Paving The Road to Exascale Computing

HPC Technology Update

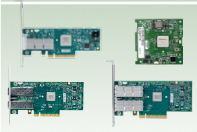
PC Cluster Consortium

Todd Wilde, Director of Technical Computing and HPC

Mellanox HPC Technology Accelerates the TOP500

- InfiniBand has become the de-factor interconnect solution for High Performance Computing
- InfiniBand is the most used interconnect on the TOP500 list with 224 systems
- FDR InfiniBand connected systems doubled to 45 systems since June '12
- Mellanox InfiniBand is the interconnect of choice for Petascale systems

Comprehensive End-to-End 10/40/56Gb/s Ethernet and 56Gb/s InfiniBand Portfolio

ICs	Adapter Cards	Switches/Gateways	Software	Cables
				

Scalability, Reliability, Power, Performance

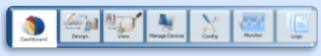
Mellanox InfiniBand Paves the Road to Exascale



PetaFlop
Mellanox Connected


The Mellanox Advantage

Host/Fabric Software Management



- UFM, MLNX-OS
- Integration with job schedulers
- Inbox drivers from major distributions

Application Accelerations

- Collectives Accelerations (FCA/CORE-Direct)
- GPU Accelerations (RDMA for GPUDirect)
- MPI/SHMEM/PGAS
- RDMA
- Quality of Service

Networking Efficiency/Scalability

- Dynamically Connected Transport
- Adaptive Routing
- Congestion Management

Server and Storage High-Speed Connectivity

- Latency
- Bandwidth
- CPU Utilization
- Message rate

FDR INFINIBAND TECHNOLOGY

THE NEXT GENERATION OF
HIGH-PERFORMANCE SCALABLE CONNECTIVITY

FDR InfiniBand New Features and Capabilities

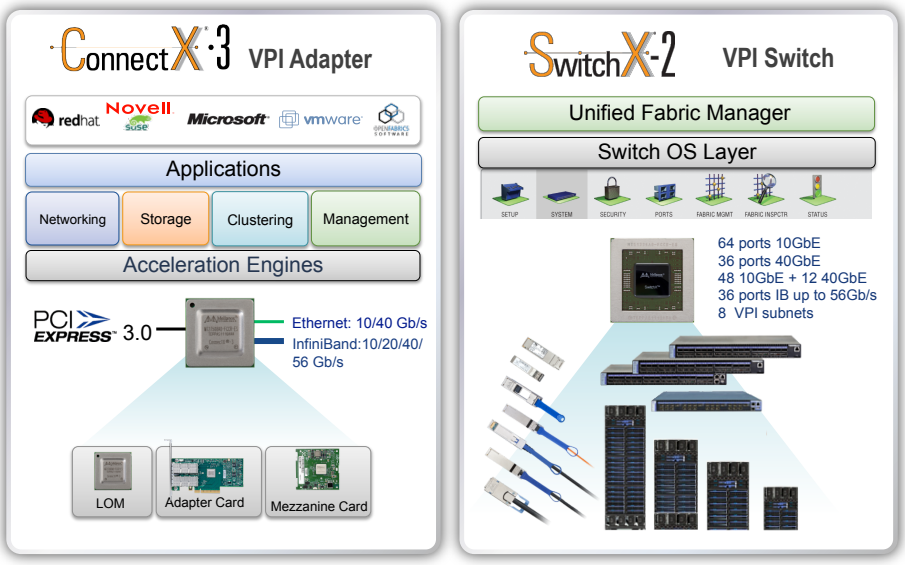
Performance / Scalability

- >100Gb/s bandwidth, <0.7usec latency
- PCI Express 3.0
- InfiniBand Routing and IB-Ethernet Bridging

