Interconnect Your Future with Mellanox

Journées Mesocentres 2016

Paving the Road to Exascale Computing

Saddik El Arguioui – Mellanox

Mellanox TECHNOLOGIES

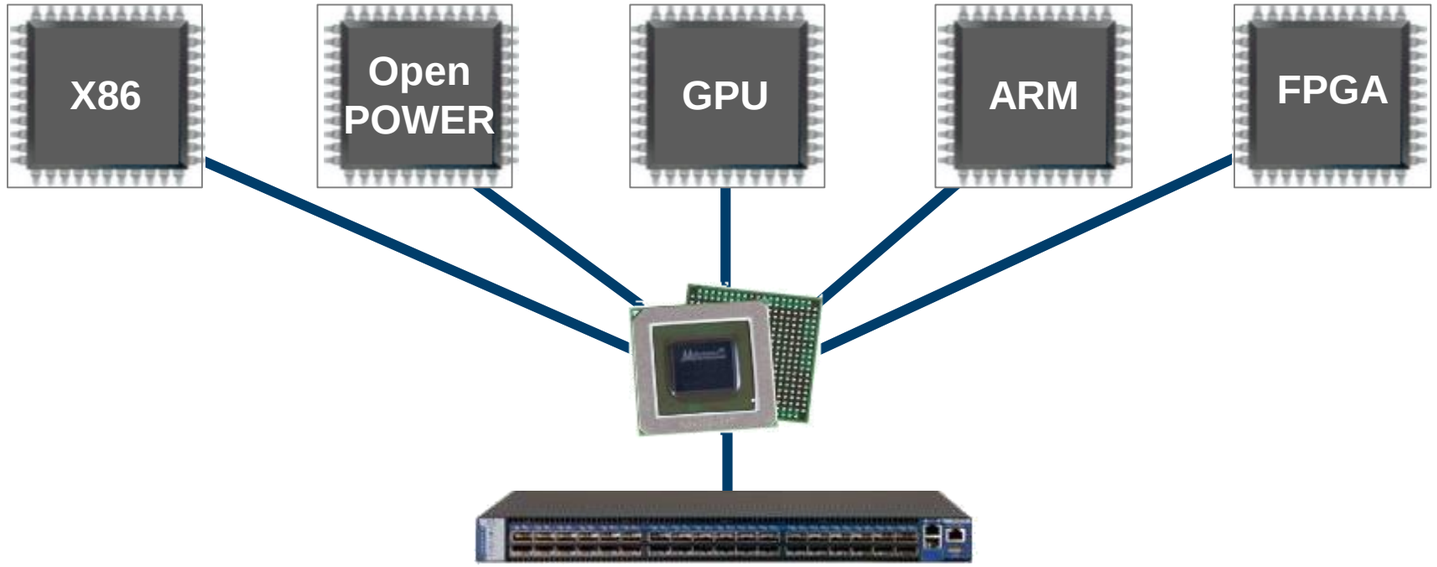
Connect. Accelerate. Outperform.™

End-to-End Interconnect Solutions for All Platforms

Highest Performance and Scalability for

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

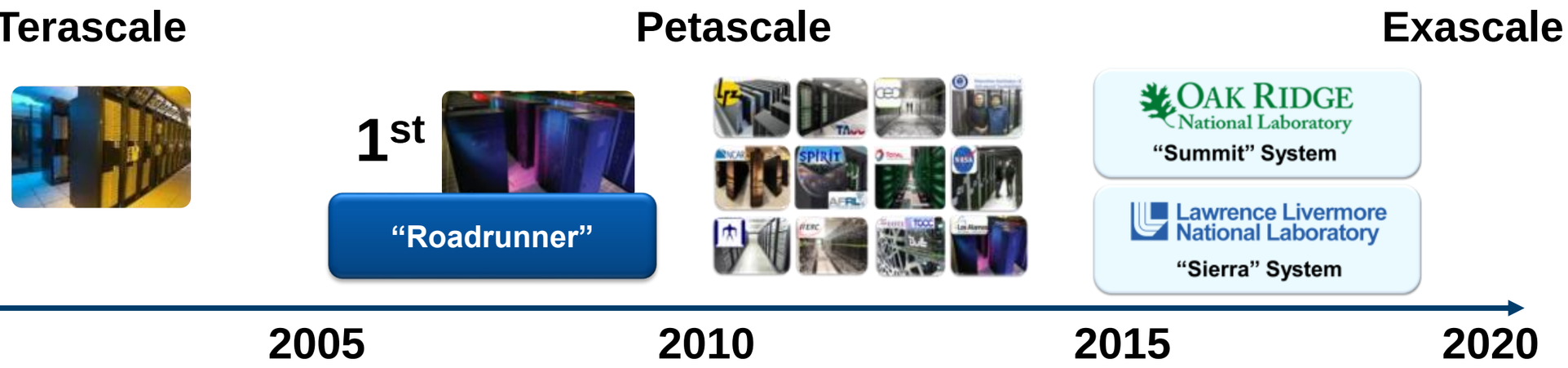
10, 20, 25, 40, 50, 56 and 100Gb/s Speeds



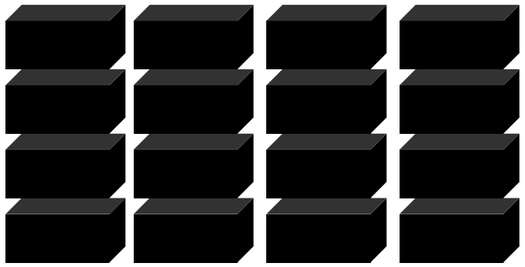
Smart Interconnect to Unleash The Power of All Compute Architectures

The Ever Growing Demand for Higher Performance

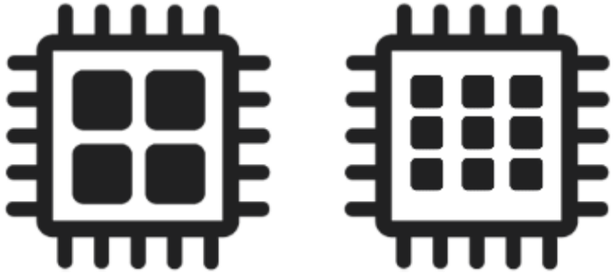
Performance Development



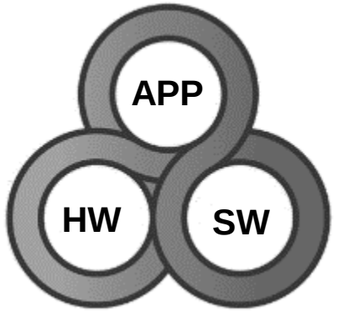
The Interconnect is the Enabling Technology



