RDMA over Converged Ethernet (RoCE)

Workshop on Data Center
Converged and Virtual Ethernet Switching
(DC CAVES)

Ali Ayoub
Mellanox Technologies

Agenda

- Background
- RoCE Overview
  - Protocol stack
  - Packet format
- Performance
  - Low Level Benchmarks
  - Application Level Benchmarks
- RoCE Benefits
- Q&A

Background

- RDMA: Remote Direct Memory Access
  - RDMA provides high-throughput, low latency
    - Peer-to-peer, memory-to-memory access, zero-copy
    - Reduces consumption of CPU cycles
    - Reduces communication latency
- InfiniBand: is a switched fabric communication link
  - Originally designed for HPC:
    - High throughput, Low latency, Quality of service
    - Failover, Scalability, Reliable transport
  - How do we interface this high performance link with existing Ethernet infrastructure? RoCE!
- RoCE: RDMA over Converged Ethernet
  - Provide InfiniBand-like performance and efficiency to ubiquitous Ethernet infrastructure.
  - Utilize the same transport and network layers from IB
  - InfiniBand stack and swap the link layer for Ethernet.
  - Implement IB verbs over Ethernet.

RoCE Overview

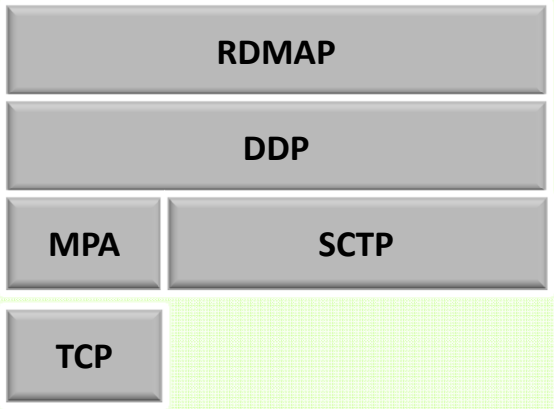
- Most efficient low latency Ethernet solution
  - 1.3usec end-to-end RDMA latency
  - Very low CPU overhead
  - Takes advantage of PFC (Priority Flow Control) in DCB Ethernet
- Standards based
  - Implements the RoCE (RDMA over Converged Ethernet) specification
  - IBTA standard, provide InfiniBand API over standard Ethernet
- All proven OFA verbs supported semantics
  - Kernel bypass, SEND/RCV, atomic operations
  - UDP, multicast
  - Existing low latency (RDMA) apps run seamlessly over RoCE
- Proven, most deployed RDMA transport
  - Server efficiency and scaling to 1000s of nodes
  - Scales to 10GE/40GigE support and beyond

RoCE

- **Simple, purpose built for data center traffic**
- **Utilizes lossless data-link (Priority Flow Control)**

iWARP

Internet Wide Area
RDMA Protocol



- **Multiple, complicated layers to patch up best-effort LAN/WAN networks**
- **Requires TCP offloading**
- **SCTP is unproven**

