Infiniband and 10GbE Low latency networks

September 2010

Presented By:

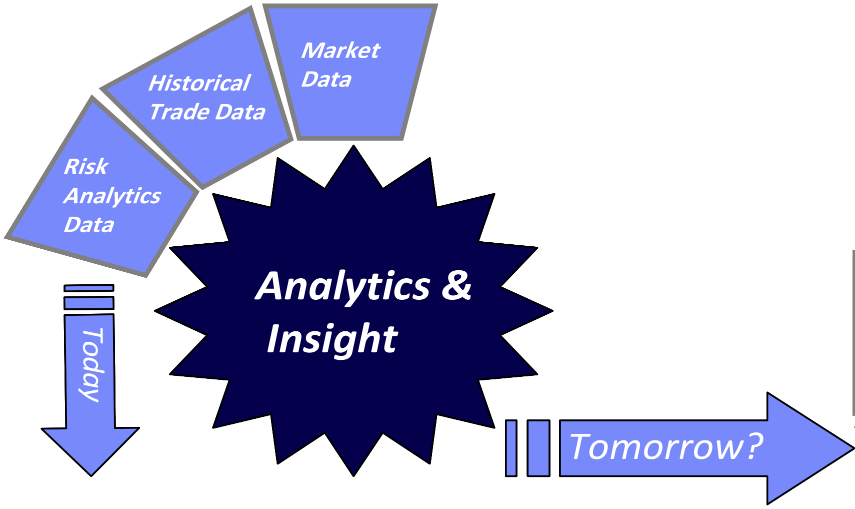
Michael Kagan

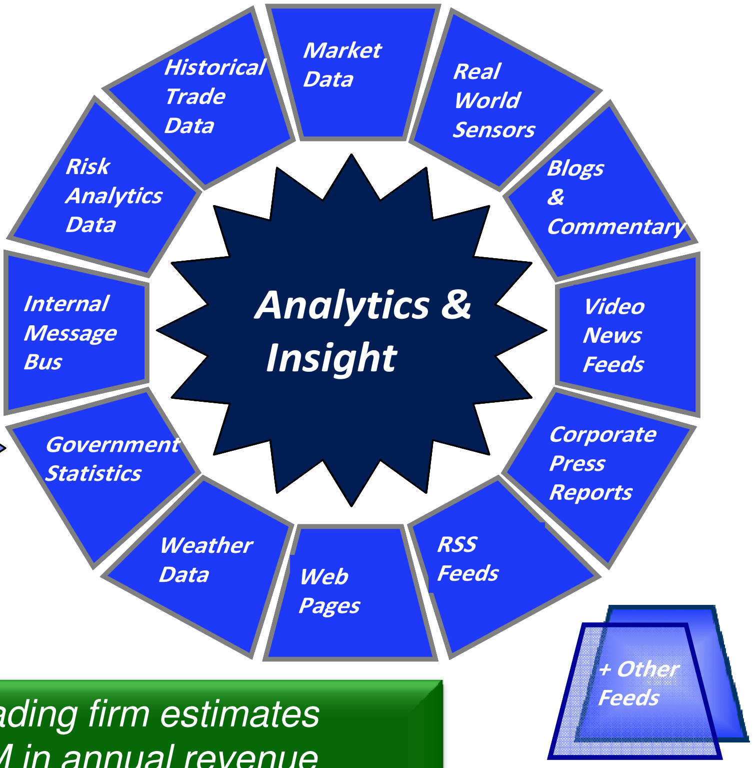
Chief Technology Officer

Mellanox TECHNOLOGIES

Financial Trading Market Trends

- **Explosive growth in messages that must be processed REAL TIME**
  - The volume, complexity & semantic depth of data that will be required to be analyzed will continue to increase significantly*
- **Capacity and latency performance is a serious and a real reliability concern**
  - Slow response → Lost revenue**





***One US equity trading firm estimates 1 millisecond = $4M in annual revenue*

Source: Tabb Group

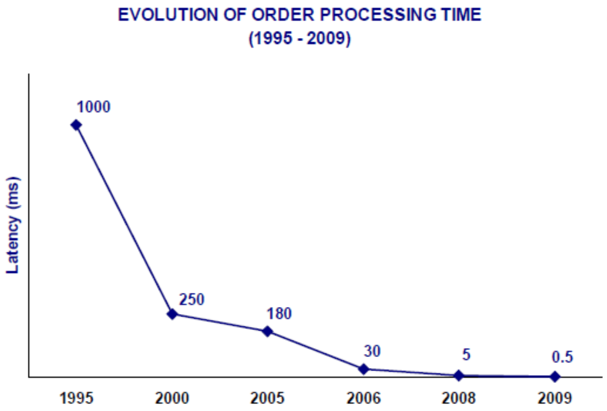
***Source: IBM**

Fabric Performance is critical

Achieve competitive advantage through fabric performance

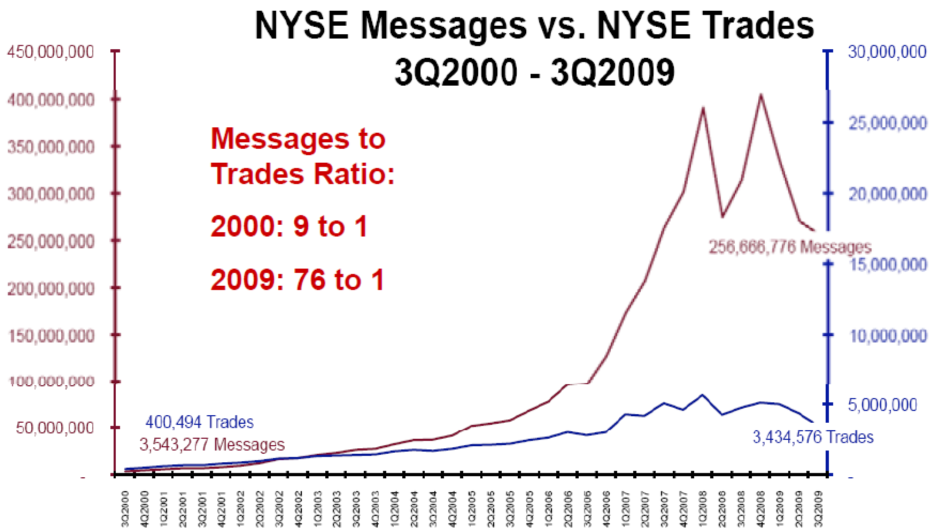
- High availability
  - Network data loss & downtime are not options
- Lowest latency
  - Every microsecond counts
- Highest throughput
  - Higher messages per second

NASDAQ MARKET CENTER			
Peak day		Peak second	
Message Volume	1,684,103,265	Messages	411,816
Order Volume	821,808,375	Orders	194,205
Share Volume	12,814,454,760	Executions	44,490



NYSE Euronext trading engine (various releases)

Between 1995 and 2009, the latency has been divided by 2,000, the throughput has been multiplied by 10,000 and the total number of orders processed by the trading engine has been multiplied by 100.


NYSE Messages vs. NYSE Trades 3Q2000 - 3Q2009

Messages to Trades Ratio:

2000: 9 to 1

2009: 76 to 1

2008/09: increased algorithmic trading, latency in microseconds, co-location

Connectivity Solutions must meet Market Needs

500% increase in capital market data volume

600% increase in share volume

Size of share trades shrink to 1/4

Source: NASDAQTrader.com 1997 to 2009 trend

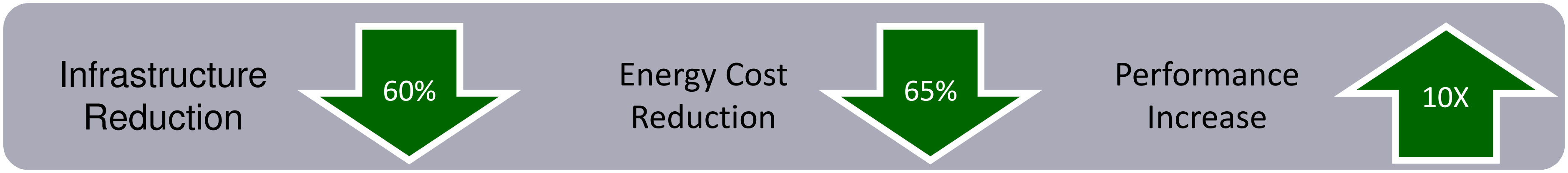
InfiniBand + Ethernet

- 1usec server-to-server latency
- 40Gb/s server-to-server throughput
- 3usec 10GigE server to InfiniBand server latency**

