Paving the Road to Exascale

Gilad Shainer

August 2015 , MVAPICH User Group (MUG) Meeting

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

The Ever Growing Demand for Performance

Performance

Terascale

Petascale

1st

"Roadrunner"

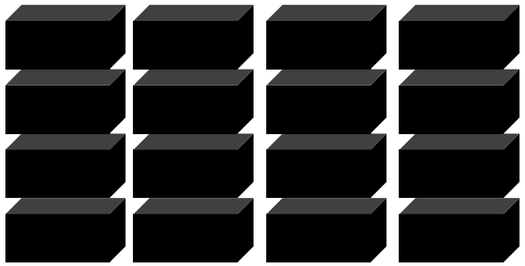
Exascale

OAK RIDGE National Laboratory
"Summit" System

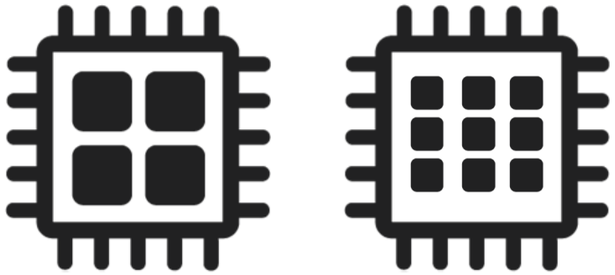
Lawrence Livermore National Laboratory
"Sierra" System

2000 2005 2010 2015 2020

Technology Development

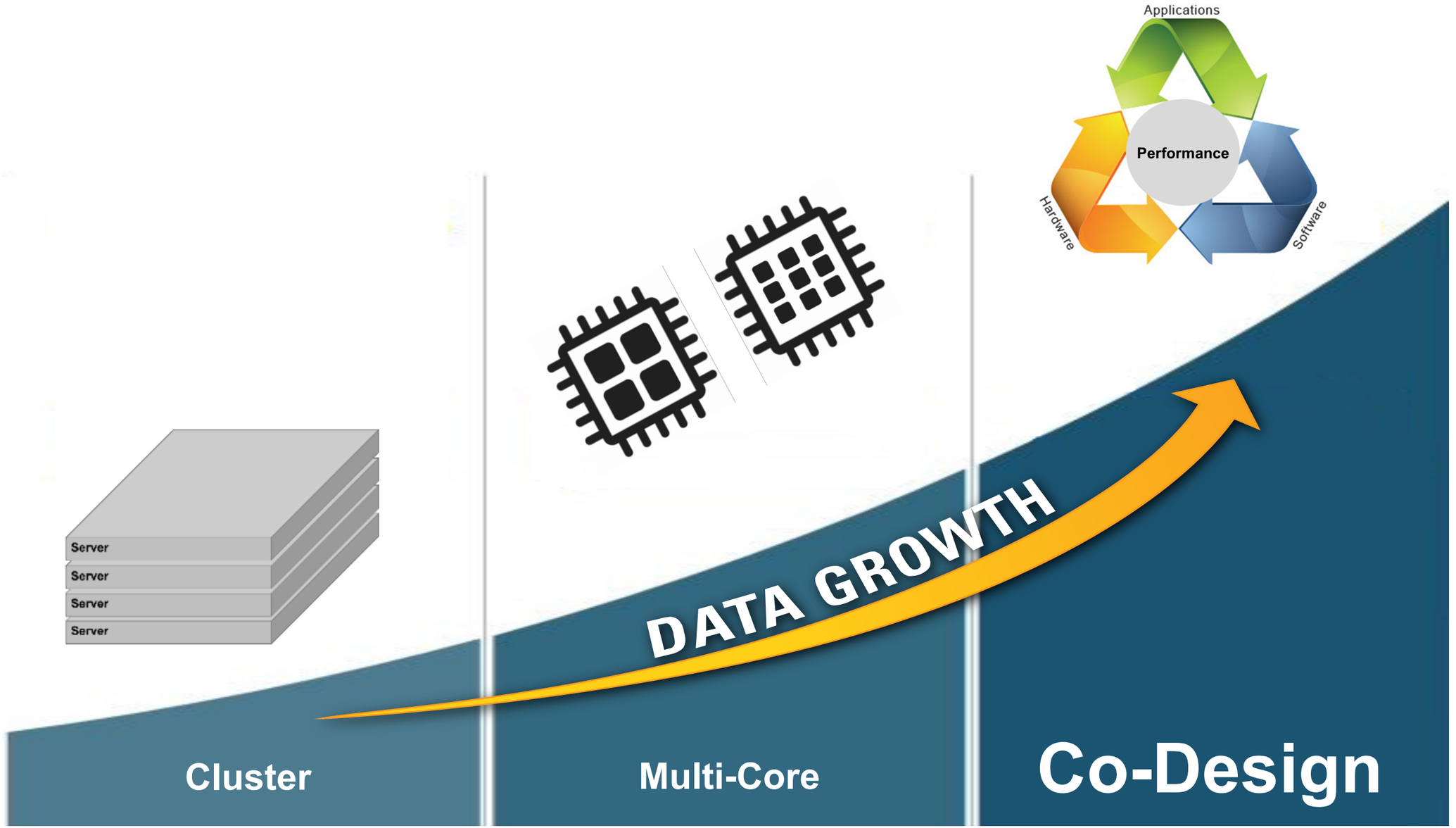


SMP to Clusters



Single-Core to Multi-Core

The Road to Exascale Computing



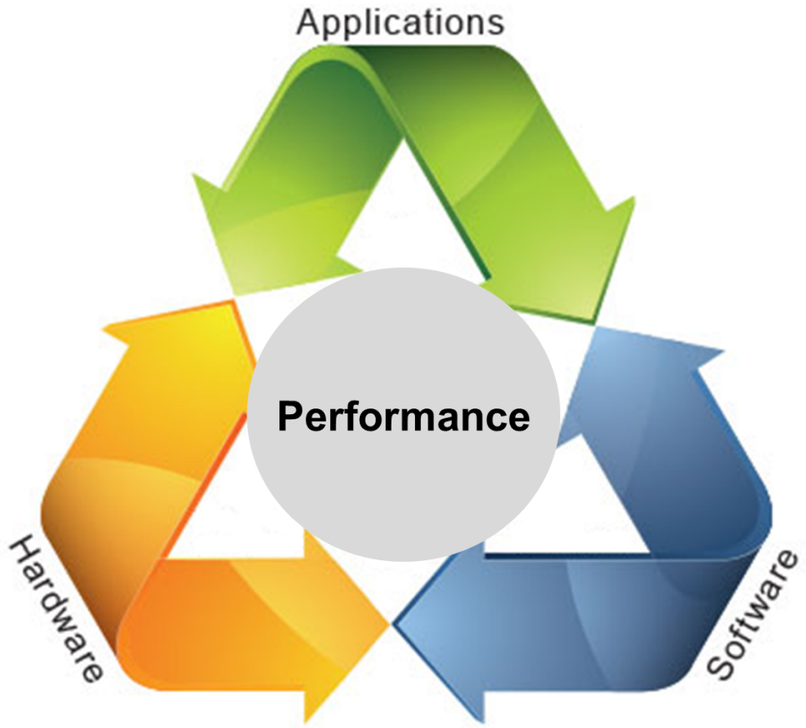
Co-Design Architecture – From Discrete to System Focused

