Paving the Road to Exascale

High Performance Computing

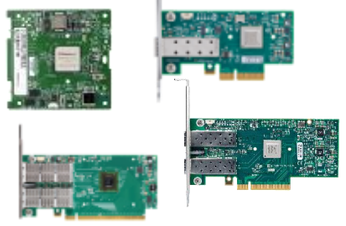
RMACC ‘15 | July 2015

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Leading Supplier of End-to-End Interconnect Solutions

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs	Adapter Cards	Switches/Gateways	Software and Services	Metro / WAN	Cables/Modules
	ConnectX-4 ConnectX-3 ConnectIB 	Spectrum SwitchIB SwitchX-2 			

At the Speeds of 10, 25, 40, 50, 56 and 100 Gigabit per Second

Technology Roadmap – One-Generation Lead over the Competition

Mellanox → 20Gbs → 40Gbs → 56Gbs → 100Gbs → 200Gbs →

Terascale | **Petascale** | **Exascale**

3rd

TOP500 2003
Virginia Tech (Apple)

1st

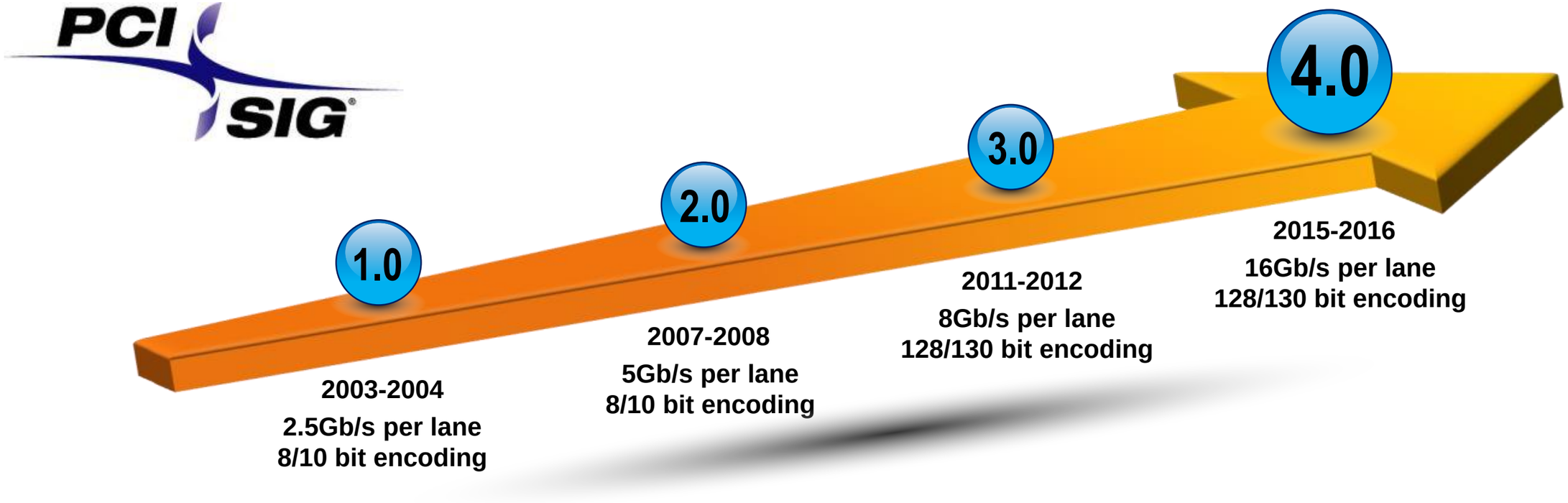
"Roadrunner"
Mellanox Connected

OAK RIDGE National Laboratory
"Summit" System

Lawrence Livermore National Laboratory
"Sierra" System

2000 | **2005** | **2010** | **2015** | **2020**

PCIe 4.0 Accelerating CPU / Memory - Interconnect Performance



PCI EXPRESS 4.0

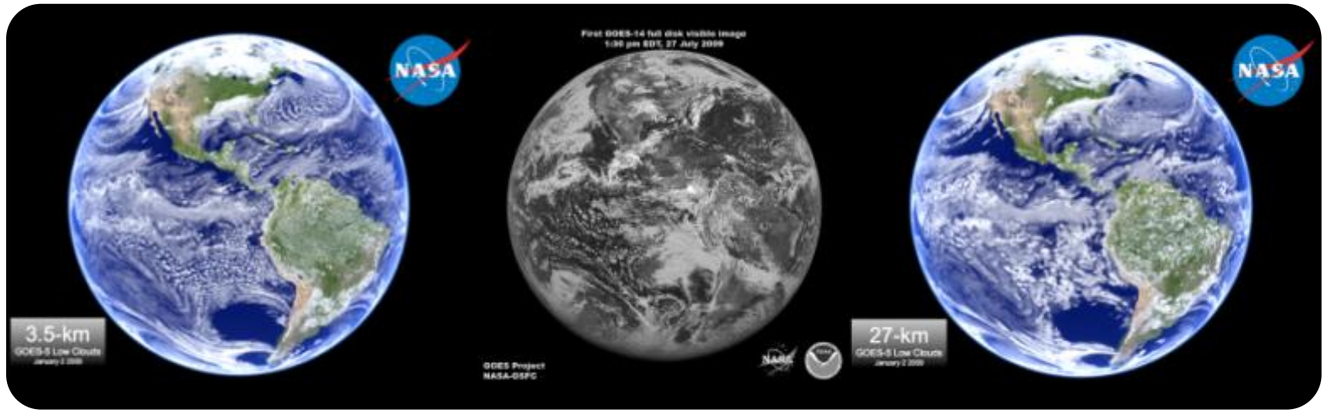
New Capabilities

- **Higher Bandwidth (16 to 25Gb/s per lane)**
- **Cache Coherency**
- **Atomic Operations**
- **Advanced Power Management**
- **Memory Management…**

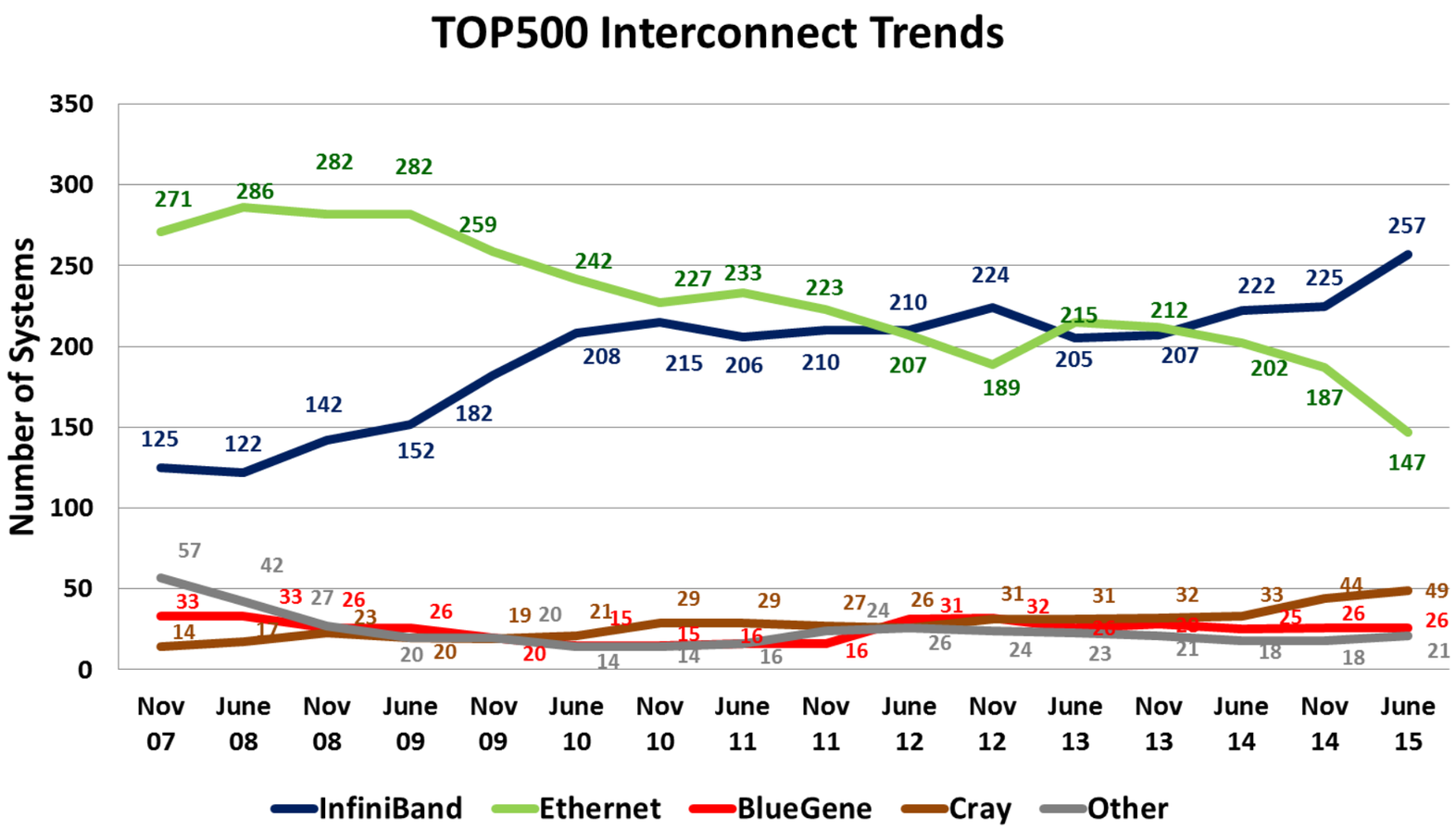
System Example: NASA Ames Research Center Pleiades

