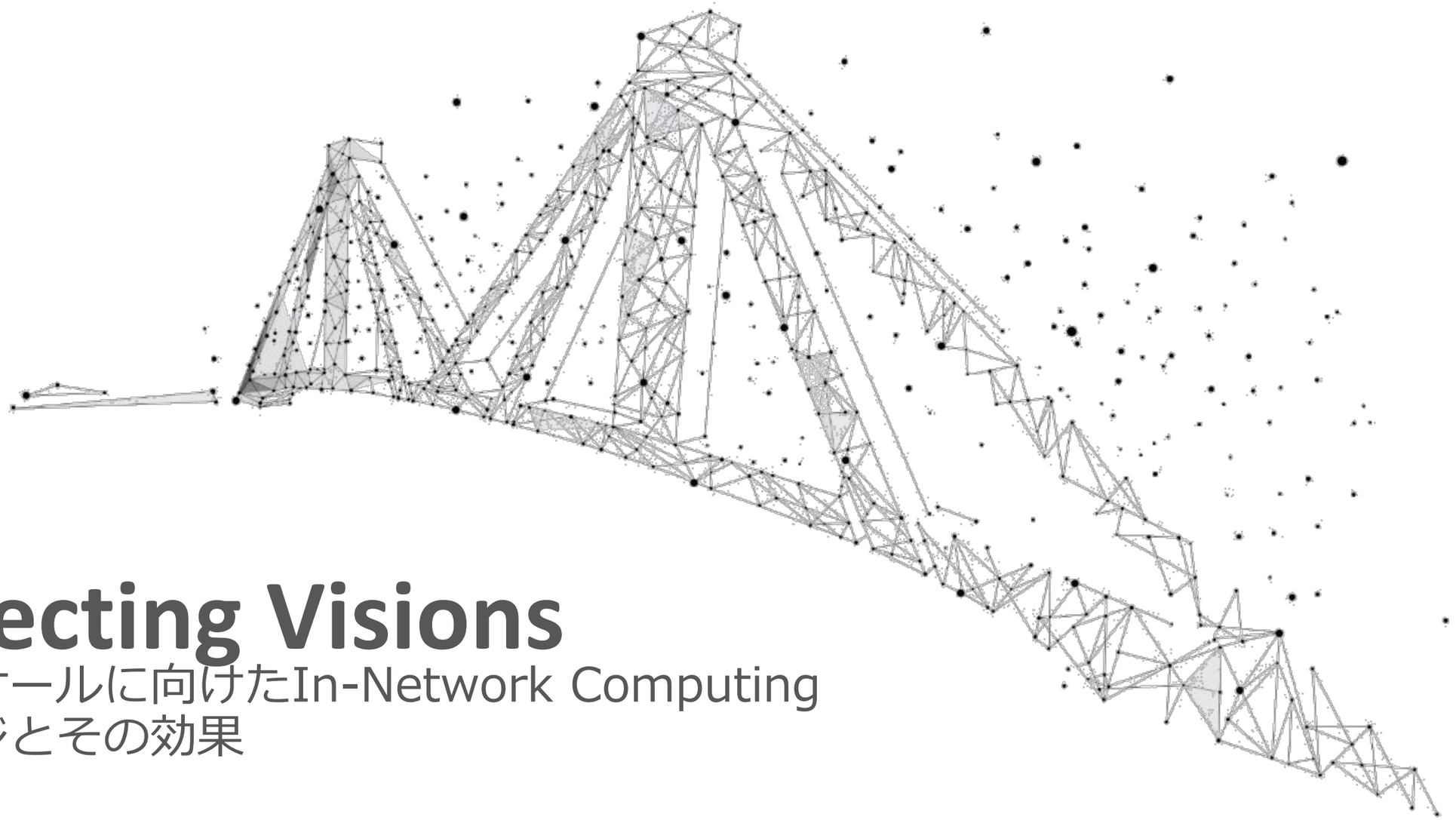


Connecting Visions

エクサスケールに向けたIn-Network Computing テクノロジとその効果

December 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は上位10システム中6つのスーパーコンピュータで採用されました。

HDR 200G InfiniBandが次世代スパコンで続々採用に

TACC
TEXAS ADVANCED COMPUTING CENTER

23.5 Petaflops
8K HDR InfiniBand Nodes
Fat-Tree Topology

KMA Korea Meteorological Administration

50 Petaflops
7.2K HDR InfiniBand Nodes
Dragonfly+ Topology

Australian National University

3K HDR InfiniBand Nodes
Dragonfly+ Topology

MISSISSIPPI STATE UNIVERSITY

3.1 Petaflops
1.8K HDR InfiniBand Nodes
Fat-Tree Topology

CSC
FINNISH METEOROLOGICAL INSTITUTE

1.7 Petaflops
2K HDR InfiniBand Nodes
Dragonfly+ Topology

Microsoft Azure

Highest Performance Cloud

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

PITTSBURGH SUPERCOMPUTING CENTER

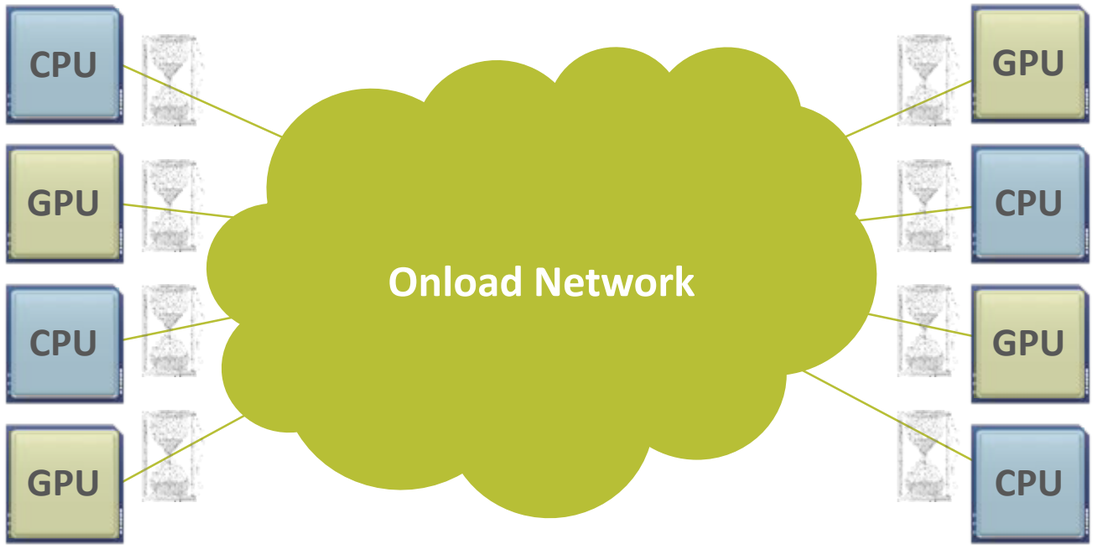
筑波大学
University of Tsukuba

1.6 Petaflops
Hybrid CPU-GPU-FPGA
Fat-Tree Topology

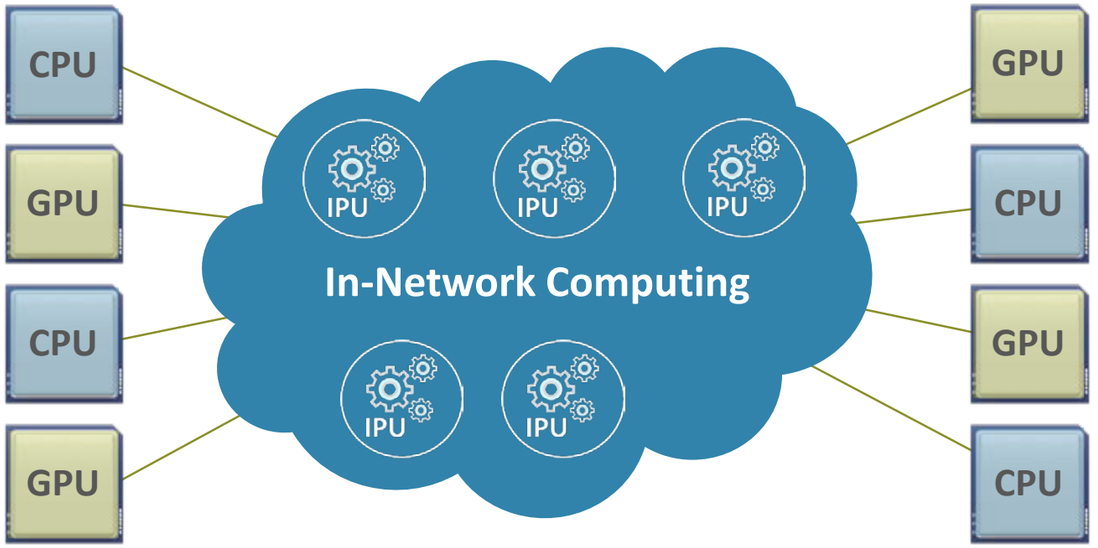
よりインテリジェントで高速なネットワーク

高パフォーマンスとスケールを実現するには、
より高速なネットワークとIn-Network Computingが必要

CPU中心 (Onload)