SMP to Clusters



Single-Core to Many-Core

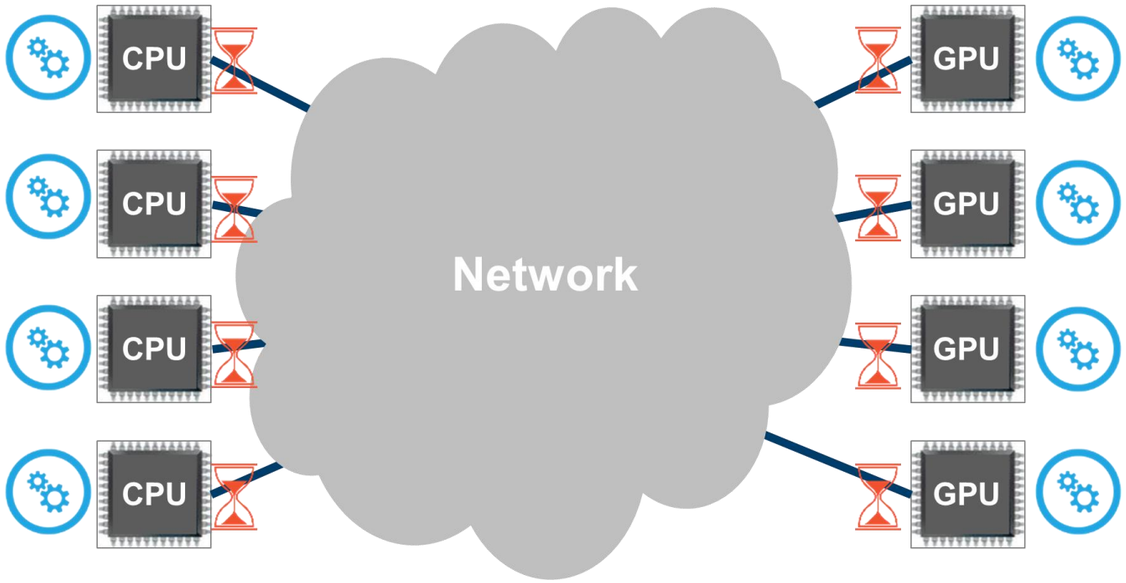


Application
Software
Hardware

Co-Design

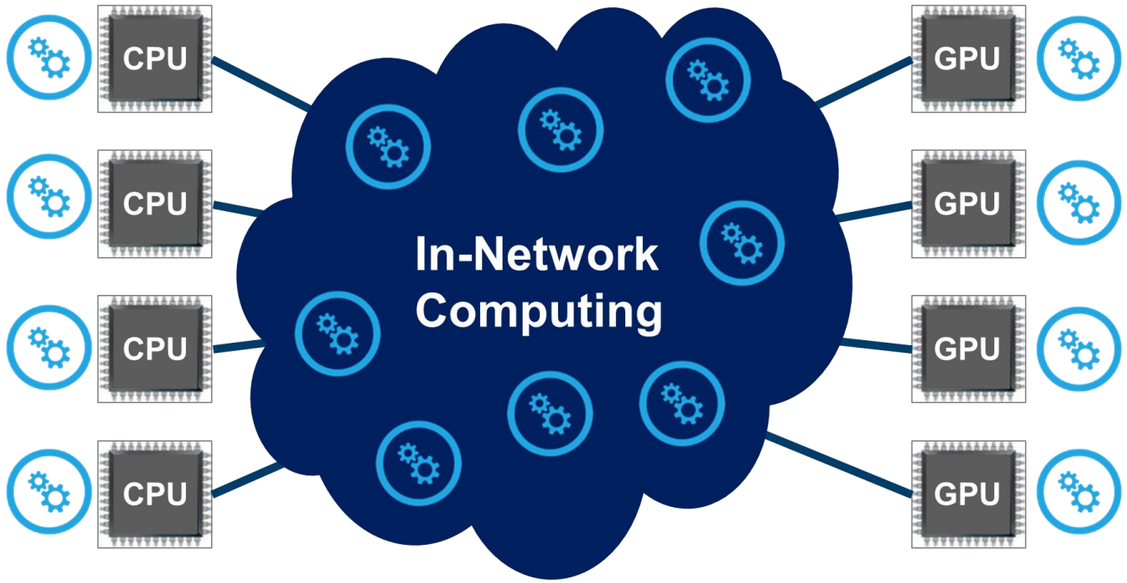
The Intelligent Interconnect to Enable Exascale Performance

CPU-Centric



Limited to Main CPU Usage
Results in Performance Limitation

Co-Design

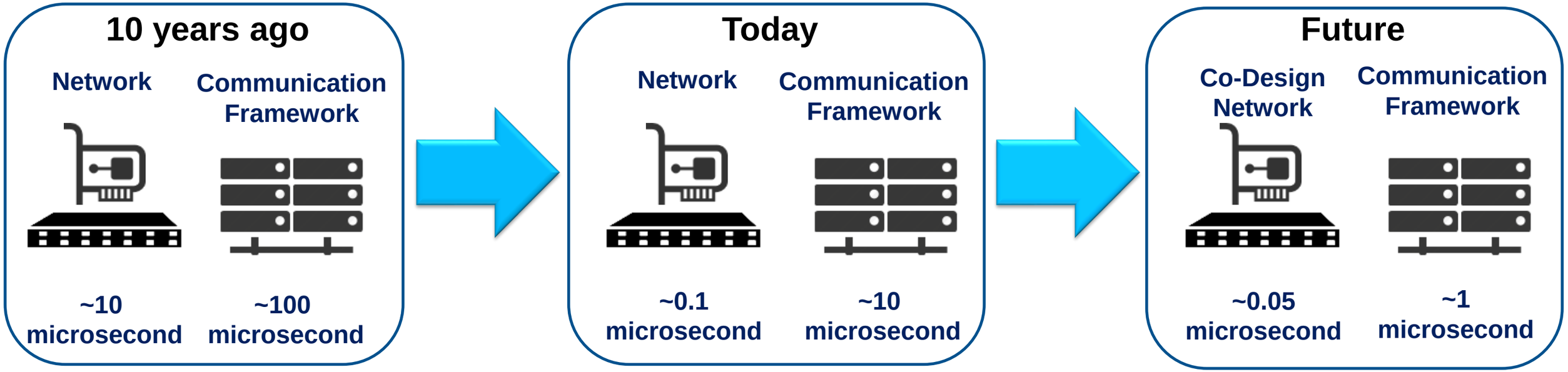


Creating Synergies
Enables Higher Performance and Scale

Must Wait for the Data
Creates Performance Bottlenecks

Work on The Data as it Moves
Enables Performance and Scale

Breaking the Application Latency Wall



- Today: Network device latencies are on the order of 100 nanoseconds
- Challenge: Enabling the next order of magnitude improvement in application performance
- Solution: Creating synergies between software and hardware – intelligent interconnect

