Paving the Road to Exascale

New Era of Performance through Co-Design
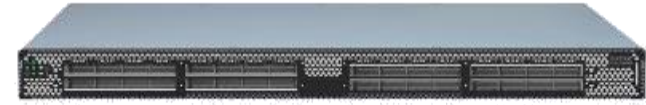
Oklahoma Supercomputing Symposium | 2015

Mellanox® TECHNOLOGIES

Connect. Accelerate. Outperform.™

High-Performance Designed 100Gb/s Interconnect Solutions

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

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
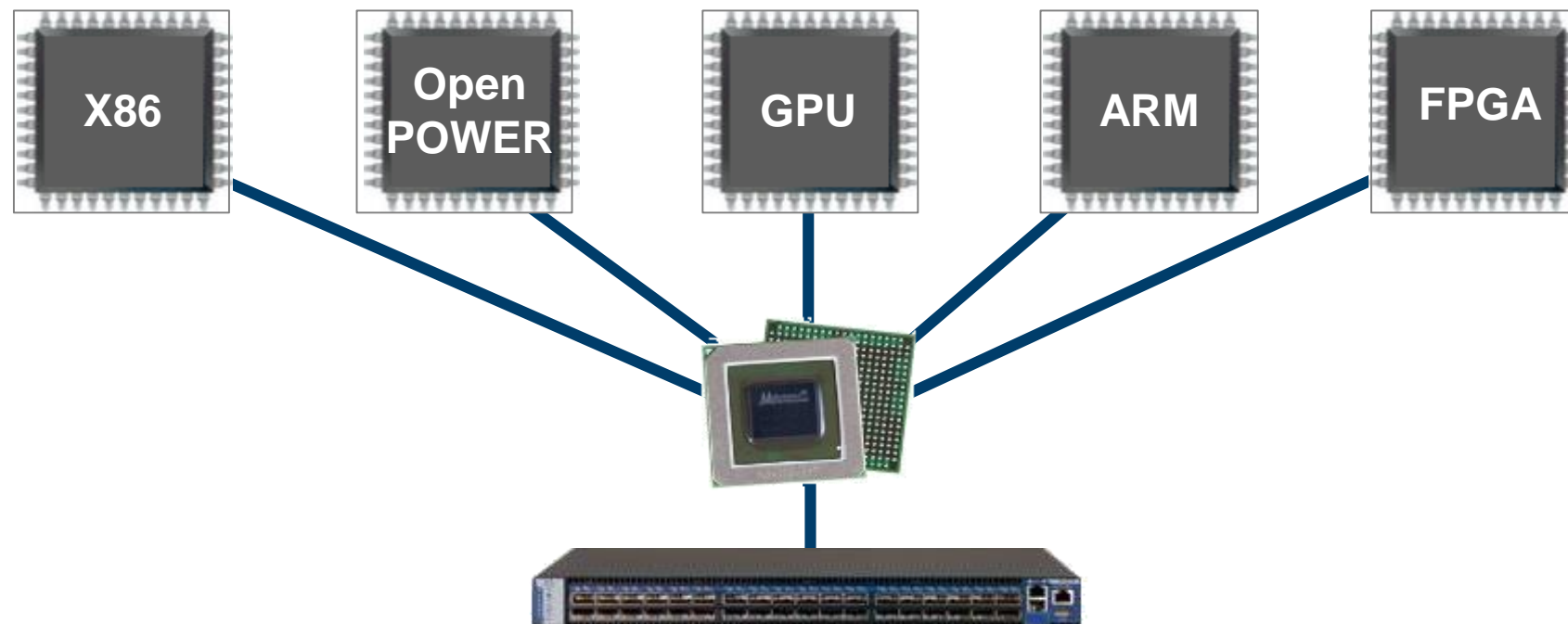
Smart Interconnect to Unleash The Power of All Compute Architectures

Technology Roadmap – One-Generation Lead over the Competition

Mellanox ➔ 20Gbs ➔ 40Gbs ➔ 56Gbs ➔ 100Gbs ➔ 200Gbs ➔
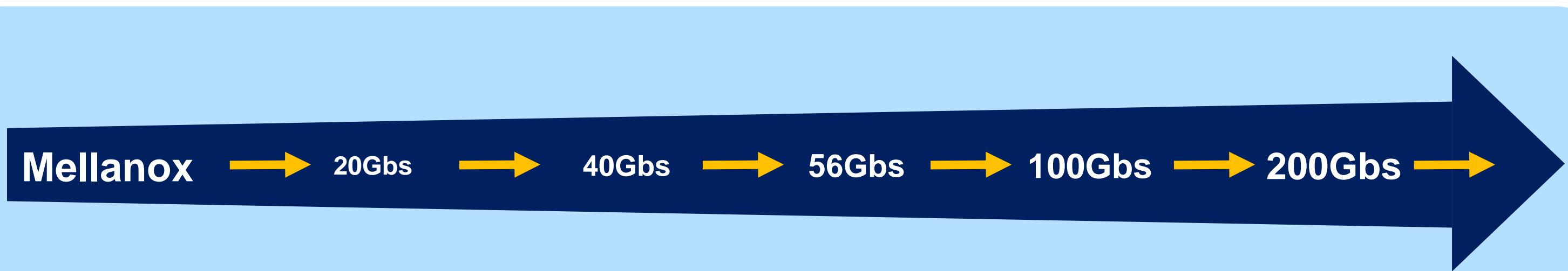
Terascale   **Petascale**   **Exascale**