** when using BridgeX based Gateway

Mellanox Network Connectivity Aims & Benefits*

* Based on end-users testimonies



Infrastructure Reduction 60%

Energy Cost Reduction 65%

Performance Increase 10X

Typical Deployment Configurations - IB

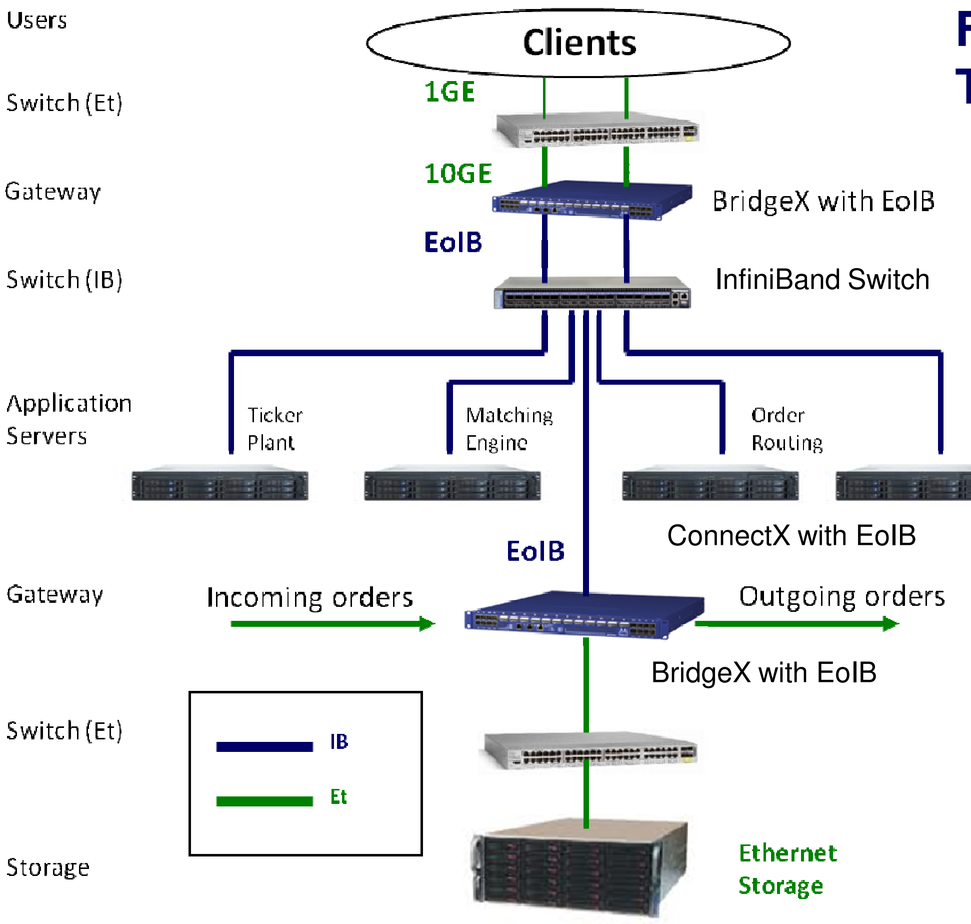
Financial

Ticker plant, order processing

Risk analysis

- High frequency trading
- Securities and investment services

High Frequency Trading



BridgeX™

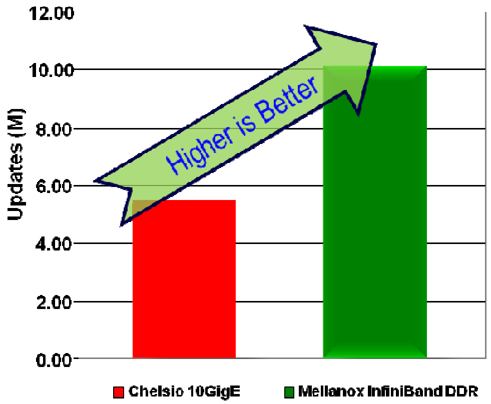
ConnectX-2

BridgeX™

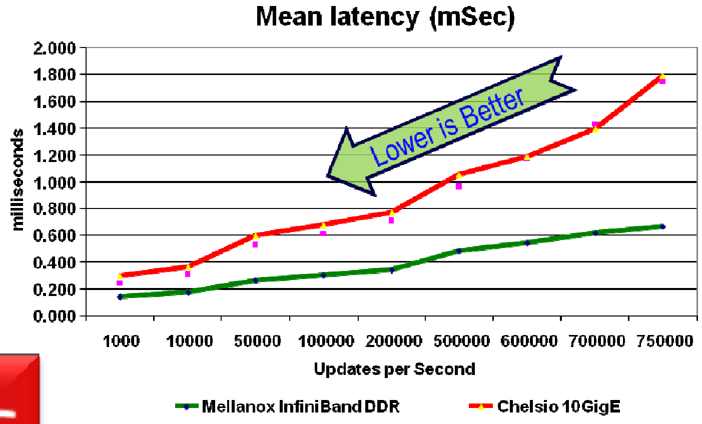
Highest Performance at Lowest TCO

Case Study: Reuters RMDS* Financial Application

- **82% higher updates/sec**

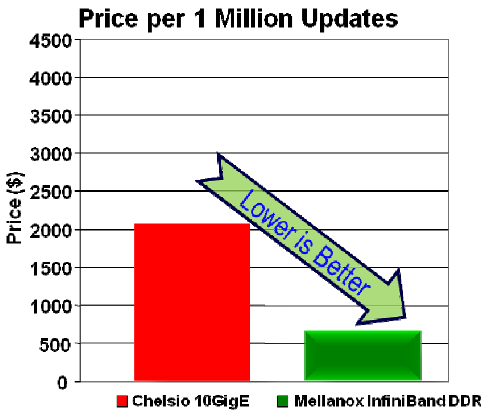


- **62% lower mean latency**

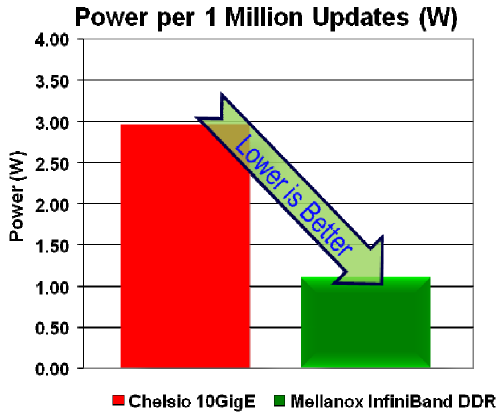


REUTERS

- **Costs 70% lower**



- **3X less power consumption**



Source: STAC

*Reuters Market Data System

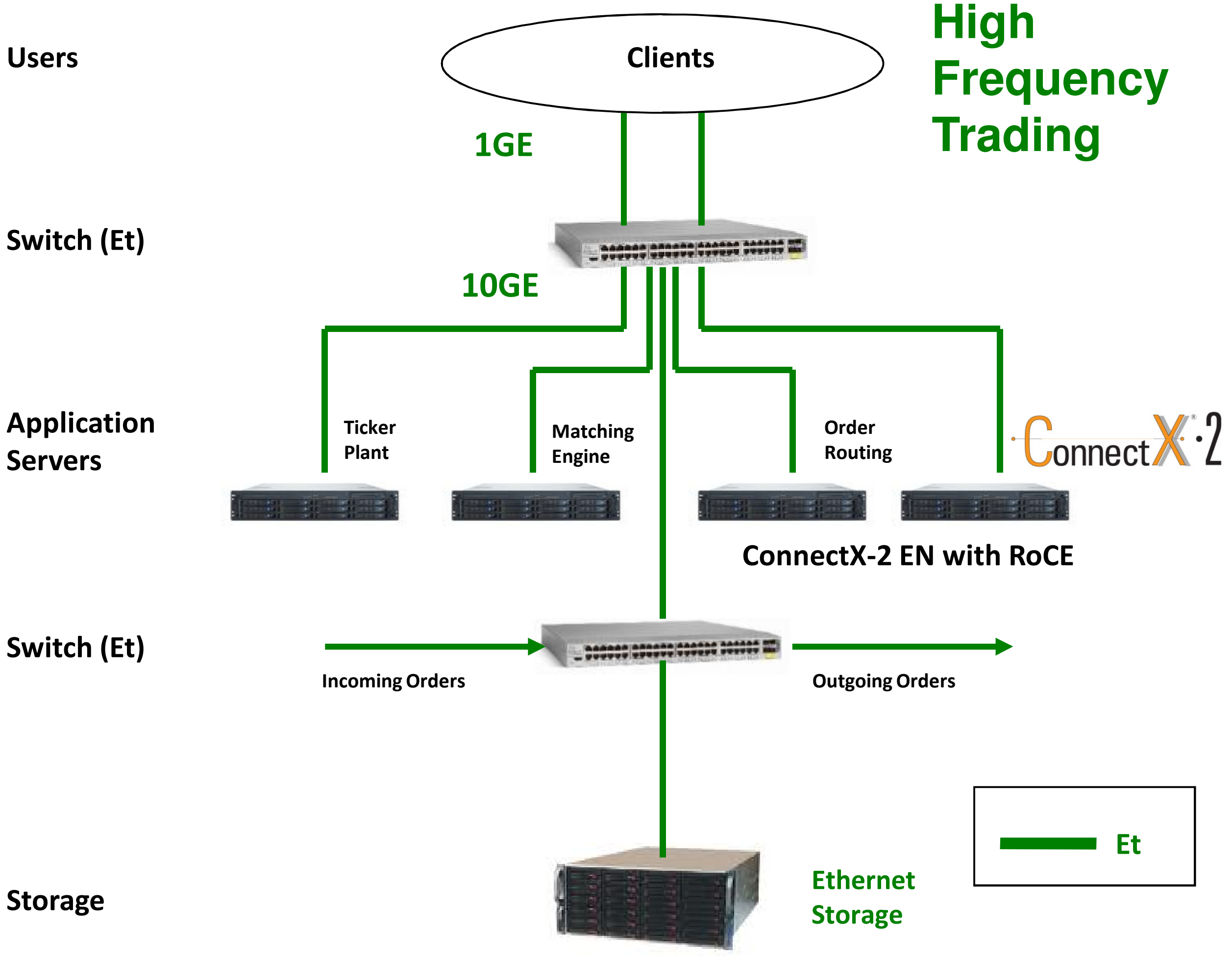
Typical Deployment Configurations - Ethernet

