InfiniBand In-Network Computing

Paving the Road to Exascale

June 2019

Mellanox TECHNOLOGIES

SUPERCONNECTING the #1 Supercomputers

OAK RIDGE National Laboratory

1 TOP 500 The List.

Lawrence Livermore National Laboratory

2 TOP 500 The List.

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

3 TOP 500 The List.

TACC TEXAS ADVANCED COMPUTING CENTER

5 TOP 500 The List.

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

8 TOP 500 The List.

Lawrence Livermore National Laboratory

10 TOP 500 The List.

InfiniBand Accelerates 6 of Top 10 Supercomputers

Mellanox TECHNOLOGIES

SUPERCONNECTING the #1 Supercomputers

TACC TEXAS ADVANCED COMPUTING CENTER

NSF FRONTERA

5 TOP 500 The List.

MISSISSIPPI STATE UNIVERSITY

NOAA NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION U.S. DEPARTMENT OF COMMERCE

62 TOP 500 The List.

CSC

FINNISH METEOROLOGICAL INSTITUTE

166 TOP 500 The List.

筑波大学 University of Tsukuba

264 TOP 500 The List.

UNIVERSITY OF MICHIGAN

World's First HDR InfiniBand Supercomputer

सी डैक CDAC

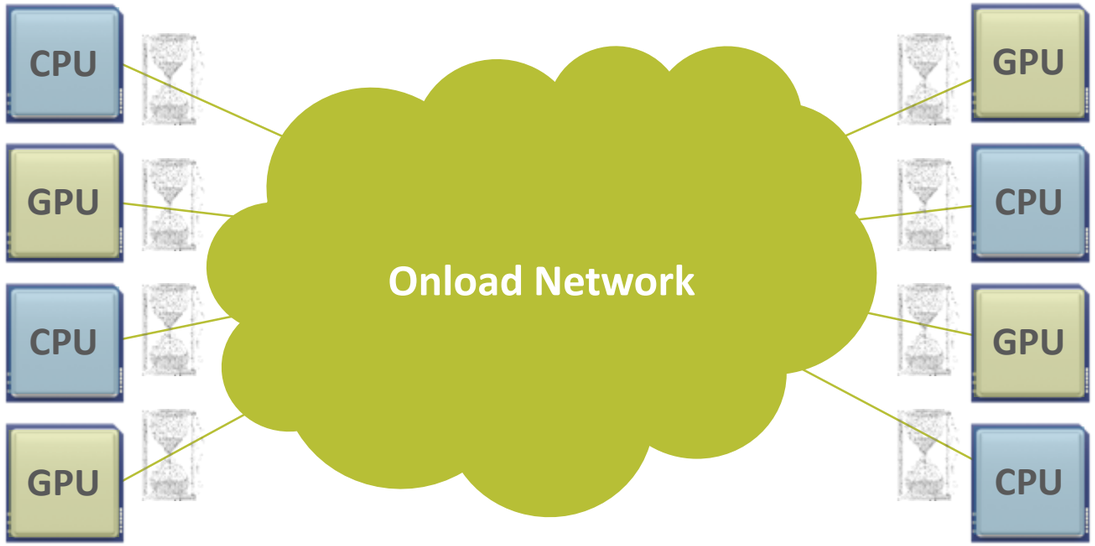
India's National Supercomputing Program

HDR 200G InfiniBand Accelerated Supercomputers

The Need for Intelligent and Faster Interconnect

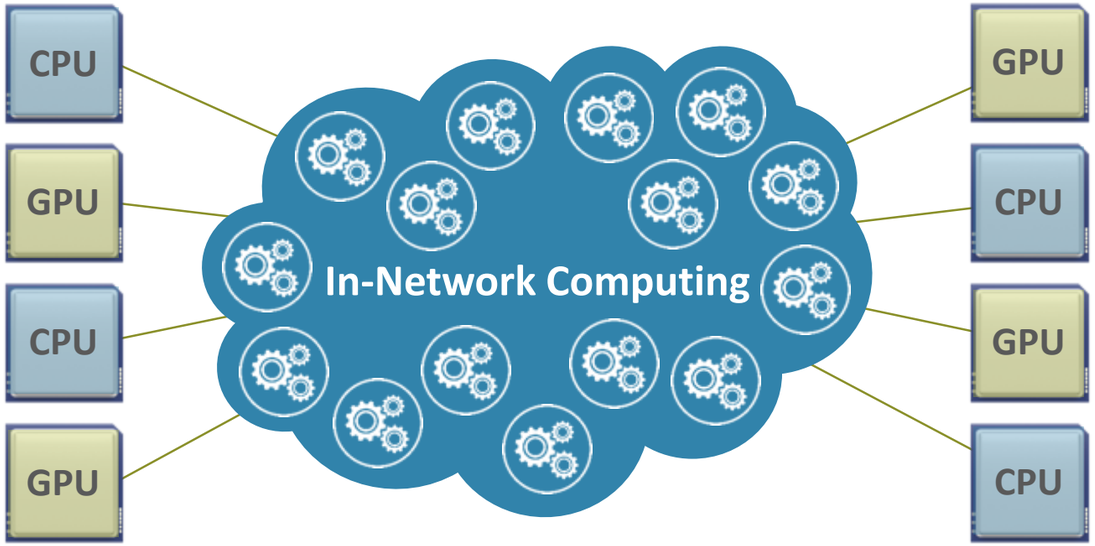
Faster Data Speeds and In-Network Computing
Enable Higher Performance and Scale