データセントリック (Offload)



データ待ちが生じる
パフォーマンスボトルネックに

データの転送中でもデータ処理を実施
高速でよりスケーラブルに

全レベルでHPC/AIフレームワークを加速

アプリケーション

- データ解析
- リアルタイム
- 深層学習

通信

- Mellanox SHARP In-Network Computing
- MPI Tag Matching
- MPI Rendezvous

MPI HPC-X™

UCF Unified Communication Framework

SHARP Scalable Hierarchical Aggregation and Reduction Protocol

ネットワーク

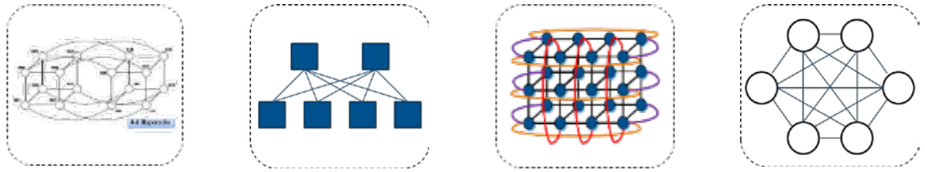
- Network トランスポートオフロード
- RDMA および GPU-Direct RDMA
- SHIELD (Self-Healing Network)
- アダプティブルーティング および輻輳コントロール

RDMA GPUDirect

SHIELD

接続性

- Multi-Host テクノロジ
- Socket-Direct テクノロジ
- 様々なトポロジ



最高の性能を提供する200Gb/sソリューション

Adapters

ConnectX-6

200Gb/s アダプタ
215 M メッセージ/秒
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

Switch

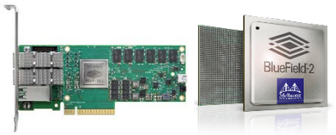
Mellanox Quantum

40 ポートのHDR (200Gb/s) InfiniBand
80 ポートのHDR100 InfiniBand
16Tb/sのスループット, 130nsの低遅延

SoC

BlueField-2

System on ChipおよびスマートNIC
プログラマブルアダプタ
スマートオフロード



Interconnect

LinkX

Transceivers
Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

Software

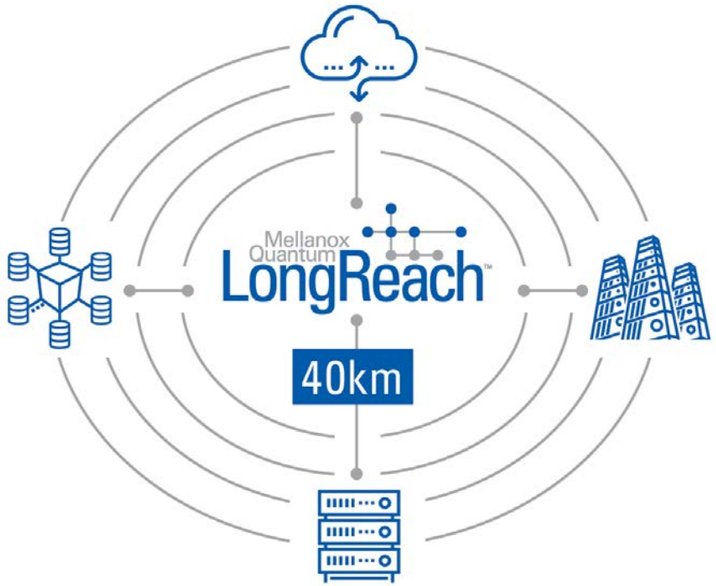
HPC-X

MPI, SHMEM/PGAS, UPC
商用およびオープンソースのアプリケーションに対応
ハードウェアアクセラレーションを活用

新製品 Mellanox Quantum LongReach™

InfiniBandで40kmの長距離接続を実現

- InfiniBandで40kmまでのデータセンター間接続
- データセンターをまたいでのロードバランスと拡張を実現
- データセンター障害時でも計算サービスの継続が可能に
- エンドツーエンドで標準のHDRおよびEDR接続
- 進化したIn-Network Computingをサポート



新製品 Mellanox Skyway™ Ethernet Gateway

NEW

- 100G EDR / 200G HDR InfiniBand から100G / 200G Ethernet へのゲートウェイ
- 将来を見据えた400G NDR / 800G XDR InfiniBand レディー
- EDR/HDR100/HDR InfiniBandが8ポート、及び100/200G Ethernetが8ポート
- 最大スループット1.6 Tbit/s
- 高可用性およびロードバランス機能
- 専用のMellanox Gateway operating system
- 高い拡張性、高効率

