Paving the Road to Exascale

Dror Goldenberg

ExaComm 2015

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Proud to Accelerate Future DOE Leadership Systems (“CORAL”)

U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

“Summit” System

Lawrence Livermore National Laboratory

“Sierra” System

5X – 10X Higher Application Performance versus Current Systems

Mellanox InfiniBand, IBM POWER CPUs, NVIDIA Tesla GPUs

Paving the Road to Exascale

Mellanox Accelerated World-Leading HPC Systems

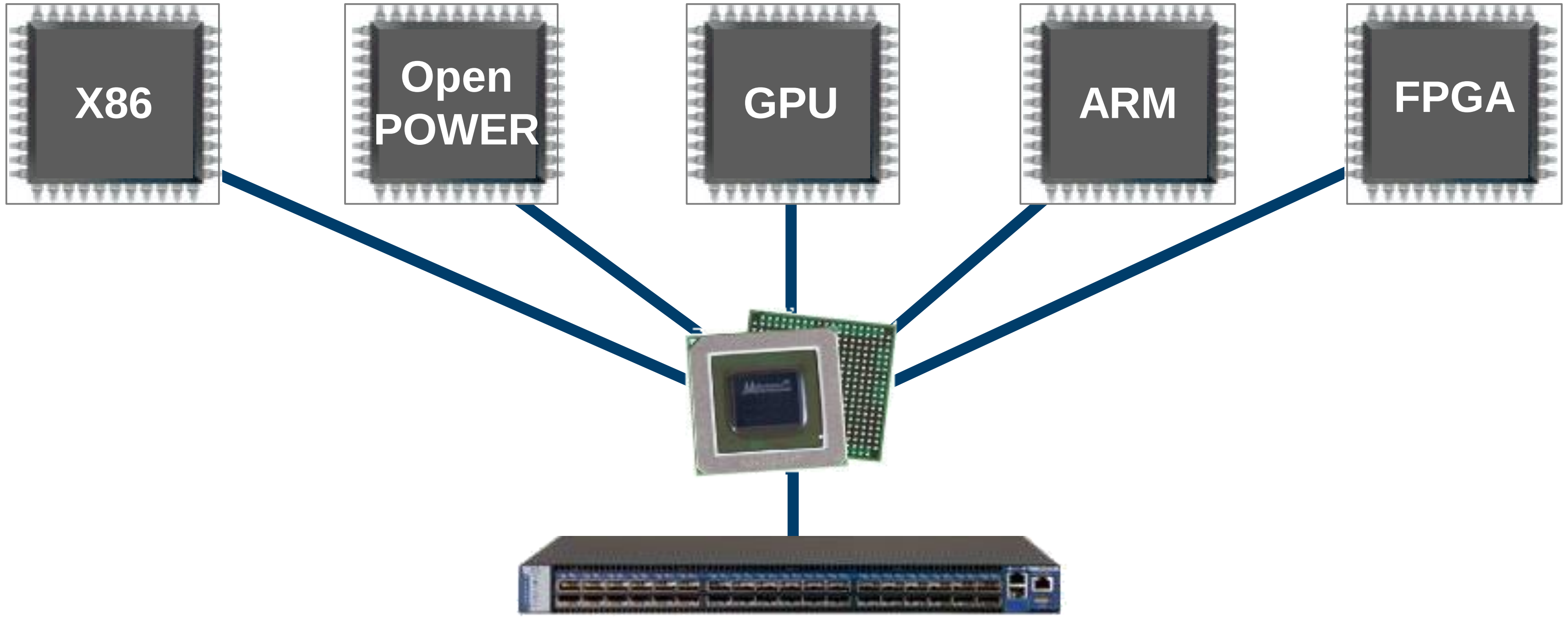
Connected the First Petaflop System

Now Connecting Many of the World’s Leading Petascale Systems

The Future is Heterogeneous

Highest Performance and Scalability for

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



Smart Interconnect to Unleash The Power of All Compute Architectures

Today Latest Technology

100Gb/s InfiniBand

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

Interconnect

LinkX™

Copper (Passive, Active)

Optical Cables (VCSEL)

Silicon Photonics

Enter the World of Scalable Performance – 100Gb/s Adapter

ConnectX-4: Highest Performance Adapter in the Market

InfiniBand: SDR / DDR / QDR / FDR / EDR

Ethernet: 10 / 25 / 40 / 50 / 56 / 100GbE

100Gb/s, <0.7us latency

150 million messages per second

OpenPOWER CAPI technology

CORE-Direct technology

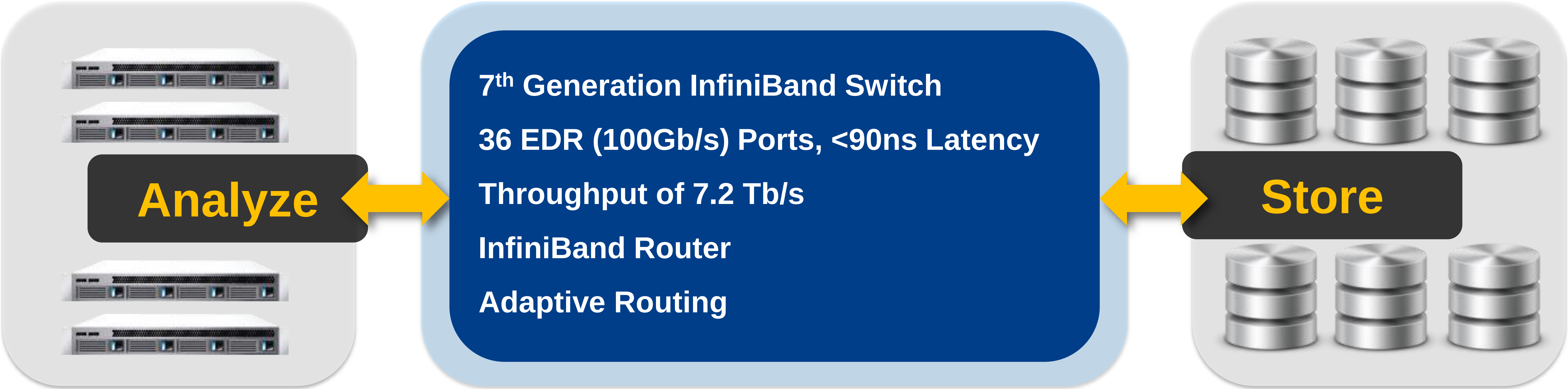
GPUDirect RDMA

Dynamically Connected Transport (DCT)

Ethernet /IPoIB offloads (HDS, RSS, TSS, LRO, LSOv2)

