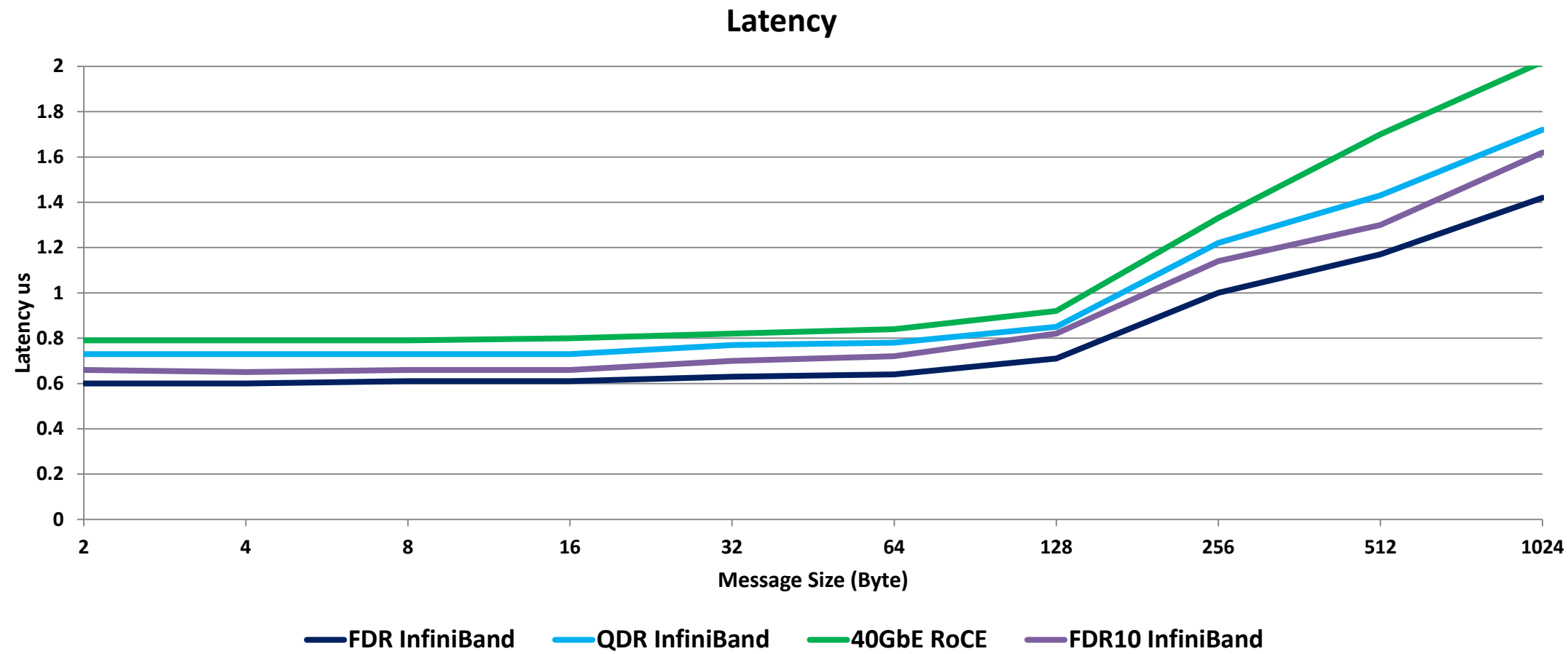
Infiniband



RoCE