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**

TOP500 Interconnect Trends
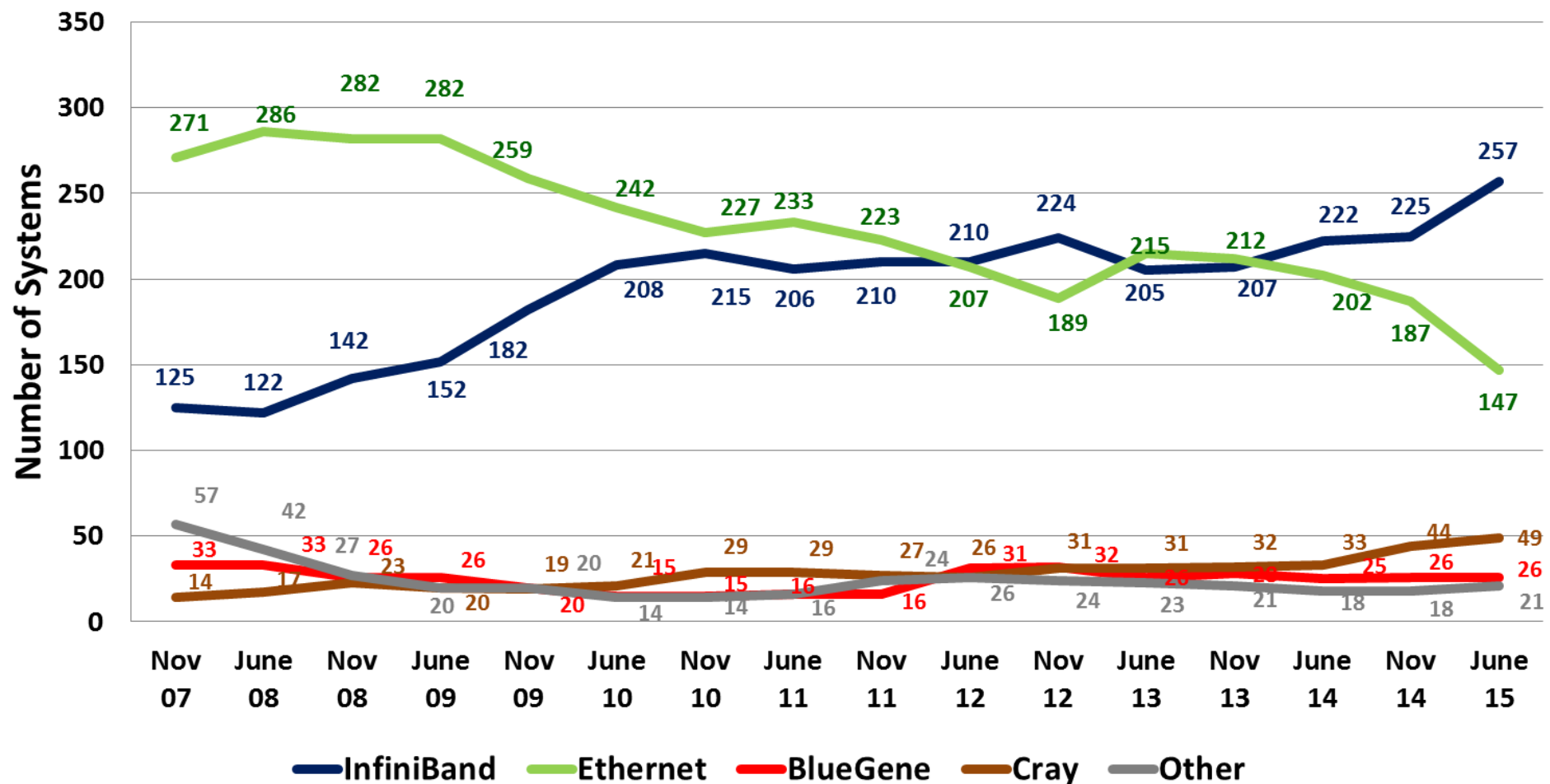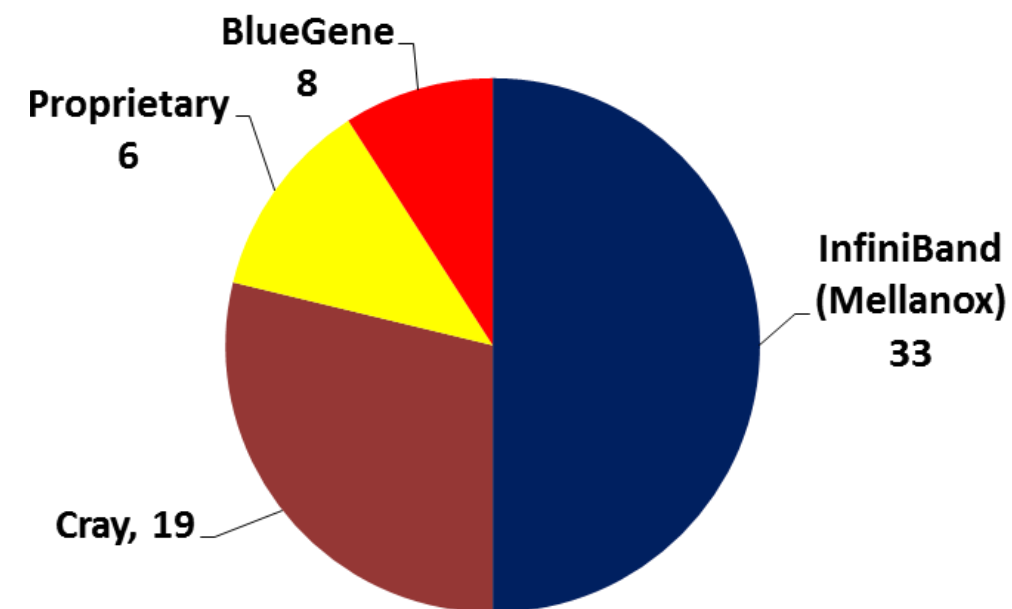
- **InfiniBand is the de-facto interconnect solution for performance demanding applications**