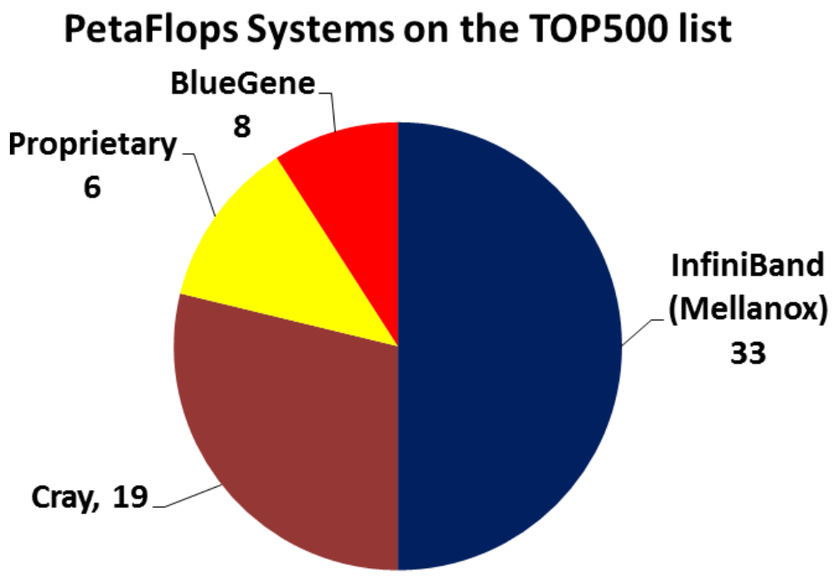
- 20K InfiniBand nodes
- Mellanox end-to-end scalable FDR and QDR InfiniBand
- Supports variety of scientific and engineering projects
  - Coupled atmosphere-ocean models
  - Future space vehicle design
  - Large-scale dark matter halos and galaxy evolution
- Leveraging InfiniBand backward and future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

High-Resolution Climate Simulations



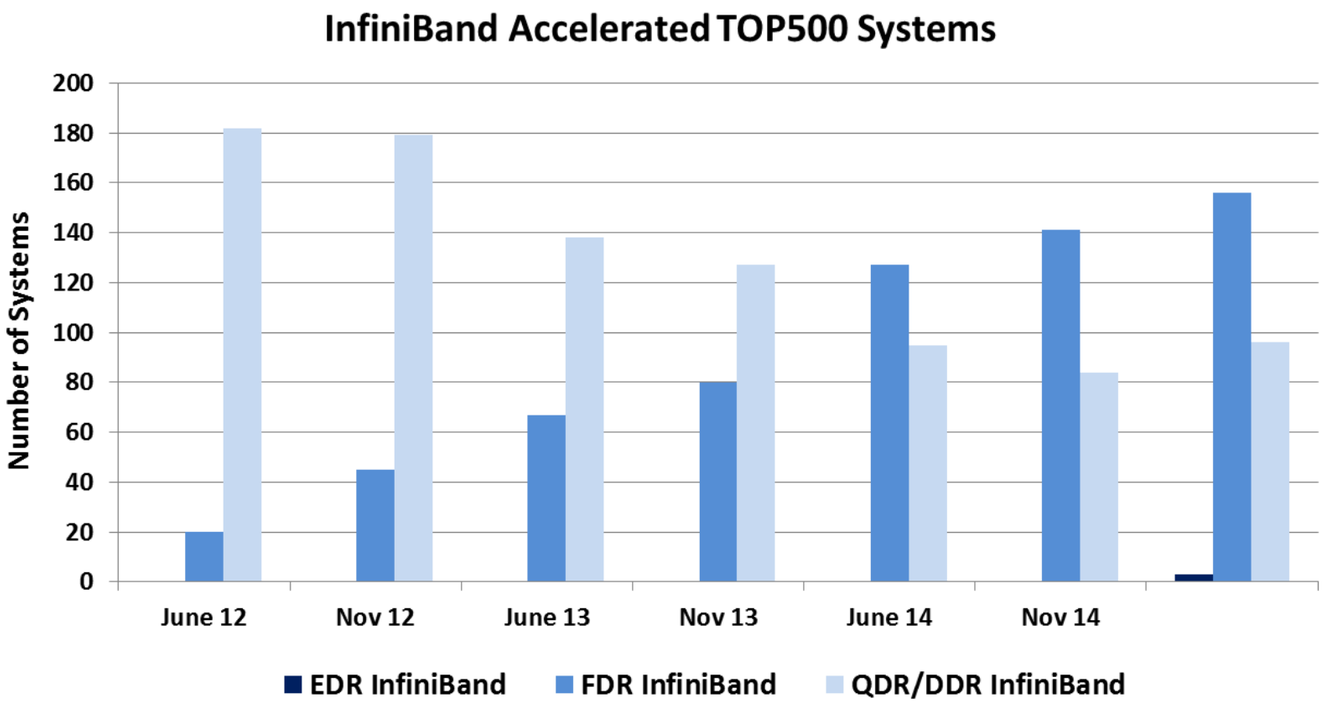
TOP500 Interconnect Trends

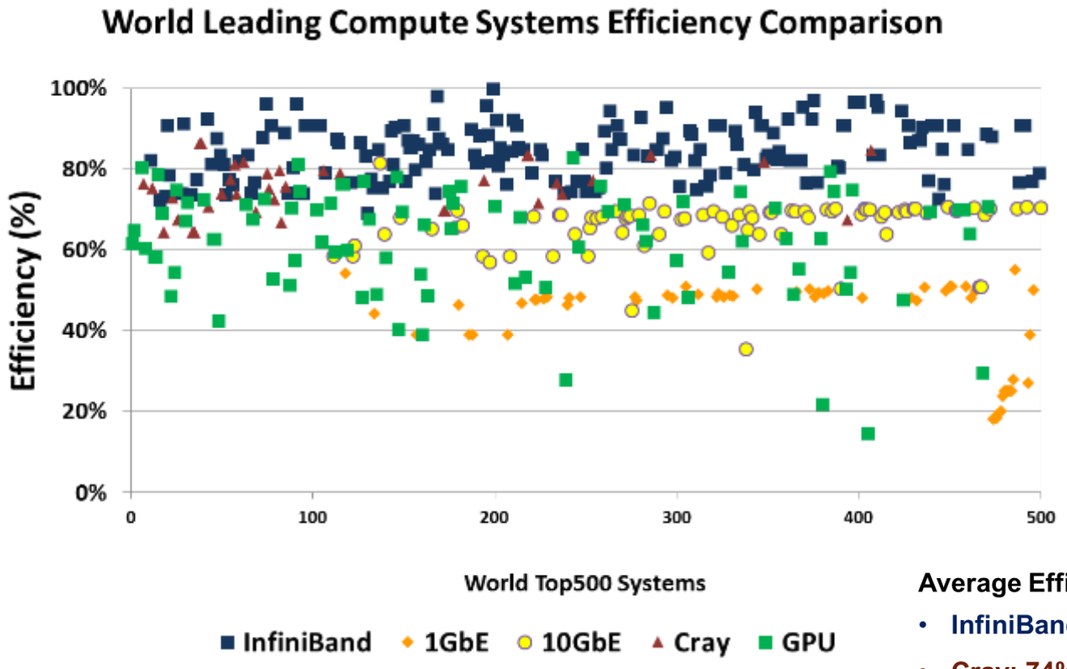




- InfiniBand is the de-facto interconnect solution for performance demanding applications

TOP500 InfiniBand Accelerated Systems





Average Efficiency

- InfiniBand: 85%
- Cray: 74%
- 10GbE: 66%
- GigE: 43%

- Number of Mellanox FDR InfiniBand systems grew 23% from June'14 to June'15
- EDR InfiniBand entered the list with 3 systems

Entering the Era of 100Gb/s

Adapters

ConnectX®-4

100Gb/s Adapter, 0.7us latency

150 million messages per second

(10 / 25 / 40 / 50 / 56 / 100Gb/s)