Financial

Ticker plant, order processing

Risk analysis

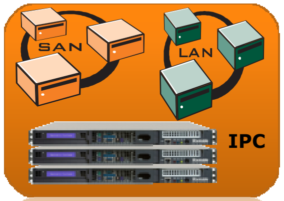
- High frequency trading
- Securities and investment services



RoCE, lowest latency over Ethernet

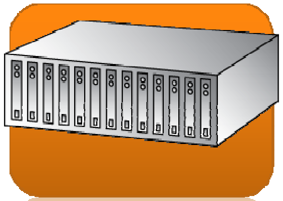
RoCE (RDMA over Converged Ethernet)

- Efficient RDMA & Send/Receive semantics over Ethernet
- Provides low-latency and line-rate bandwidth
- Adds efficient and reliable memory management
- Improved Ethernet performance with data center bridging
- Enhanced data center I/O consolidation

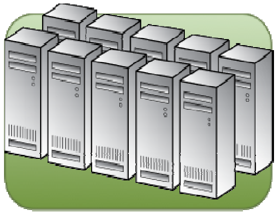


I/O Consolidation

Cloud Computing



Storage Applications



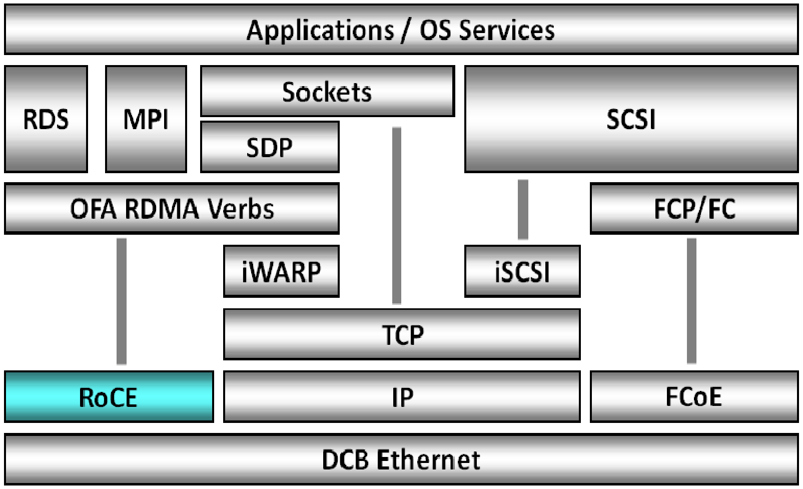
Consolidation/Green

RoCE (RDMA over Converged Ethernet)

InfiniBand transport over Ethernet

- Efficient, light-weight transport, layered directly over Ethernet L2
- Takes advantage of PFC (Priority Flow Control) in DCB Ethernet
- IBTA standard, supported in OFED 1.5.1, Support for commonly used Linux releases

Rich communication services (full verbs support)



Applications / OS Services

RDS

MPI

Sockets

SDP

SCSI

OFA RDMA Verbs

FCP/FC

iWARP

iSCSI

TCP

RoCE

IP

FCoE

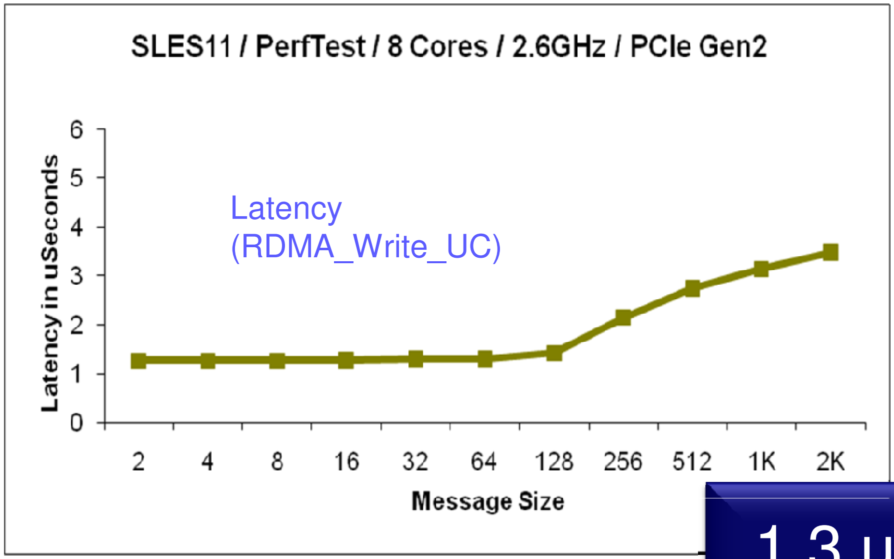
DCB Ethernet

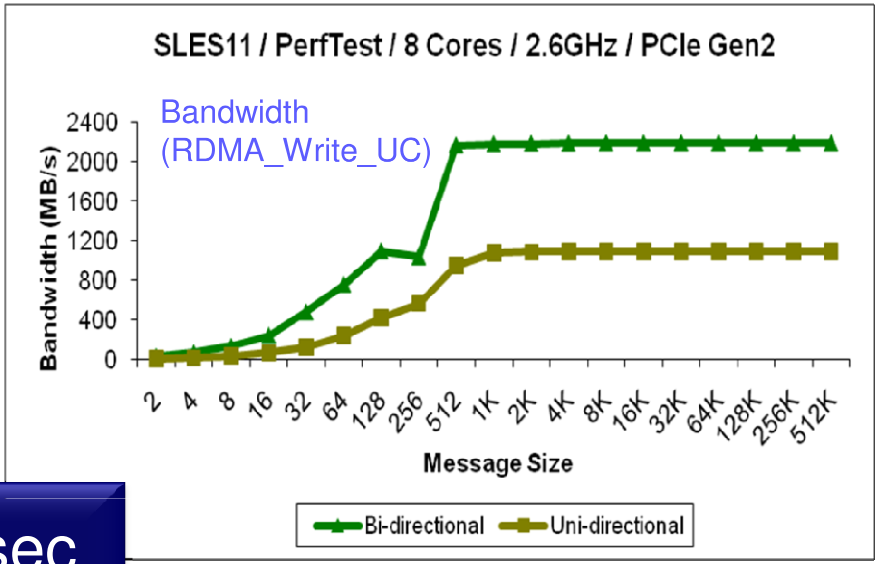
Most comprehensive low latency features

Feature	CX2 RoCE	iWARP
OFA Verbs Compliant	X	X
Ubiquitous Ethernet Management	X	X
Most Proven and Cost-Effective RDMA Transport Protocol	X	
Reliable Connected Service	X	X
Datagram Service	X	
RDMA and Send/Receive Semantics	X	X
Atomic Operations	X	
User Level Multicast	X	
User Level IO Access / Kernel Bypass / Zero Copy	X	X
Stateless Traffic De-multiplexing, dedicated QoS for RDMA flows	X	
Can operate over lossy Ethernet (without PFC enabled)		X
IP Routing	Future	X
Latency	1.3usec	10+usec

