



200Gbps HDR InfiniBandが登場！

～インターコネクトで加速！最新製品技術のご紹介～

Dec 13, 2018

Mellanox と TOP500

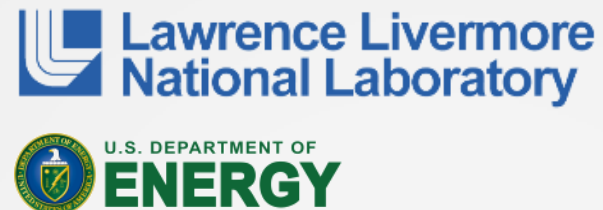
Mellanox は最先端のHPC や AI をサポート

World's Top 3 Supercomputers



1

Summit CORAL System
World's Fastest HPC / AI System
9.2K InfiniBand Nodes



2

Sierra CORAL System
#2 USA Supercomputer
8.6K InfiniBand Nodes



3

Wuxi Supercomputing Center
Fastest Supercomputer in China
41K InfiniBand Nodes



Mellanox は最先端のHPC や AI をサポート



Fastest HPC / AI System in Japan
1.1K InfiniBand Nodes



The world's Fastest Industry
Supercomputer
1.6K InfiniBand Nodes

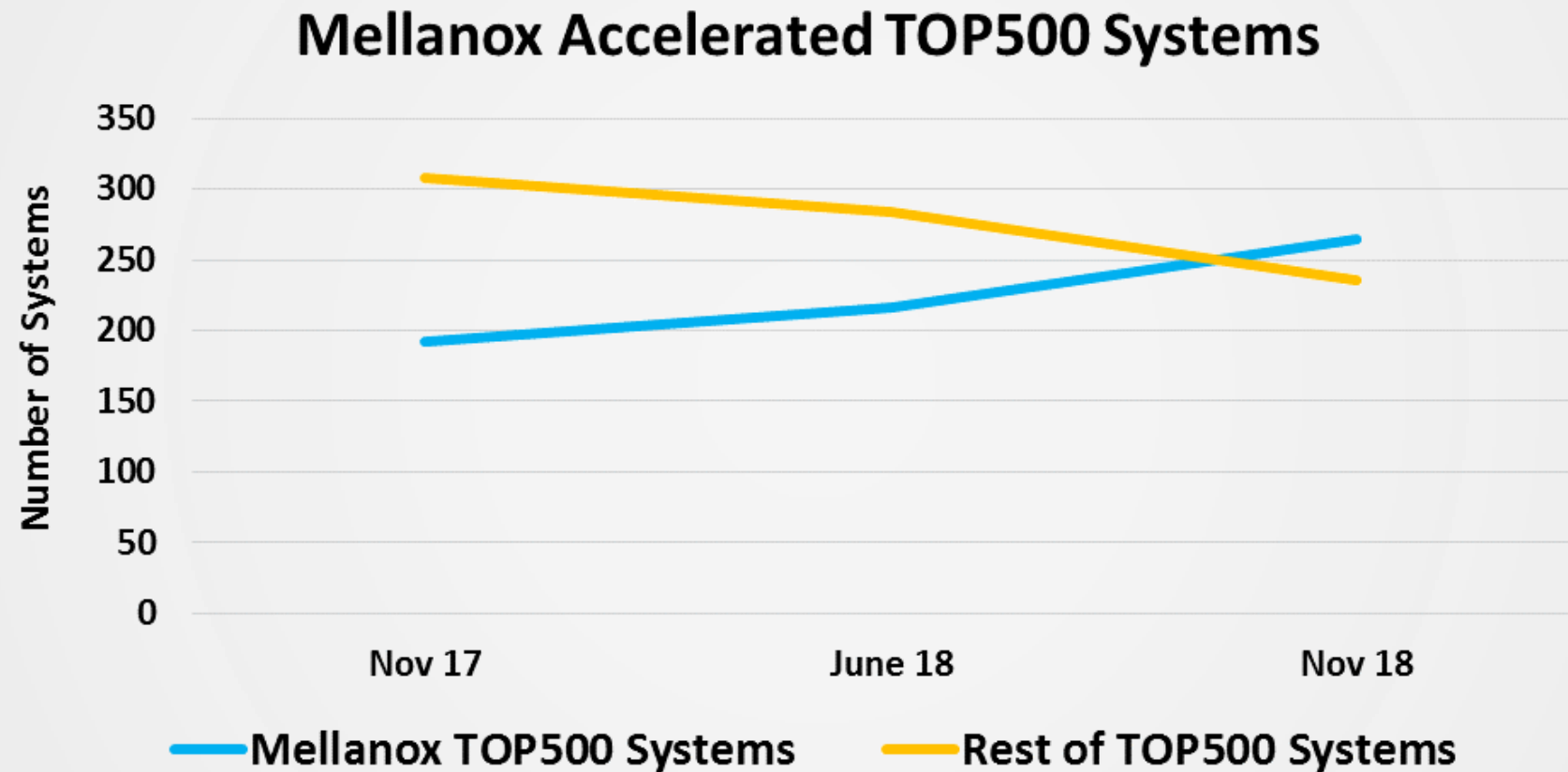


JUWELS Supercomputer
2.6K InfiniBand Nodes