Discrete Level

System Level

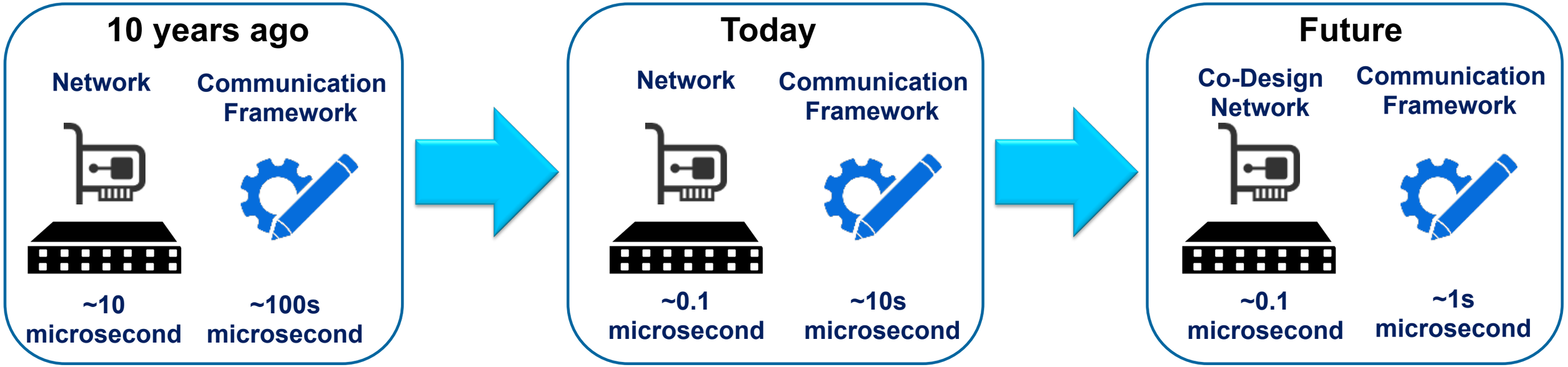
Exascale will be Enabled via Co-Design Architecture

- 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

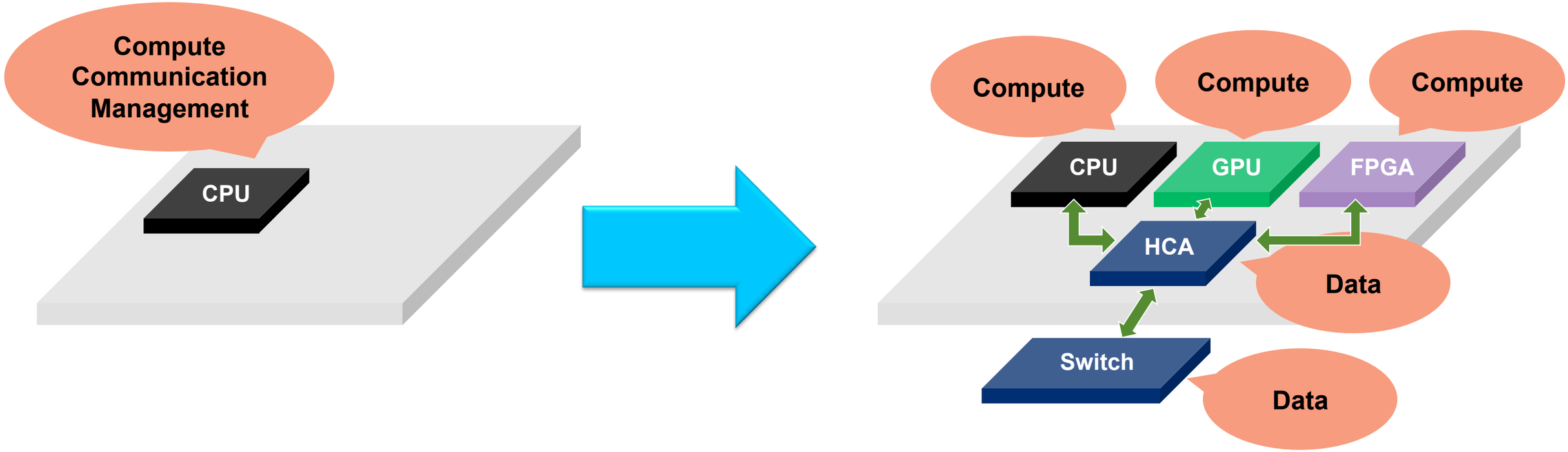


- Today: Network devices are in 100ns latency today
- Challenge: How to enable the next order of magnitude performance improvement?
- Solution: Co-Design - mapping the communication frameworks on all active devices
- Result: reduce HPC communication frameworks latency by an order of magnitude

Co-Design Architecture Paves the Road to Exascale Performance

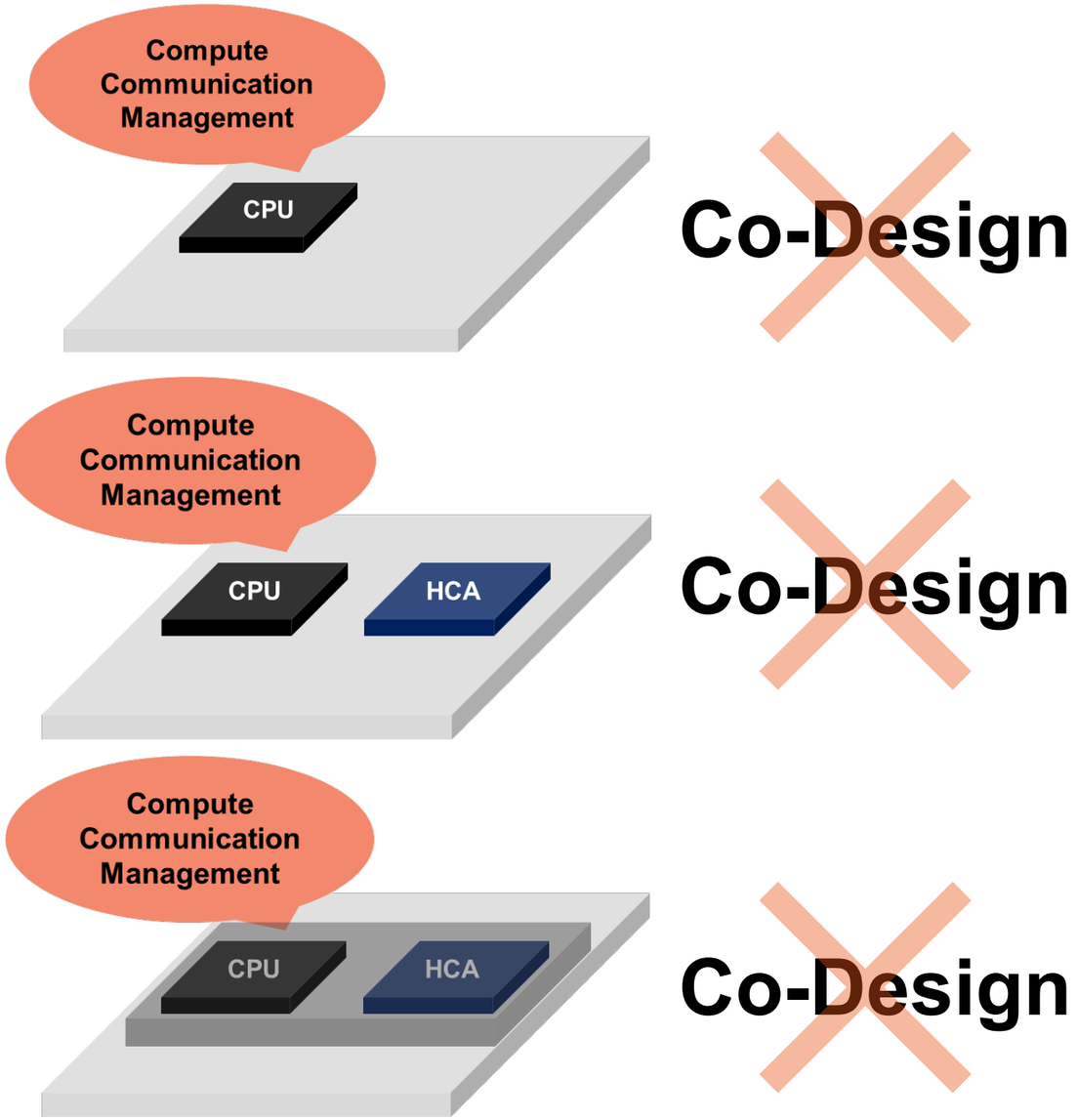
The Road to Exascale – Co-Design System Architecture