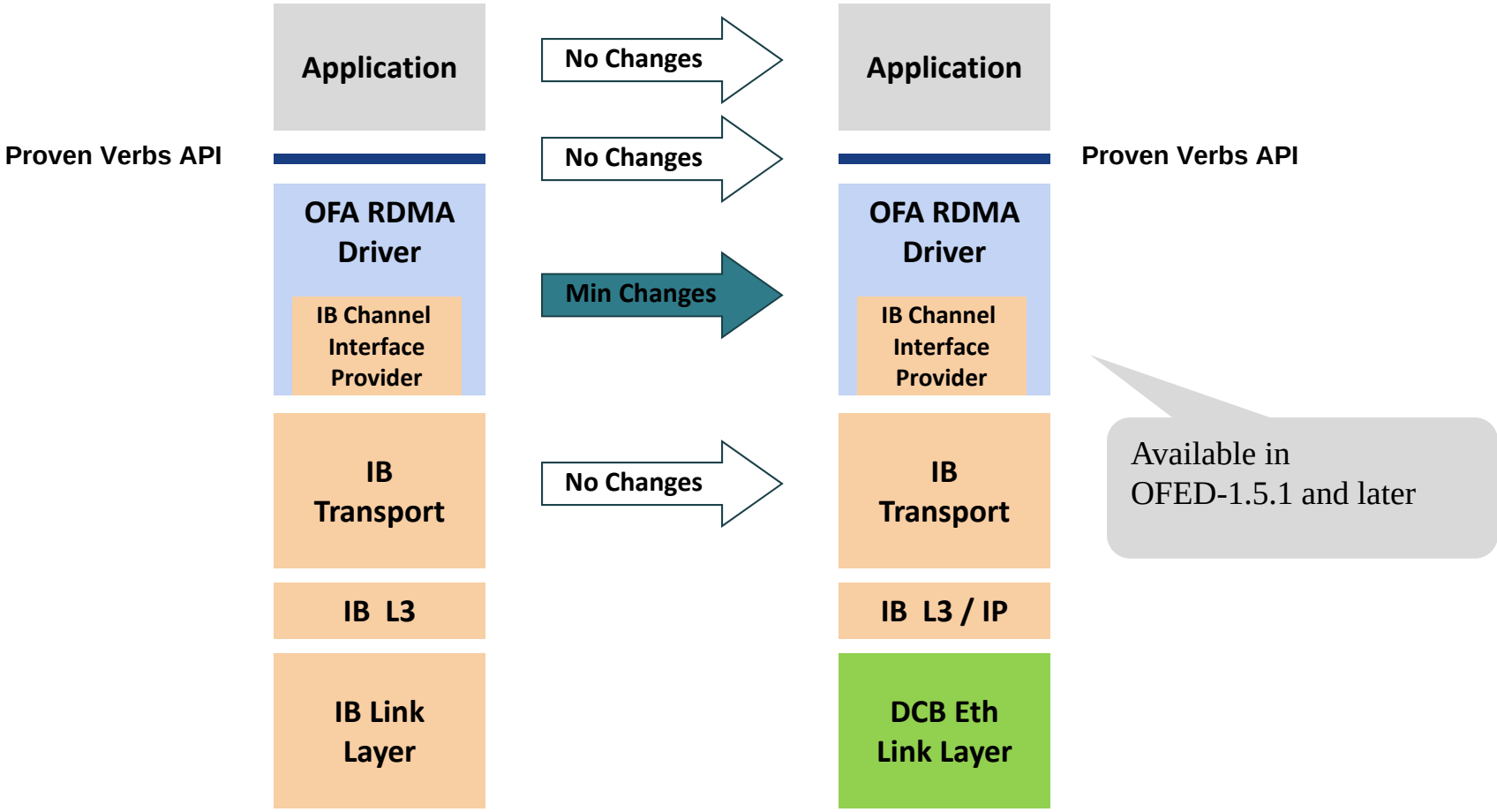
Improve Data Center Efficiency

- RDMA transport paradigm depends on a set of characteristics
  - No dropped packets
  - Arbitrary topologies
  - Traffic class types

- So What Changed? **Ethernet!**

Ethernet	802.1	IEEE 802.1Qx
Lossless	No	Yes (802.1Qaz) PFC
Classes of service	No	Yes (802.1Qbb) ETS
Congestion management	No	Yes (802.1Qau) QCN