Enter the World of Scalable Performance – 100Gb/s Switch

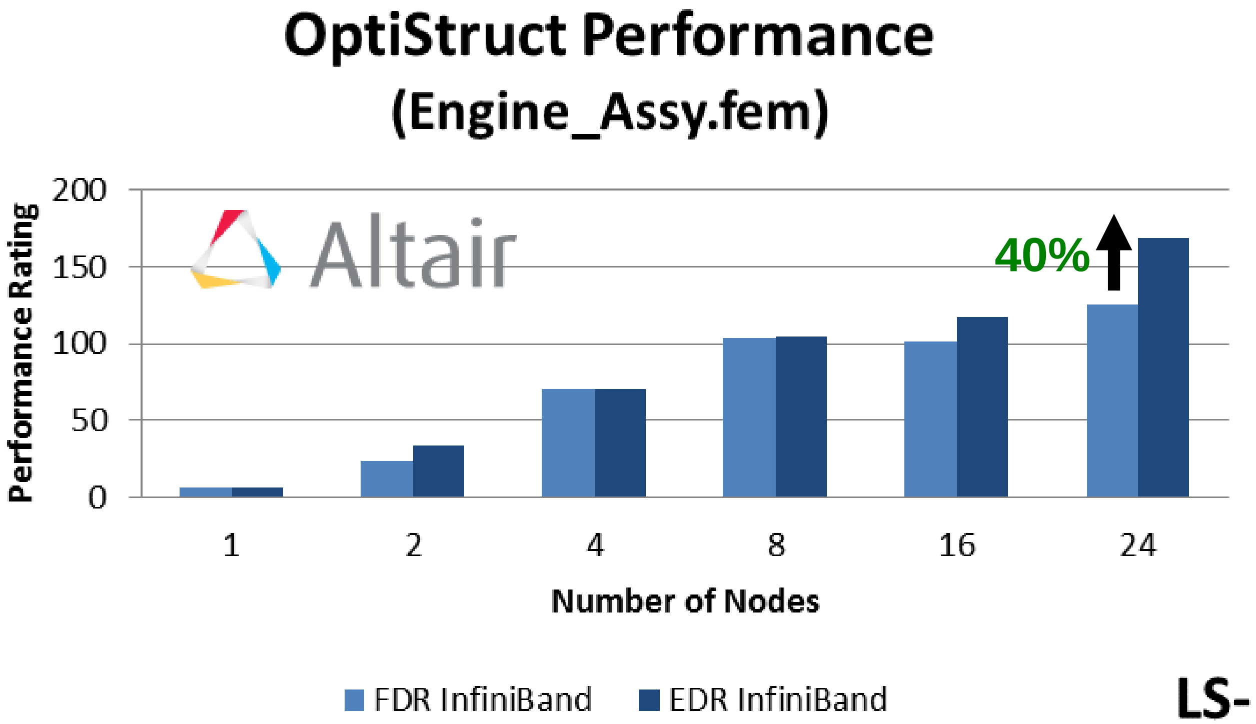
Switch-IB: Highest Performance Switch in the Market

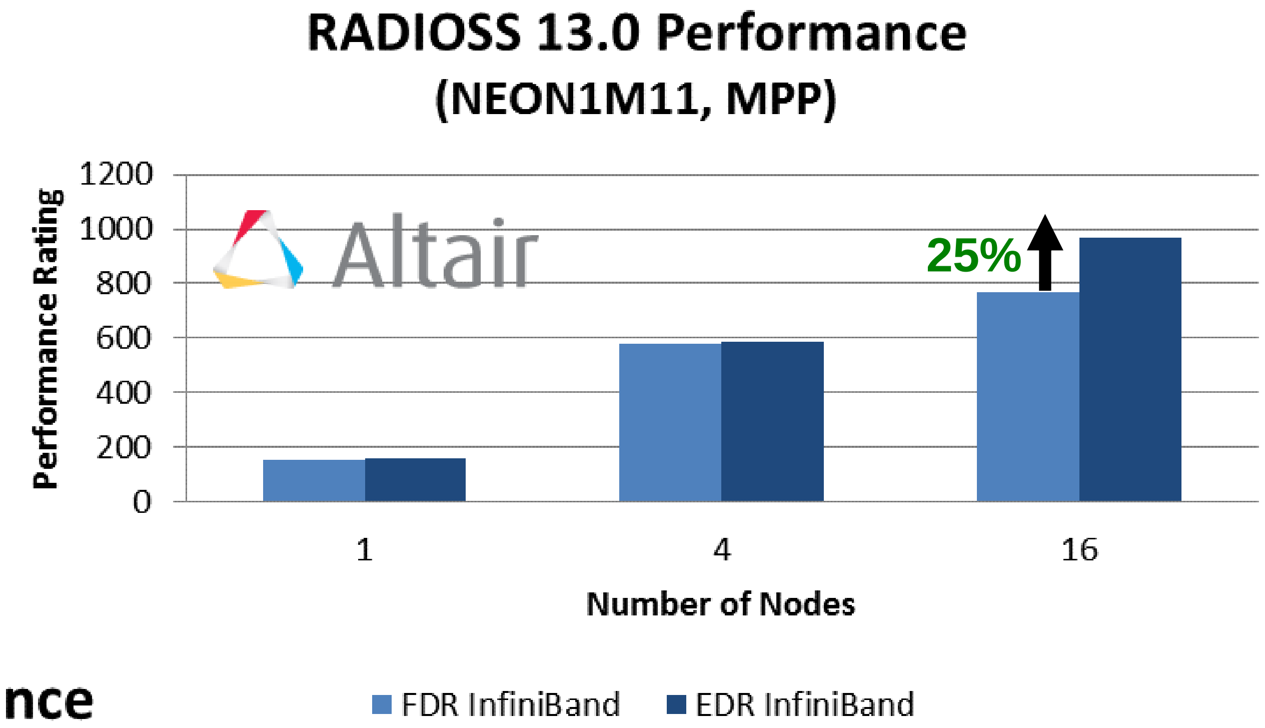


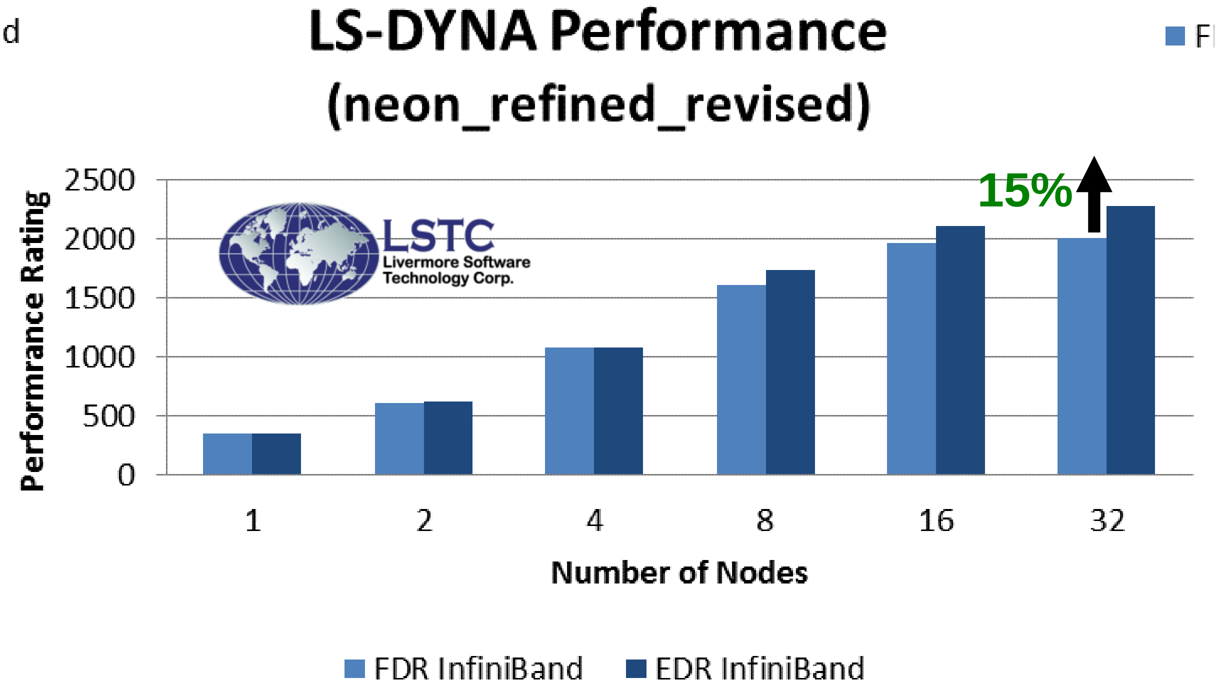
- 7th Generation InfiniBand Switch
- 36 EDR (100Gb/s) Ports, <90ns Latency
- Throughput of 7.2 Tb/s
- InfiniBand Router
- Adaptive Routing