InfiniBand Accelerated TOP500 Systems

Legend: EDR InfiniBand, FDR InfiniBand, QDR/DDR InfiniBand
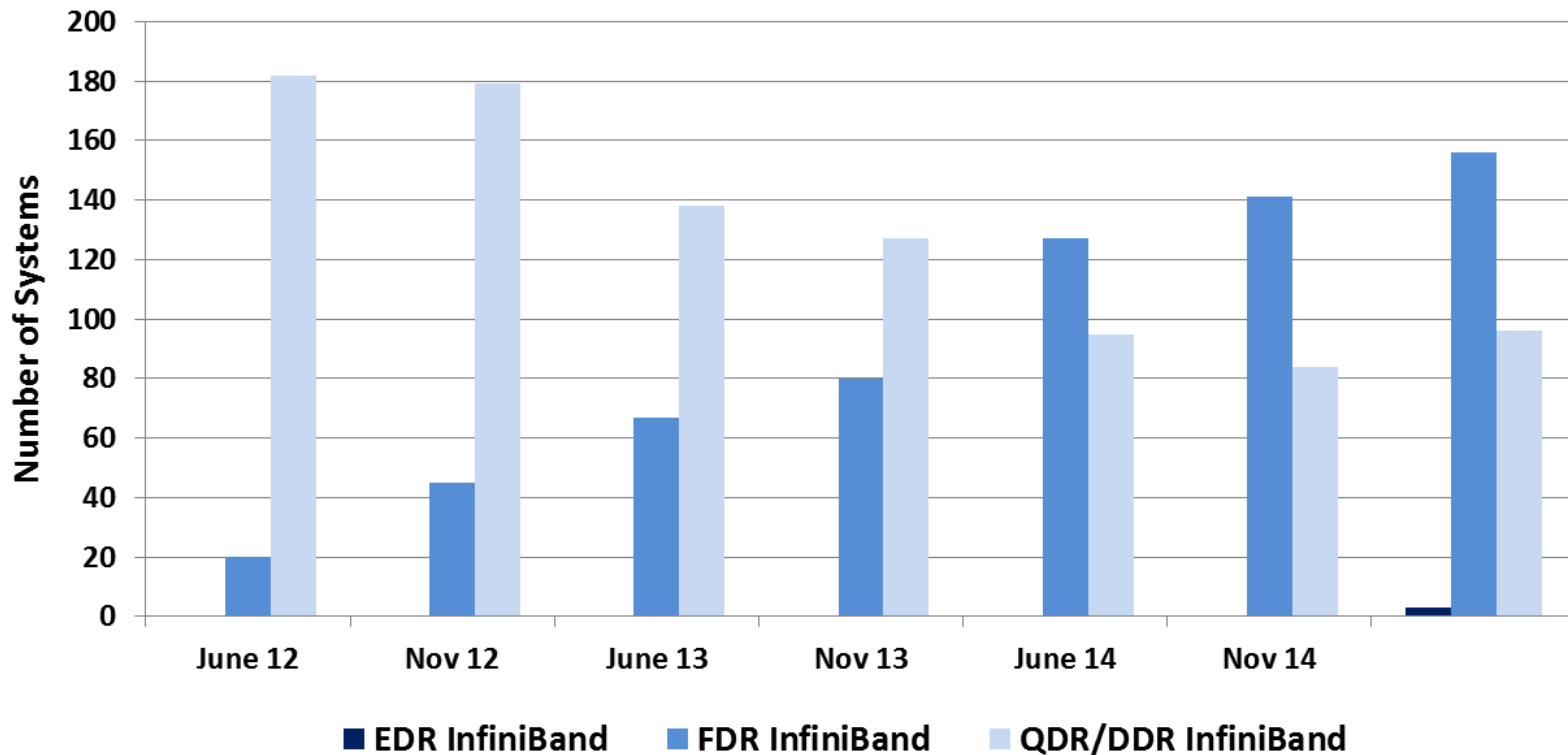
World Leading Compute Systems Efficiency Comparison