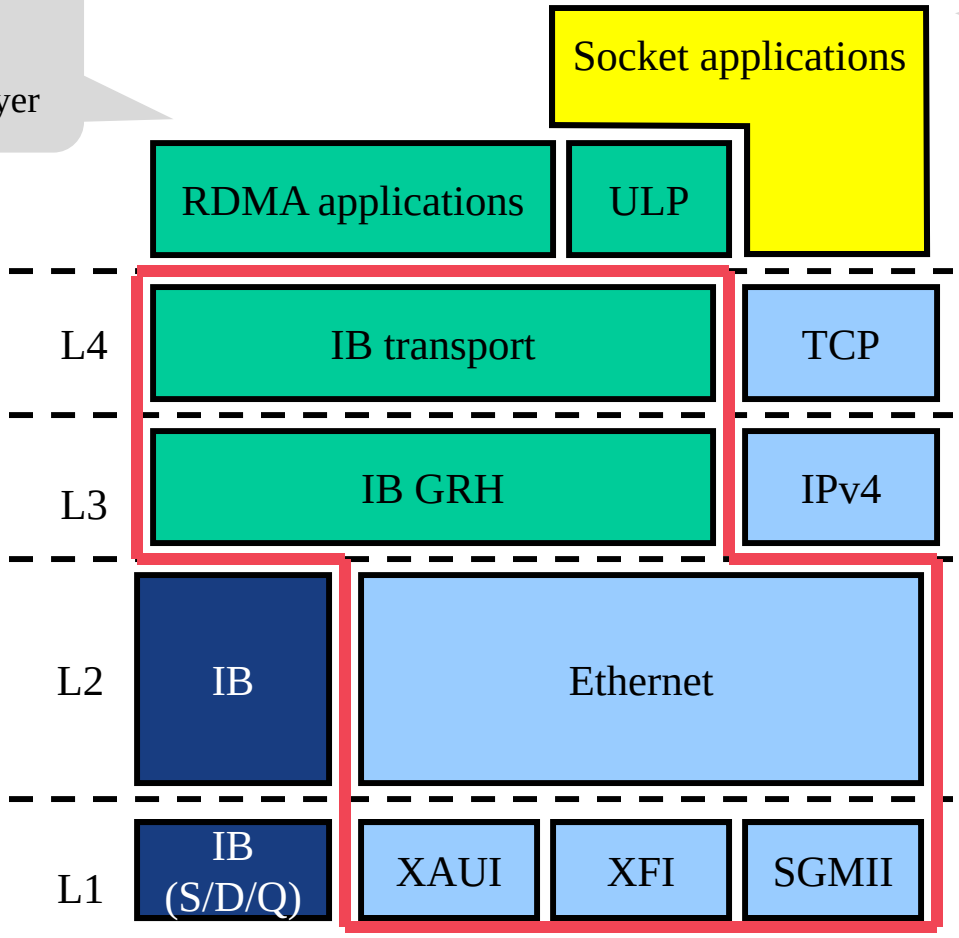
Protocol Stack



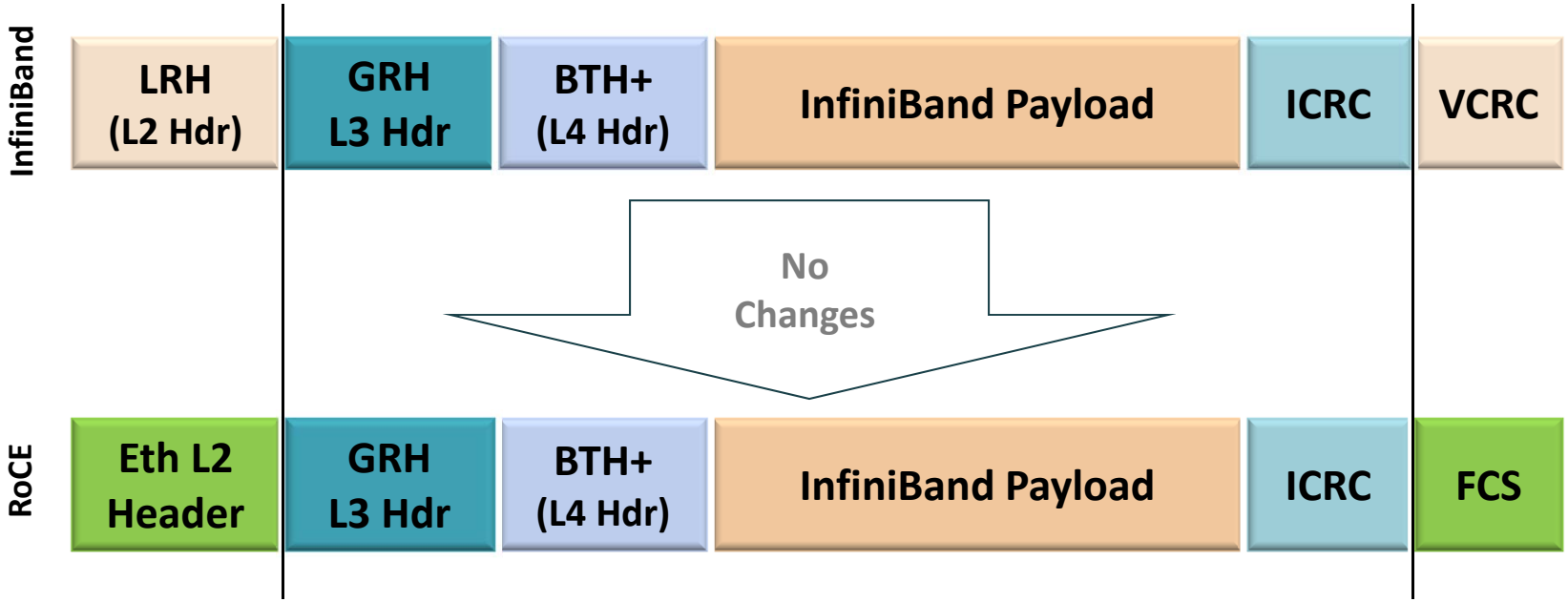
Protocol Stack

RoCE Based Applications written over IB Transport Layer

Standard Ethernet applications written over Sockets API

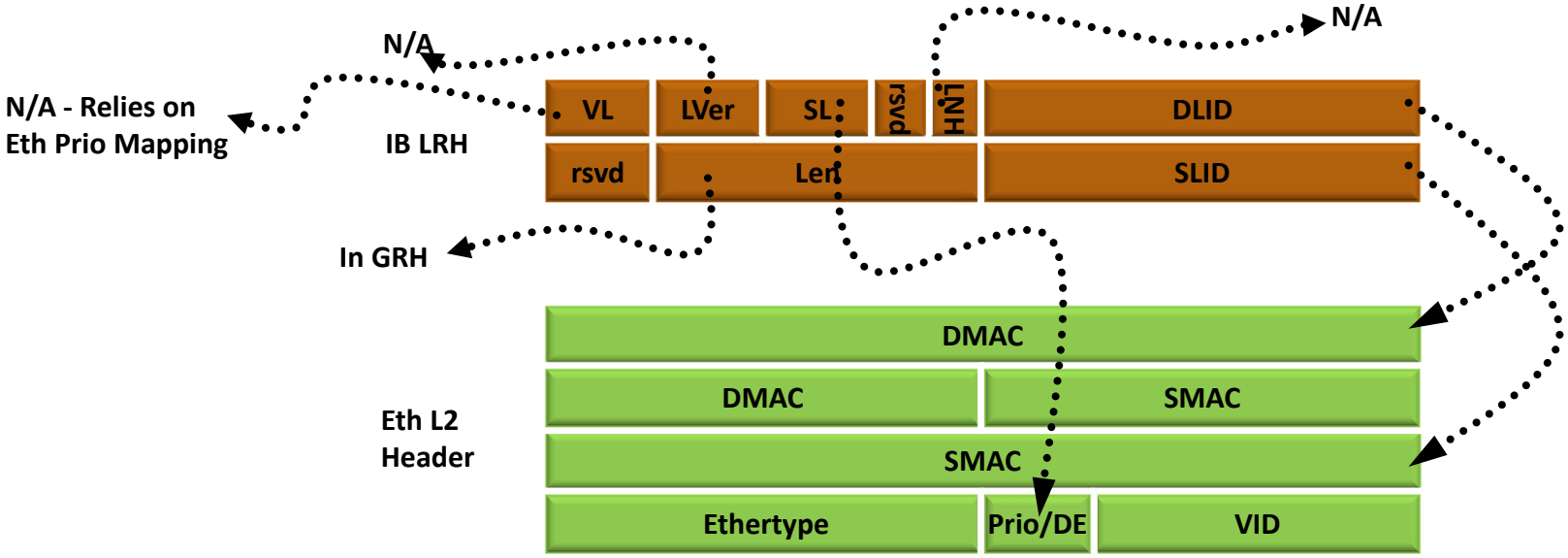


Packet Format



Data Link Implications

- DMAC,SMAC replace DLID,SLID
- 802.1Q header priority field replaces SL
  - Link Level Flow Control is Ethernet PFC
- 802.1Q header VLAN ID field allows for L2 partitioning
- IEEE Assigned Ethertype for RoCE (0x8915)



