Advancing Applications Performance With InfiniBand

Pak Lui, Application Performance Manager

September 12, 2013

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Mellanox Overview

Ticker: MLNX

- Leading provider of high-throughput, low-latency server and storage interconnect
  - FDR 56Gb/s InfiniBand and 10/40/56GbE
  - Reduces application wait-time for data
  - Dramatically increases ROI on data center infrastructure

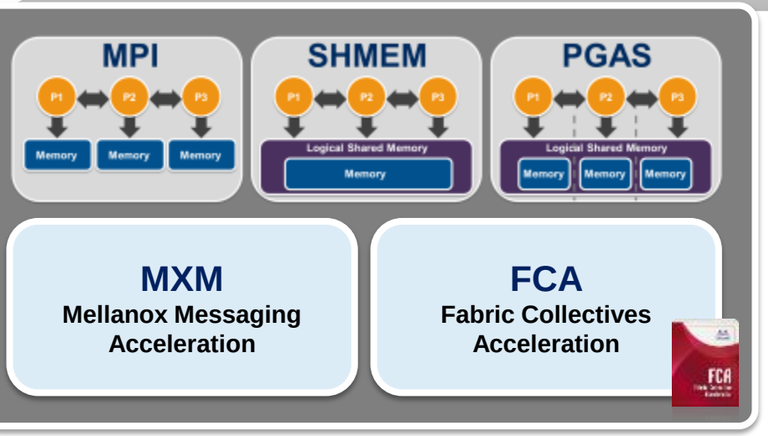
- Company headquarters:
  - Yokneam, Israel; Sunnyvale, California
  - ~1,200 employees* worldwide

- Solid financial position
  - Record revenue in FY12; $500.8M, up 93% year-over-year
  - Q2'13 revenue of $98.2M
  - Q3'13 guidance ~$104M to $109M
  - Cash + investments @ 6/30/13 = $411.3M

* As of June 2013

Providing End-to-End Interconnect Solutions

Comprehensive End-to-End Software Accelerators and Managment



MXM
Mellanox Messaging Acceleration

FCA
Fabric Collectives Acceleration

Management

UFM
Unified Fabric Management

Storage and Data

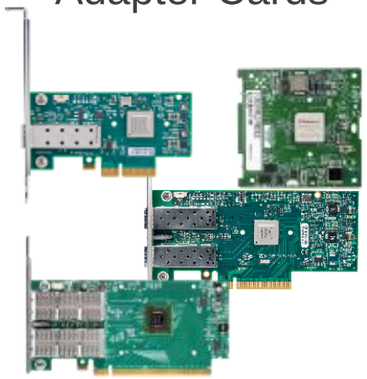
VSA
Storage Accelerator (iSCSI)

UDA
Unstructured Data Accelerator

Comprehensive End-to-End InfiniBand and Ethernet Solutions Portfolio

ICs

Adapter Cards



Switches/Gateways

Long-Haul Systems

Cables/Modules

Virtual Protocol Interconnect (VPI) Technology

ConnectX®-3 VPI Adapter

redhat. Novell SUSE Microsoft vmware OPENFABRICS SOFTWARE

Applications

Networking | **Storage** | **Clustering** | **Management**

Acceleration Engines

PCI EXPRESS™ 3.0

- Ethernet: 10/40/56 Gb/s
- InfiniBand:10/20/40/56 Gb/s


LOM

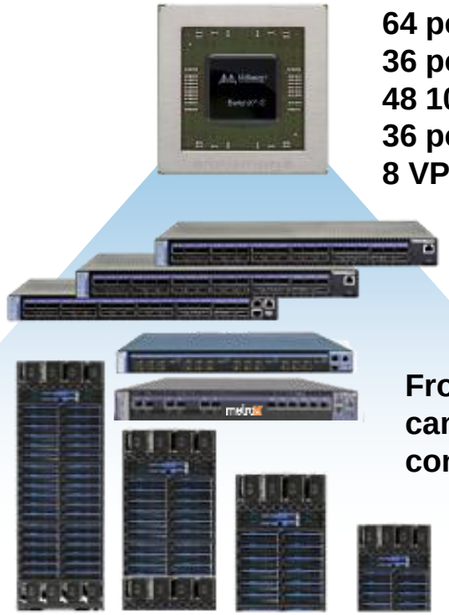

Adapter Card


Mezzanine Card

SwitchX®-2 VPI Switch

Unified Fabric Manager

Switch OS Layer

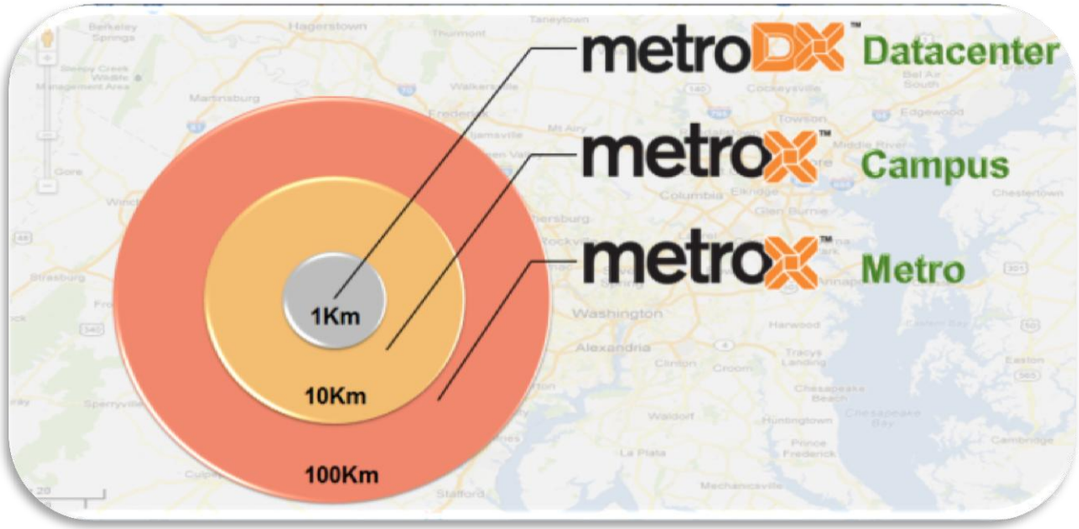


- 64 ports 10GbE
- 36 ports 40/56GbE
- 48 10GbE + 12 40/56GbE
- 36 ports IB up to 56Gb/s
- 8 VPI subnets

From data center to campus and metro connectivity

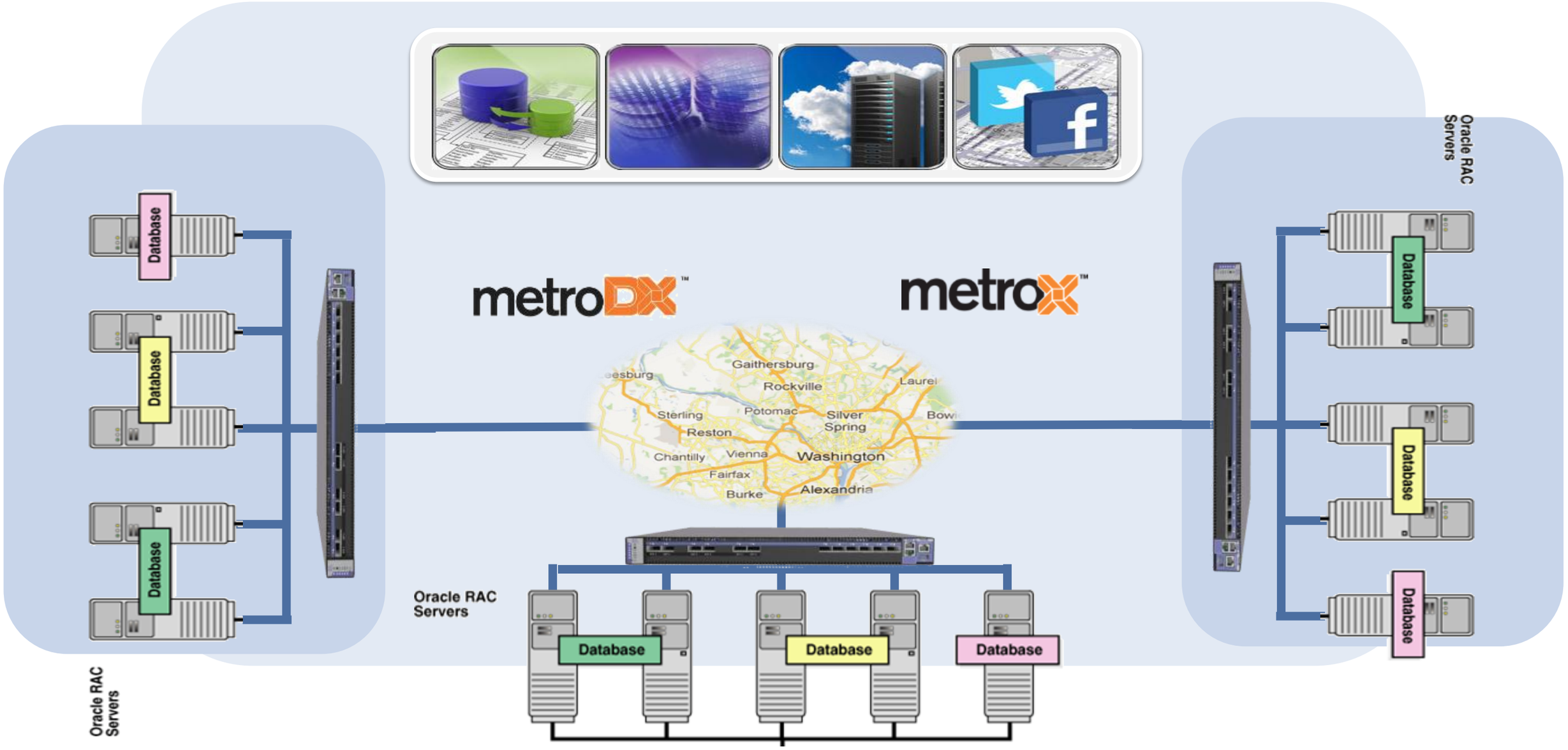
MetroDX™ and MetroX™

- MetroX™ and MetroDX™ extends InfiniBand and Ethernet RDMA reach
- Fastest interconnect over 40Gb/s InfiniBand or Ethernet links
- Supporting multiple distances
- Simple management to control distant sites
- Low-cost, low-power , long-haul solution