SwitchIB™

EDR InfiniBand Performance Leadership







ConnectX-4®

SwitchIB™

HPC-X™

For all graphs: higher is better

InfiniBand Adapters Performance Evaluation

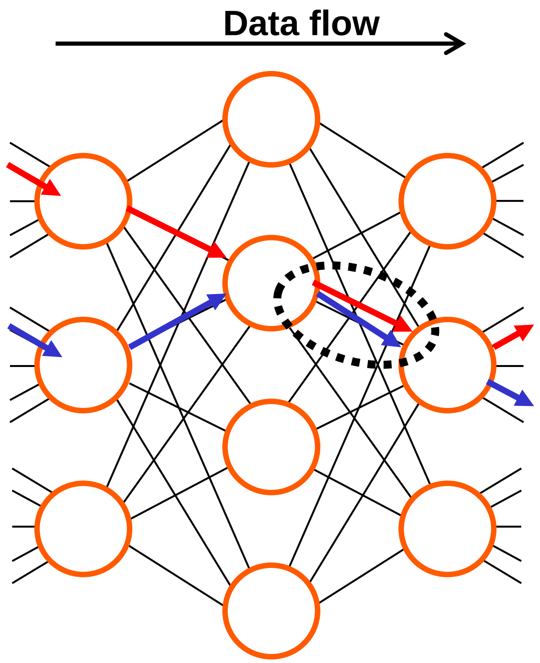
Mellanox Adapters Single Port Performance	ConnectX-4 EDR 100G	Connect-IB FDR 56G	ConnectX-3 Pro FDR 56G
Uni-Directional Throughput	100 Gb/s	54.24 Gb/s	51.1 Gb/s
Bi-Directional Throughput	195 Gb/s	107.64 Gb/s	98.4 Gb/s
Latency	0.61 us	0.63 us	0.64 us
Message Rate	149.5 Million/sec	105 Million/sec	35.9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

Destination Based Forwarding and Non-Blocking Clos Networks

- Non-Blocking
  - A network is Non-Blocking if it can be routed to support any possible source destination pairing (a.k.a. “Permutation”) with no network contention

- Strictly Non-Blocking
  - When routing of new pairs does not interfere with previously routed pairs

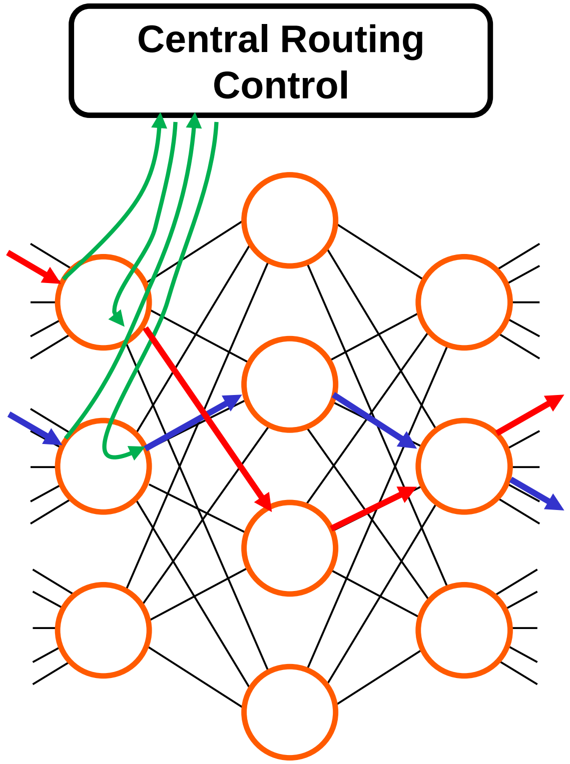
- Rearrangeable Non-Blocking
  - When routing of new pairs may require re-routing of previously routed pairs



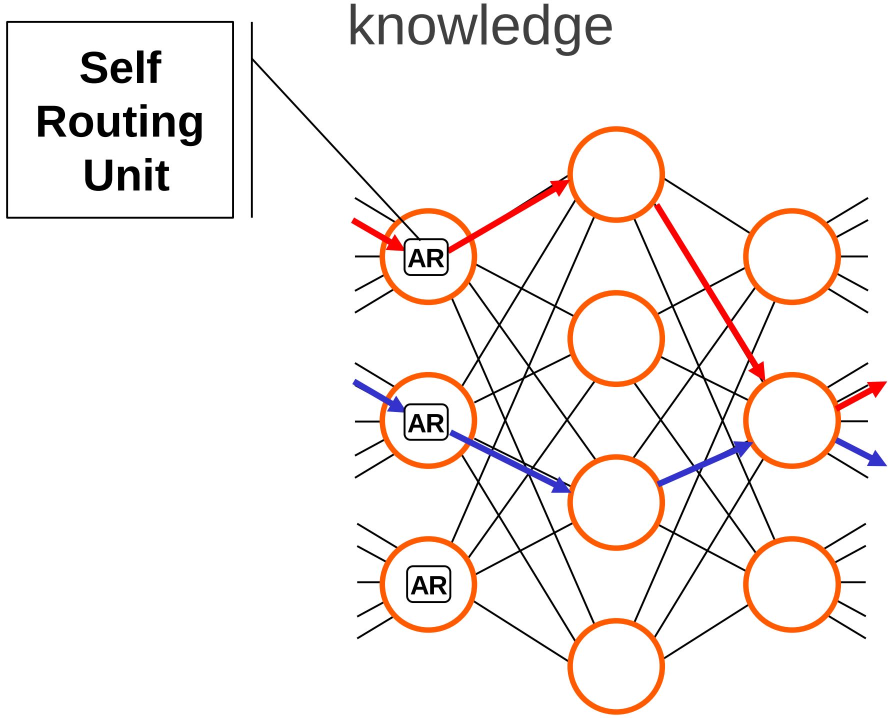
Traffic Aware Load Balancing Systems

Property	Central Adaptive Routing	Distributed Adaptive Routing
Scalability	Low	High
Knowledge	Global	Local (to keep scalability)
Non-Blocking	Yes	Good

