In-Network Computing Technology and Performance Advantages

September 2019

Mellanox TECHNOLOGIES

SUPERCONNECTING the #1 Supercomputers

OAK RIDGE National Laboratory

1 TOP500 The List.

Lawrence Livermore National Laboratory

2 TOP500 The List.

国家超级计算无锡中心 National Supercomputing Center in Wuxi

3 TOP500 The List.

TACC TEXAS ADVANCED COMPUTING CENTER

5 TOP500 The List.

AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

8 TOP500 The List.

Lawrence Livermore National Laboratory

10 TOP500 The List.

InfiniBand Accelerates 6 of Top 10 Supercomputers

Mellanox TECHNOLOGIES

SUPERCONNECTING the #1 Supercomputers

TACC
TEXAS ADVANCED COMPUTING CENTER

FRONTERA

5 TOP500 The List.

MISSISSIPPI STATE UNIVERSITY

NOAA

62 TOP500 The List.

CSC

FINNISH METEOROLOGICAL INSTITUTE

166 TOP500 The List.

筑波大学
University of Tsukuba

264 TOP500 The List.

UNIVERSITY OF MICHIGAN

World's First HDR InfiniBand Supercomputer

सी डैक CDAC

India's National Supercomputing Program

HDR 200G InfiniBand Accelerated Supercomputers

HPC and AI Needs the Most Intelligent Interconnect

SmartNIC

System on a Chip

Innova ConnectX

Adapters

Spectrum Mellanox Quantum

Switches

LinkX

Cables & Transceivers

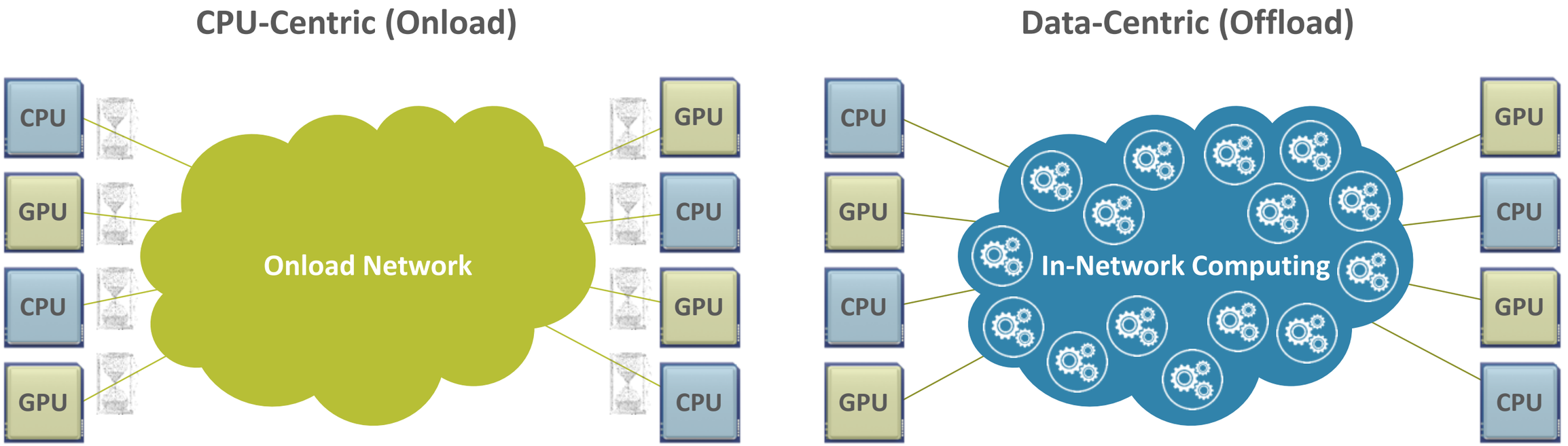
Higher Data Speeds

Faster Data Processing

Better Data Security

The Need for Intelligent and Faster Interconnect

Faster Data Speeds and In-Network Computing
Enable Higher Performance and Scale



Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale

Accelerating All Levels of HPC / AI Frameworks

Application

- Data Analysis
- Real Time
- Deep Learning

Communication

- Mellanox SHARP In-Network Computing
- MPI Tag Matching
- MPI Rendezvous
- Software Defined Virtual Devices

MPI HPC-X™ UCF Unified Communication Framework SHARP Scalable Hierarchical Aggregation and Reduction Protocol

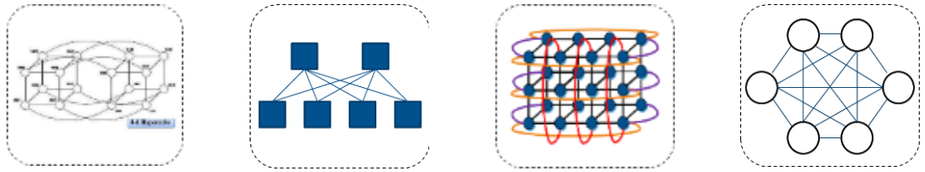
Network

- Network Transport Offload
- RDMA and GPU-Direct RDMA
- SHIELD (Self-Healing Network)
- Enhanced Adaptive Routing and Congestion Control