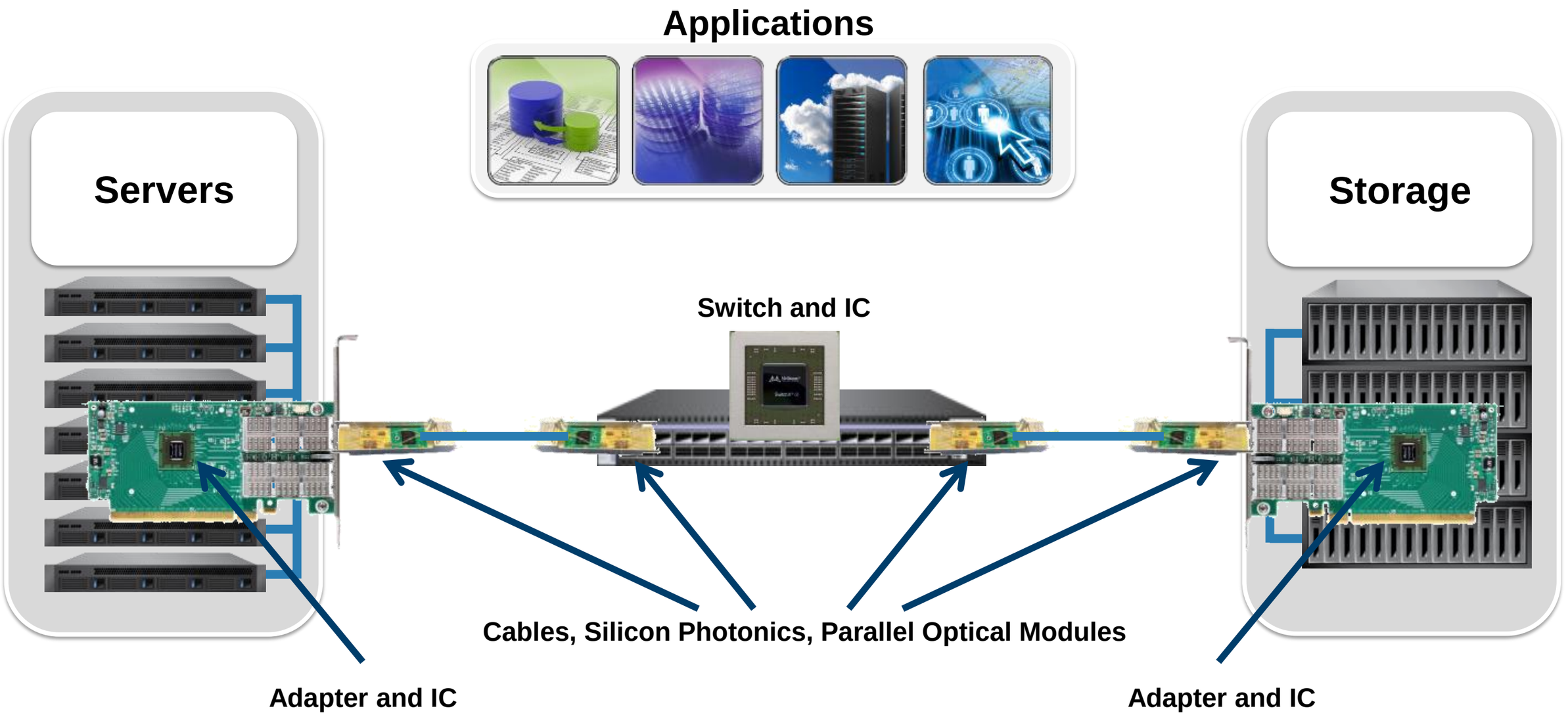


40Gb/s over Campus and Metro

Data Center Expansion Example – Disaster Recovery

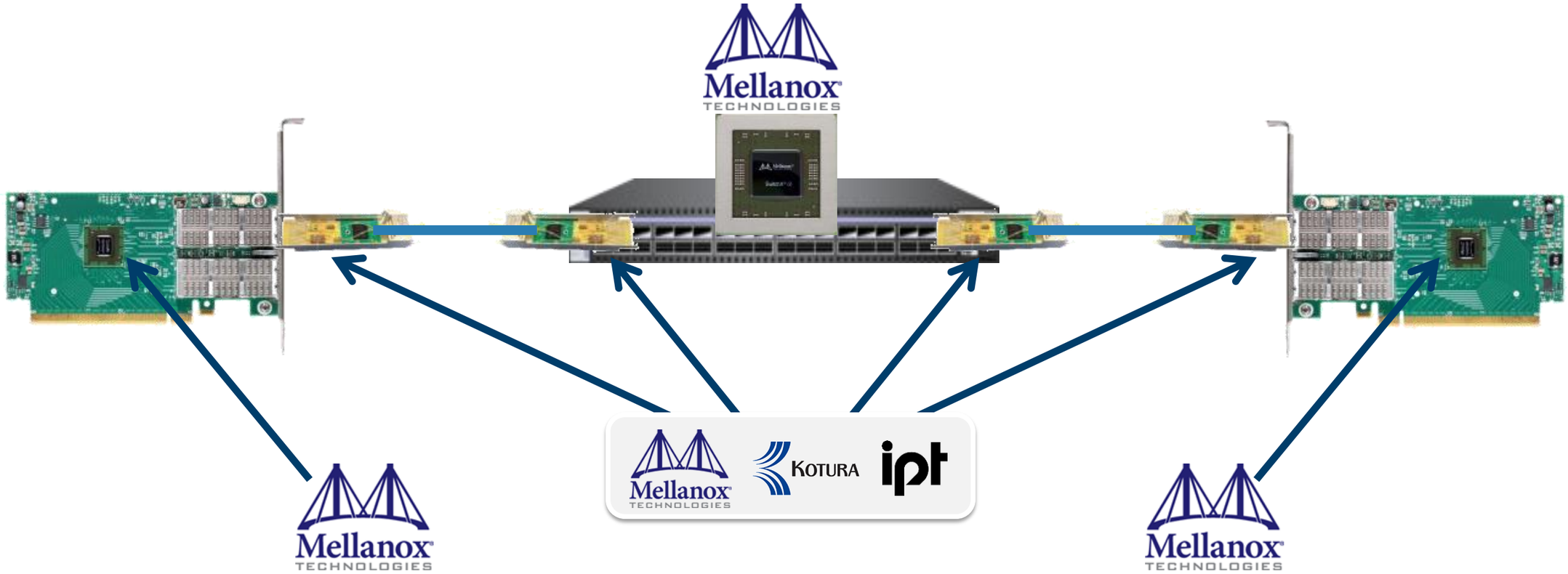


Key Elements in a Data Center Interconnect



IPtronics and Kotura Complete Mellanox's 100Gb/s+ Technology

Recent Acquisitions of Kotura and IPtronics Enable Mellanox to Deliver Complete High-Speed Optical Interconnect Solutions for 100Gb/s and Beyond



Mellanox InfiniBand Paves the Road to Exascale

Meteo France

Centre National de Calcul
Météopole, Toulouse

522 TFlops peak performance
56 racks **bullx DLC**
1008 nodes
Fat Tree InfiniBand FDR
Lustre **2 Po, 69 GB/s**
Disks storage 209 TB

Computer C1 (03/2013)

2,85 PFlops peak performance
56+45 racks **bullx DLC**
1800 nodes
Fat Tree InfiniBand FDR
Lustre **3,57 Po, 138 GB/s**
Disks storage 400 TB

Computer C3 (08/2015)

Research

2013 2014 **2015** 2016

Operational

522 TFlops peak performance
56 racks **bullx DLC**
1008 nodes
Fat Tree InfiniBand FDR
Lustre **1,53 Po, 46 GB/s**
Disks storage 135 TB

Computer C2 (11/2013)

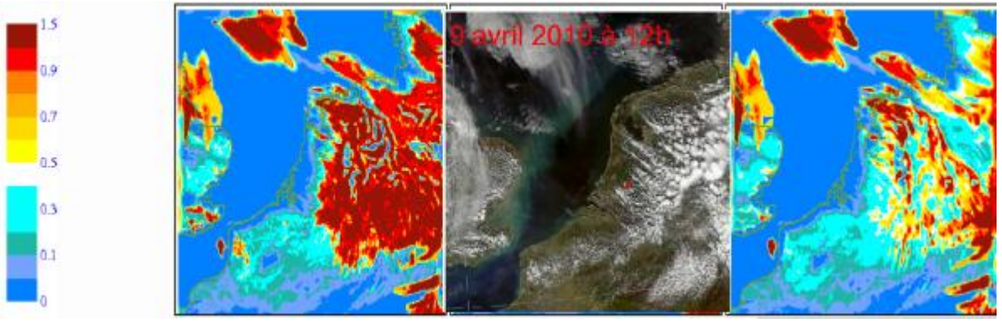
Espace Clément Ader
Montaudran

2,85 PFlops peak performance
56+45 racks **bullx DLC**
1800 nodes
Fat Tree InfiniBand FDR
Lustre **2,55 Po, 92 GB/s**
Disks storage 135 TB

Computer C4 (04/2016)