- Centralized
  - Flows are routed according to a “global” knowledge



- Distributed
  - Each flow is routed by its input switch with “local” knowledge

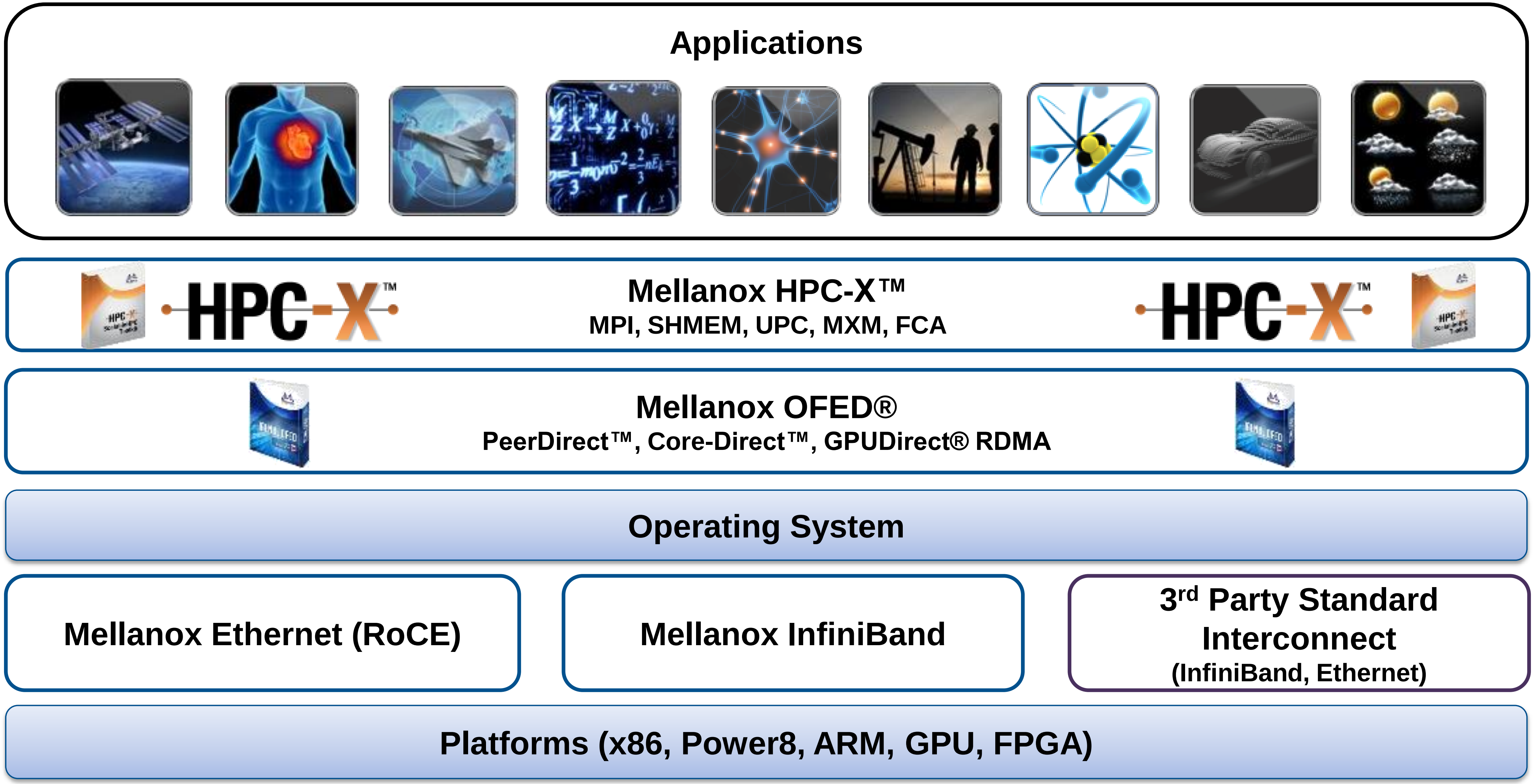


Mellanox HPC-X™ Scalable HPC Software Toolkit

HPC-X™

- MPI, PGAS OpenSHMEM and UPC package for HPC environments
- Fully optimized for Mellanox InfiniBand and 3rd party interconnect solutions
- Maximize application performance
- Mellanox tested, supported and packaged
- For commercial and open source usage

Enabling Highest Applications Scalability and Performance



Applications

Mellanox HPC-X™
MPI, SHMEM, UPC, MXM, FCA

Mellanox OFED®
PeerDirect™, Core-Direct™, GPUDirect® RDMA

Operating System

Mellanox Ethernet (RoCE)

Mellanox InfiniBand

3rd Party Standard Interconnect
(InfiniBand, Ethernet)

Platforms (x86, Power8, ARM, GPU, FPGA)

Comprehensive MPI, PGAS/OpenSHMEM/UPC Software Suite

Software Development - Unified Communication – X Framework

Collaboration between industry, laboratories, and academia, to create open-source production grade communication framework for data centric and HPC applications

www.openucx.org

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

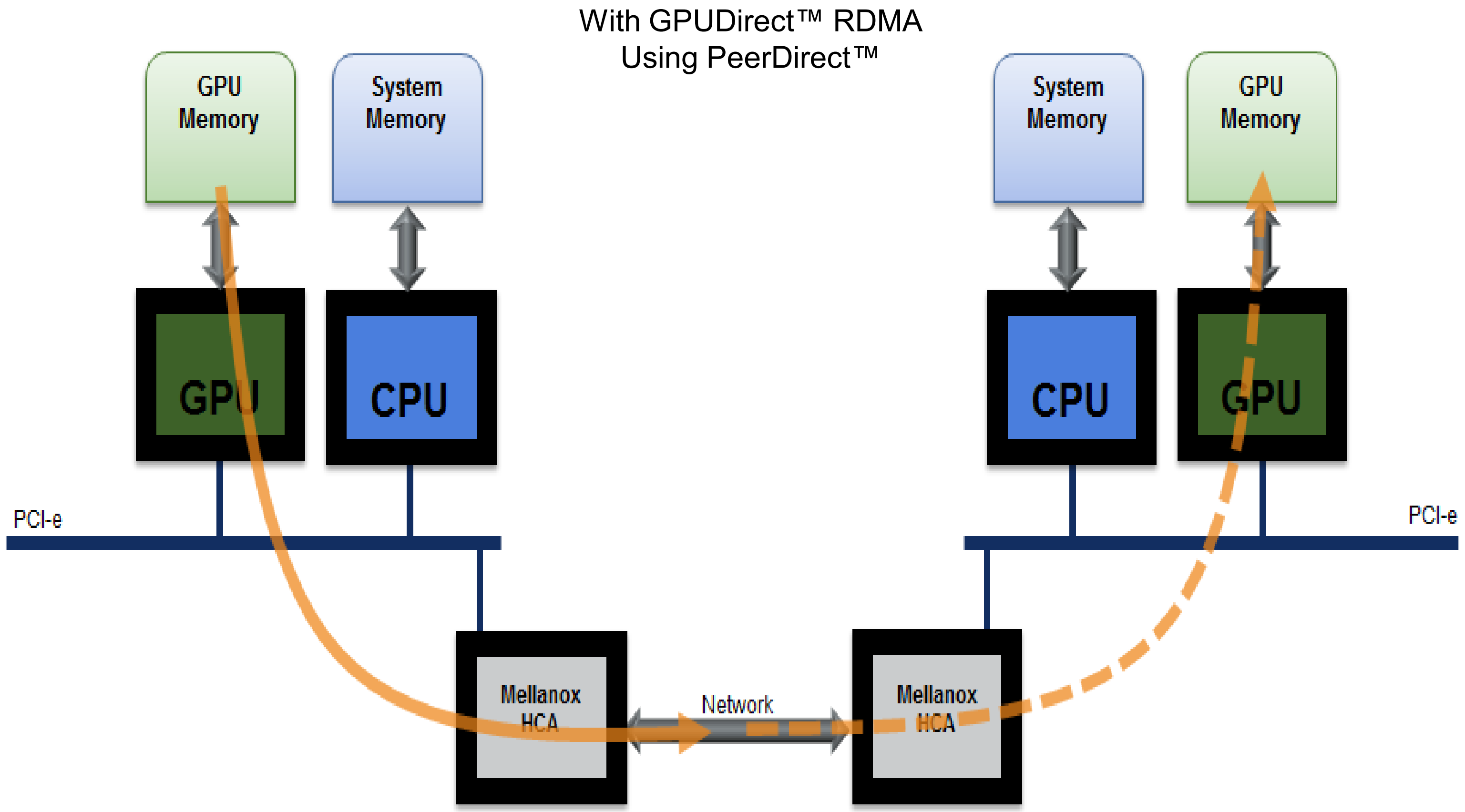
High-level API uses UCT framework to construct protocols commonly found in applications