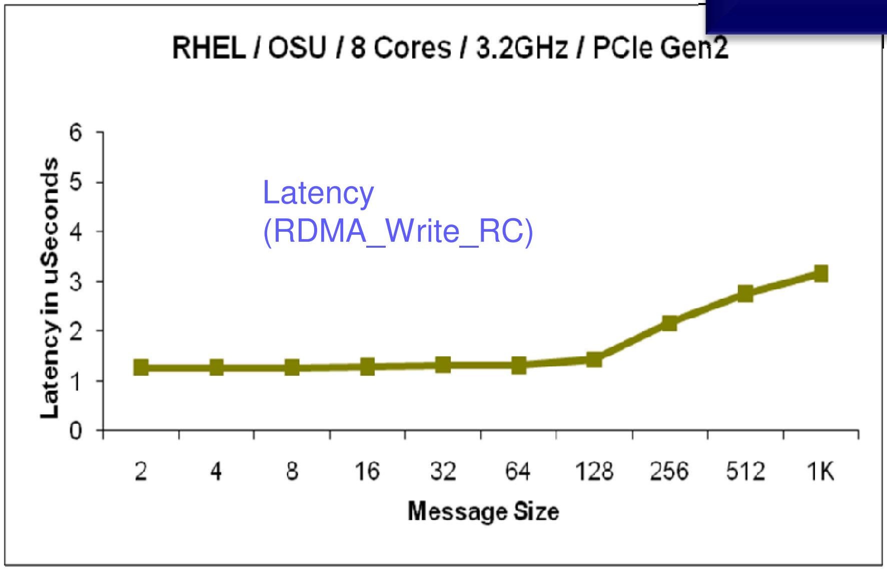
RoCE App-level Benchmarks

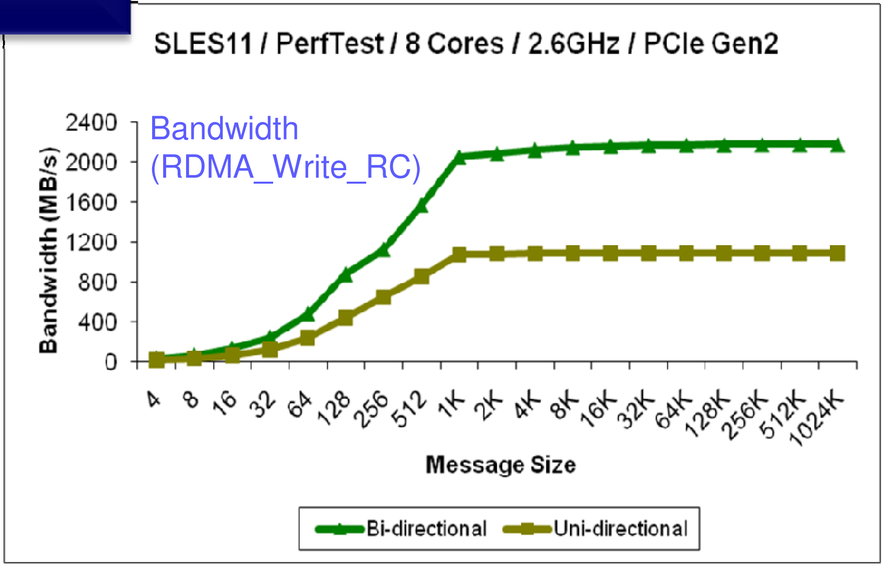
1/5th the Latency Delivered by other Ethernet solutions



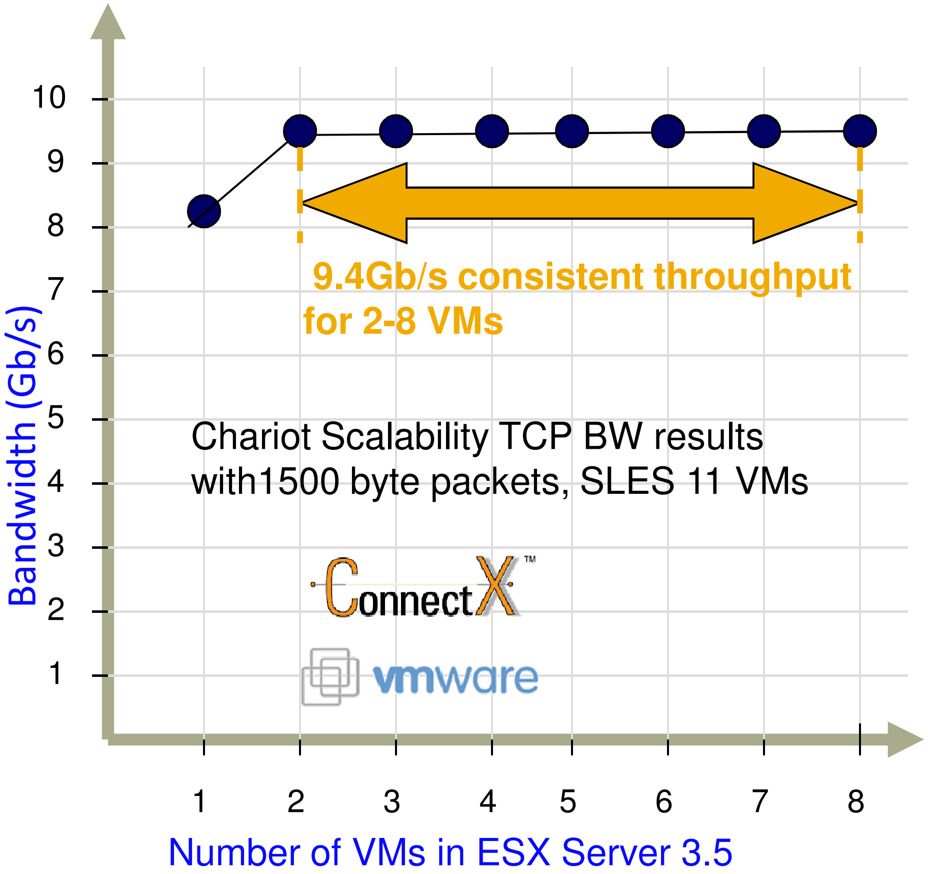


1.3 usec



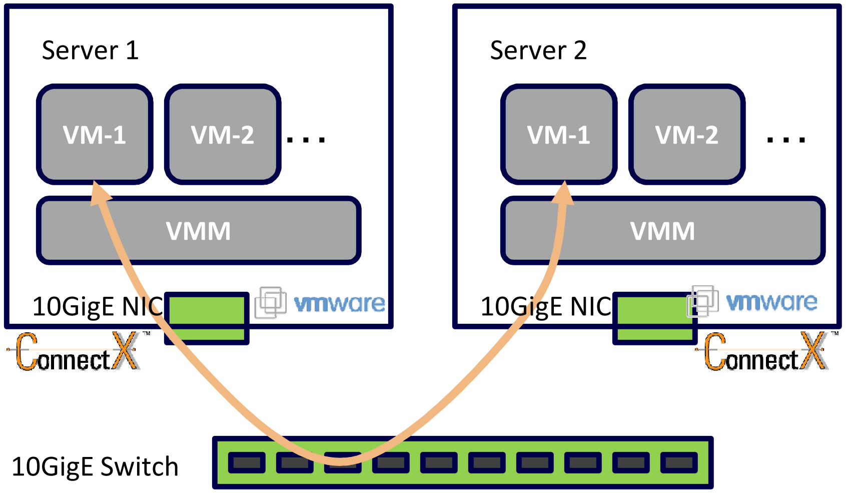


Superior Virtualization Performance



9.4Gb/s consistent throughput for 2-8 VMs

Chariot Scalability TCP BW results with1500 byte packets, SLES 11 VMs



Minimum latency 22.3usec compared to other solutions minimimum31.4usec

TCP Latency results with1500 byte packets, RHEL 5.3 VMs

More Virtual Machines per Server

More VM applications serviced faster

Faster VM migration (vMotion)

Future proof - RoCE

NYSE Data Fabric Financial Services App

NYSE Euronext

ConnectX-2 EN with RoCE

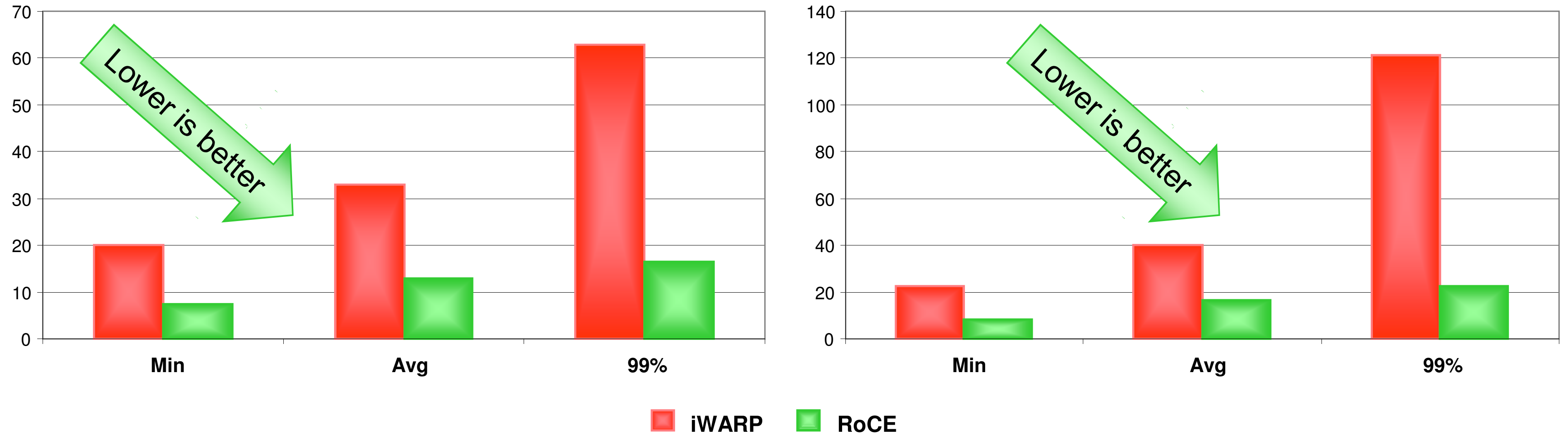
Average latency for 100-200 bytes messages
12 – 16 microseconds