Legend: InfiniBand, 1GbE, 10GbE, Cray, GPU

Average Efficiency
- **InfiniBand: 85%**
- **Cray: 74%**
- **10GbE: 66%**
- **GigE: 43%**
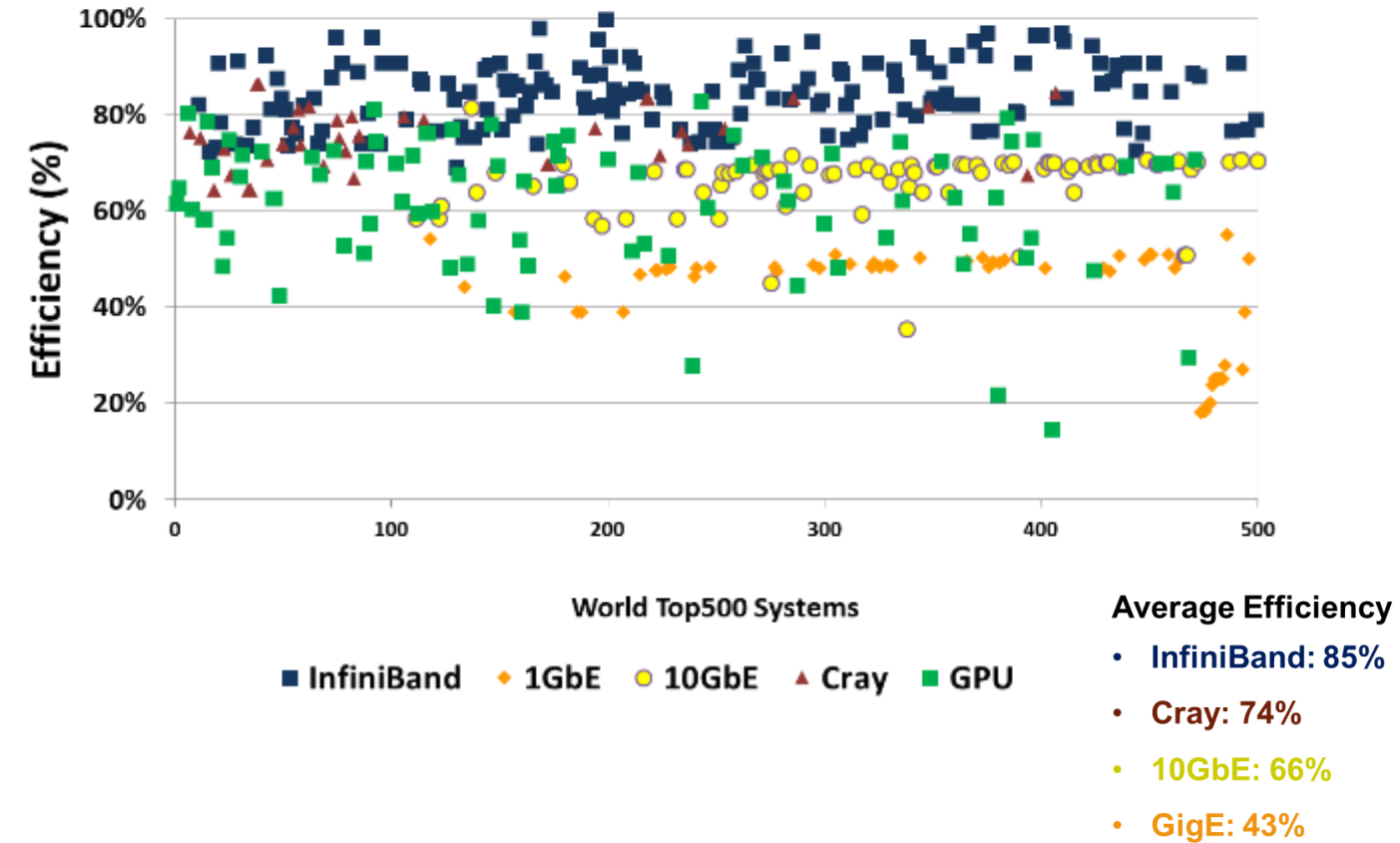
- Number of Mellanox FDR InfiniBand systems grew 23% from June'14 to July'15
- EDR InfiniBand entered the list with 3 systems

Mellanox Accelerated World-Leading HPC Systems

Connecting Half of the World's Petascale Systems (examples)