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

Co-Design Collaboration

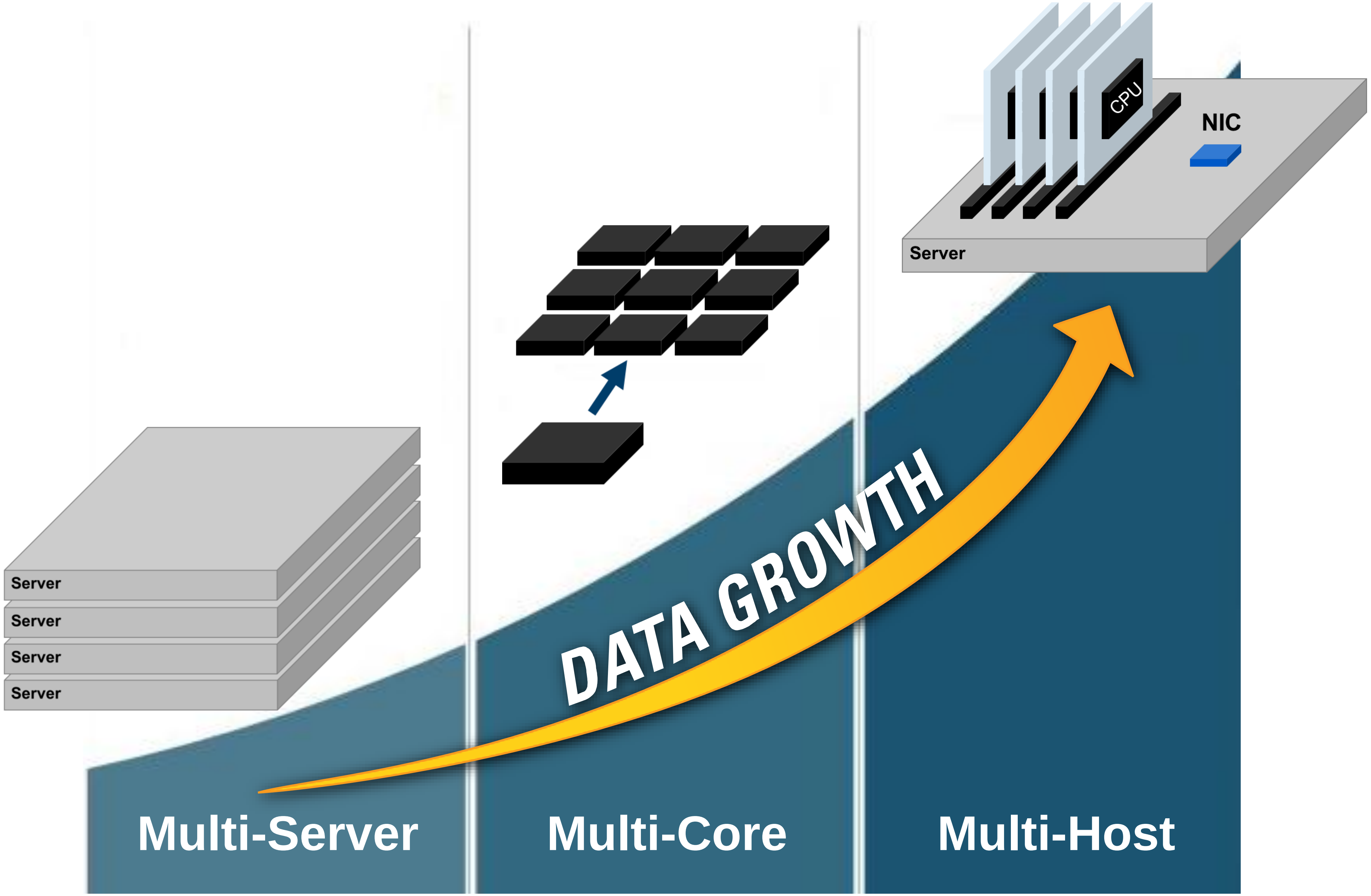
- Mellanox co-designs network interface and contributes MXM technology
  - Infrastructure, transport, shared memory, protocols, integration with OpenMPI/SHMEM, MPICH
- ORNL co-designs network interface and contributes UCCS project
  - InfiniBand optimizations, Cray devices, shared memory
- NVIDIA co-designs 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

GPUDirect™ RDMA (GPUDirect 3.0)

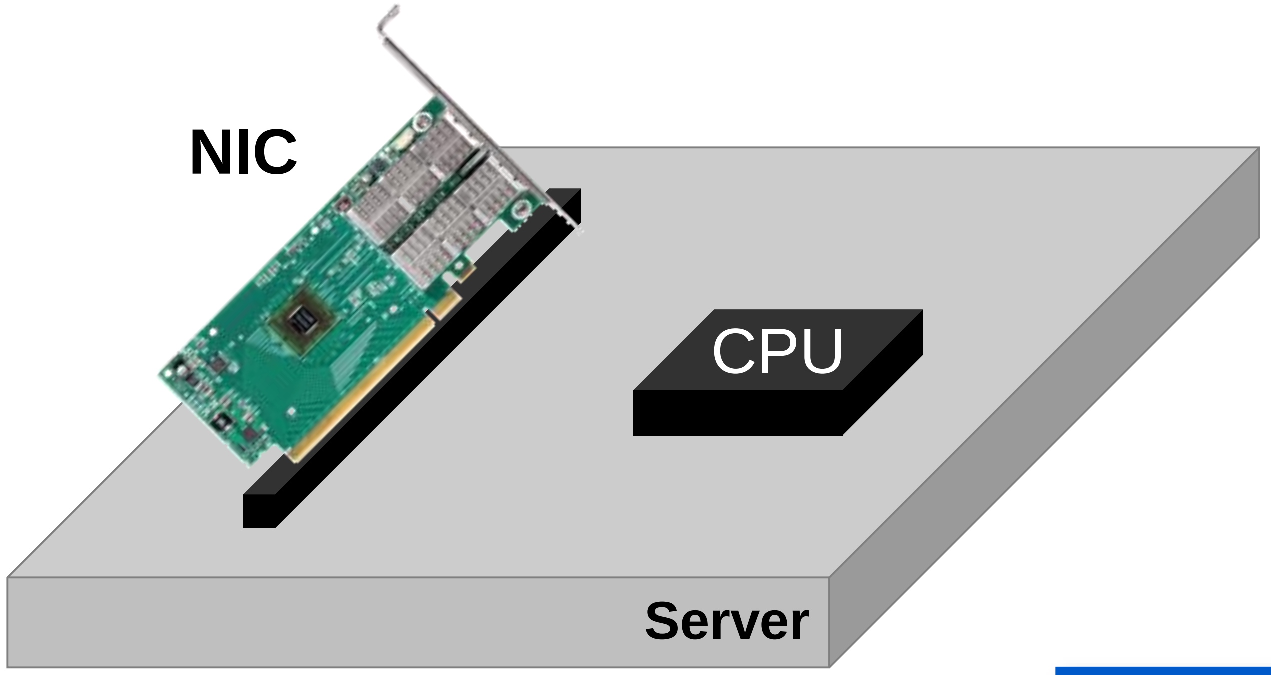


- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPI SendRecv efficiency between GPUs in remote nodes
- Based on PeerDirect technology