Alternative Solution
10GigE NIC with iWARP

Average latency for 100-200 bytes messages
33 – 40 microseconds

RoCE vs. iWARP Latency @ 100B Message Size (usec)

RoCE vs. iWARP Latency @ 200B Message Size (usec)

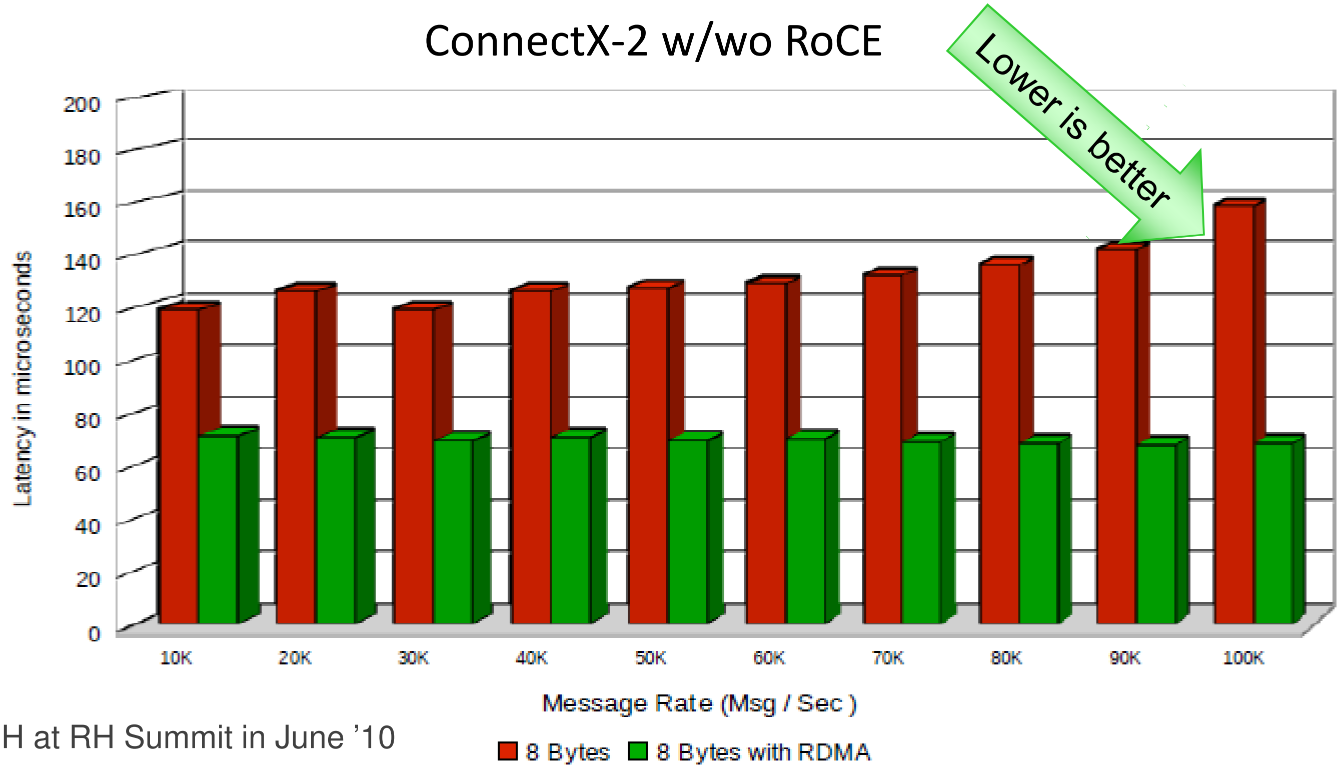


62% better on execution time vs. 10GigE with iWARP

RoCE: Performance and Profitability
Latency remains constant as msg rate increases

MRG 1.3 Red Hat Enterprise 6.0 over RoCE*

>42%


ConnectX-2 w/wo RoCE

Lower is better

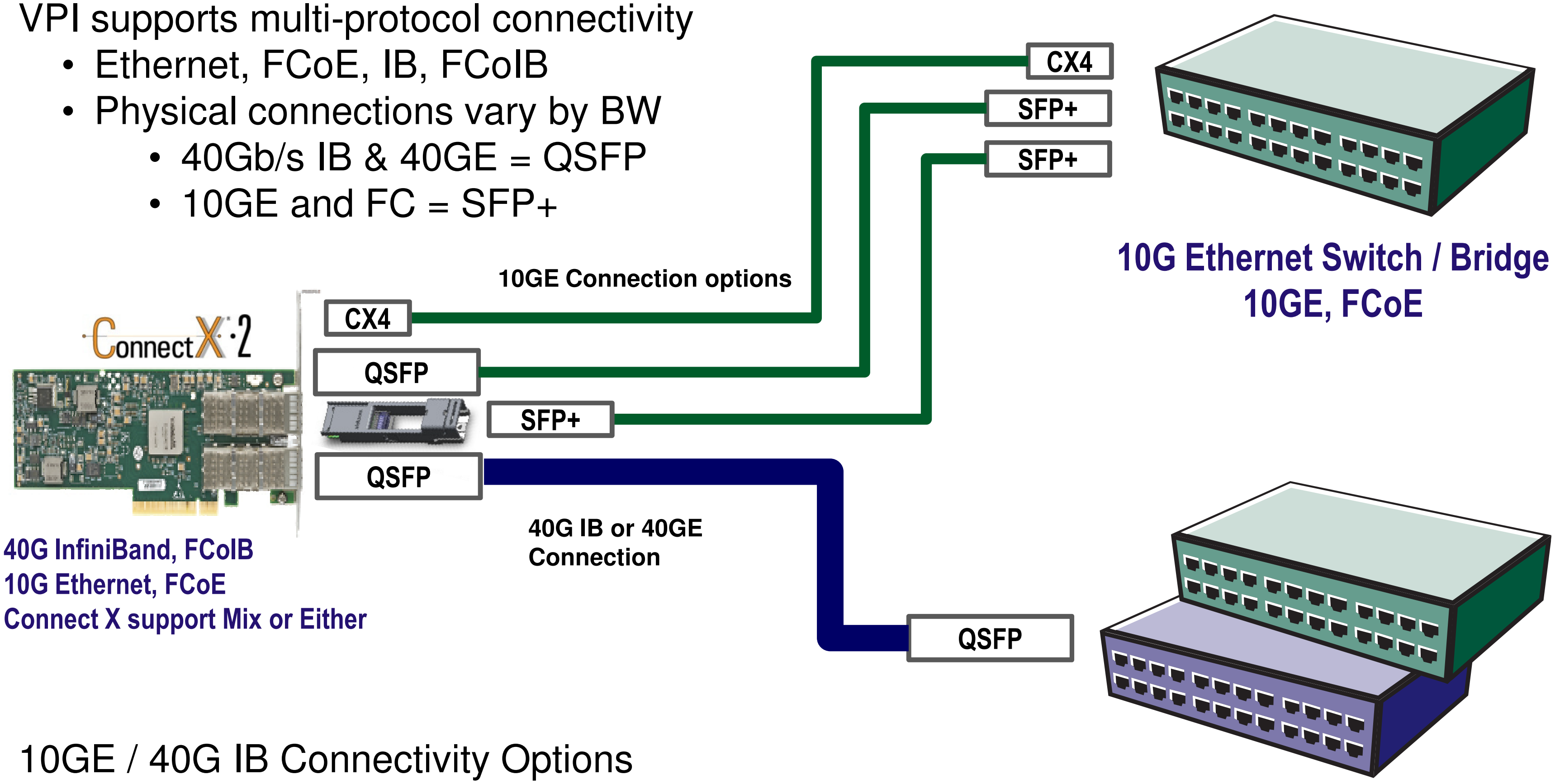
*Presented by RH at RH Summit in June '10

1.2 Million Acknowledged Messages per Second

VPI Connectivity Options

VPI supports multi-protocol connectivity

- Ethernet, FCoE, IB, FCoIB
- Physical connections vary by BW
  - 40Gb/s IB & 40GE = QSFP
  - 10GE and FC = SFP+