Cluster

Multi-Core

Co-Design

DATA GROWTH

Applications

Performance

Hardware

Software

Server
Server
Server
Server
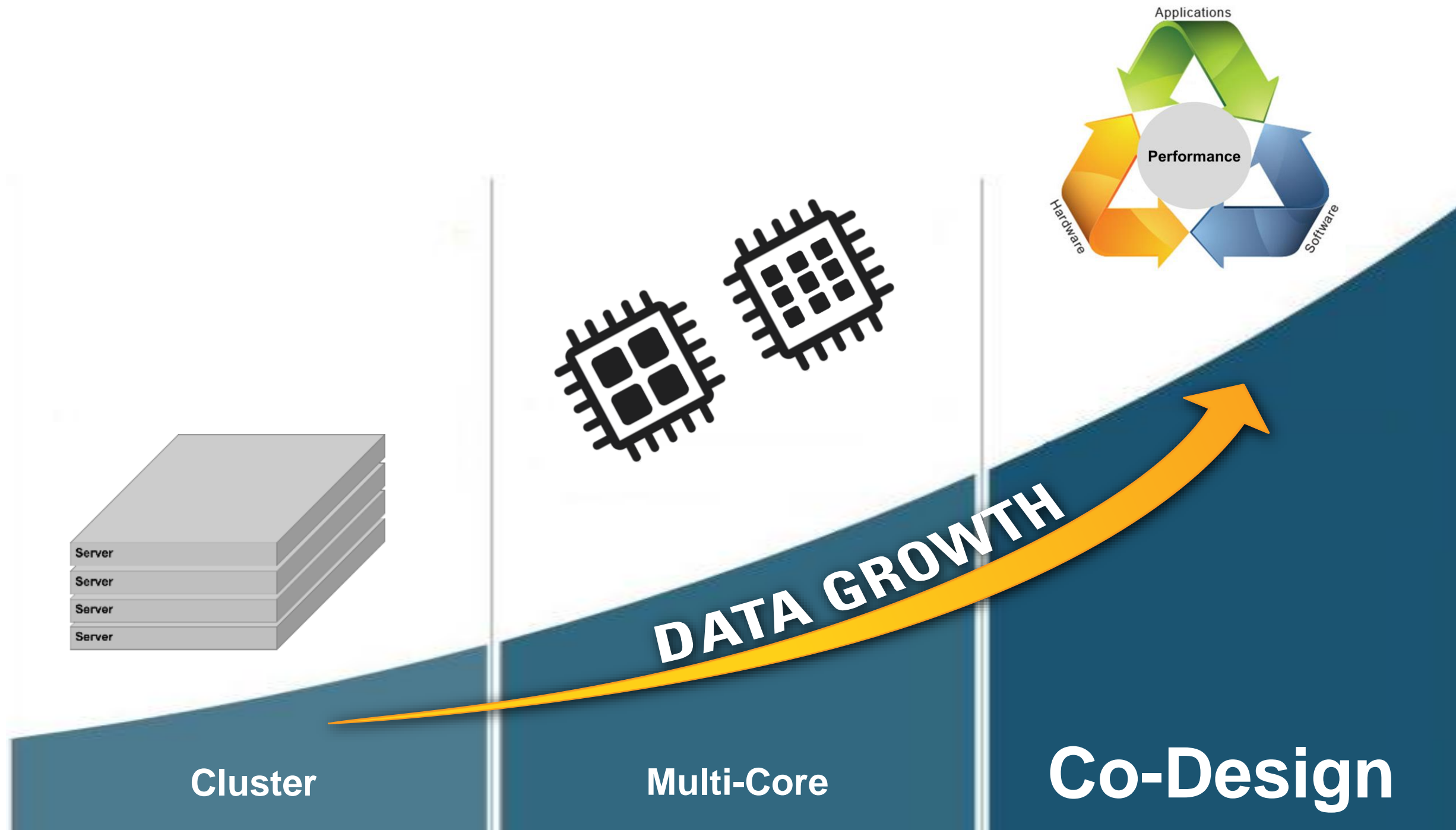
Discrete Level
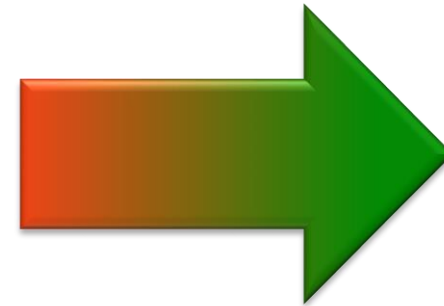
System Level

Software – Hardware

Hardware – Hardware (e.g. GPU-Direct)

Software – Software (e.g. OpenUCX)