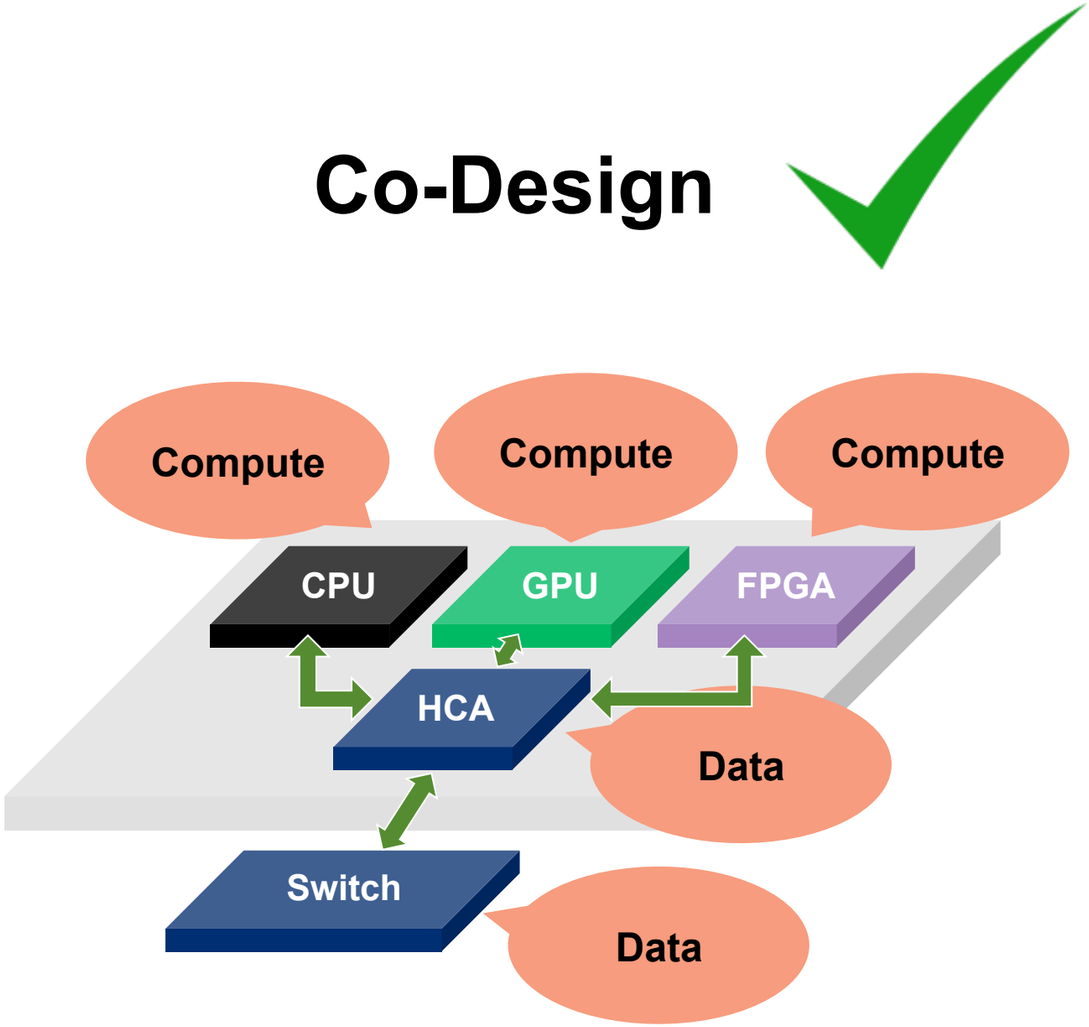
- The road to Exascale requires order of magnitude performance improvements
- Co-Design architecture enables all active devices to become co-processors



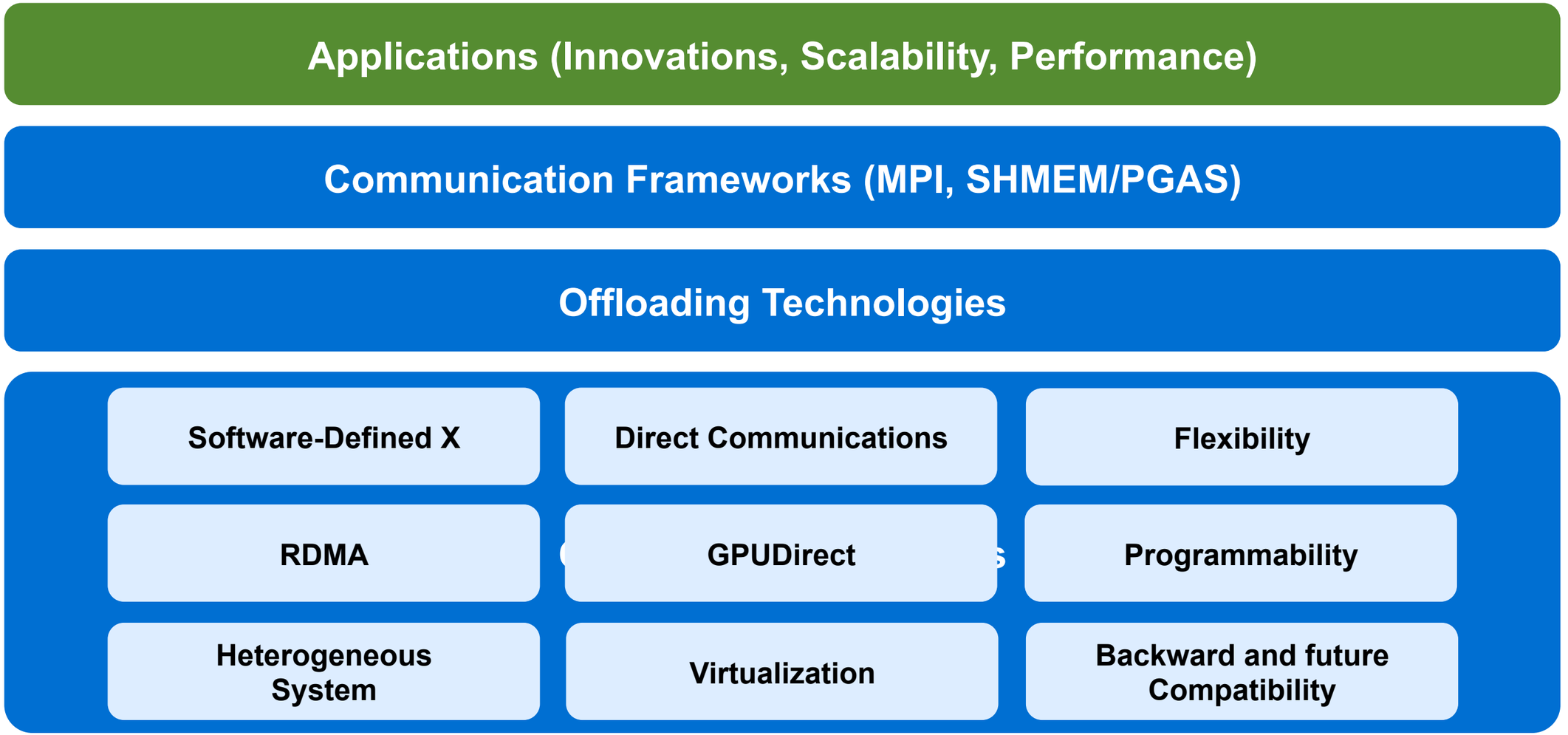
Mapping Communication Frameworks on All Active Devices