Switch

SwitchIB™

36 EDR (100Gb/s) Ports, <90ns Latency

Throughput of 7.2Tb/s

Switch

Spectrum™

32 100GbE Ports, 64 25/50GbE Ports

(10 / 25 / 40 / 50 / 100GbE)

Throughput of 6.4Tb/s

Interconnect

LinkX™

Copper (Passive, Active)

Optical Cables (VCSEL)

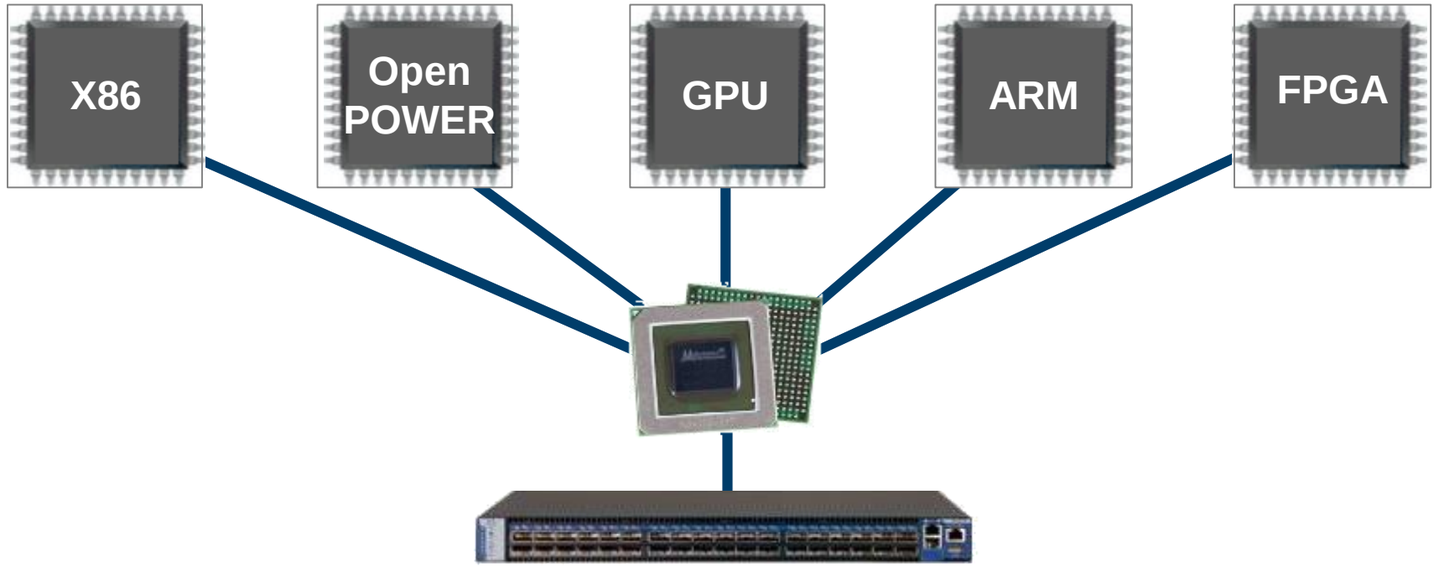
Silicon Photonics

End-to-End Interconnect Solutions for All Platforms

Highest Performance and Scalability for

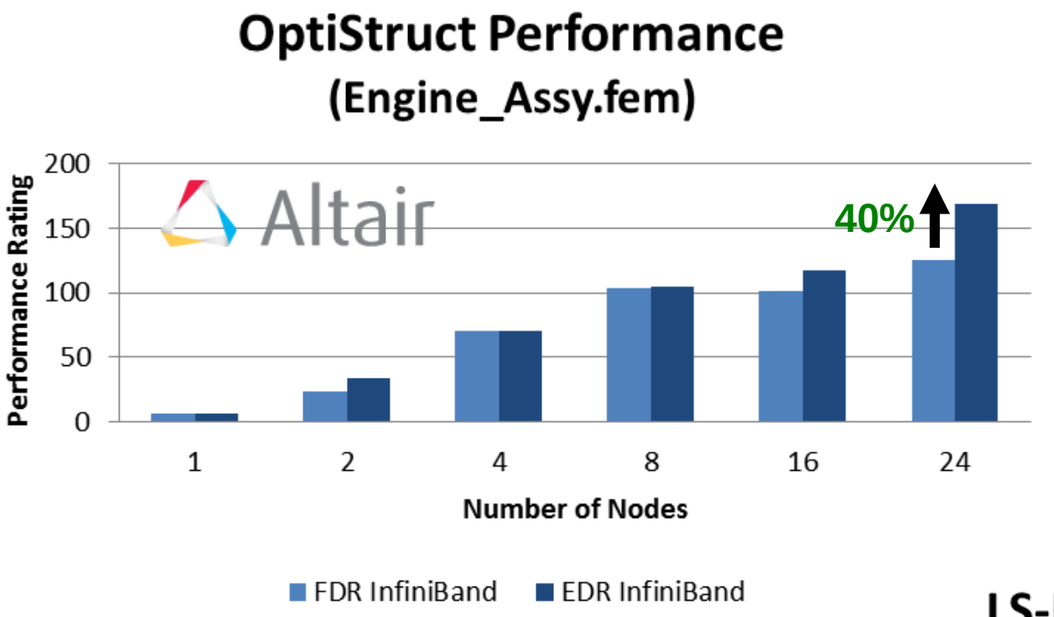
X86, Power, GPU, ARM and FPGA-based Compute and Storage Platforms

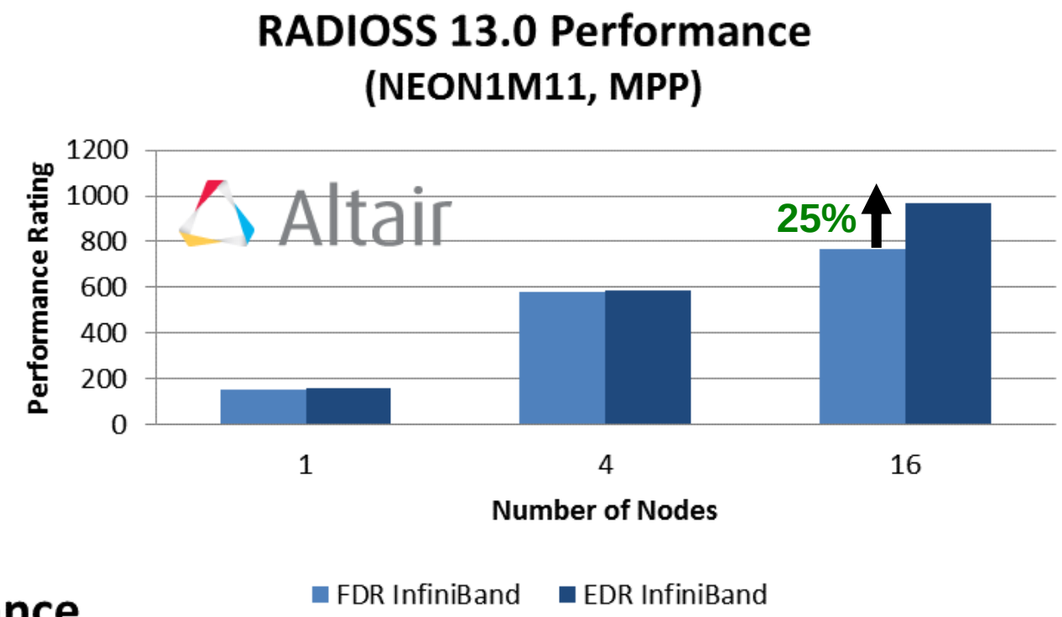
10, 20, 25, 40, 50, 56 and 100Gb/s Speeds