Data Center Evolution Over Time



New Compute Rack / Data Center Architecture

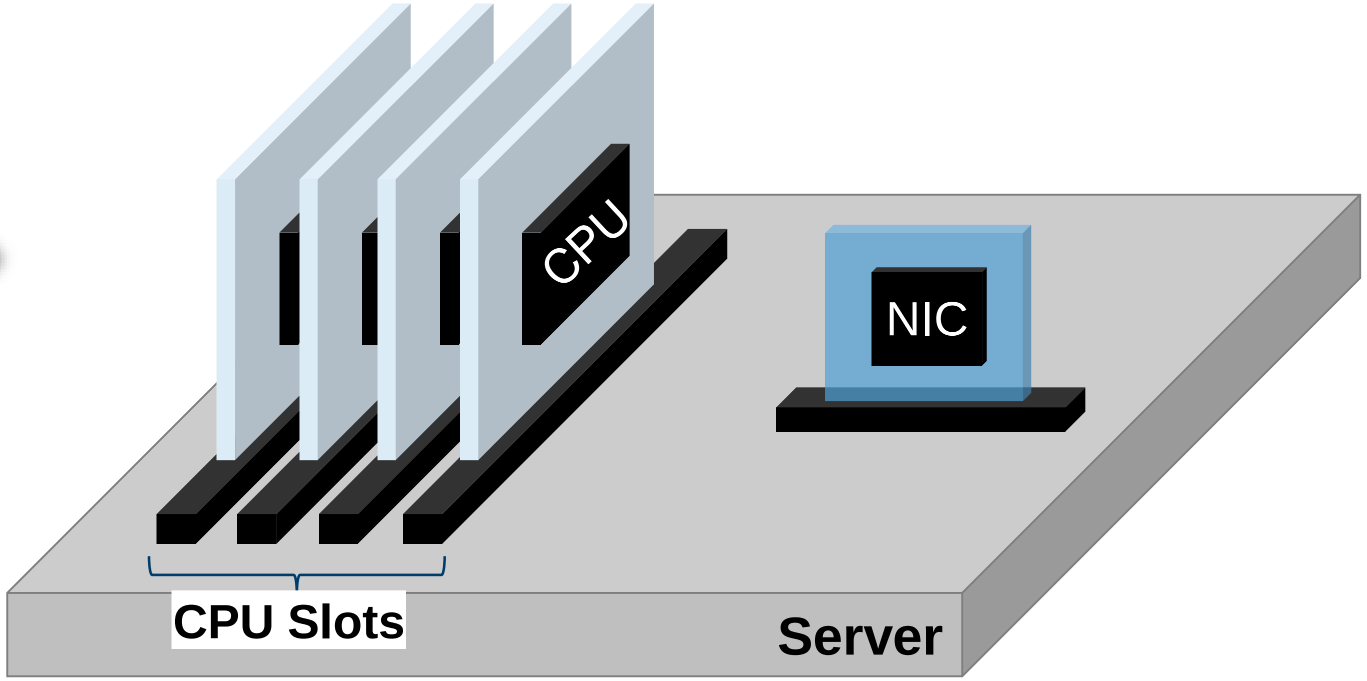


Traditional Data Center

- 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

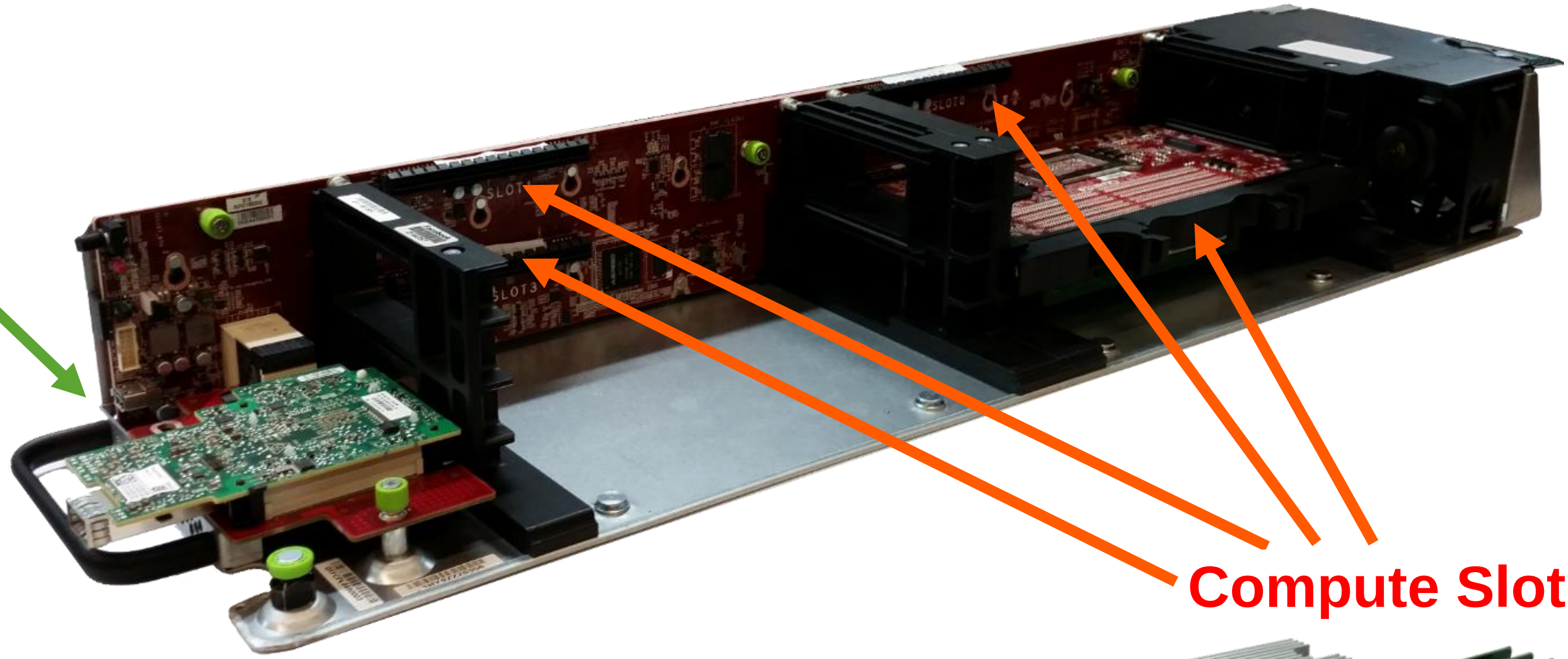


The Network is The Computer

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

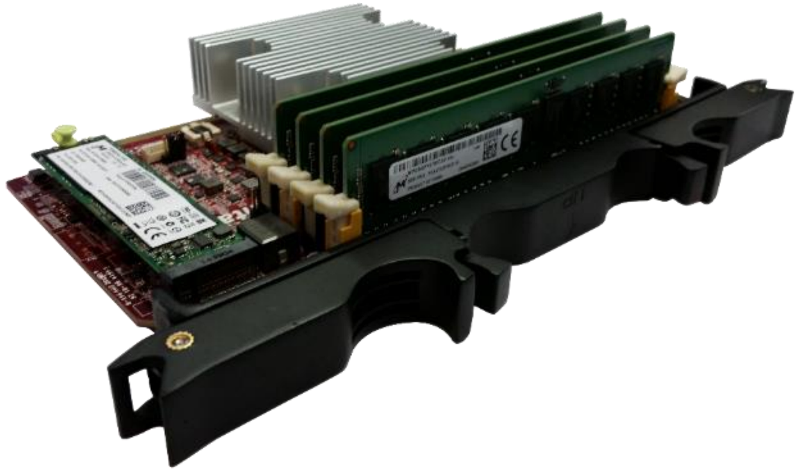
Mellanox TECHNOLOGIES

OPEN Compute Project



ConnectX-4 Multi-Host Adapter

Compute Slots



Mellanox TECHNOLOGIES

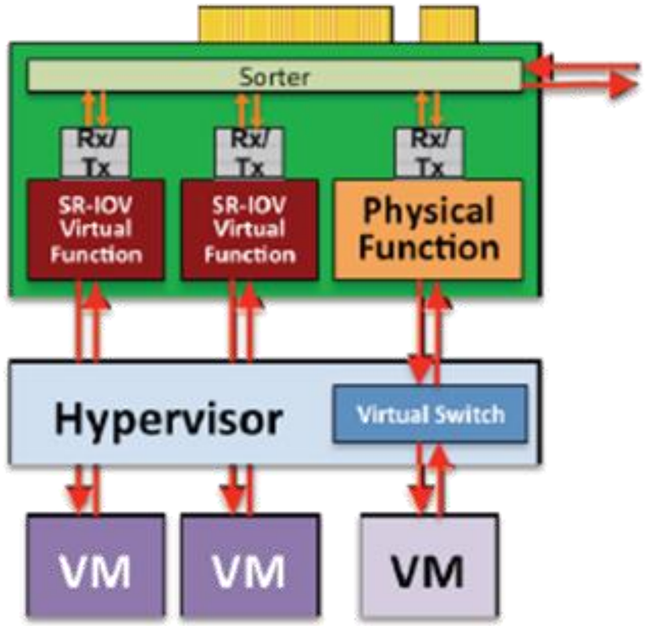
facebook

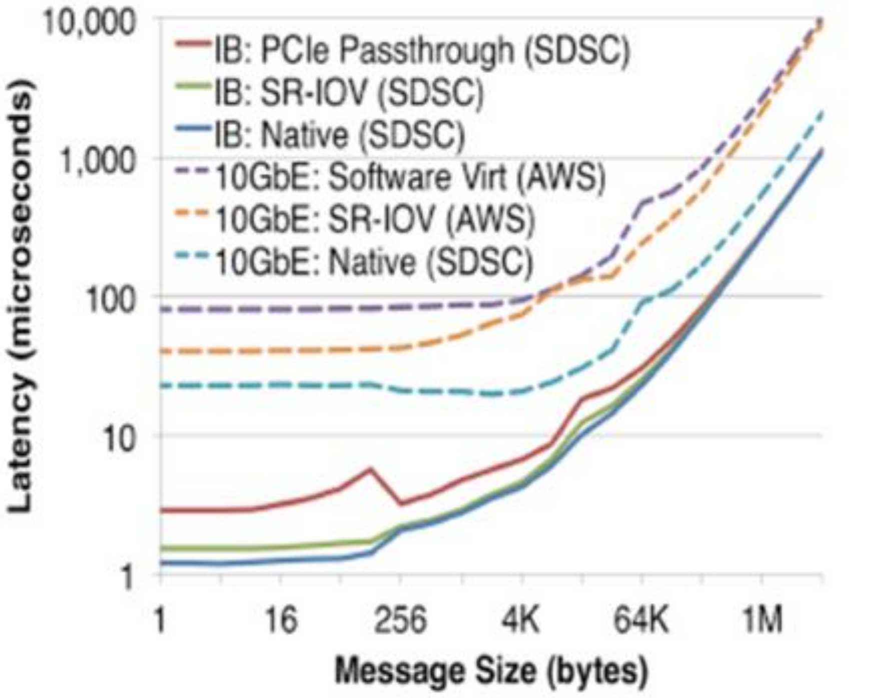
The Next Generation Compute and Storage Rack Design

Virtualization for HPC: Mellanox SRIOV

Single Root I/O Virtualization in HPC