Industry – Users – Academia

Standard, Open Source, Eco-System Programmable, Configurable, Innovative

Software-Hardware Co-Design? Example: Breaking the Latency Wall

10 years ago

Network | Communication Framework

~10 microsecond | ~100s microsecond

Today

Network | Communication Framework

~0.1 microsecond | ~10s microsecond

Future

Co-Design Network | Communication Framework

~0.1 microsecond | ~1s microsecond
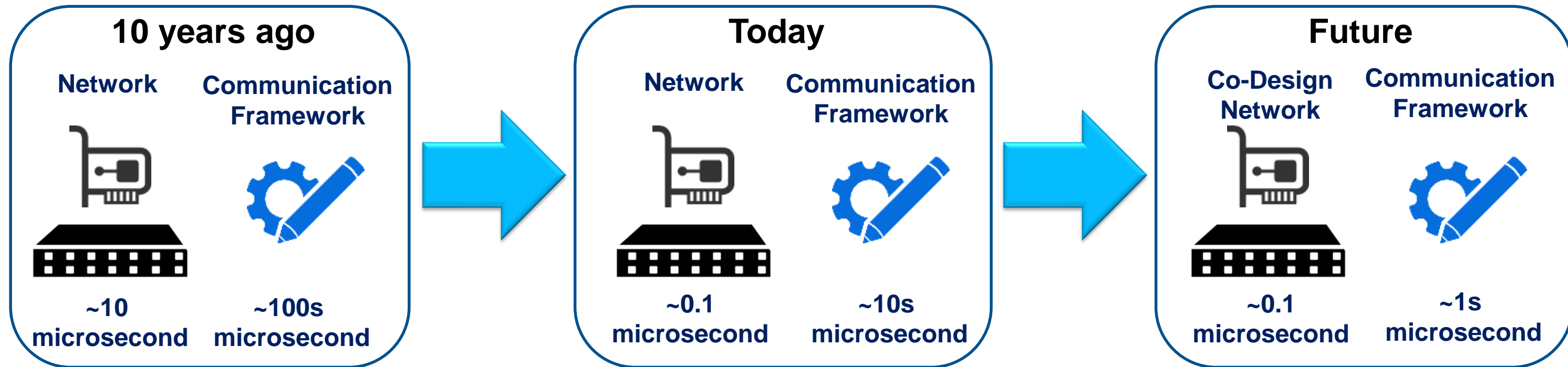
- Today: Network devices are in 100ns latency today
- Challenge: How to enable the next order of magnitude performance improvement?
- Solution: Co-Design - mapping the communication frameworks on all active devices
- Result: reduce HPC communication frameworks latency by an order of magnitude

Co-Design Architecture Paves the Road to Exascale Performance

The Next Generation

HPC Software Framework

Collaborative Effort

Industry, National Laboratories and Academia

Applications

MPI

OpenSHMEM, UPC, CAF, X10, Chapel, etc.

Parsec, OCR, Legions, etc.

Burst buffer, ADIOS, etc.
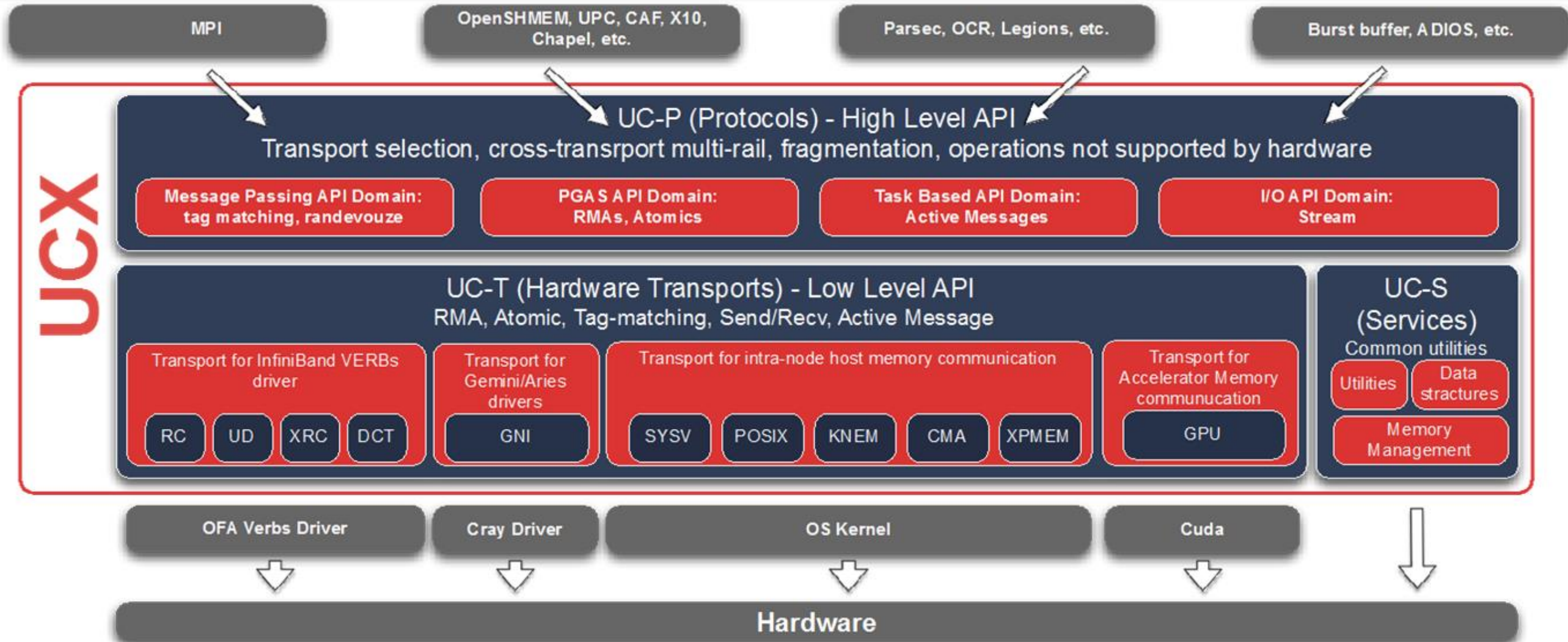
UCX

UC-P (Protocols) - High Level API

Transport selection, cross-transrport multi-rail, fragmentation, operations not supported by hardware