Intelligent Interconnect Paves the Road to Exascale Performance

ISC'16: Introducing ConnectX-5 World's Smartest Adapter

State of the Smart

a new generation of co-processors emerges

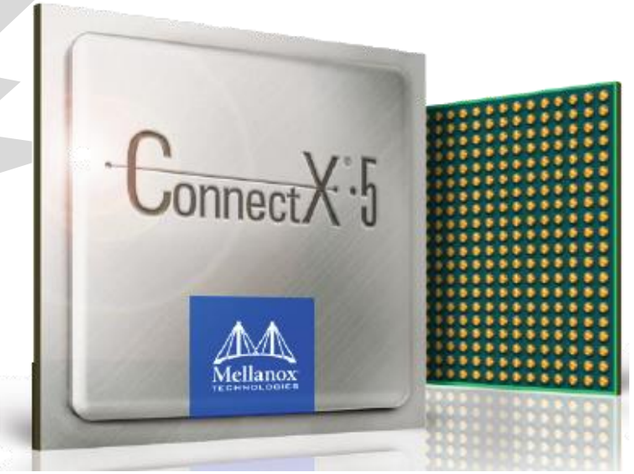
Mellanox Smart Interconnect

Switch-IB 2

SHArP

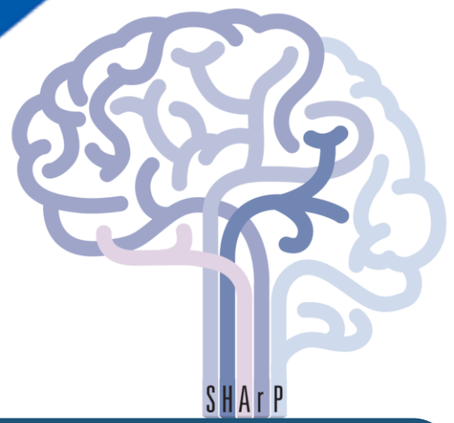
NEW!

ConnectX-5



Switch-IB 2 EDR 100G Advantages

State of the Smart



Switch-IB 2 SHArP

SHArP Enables Switch-IB 2 to Manage and Execute MPI Operations in the Network

Switch-IB 2 Enables the Switch Network to Operate as a Co-Processor

Delivering 10X Performance Improvement for MPI and SHMEM/PAGS Communications

ConnectX-5 EDR 100G Advantages

State of the Smart

NEW!

Performance
- 100Gb/s Throughput
- 0.6usec Latency (end-to-end)
- 200M Messages per Second

Smart
- MPI Collectives in Hardware
- MPI Tag Matching in Hardware
- In-Network Memory

Platform
- PCIe Gen3 and Gen4
- Integrated PCIe Switch
- Advanced Dynamic Routing