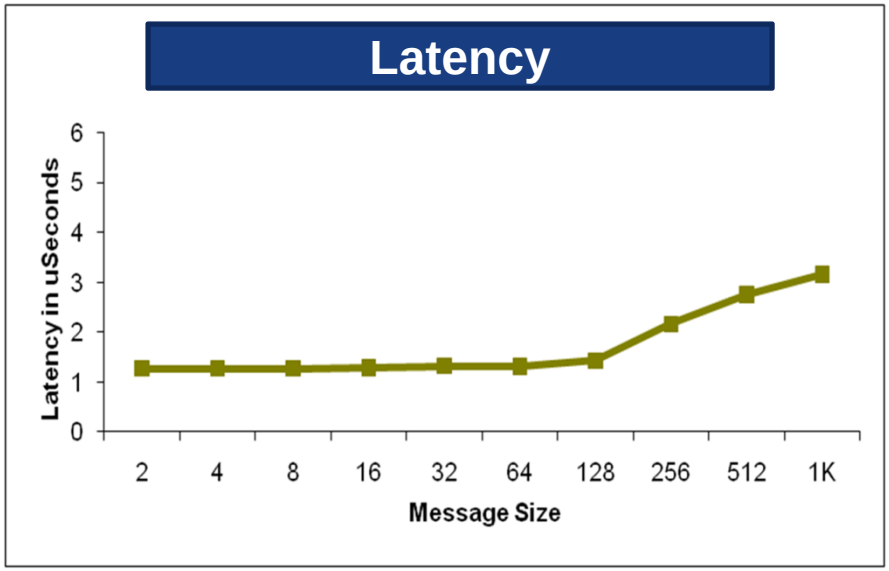
API

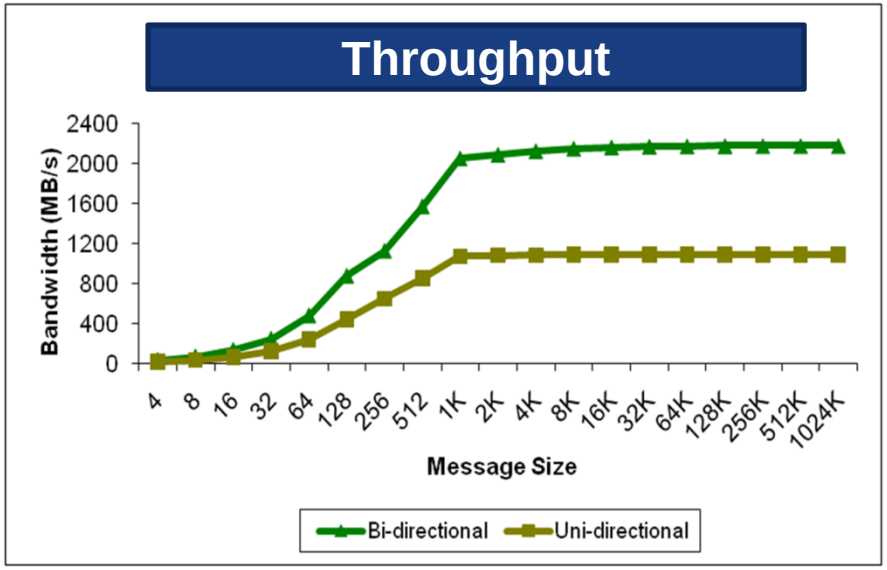
- Verbs
  - No syntax changes
  - Address handles must contain GIDs
- Connection management abstraction (CMA) – a.k.a RDMACM
  - OS IP addresses resolution is used to resolve remote MAC and outgoing Ethernet interface (or VLAN device)
    - RoCE device is selected accordingly
    - SGID is set to local interface IP
    - DGID is set to remote IP
  - Fills network parameters (MTU, SL, timeout) according to IP stack
  - Connection proceeds with CM as in IB

Compatibility

- iWARP
  - Complete compatibility at the binary level
    - Applications can run unmodified (no recompilation needed)
      - IB / RoCE / iWARP device selected based on IP address resolution
    - However, iWARP and RoCE are not interoperable
      - Different wire protocol
- Ethernet (management)
  - Ethernet host stack management is unchanged
    - E.g., ifconfig, ethtool, vconfig

Low Level Benchmarks



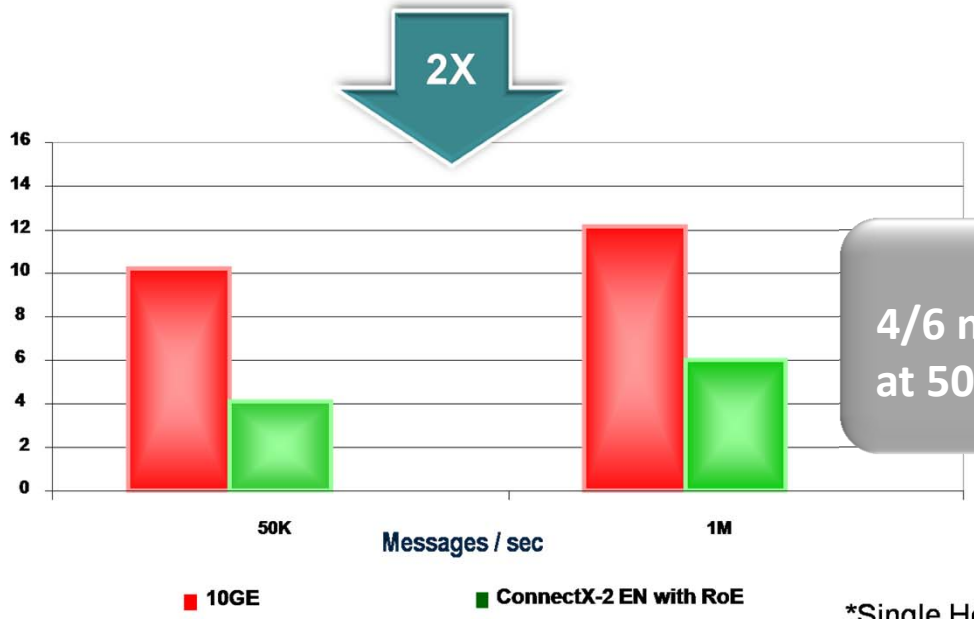


1.3µs Latency

Full Bandwidth with Small Message Sizes

Application Level Benchmarks (Messaging Services)

