Low Latency Infrastructure for 25GbE and Above

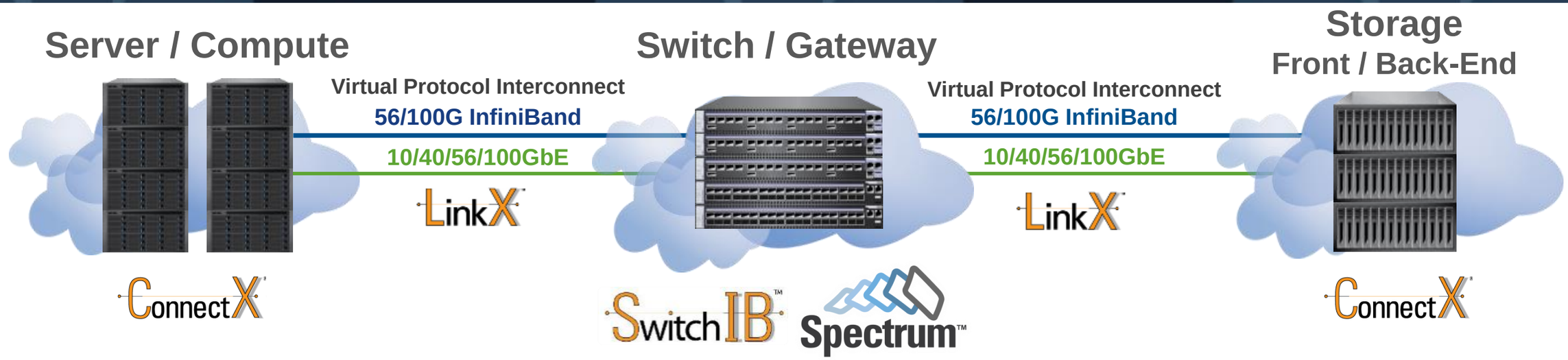
Asaf Wachtel, Sr. Director Enterprise Business Development