10G Ethernet Switch / Bridge
10GE, FCoE

40G InfiniBand, FCoIB
10G Ethernet, FCoE
Connect X support Mix or Either

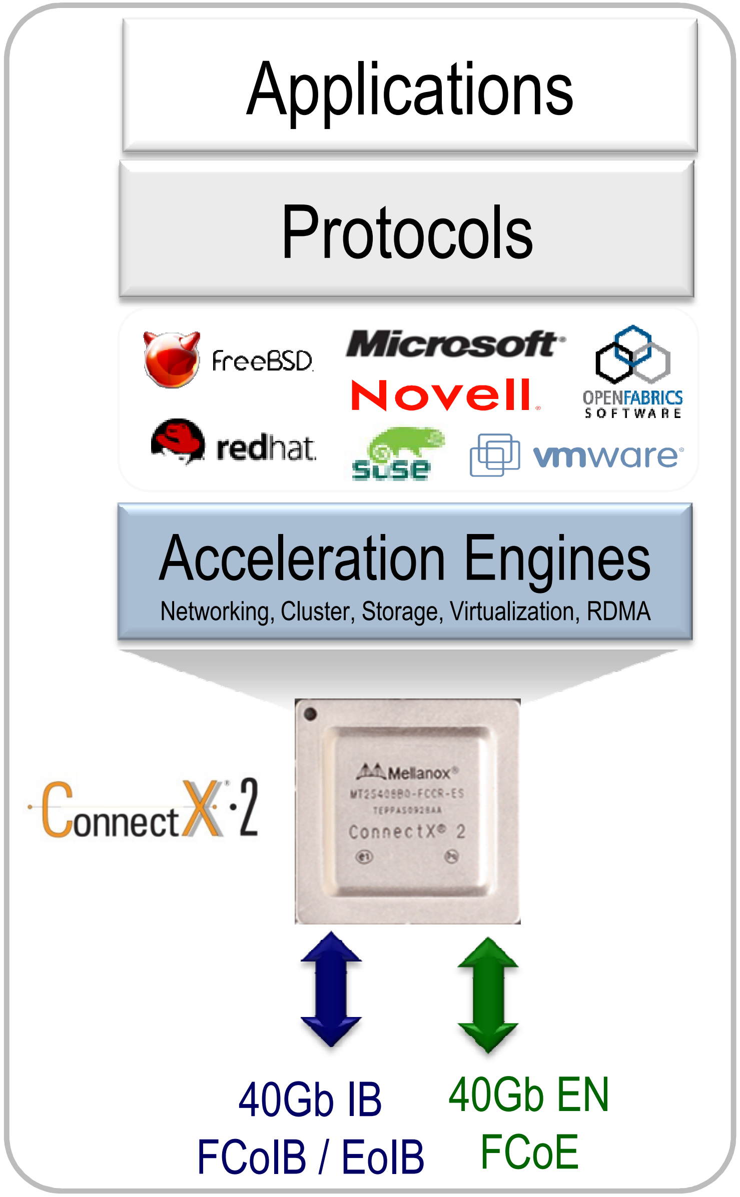
40Gb/s InfiniBand Switch
0r 40GE Switch

10GE / 40G IB Connectivity Options

- CX-4 connects either 10G or 20/40G IB
- Hybrid cables (QSFP to SFP+)
- Standard SFP+ for 10G using QSFP to SFP+ Adapter

Flexibility / Consolidation:
Virtual Protocol Interconnect (VPI)

- **Broad OS / Virtualization support**
  - Strong software ecosystem foundation
- **Consolidation / Extensive connectivity options and features**
  - Cost-Effective convergence over:
    - InfiniBand - FCoIB and EoIB
    - Ethernet - FCoE
- **Performance**
  - Application acceleration, PCIe 2.0, low-latency, high-bandwidth



Applications

Protocols

Acceleration Engines

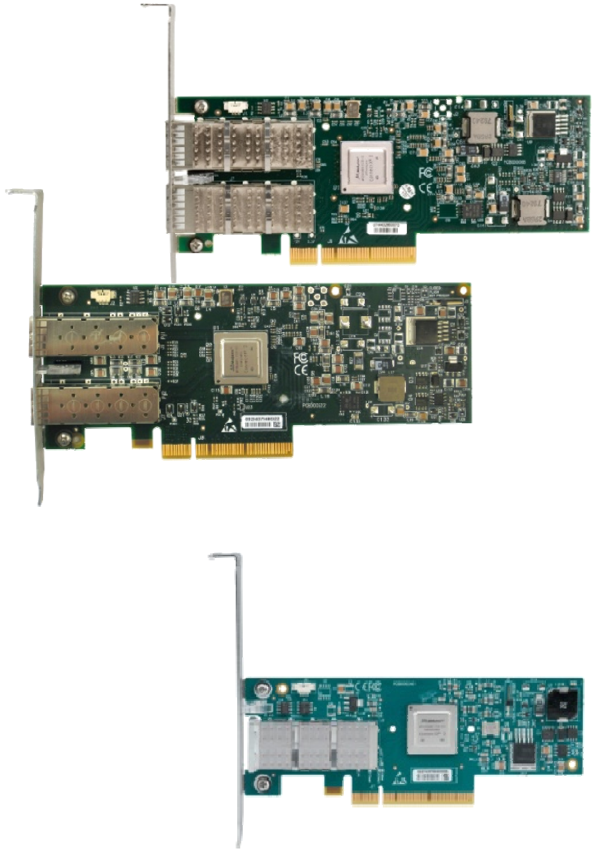
Networking, Cluster, Storage, Virtualization, RDMA

40Gb IB
FCoIB / EoIB

40Gb EN
FCoE

Common solutions for InfiniBand & 10/40GigE

- **Solutions to address IB & GE deployment options**
- **40Gb/s InfiniBand**
  - Latency using IB Verbs is ~1µs
  - Bandwidth of 6.6GB/s
  - SR-IOV supported for Virtualization
  - RDMA hardware offload with zero copy
- **10 and 40 Gigabit Ethernet**
  - RDMA hardware offload with zero copy
    - Now made available with RoCE
  - Latency using RoCE Verbs is ~1.3µs
  - Latency using standard sockets is ~6µs
  - SR-IOV supported for Virtualization
  - Data Center Bridging (DCB) for PFC and CC
  - T-11 FCoE



Thank You

Contacts:

Michael Kagan, Chief Technology Officer	michael@mellanox.com
Colin Bridger, Region Manager EMEA	colin@mellanox.com
Yossi Avni, VP EMEA	yossia@mellanox.co.il

Mellanox®
TECHNOLOGIES

Reference Material

Mellanox®
TECHNOLOGIES

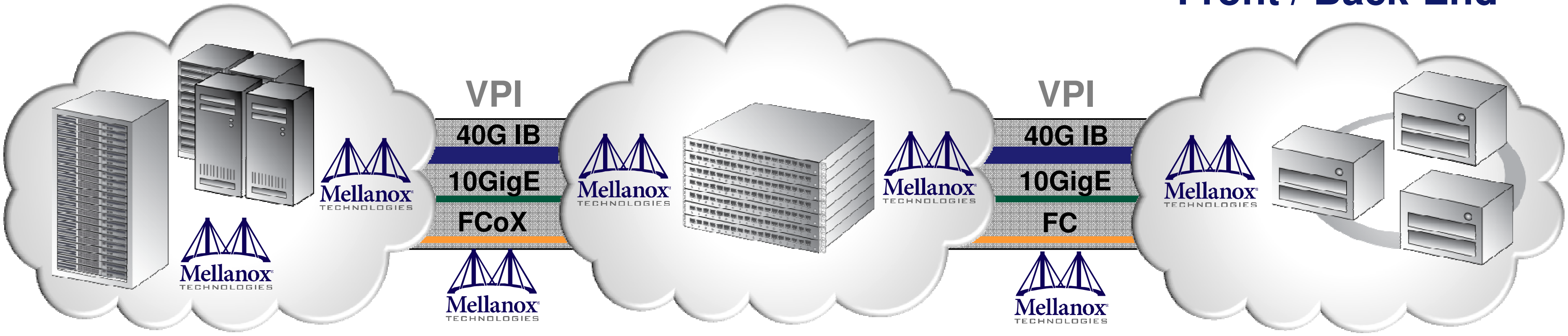
Financial Markets choose Mellanox for Performance

- **3 out of the 5 largest banks worldwide are Mellanox customers.**
- **3 out of the 5 largest stock exchanges worldwide are customers.**
- **A majority of the Algorithmic-Trading and Hedge-Funds world wide are using Mellanox products.**