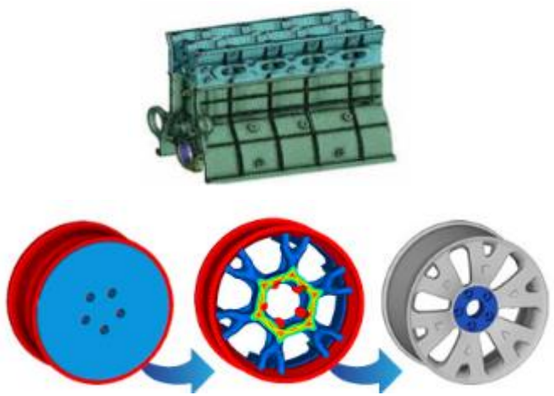


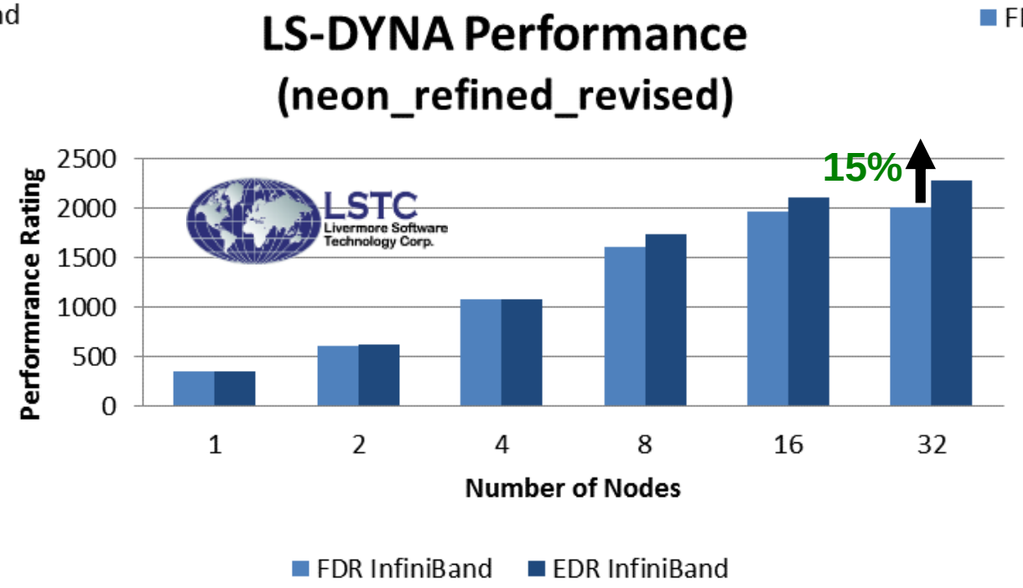
Smart Interconnect to Unleash The Power of All Compute Architectures

EDR InfiniBand Performance – Commercial Applications



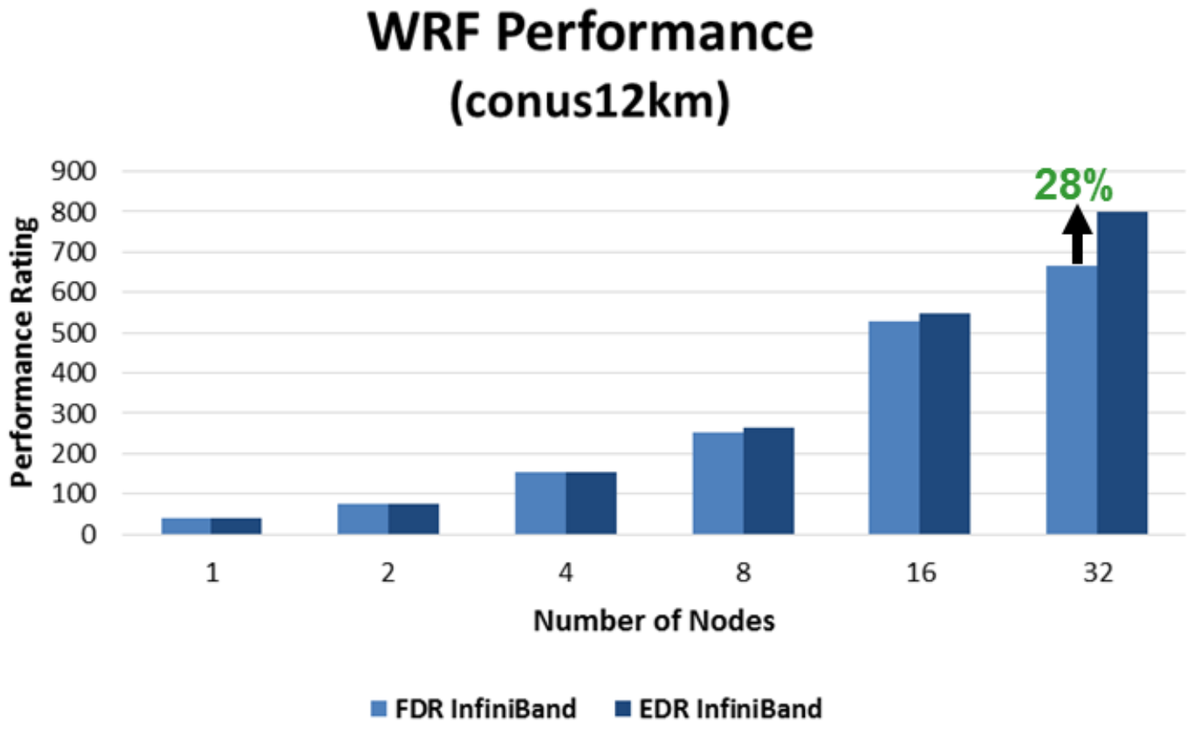




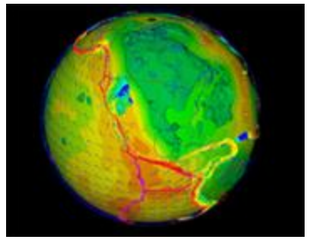


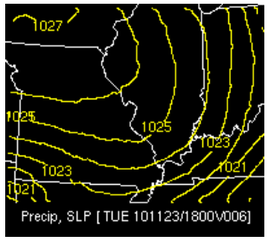
EDR InfiniBand Performance – Weather Simulation

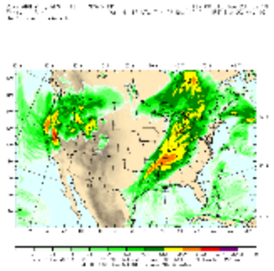
- Weather Research and Forecasting Model
- Optimization effort with the HPCAC
- EDR InfiniBand delivers 28% higher performance
  - 32-node cluster
  - Performance advantage increase with system size



WRF THE WEATHER RESEARCH & FORECASTING MODEL



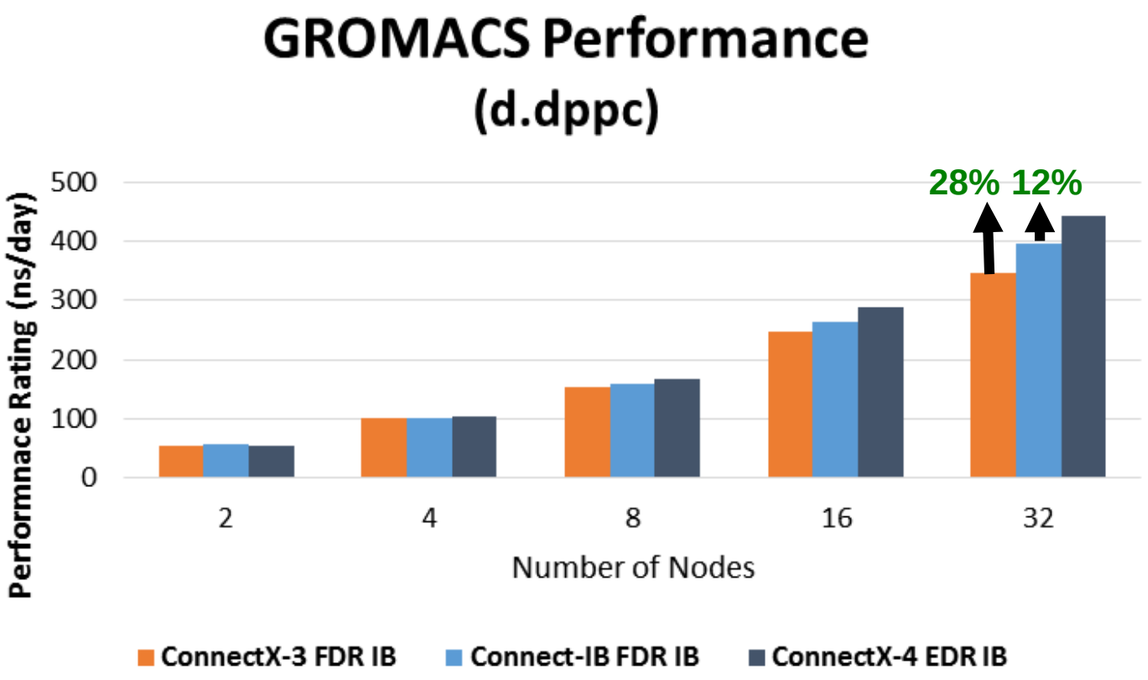




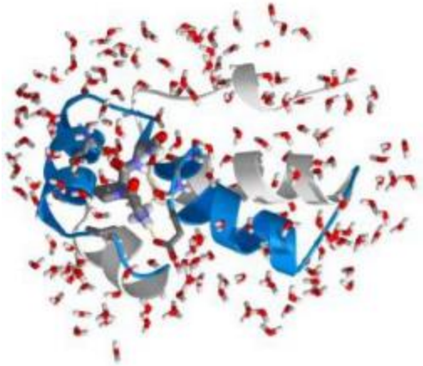
HPC ADVISORY COUNCIL

EDR InfiniBand Performance – CFD Applications

- A molecular dynamics simulation software
- ConnectX-4 EDR IB delivers highest performance
  - 28% higher performance versus ConnectX-3 FDR
  - 12% higher performance versus Connect-IB FDR
  - 32-node cluster
  - Performance advantage increase with system size



GROMACS FAST. FLEXIBLE. FREE.