Co-Design Architecture



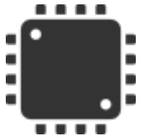


The Elements of the Co-Design Architecture



Mellanox Co-Design Architecture (Collaborative Effort)

Communication Frameworks (MPI, SHMEM/PGAS)



Applications

Transport
RDMA
SRIOV
Collectives
Peer-Direct
GPU-Direct
More…

Future

The Only Approach to Deliver 10X Performance Improvements

Mellanox InfiniBand Proven and Most Scalable HPC Interconnect

OAK RIDGE National Laboratory

"Summit" System

U.S. DEPARTMENT OF ENERGY

Lawrence Livermore National Laboratory

"Sierra" System

Paving the Road to Exascale

High-Performance Designed 100Gb/s Interconnect Solutions

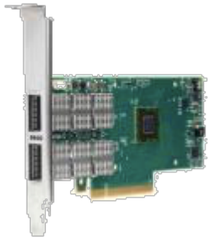
Adapters

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

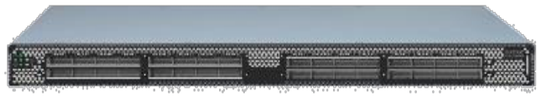
Switch

Spectrum™

32 100GbE Ports, 64 25/50GbE Ports

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

Throughput of 6.4Tb/s

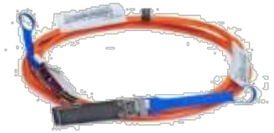


Interconnect

LinkX™

Copper (Passive, Active)

Optical Cables (VCSEL)

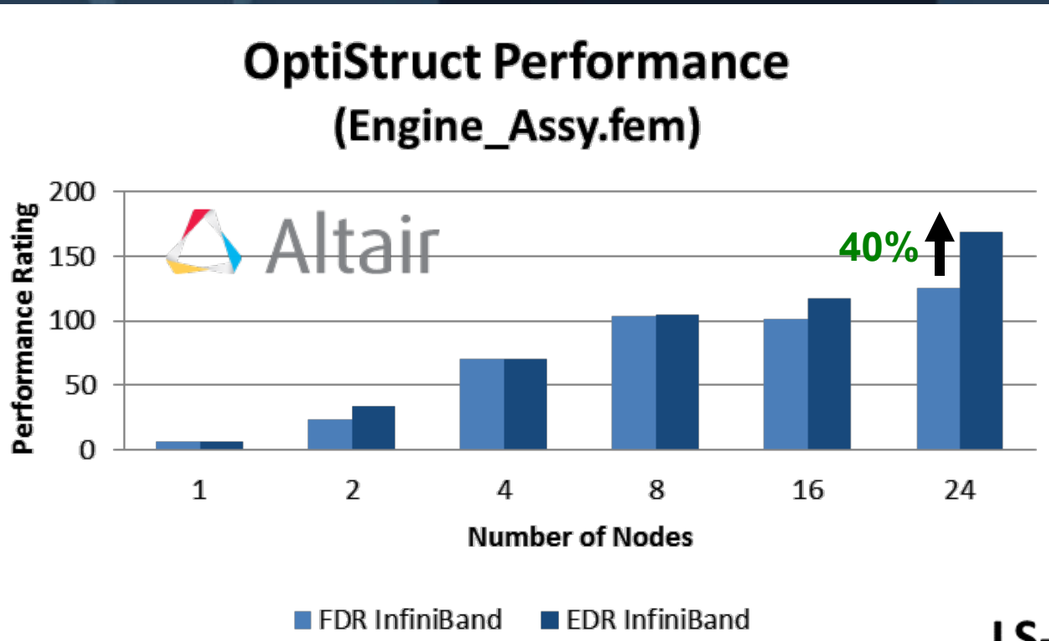


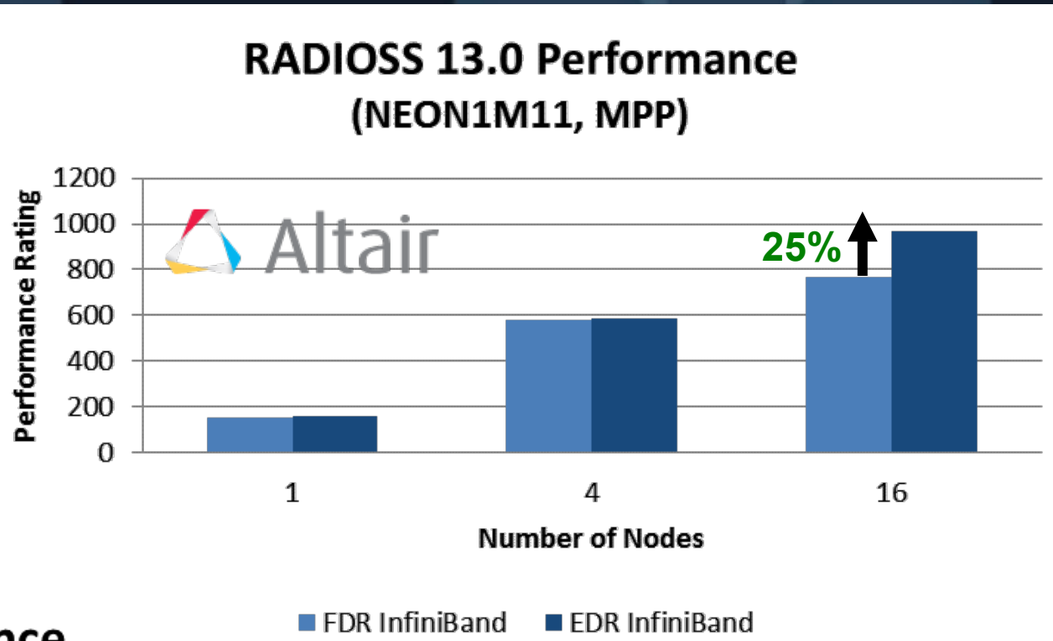
Silicon Photonics

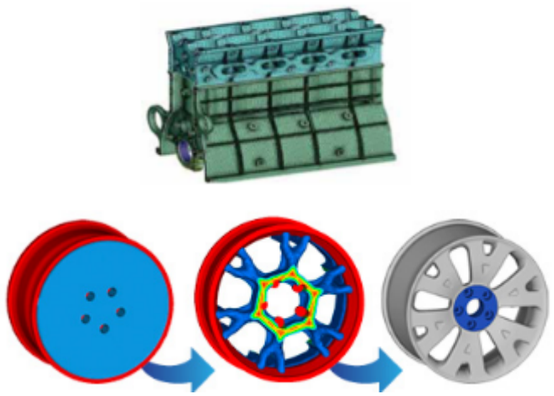
InfiniBand Adapters Performance Comparison