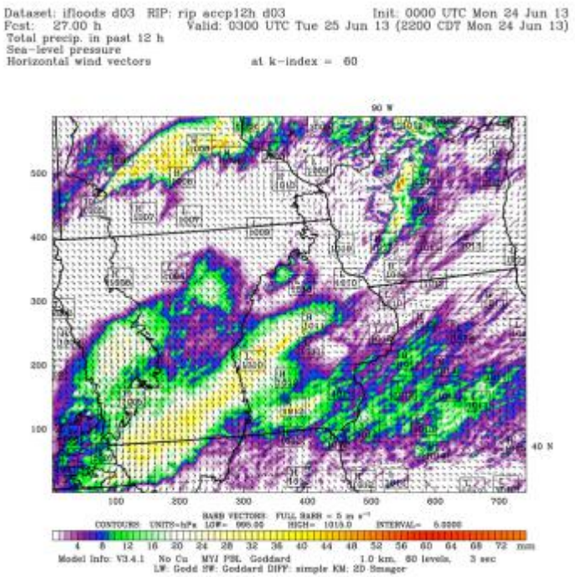
METEO FRANCE
Toujours un temps d'avance

BULL

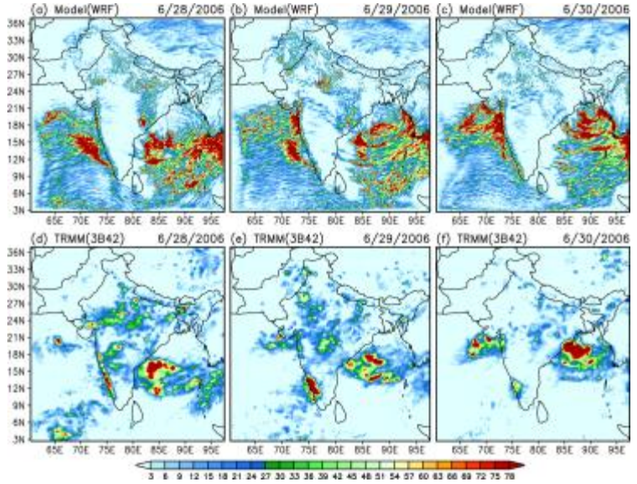


NASA Ames Research Center Pleiades

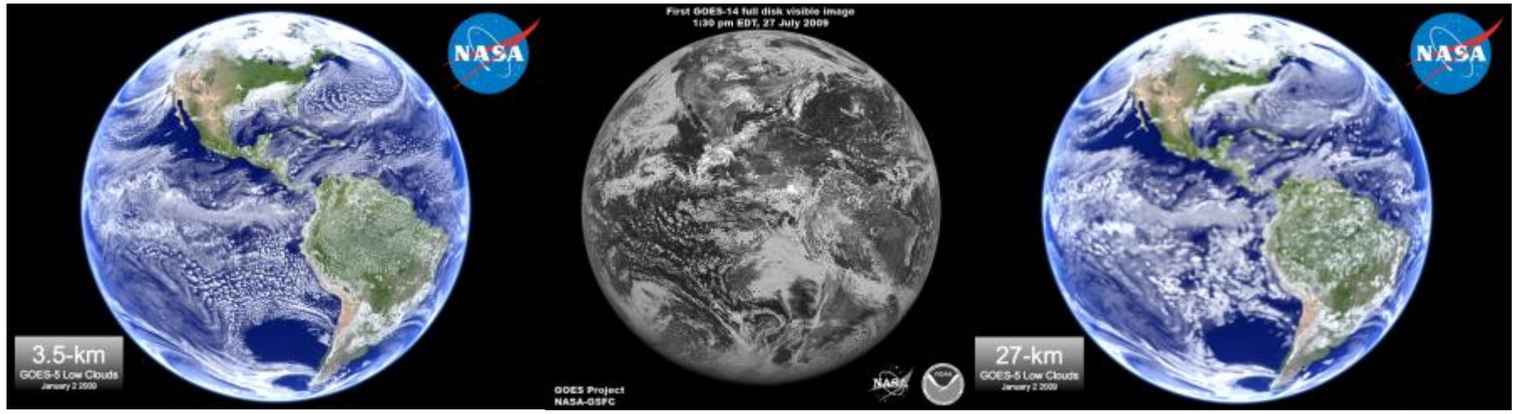
- 20K InfiniBand nodes
- Mellanox end-to-end FDR and QDR InfiniBand
- Supports variety of scientific and engineering projects
  - Coupled atmosphere-ocean models
  - Future space vehicle design
  - Large-scale dark matter halos and galaxy evolution



Asian Monsoon Water Cycle

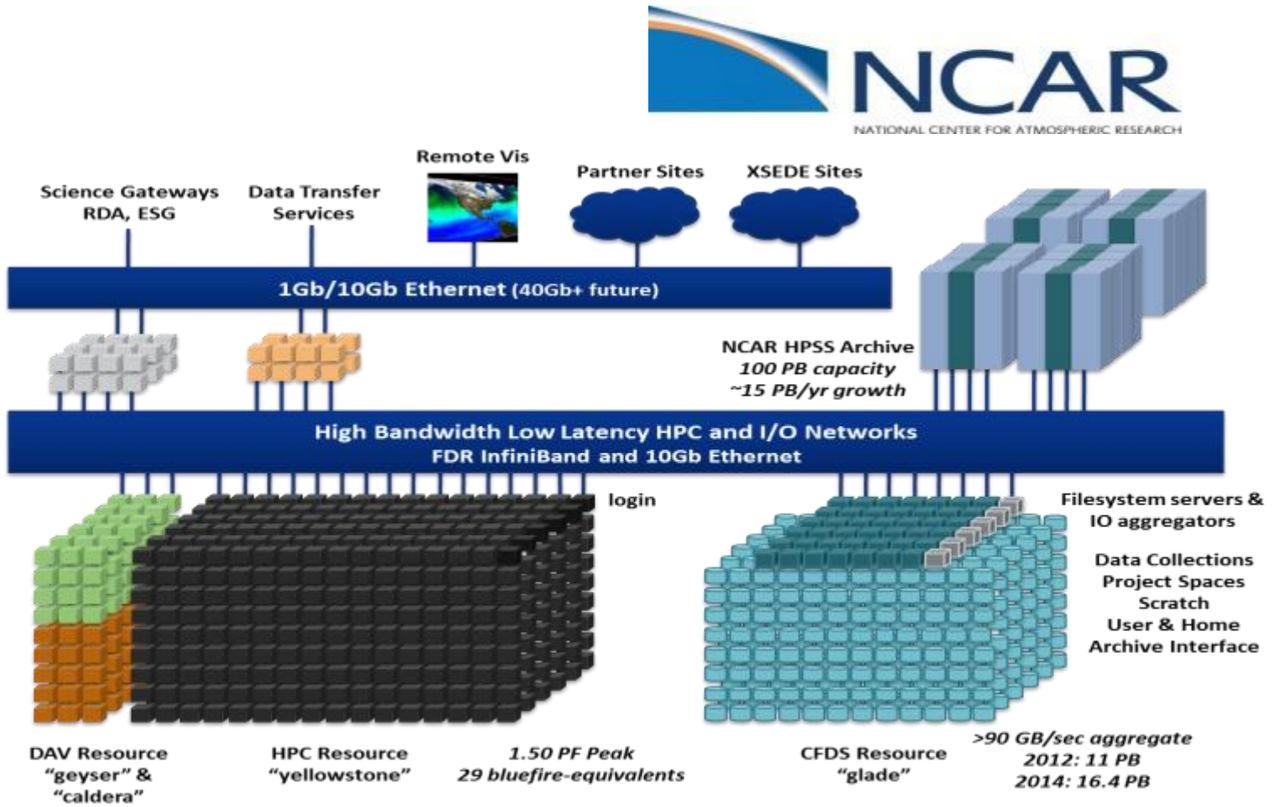


High-Resolution Climate Simulations



NCAR (National Center for Atmospheric Research)

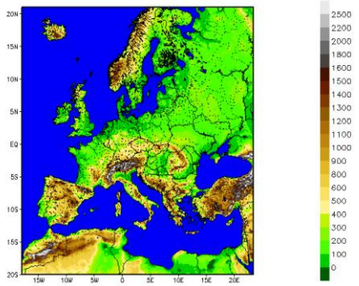
- “Yellowstone” system
- 72,288 processor cores, 4,518 nodes
- Mellanox end-to-end FDR InfiniBand, CLOS (full fat tree) network, single plane

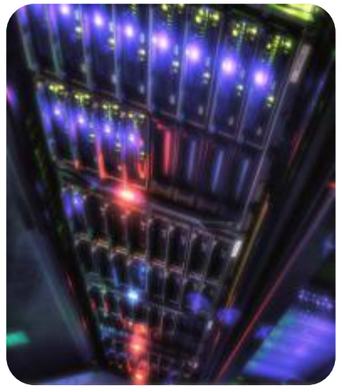


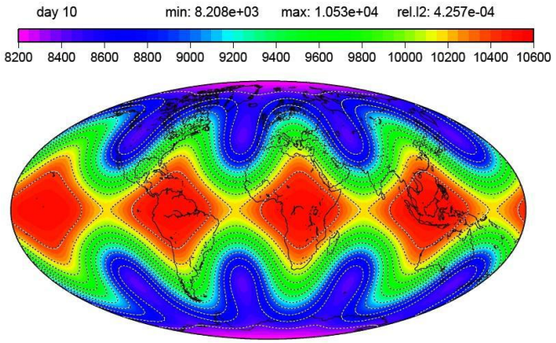
Applications Performance (Courtesy of the HPC Advisory Council)

