
# Designing Towards Exascale Compute and Extreme Performance in the Cloud

BEAR Cloud Launch

October 21, 2016

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

# Mellanox Connects the World's Fastest Supercomputer

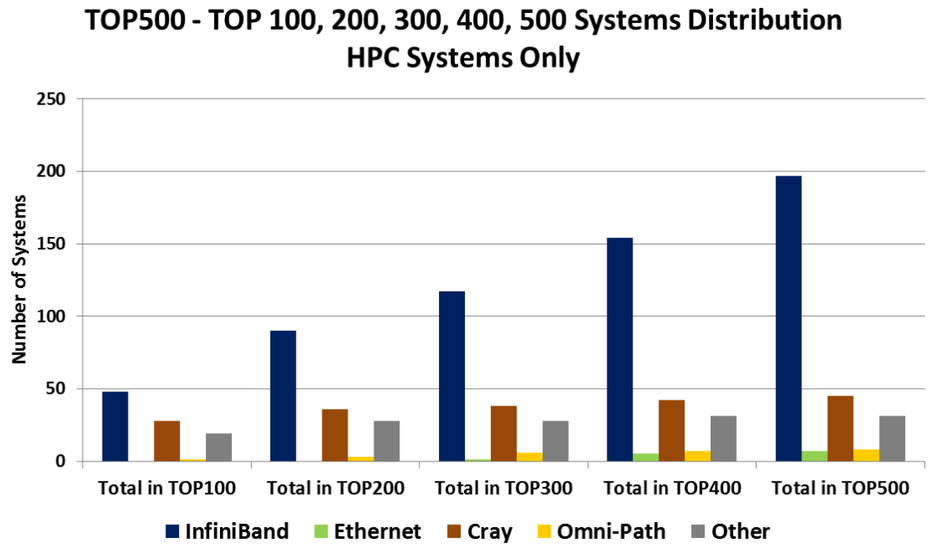
**National Supercomputing Center in Wuxi, China**

**#1 on the TOP500 Supercomputing List**

- 93 Petaflop performance, 3X higher versus #2 on the TOP500
- Nearly 41K nodes, 10 million cores, 256 cores per CPU
- Mellanox adapter and switch solutions

- The TOP500 list has evolved, includes HPC & Cloud / Web2.0 Hyperscale systems
- Mellanox connects 41.2% of  overall TOP500 systems
- Mellanox connects 70.4% of the TOP500 HPC platforms
- Mellanox connects 46 Petascale systems, Nearly 50% of the total Petascale systems

**InfiniBand is the Interconnect of Choice for HPC Compute and Storage Infrastructures**



# Accelerating HPC Leadership Systems

OAK RIDGE National Laboratory

**“Summit” System**

U.S. DEPARTMENT OF ENERGY

Lawrence Livermore National Laboratory

**“Sierra” System**


**Proud to Pave the Path to Exascale**

# 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**

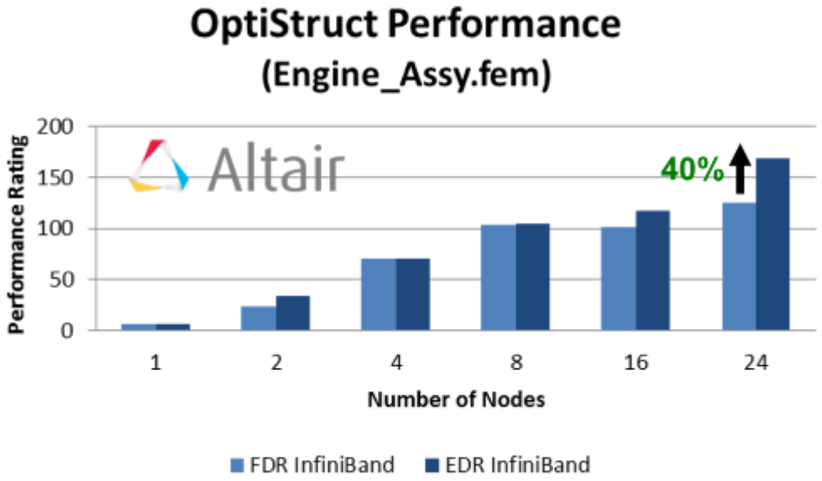
## Exascale

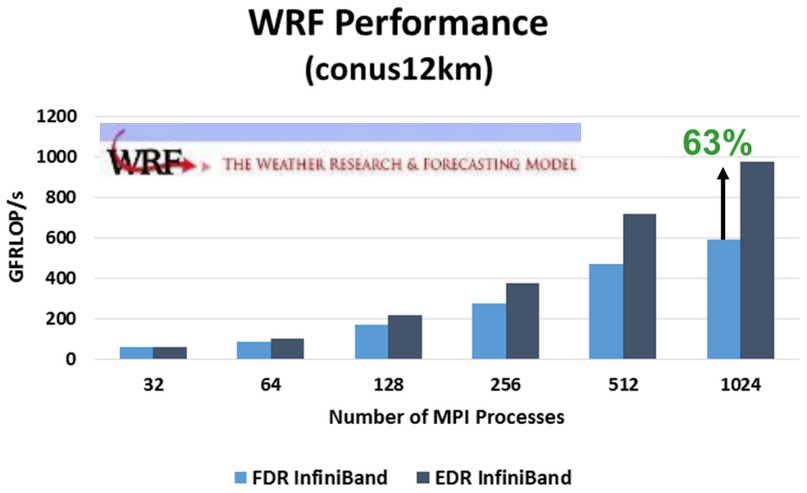
**OAK RIDGE National Laboratory**
**“Summit” System**