HPC ADVISORY COUNCIL

Mellanox HPC-X™ Scalable HPC Software Toolkit

- MPI, PGAS OpenSHMEM and UPC package
- Maximize application performance
- For commercial and open source applications
- Based on UCX (Unified Communication – X Framework)

Mellanox Delivers Highest Applications Performance (HPC-X)

- Bull (Atos) testing results – Quantum Expresso application

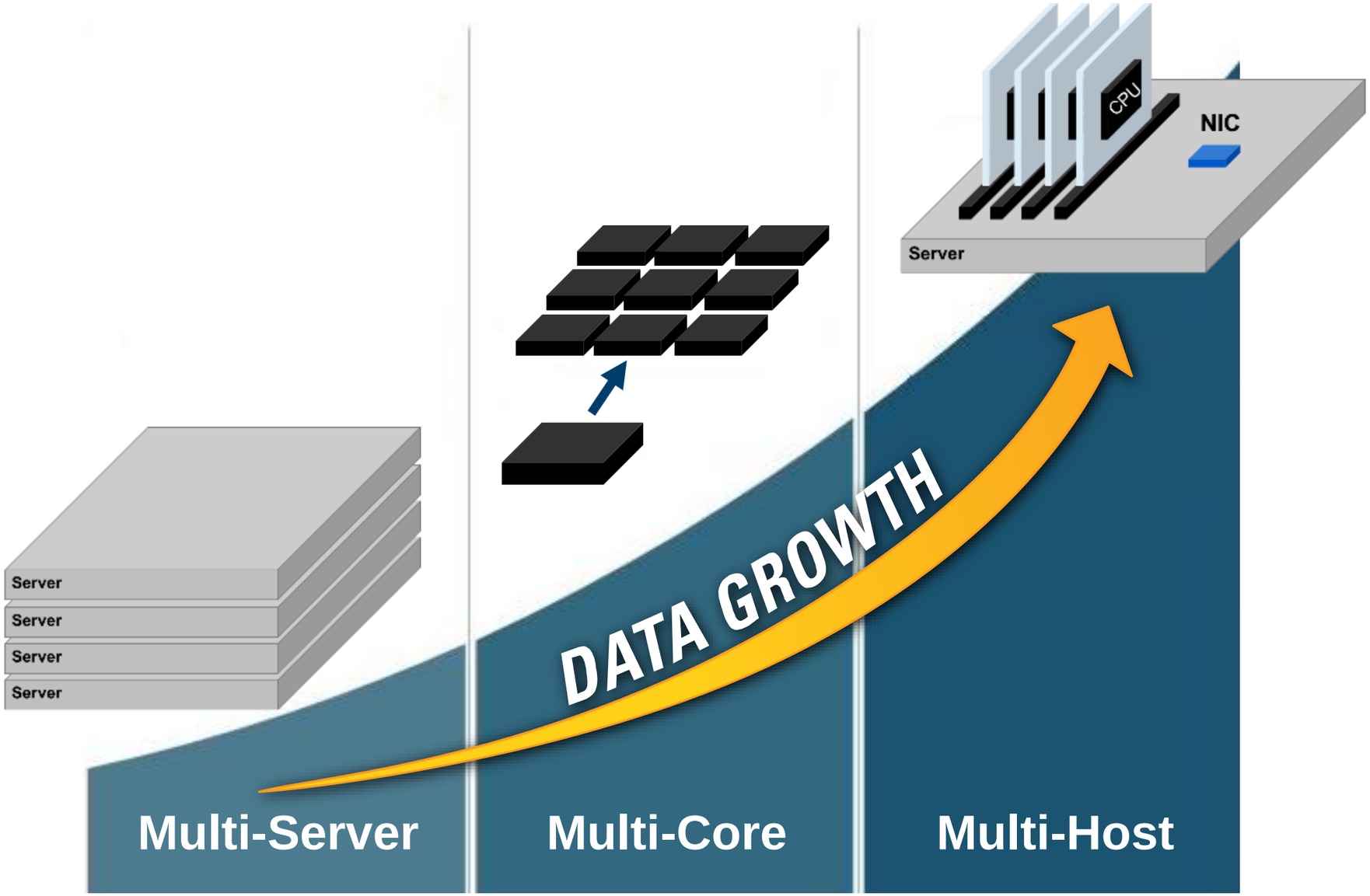
			Intel MPI	Bull MPI (HPC-X)	
Quantum Espresso	Test Case	# nodes	time (s)	time (s)	Gain
	A	43	584	368	**37%**
	B	196	2592	998	**61%**

Enabling Highest Applications Scalability and Performance

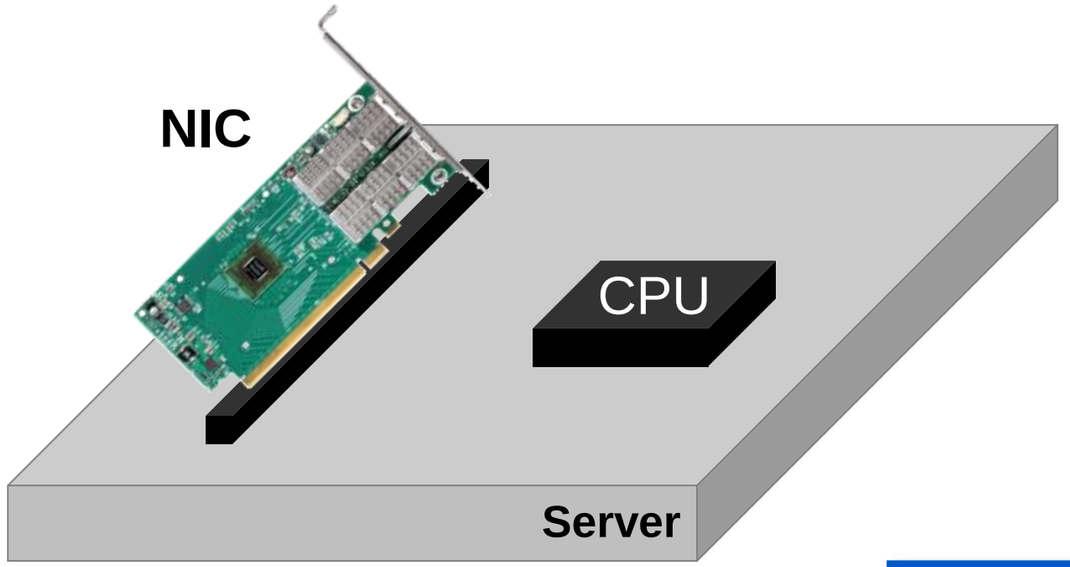
Mellanox Multi-Host™ Technology

Next Generation Data Center Architecture

Data Center Evolution Over Time



New Compute Rack / Data Center Architecture

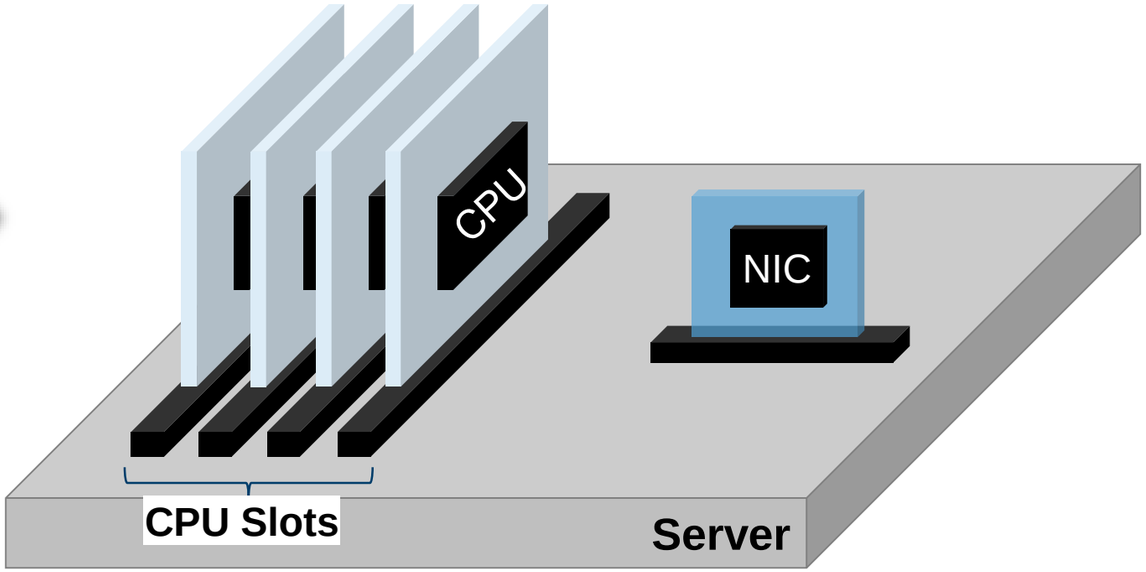


Traditional Data Center

- Expensive design for fixed data centers
- Requires many ports on top-of-rack switch
- Dedicated NIC / cable per server