Highest-Performance 100Gb/s Interconnect Solutions

Adapters

ConnectX®-5

100Gb/s Adapter, 0.6us latency

200 million messages per second

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

Switch

SwitchIB™-2

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

Throughput of 7.2Tb/s

7.02 Billion msg/sec (195M msg/sec/port)

Switch

Spectrum™

32 100GbE Ports, 64 25/50GbE Ports

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

Throughput of 6.4Tb/s

Interconnect

LinkX™

Transceivers

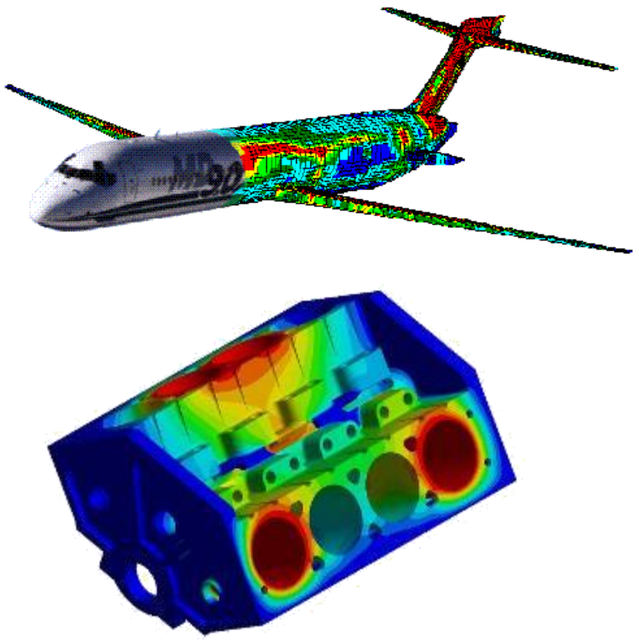
Active Optical and Copper Cables

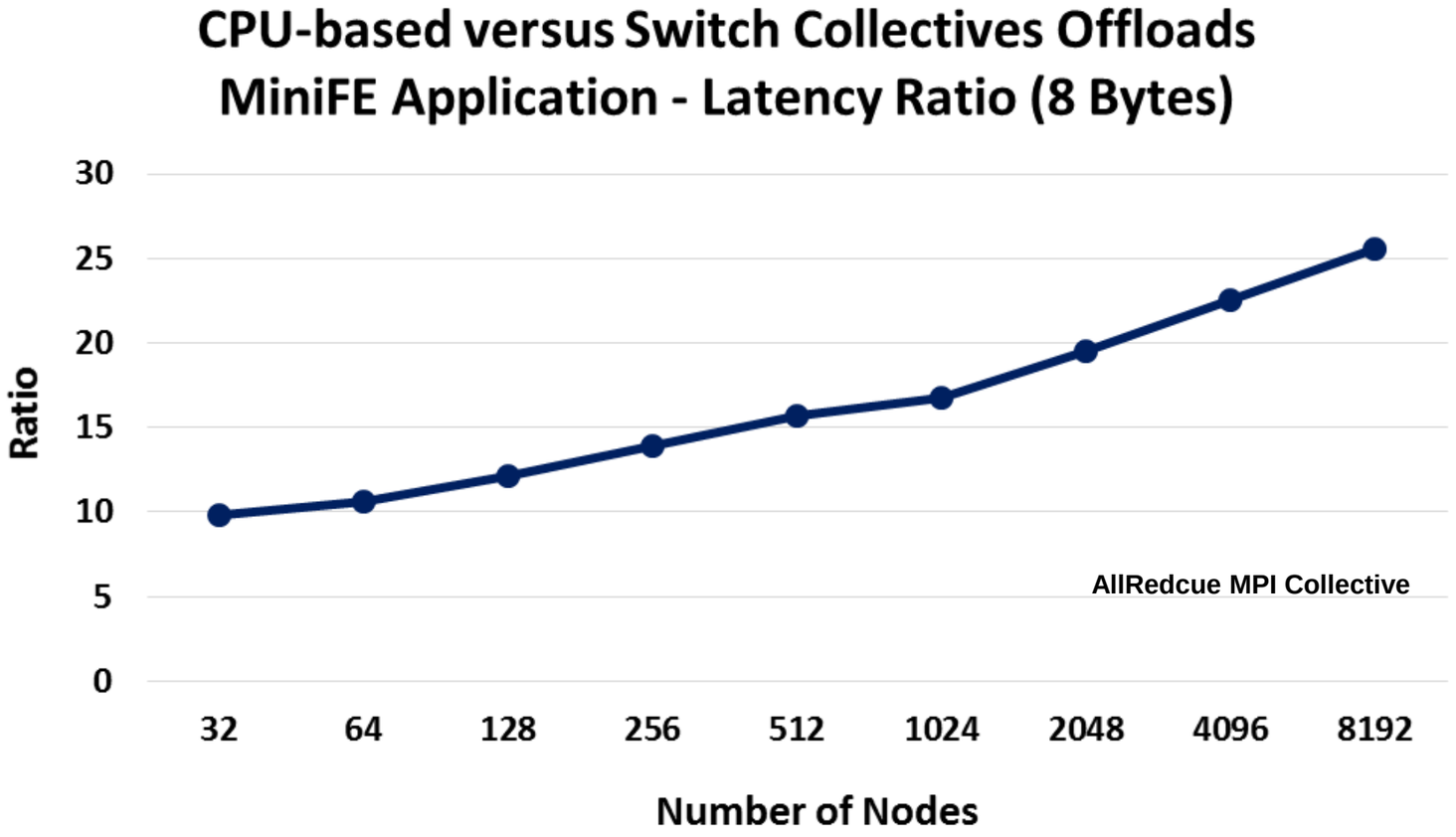
(10 / 25 / 40 / 50 / 56 / 100Gb/s)

VCSELs, Silicon Photonics and Copper

SHArP Performance Advantage

- MiniFE is a Finite Element mini-application
  - Implements kernels that represent implicit finite-element applications



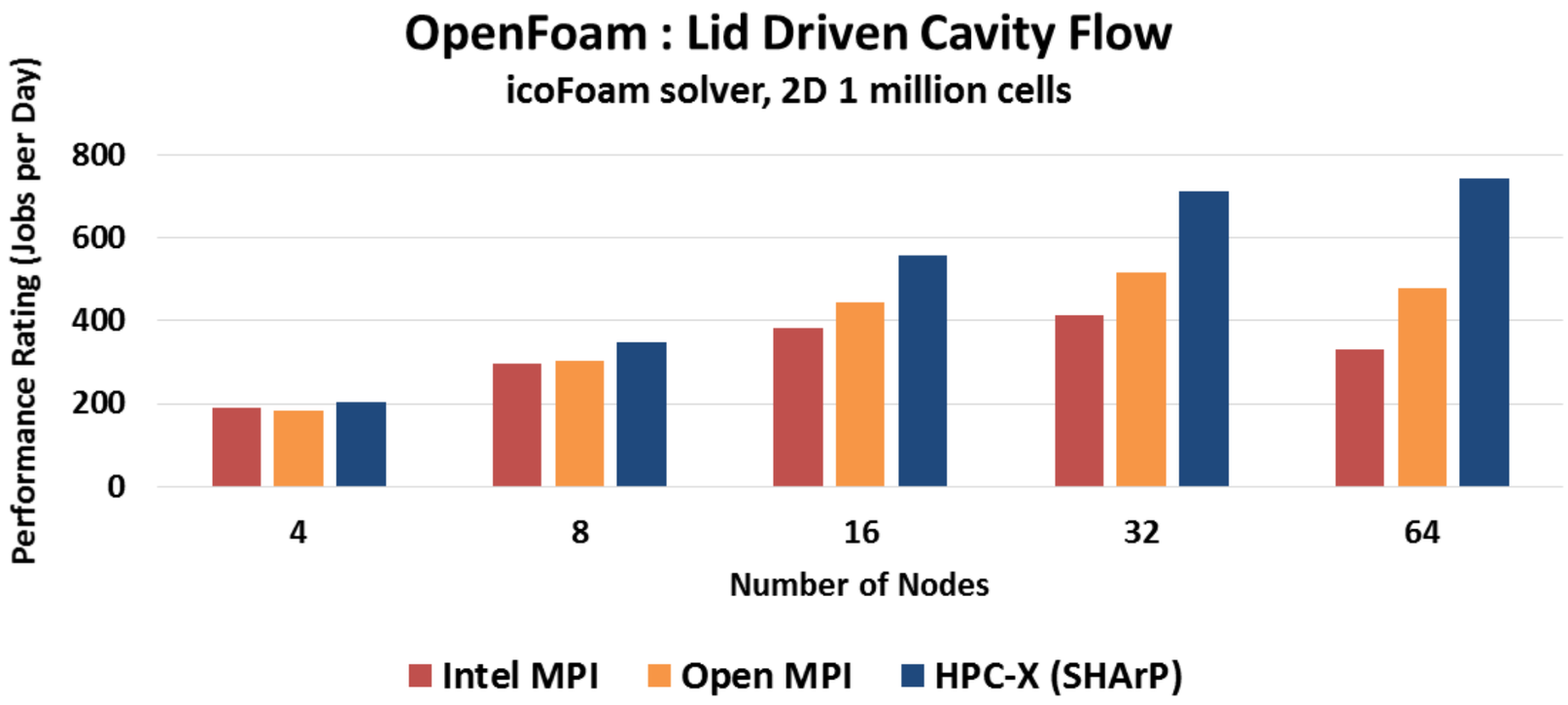


10X to 25X Performance Improvement!