CONSORTIUM FOR SMALL SCALE MODELING

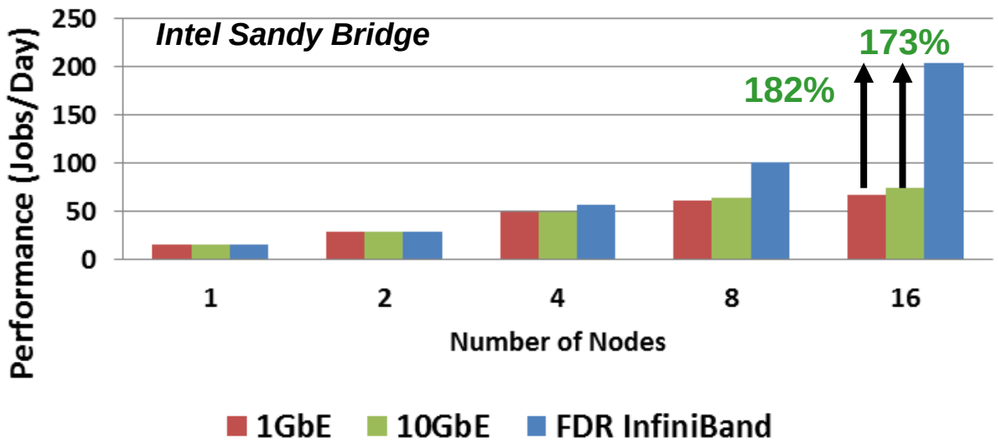
COSMO



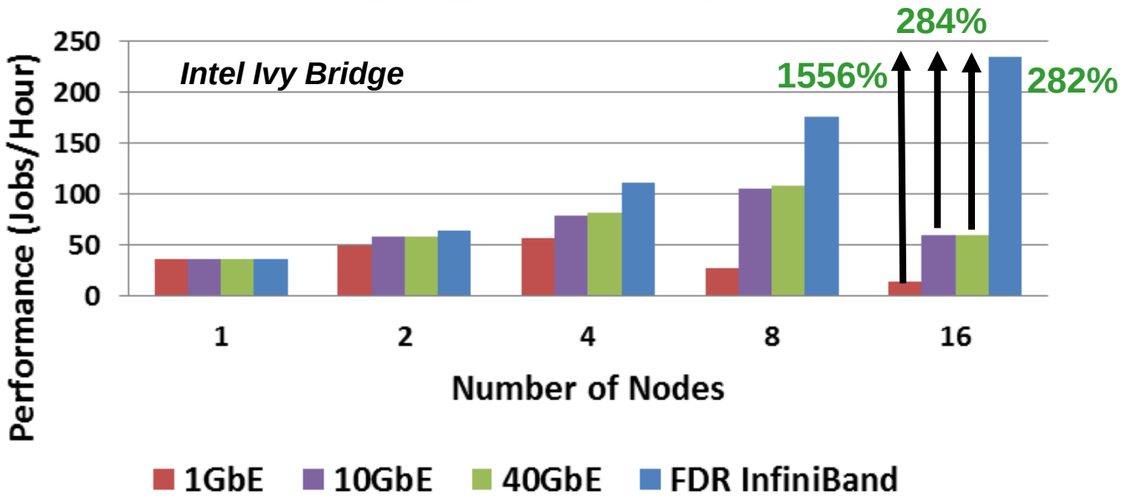




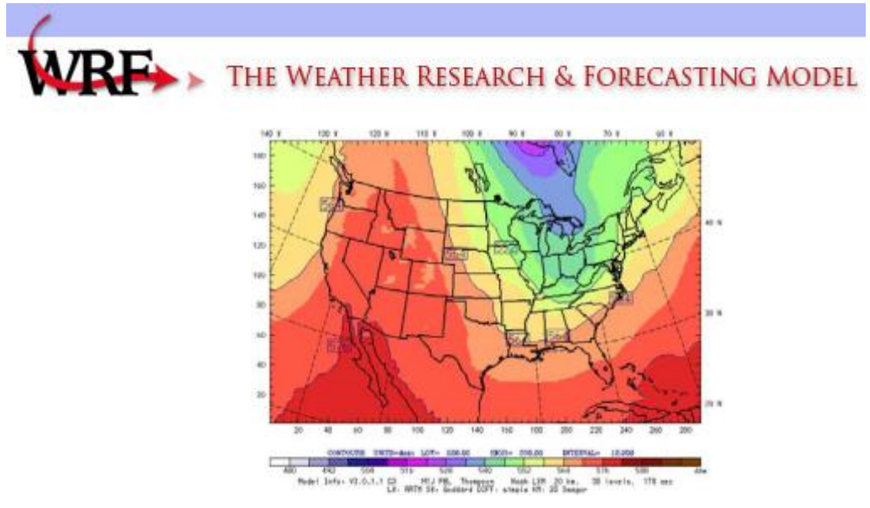
COSMO RAPS 5.1 Performance (COSMO_EU)

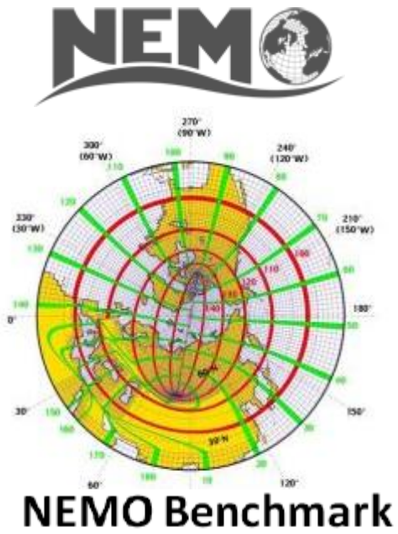


ICON RAPS 2.0 Performance (test_hat_jww, Ivy Bridge)



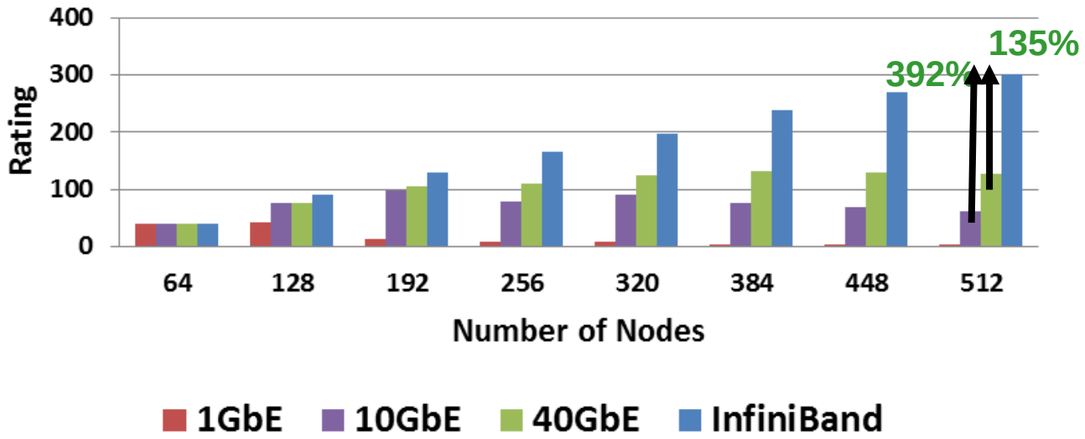
Applications Performance (Courtesy of the HPC Advisory Council)



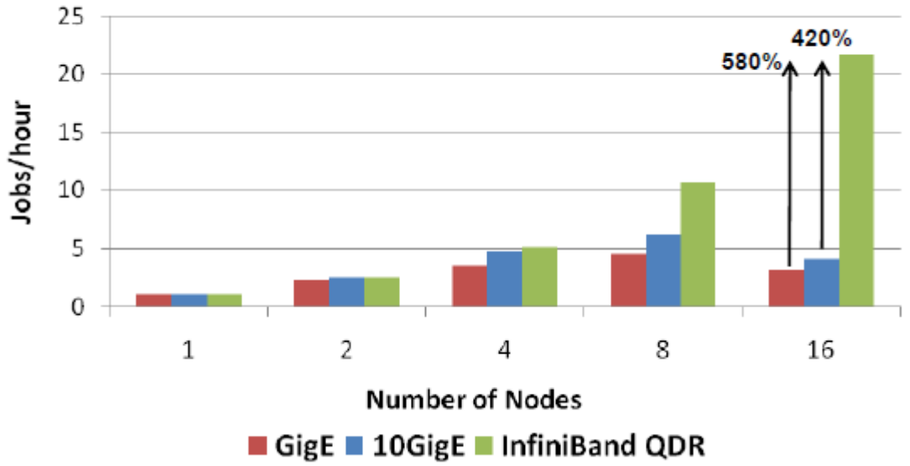


WRF Benchmark

(conus12km)