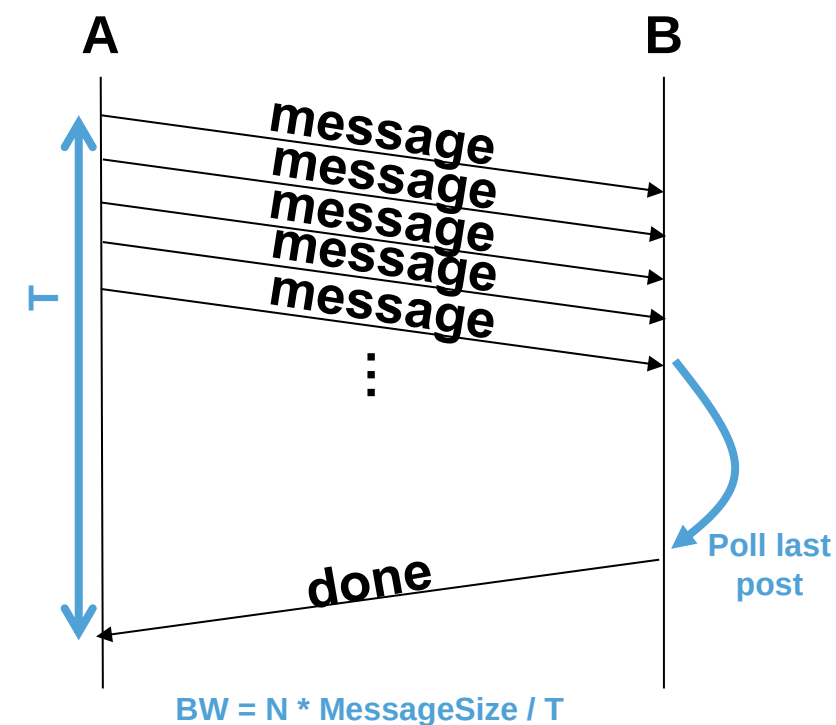
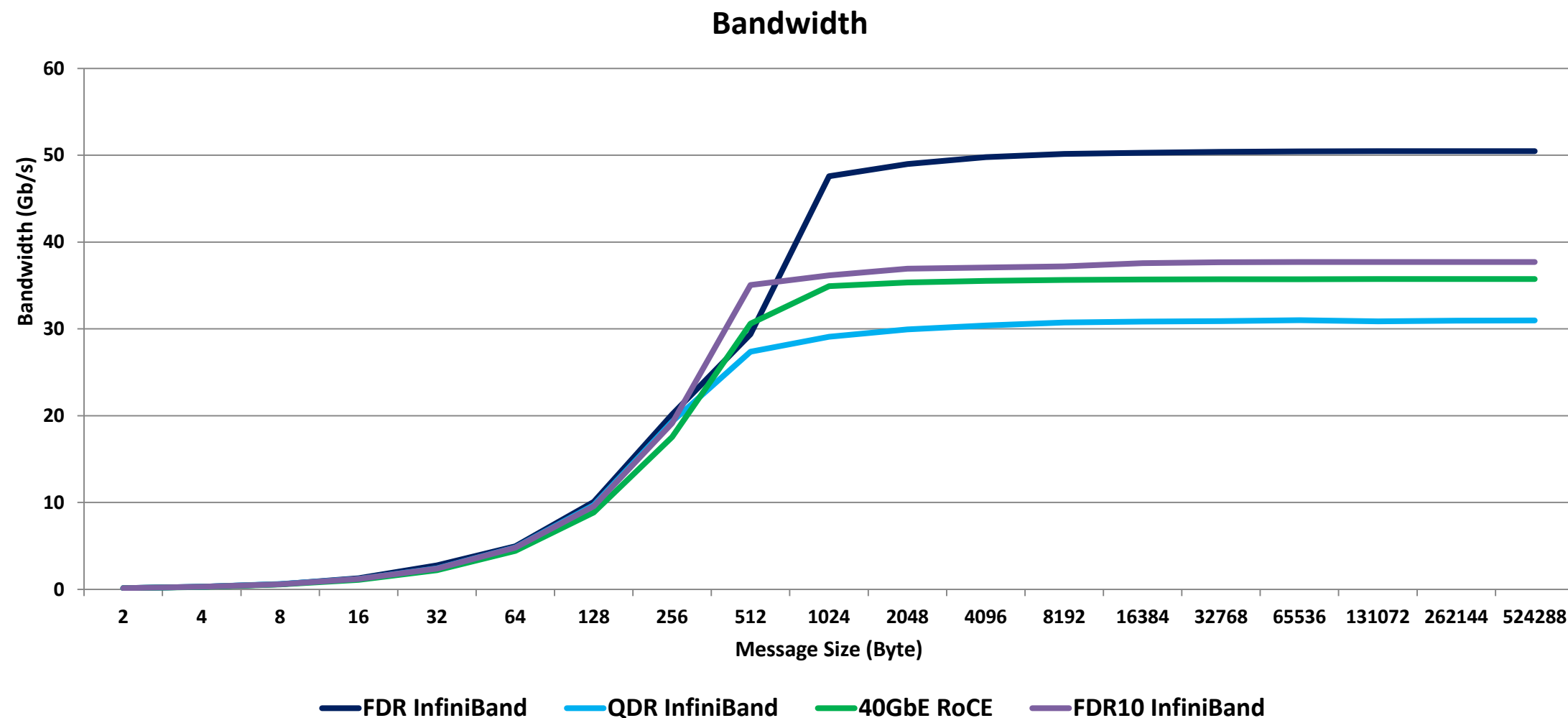