HPC-X with SHArP Technology

OpenFOAM

OpenFOAM is a popular computational fluid dynamics application



HPC-X™

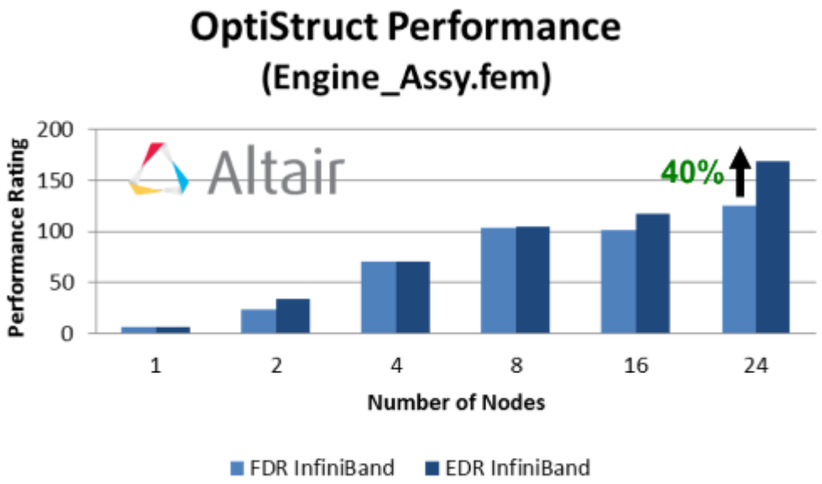
HPC-X ScalableHPC Toolkit

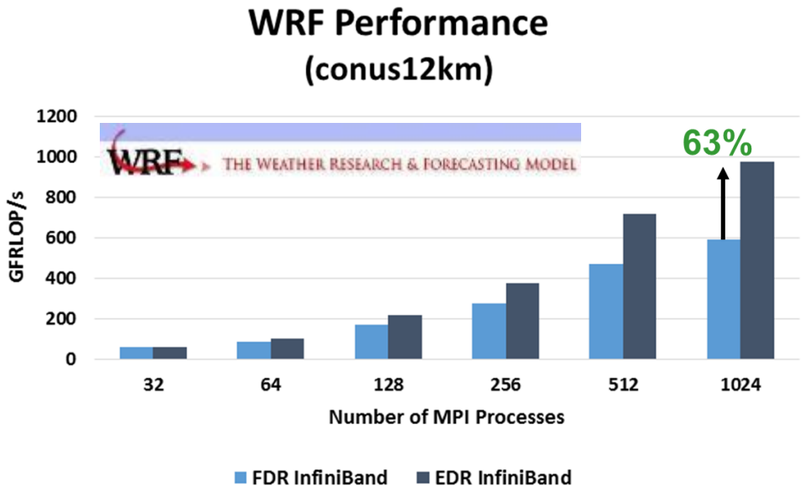
SHArP

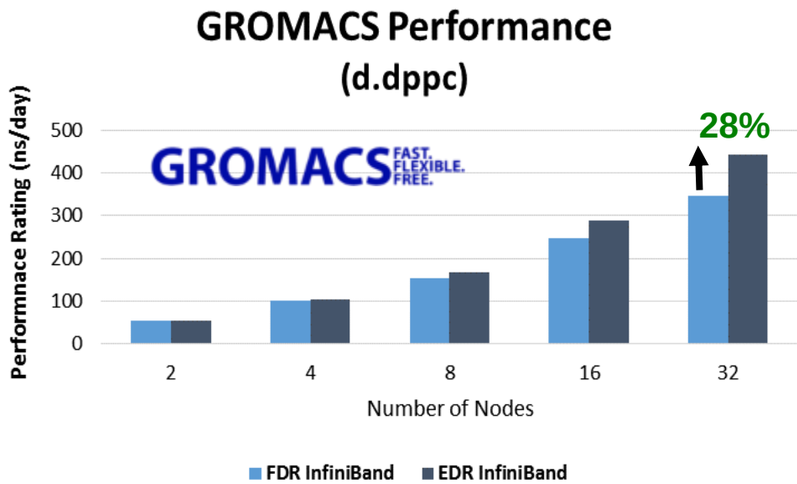
SwitchIB™-2

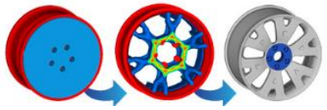
HPC-X with SHArP Delivers 2.2X Higher Performance over Intel MPI

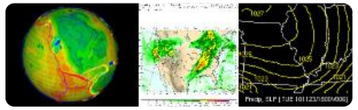
The Performance Advantage of EDR 100G InfiniBand (28-80%)