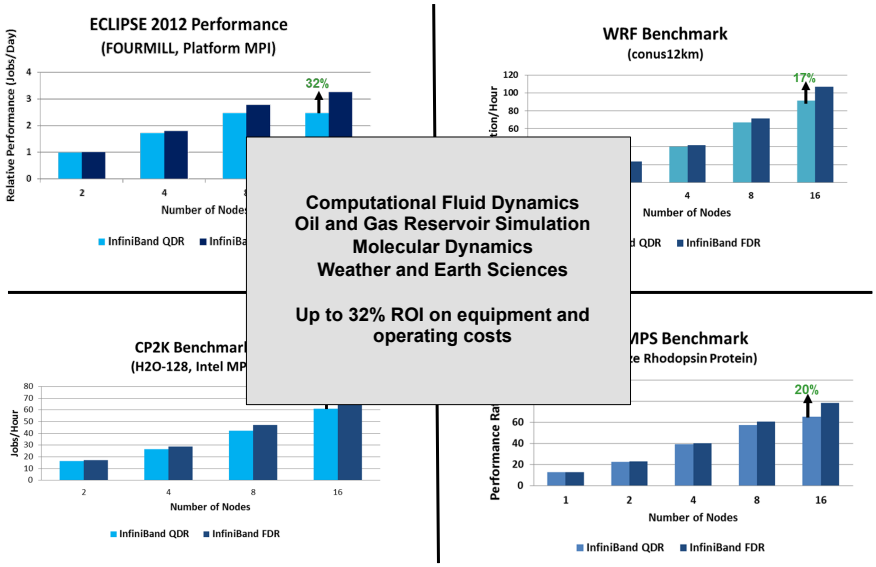
Reliability / Efficiency

- Unbreakable Link Technology
  - Forward Error Correction
  - Link quality auto-sensing
  - Link Level Retransmission
- Link bit encoding – 64/66
- Lower power consumption

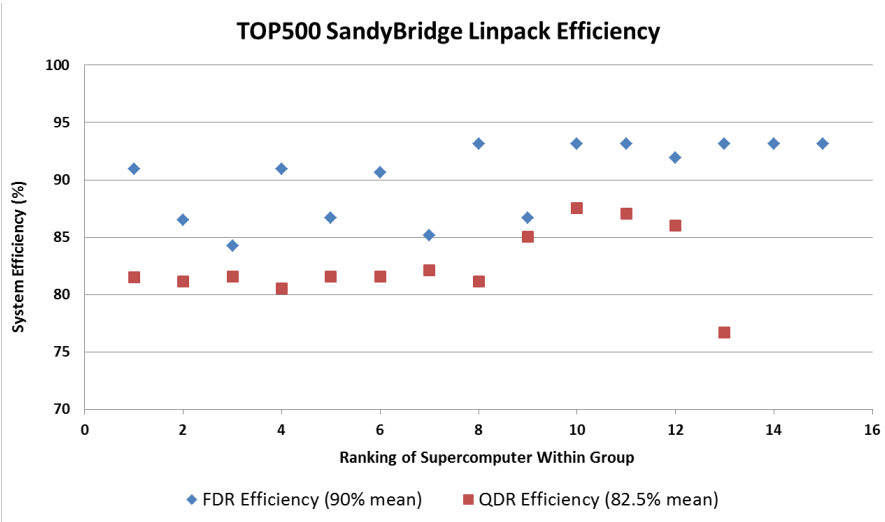
Virtual Protocol Interconnect (VPI) Technology



FDR Application Benchmarks



FDR/QDR InfiniBand Comparisons – Linpack Efficiency



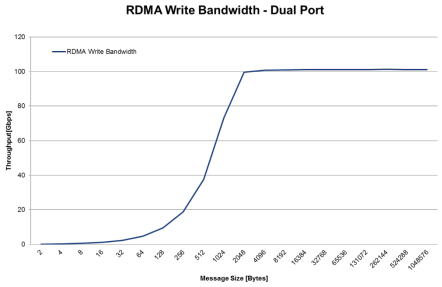
- Derived from 6/12 TOP500 List
- Highest &Lowest Outlier Removed from each group

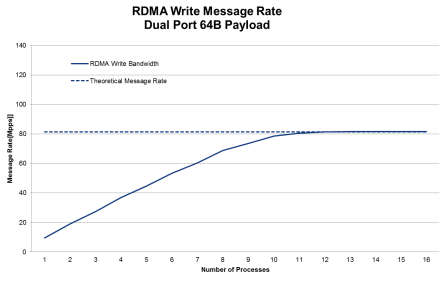
Roadmap of Interconnect Innovations

InfiniHost	InfiniHost III	ConnectX (1,2,3)	Connect-IB
World's first InfiniBand HCA	World's first PCIe InfiniBand HCA	World's first Virtual Protocol Interconnect (VPI) Adapter	Built from the Ground up for the Exascale Foundation
10Gb/s InfiniBand PCI-X host interface 1 million msg/sec	20Gb/s InfiniBand PCIe 1.0 2 million msg/sec	40/56Gb/s InfiniBand PCIe 2.0, 3.0 x8 33 million msg/sec	>100 Gb/s InfiniBand PCIe 3.0 x16 >135 million msg/sec
TOP500-#3	TOP500-#2		
			