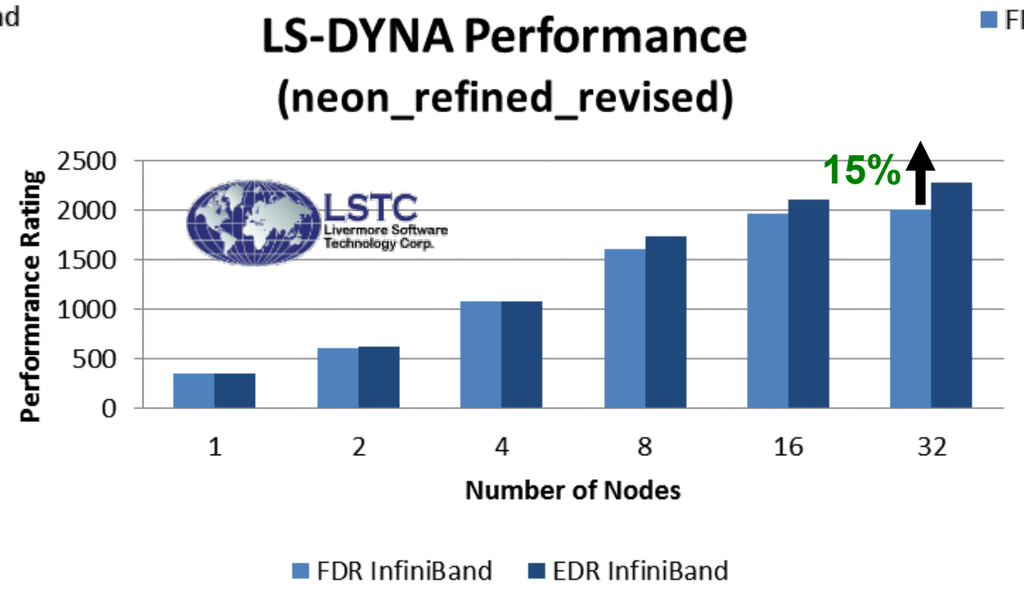
Mellanox Adapters Single Port Performance	ConnectX-4 EDR 100G	Connect-IB FDR 56G	ConnectX-3 Pro FDR 56G
Uni-Directional Throughput	100 Gb/s	54.24 Gb/s	51.1 Gb/s
Bi-Directional Throughput	195 Gb/s	107.64 Gb/s	98.4 Gb/s
Latency	0.61 us	0.63 us	0.64 us
Message Rate (Uni-Directional)	149.5 Million/sec	105 Million/sec	35.9 Million/sec

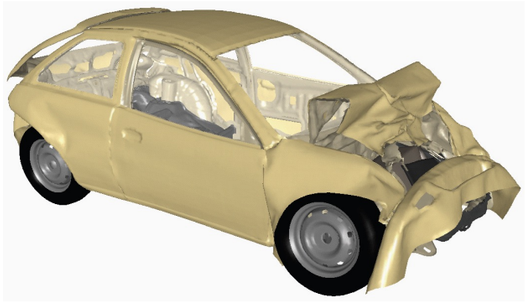
EDR InfiniBand Performance – Commercial Applications





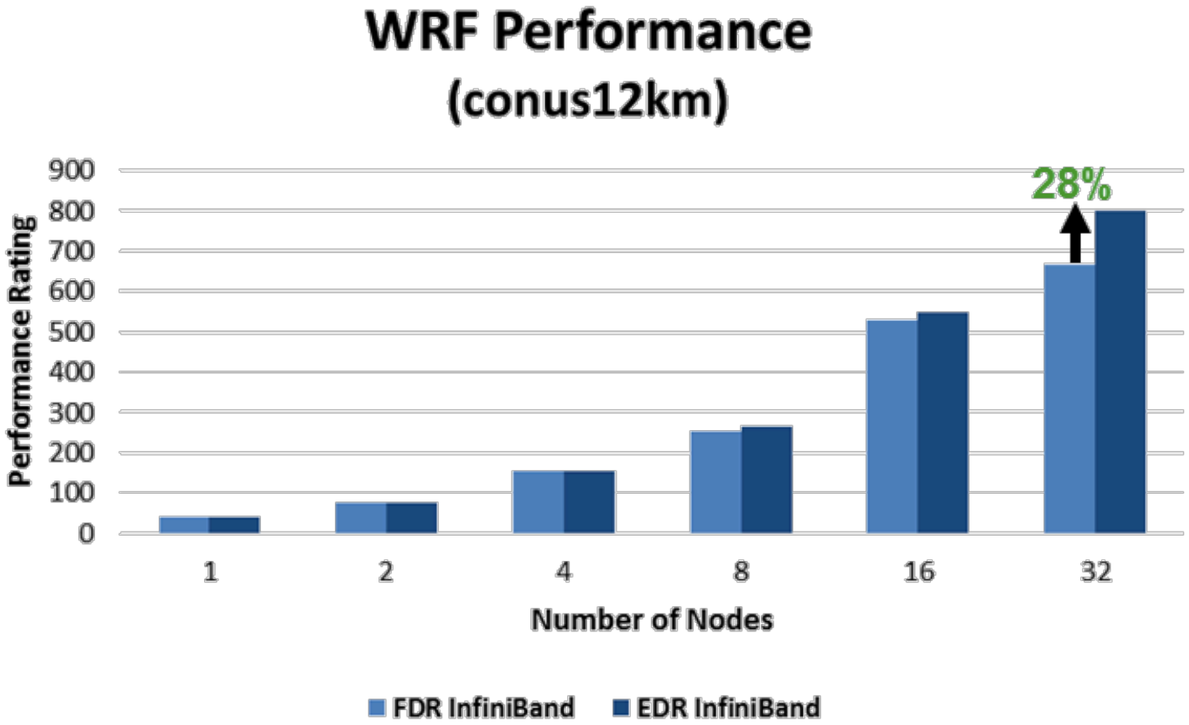




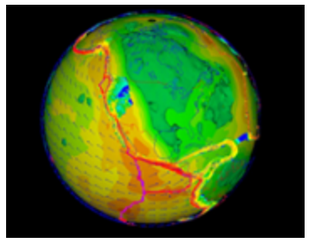


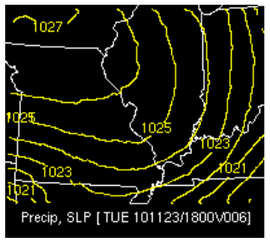
EDR InfiniBand Performance – Weather Simulation

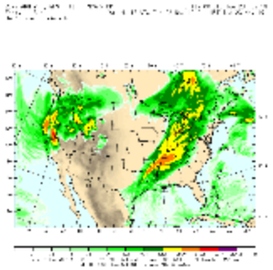
- Weather Research and Forecasting Model
- Optimization effort with the HPCAC
- EDR InfiniBand delivers 28% higher performance
  - 32-node cluster
  - Performance advantage increase with system size



WRF THE WEATHER RESEARCH & FORECASTING MODEL







HPC ADVISORY COUNCIL

Lenovo EDR InfiniBand System (TOP500)

- “LENOX“ EDR InfiniBand connected system at the Lenovo HPC innovation center

- EDR InfiniBand provides >20% higher performance versus over FDR on Graph500
  - At 128nodes

Unified Communication – X Framework (UCX)

The Next Generation HPC Software Framework
To Meet the Needs of Future Systems / Applications

The Next Generation HPC Software Framework

Collaborative Effort

Industry, National Laboratories and Academia

UCX Framework Mission

- Collaboration between industry, laboratories, and academia
- Create open-source production grade communication framework for HPC applications
- To enable the highest performance through co-design of software-hardware interfaces
- To unify industry - national laboratories - academia efforts

API

Exposes broad semantics that target data centric and HPC programming models and applications

Performance oriented

Optimization for low-software overheads in communication path allows near native-level performance

Production quality

Developed, maintained, tested, and used by industry and researcher community

Community driven

Collaboration between industry, laboratories, and academia

Research

The framework concepts and ideas are driven by research in academia, laboratories, and industry

Cross platform

Support for Infiniband, Cray, various shared memory (x86-64 and Power), GPUs

Co-design of Exascale Network APIs

The UCX Framework

UC-S for Services

This framework provides basic infrastructure for component based programming, memory management, and useful system utilities

Functionality:
Platform abstractions and data structures

UC-T for Transport

Low-level API that expose basic network operations supported by underlying hardware