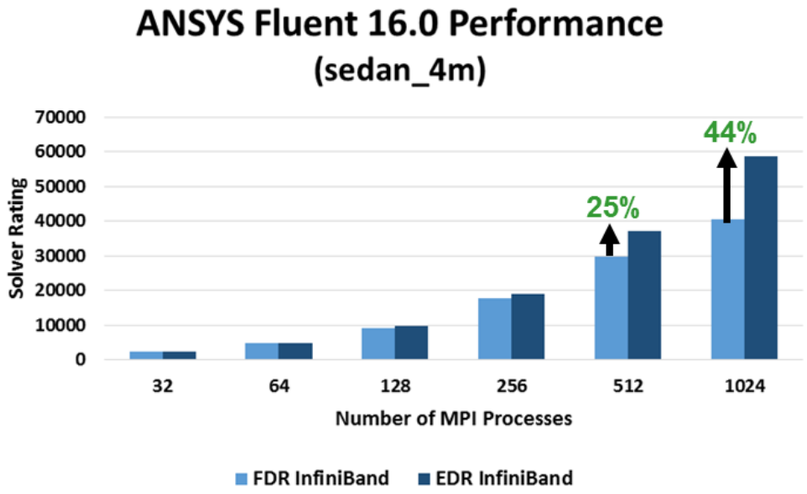


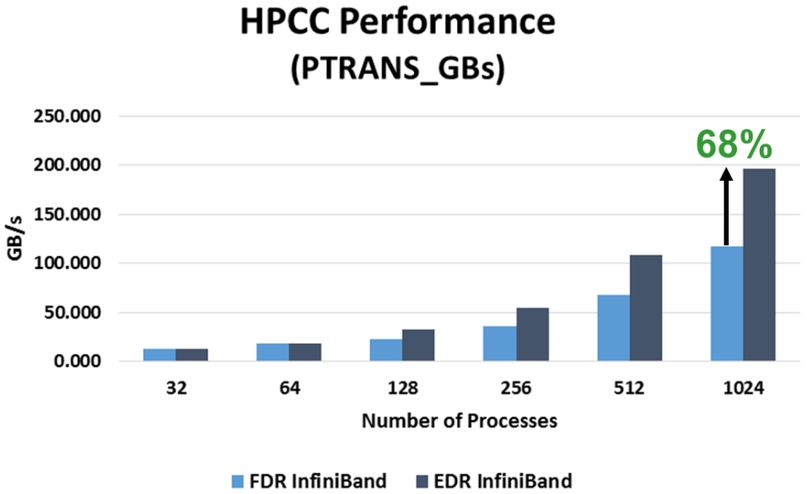


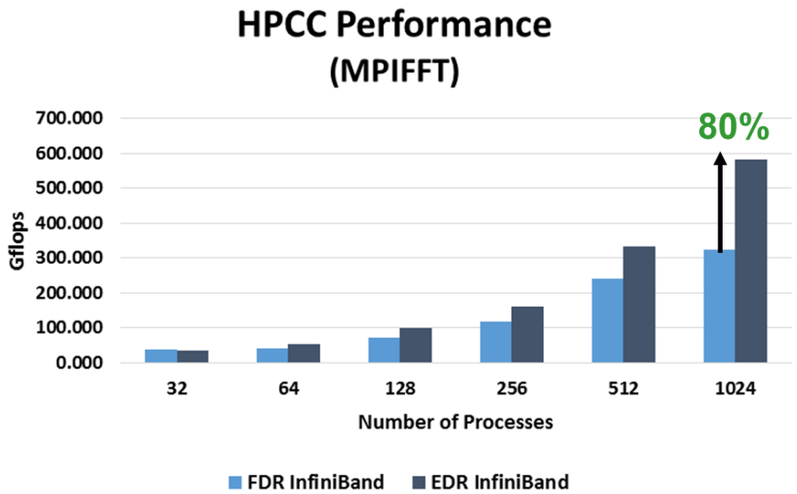












InfiniBand The Smart Choice for HPC Platforms and Applications

- *"We chose a co-design approach. This system was of course targeted at supporting in the best possible manner our key applications. The only interconnect that really could deliver that was Mellanox InfiniBand."*

JÜLICH FORSCHUNGSZENTRUM

Watch Video

- *"One of the big reasons we use InfiniBand and not an alternative is that we've got backwards compatibility with our existing solutions."*

UNIVERSITY OF BIRMINGHAM

Watch Video

- *"InfiniBand is the most advanced high performance interconnect technology in the world, with dramatic communication overhead reduction that fully unleashes cluster performance."*

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Watch Video

- *"InfiniBand is the best that is required for our applications. It enhancing and unlocking the potential of the system."*

CHPC CENTRE FOR HIGH PERFORMANCE COMPUTING

Watch Video

Technology Roadmap – One-Generation Lead over the Competition

Mellanox → 20G → 40G → 56G → 100G → 200G → 400G

Terascale

3rd

TOP500 2003
Virginia Tech (Apple)

Petascale

1st

“Roadrunner”
Mellanox Connected

Exascale

OAK RIDGE National Laboratory
“Summit” System

Lawrence Livermore National Laboratory
“Sierra” System

2000 2005 2010 2015 2020

Offload versus Onload (Non-Offload)

Interconnect Architecture Comparison

Offload versus Onload (Non-Offload)

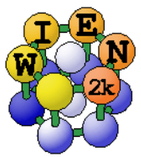
- Two interconnect architectures exist – Offload-based and Onload-based

- Offload Architecture
  - The Interconnect manages and executes all network operations
  - The interconnect is capable of including application acceleration engines
  - Offloads the CPU and therefore free CPU cycles to be used by the applications
  - Development requires large R&D investment
  - Higher data center ROI

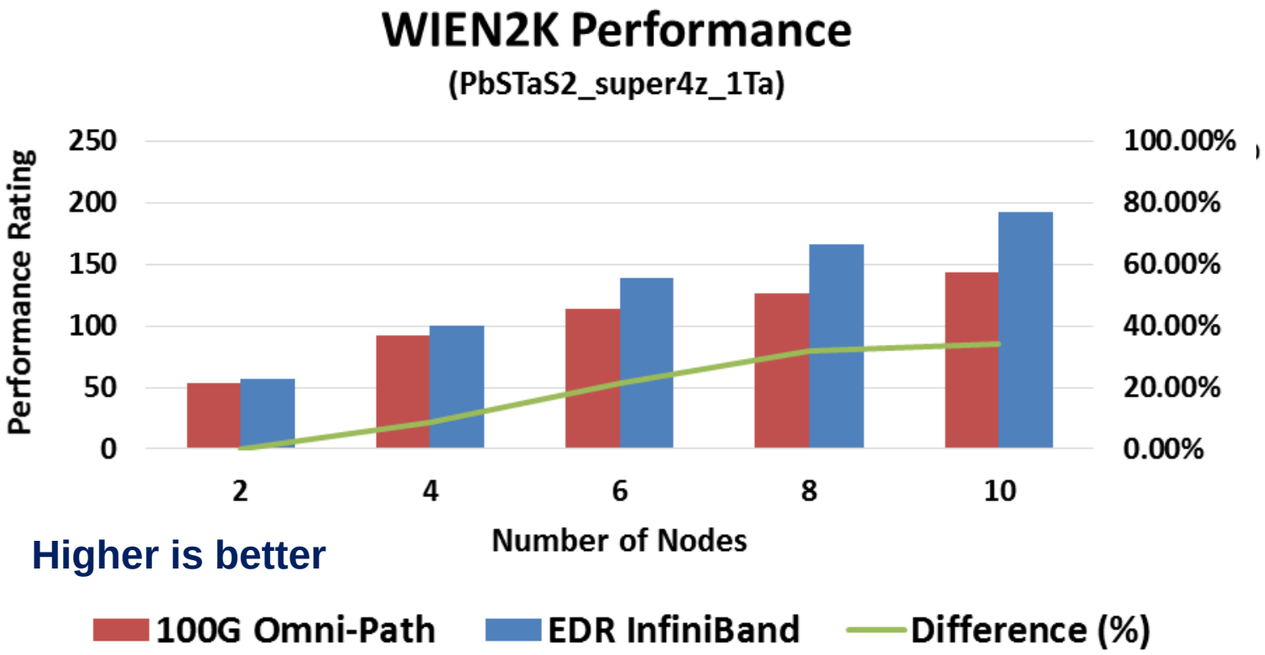
- Onload architecture
  - A CPU-centric approach – everything must be executed on and by the CPU
  - The CPU is responsible for all network functions, the interconnect only pushes the data into the wire
  - Cannot support acceleration engines, no support for RDMA, and network transport is done by the CPU
  - Onload the CPU and reduces the CPU cycles available for the applications
  - Does not require R&D investments or interconnect expertise