Mellanox Skyway™

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

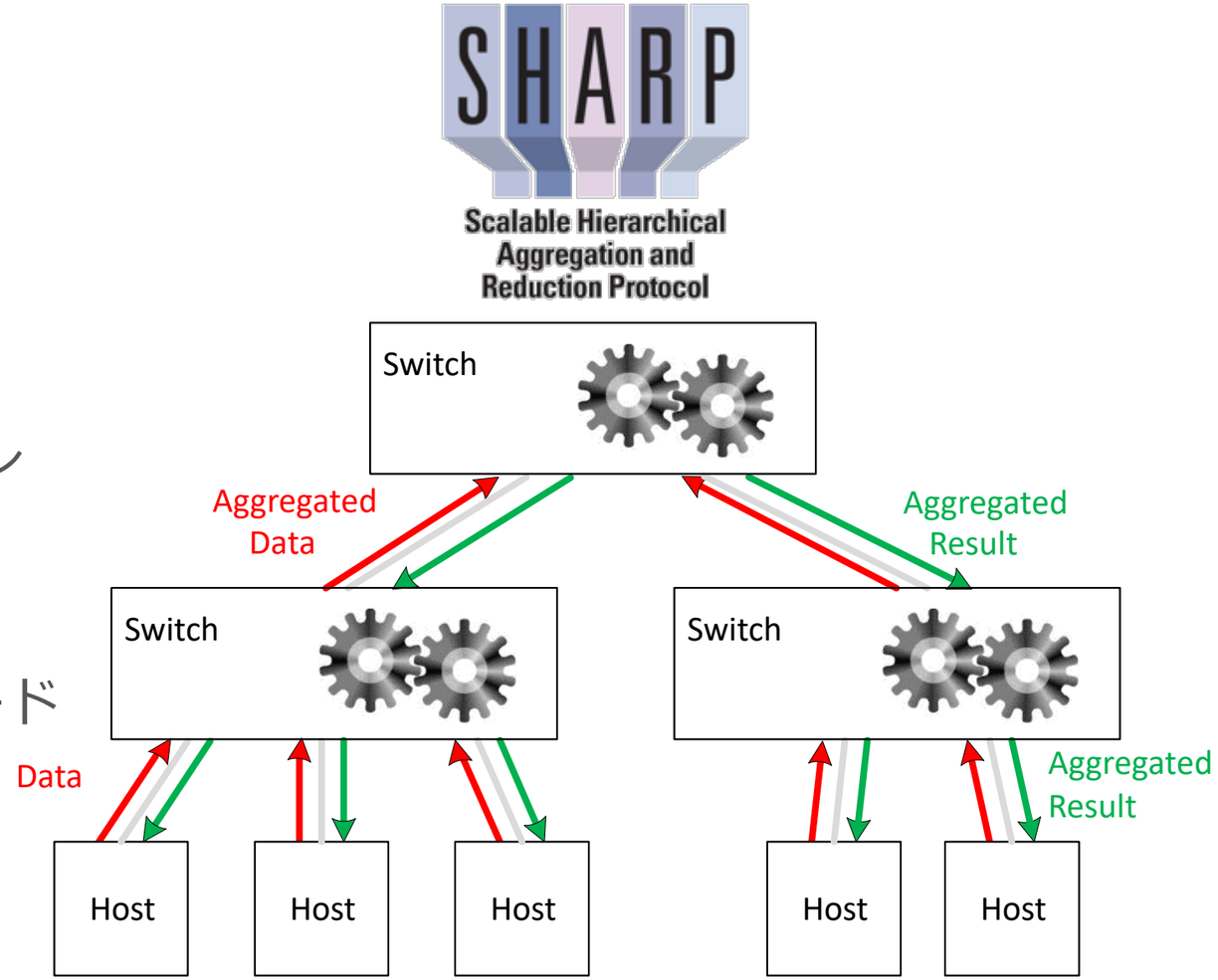
SHARP

Scalable Hierarchical Aggregation and Reduction Protocol

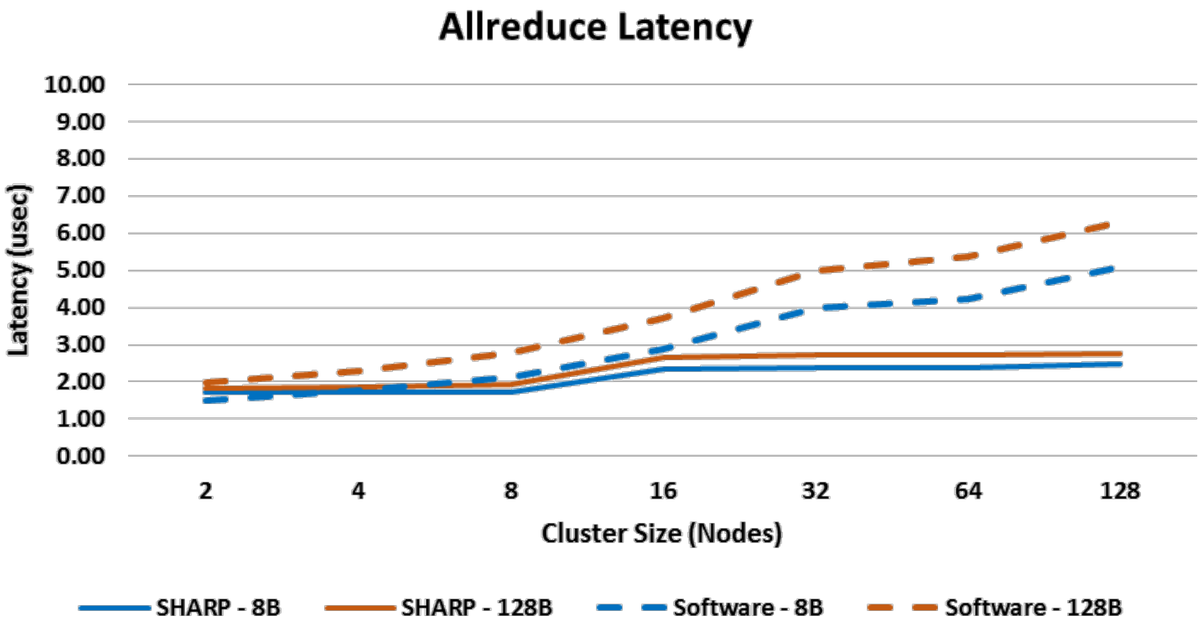
Mellanox SHARPとは

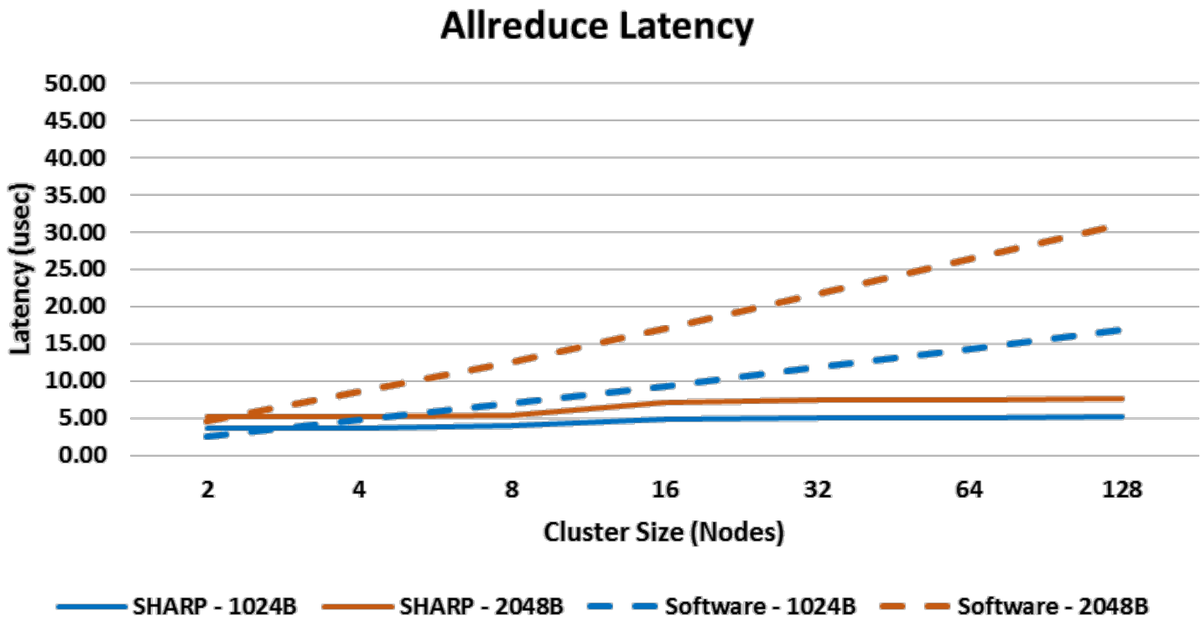
Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

- 高信頼でスケーラブルな汎用プリミティブ
  - In-Network ツリーベースの集合メカニズム
  - 多数のグループをサポート
  - 複数の処理を同時に実行可能
- 幅広いユースケースに適用可能
  - MPIやSHMEMを用いるHPCアプリケーション
  - 分散学習を行うアプリケーション
- 高性能で拡張性の高い、コレクティブオフロード
  - Barrier, Reduce, All-Reduce, Broadcast 他
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - 整数 および浮動小数点：16/32/64 ビット



SHARP AllReduce 性能改善 (128 Nodes)