Message Passing API Domain: tag matching, randevouze

PGAS API Domain: RMAs, Atomics

Task Based API Domain: Active Messages

I/O API Domain: Stream

UC-T (Hardware Transports) - Low Level API

RMA, Atomic, Tag-matching, Send/Recv, Active Message

Transport for InfiniBand VERBs driver

RC | UD | XRC | DCT

Transport for Gemini/Aries drivers

GNI

Transport for intra-node host memory communication

SYSV | POSIX | KNEM | CMA | XPMEM

Transport for Accelerator Memory communucation

GPU

UC-S (Services)

Common utilities

Utilities | Data stractures

Memory Management

OFA Verbs Driver | Cray Driver | OS Kernel | Cuda

Hardware

Unified, Light-Weight, High-Performance Communication Framework

- Complete MPI, PGAS OpenSHMEM and UPC package

- Maximize application performance

- For commercial and open source applications
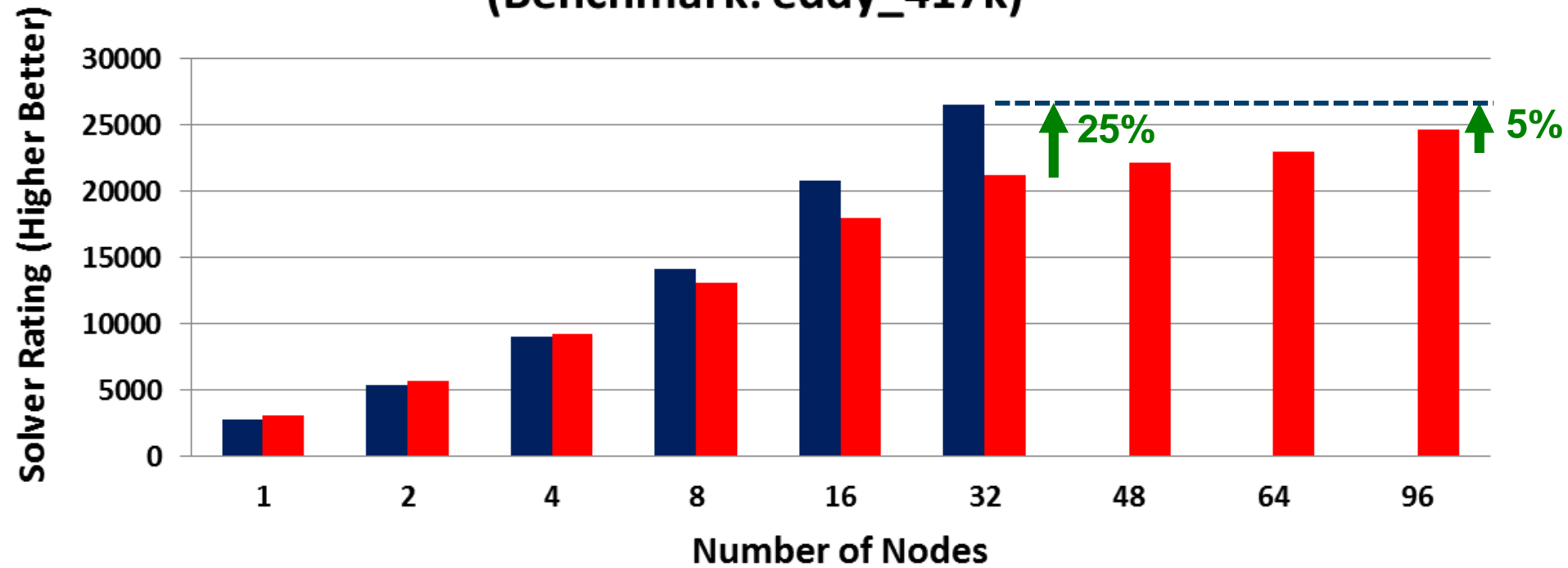
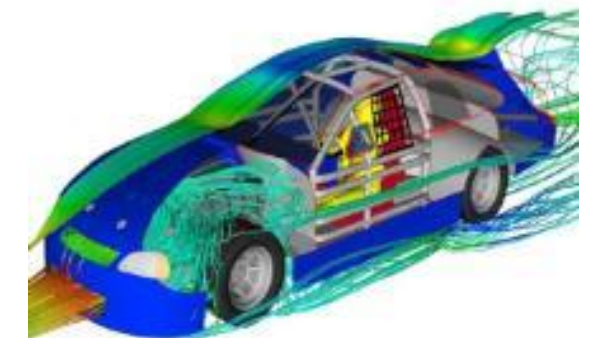- Based on UCX (Unified Communication – X Framework)
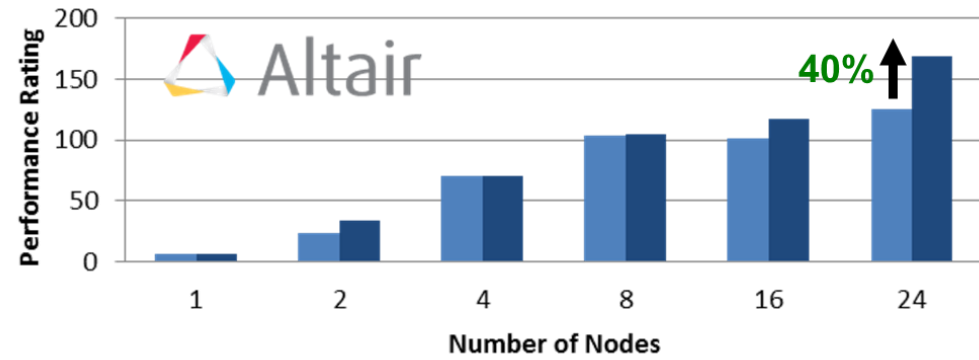
ANSYS Fluent 15.0.7 Performance
(Benchmark: eddy_417k)