STAC Conference, Oct-Nov 2016

Mellanox® TECHNOLOGIES

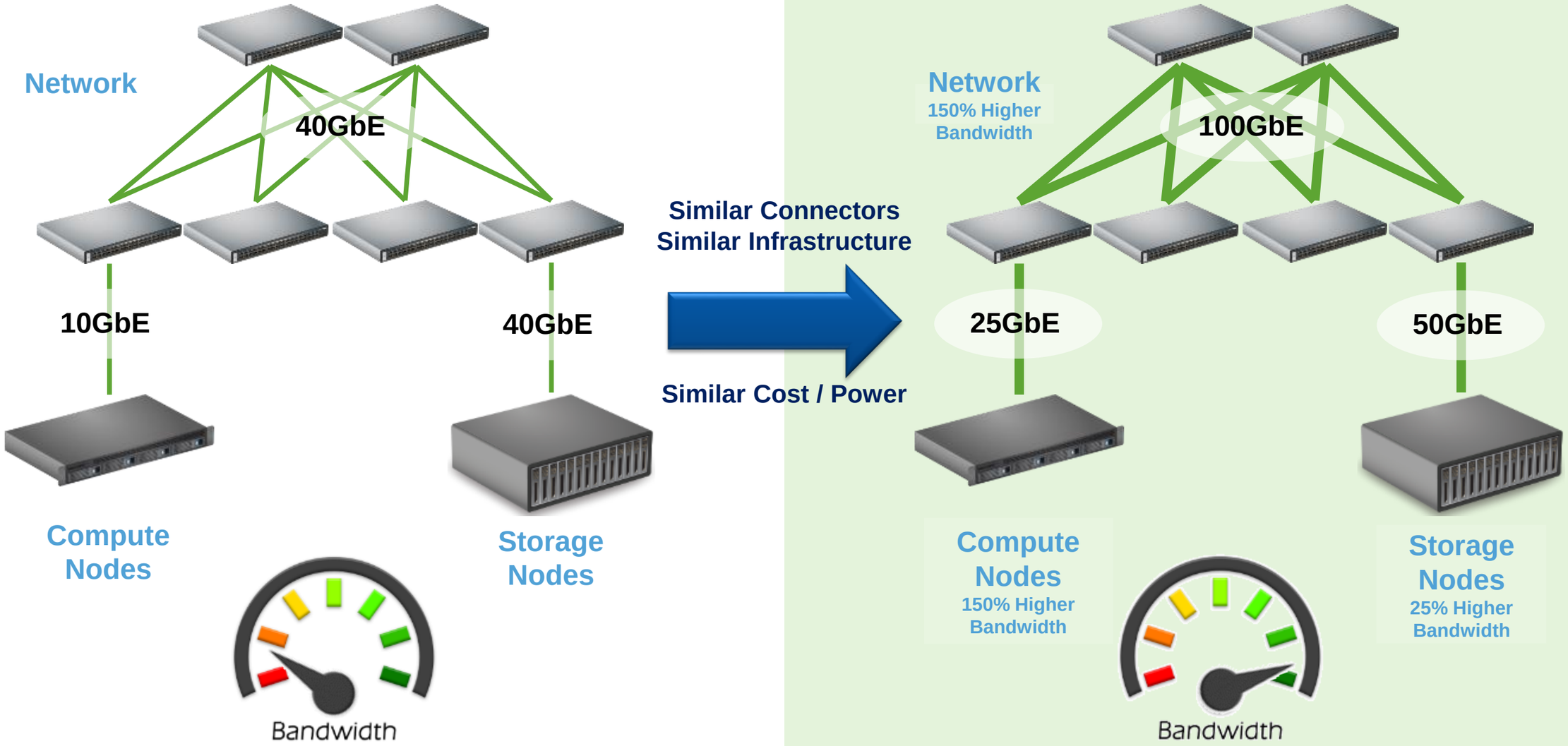
Connect. Accelerate. Outperform.®

Leading Supplier of End-to-End Interconnect Solutions



Comprehensive End-to-End InfiniBand and Ethernet Portfolio (VPI)

25/50/100GbE is Here



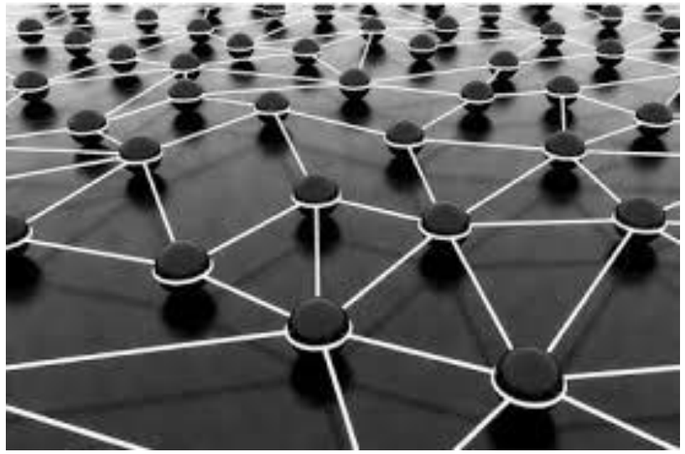
25GbE is Being Deployed on Wall Street Today!