GPUDirect RDMA SHIELD

Connectivity

- Multi-Host Technology
- Socket-Direct Technology
- Enhanced Topologies

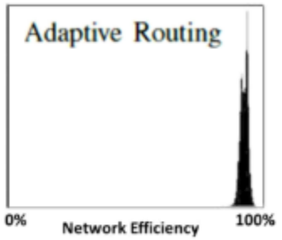


Highest Performance and Scalability for Exascale Platforms

OAK RIDGE National Laboratory

SHARP SHIELD

96%
Network Utilization



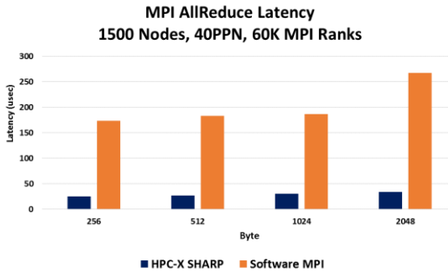
Summit's MpiGraph

UNIVERSITY OF TORONTO

SciNet

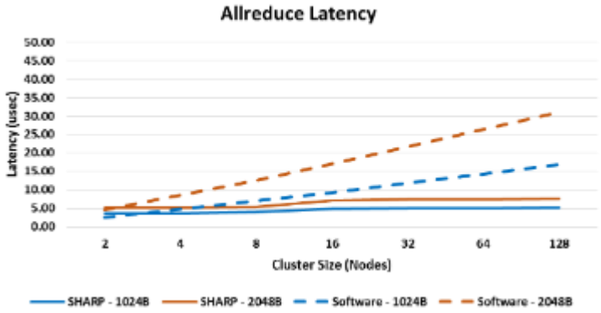
SHARP

7X
Higher Performance

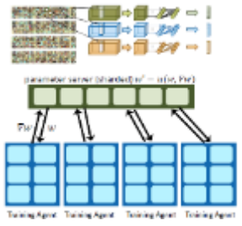


SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

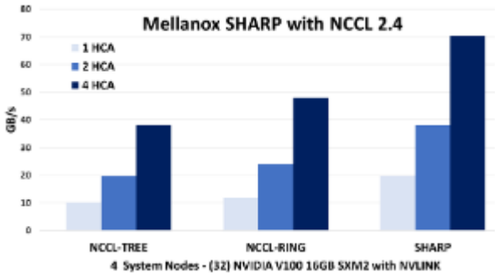
Flat Latency



Deep Learning

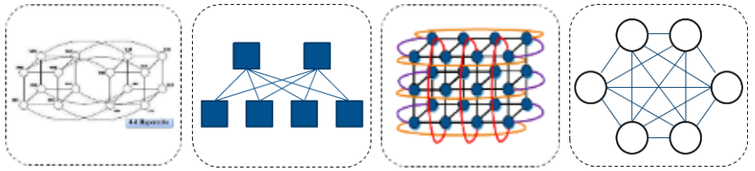


2X
Higher Performance



SHIELD SELF-HEALING

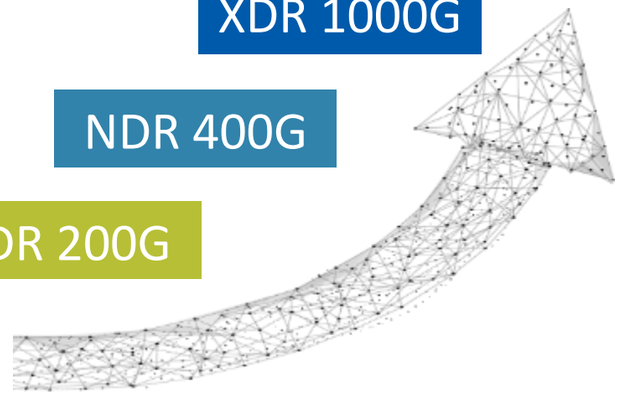