NEMO Benchmark



Dominant in Enterprise Back-End Storage Interconnects

Microsoft

SMB Direct

NetApp

xyratex

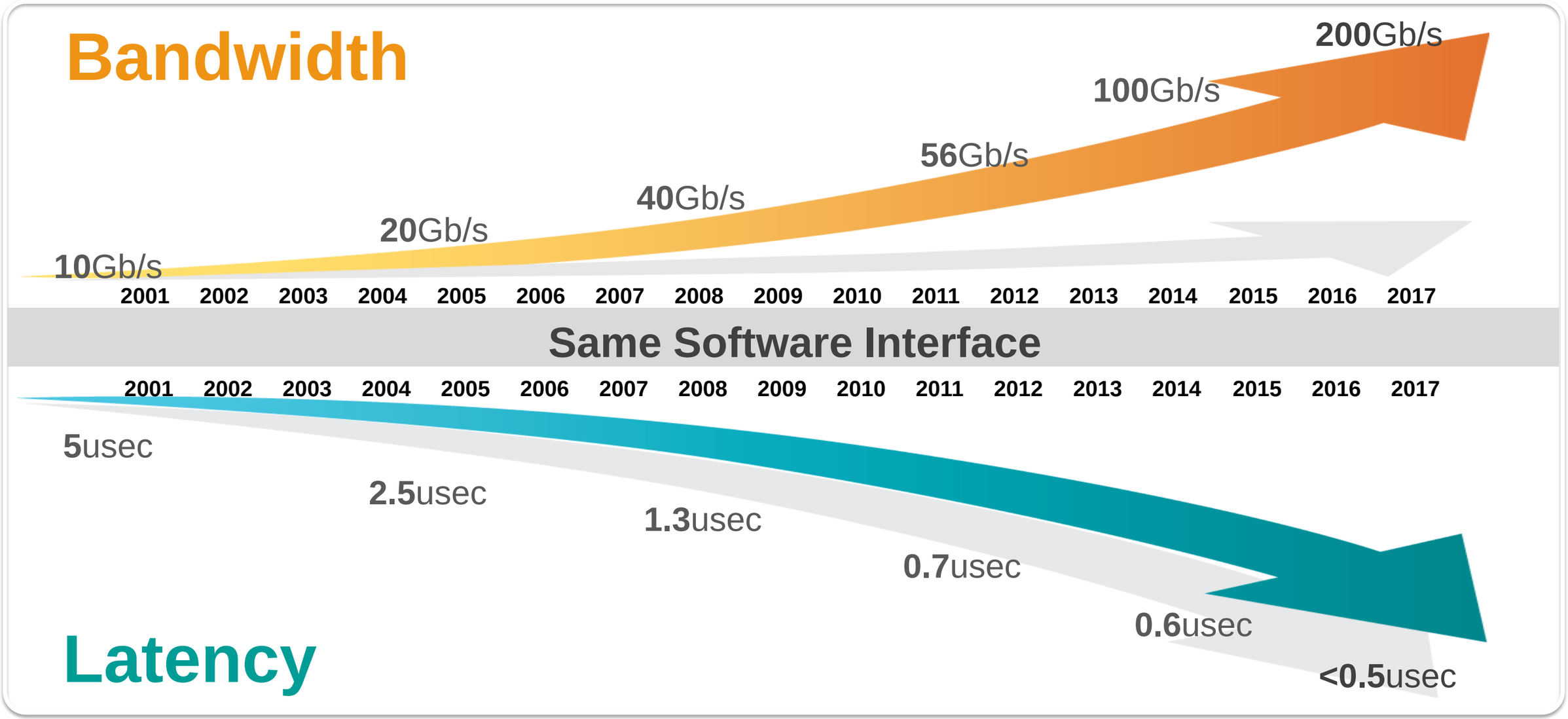
EMC²

DataDirect NETWORKS

IBM xiv

ORACLE

Leading Interconnect, Leading Performance



Connect-IB

Architectural Foundation for Exascale Computing

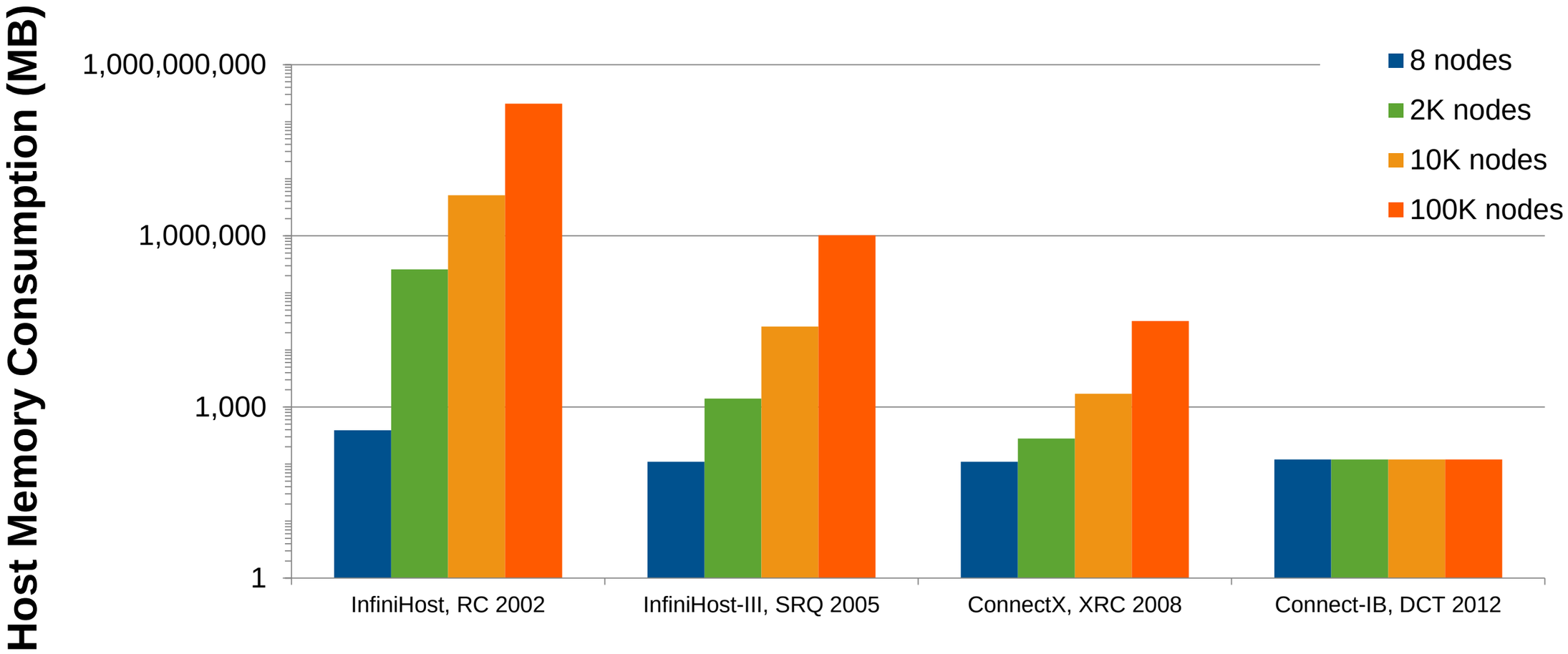
Connect-IB : The Exascale Foundation

- World's first 100Gb/s interconnect adapter
  - PCIe 3.0 x16, dual FDR 56Gb/s InfiniBand ports to provide >100Gb/s
- Highest InfiniBand message rate: 137 million messages per second
  - 4X higher than other InfiniBand solutions
- <0.7 micro-second application latency
- Supports GPUDirect RDMA for direct GPU-to-GPU communication
- Unmatchable Storage Performance
  - 8,000,000 IOPs (1QP), 18,500,000 IOPs (32 QPs)
- New Innovative Transport – Dynamically Connected Transport Service
- Supports Scalable HPC with MPI, SHMEM and PGAS/UPC offloads

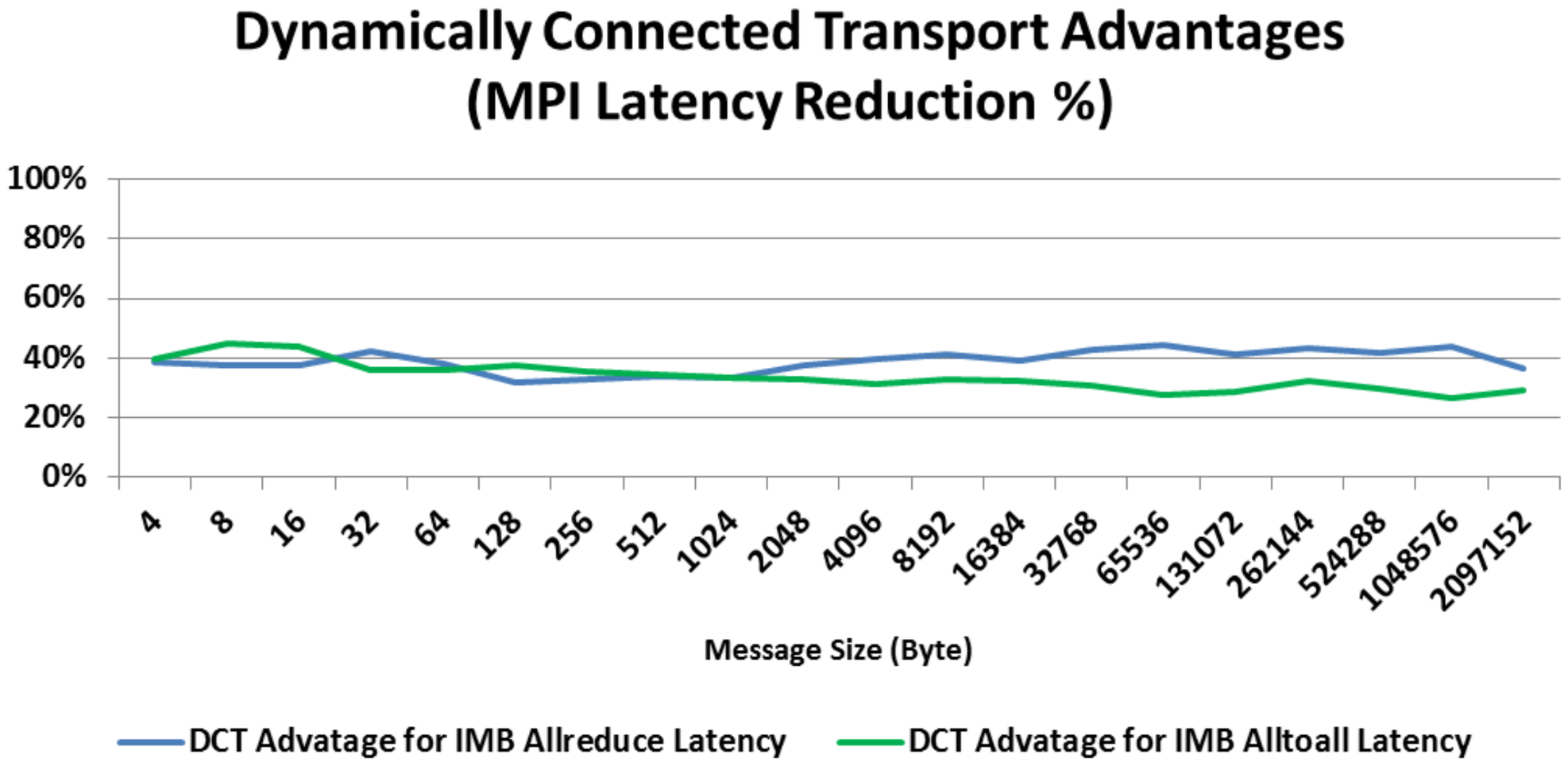
ConnectIB™

Enter the World of Boundless Performance

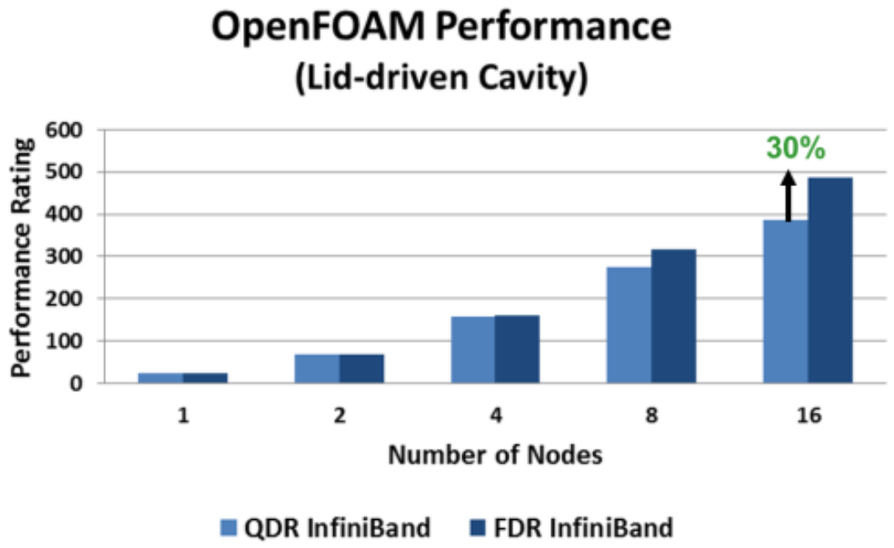
Connect-IB Memory Scalability

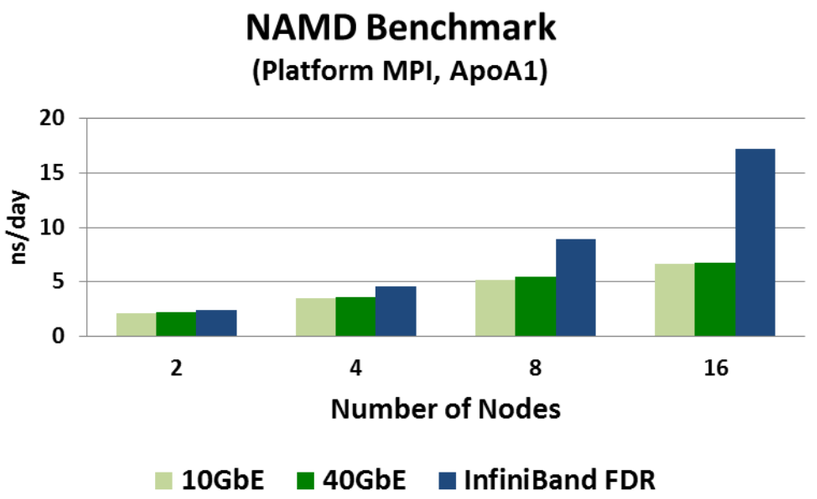


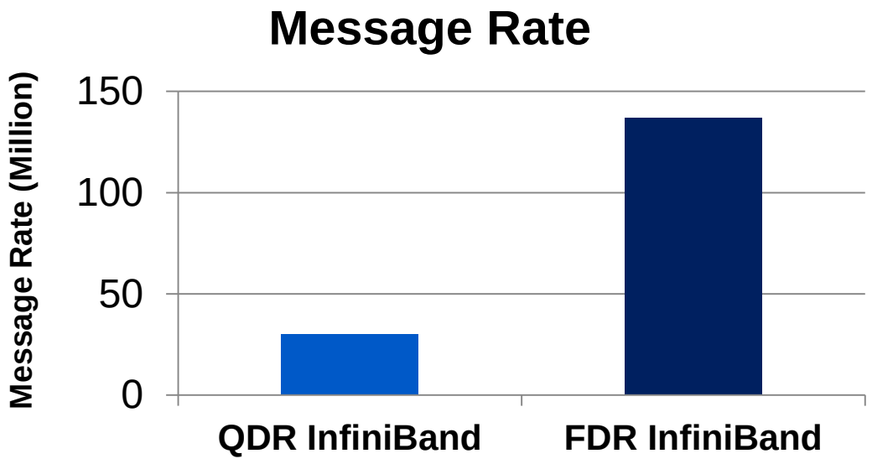
Dynamically Connected Transport Advantages