Private Cloud

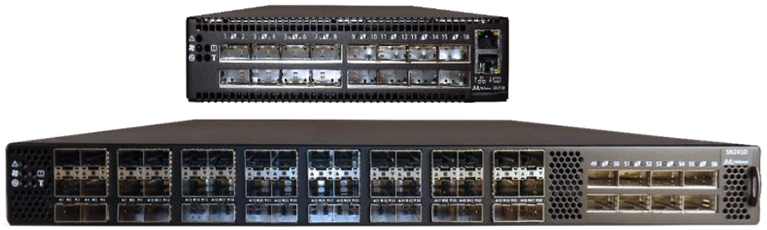
Storage

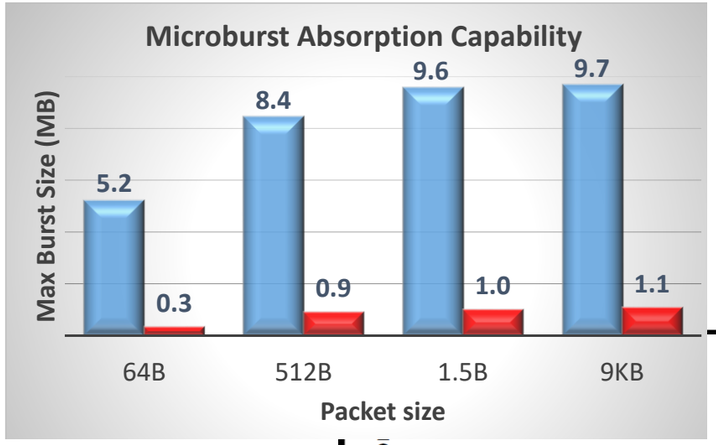
nvm EXPRESS

Research Grid



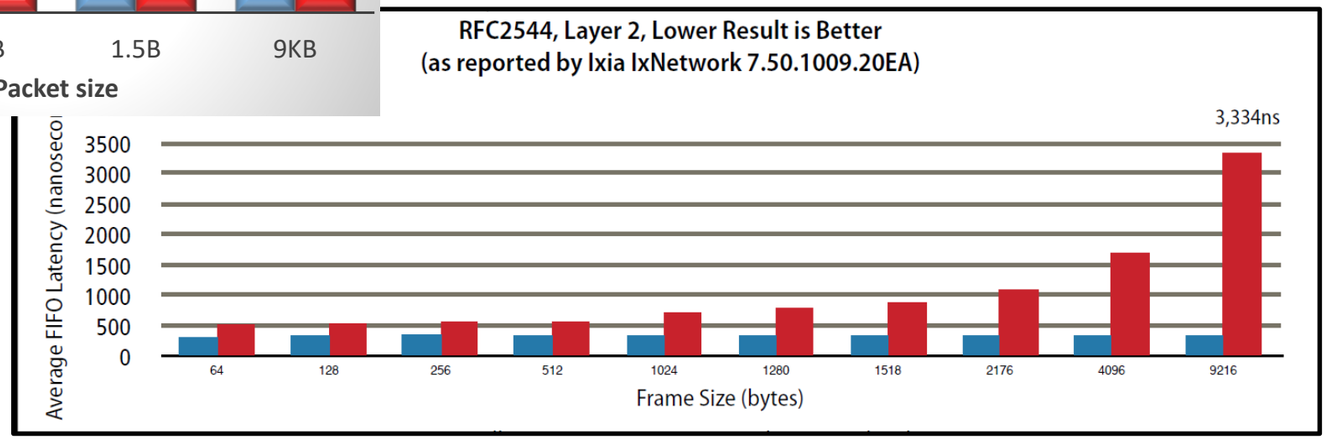
Spectrum: Industry's First & Only Low Latency 25GbE Switch





Mellanox Spectrum (not STAC benchmark)