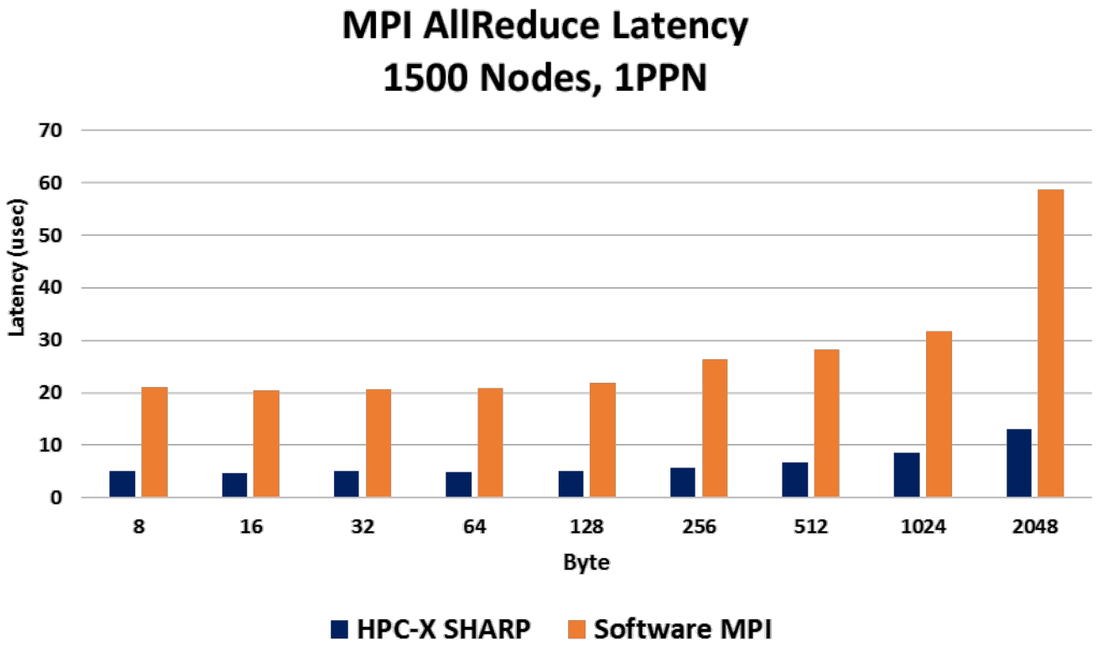


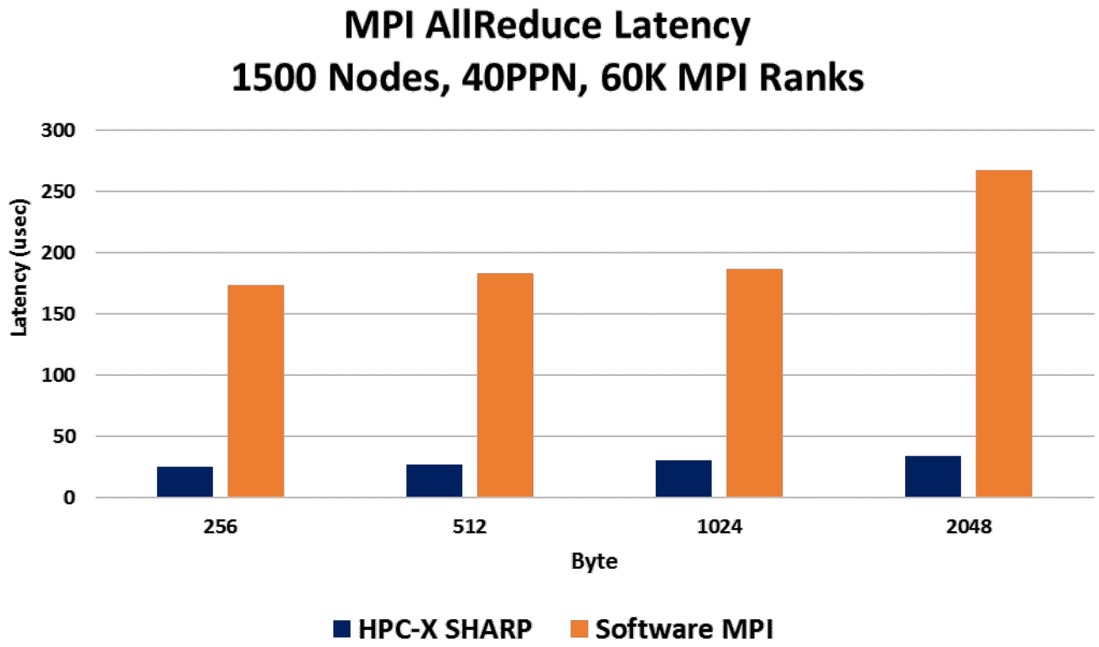
SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP により75%も遅延を削減
スケーラブルでフラットな遅延

SHARP AllReduce 性能改善

1500 Nodes, 60K MPI Ranks, Dragonfly+ トポロジ





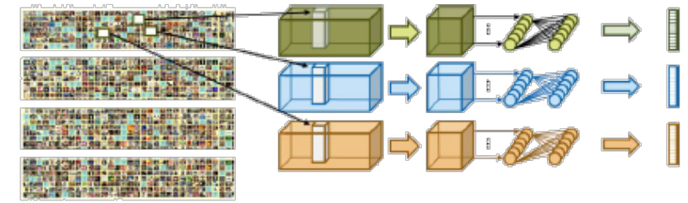
Scalable Hierarchical Aggregation and Reduction Protocol

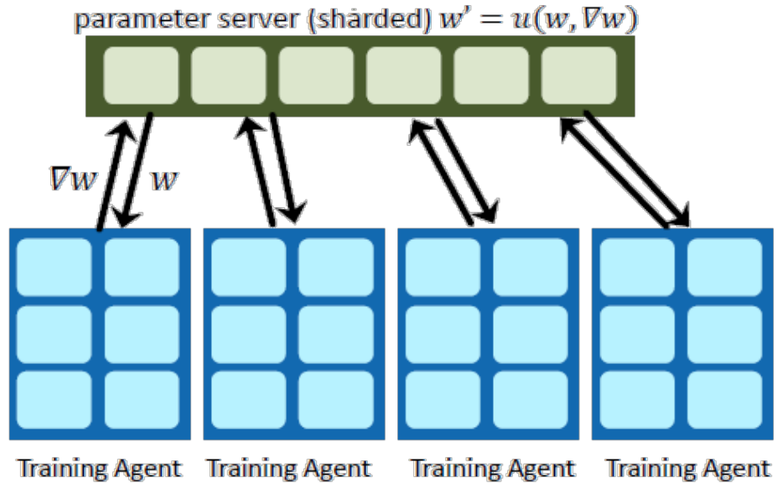
SHARP により最高のパフォーマンスが実現

SHARP はAIの性能を加速します

パラメータサーバのCPU が
ボトルネックに

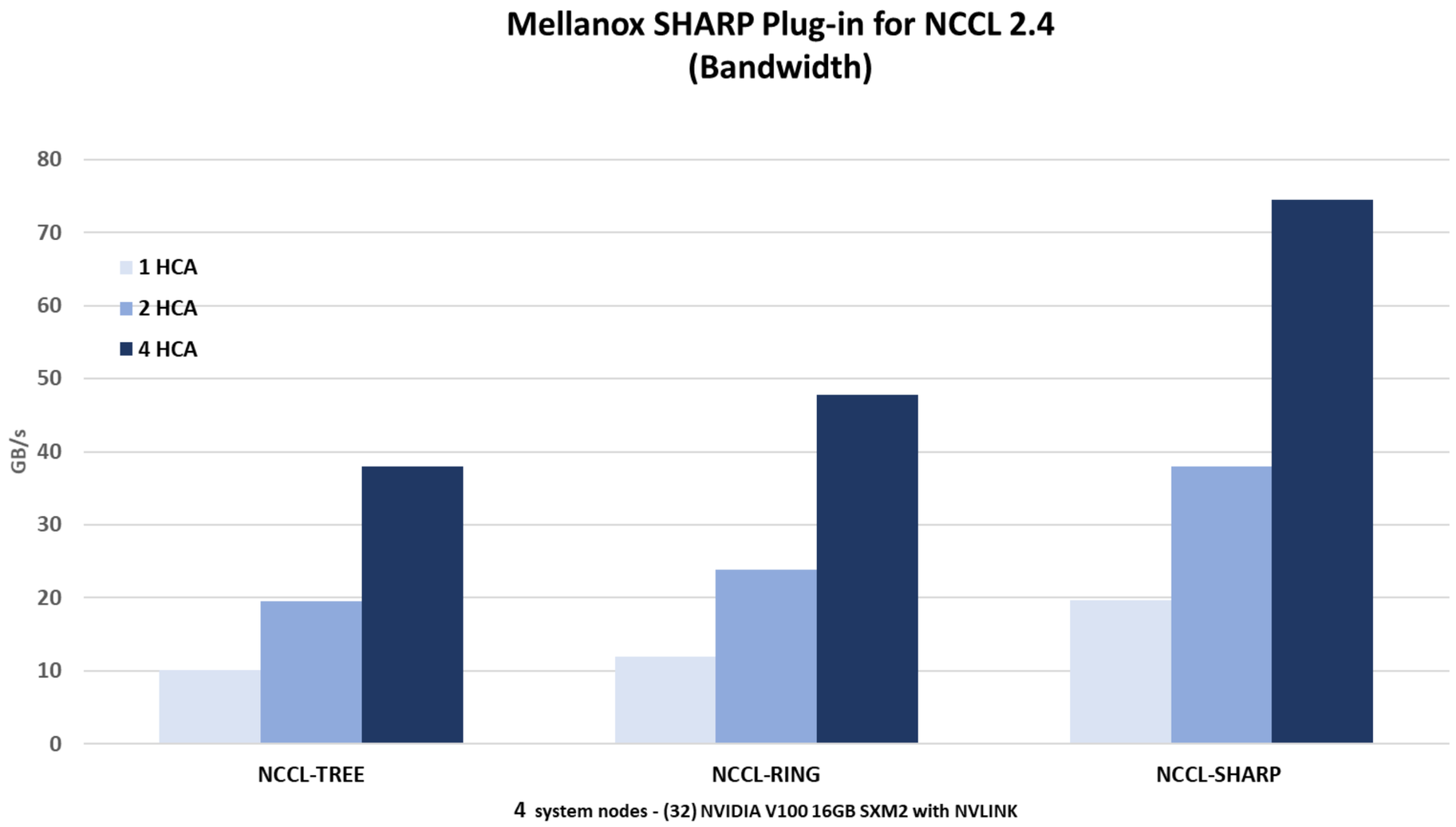
SHARP Scalable Hierarchical Aggregation and Reduction Protocol