CPU-Centric (Onload)



Must Wait for the Data
Creates Performance Bottlenecks

Data-Centric (Offload)



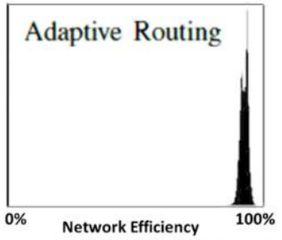
Analyze Data as it Moves!
Higher Performance and Scale

Highest Performance and Scalability for Exascale Platforms

OAK RIDGE National Laboratory

SHARP SHIELD

96%
Network Utilization



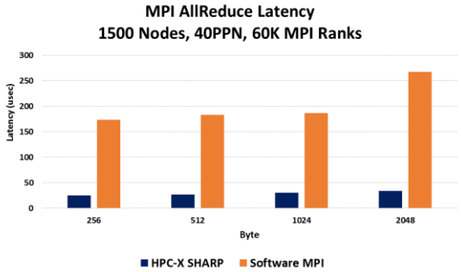
Summit's MpiGraph

UNIVERSITY OF TORONTO

SciNet

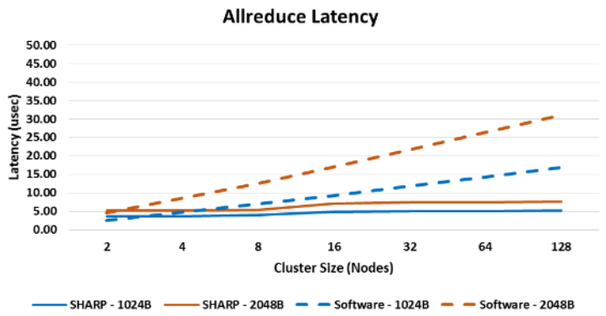
SHARP

7X
Higher Performance

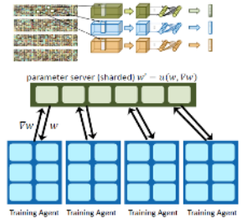


SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

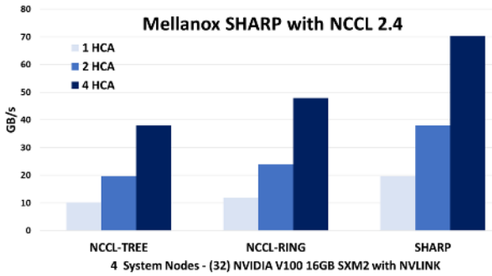
Flat Latency



Deep Learning

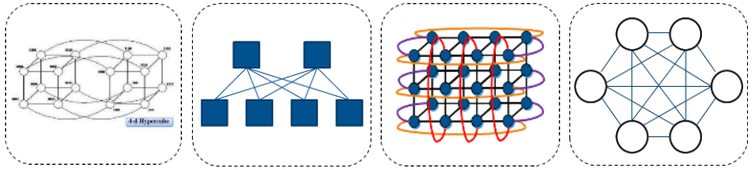


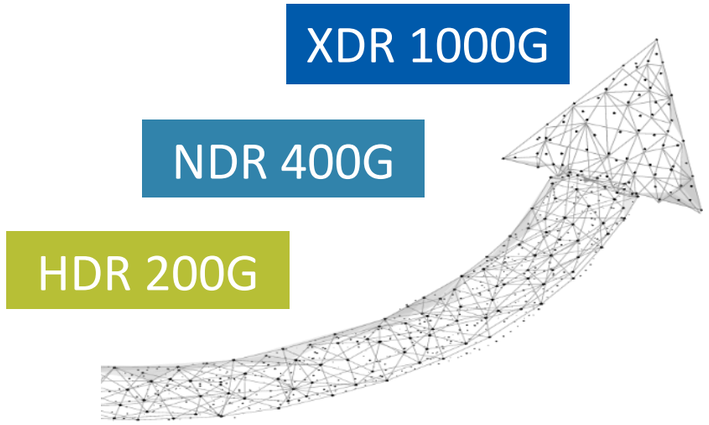
2X
Higher Performance



SHIELD SELF-HEALING

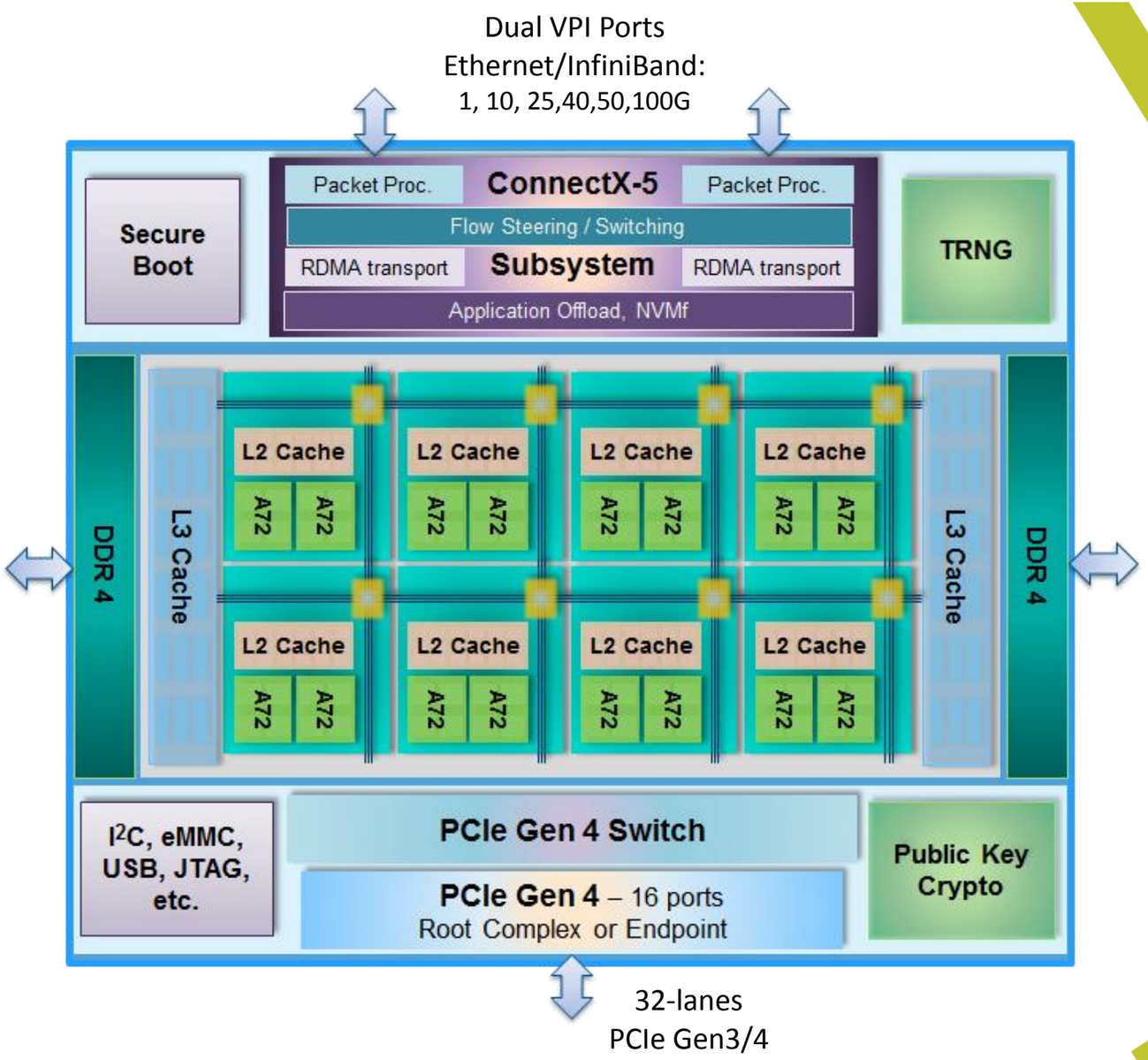
5000X
Higher Resiliency





BlueField Block Diagram

- Tile Architecture - 16 ARM® A72 CPUs subsystem
- Dual Port 100g IO Controller, based on ConnectX-5
  - Dual 100Gb/s Ethernet/InfiniBand, compatible with ConnectX-5
  - NVMe-oF hardware accelerator
  - High-end Networking Offloads: RDMA, Erasure Coding, T10-DIF
- Fully Integrated PCIe switch
  - 32 PCI Gen3/4 lanes
  - Root Complex or Endpoint modes
  - 2x16, 4x8, 8x4 or 16x2 configurations
- Memory Controllers
  - 2x Channels DDR4 Memory Controllers w/ ECC
  - NVDIMM-N Support



BlueField for Smart Solutions

BlueField SoC (System on Chip)

- SoC: Compute, networking and PCIe connectivity
  - Dual port VPI EDR/100GbE
  - 16 Arm cores
  - 32 lanes of PCIe switch gen3/4

Storage Solutions

- NVMe-based storage platforms
  - RDMA, NVMe over Fabrics, RAID, Signature offload
- Partner's solutions based on BlueField storage controller

Smart Adapters

- In-network computing and collective offloads
- Co-processor running proprietary smart algorithms
- Security and privacy algorithms

Proud to be Part of Astra