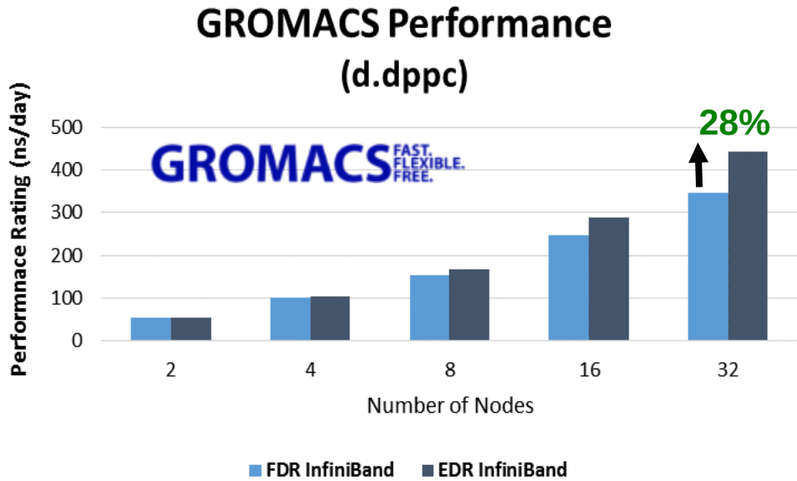
**Lawrence Livermore National Laboratory**
**“Sierra” System**

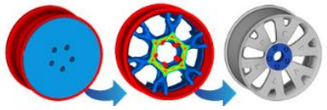
2000 2005 2010 2015 2020

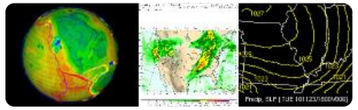
# The Performance Advantage of EDR 100G InfiniBand (28-80%)