5000X
Higher Resiliency



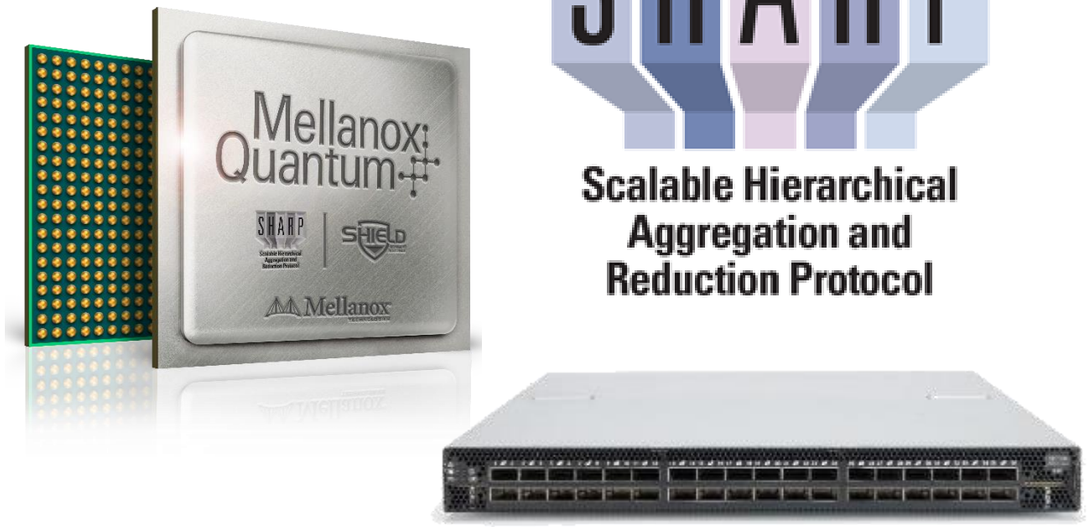
XDR 1000G

NDR 400G

HDR 200G

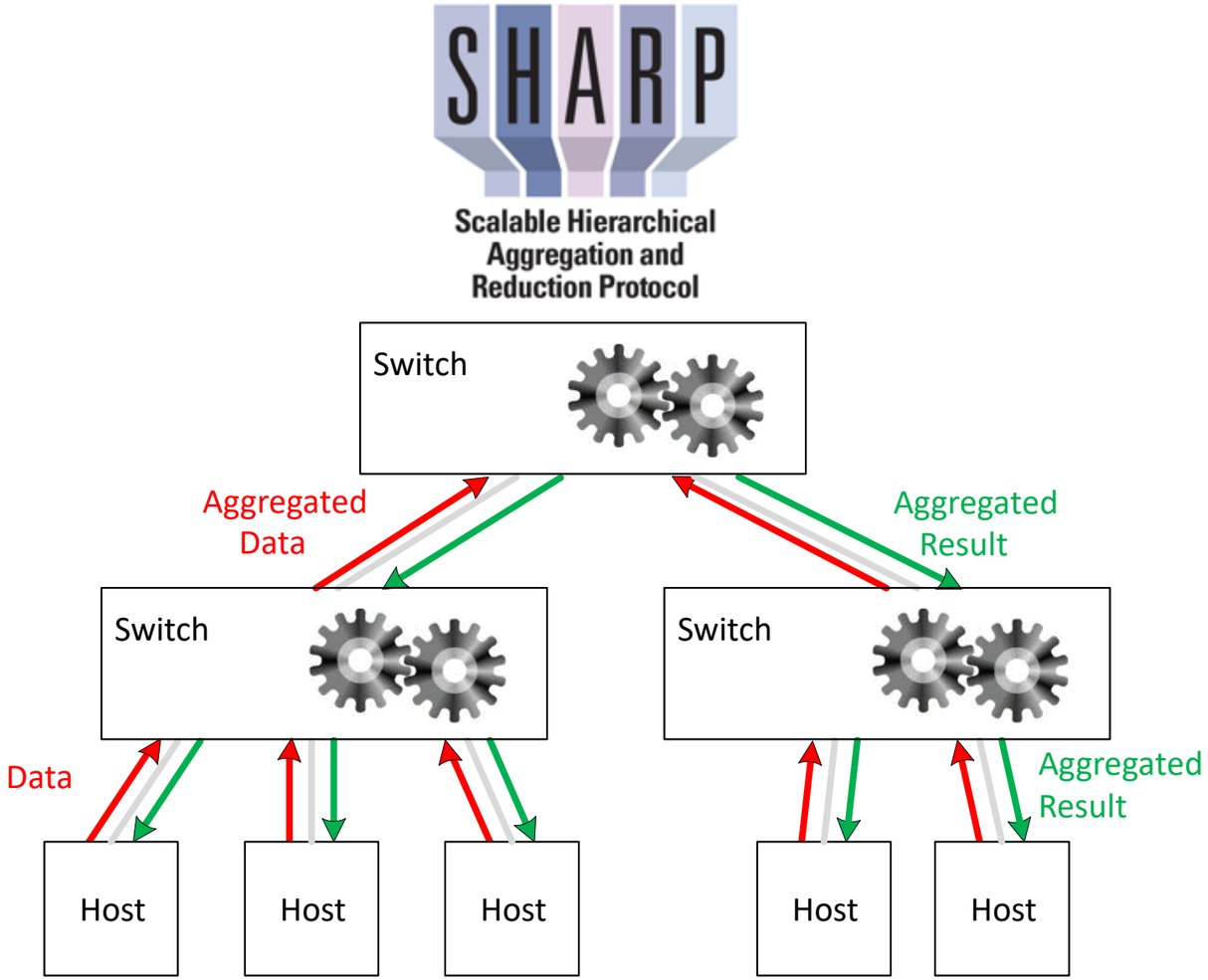


Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

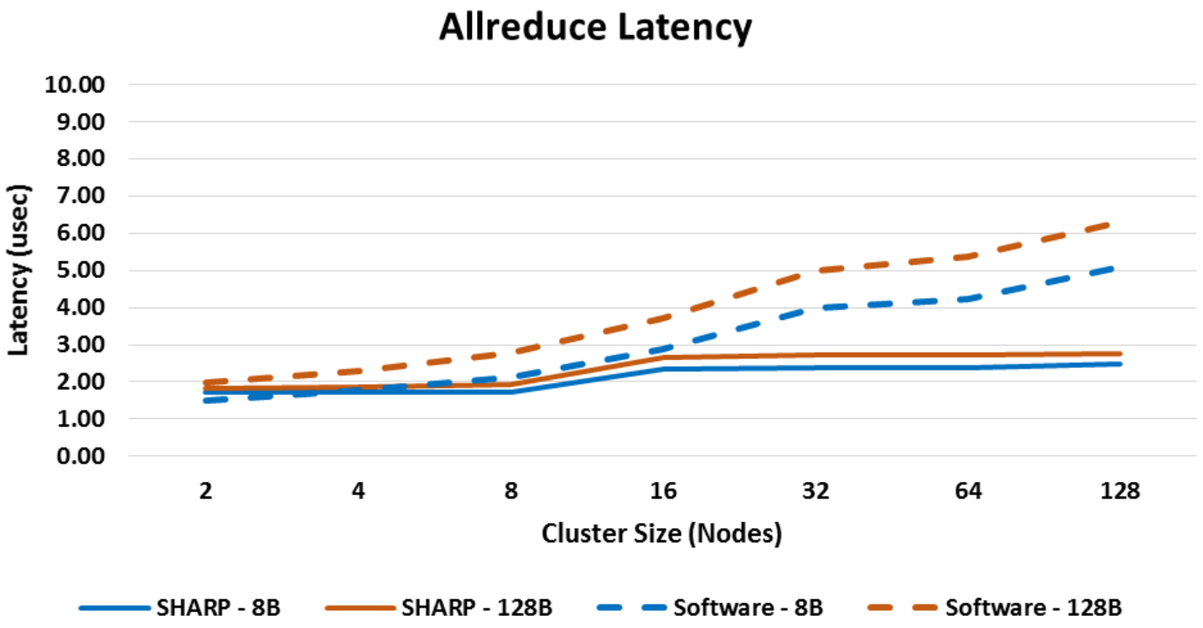


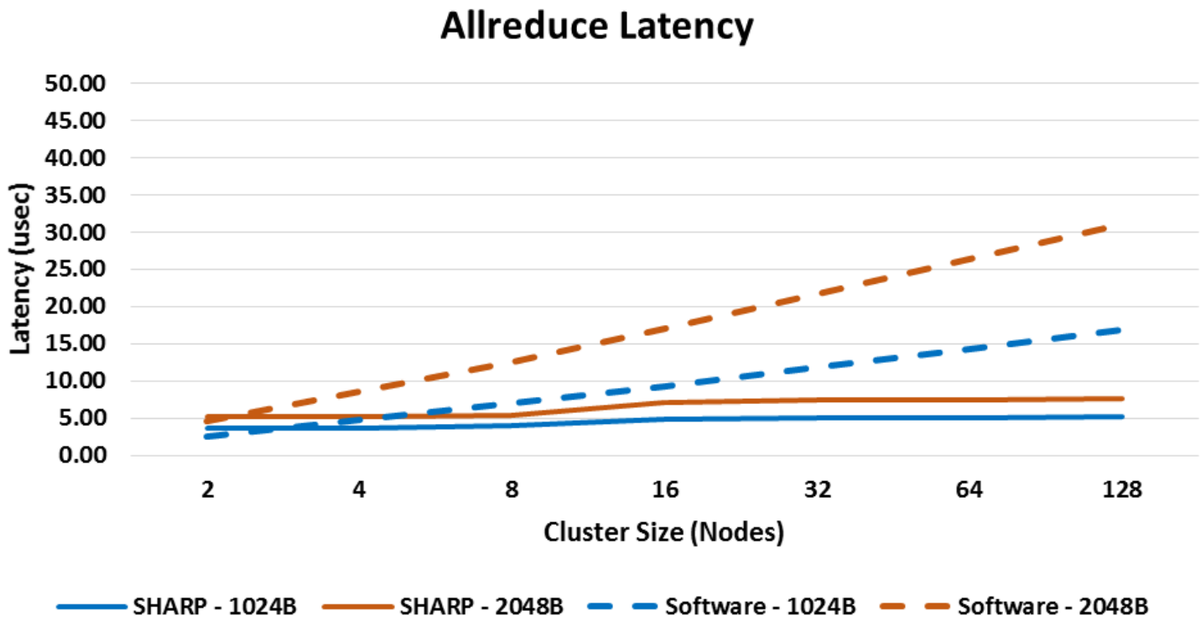
Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations
- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications
- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast and more
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16/32/64 bits



SHARP AllReduce Performance Advantages (128 Nodes)



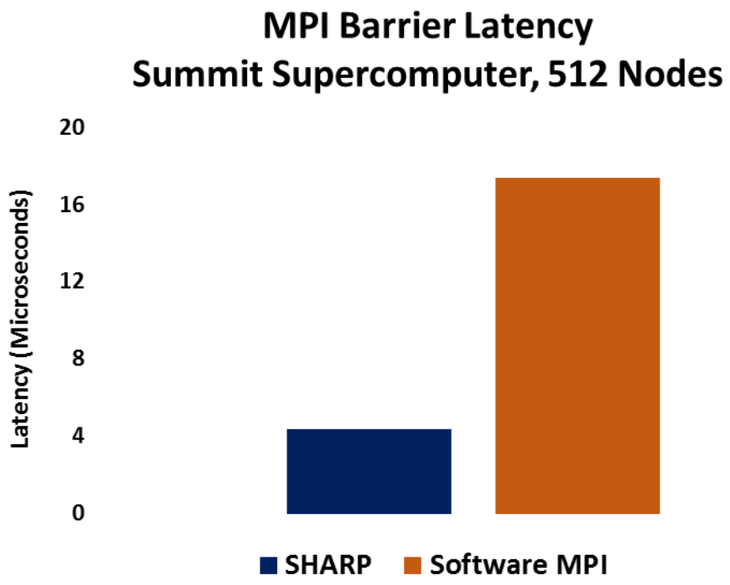


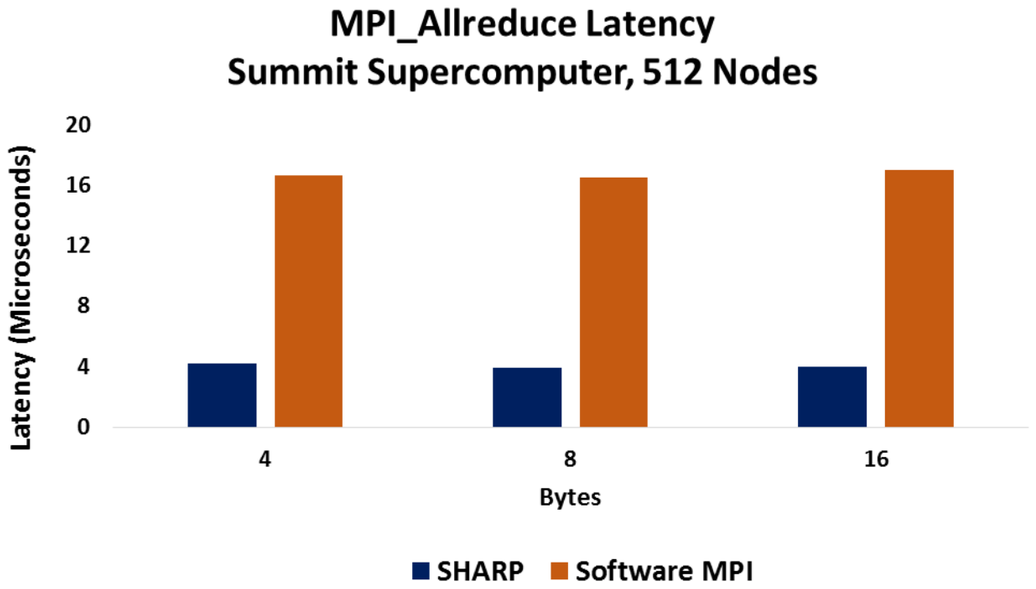
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