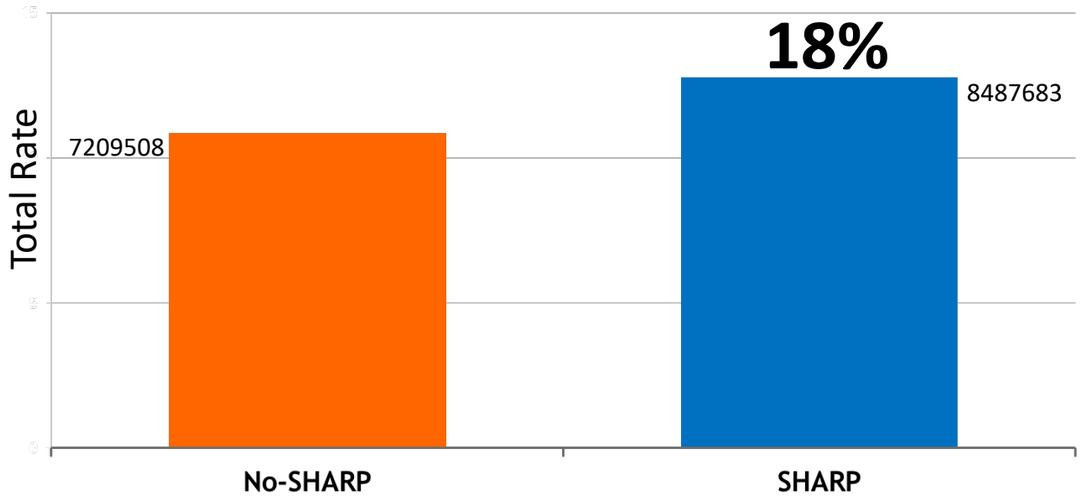
勾配平均の演算に効果
パラメータサーバの置き換えが可能
AIのパフォーマンスを改善

SHARP により最高のAI性能を実現



SHARP により最高のAI性能を実現

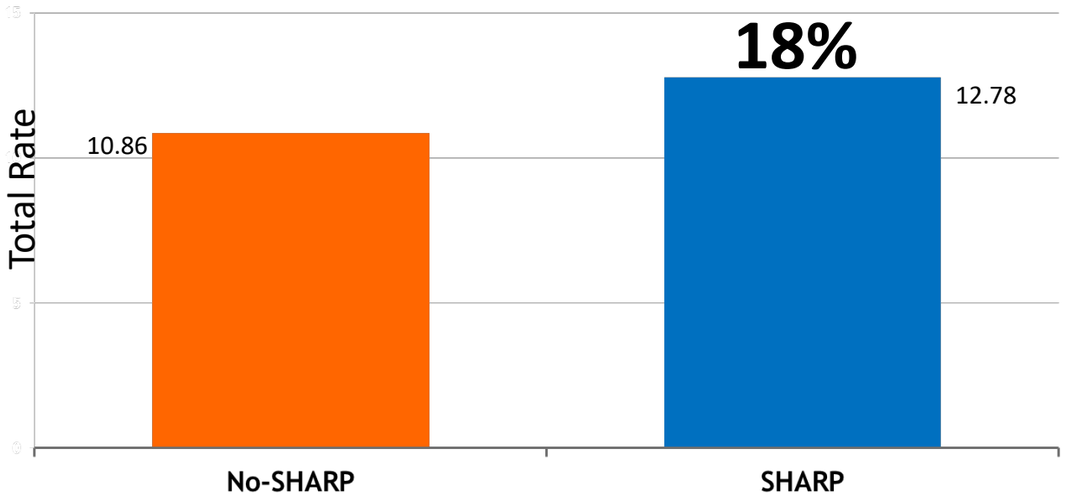
GNMT MLPerf Benchmark
Neural Machine Translation



24xDGX1V + 4xMellanox ConnectX-6

GNMT MLPerf 0.6 benchmark: Batch Size=32, Overlap=0.15

VAE Benchmark
Variable Auto-Encoder



32xDGX1V + 4xMellanox ConnectX-6

VAE benchmark: Model=3, BS=512

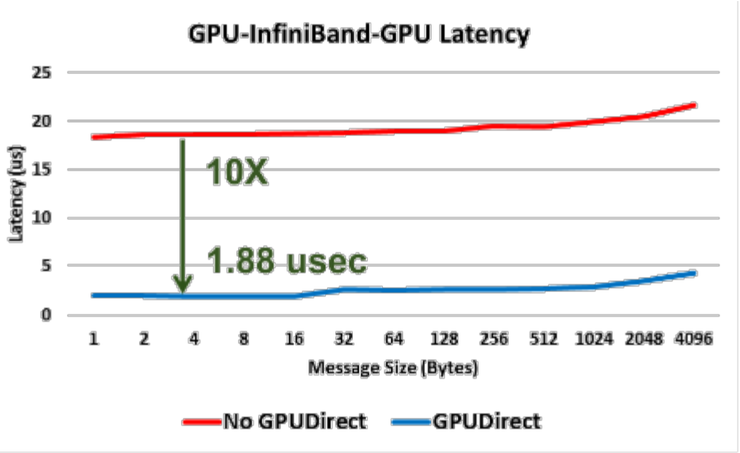
SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

SHARPにてAIのパフォーマンスが改善

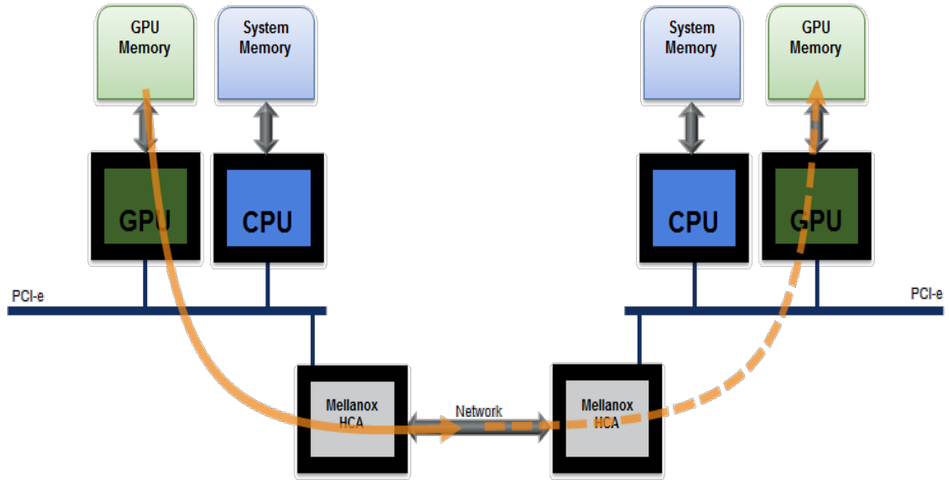
GPUDirect

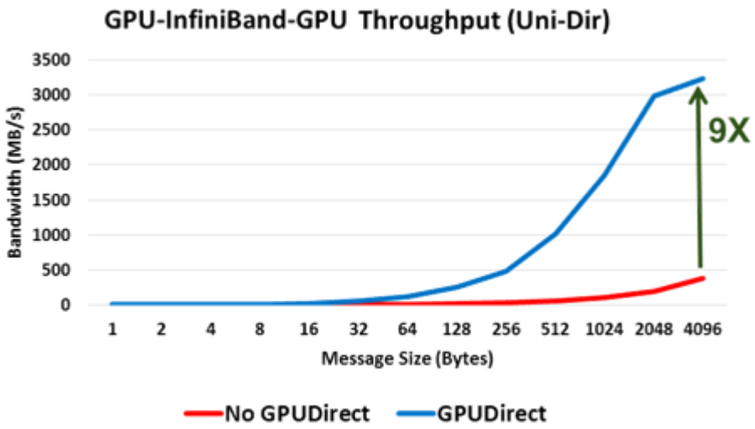
GPUDirect™ RDMAで10倍の性能改善

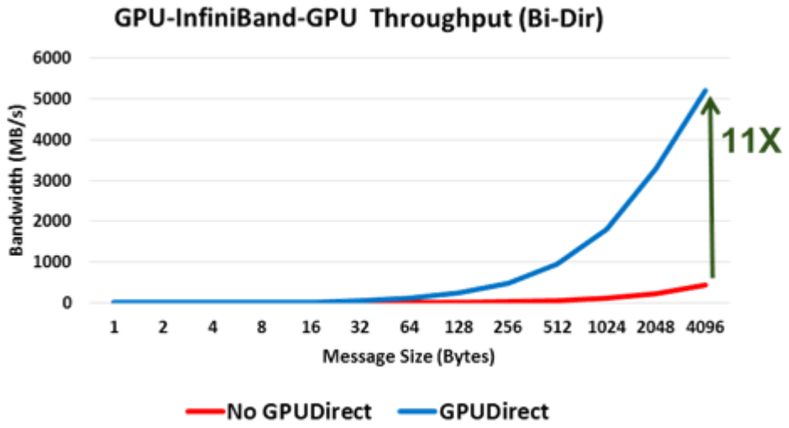
- HPCやDeep Learningの性能を改善
- GPU間のコミュニケーション遅延を最低に