IBM WebSphere® MQ Low Latency Messaging
RoCE vs. TCP/IP Latency



IBM Results:
4/6 microseconds latency
at 50K/1M messages / sec

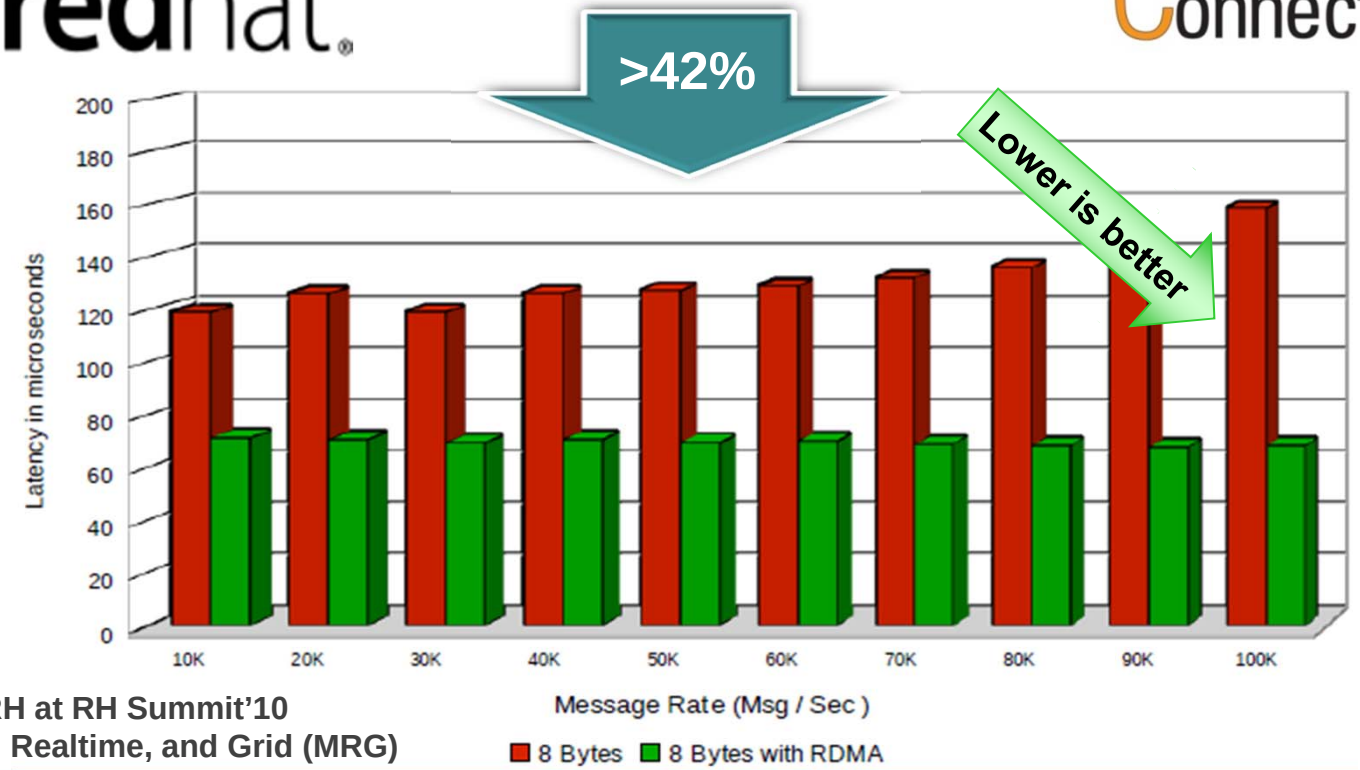
*Single Hop Latency (usec)

200% better latency helping faster execution time on large data volumes*

***Compared to other standards-based low latency NICs**

Application Level Benchmarks (Realtime)

MRG 1.3 Red Hat Enterprise 6.0 over RoCE*
RoCE vs. TCP/IP Latency



*Presented by RH at RH Summit'10
For Messaging, Realtime, and Grid (MRG)

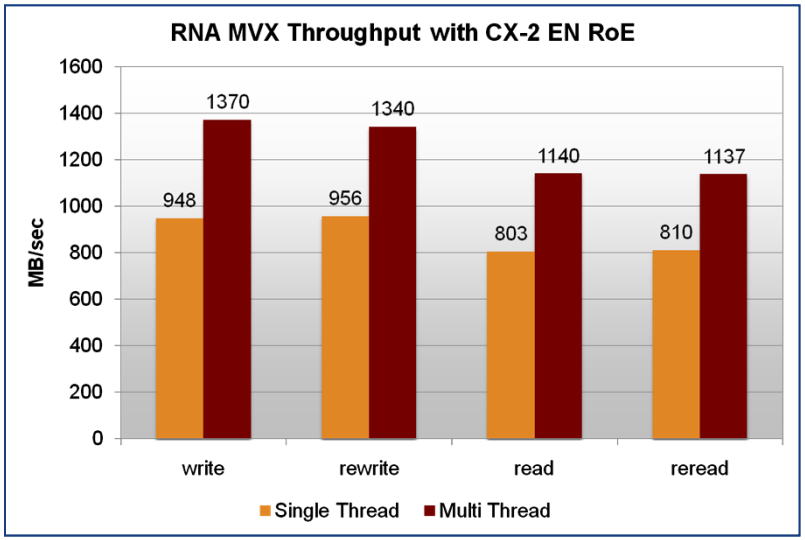
1.2 Million Acknowledged Messages per Second
Consistent latency across message rate