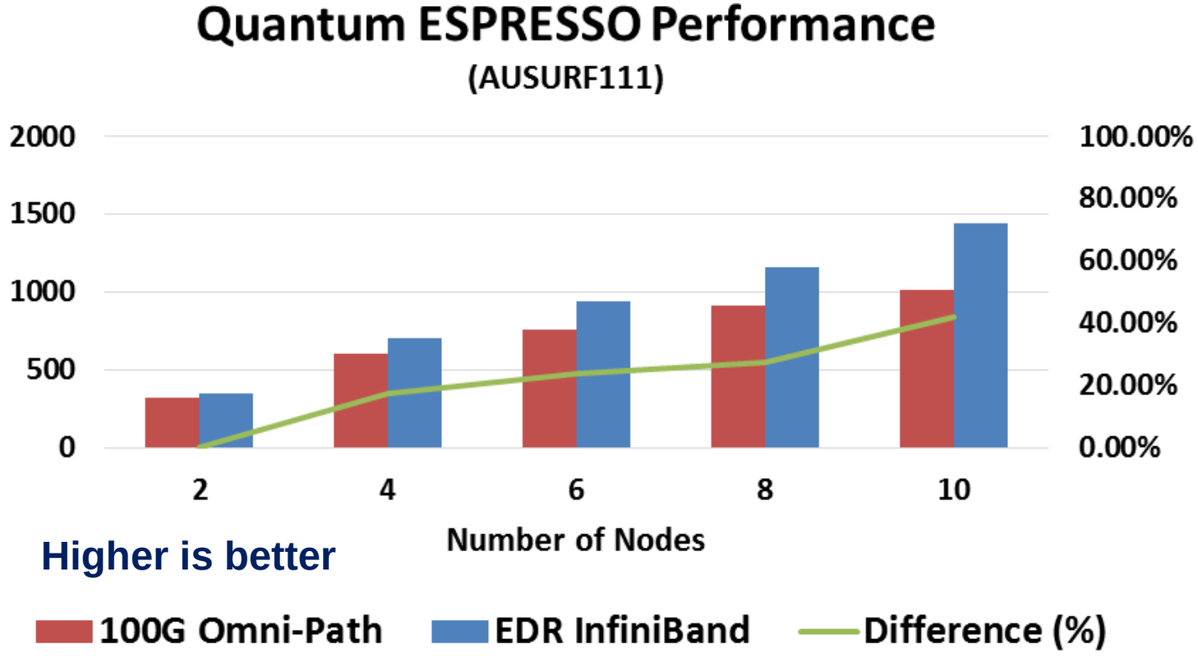
Application Performance Comparison – Quantum ESPRESSO