GPUDirect™ RDMA







Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

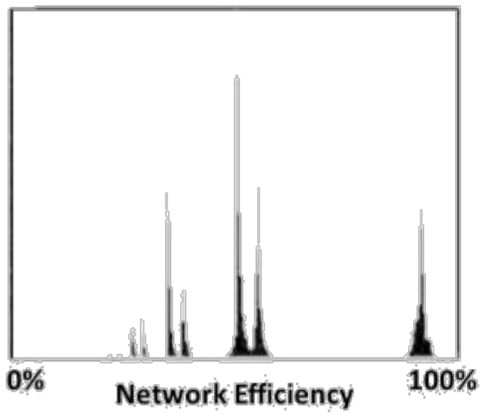
Adaptive Routing

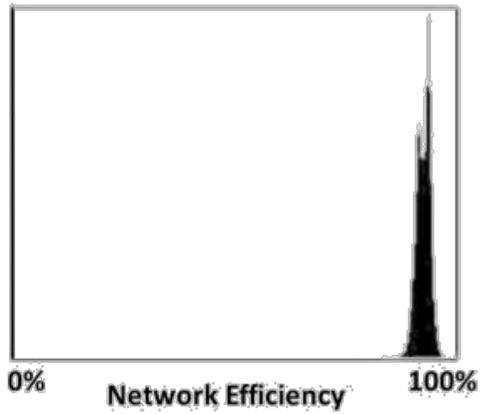
確かな実績のInfiniBand Adaptive Routing

- Oak Ridge National Laboratory – Coral Summit supercomputer
- mpiGraphをベースにした双方向帯域ベンチマーク
  - 可能性のあるすべてのMPIプロセスペア間の帯域を探索
- アダプティブルーティングを使用すると、**理論最大帯域の96%もの平均実効帯域**を観測

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.

InfiniBand High Network Efficiency - mpiGraph





Static Routing

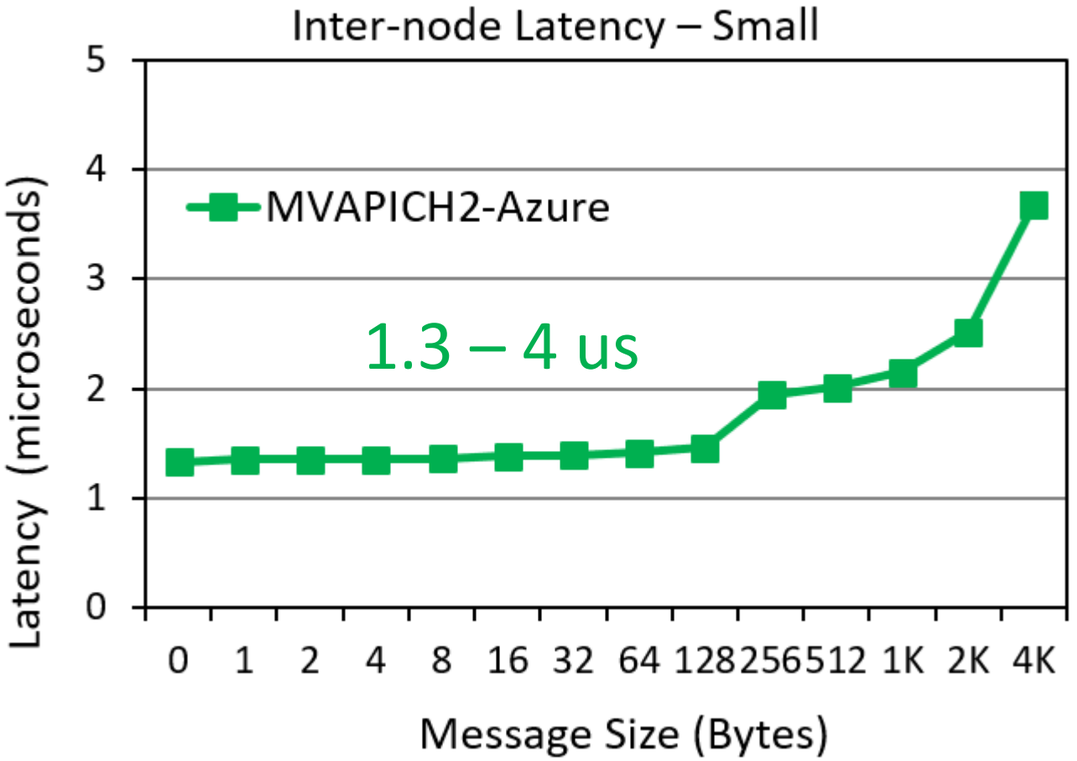
Adaptive Routing

Oak Ridge National Lab Summit Supercomputer

"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)