Latency – RoCE 40GbE vs. QDR vs. FDR (ConnectX®-3)



Lower is Better

Throughput – FDR vs. QDR vs. RoCE 40GbE (ConnectX-3)



Higher is Better

■ RDMA Goodies

- Transport offload
- Kernel bypass
- RDMA and Atomic operations

■ InfiniBand Goodies

- True SDN
- Cheaper
- Lower latency
- Higher density (roadmap)

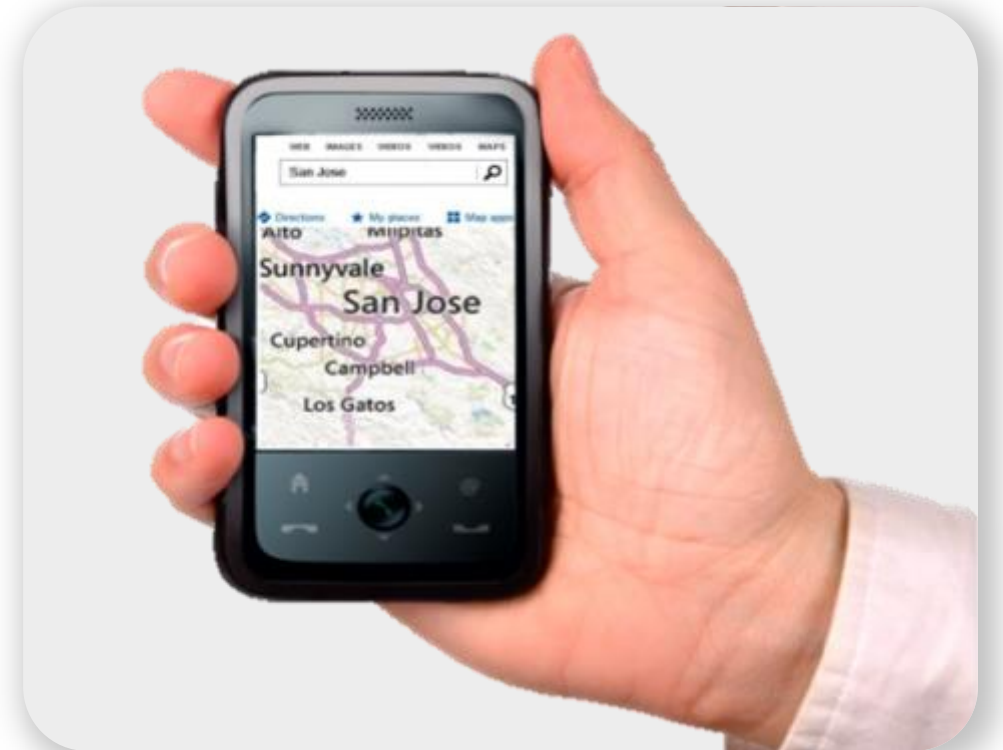
RDMA Inside ... Best Kept Secret in the Data Center



- Many of the World's Largest Web 2.0 Data Centers are Running on Mellanox Interconnects
 - InfiniBand and Ethernet (RoCE) RDMA Interconnects are connecting millions of servers

50% CAPEX Reduction for Bing Maps

- High-performance system to support map image processing
- 10X performance improvement compared to previous systems
- Half the cost compared to 10GbE
- Mellanox end-to-end InfiniBand 40Gb/s interconnect solutions