- **Problem**: Virtualization generally has resulted in significant I/O performance degradation (e.g., excessive DMA interrupts)
- **Solution**: SR-IOV and Mellanox InfiniBand host channel adapters
    - One physical function → multiple virtual functions, each light weight but with its own DMA streams, memory space, interrupts
    - Allows DMA to bypass hypervisor to VMs
- ***SRIOV enables virtual HPC cluster w/ near-native InfiniBand latency/bandwidth and minimal overhead***

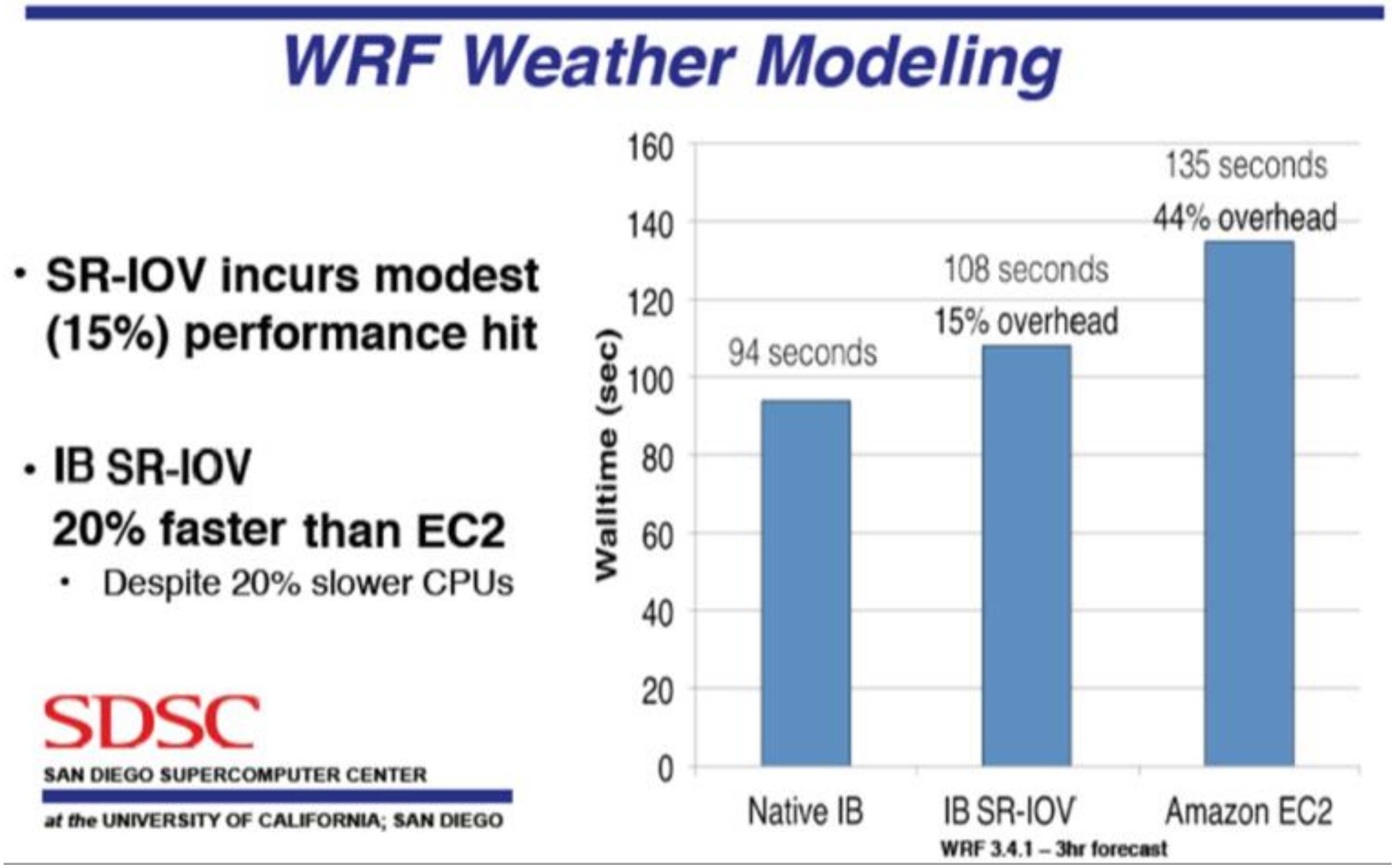




MPI point-to-point latency measured by osu_latency for QDR InfiniBand. Included for scale are the analogous 10GbE measurements from Amazon (AWS) and non-virtualized 10GbE.