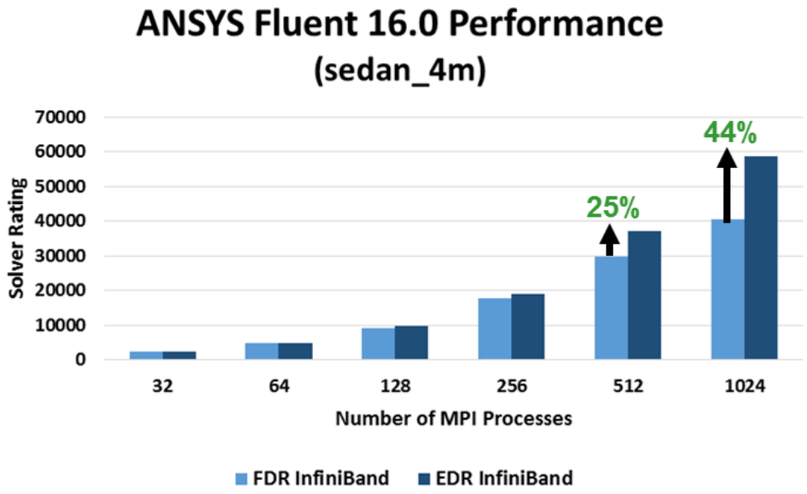


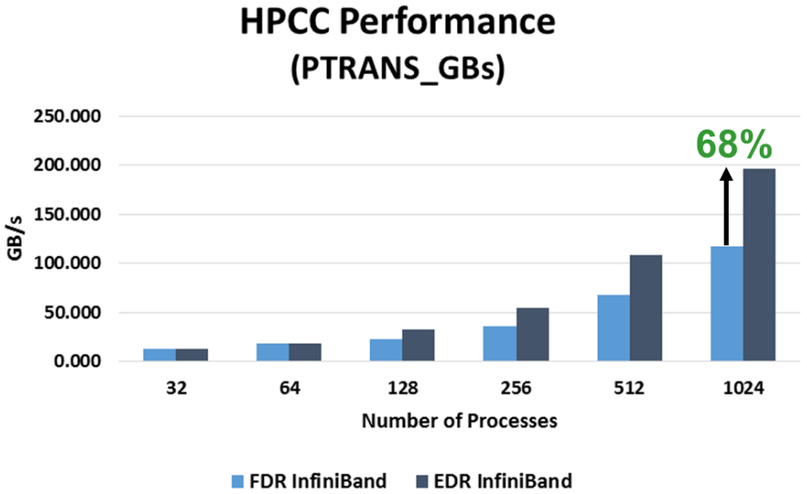



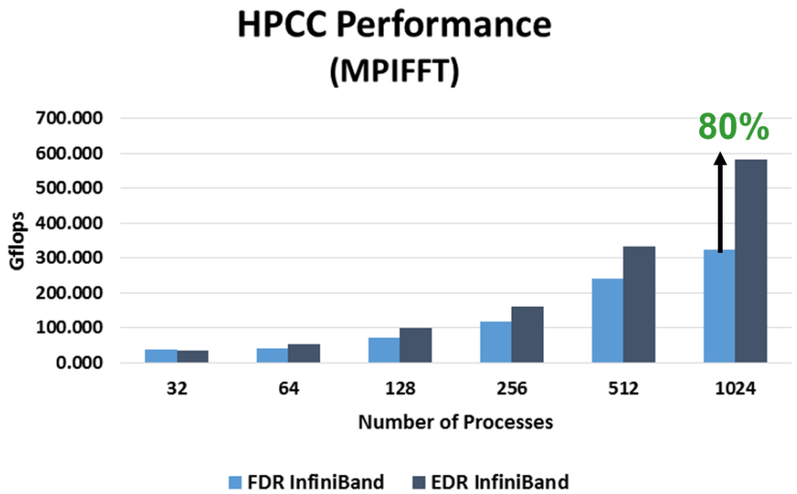







# The Ever Growing Demand for Higher Performance

## Performance Development

Terascale

Petascale

1st

"Roadrunner"

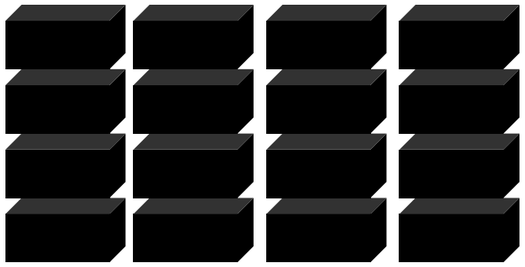
Exascale

OAK RIDGE National Laboratory
"Summit" System

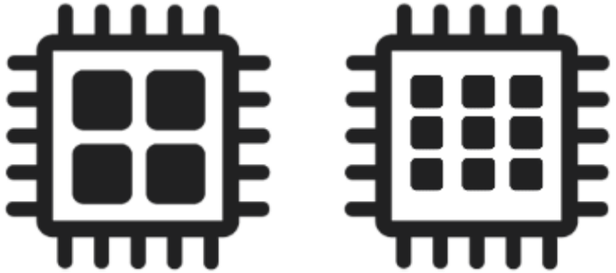
Lawrence Livermore National Laboratory
"Sierra" System

2000 2005 2010 2015 2020

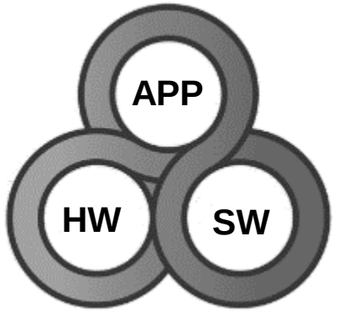
## The Interconnect is the Enabling Technology