SHARP AllReduce Performance Advantages

Oak Ridge National Laboratory – Coral Summit Supercomputer





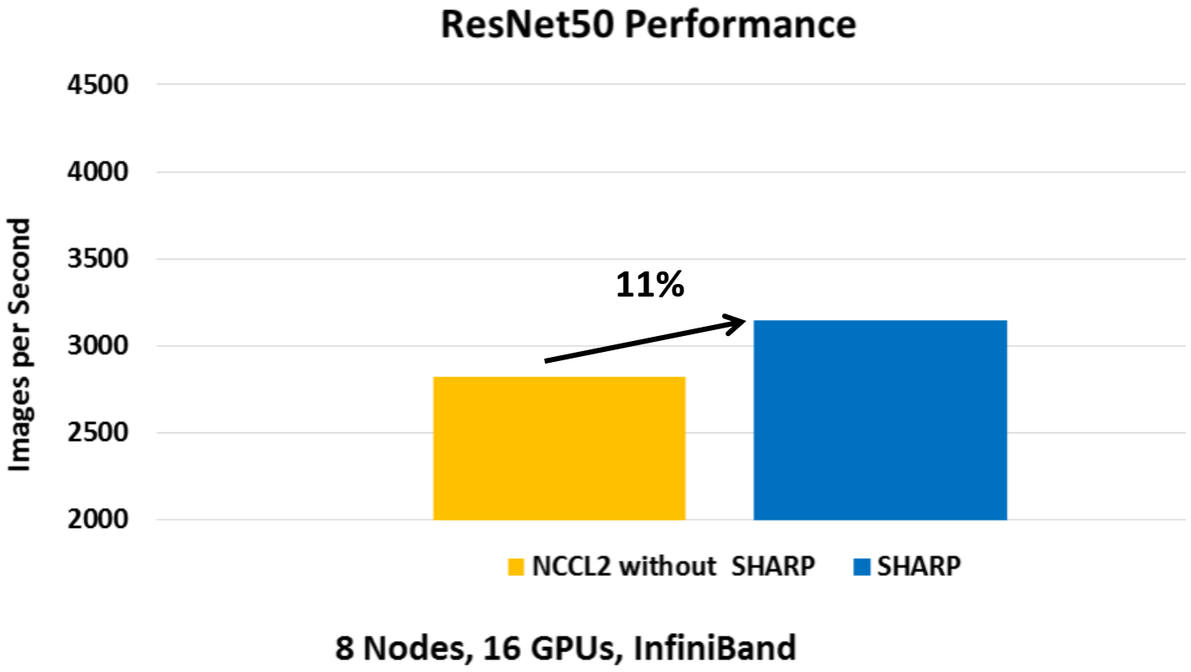
SHARP

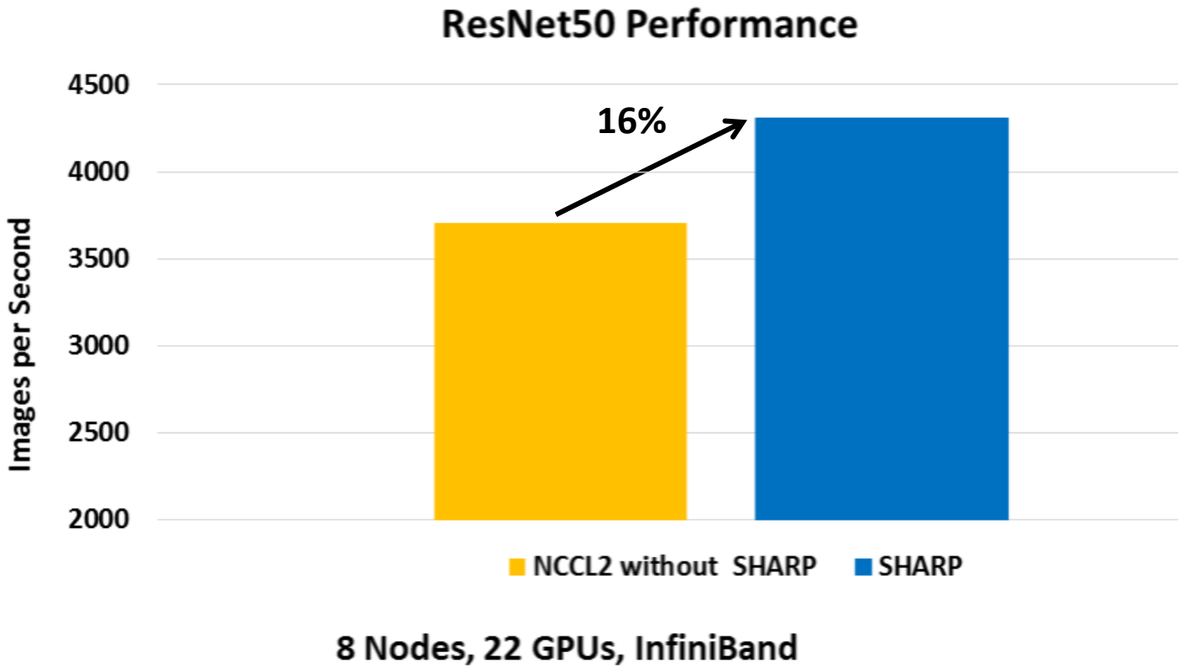
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP Enables Highest Performance

SHARP Performance Advantage for AI

- SHARP provides 16% Performance Increase for deep learning, initial results
- TensorFlow with Horovod running ResNet50 benchmark, HDR InfiniBand (ConnectX-6, Quantum)





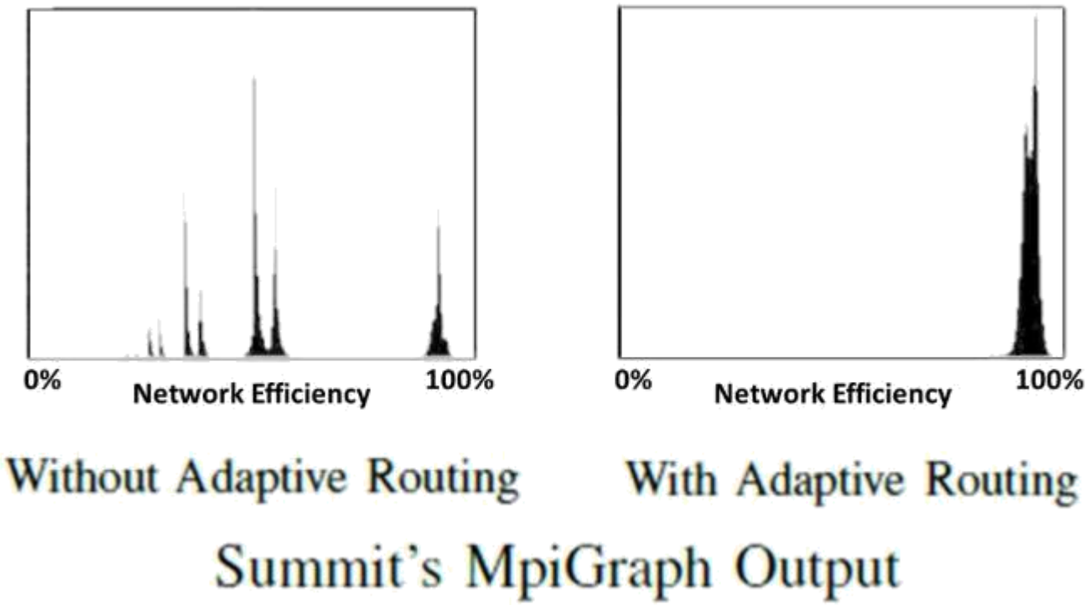
P100 NVIDIA GPUs, RH 7.5, Mellanox OFED 4.4, HPC-X v2.3, TensorFlow v1.11, Horovod 0.15.0

Adaptive Routing

InfiniBand Proven Adaptive Routing Performance

- Oak Ridge National Laboratory – Coral Summit supercomputer
- Bisection bandwidth benchmark, based on mpiGraph
  - Explores the bandwidth between possible MPI process pairs
- AR results demonstrate an average performance of 96% of the maximum bandwidth measured

mpiGraph explores the bandwidth between possible MPI process pairs. In the histograms, the single cluster with AR indicates that all pairs achieve nearly maximum bandwidth while single-path static routing has nine clusters as congestion limits bandwidth, negatively impacting overall application performance.