2002	2005	2008-11	

Connect-IB Performance Highlights

- World's first 100Gb/s InfiniBand interconnect adapter
  - PCIe 3.0 x16, dual FDR 56Gb/s InfiniBand ports to provide >100Gb/s
- Highest InfiniBand message rate: 130 million messages per second
  - 4X higher than other InfiniBand solutions



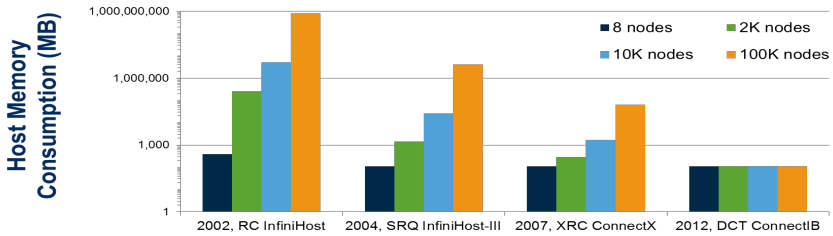


Enter the World of Boundless Performance

Connect-IB Scalability Features

Inter-node Scalability (Scale Out)

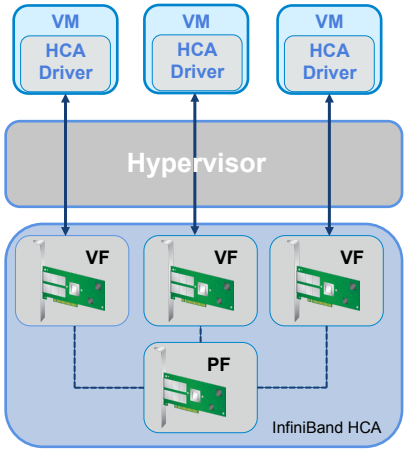
- New innovative transport – Dynamically Connected Transport service
  - The new transport service combines the best of:
    - Unreliable Datagram (UD) – no resources reservation
    - Reliable Connected Service – transport reliability
  - Scale out for unlimited clustering size of compute and storage
  - Eliminates overhead and reduces memory footprint



Enter the World of Unlimited Scalability

Connect-IB Virtualization Enhancements

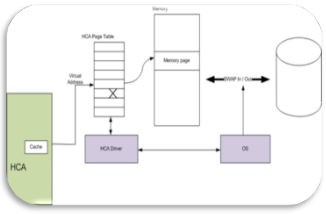
- Scale out virtualization solution
- Full dual-port virtual HCA support for each guest VM
- SRIOV with 256 Virtual Functions (VFs)
  - 2X higher than previous solutions
- 1K egress QoS levels
  - Guaranteed Quality of Service for VMs
- IB virtualization
  - Embedded virtual IB leaf switch
  - LID/GID based forwarding
  - Up to 1K virtual ports