Scalable Data Center with Multi-Host

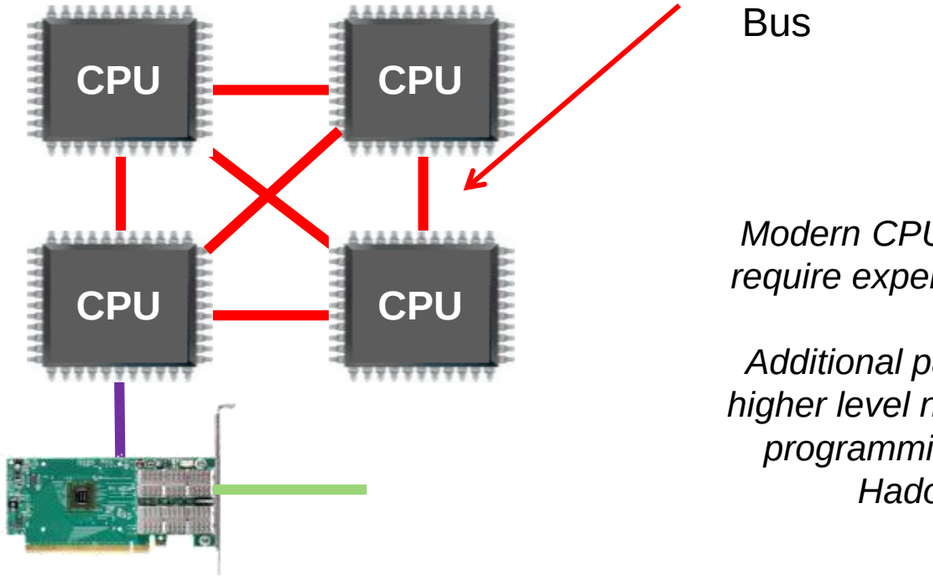
- Flexible, configurable, application optimized
- Optimized top-of-rack switches
- Takes advantage of high-throughput network



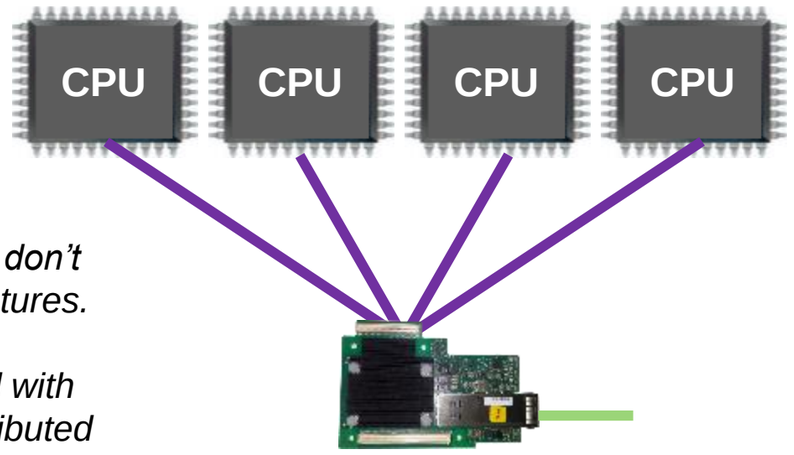
The Network is The Computer

Multi-Host Dramatically Reduces Server Cost

Traditional 4-Way SMP Machine



Multi-Host 4-Socket Architecture



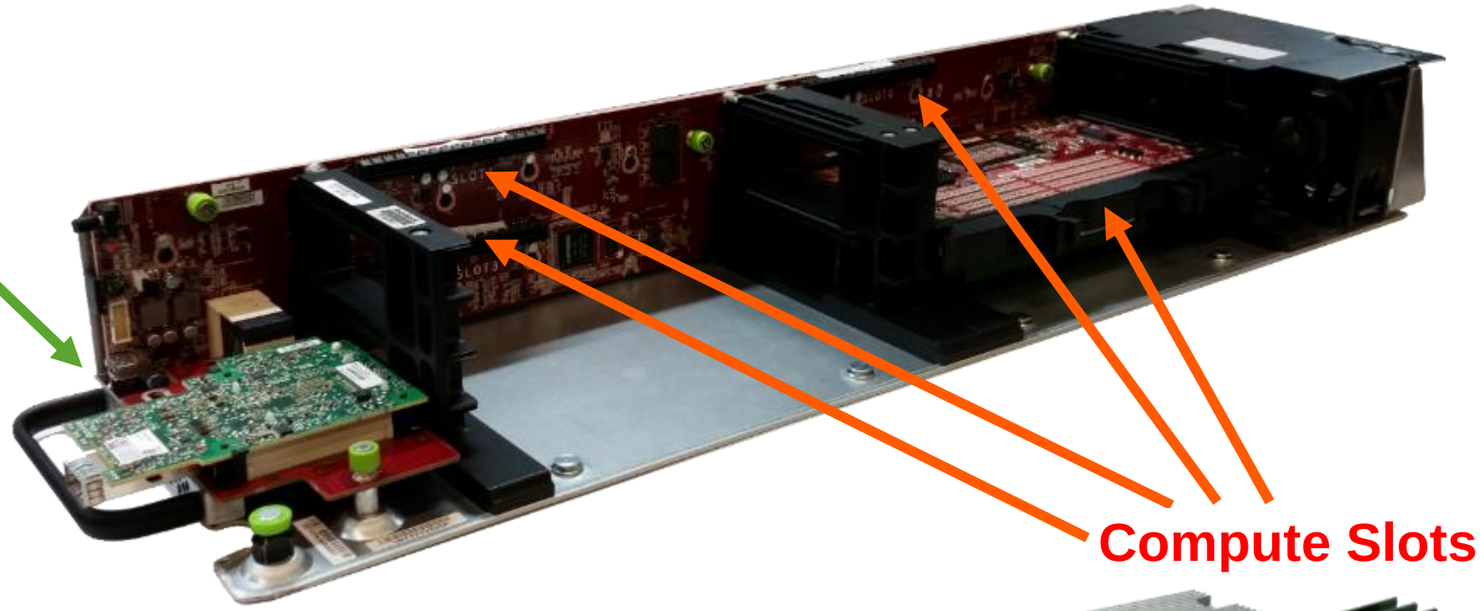
Modern CPUs with 8-20 cores don't require expensive SMP architectures.

Additional parallelism achieved with higher level network based distributed programming techniques such as Hadoop Map-Reduce

- Expensive 4-Way CPU
  - Massive but unused cache-coherent domain
- High overhead but un-necessary CPU bus
  - High pin count ,high power, complex layout
- Asymmetric (NUMA) of data access

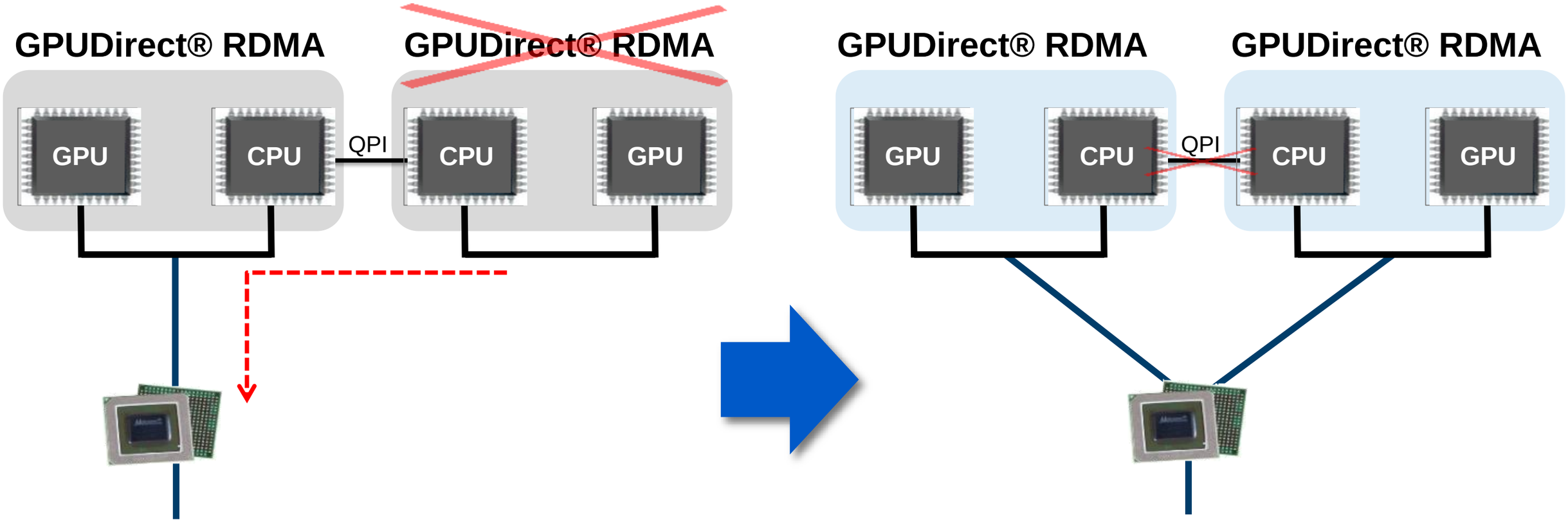
- Low cost single-socket CPU
  - Clean, simple, cost-effective, software transparent
- Cache coherent domain: Multi-Core CPU
  - Eliminates pins, Lower power, Simpler layout
- Symmetric Data Access