WIEN2k is a Quantum Mechanical Simulation



Quantum ESPRESSO is an electronic structure and materials modeling Simulation

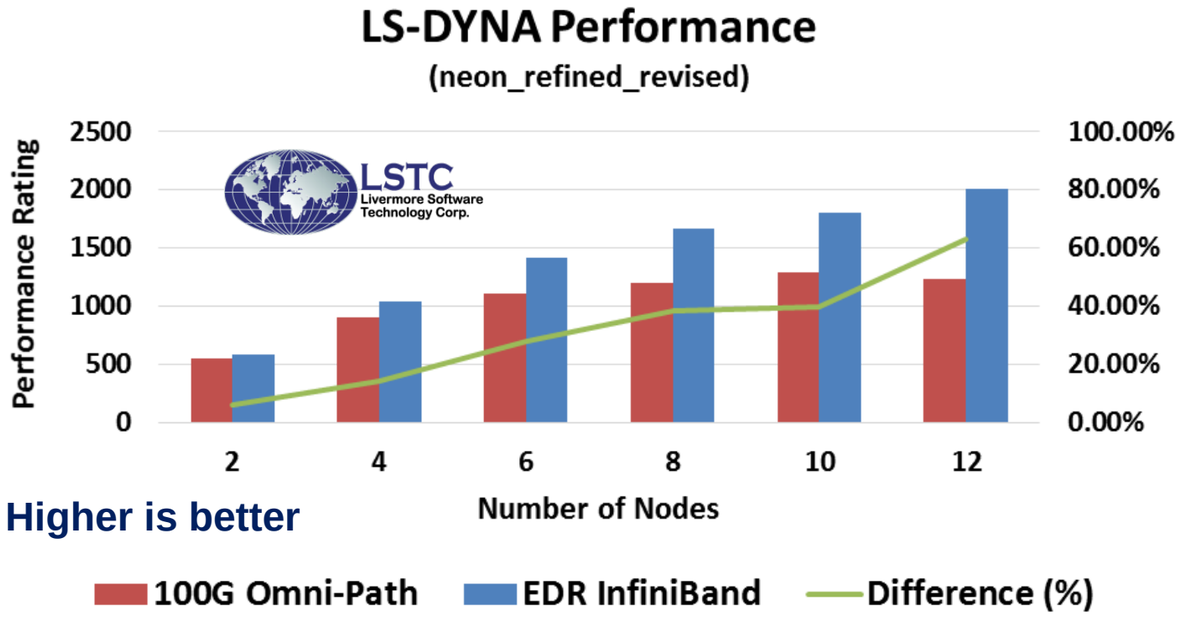


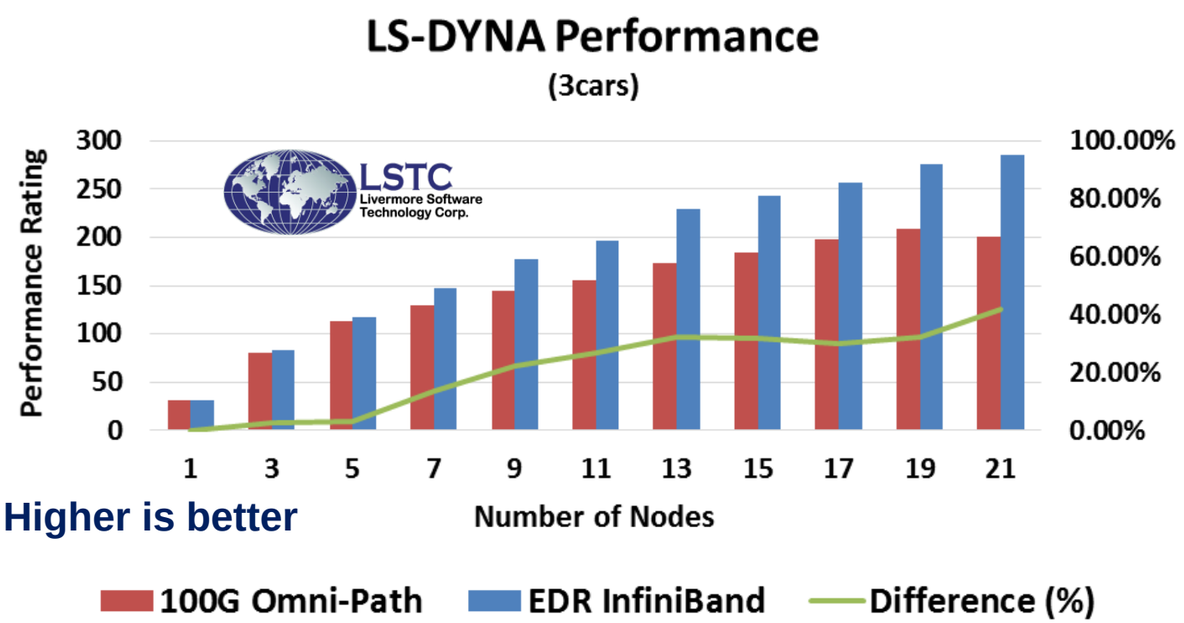


InfiniBand Delivers Higher Performance and Scaling

Application Performance Comparison – LS-DYNA

A structural and fluid analysis software, used for automotive, aerospace, manufacturing simulations and more





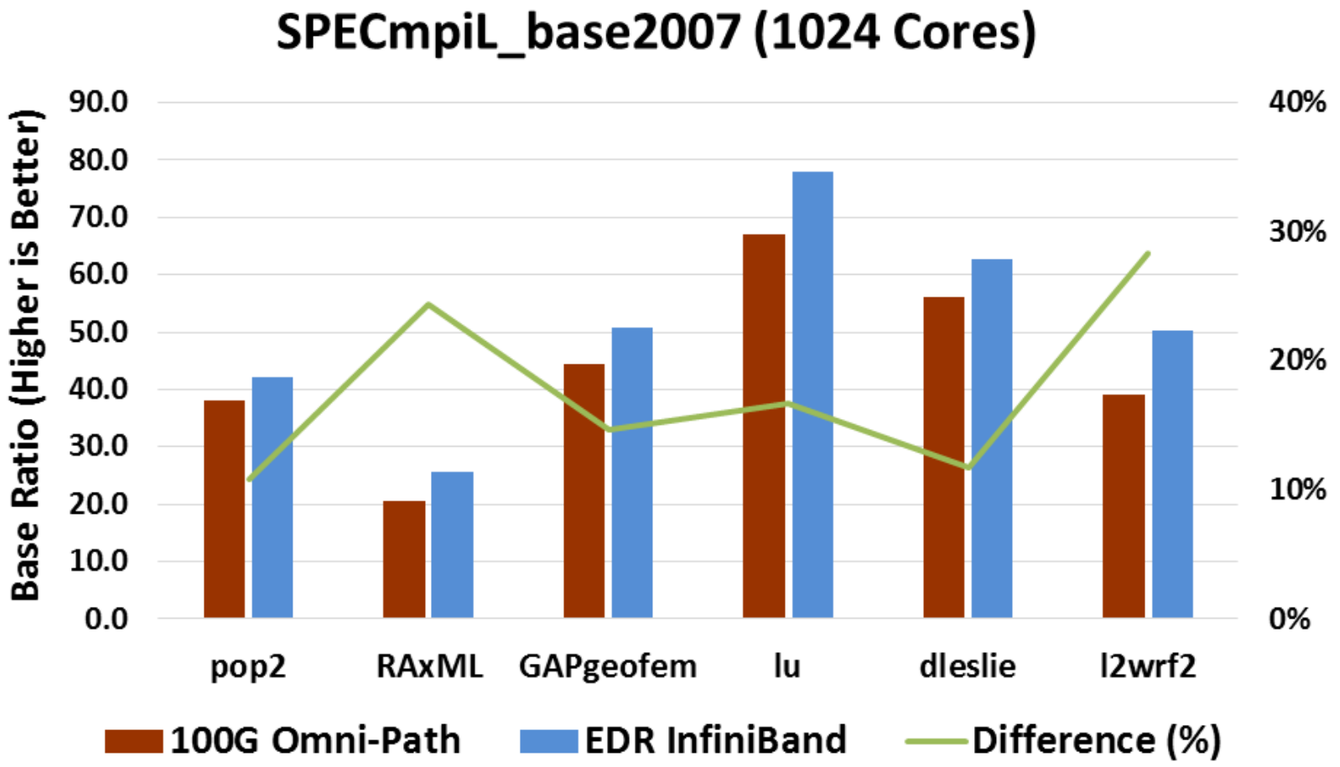
InfiniBand Delivers 42-63% Higher Performance With Only 12 Nodes

Omni-Path Does Not Scale Beyond 10 Nodes

Application Performance Comparison – SPEC MPI Benchmark Suite

The SPEC MPI benchmark suite evaluates MPI-parallel, floating point, compute intensive performance, across a wide range of compute intensive applications using the Message-Passing Interface (MPI)

The Standard Performance Evaluation Corporation (SPEC) is a non-profit corporation formed to establish, maintain and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers



InfiniBand Delivers Superior Performance and Scaling

Offload versus Onload – CPU Overhead

Operation	InfiniBand		Omni-Path	
	CPU Utilization	CPU Frequency at Operation Time	CPU Utilization	CPU Frequency at Operation Time
100Gb/s Data Throughput (Send-Receive)	0.8%	59%	59.6%	100%

Intel Performance Counter Monitor Tool - Output				
	Data Throughput (Gb/s)	AFREQ (relation to nominal CPU frequency while in active state)	CPU Instructions	Active Cycles
InfiniBand	99.5	0.59	39M	163M
Omni-Path	95	1	3725M	12000M

InfiniBand Guarantees Lowest CPU Overhead and Enables OPEX Saving!

Mellanox InfiniBand Leadership

Smart Network For Smart Systems

RDMA, Acceleration Engines, Programmability

Higher Performance

Unlimited Scalability

Higher Resiliency

Proven!

100 Gb/s
Link Speed

2014

200 Gb/s
Link Speed

2017

Gain Competitive Advantage Today
Protect Your Future

Message Rate

68% Higher

Latency

20% Lower

Application Performance

34-62% Higher

Cost/Performance
(CPAR - $/Performance)

20-35% Lower

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™