SMP to Clusters



Single-Core to Many-Core

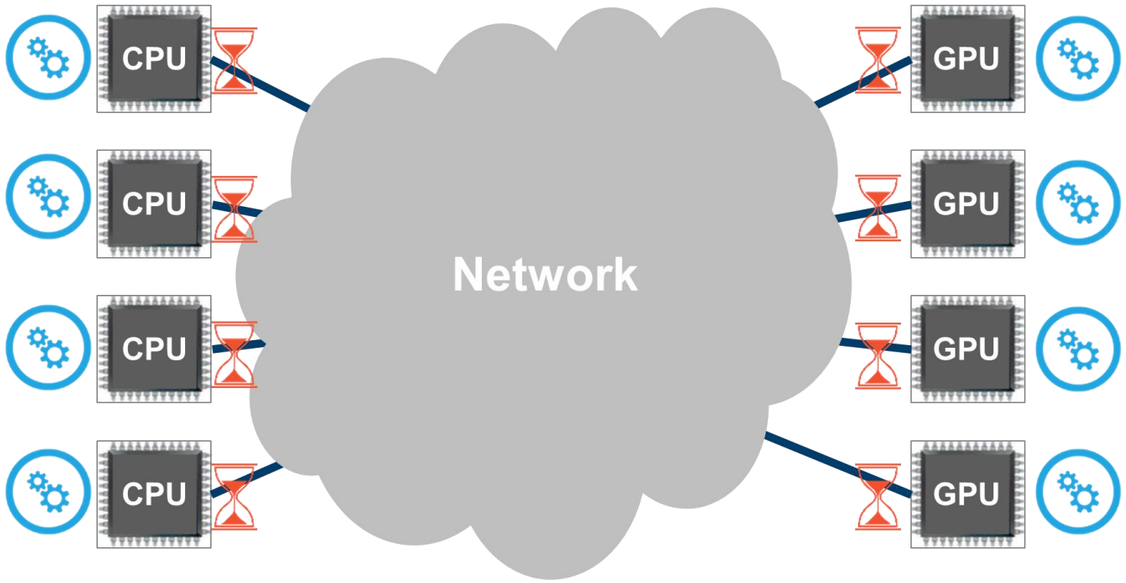



Application
Software
Hardware

Co-Design

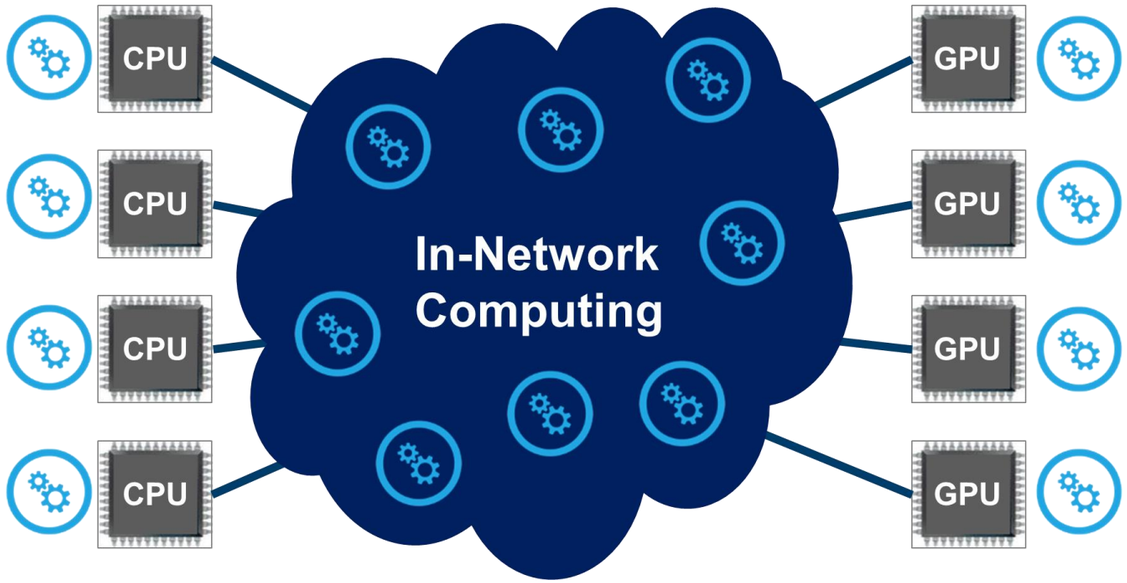
# The Intelligent Interconnect to Enable Exascale Performance

## CPU-Centric




Limited to Main CPU Usage
Results in Performance Limitation

## Co-Design

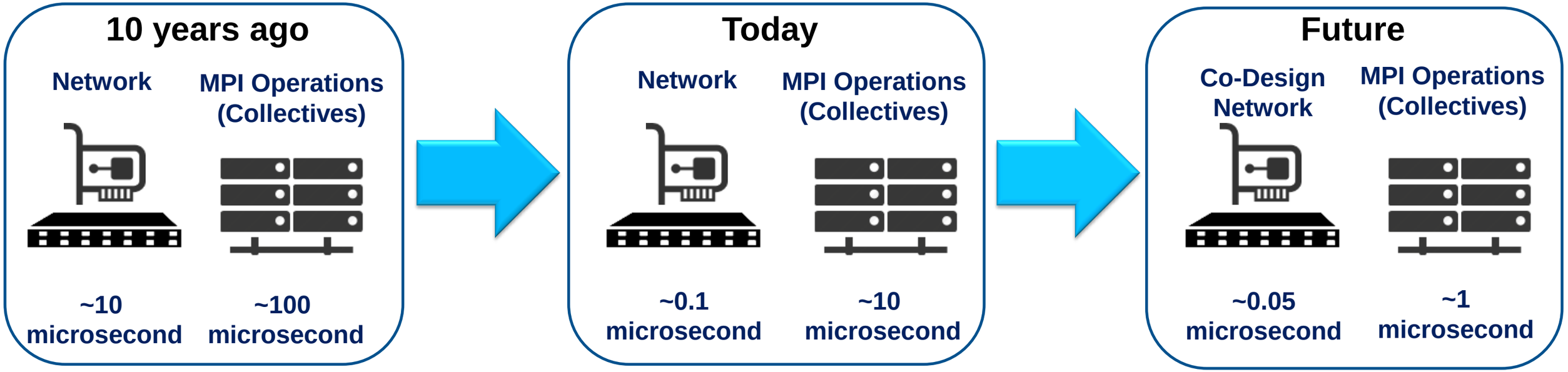



Creating Synergies
Enables Higher Performance and Scale

**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**

# Switch-IB 2 and ConnectX-5 Smart Interconnect Solutions

**SHArP Enables Switch-IB 2 to Execute Data Aggregation / Reduction Operations in the Network**

**Barrier, Reduce, All-Reduce, Broadcast**
**Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND**
**Integer and Floating-Point, 32 / 64 bit**

**Delivering 10X Performance Improvement for MPI and SHMEM/PGAS Communications**

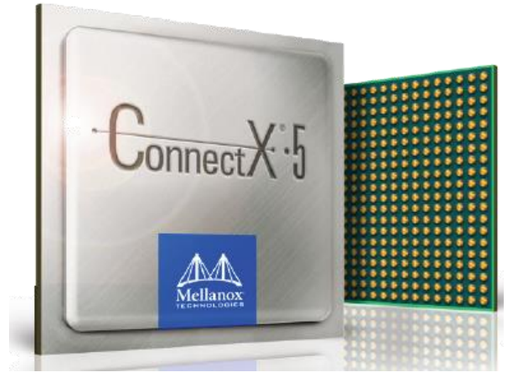
**100Gb/s Throughput**
**0.6usec Latency (end-to-end)**
**200M Messages per Second**