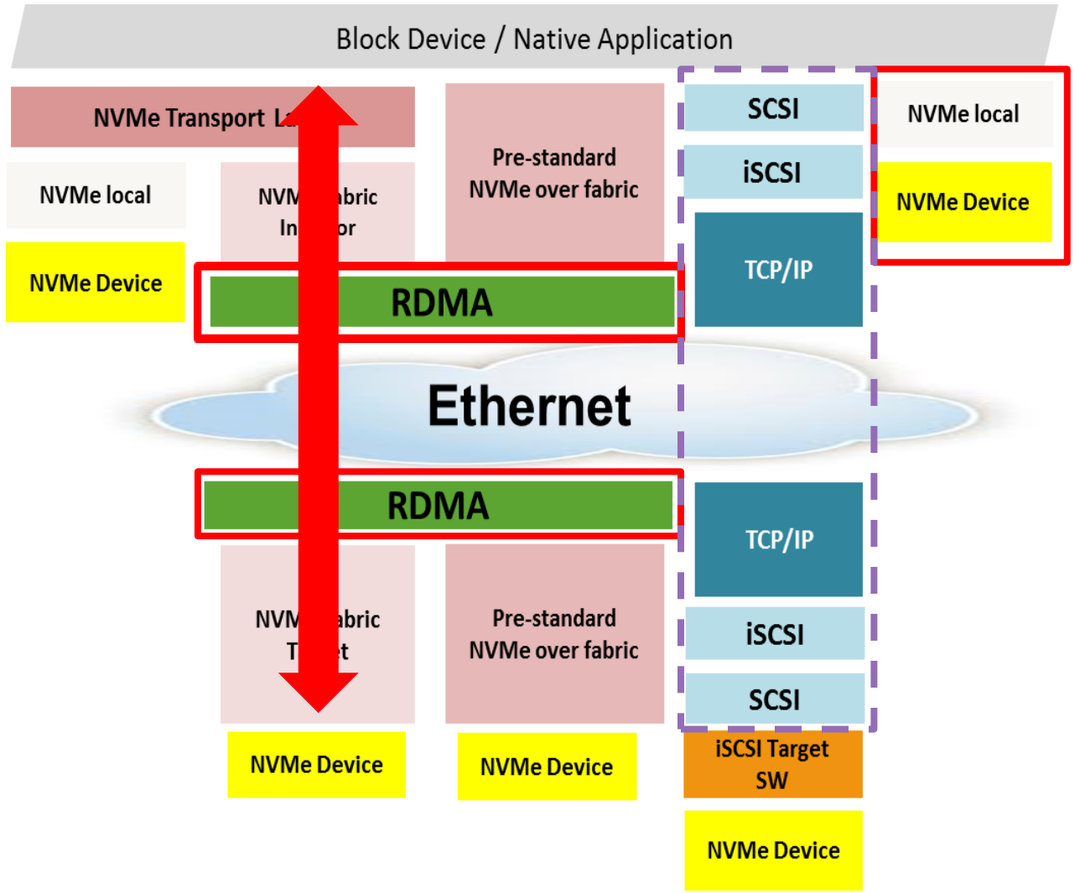
Not STAC Benchmarks



- **Full Cut-through Switching**
- **Consistent 300nsec latency on all ports, all message sizes, L2 & L3**
- **Superior microburst absorption (fully shared buffer)**
- **Full L3 stack including multicast routing**

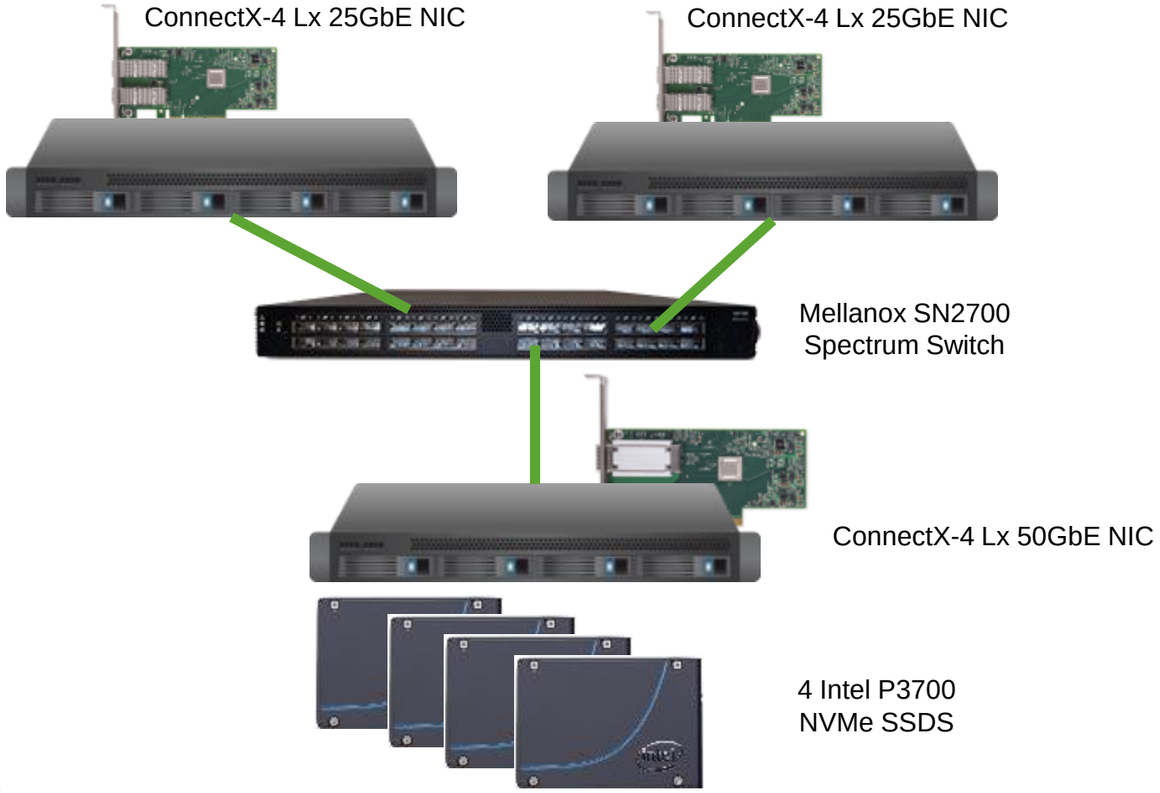
“NVMe Over Fabrics” Enables Storage Networking of NVMe