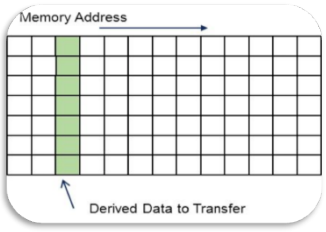


Enter the World of Hardware Virtualization

Connect-IB Memory Management Enhancements

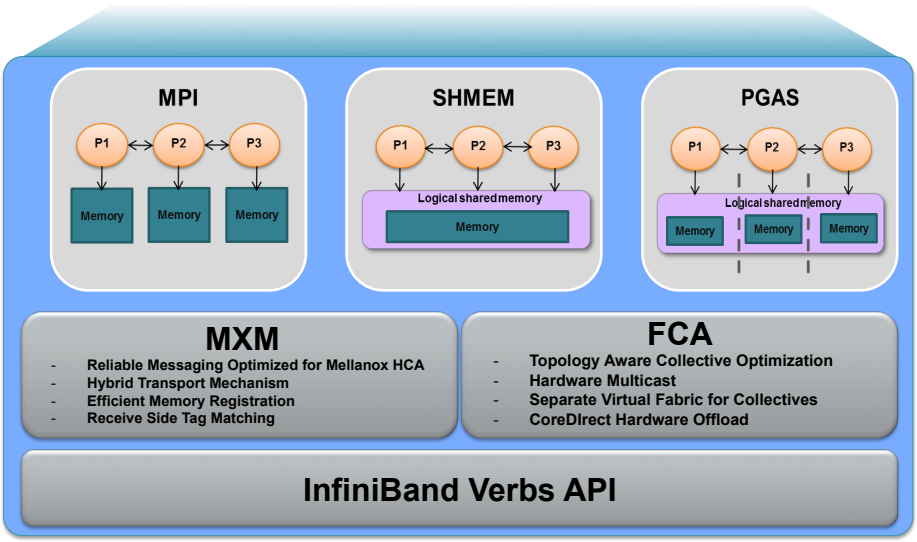
- Extended atomics
  - 4 - 256-byte argument
- On Demand Paging
  - Save pinning of HCA registered memory
- Derived Data Types - Noncontiguous Data Elements
  - Allows noncontiguous data elements to be grouped together in a single message
  - Eliminates unnecessary data copy operations and multiple I/O transactions



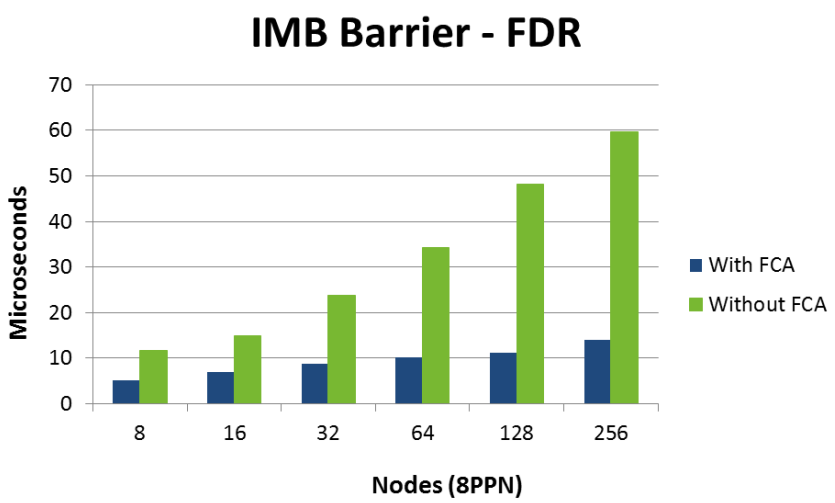


Hardware-Based Sophisticated Memory Operations

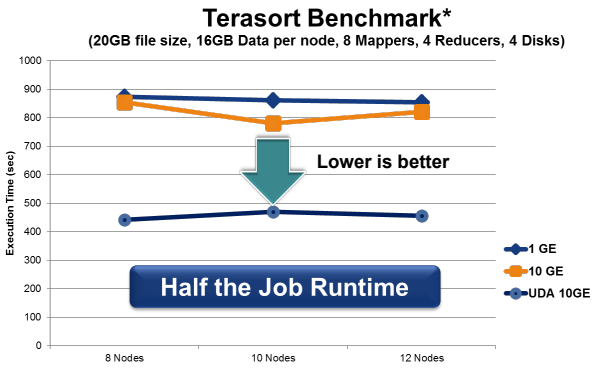
Mellanox ScalableHPC Accelerate Parallel Applications



FCA Collective Performance with OpenMPI



Double Hadoop Performance with UDA



AUBURN UNIVERSITY

Disk Writes 40%

Disk Reads 15%

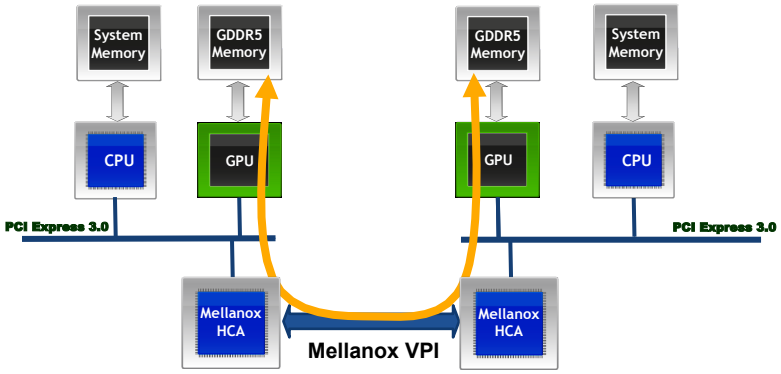
CPU Utilization 2.5X

*TeraSort is a popular benchmark used to measure the performance of Hadoop cluster