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

**PCIe Gen3 and Gen4**
**Integrated PCIe Switch**
**Advanced Dynamic Routing**

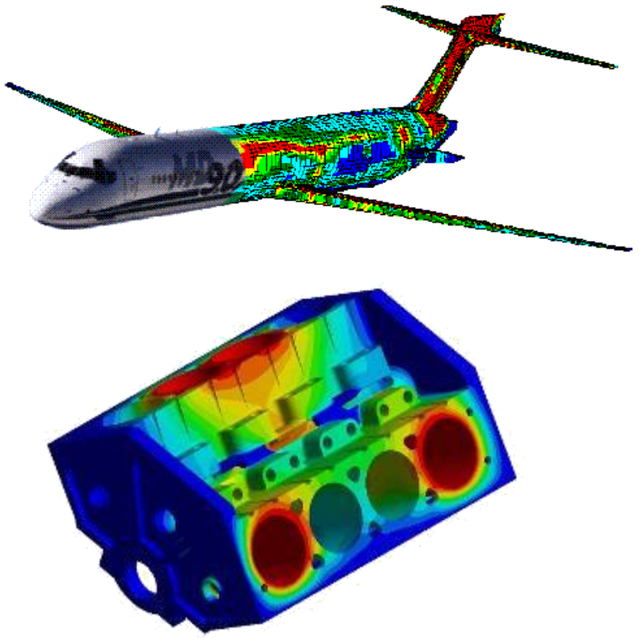
Switch-IB 2 SHArP

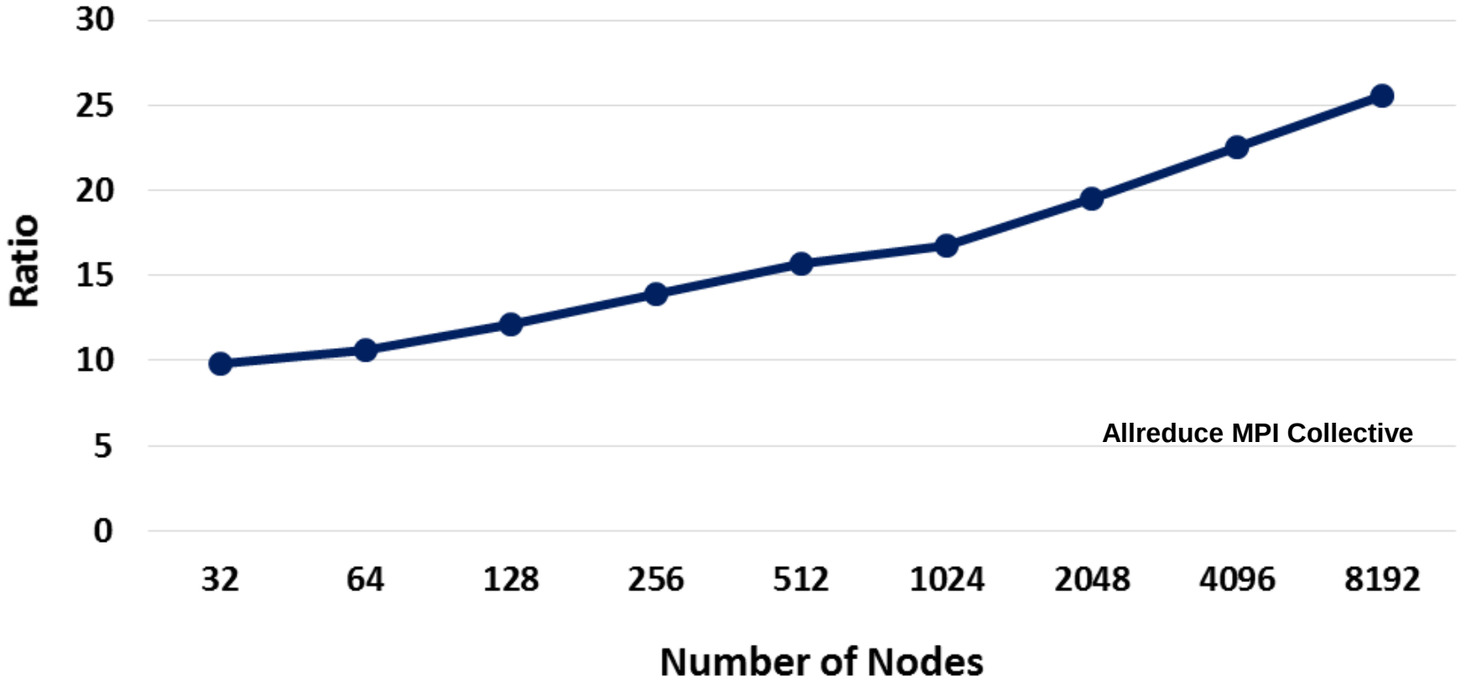
ConnectX-5



# Switch-IB 2 SHArP Performance Advantage

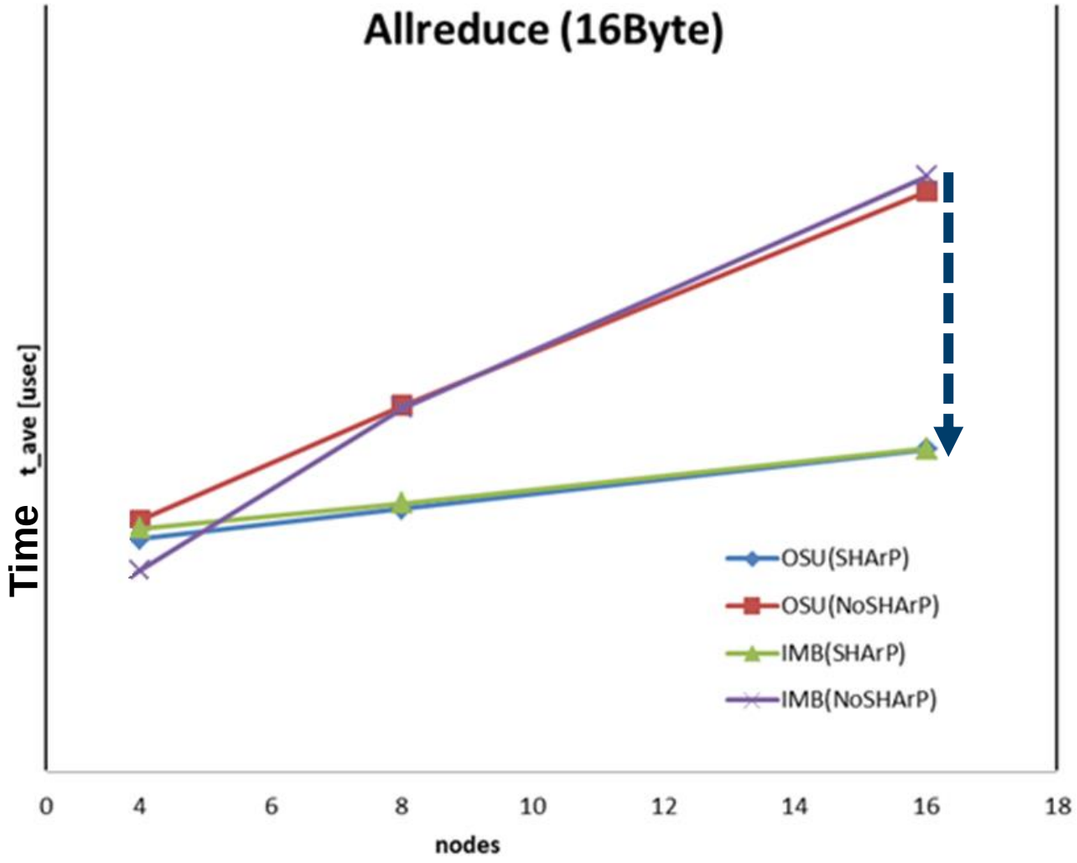