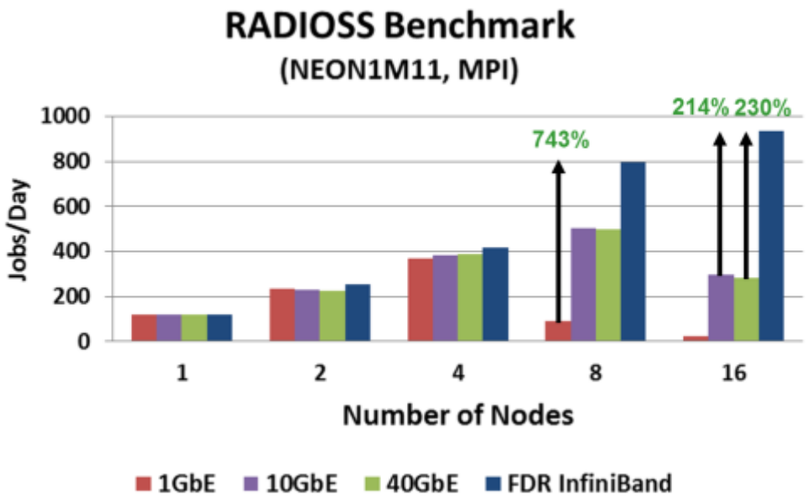


FDR InfiniBand Delivers Highest Application Performance





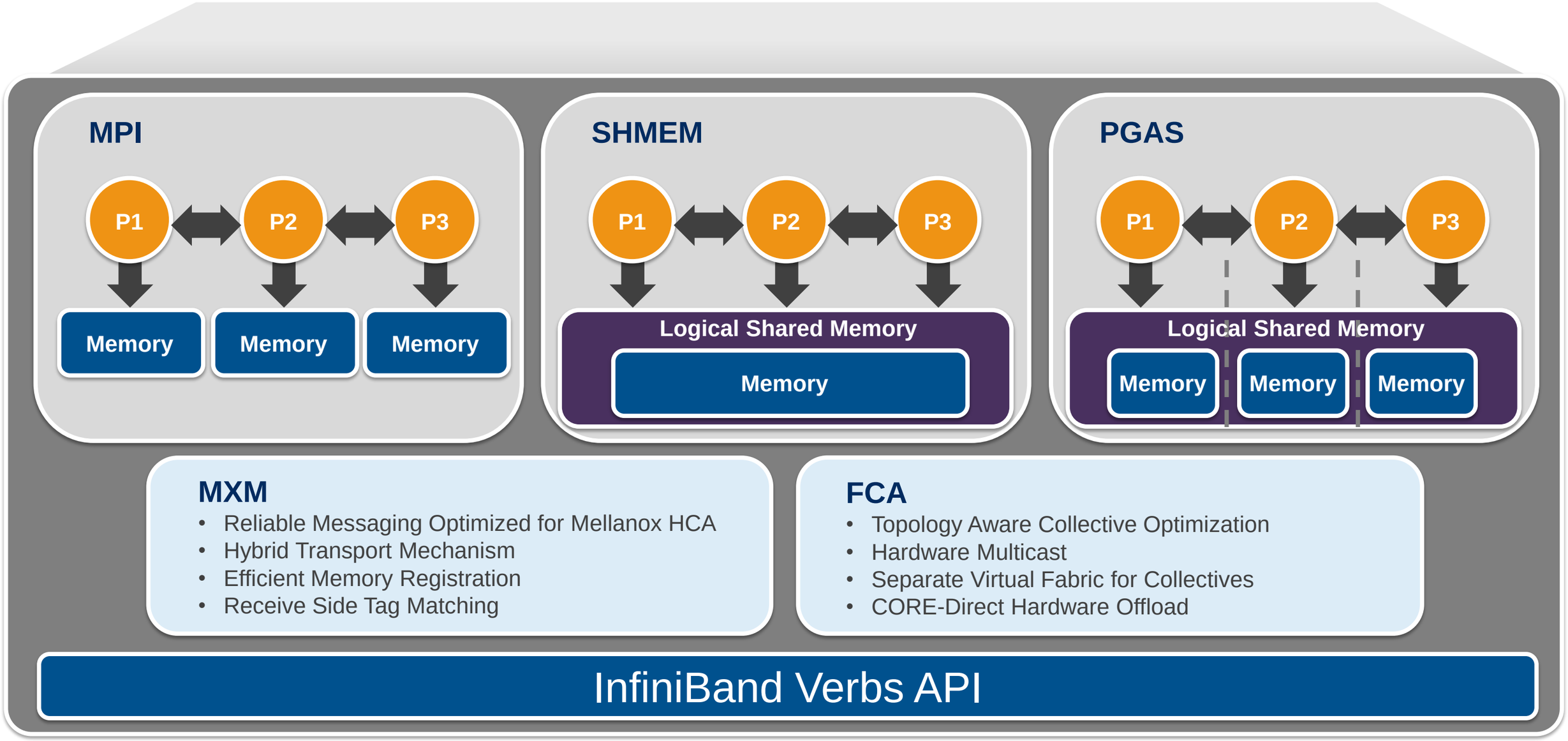




Scalable Communication

MXM, FCA

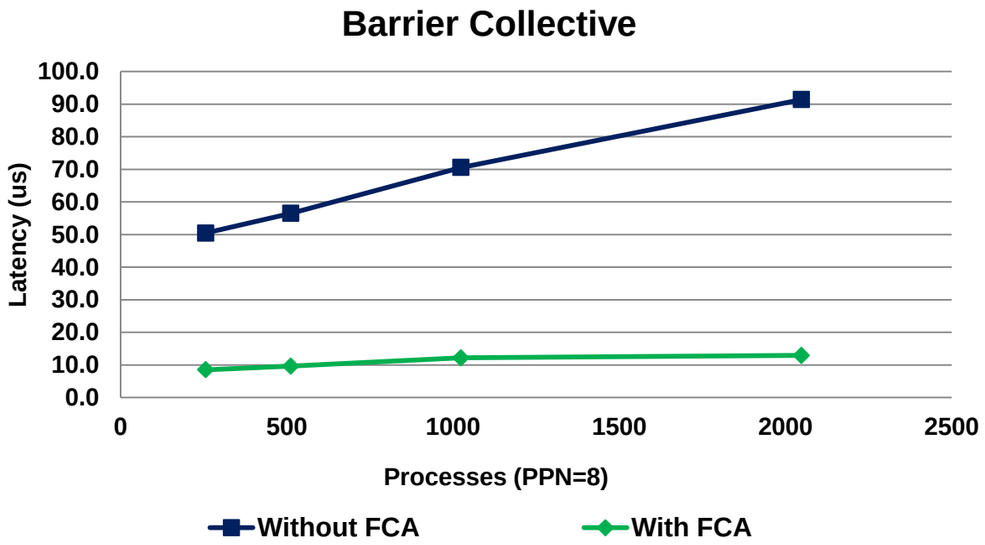
Mellanox ScalableHPC Accelerate Parallel Applications