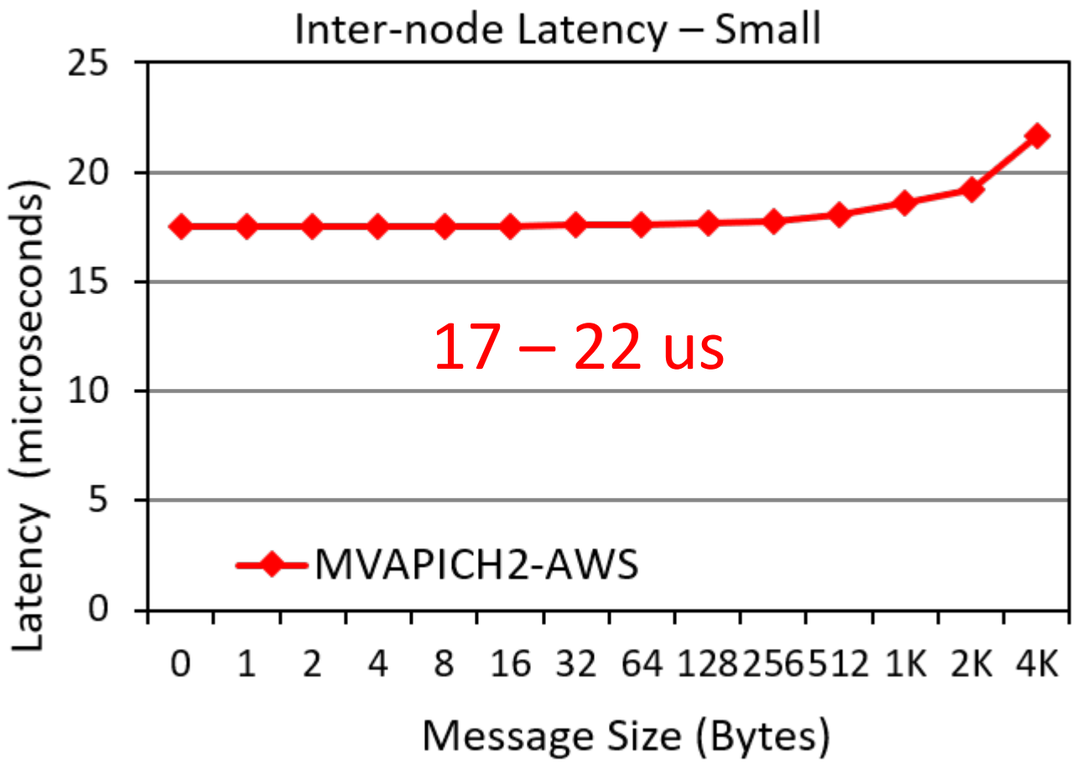
InfiniBandによる HPCクラウド

Azure (100G IB) vs Major Cloud(100G) - MPI 性能



Microsoft Azure

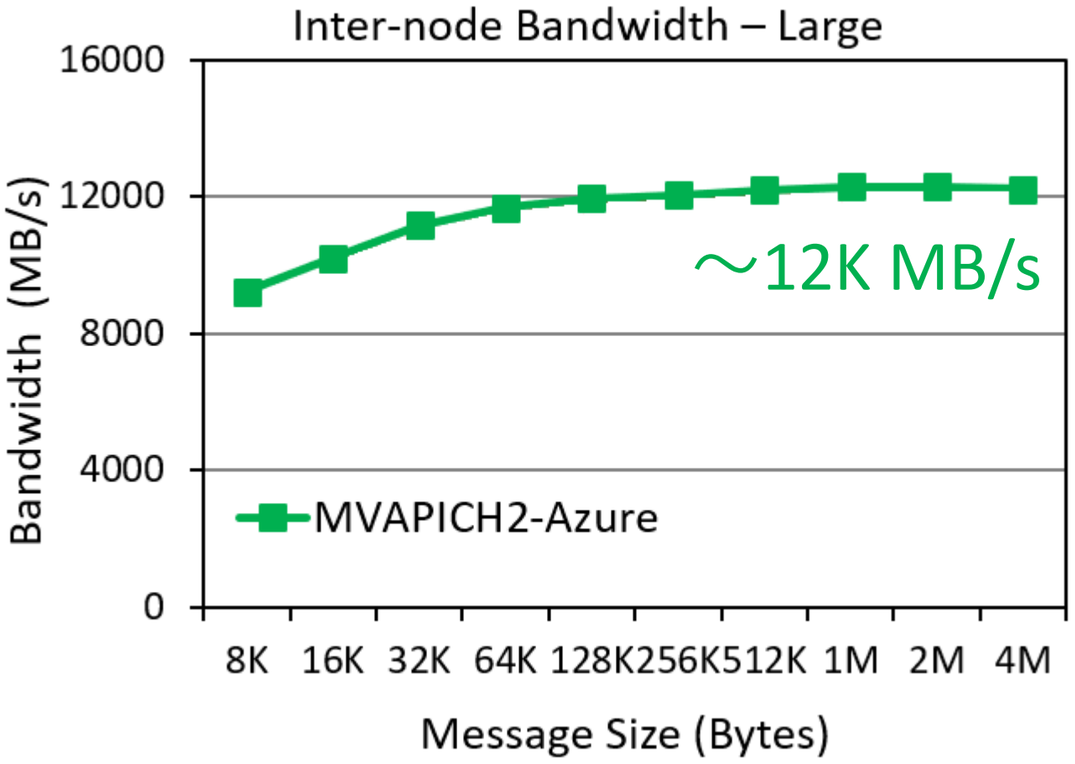
INFINIBAND™



Major Brand A-Cloud

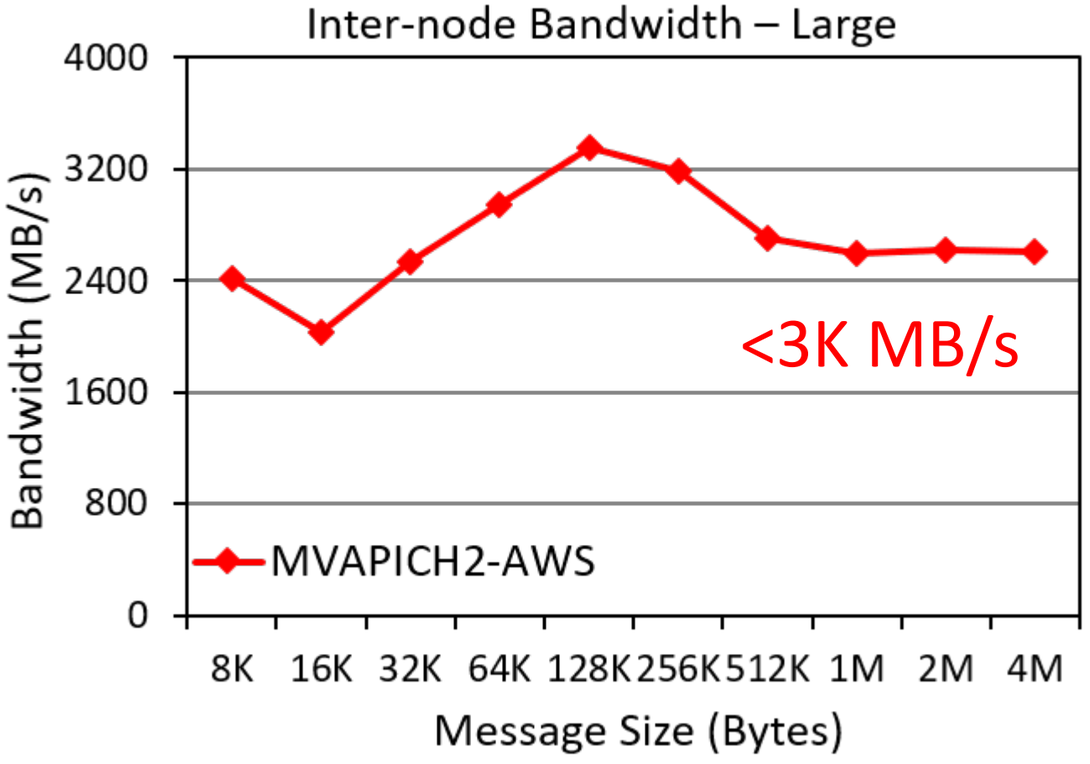
InfiniBand は スモールメッセージで**13.5X** もの低遅延を実現

Azure (100G IB) vs Major Cloud(100G) - MPI 性能



Microsoft Azure

INFINIBAND™



Major Brand A-Cloud

InfiniBandはラージメッセージで**3.6X**もの帯域を提供

最高の性能と拡張性を提供する InfiniBand

Mellanox Supercomputing 2019 News

ConnectX-6

SHIELD SHARP Scalable Hierarchical Aggregation and Reduction Protocol

Mellanox Quantum

Mellanox Quantum LongReach

Mellanox Skyway

HPC-X™ Message Passing Interface (MPI)

HPC-X™

200G HDR InfiniBand は最高の性能と拡張性を提供します

InfiniBand H·D·R 200

eni · CSC · Pittsburgh Supercomputing Center · NASA · Meteo France · Microsoft · Jülich Forschungszentrum · KMA Korea Meteorological Administration · EuroHPC Joint Undertaking · TACC Texas Advanced Computing Center · GENCI · TGCC · INL Idaho National Laboratory · SDSC San Diego Supercomputer Center · Australian National University · Harvard University

HDR InfiniBand は新規のシステムの31%で採用

InfiniBand の採用が12%増加

297 システムで

メラノックスの採用

TOP 500 The List.

PRIMEHPC FX700, A64FX Arm CPU
With InfiniBand (Post-K)

FUJITSU

2019 R&D 100 WINNER

Los Alamos NATIONAL LABORATORY EST. 1943 · Argonne NATIONAL LABORATORY · Mellanox TECHNOLOGIES · OAK RIDGE National Laboratory · AMD · NVIDIA · arm

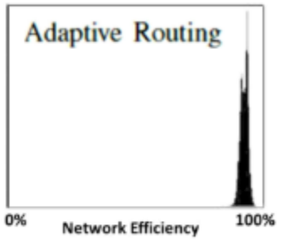
UCF Unified Communication Framework

エクサスケールに向けた最高の性能と拡張性

OAK RIDGE National Laboratory

SHARP SHIELD

96%
の帯域効率

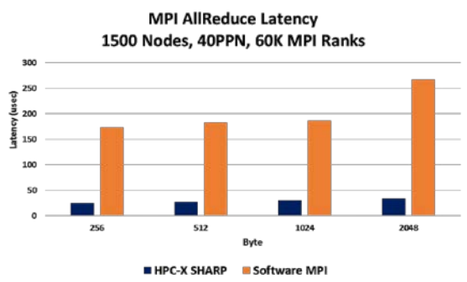


Summit's MpiGraph

UNIVERSITY OF TORONTO
SciNet

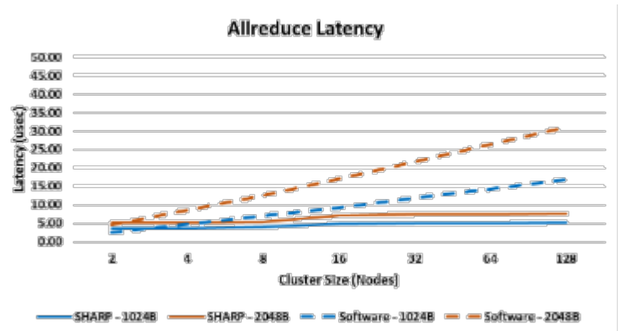
7倍
高い性能

SHARP



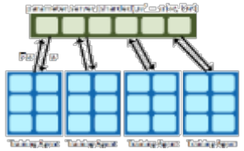
SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