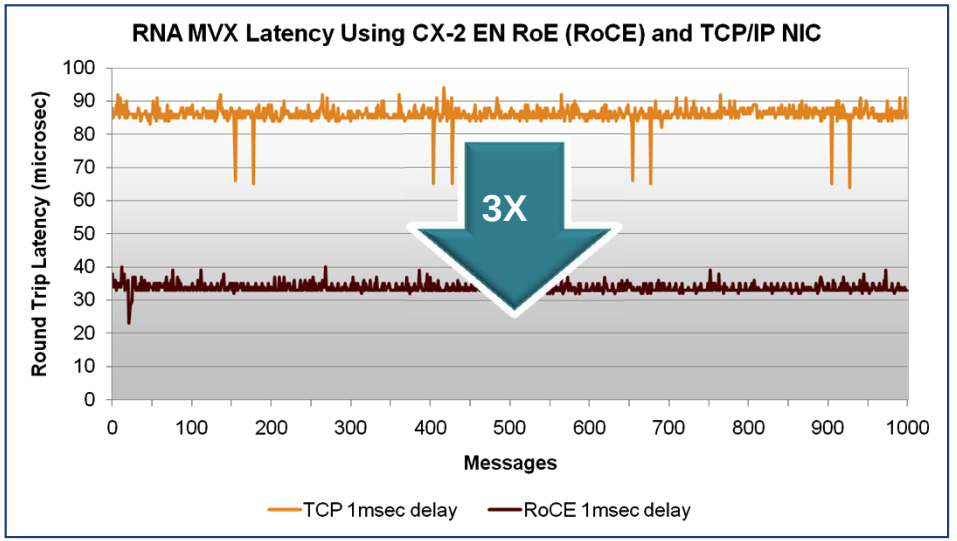
Application Level Benchmarks (Virtual Memory)

RNA networks MVX

RNA Memory Virtualization
RoCE vs. TCP/IP Latency

Mellanox TECHNOLOGIES
ConnectX-2 EN with RoCE



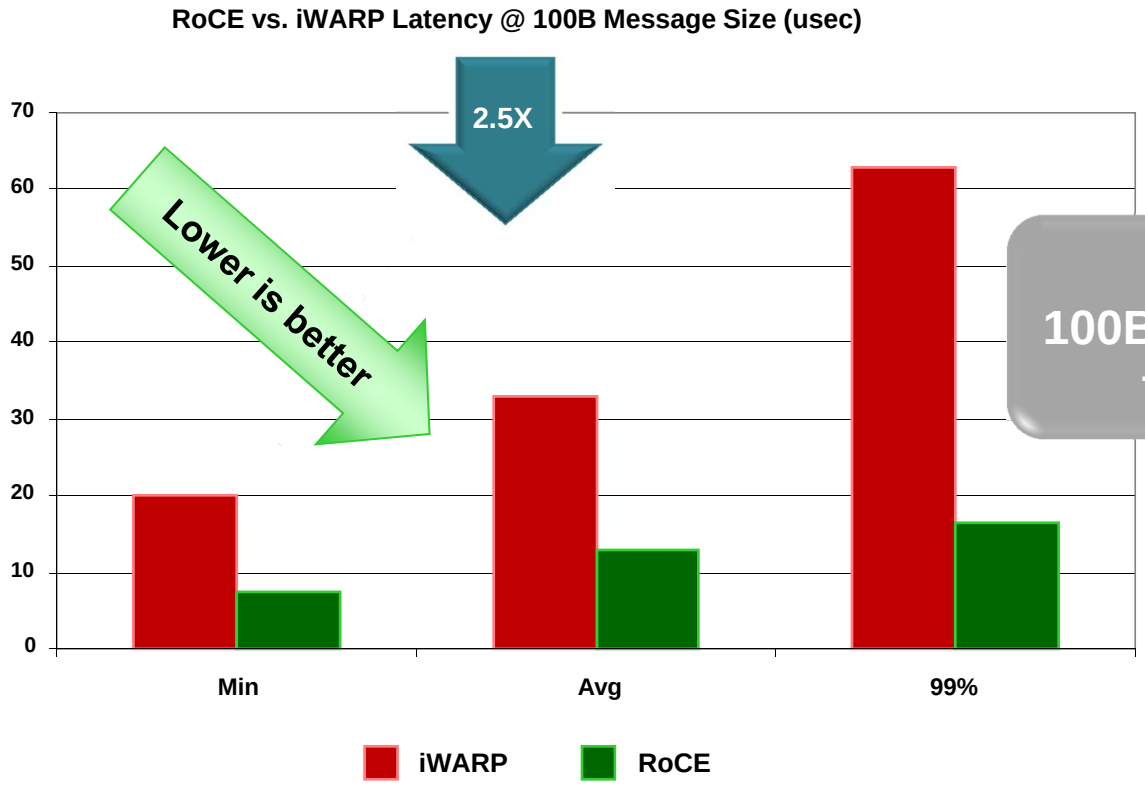


Max throughput and 1/3rd the latency
Consistent latency across message rate

Application Level Benchmarks (Financial Services)