SDSC SAN DIEGO SUPERCOMPUTER CENTER
at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

HPC Clouds – Performance Demands Mellanox Solutions

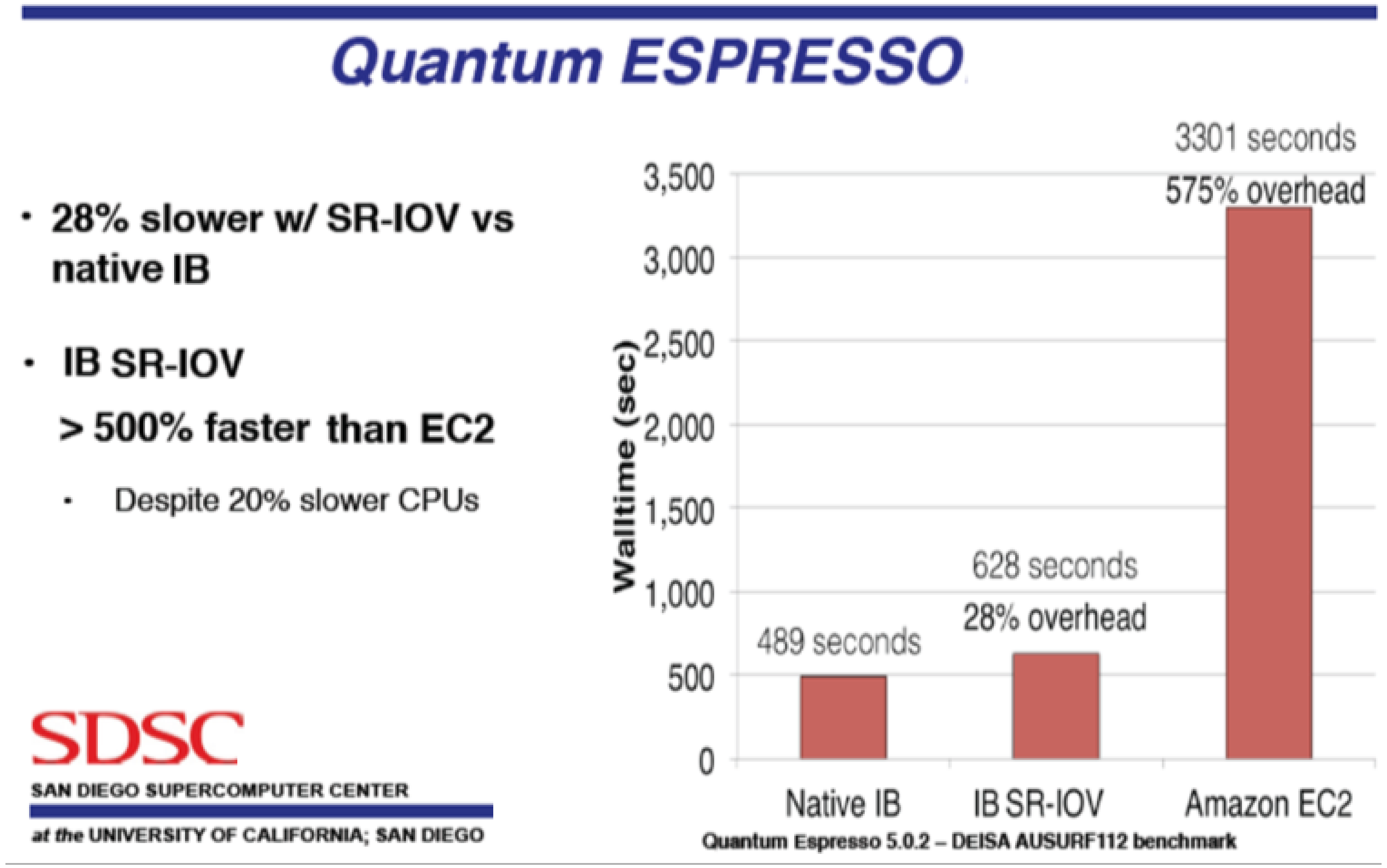


WRF Weather Modeling

- SR-IOV incurs modest (15%) performance hit
- IB SR-IOV 20% faster than EC2
  - Despite 20% slower CPUs

SDSC SAN DIEGO SUPERCOMPUTER CENTER at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

WRF 3.4.1 – 3hr forecast



Quantum ESPRESSO

- 28% slower w/ SR-IOV vs native IB
- IB SR-IOV
  > 500% faster than EC2
  - Despite 20% slower CPUs

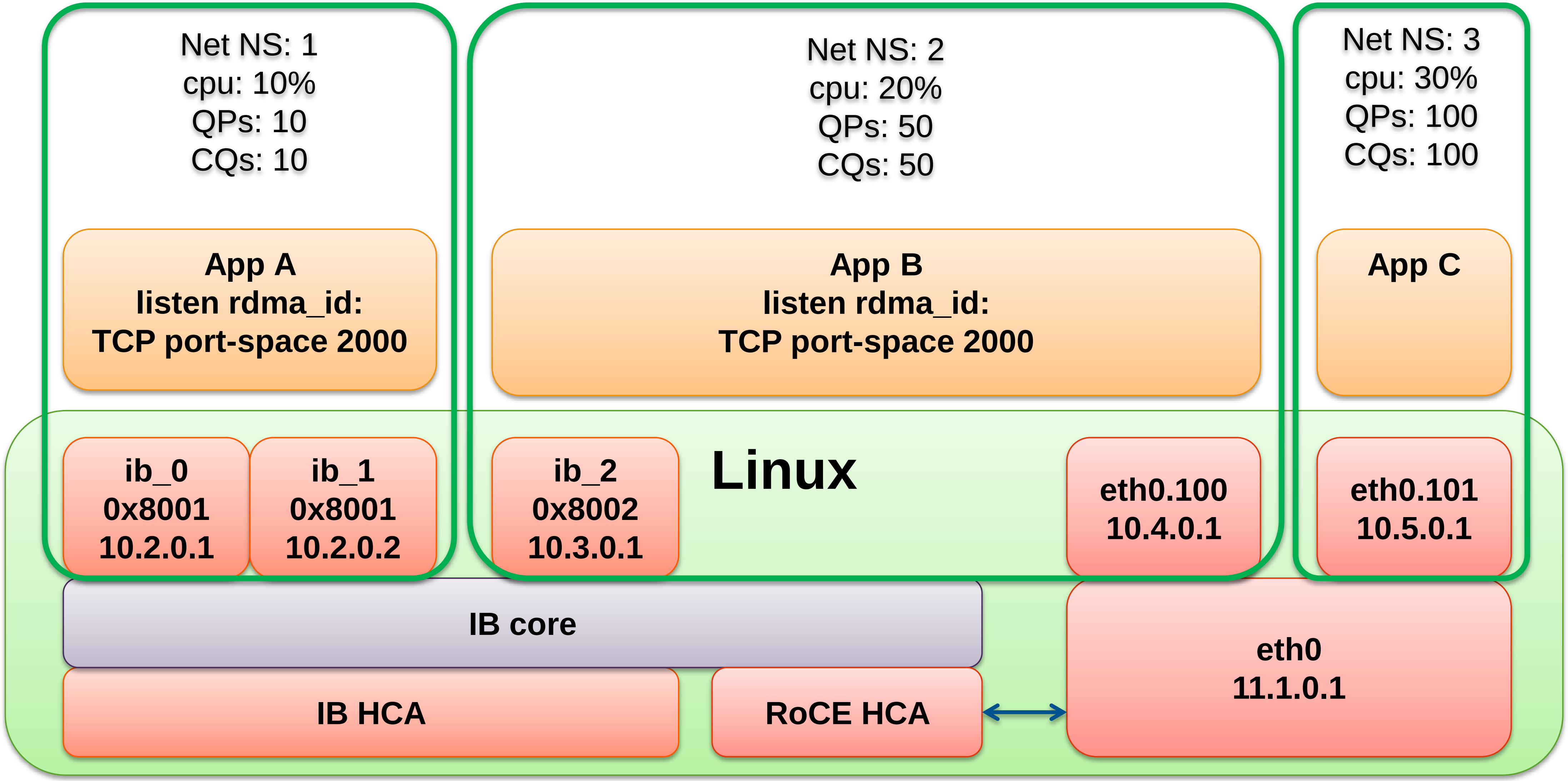
SDSC SAN DIEGO SUPERCOMPUTER CENTER at the UNIVERSITY OF CALIFORNIA; SAN DIEGO

Quantum Espresso 5.0.2 – DEISA AUSURF112 benchmark

San Diego Supercomputing Center "Comet" System (2015) to Leverage Mellanox Solutions and Technology to Build HPC Cloud

RDMA Container Support – Coming Up

- Secure and isolated access to high performance networking
- Extend network namespaces to support RDMA
- Fine grained control of HCA RDMA resources (cgroup)



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 ecosystem
- Higher performance: 100Gb/s, sub 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.™