ConnectX-4 on Facebook OCP Multi-Host Platform (Yosemite)



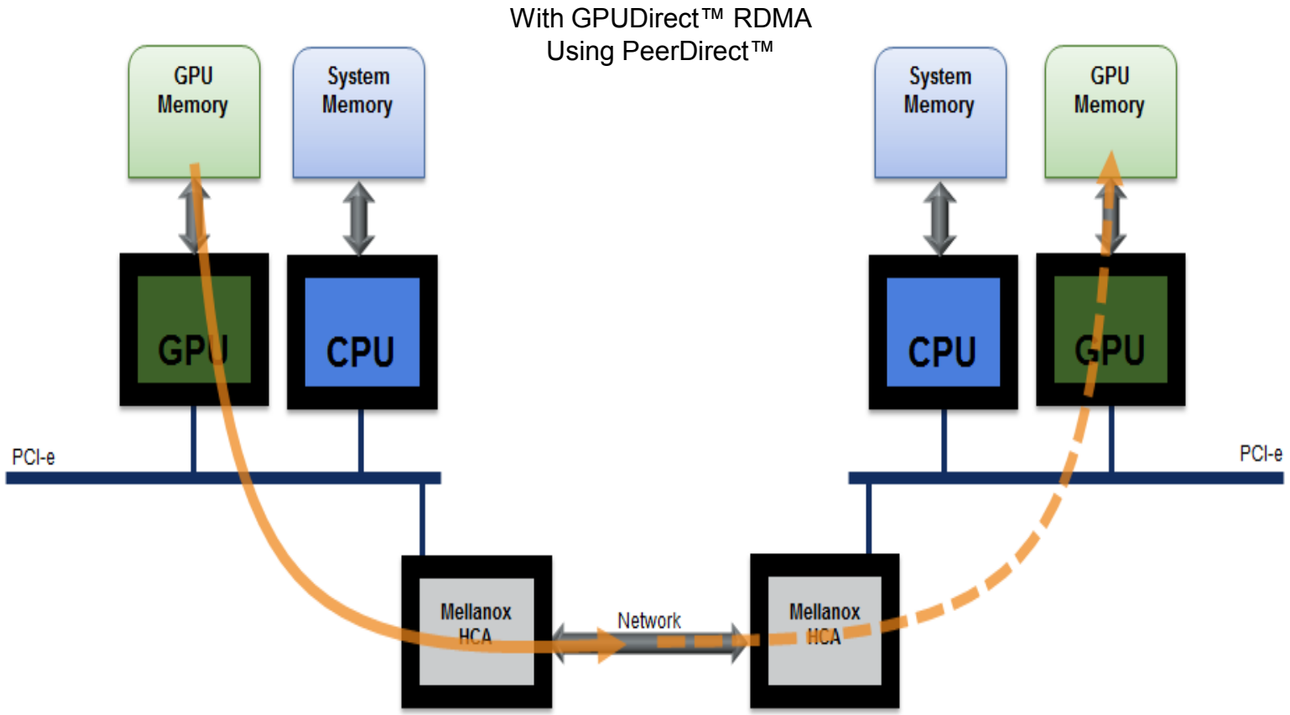
The Next Generation Compute and Storage Rack Design

Enabling GPUDirect RDMA Across all Available GPUs



Smart Interconnect to Unleash the Power of All Compute Architectures

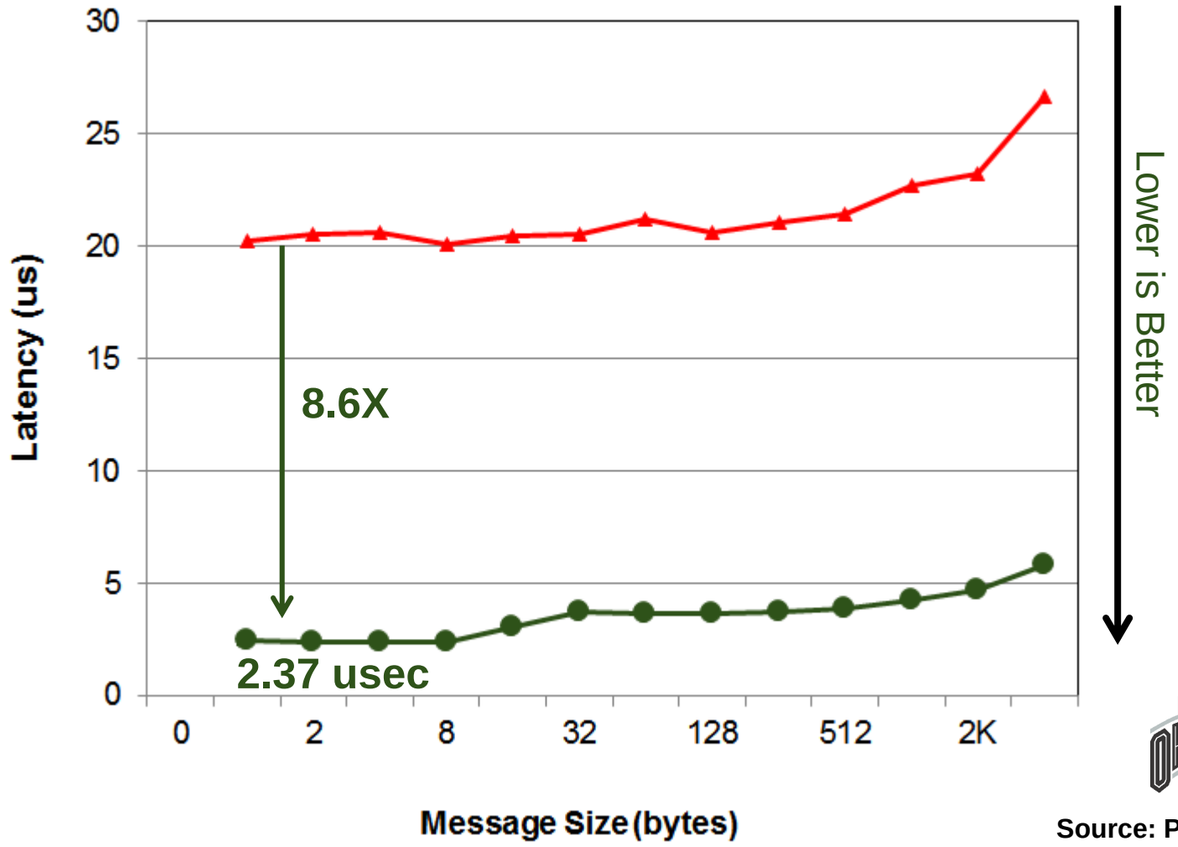
GPUDirect™ RDMA (GPUDirect 3.0)