- MiniFE is a Finite Element mini-application
  - Implements kernels that represent implicit finite-element applications






**10X to 25X Performance Improvement**

# SHArP Performance Advantage with Intel Xeon Phi Knight Landing




**Lower is better**

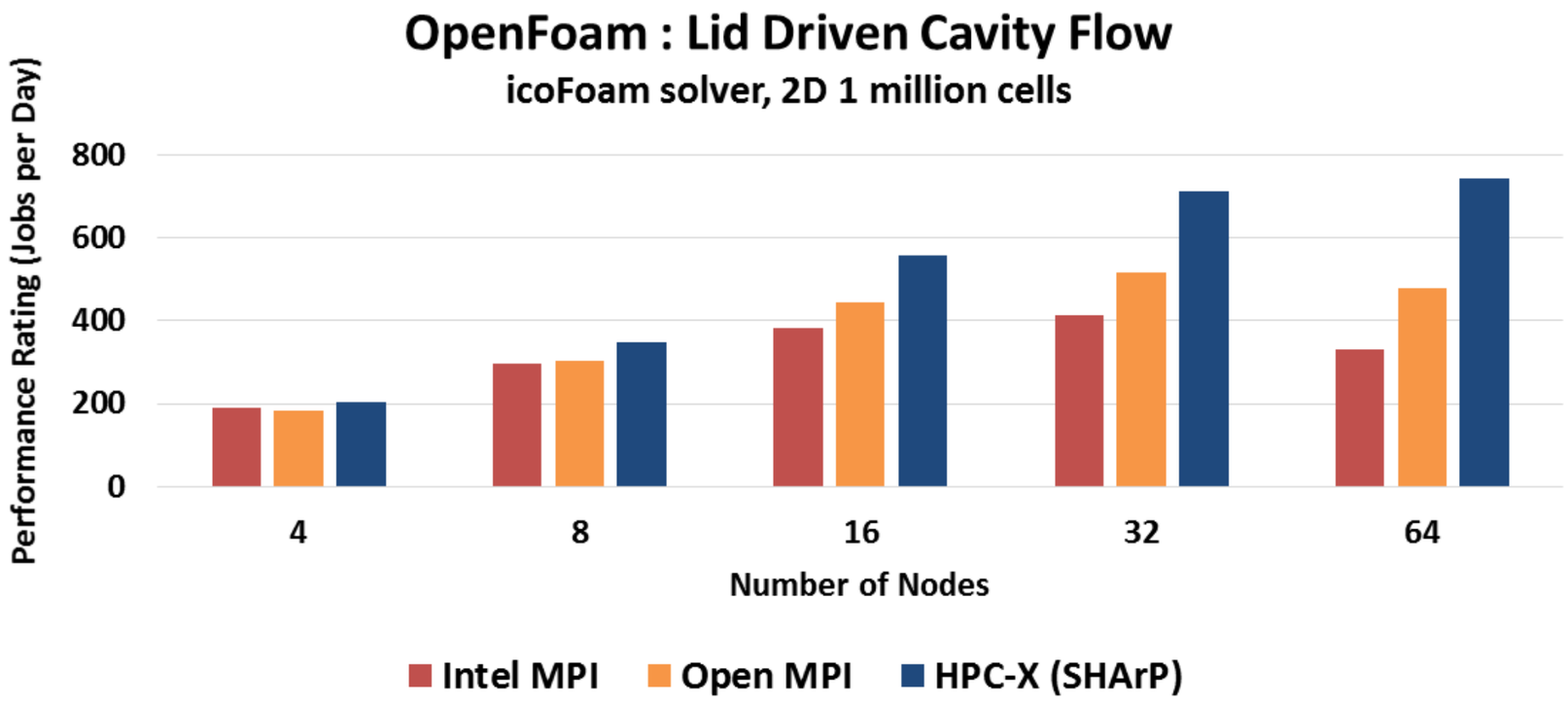
**OSU - OSU MPI benchmark; IMB – Intel MPI Benchmark**