Financial Benchmark Examples

Benchmarks Comparison Criteria	Performance Results
Latency on RoCE	1.3usec over OpenFabrics verbs API , Linux
Latency over TCP sockets	6.4usec (without kernel bypass), Linux
Latency over UDP sockets	5.9usec (without kernel bypass) 2-3usec (expected with kernel bypass in Q4 2010), Linux
Highest throughput over TCP sockets , unidirectional (CPU utilization)	9.4Gb/s for 1500 byte packets (5%), Linux
Highest throughput over UDP sockets unidirectional (CPU utilization)	9.4Gb/s for 1500 byte packets (3%), Linux
NYSE Data Fabric Performance	12-16usec average latency (100-200byte msgs), 1.2M messages/sec
IBM WebSphere LLM Performance	Latency – 4usec, 1M messages per second). Record with any 10GigE NIC
Red Hat Enterprise MRG	Latency – 70usec. 1.2M messages per second. Record with any 10GigE NIC
29West	Coming soon (expected <10usec mean latency, 1.3M messages per second)
TIBCO	Coming soon

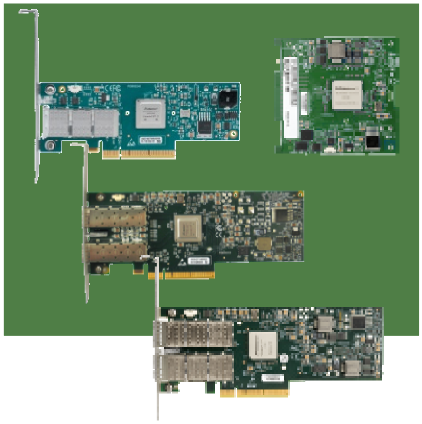
End-to-End Network Connectivity



Server / Compute

Switch / Gateway

Storage
Front / Back-End

Silicon Adapter, Switch, Gateway	Host Adapters	Network Switches and Gateways	Cables
			

Comprehensive System Products Portfolio

Modular Switch IS5x00 Series

IS5X00
108 to 648 ports modular switch

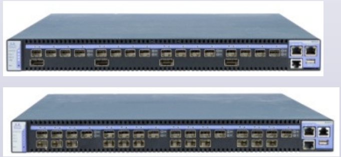
Gateways

BX4010
QDR to 10GbE and/or 2/4/8G FC

BX5020
40Gb/s IB to Eth or FC Gateway

BX1020
10GbE to 2/4/8G FC FCoE to FC Gateway

Dec'10



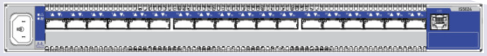
Edge Switch IS50xx Series

IS5031
1U 18/36 port QSFP
Managed 108 Nodes

IS5035
1U 36 port QSFP
Managed 2000 Nodes

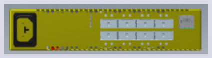
IS5030
1U 36 port QSFP
Managed 108 Nodes

IS5025
1U 36 port QSFP
Unmanaged



Nov'10

IS50231U 18 port QSFP Unmanaged – No FRUs



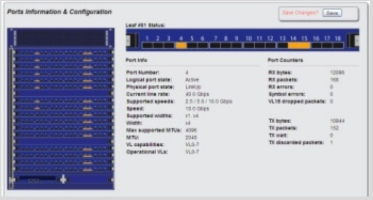
Nov'10

IS5022 1U 8 port QSFP Unmanaged – No FRUs

Mellanox M-1
E-2-E Cluster
Support Services

Fabric Management

FabricIT



Ports Information & Configuration

Save Changes? Save

Leaf #01 Status:

Port Info

Port Number: 4
Logical port state: Active
Physical port state: LinkUp
Current line rate: 40.0 Gbps
Supported speeds: 2.5 / 5.0 / 10.0 Gbps
Speed: 10.0 Gbps
Supported widths: x1, x4
Width: x4
Max supported MTUs: 4096
MTU: 2048
VL capabilities: VL0-7
Operational VLs: VL0-7

Port Counters

RX bytes: 12096
RX packets: 168
RX errors: 0
Symbol errors: 0
VL15 dropped packets: 0

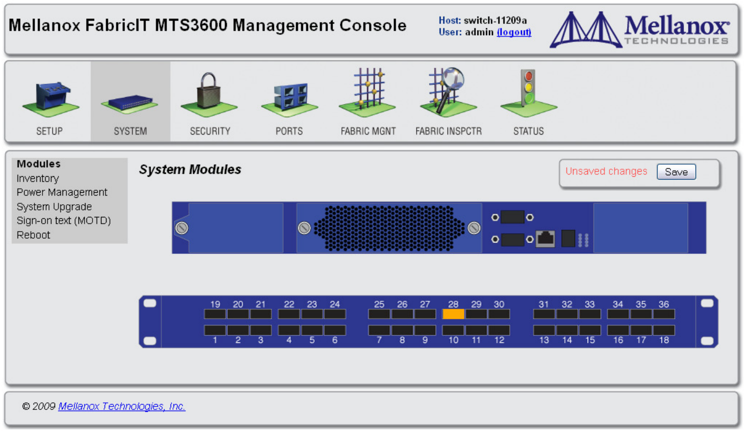
TX bytes: 10944
TX packets: 152
TX wait: 0
TX discarded packets: 1

Cables

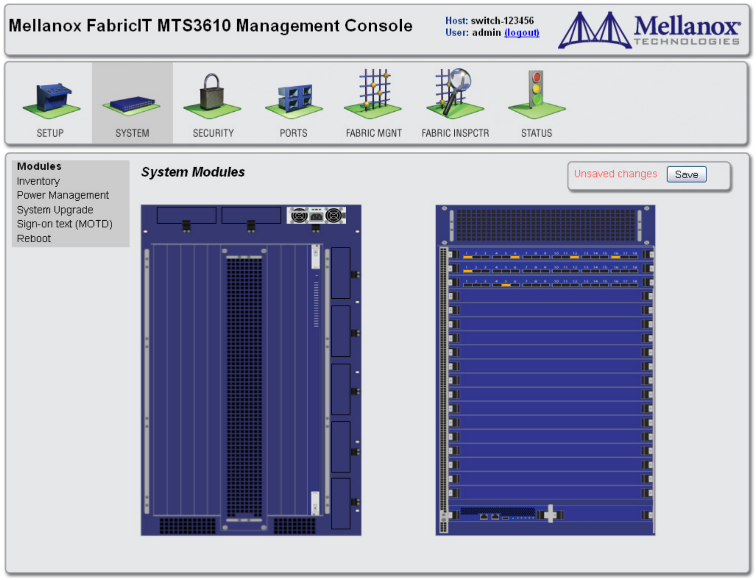
SFP+
QSFP

FabricIT Management Suite

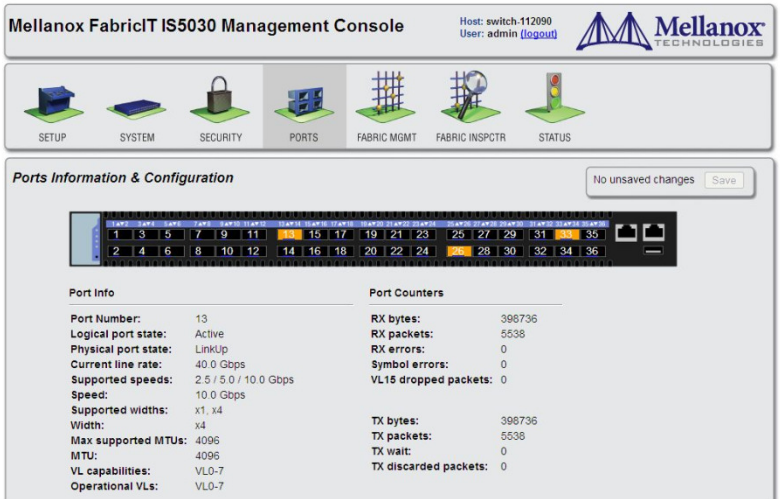
- **Management (CLI, WebUI) unified access**
  - RS232 Console (CLI only)
  - 10/100 Management Port
  - IPoIB in-band interface
- **FabricIT Chassis Manager (SCM)**
  - Chassis management: sensor reading, alerts, firmware update, counters reading
- **FabricIT Fabric Manager (EFM)**
  - SM, diagnostics, Adaptive Routing & Congestion Managers, Cluster diagnostics
  - Upgradeable ordering option (license)



Shark Rev2 WebUI



Mammoth Rev2 WebUI



Ports Information & Configuration

Port Info

Port Number: 13
Logical port state: Active
Physical port state: LinkUp
Current line rate: 40.0 Gbps
Supported speeds: 2.5 / 5.0 / 10.0 Gbps
Speed: 10.0 Gbps
Supported widths: x1, x4
Width: x4
Max supported MTUs: 4096
MTU: 4096
VL capabilities: VL0-7
Operational VLs: VL0-7

Port Counters

RX bytes: 398736
RX packets: 5538
RX errors: 0
Symbol errors: 0
VL15 dropped packets: 0