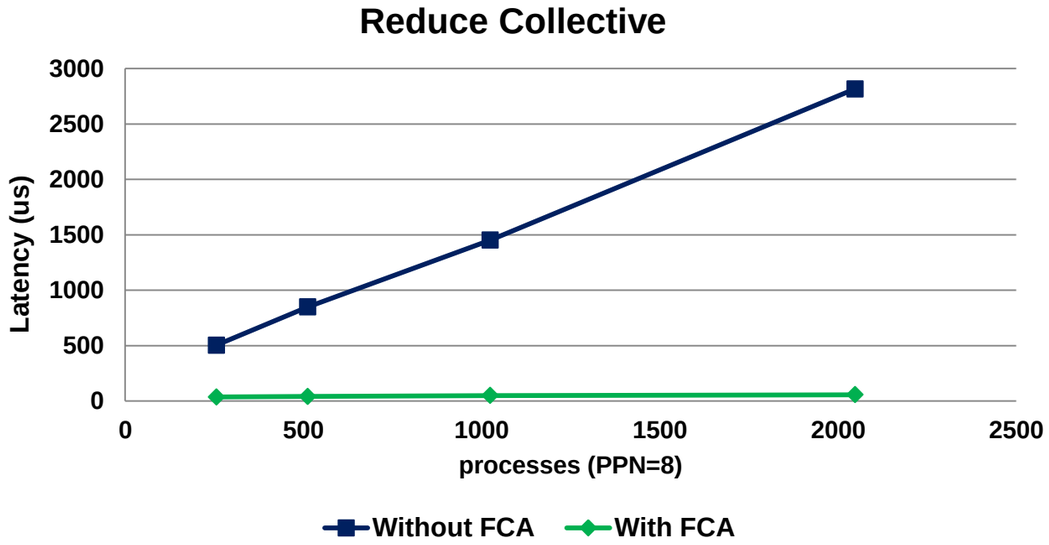


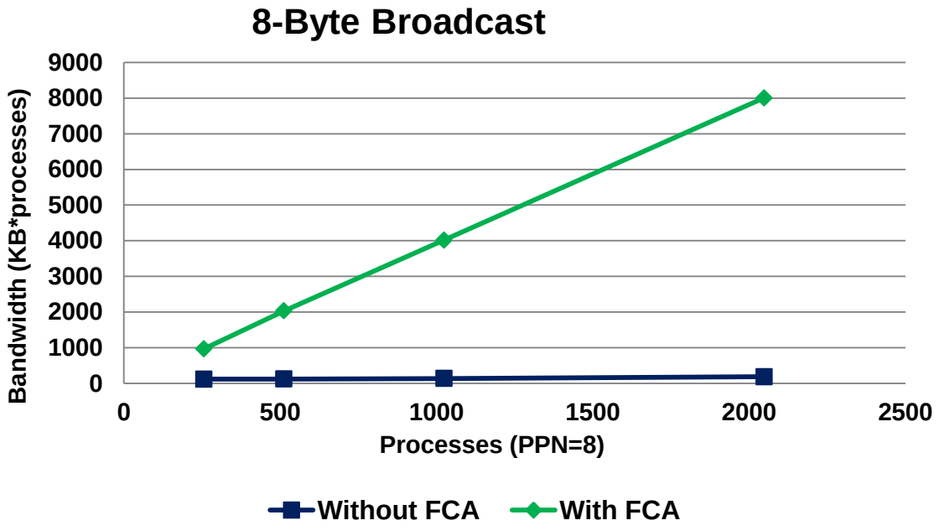
MXM v2.0 - Highlights

- Transport library integrated with OpenMPI, OpenSHMEM, BUPC, Mvapich2
  - More solutions will be added in the future
  - Utilizing Mellanox offload engines
- Supported APIs (both sync/async): AM, p2p, atomics, synchronization
- Supported transports: RC, UD, DC, RoCE, SHMEM
- Supported built-in mechanisms: tag matching, progress thread, memory registration cache, fast path send for small messages, zero copy, flow control
- Supported data transfer protocols: Eager Send/Recv, Eager RDMA, Rendezvous

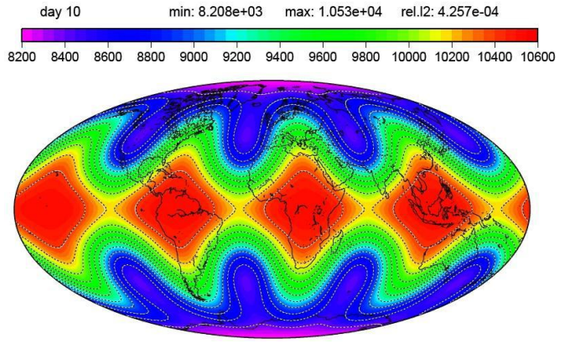
Mellanox FCA Collective Scalability

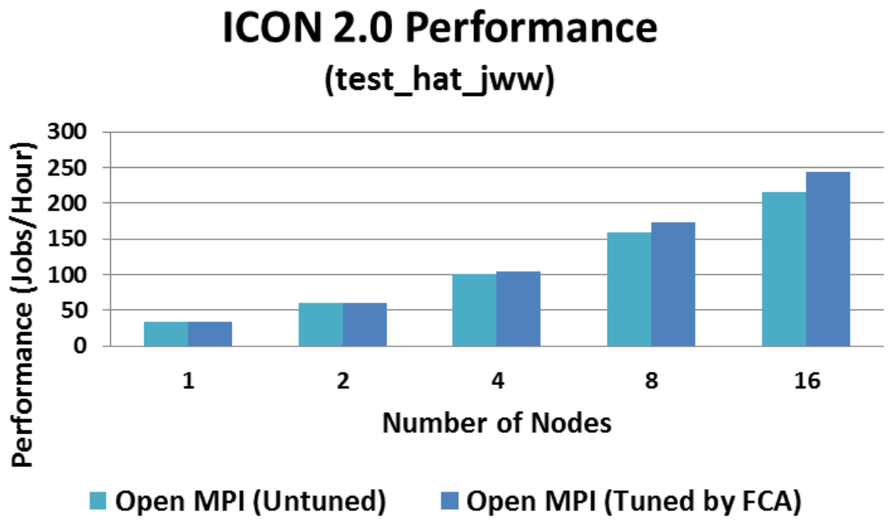




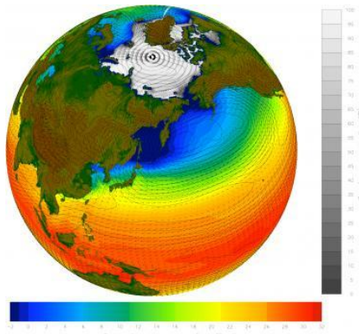


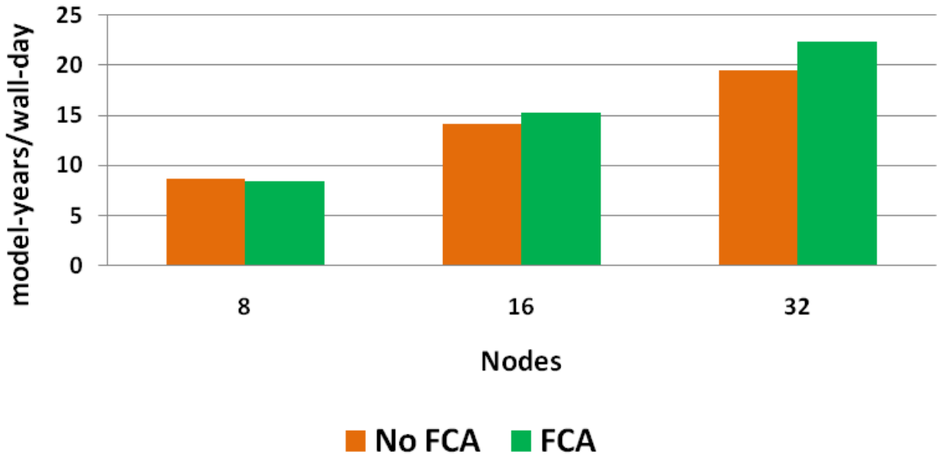
FDR InfiniBand Delivers Highest Application Performance