**Maximizing KNL Performance – 50% Reduction in Run Time (Customer Results)**

# HPC-X with SHArP Technology

OpenFOAM

**OpenFOAM is a popular computational fluid dynamics application**




HPC-X™


SHArP

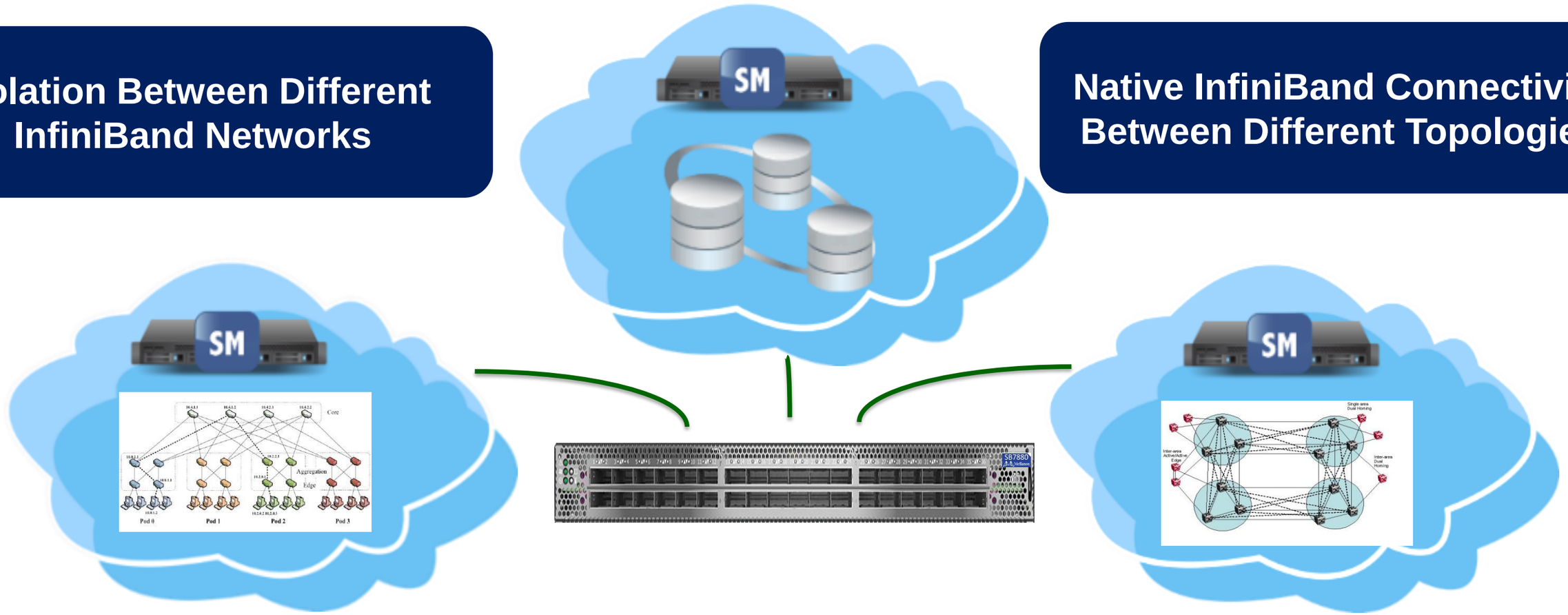
SwitchIB™-2

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

# InfiniBand Router Solutions

**Isolation Between Different InfiniBand Networks**

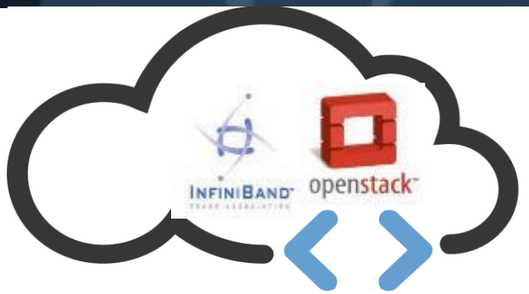
**Native InfiniBand Connectivity Between Different Topologies**