TX bytes: 398736
TX packets: 5538
TX wait: 0
TX discarded packets: 0

40Gb/s Switch System Portfolio

IS5025

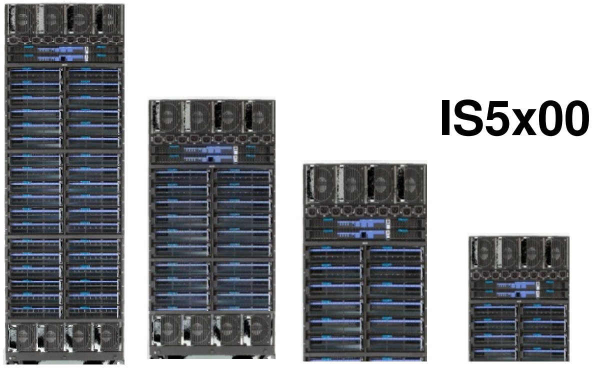
- Unmanaged (Externally Managed)
- Host Subnet Manager based on MLNX_OFED
- For cost conscious customers with their own management software

IS5030

- Chassis Management
- Fabric Management for small clusters (up to 108)
- Low cost entry level managed switch

IS5035

- Fully managed
- Fabric Management for large clusters (up to 2000)

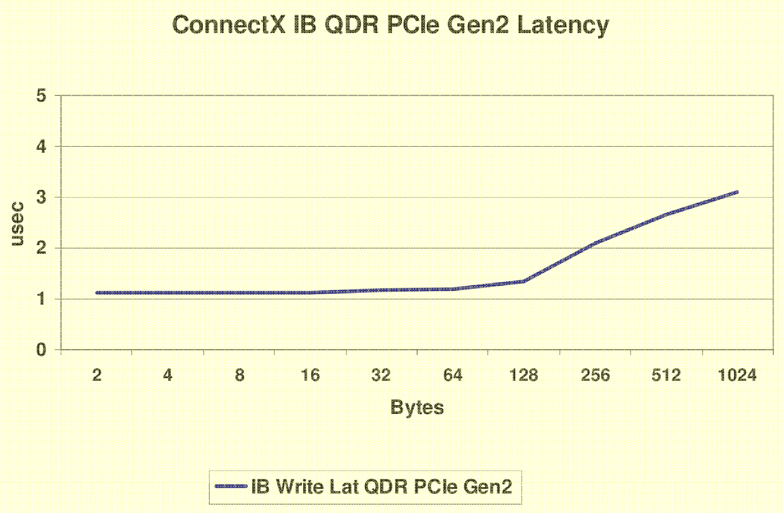


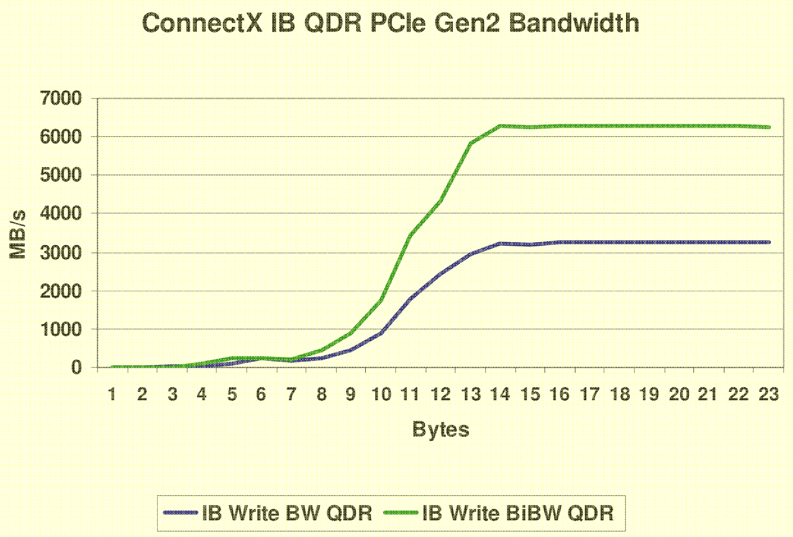
IS5x00

- Modular chassis systems
- Designed for large to Peta-scale computing
- Redundant components for high availability

InfiniBand Fabric Performance

- **High throughput**
- **Low Latency**
- **Lowest CPU utilization**
  - RDMA hardware offload
- **Lossless transport**
- **Lowest power per 1Gb/s**


ConnectX IB QDR PCIe Gen2 Latency

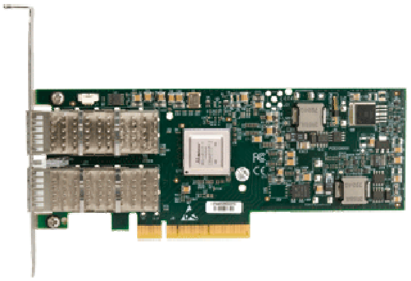

ConnectX IB QDR PCIe Gen2 Bandwidth

*

Industry-Leading End-to-End InfiniBand

Mellanox TECHNOLOGIES

- **Mellanox is the InfiniBand market and performance leader**
  - First to market with 40Gb/s adapters and 120Gb/s switches
    - Roadmap to 100Gb/s adapters in 2011
  - Strong industry adoption of 40Gb/s InfiniBand
    - ~57% of revenue
  - Industry's highest density switches at 51.8TB
    - With comprehensive fabric management software
  - BridgeX Gateways provide I/O convergence and flexibility
  - Robust, certified copper and optical cables
  - 100% of IB-connected Top500 systems



(Worldwide Tier-1 Server OEM Availability)

Bull
曙光 DAWNING
DELL
FUJITSU
hp invent
IBM
Sun microsystems
PLATFORMS

Breadth and Leadership: 10 Gigabit Ethernet Innovation

Ethernet Leadership

- First to market with dual-port PCIe Gen2 10GigE adapter
- First to market with 10GigE w/FCoE with hardware offload
- Industry’s lowest latency Ethernet ~ 1.3us
- First to market with 40GigE adapter
- Industry’s most flexible FCoE bridge
  - E to FC, IB to FC, IB to E

BEST OF INTEROP
WINNER
2008
Presented by InformationWeek

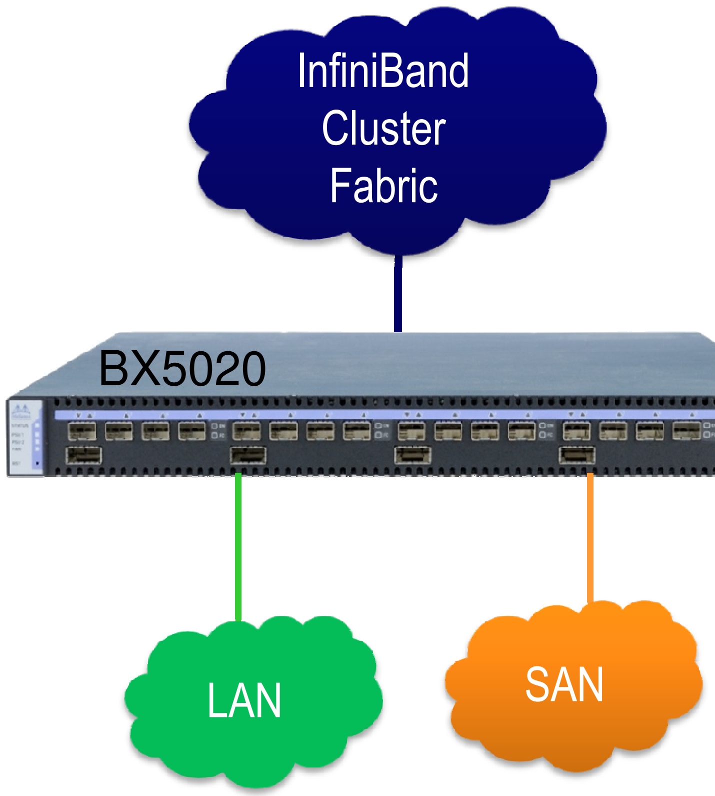
Industry-wide Acceptance and Certification

- Multiple tier-1 server OEM design wins
  - Servers, LAN on Motherboard (LOM), and storage systems
- VMware Virtual Infrastructure 3.5 & vSphere
- Citrix XenServer 4.1 in-the-box support
- Windows Server 2003 & 2008, RedHat 5, SLES 11

CITRIX · CERTIFIED FOR Windows Server 2008 · Certified for Microsoft® Windows Server® 2003 · Novell. · redhat. · vmware®

BX5020 VPI Gateway

- **Server facing ports**
  - Four 40Gb/s IB ports at line rate
  - Connects to InfiniBand Switch
- **LAN/SAN ports**
  - Up to 12 10GigE ports at line rate
  - Up to 16 1/2/4/8G FC ports at line rate
- **Lowest server to LAN/SAN latency**
  - Less than 200nsec
- **Seamless integration**
  - Applications run over standard Ethernet and FC API



BridgeX™