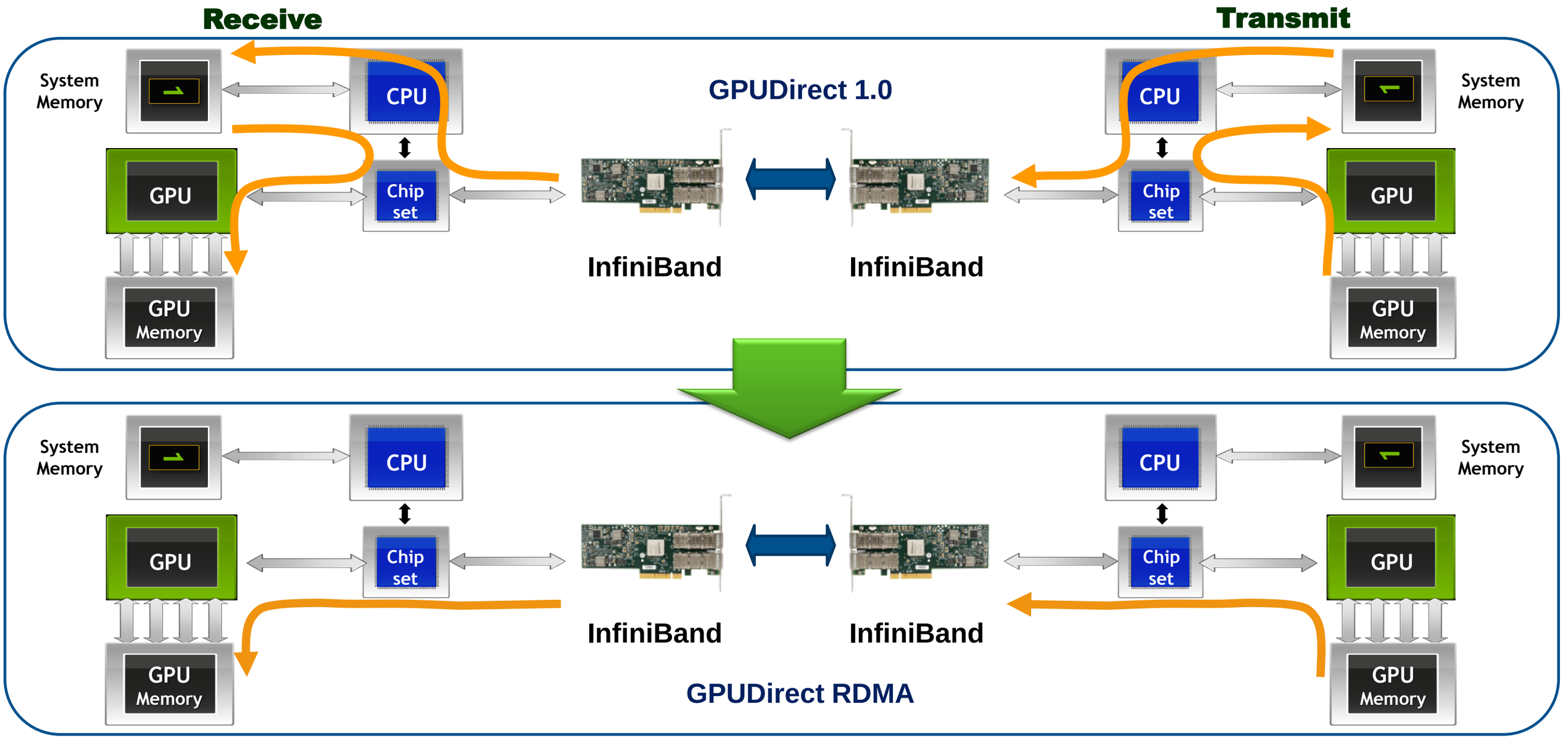
CESM
COMMUNITY EARTH SYSTEM MODEL





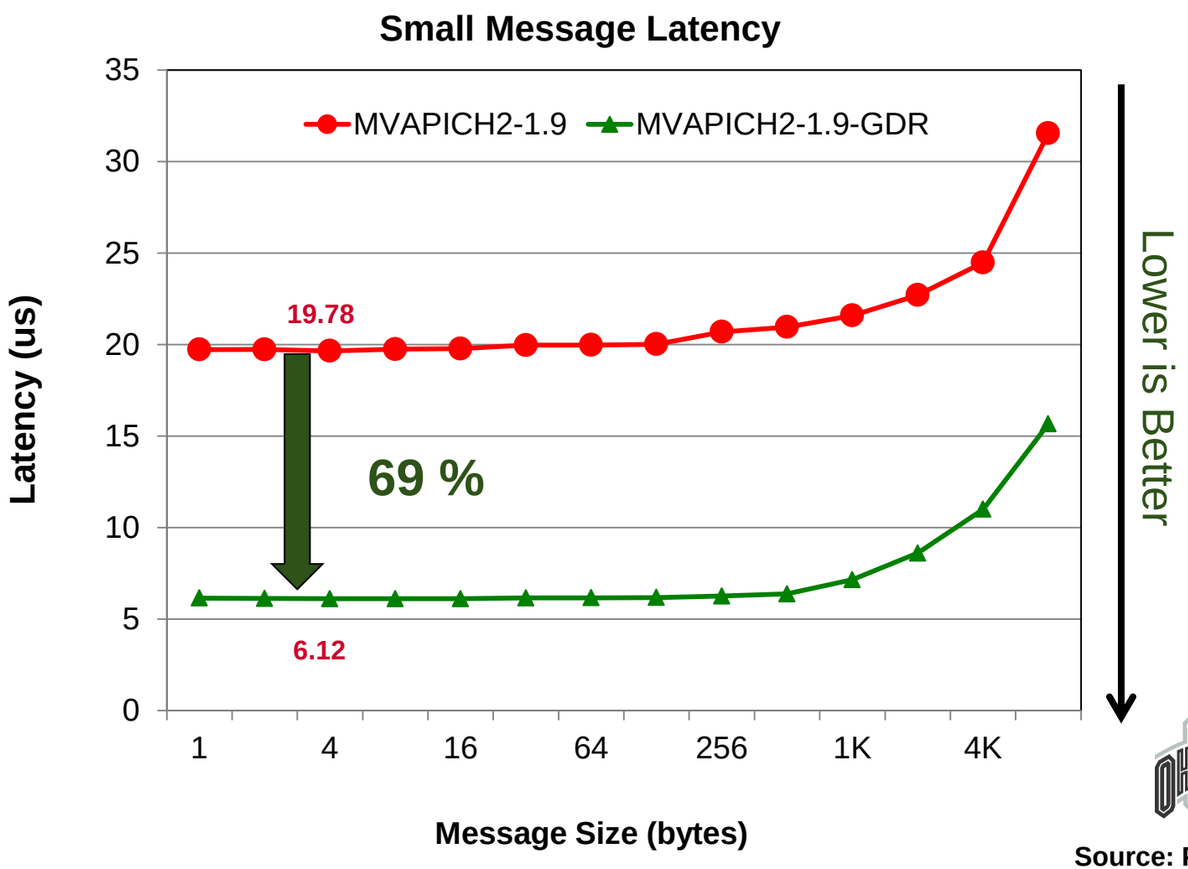
GPU Direct

GPUDirect RDMA

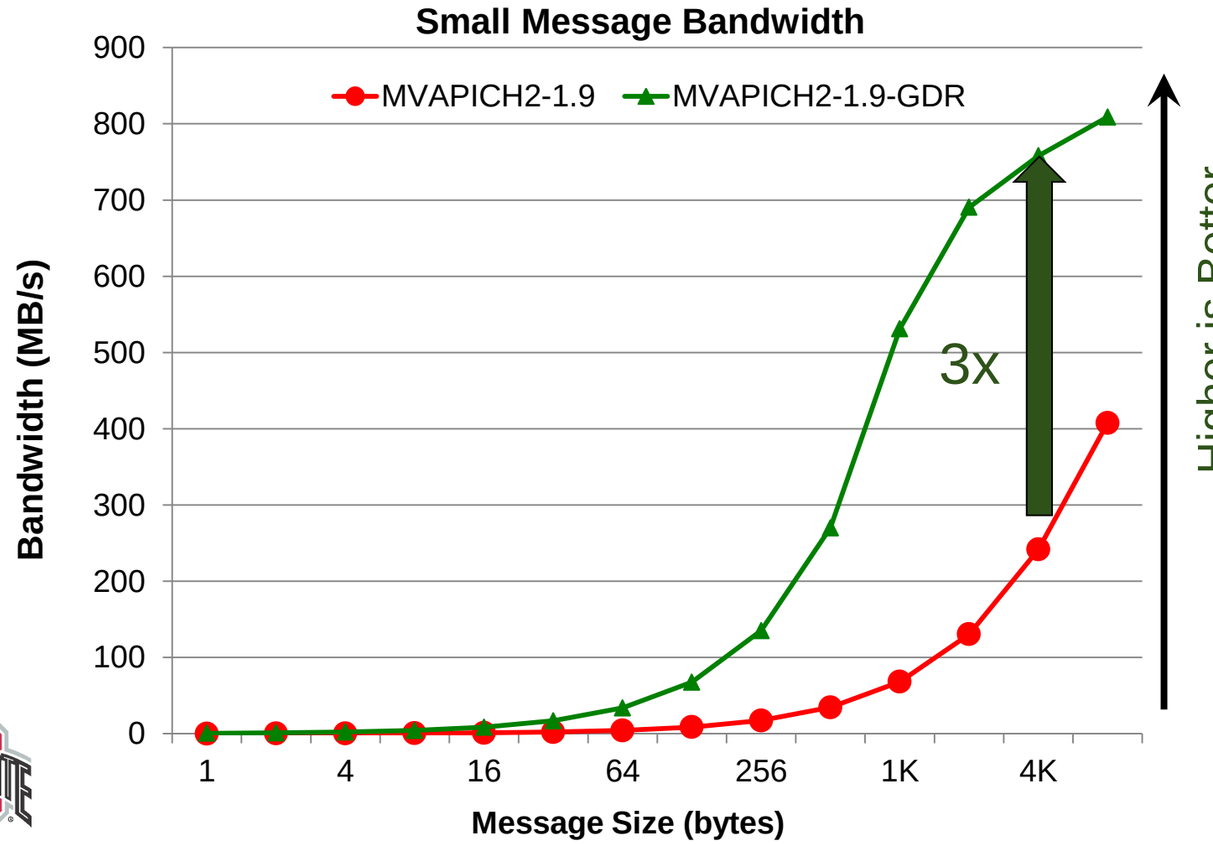


Preliminary Performance of MVAPICH2 with GPUDirect RDMA

GPU-GPU Internode MPI Latency



GPU-GPU Internode MPI Bandwidth



Source: Prof. DK Panda

69% Lower Latency

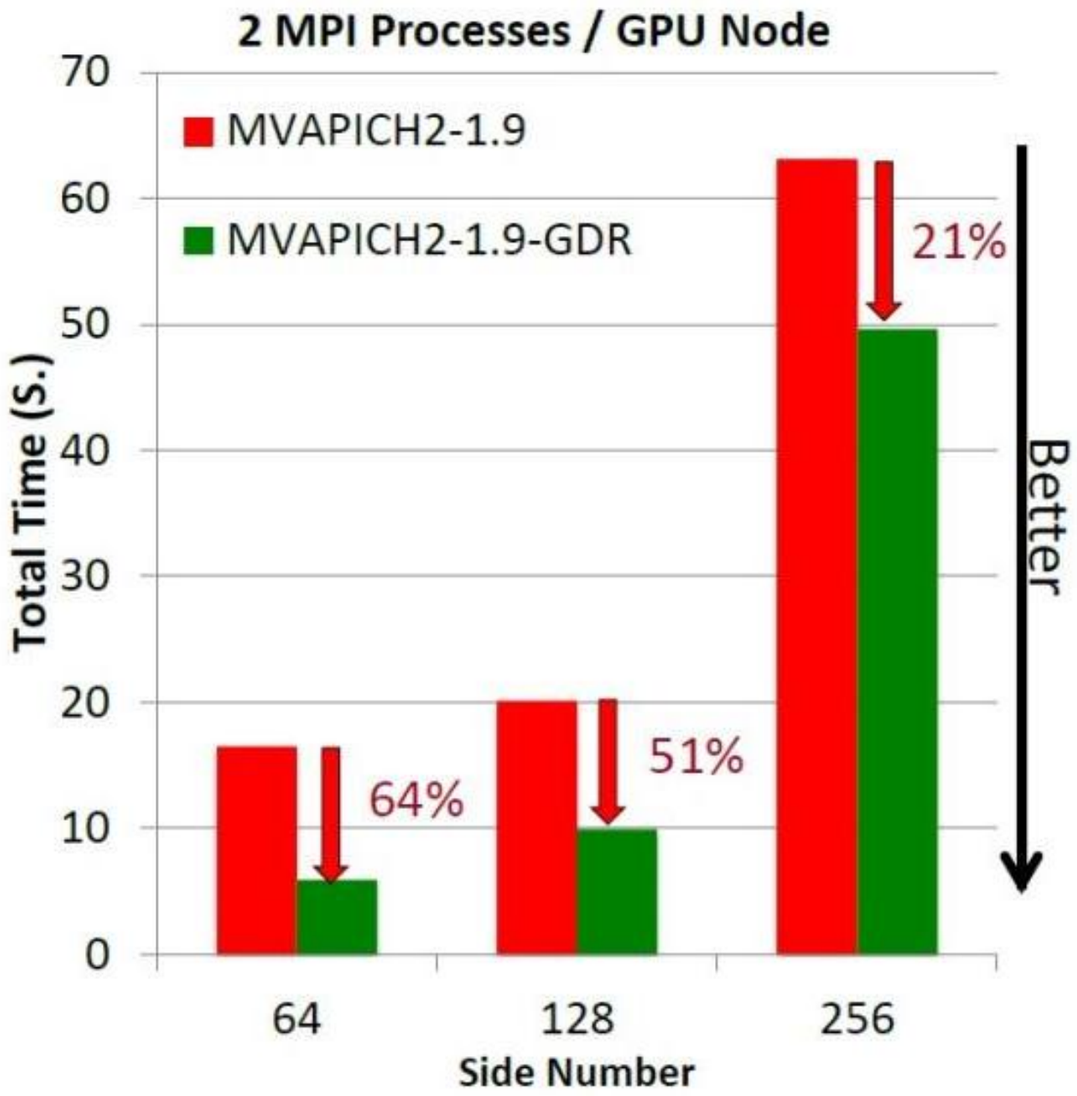
3X Increase in Throughput

Preliminary Performance of MVAPICH2 with GPU-Direct-RDMA

Execution Time of HSG

(Heisenberg Spin Glass)

Application with 2 GPU Nodes



Source: Prof. DK Panda

Thank You