Cost-Effective Accelerated Web 2.0 Services

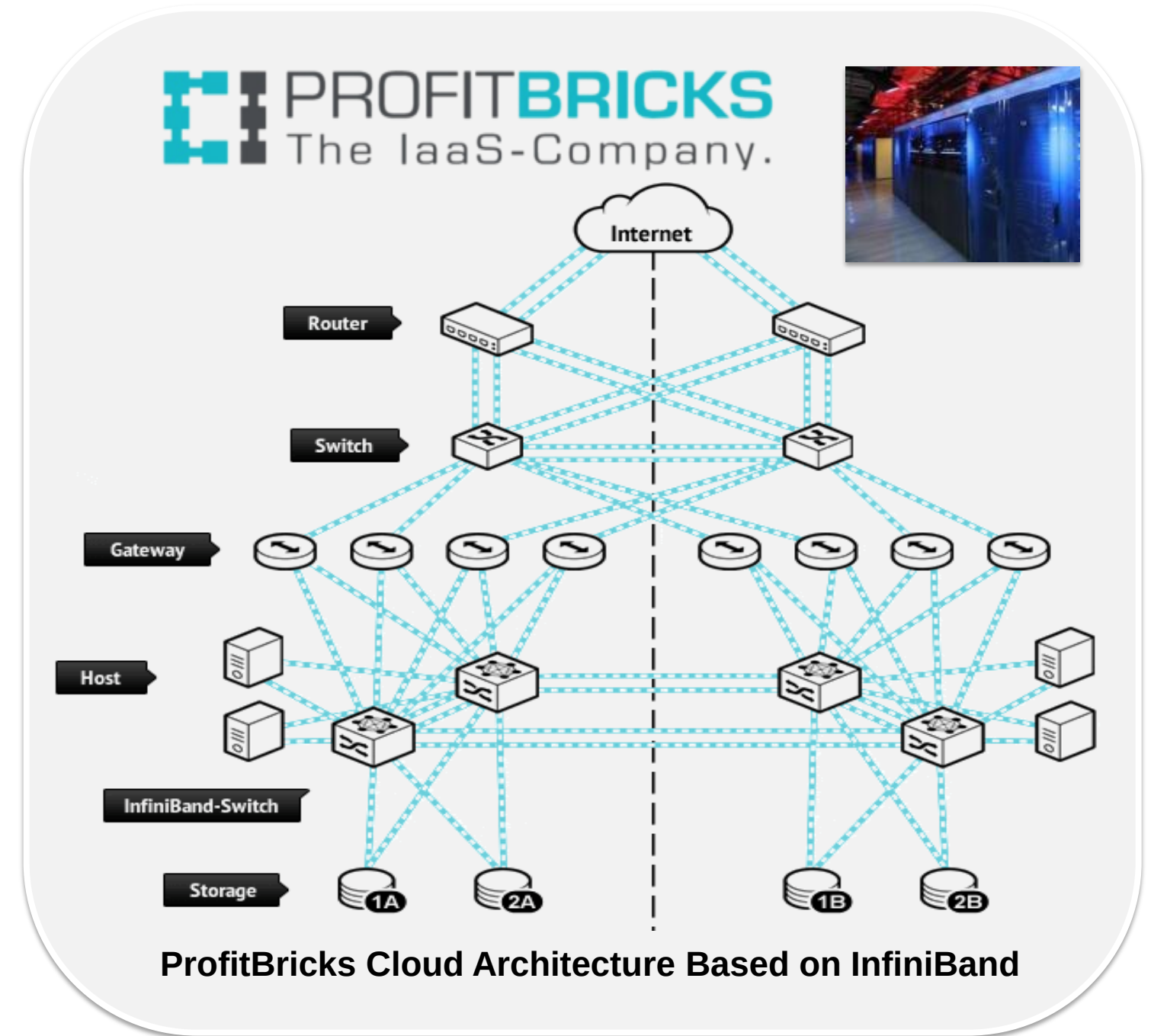
ProfitBricks Public Cloud Solution

- InfiniBand based IaaS
- Deploy more VMs per physical server
- Saving on CapEx and OpEx