"The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems", Sudharshan S. Vazhkudai, Arthur S. Bland, Al Geist, Christopher J. Zimmer, Scott Atchley, Sarp Oral, Don E. Maxwell, Veronica G. Vergara Larrea, Wayne Joubert, Matthew A. Ezell, Dustin Leverman, James H. Rogers, Drew Schmidt, Mallikarjun Shankar, Feiyi Wang, Junqi Yin (Oak Ridge National Laboratory) and Bronis R. de Supinski, Adam Bertsch, Robin Goldstone, Chris Chambreau, Ben Casses, Elsa Gonsiorowski, Ian Karlin, Matthew L. Leininger, Adam Moody, Martin Ohmacht, Ramesh Pankajakshan, Fernando Pizzano, Py Watson, Lance D. Weems (Lawrence Livermore National Laboratory) and James Sexton, Jim Kahle, David Appelhans, Robert Blackmore, George Chochia, Gene Davison, Tom Gooding, Leopold Grinberg, Bill Hanson, Bill Hartner, Chris Marroquin, Bryan Rosenburg, Bob Walkup (IBM)

HDR InfiniBand

Highest-Performance 200Gb/s InfiniBand Solutions

Adapters	ConnectX-6	200Gb/s Adapter, 0.6us latency 215 million messages per second (10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)
Switch	Mellanox Quantum	40 HDR (200Gb/s) InfiniBand Ports 80 HDR100 InfiniBand Ports Throughput of 16Tb/s, <90ns Latency
SoC	BlueField	System on Chip and SmartNIC Programmable adapter Smart Offloads
Interconnect	LinkX	Transceivers Active Optical and Copper Cables (10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)
Software	HPC-X	MPI, SHMEM/PGAS, UPC For Commercial and Open Source Applications Leverages Hardware Accelerations

ConnectX-6 HDR InfiniBand Adapter

Leading Connectivity

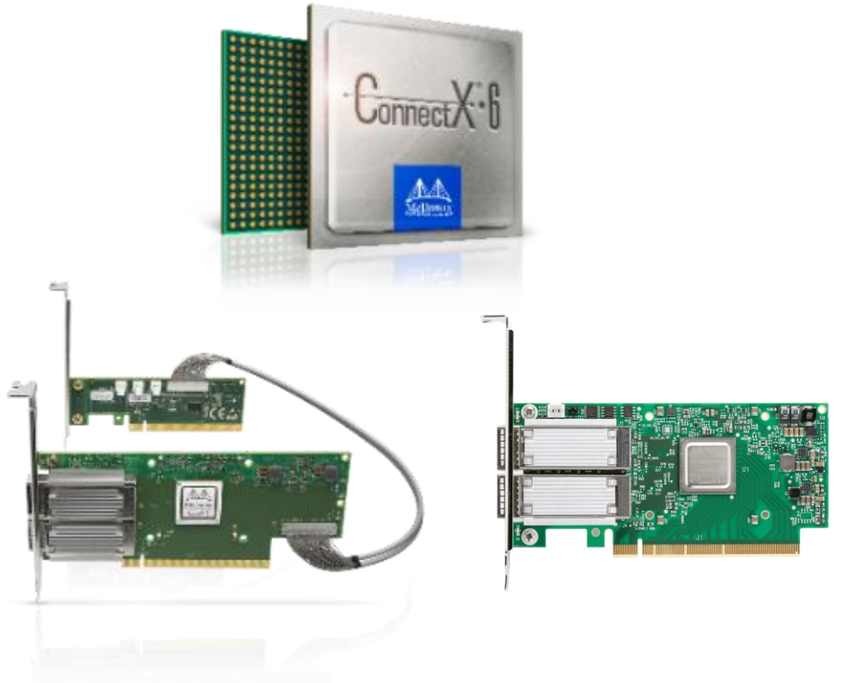
- 200Gb/s InfiniBand and Ethernet
  - HDR, HDR100, EDR (100Gb/s) and lower speeds
  - 200GbE, 100GbE and lower speeds
- Single and dual ports
- 50Gb/s PAM4 SerDes

Leading Performance

- 200Gb/s throughput, 0.6usec latency, 215 million message per second
- PCIe Gen3 / Gen4, 32 lanes
- Integrated PCIe switch
- Multi-Host - up to 8 hosts, supporting 4 dual-socket servers

Leading Features

- In-network computing and memory for HPC collective offloads
- Security – Block-level encryption to storage, key management, FIPS
- Storage – NVMe Emulation, NVMe-oF target, Erasure coding, T10/DIF