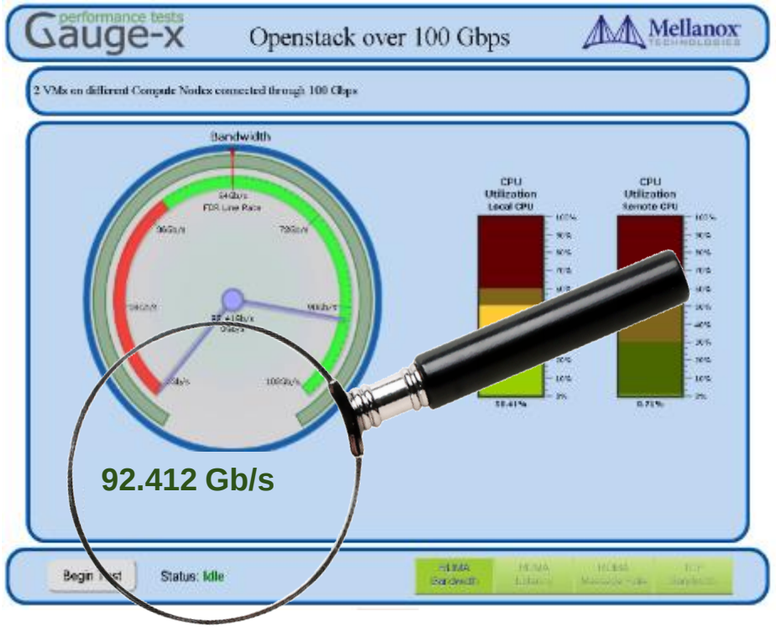


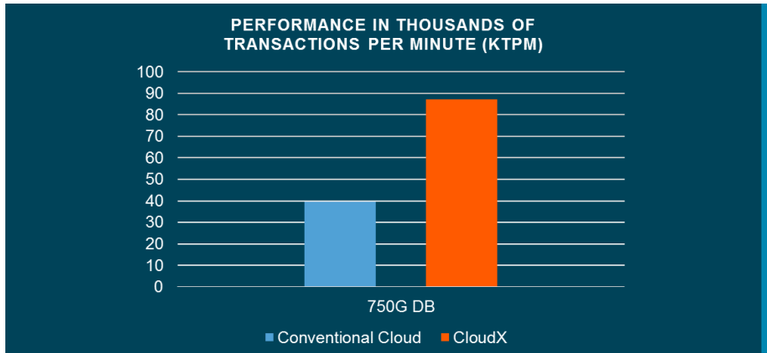

**SB7780 Router 1U**
**Supports up to 6 Different Subnets**

# OpenStack Over InfiniBand – The 100G Cloud is Here!

- Transparent InfiniBand integration into OpenStack
- InfiniBand SDN network
- RDMA directly from VM



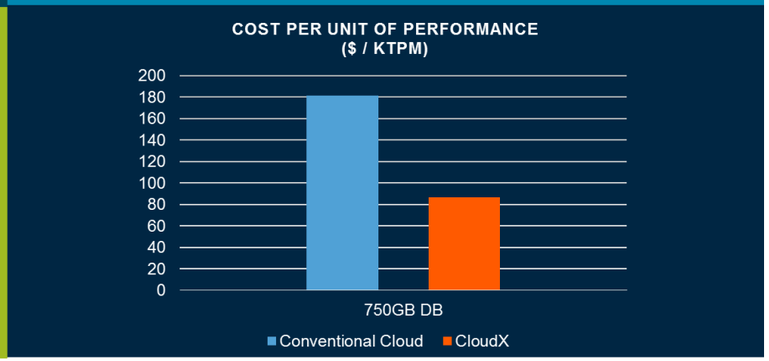




Higher Cost Efficiency;

**52%** Lower Cost per Transaction

**2.2X** OLTP Performance




**The Ideal Fit for HPC Clouds !**

# 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

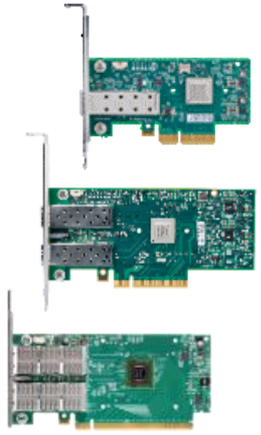
***Watch Video***

# Leading Supplier of End-to-End Interconnect Solutions

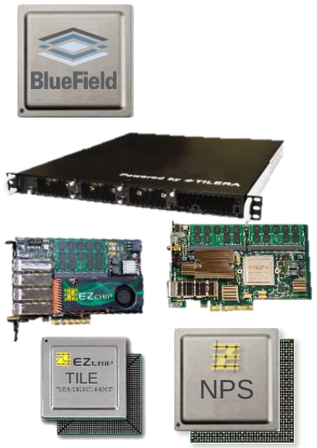

## Comprehensive End-to-End InfiniBand and Ethernet Portfolio (VPI)

ICs

Adapter Cards



NPU & Multicore




Switches/Gateways

Software


Metro / WAN


Cables/Modules