Dell PowerEdge, FDR InfiniBand, Intel Xeon 2680v2 at 2.8GHz
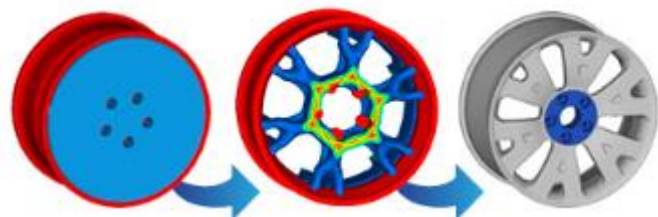Cray XC30, Aries interconnect, Intel Xeon 2690v2 at 3GHz
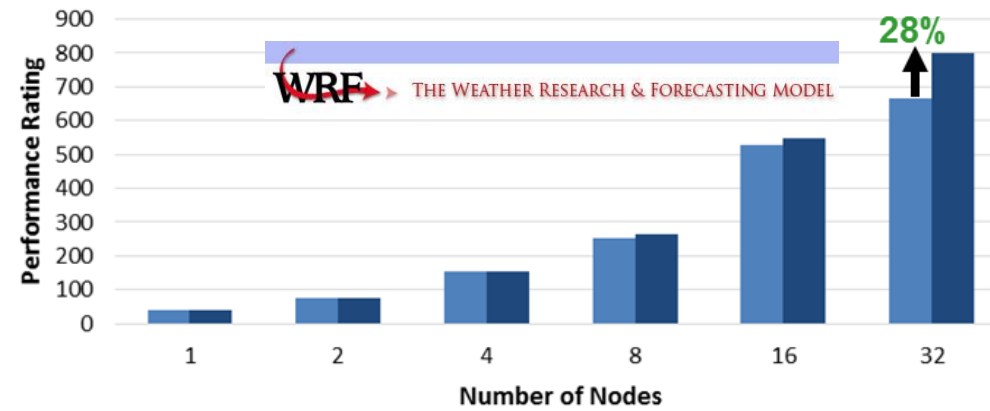
The Performance Power of Co-Design
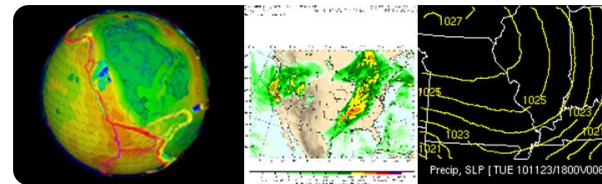

OptiStruct Performance (Engine_Assy.fem) — Performance Rating vs Number of Nodes, FDR InfiniBand vs EDR InfiniBand, 40% increase at 24 nodes.

WRF Performance (conus12km) — Performance Rating vs Number of Nodes, FDR InfiniBand vs EDR InfiniBand, 28% increase at 32 nodes.

GROMACS Performance (d.dppc) — Performance Rating (ns/day) vs Number of Nodes, FDR InfiniBand vs EDR InfiniBand, 28% increase at 32 nodes.
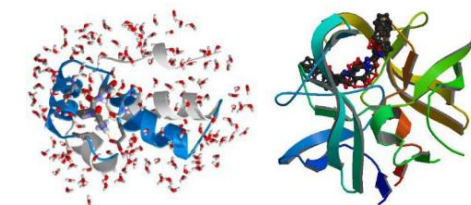
40% Increase in Overall System Performance

System: $315K, InfiniBand FDR: $35K (11%), InfiniBand EDR: $52K (16%)

For 5% System Investment ($17K), 40% Performance Increase! ($126K value, 30% - $95K)

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 is leading HPC software based on UCX, which provides leading performance for MPI, OpenSHMEM/PGAS and UPC

- Exascale will be Enabled via Co-Design Architecture

- Only Mellanox provides superior applications offloads: RDMA, Collectives, scalable transport

- Backward and future compatible

Thank You

Mellanox® TECHNOLOGIES

Connect. Accelerate. Outperform.™