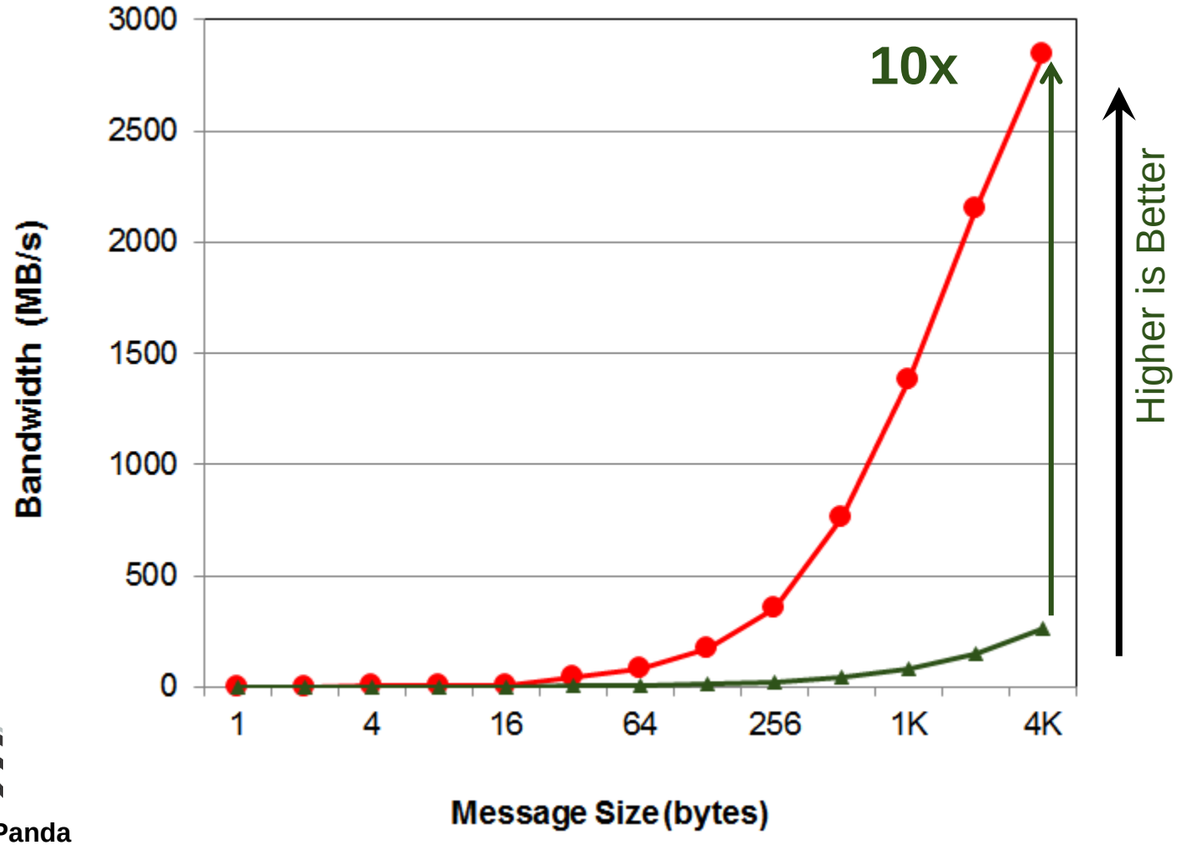
- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPI efficiency between GPUs in remote nodes
- Based on PCIe PeerDirect technology

Performance of MVAPICH2 with GPUDirect RDMA

GPU-GPU Internode MPI Latency



GPU-GPU Internode MPI Bandwidth



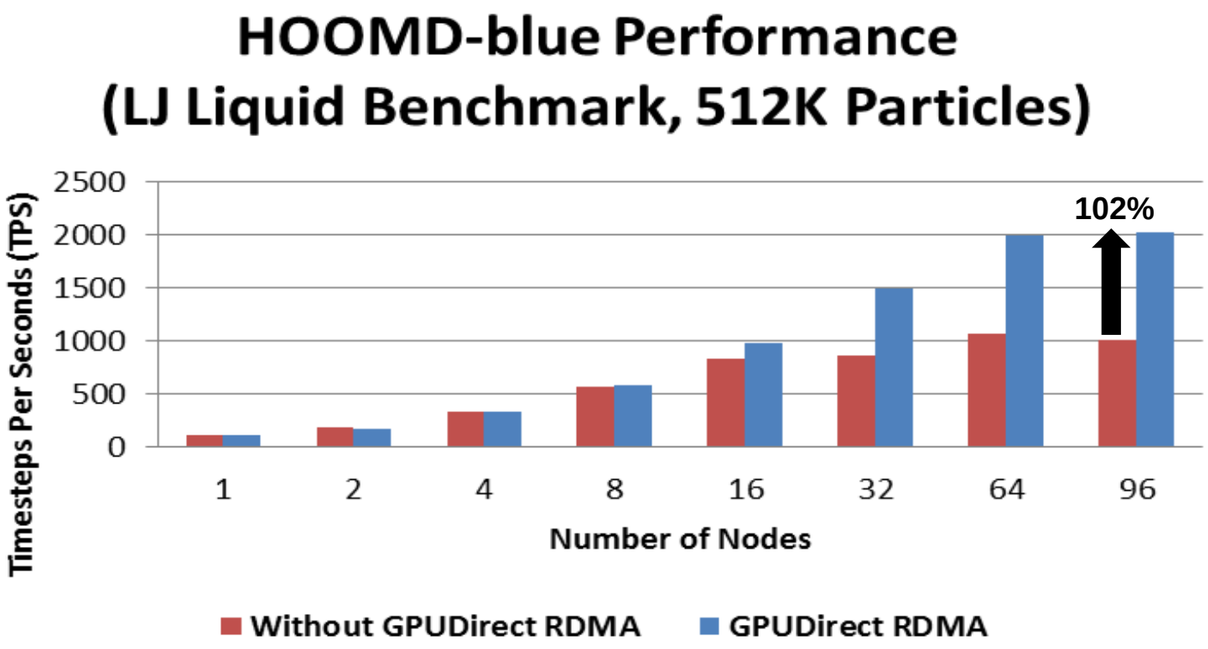
Source: Prof. DK Panda

88% Lower Latency

10X Increase in Throughput

Mellanox PeerDirect™ with NVIDIA GPUDirect RDMA

- HOOMD-blue is a general-purpose Molecular Dynamics simulation code accelerated on GPUs
- GPUDirect RDMA allows direct peer to peer GPU communications over InfiniBand
  - Unlocks performance between GPU and InfiniBand
  - This provides a significant decrease in GPU-GPU communication latency
  - Provides complete CPU offload from all GPU communications across the network
- Demonstrated up to 102% performance improvement with large number of particles

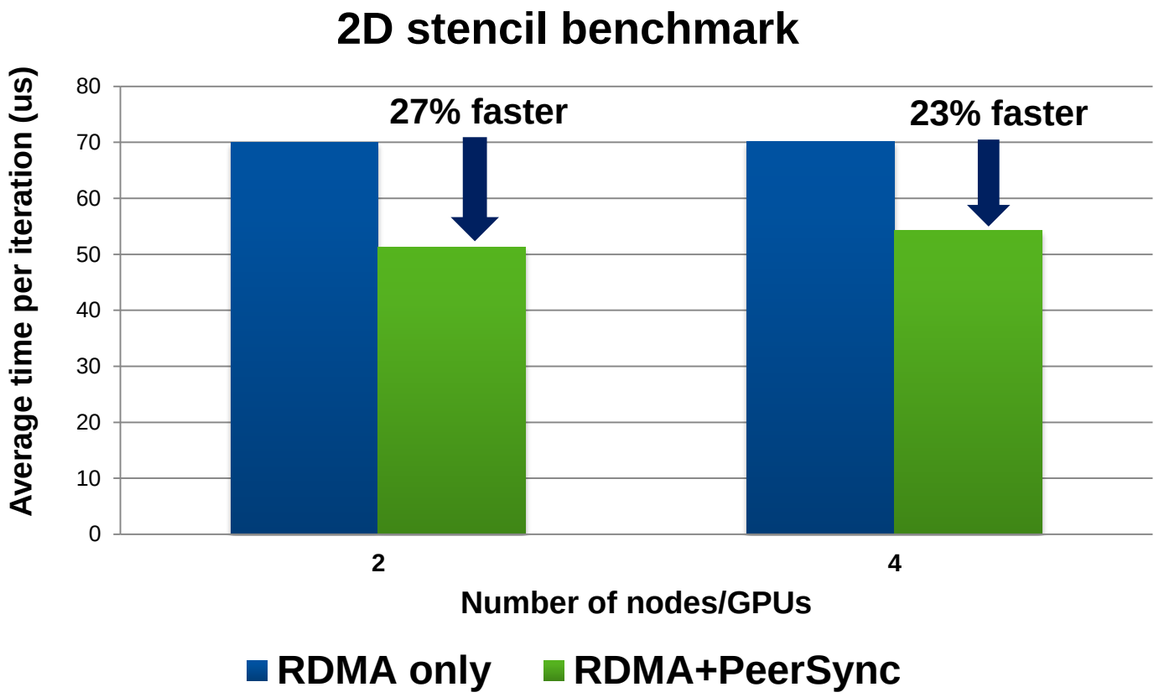


GPUDirect Async (GPUDirect 4.0)

- GPUDirect RDMA (3.0) – direct data path between the GPU and Mellanox interconnect
  - Control path still uses the CPU
    - CPU prepares and queues communication tasks on GPU
    - GPU triggers communication on HCA
    - Mellanox HCA directly accesses GPU memory

- GPUDirect Async (GPUDirect 4.0)
  - Both data path and control path go directly between the GPU and the Mellanox interconnect

Maximum Performance For GPU Clusters



Mellanox Interconnect Advantages

- Mellanox solutions provide a proven, scalable and high performance end-to-end connectivity
- Flexible, support all compute architectures: x86, Power, ARM, GPU, FPGA etc.
- Standards-based (InfiniBand, Ethernet), supported by large eco-system
- Higher performance: 100Gb/s, 0.7usec latency, 150 million messages/sec
- HPC-X software provides leading performance for MPI, OpenSHMEM/PGAS and UPC
- Superiors applications offloads: RDMA, Collectives, scalable transport
- Backward and future compatible

Speed-Up Your Present, Protect Your Future
Paving The Road to Exascale Computing Together

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™