- Extend the efficiency of the local NVMe interface over a fabric
- NVMe over Fabrics industry standard developed (v1.0, June 2016)
- RDMA (RoCE) protocol is part of the standard
  - NVMe commands and data structures are transferred end to end
- Robust ecosystem



NVMe-oF Performance With Community Drivers (Standard 1.0 Open Source Driver)

- Topology –
  - Two compute nodes
    - ConnectX4-LX 25Gbps port
  - One storage node
    - ConnectX4-LX 50Gbps port
    - 4 X Intel NVMe devices (P3700/750 series)
  - Nodes connected through switch



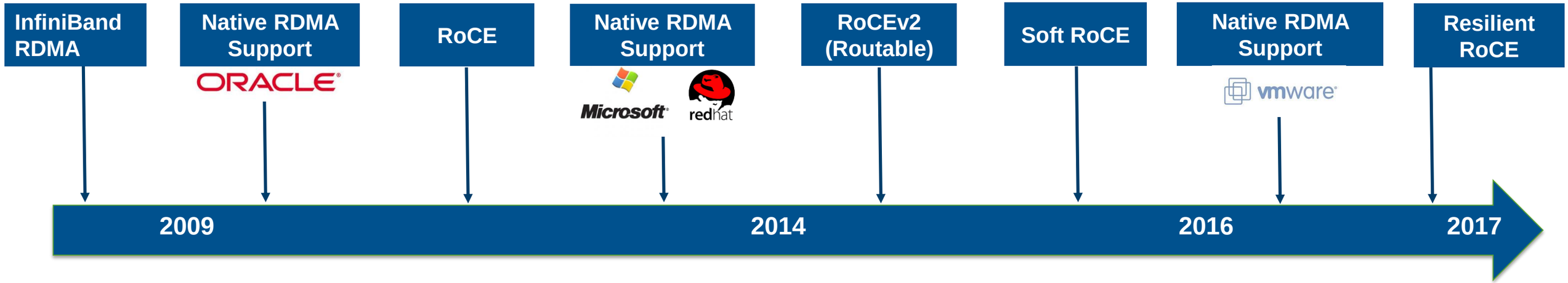
How to implement open source driver instructions:

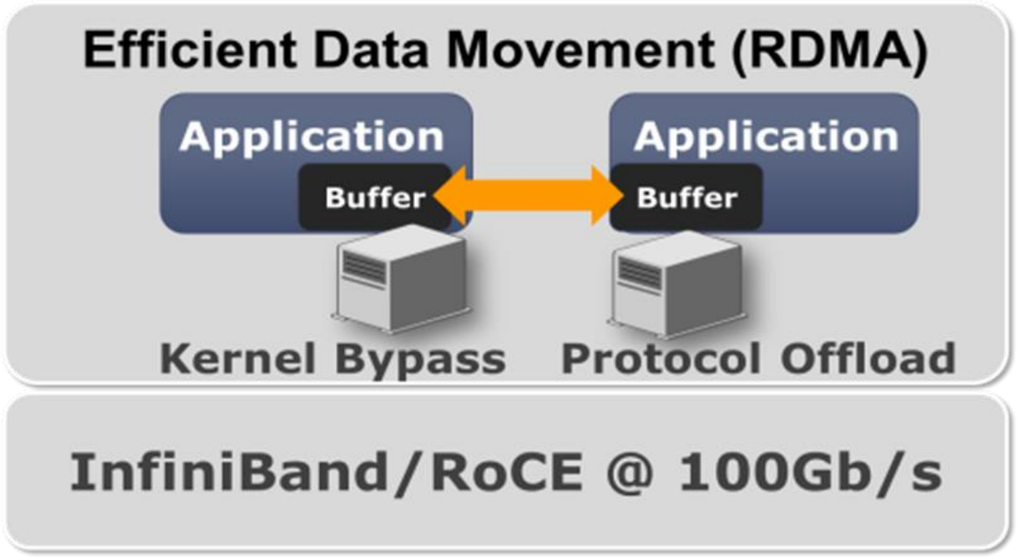
https://community.mellanox.com/docs/DOC-2504