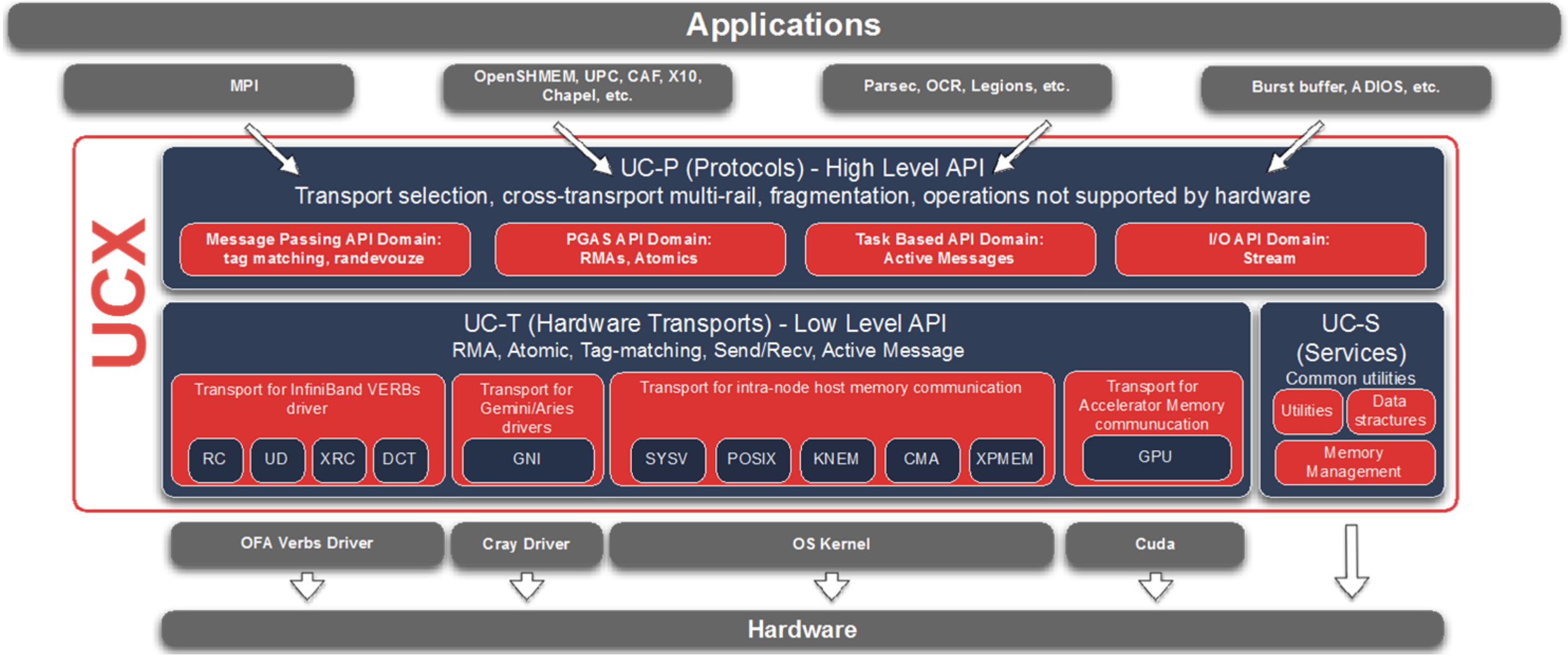
Functionality:
work request setup and instantiation of operations

UC-P for Protocols

High-level API uses UCT framework to construct protocols commonly found in applications

Functionality:
Multi-rail, device selection, pending queue, rendezvous, tag-matching, software-atomics, etc.

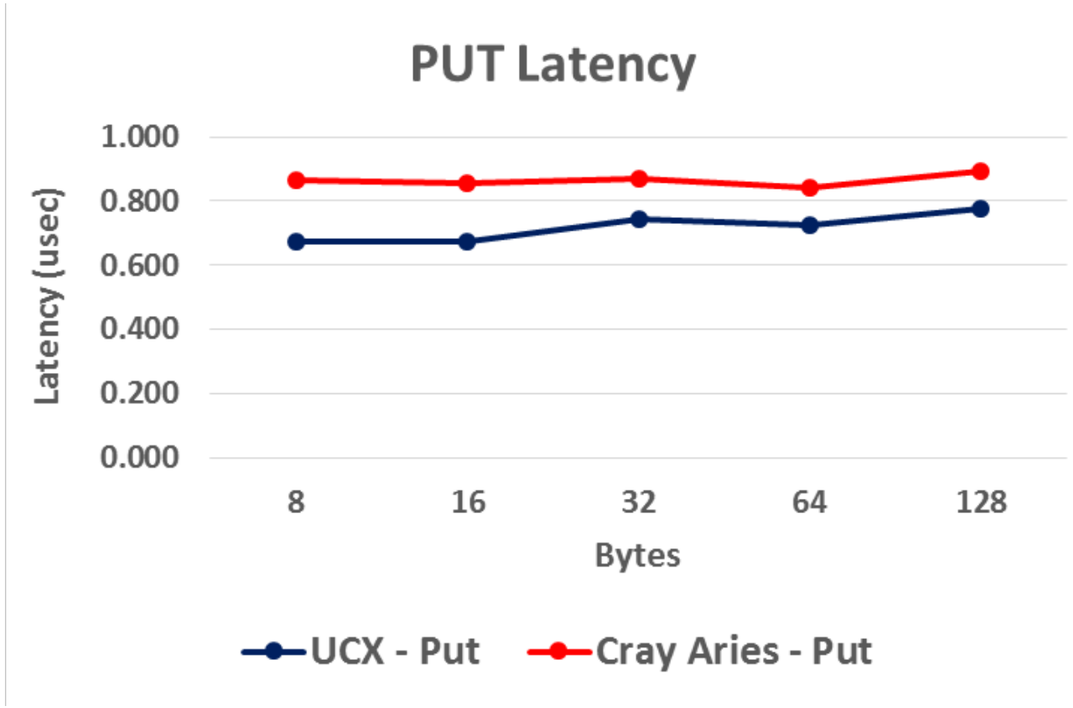
UCX High-level Overview

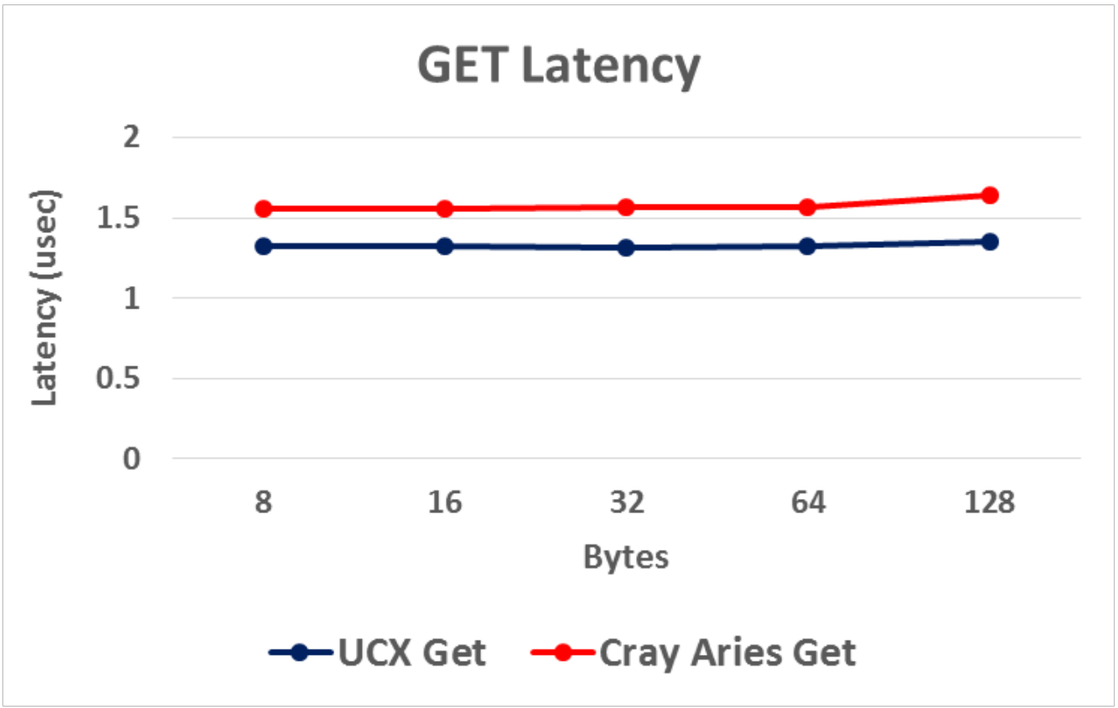


Collaboration

- Mellanox co-designs network interface and contributes MXM technology
  - Infrastructure, transport, shared memory, protocols, integration with OpenMPI/SHMEM, MPICH
- ORNL co-designs network interface and contributes UCCS project
  - InfiniBand optimizations, Cray devices, shared memory
- NVIDIA co-designs high-quality support for GPU devices
  - GPU-Direct, GDR copy, etc.
- IBM co-designs network interface and contributes ideas and concepts from PAMI
- UH/UTK focus on integration with their research platforms