New York Stock Exchange
RoCE vs. iWARP Latency

NYSE Euronext

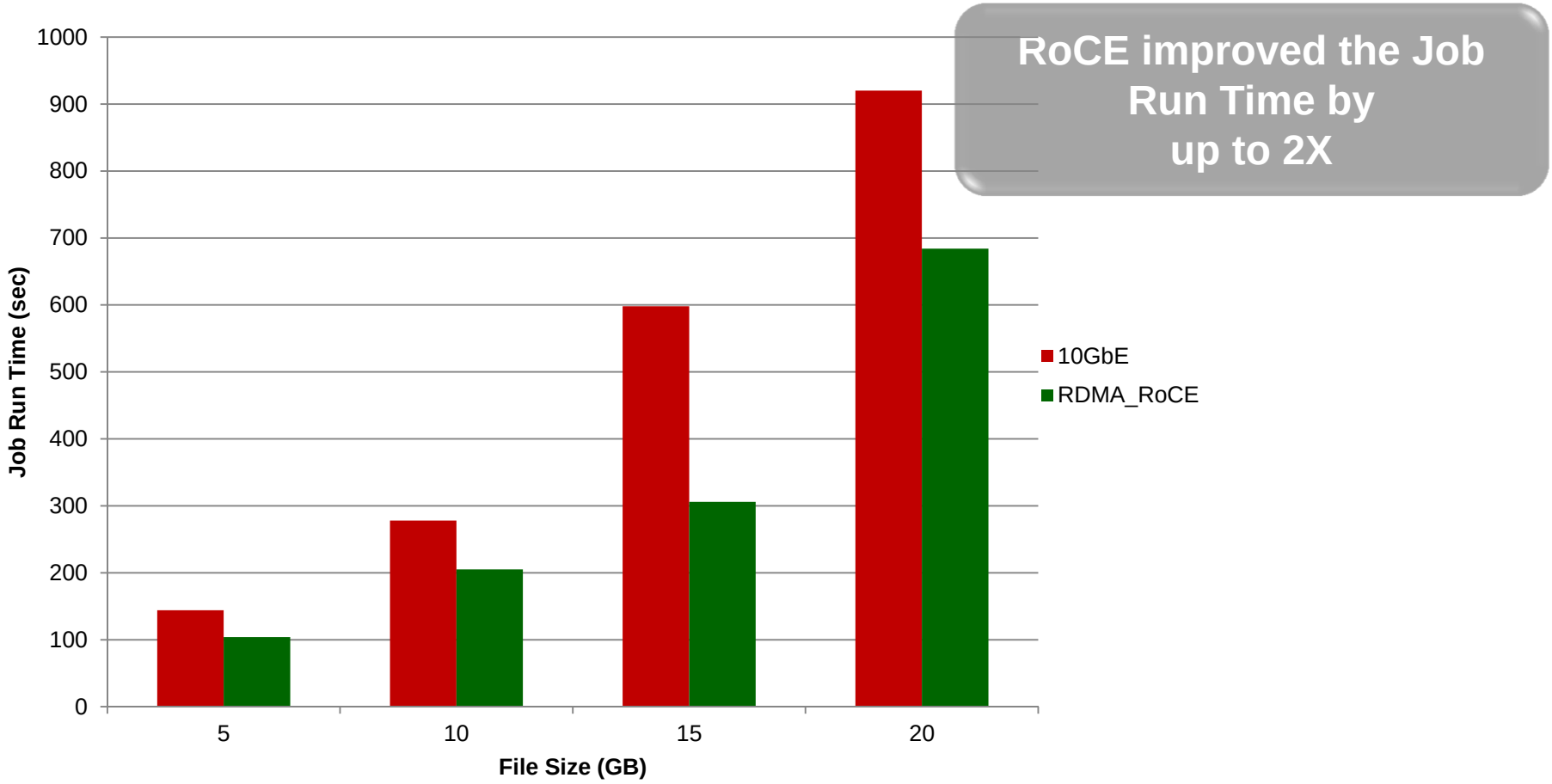


RoCE delivered 100Byte message 2.5 times faster than iWARP.

62% better on execution time vs. 10GigE iWARP adapter

Application Level Benchmarks (Hadoop)