HDR InfiniBand Switches

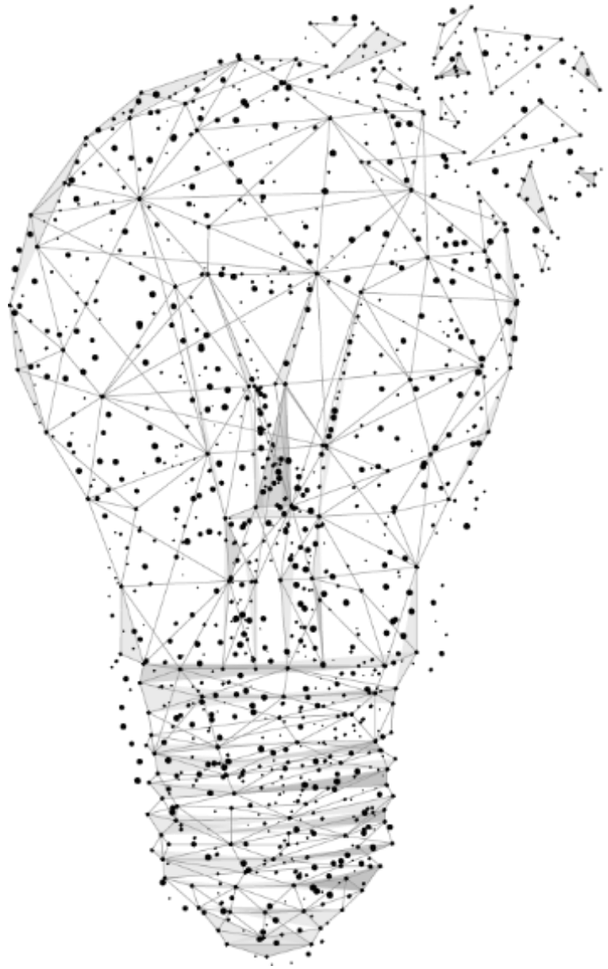
40 QSFP56 ports

- 40 ports of HDR, 200G
- 80 ports of HDR100, 100G

800 QSFP56 ports

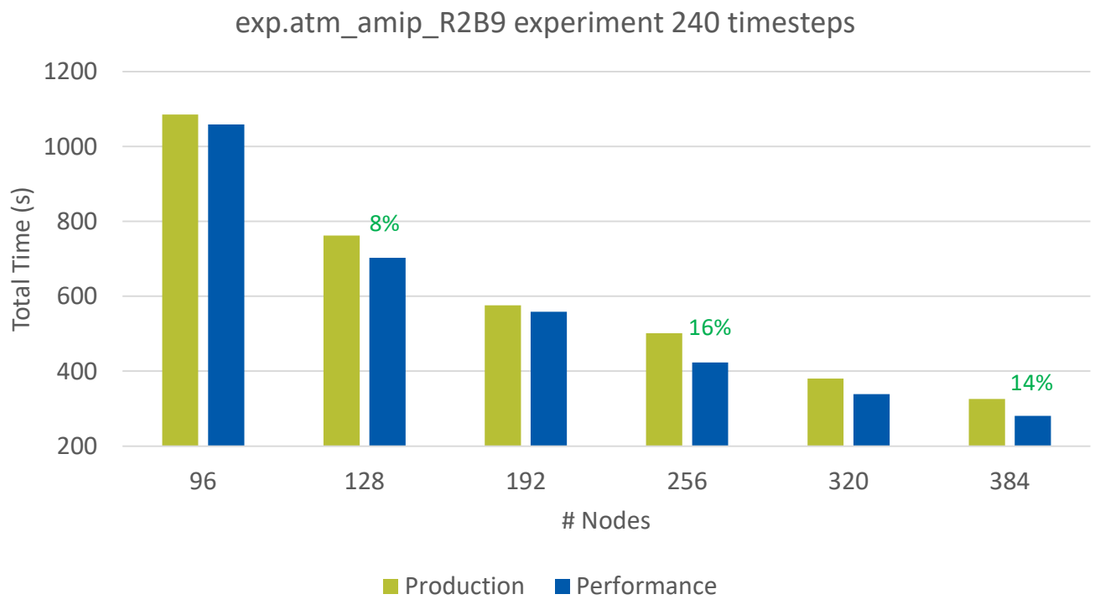
- 800 ports of HDR, 200G
- 1600 ports of HDR100, 100G

Weather and Climate Apps

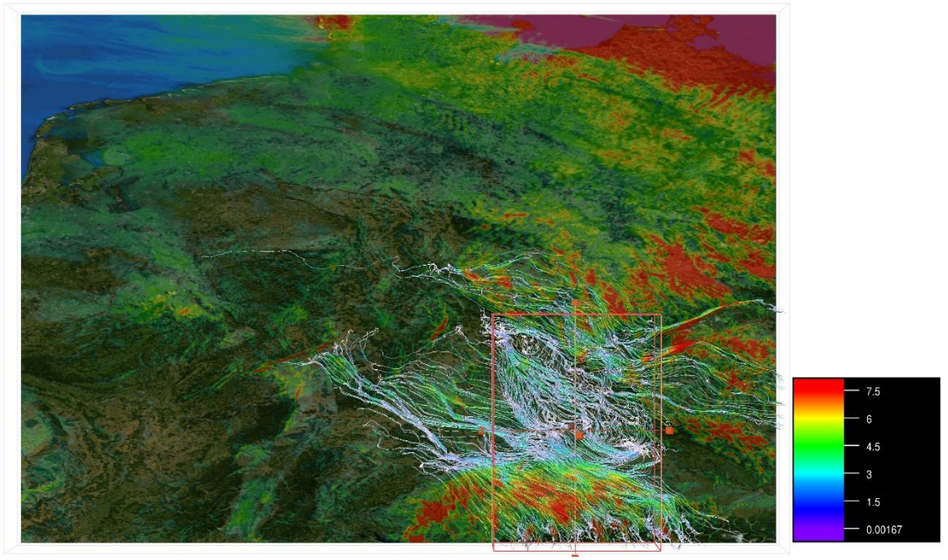


ICON (ICOsahedral Non-hydrostatic model)

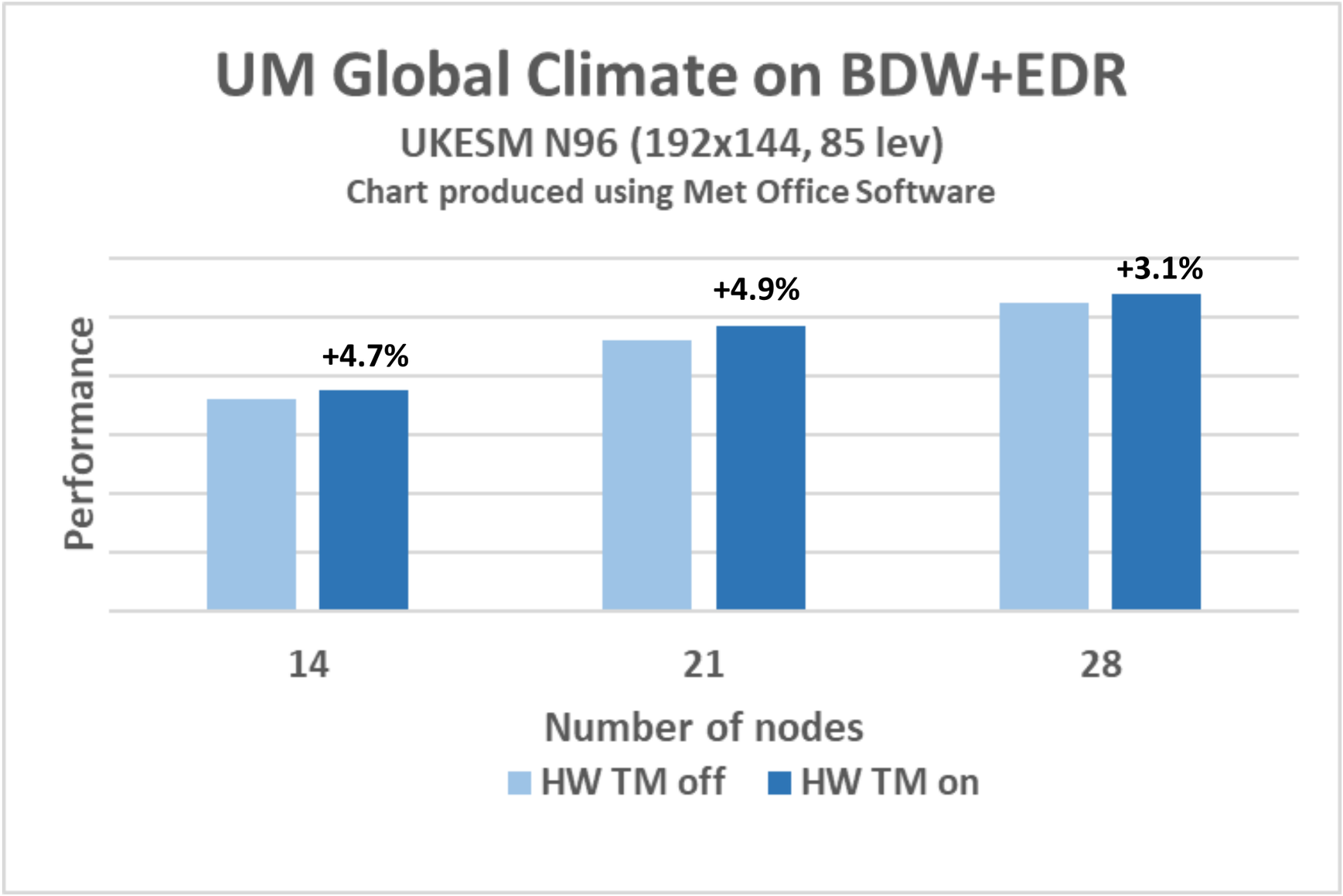
- New generation unified weather forecasting and climate model developed by MPI-M and DWD
- New data exchange module YAXT developed by DRKZ to replace traditional halo exchange mechanism
  - Main challenge lies on efficient handling of sparse data at scale
  - Improvement jointly developed by DKRZ, UTK, and Mellanox