UCX Performance

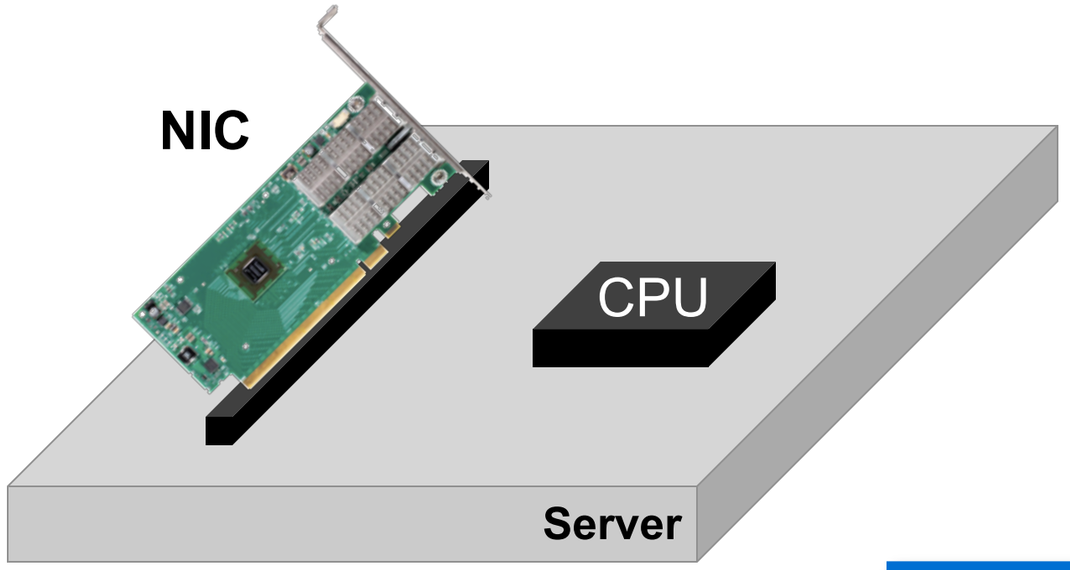




Mellanox Multi-Host™ Technology

Next Generation Data Center Architecture

New Compute Rack / Data Center Architecture

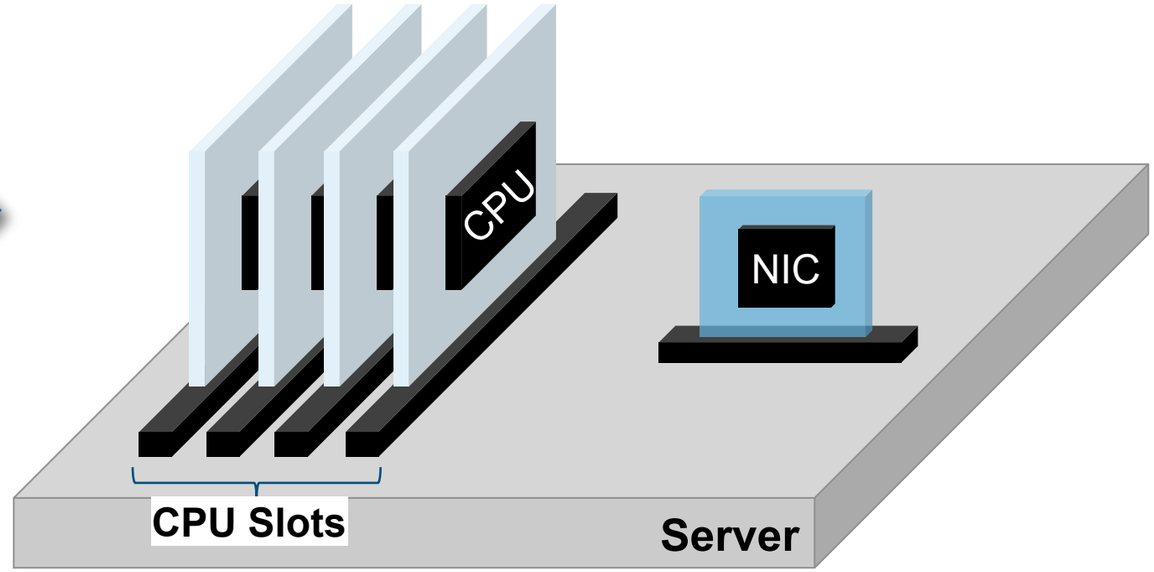


Traditional Data Center

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

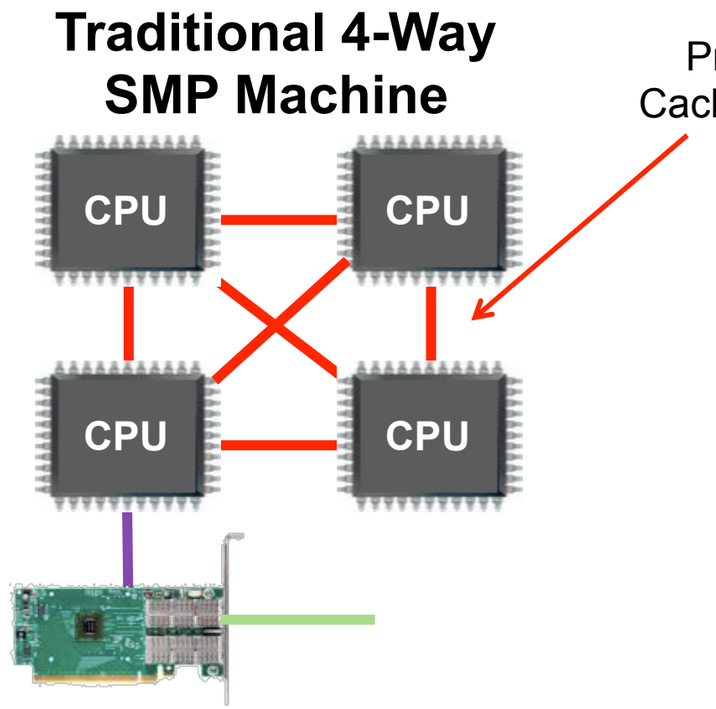
Scalable Data Center with Multi-Host

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

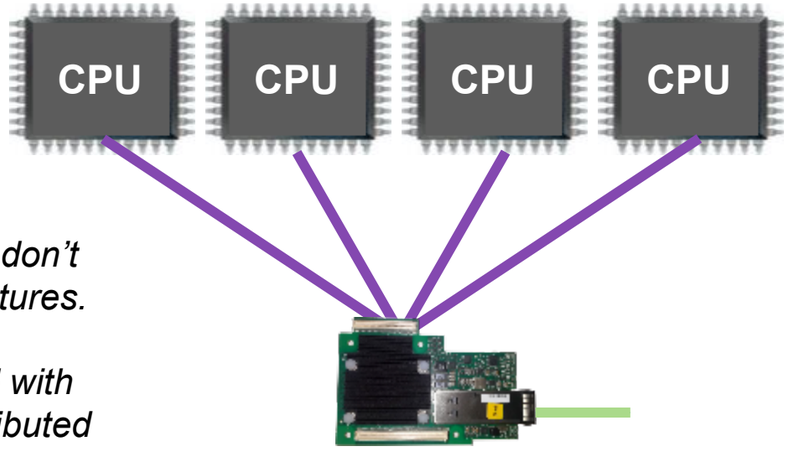


The Network is The Computer

Multi-Host Dramatically Reduces Server Cost



Proprietary Cache Coherent Bus



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

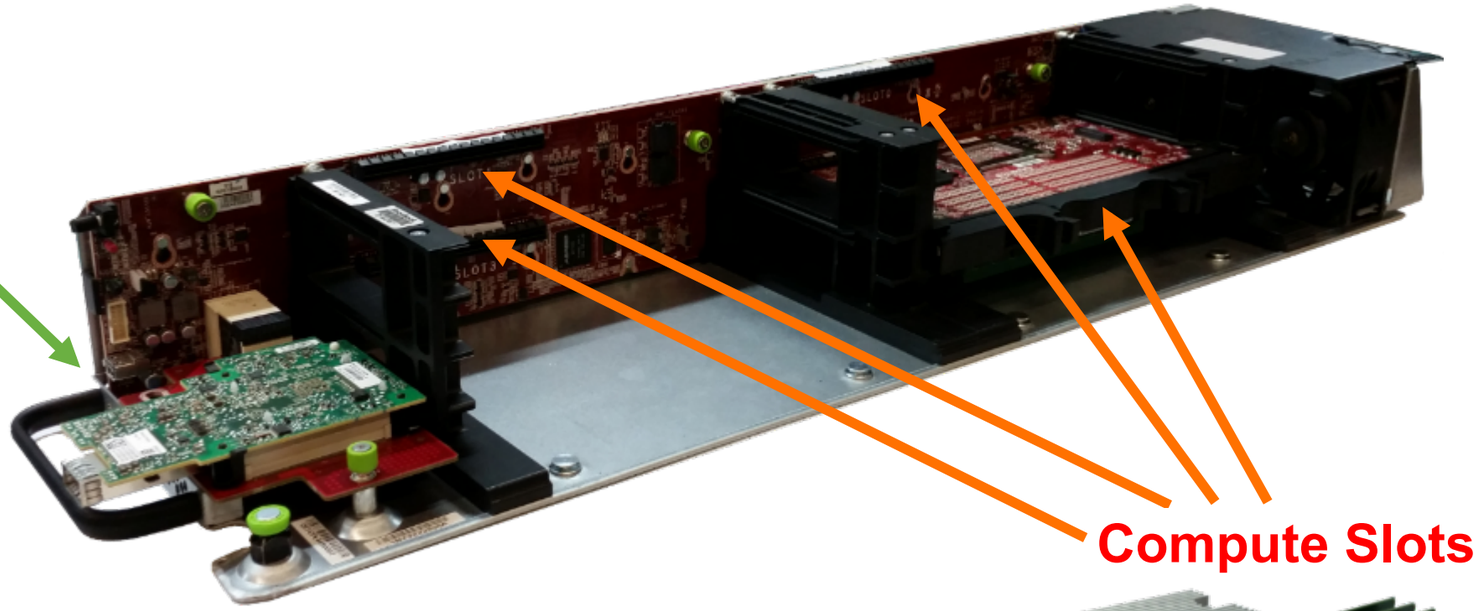
Additional parallelism achieved with higher level network based distributed programming techniques such as Hadoop Map-Reduce