~2X Faster Job Completion

GPU Direct RDMA for Fastest GPU to GPU Communications

- GPU Direct RDMA (previously known as GPU Direct 3.0)
- Enables peer to peer communication directly between HCA and GPU
- Dramatically reduces overall latency for GPU to GPU communications



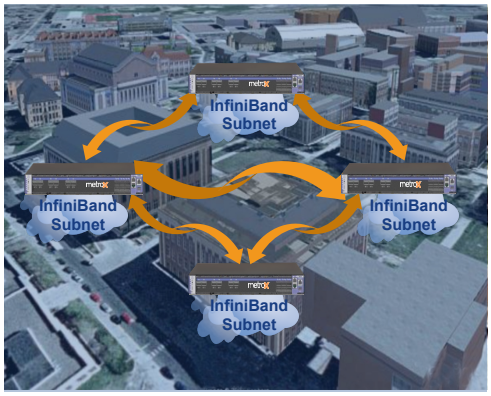
Most Efficient GPU Computing

Optimizing GPU and Accelerator Communications

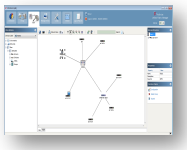
- NVIDIA GPUs
  - Mellanox were original partners in Co-Development of GPUDirect 1.0
  - Recently announced support of GPUDirect RDMA Peer-to-Peer GPU-to-HCA data path
- AMD GPUs
  - Sharing of System Memory: AMD DirectGMA Pinned supported today
  - AMD DirectGMA P2P: Peer-to-Peer GPU-to-HCA data path under development
- Intel MIC
  - MIC software development system enables the MIC to communicate directly over the InfiniBand verbs API to Mellanox devices

MetroX™– Bringing InfiniBand to Campus Wide Networks

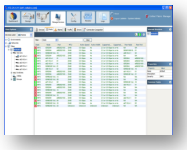
- Extends InfiniBand across campus/metro
- Low-cost, low-power
- 40Gb FDR-10 InfiniBand links
- RDMA over distant sites
- Support up to 10KM over dark fiber
- QSFP to SMF LC connectors module



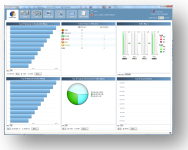
UFM – Comprehensive, Robust Management



Automatic Discovery



Central Device Management

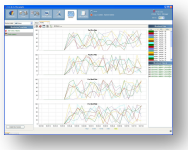


Fabric Dashboard

Congestion Analysis

Fabric Health Reports

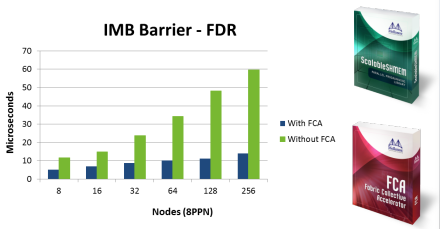
Service Oriented Provisioning



Health & Perf Monitoring

Mellanox HPC – Paving the way to Exascale Computing

ScalableHPC



Ultimate Scalability with Connect-IB

- 100Gb/s throughput to network
- Over 130-million messages/second
- Dynamically Connected Transport service for unlimited inter-node scaling

ConnectIB™

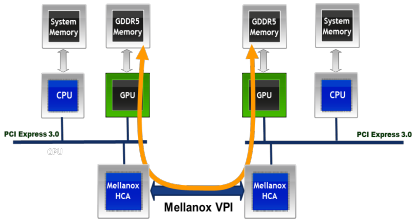
Highest Performing Interconnect

SwitchX™

ConnectX®-3

- <0.7usec latency
- 56Gb/s throughput
- Higher scalability
- Maximum Reliability

Accelerating GPU Communications



Thank You

HPC@mellanox.com

PAVING THE ROAD TO EXASCALE

ADVANCING NETWORK PERFORMANCE, EFFICIENCY, AND SCALABILITY.