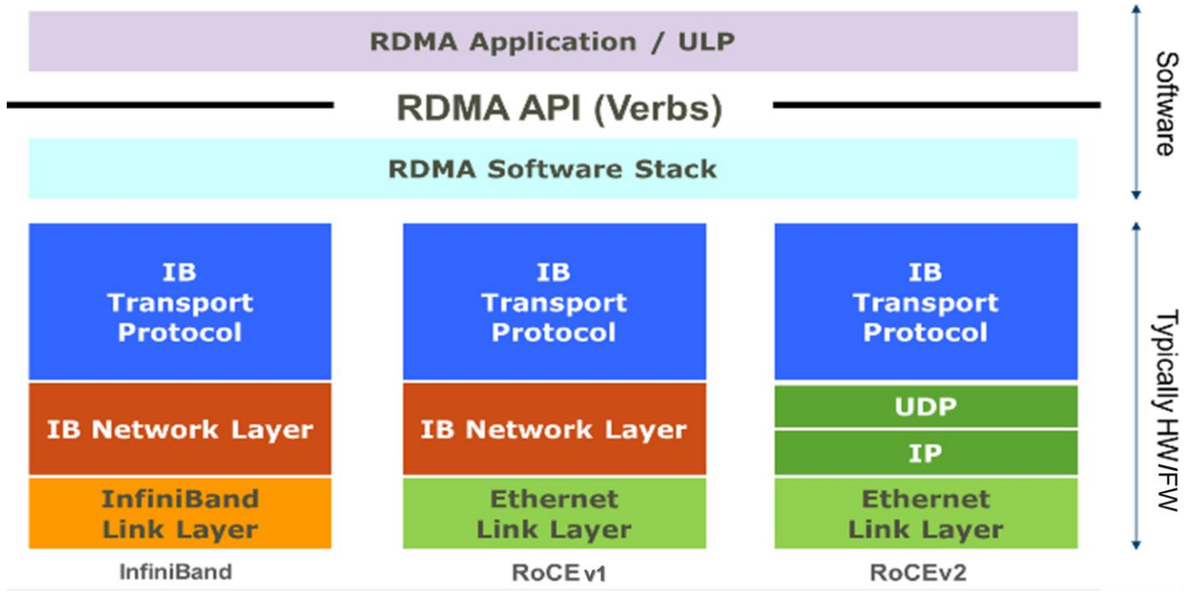
	Bandwidth (Target side)	IOPS (Target side)	Num. Online cores	Each core utilization
BS = 4KB, 16 jobs, IO depth = 64	5.2GB/sec	1.3M	4	50%

Not STAC Benchmarks

Evolution of RoCE (RDMA over Converged Ethernet)



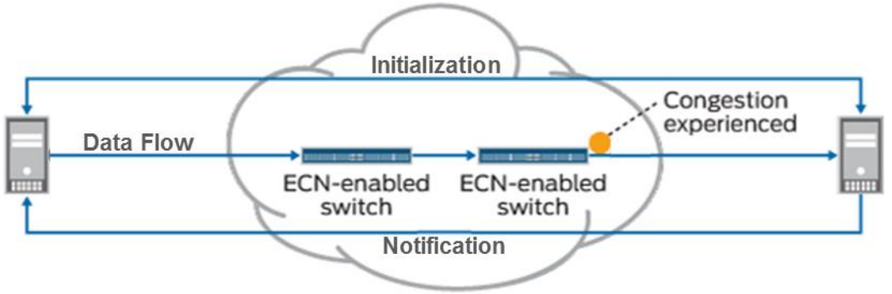




“Resilient RoCE” – No More PFC (Priority Flow Control)

- This is NOT a new version of RoCE – no change to Spec or Wire Protocol
- ConnectX-4 Upgrade enables running RoCE on a Lossy and/or Congested network
  - New RoCE congestion and enhanced lost data recovery control on NICs
  - Requires ECN enabled on switches
- ECN (Explicit Congestion Notification)
  - Standard congestion notification mechanism that is implemented in most switches
  - Non-intrusive to the host; host can be configured to respond or not
  - Used by TCP/IP in todays networks
- RoCE Congestion Control
  - Mellanox and Microsoft implemented RoCE Congestion Control
    - http://research.microsoft.com/en-us/um/people/padhye/publications/rdma-sigcomm15.pdf
  - RoCE Congestion Control controls the transmit rate to avoid congestion
  - Reacts to Congestion Notification Packets (CNP) sent as a result of receiving ECN marked packets
  - ConnectX-4 implements RoCE congestion and enhanced lost data recovery control at hardware/firmware level

RFC 3168 - The Addition of Explicit Congestion Notification (ECN) to IP



Thank You

Low Latency Switches

Low Latency NICs

25GbE

NVMe over Fabric

Kernel bypass

Resilient RoCE

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.®