- 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

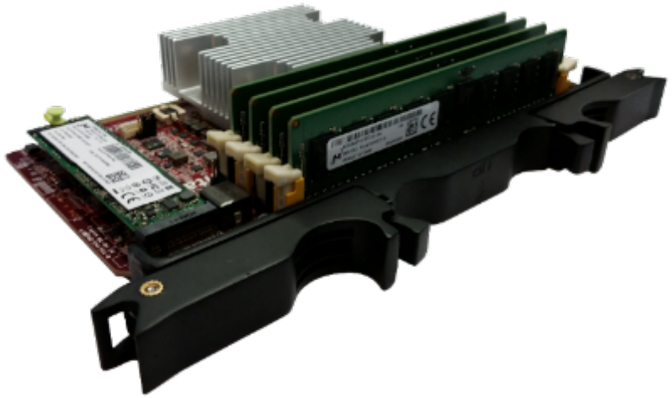
- Low cost single-socket CPU
  - Clean, simple, cost-effective, software transparent
- Cache coherent domain: Multi-Core CPU
  - Eliminates pins, Lower power, Simpler layout
- Symmetric Data Access

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

ConnectX-4 Multi-Host Adapter



Compute Slots



The Next Generation Compute and Storage Rack Design

Summary

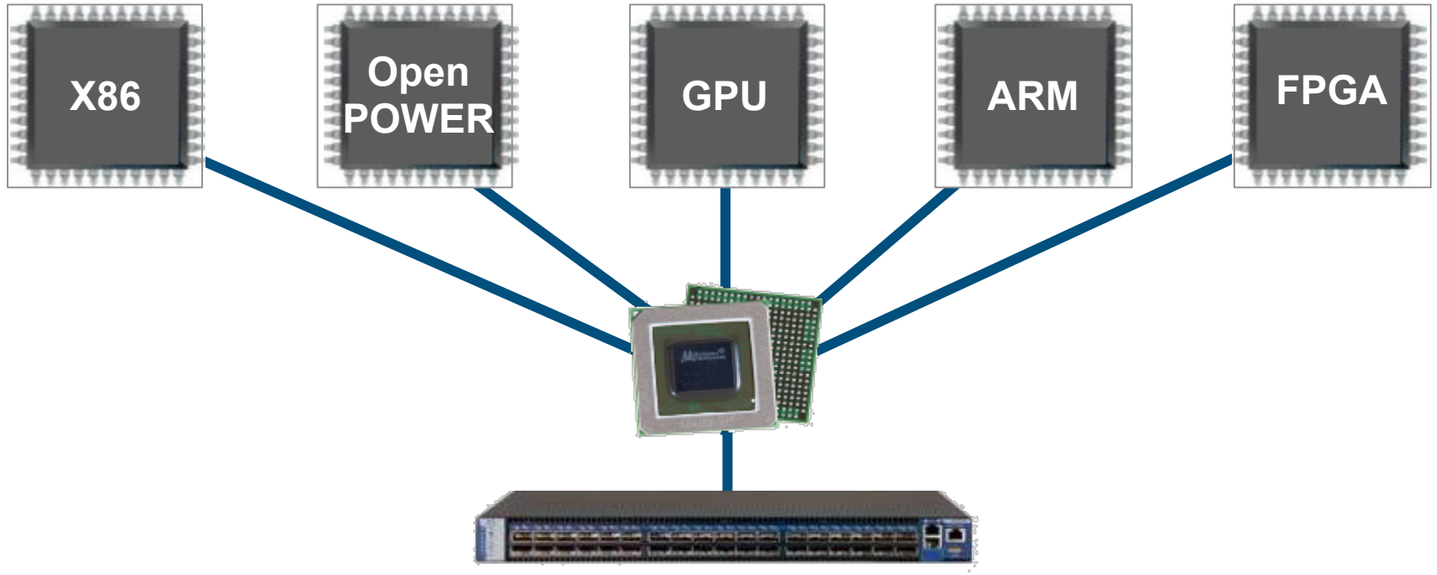
Paving the Road to Exascale Computing

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

Technology Roadmap – One-Generation Lead over the Competition

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

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

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™