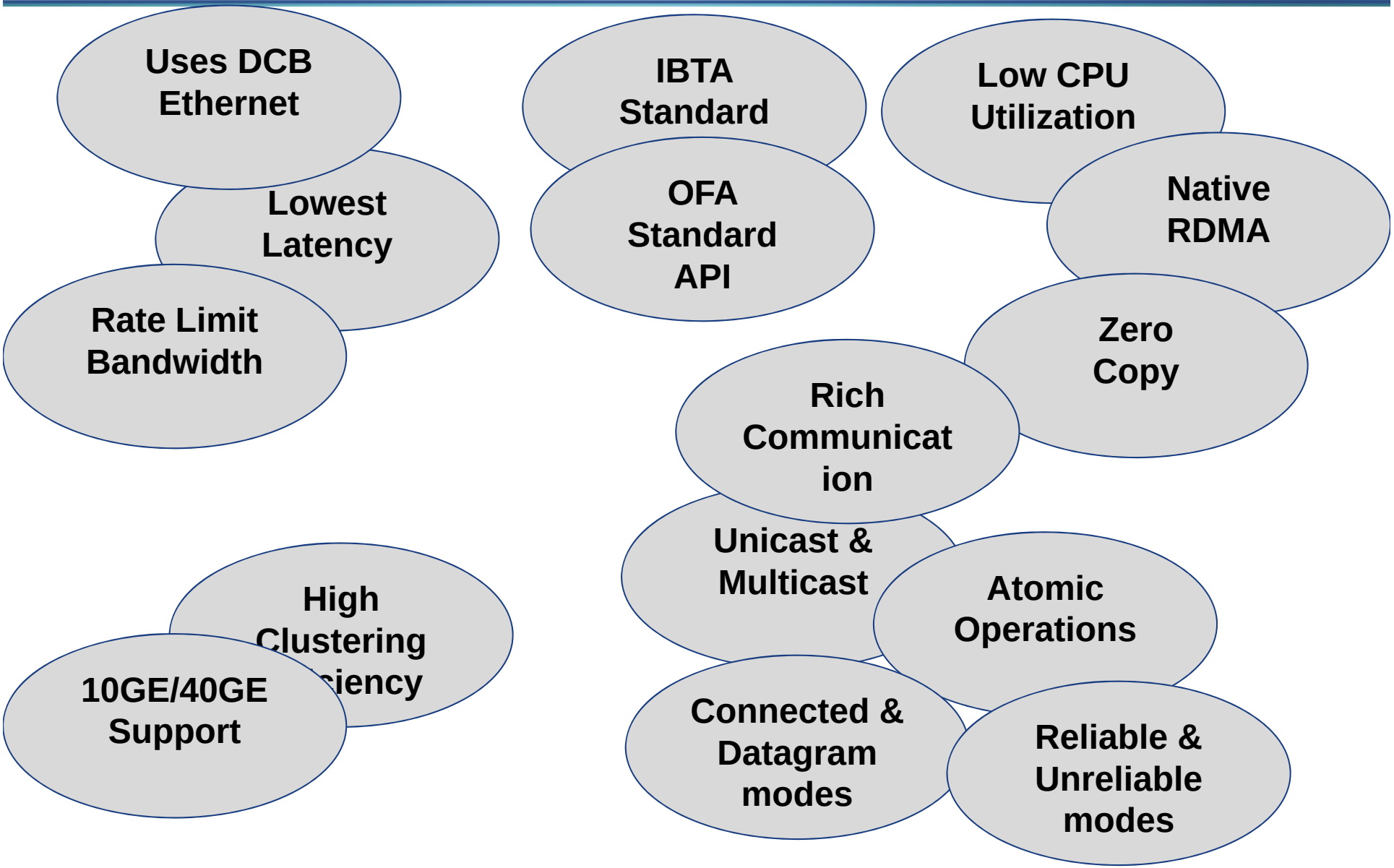
Hadoop Terasort Benchmark – RoCE vs. 10GE

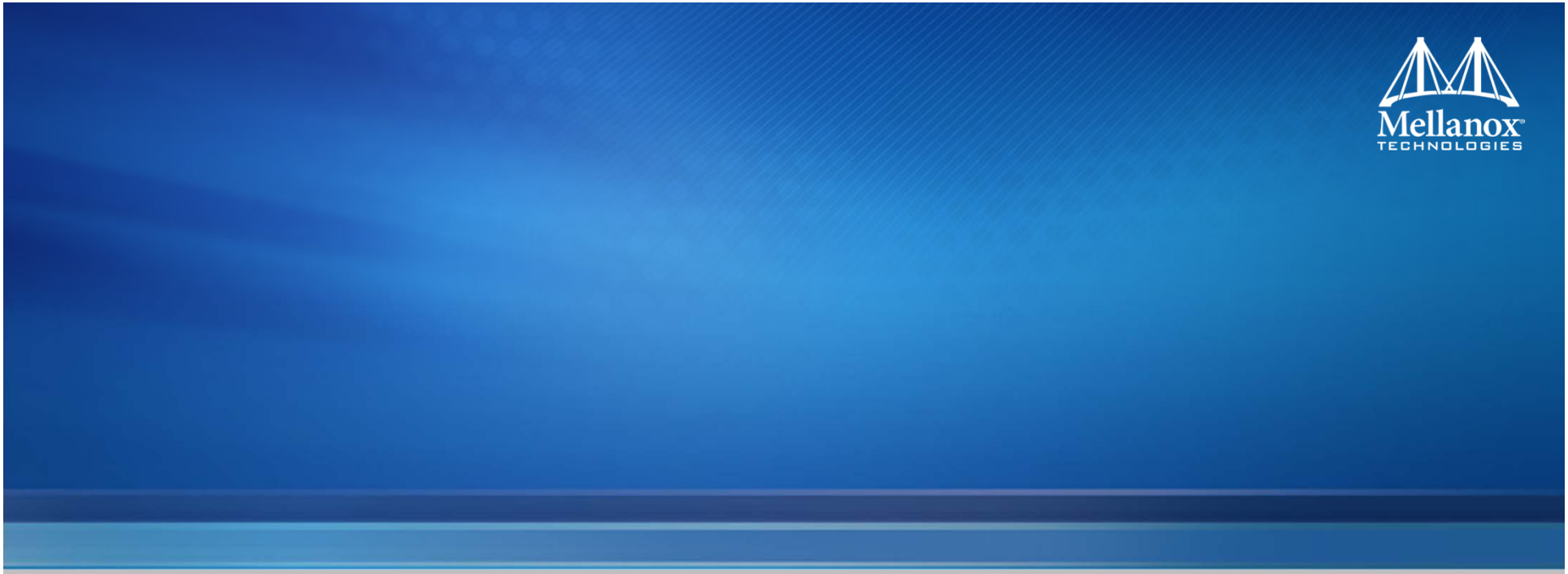


Application Level Benchmarks (others)

- Virtualization
  - Live Migration
    - Better CPU Utilization
  - Storage

RoCE Benefits





THANK YOU