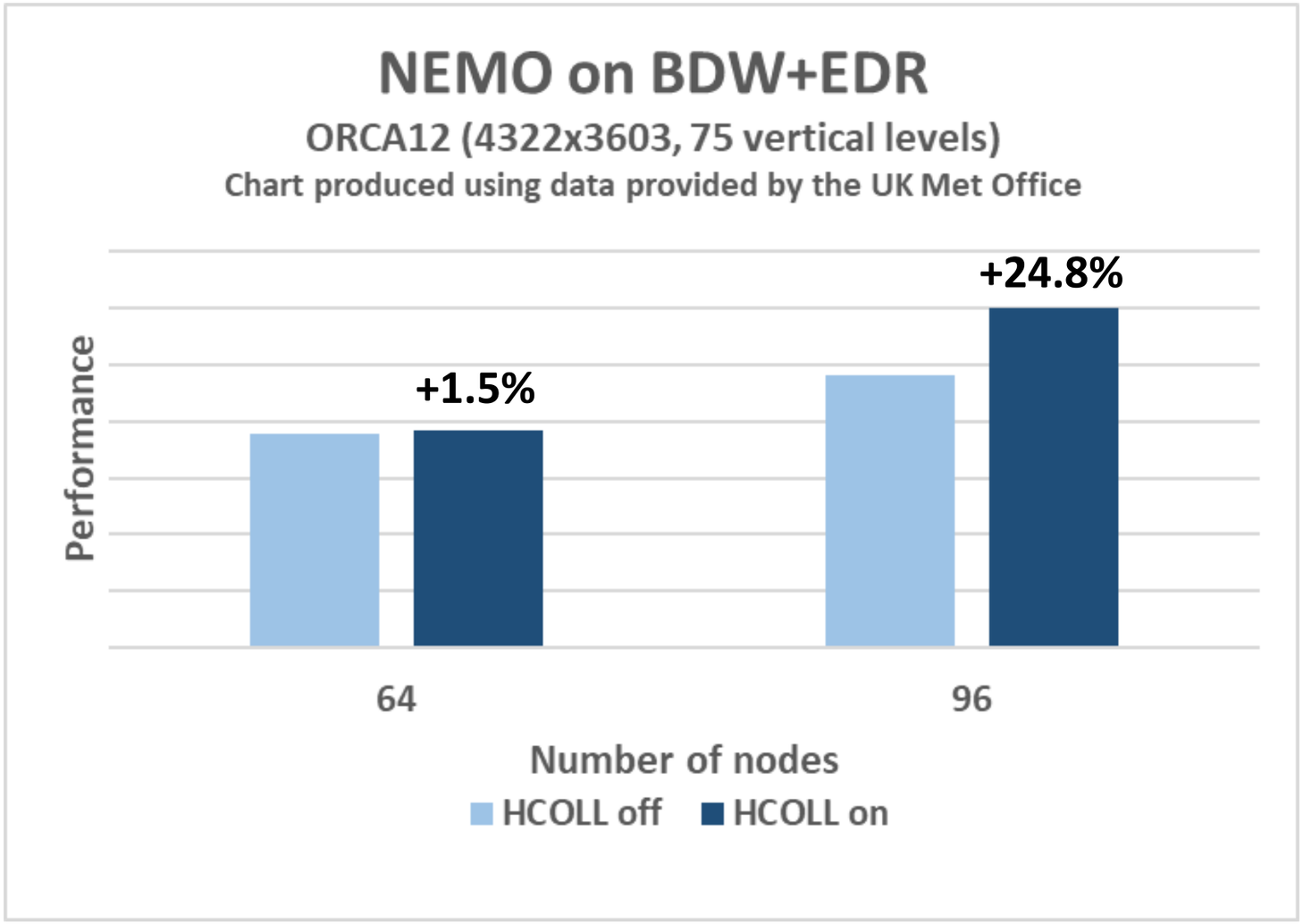
Production: Open MPI 2.0.2p1 + HPC-X 1.9.7 (MXM & HCOLL)
Performance: Open MPI master + HPC-X 2.4.0 (UCX & HCOLL)



UM

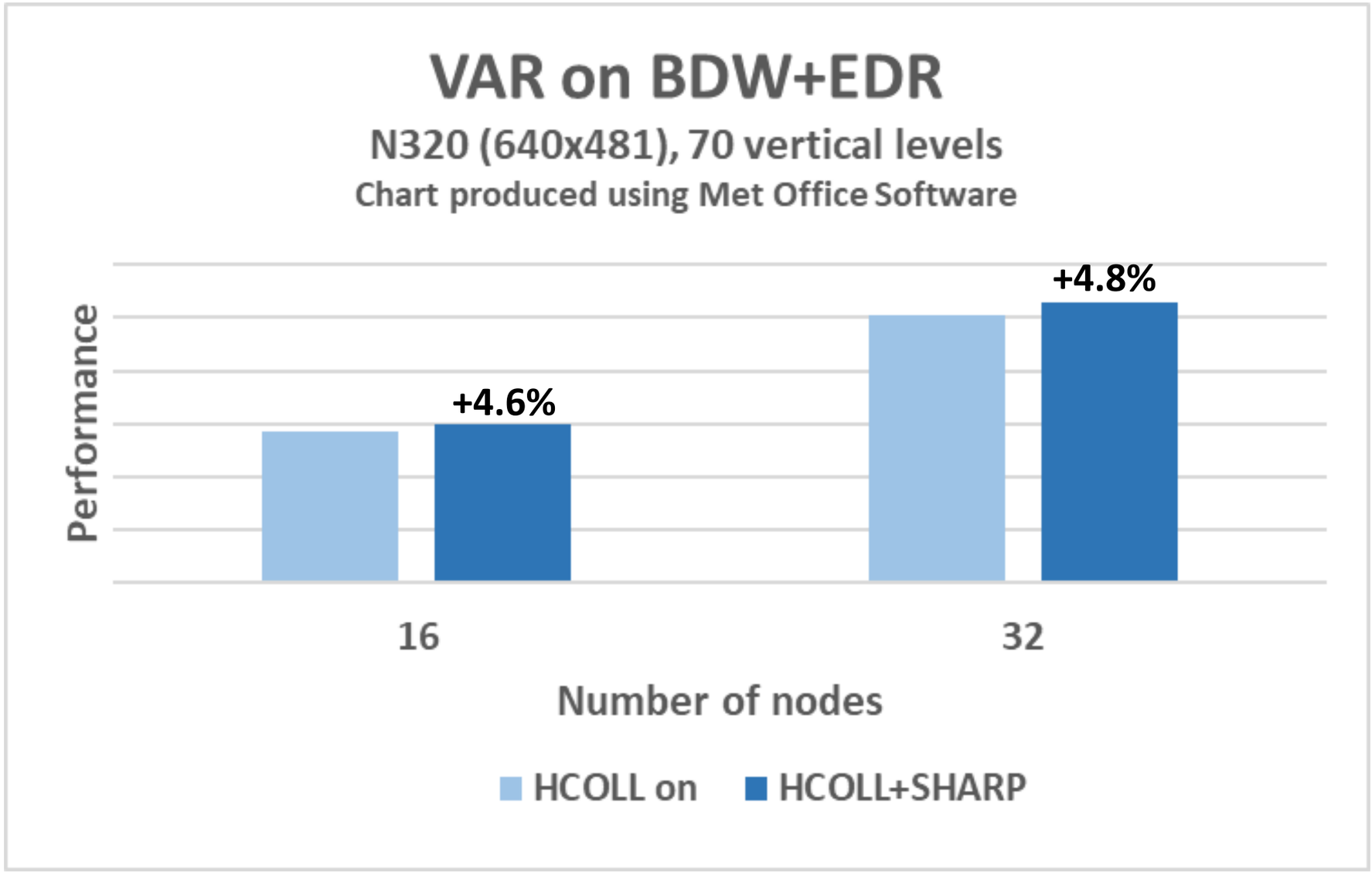

UM Global Climate on BDW+EDR
UKESM N96 (192x144, 85 lev)
Chart produced using Met Office Software
Performance
+4.7%
+4.9%
+3.1%
14
21
28
Number of nodes
HW TM off
HW TM on