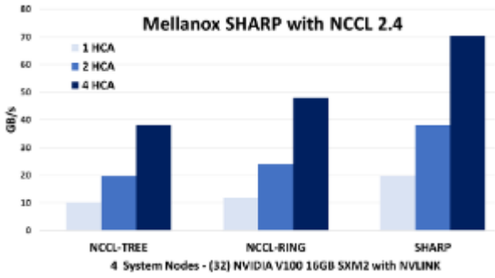
フラットで低遅延



Deep Learning

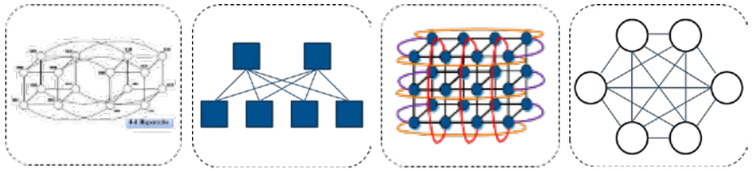
2倍
の性能

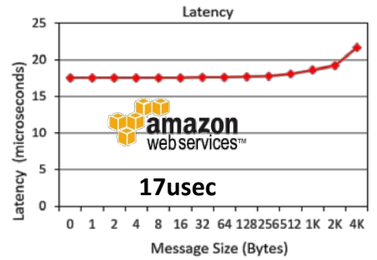




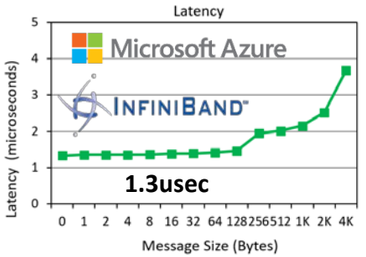
SHIELD
SELF-HEALING

5000倍
早い回復力



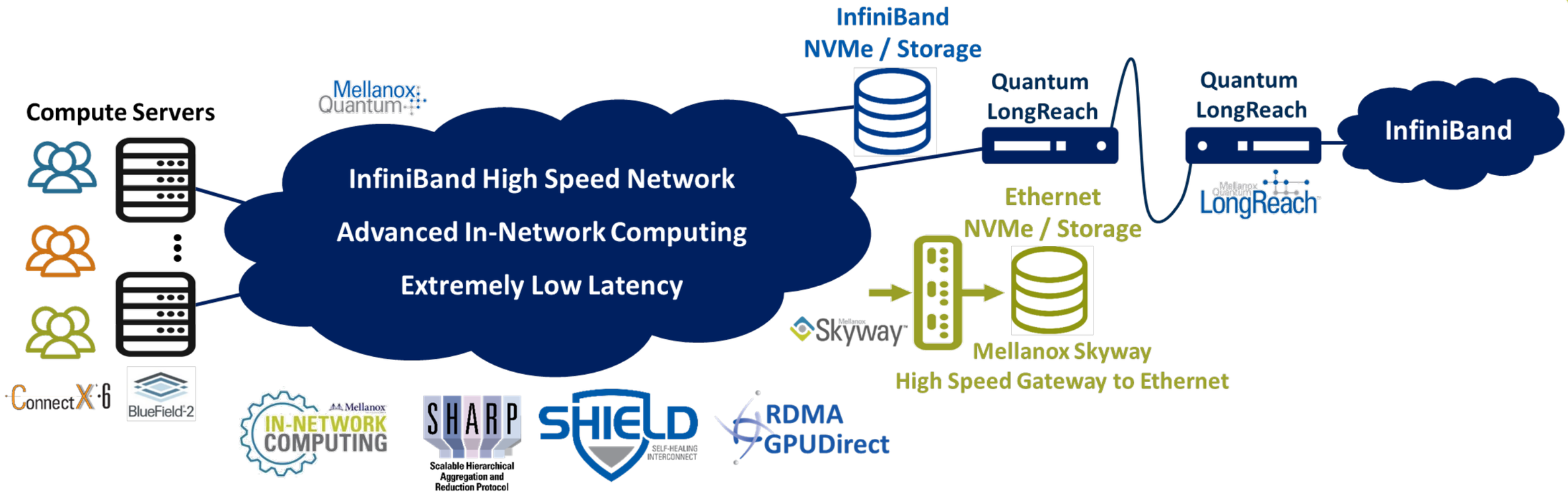


13倍
クラウド性能



Microsoft Azure

InfiniBandは最高の性能とROIを実現します



- エンドツーエンドの200G HDR InfiniBand, 超低遅延, 高メッセージレート, RDMA and GPUDirect
- 進化したエンドツーエンドのアダプティブルーティング, 輻輳制御とQoS
- In-Network Computing アクセラレーションエンジン (Mellanox SHARP, MPI オフロード)
- 高可用性を実現する、ネットワークの自己修復機能を備えたSHIELDテクノロジ
- Mellanox Skyway 高速イーサネットゲートウェイ, Mellanox Quantum LongReachで40 Kmまでの長距離InfiniBand接続
- InfiniBandは標準技術であり、後方および将来の互換性を提供

これからも最高の性能と拡張性を提供します

SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

SHIELD
SELF-HEALING INTERCONNECT

800G XDR

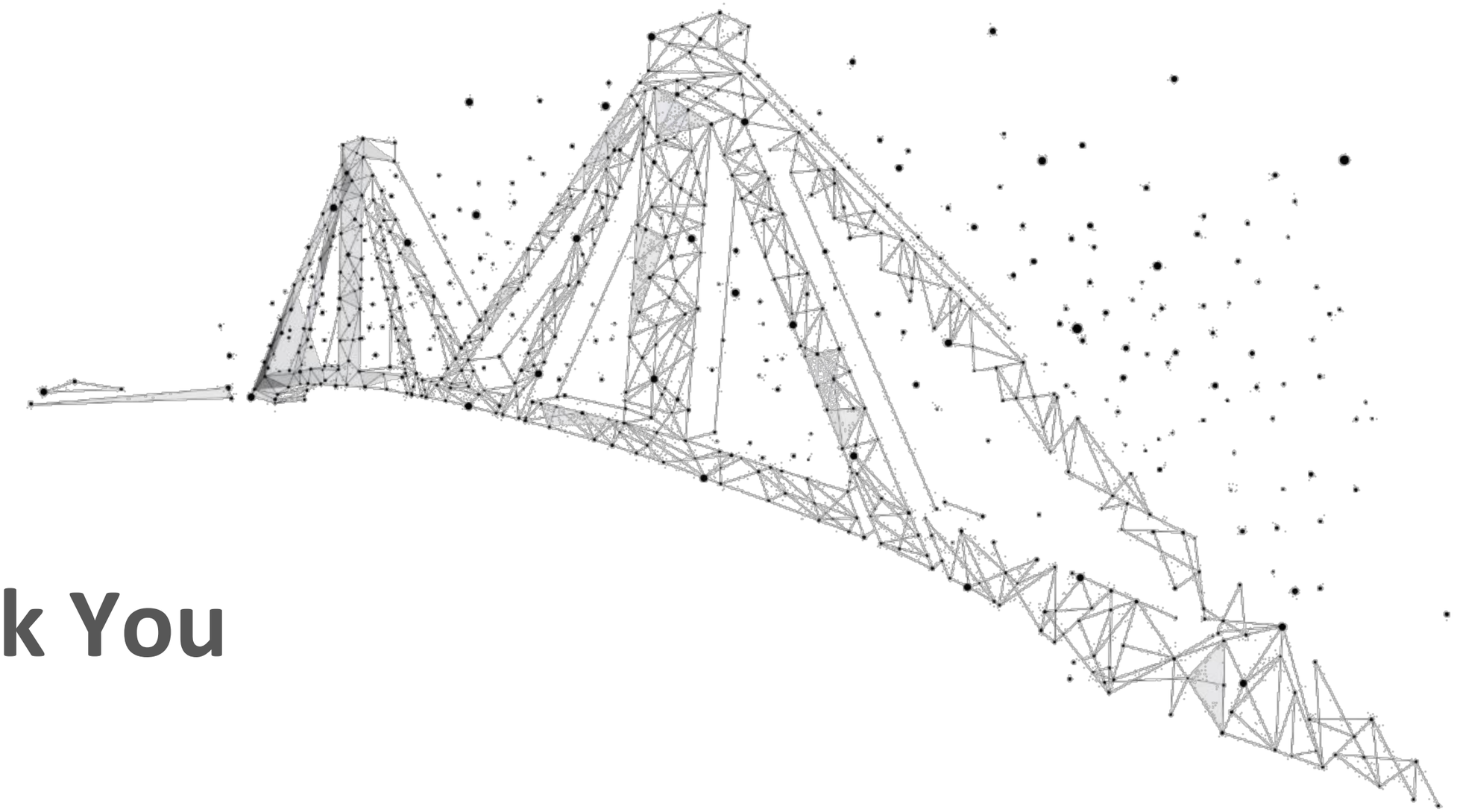
Mellanox IN-NETWORK COMPUTING

400G NDR

200G HDR

RDMA
GPUDirect

HPC-X™



Thank You