Provider	Cost Per Hour (Config 1)
ProfitBricks	\$0.07
Amazon EC2	\$0.16
Rackspace	\$0.12

Provider	Cost Per Hour (Config 2)
ProfitBricks	\$0.26
Amazon EC2	\$0.45
Rackspace	\$0.48

Config 1: 1 Core, 2GB RAM, 50 GB HDD Instance
Config 2: 1 Core, 8GB RAM, 100 GB HDD Instance

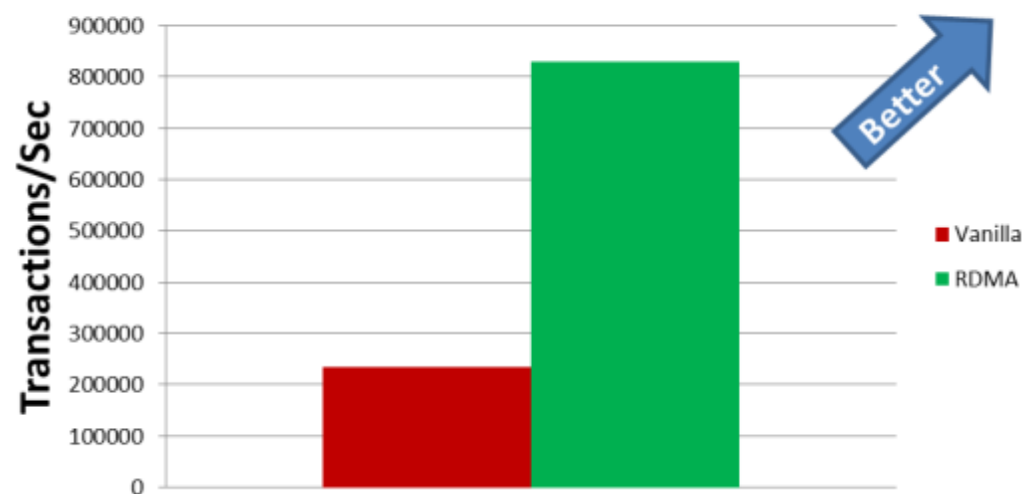


NoSQL Database Acceleration

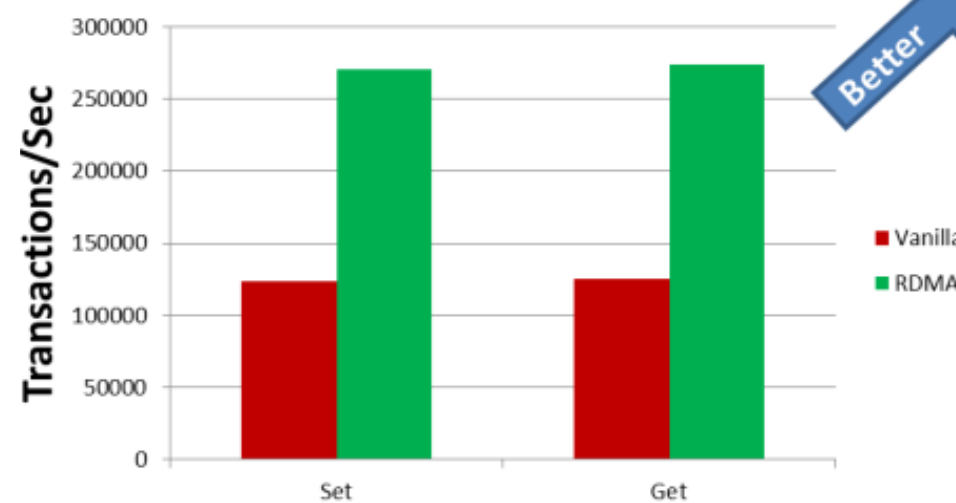
- No Change to Application Required!
- Up to 400% Increase in Number of Clients Without Adding Servers!
- Client and Server Side Installation



Memcached Transactions Per Second Testing



Redis Transactions Per Second Testing



Cassandra, Max Latency Testing



Up to 400% Acceleration Over RDMA

Mellanox InfiniBand Connected Petascale Systems



Connecting Half of the World's Petascale Systems
Mellanox Connected Petascale System Examples



Thank You