NEMO



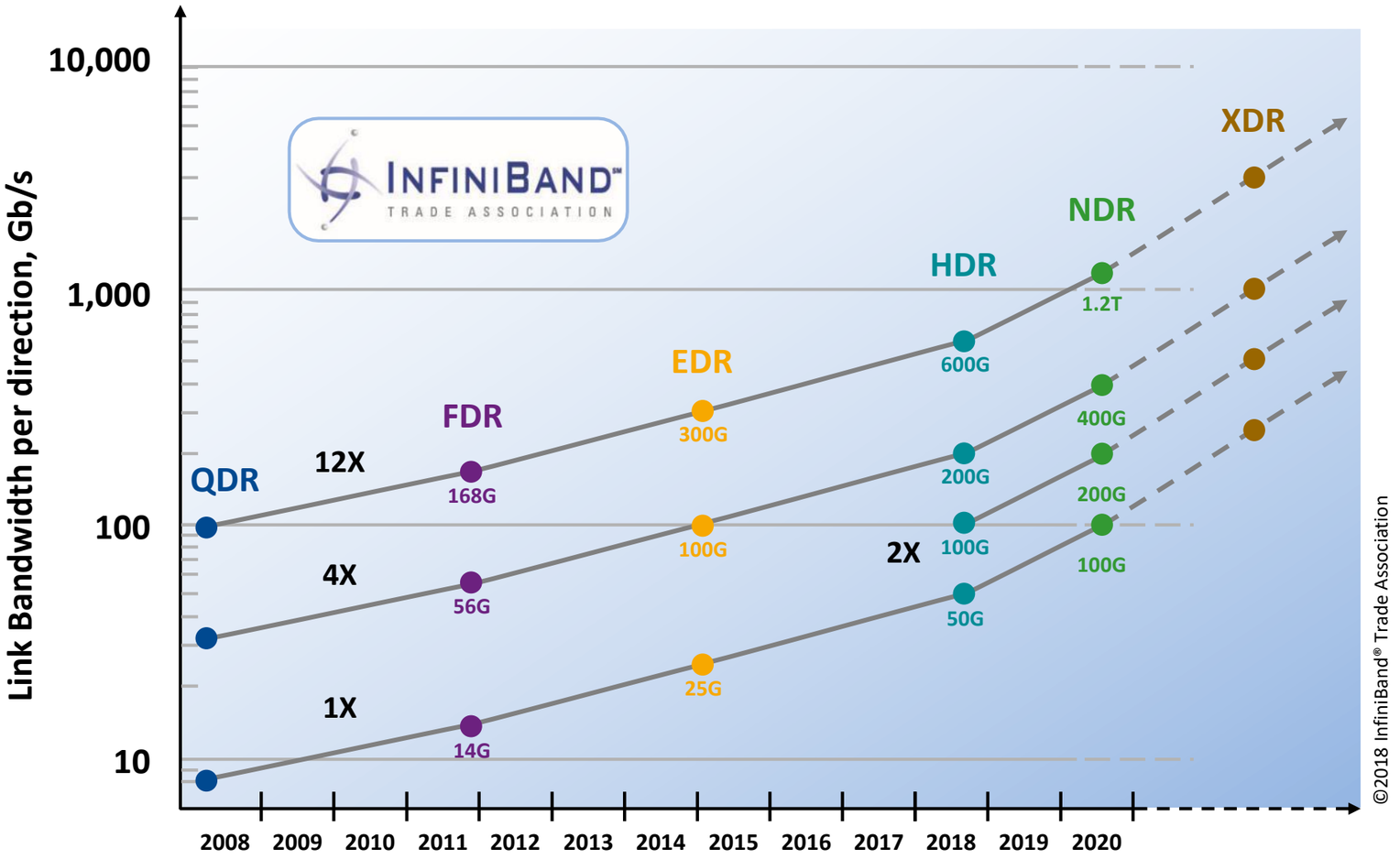
NEMO ocean engine, Madec Gurvan and NEMO System Team, *Issue 27, Scientific Notes of Climate Modelling Center*, Institut Pierre-Simon Laplace (IPSL), ISSN 1288-1619

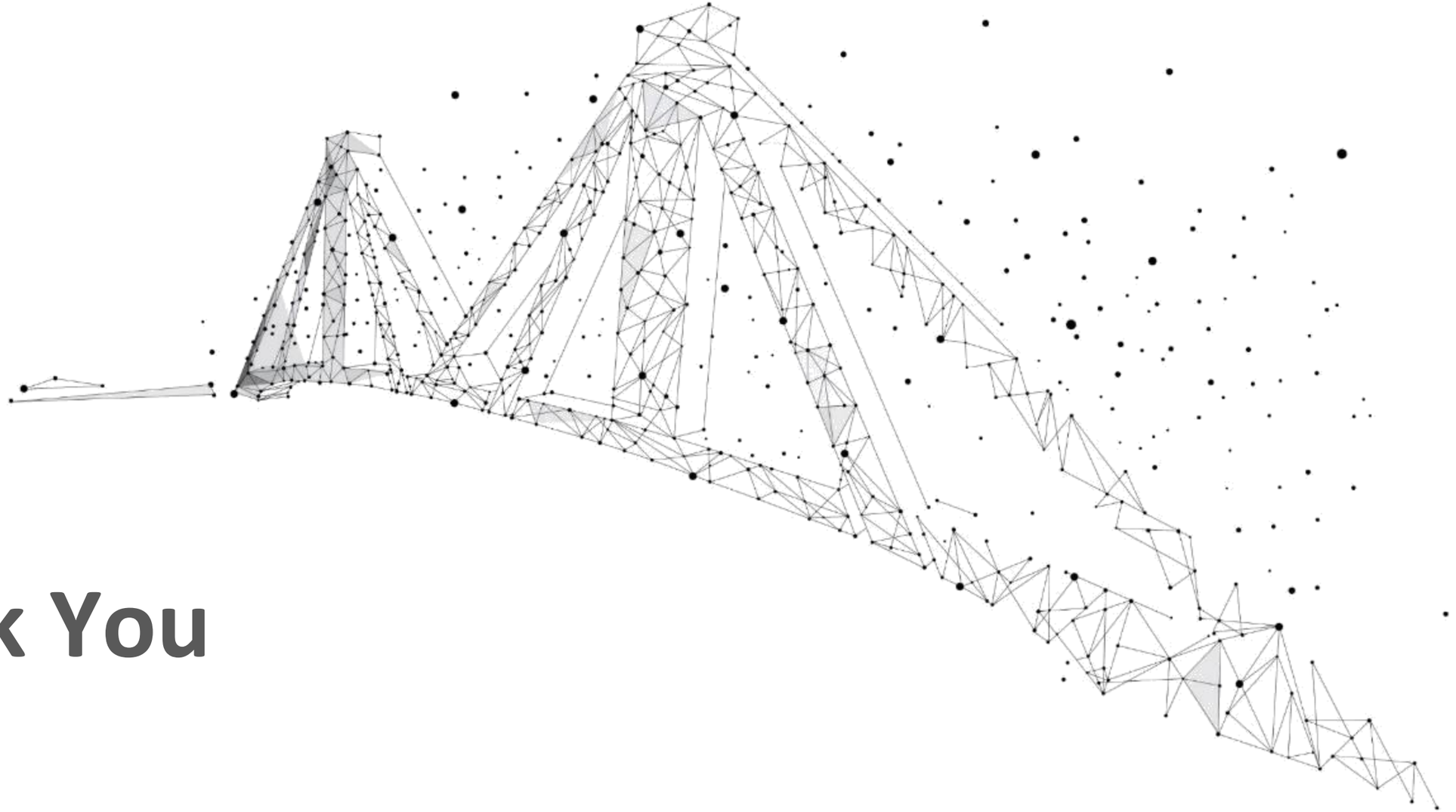
VAR


VAR on BDW+EDR
N320 (640x481), 70 vertical levels
Chart produced using Met Office Software
Performance
+4.6%
+4.8%
16
32
Number of nodes
HCOLL on
HCOLL+SHARP

InfiniBand Roadmap

InfiniBand Roadmap (IBTA)





Thank You