より多くのシステムで選ばれるMellanox製品

この一年で、TOP500ではMellanoxをご採用頂いたシステム (InfiniBandおよびEthernet)が、38%も増加しました。



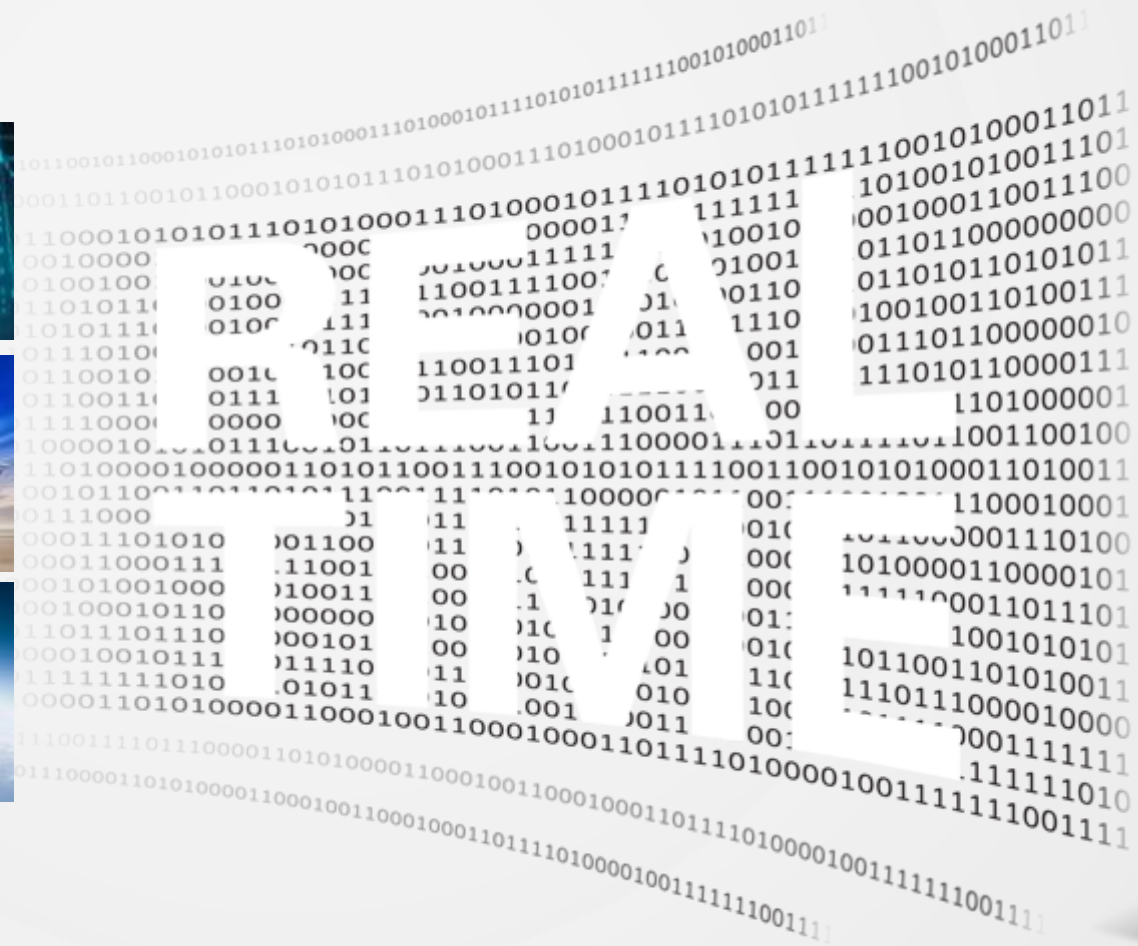
Mellanox InfiniBand and Ethernet Solutions Accelerate Majority of TOP500 Platforms
on the November TOP500 Supercomputers List

増大するデータ処理と インターコネクト

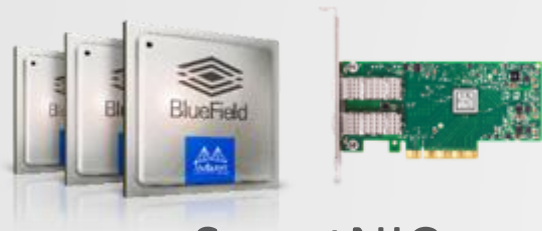
Mellanoxが実現するスマートなインターコネクト

データの力を解き放つ

加速度的に増加するデータ処理、そしてリアルタイム分析には、
今まで以上にインターコネクトが重要になります。



あらゆる場面で加速度的に増加するデータ処理には、 スマートなインターコネクトが必要



SmartNIC



System on a Chip

Higher

Data Speeds

Faster

Data Processing

Better

Data Security



Innova ConnectX

Adapters



Spectrum Quantum

Switches



LinkX

Cables &
Transceivers



より高速で高機能なインターコネクト

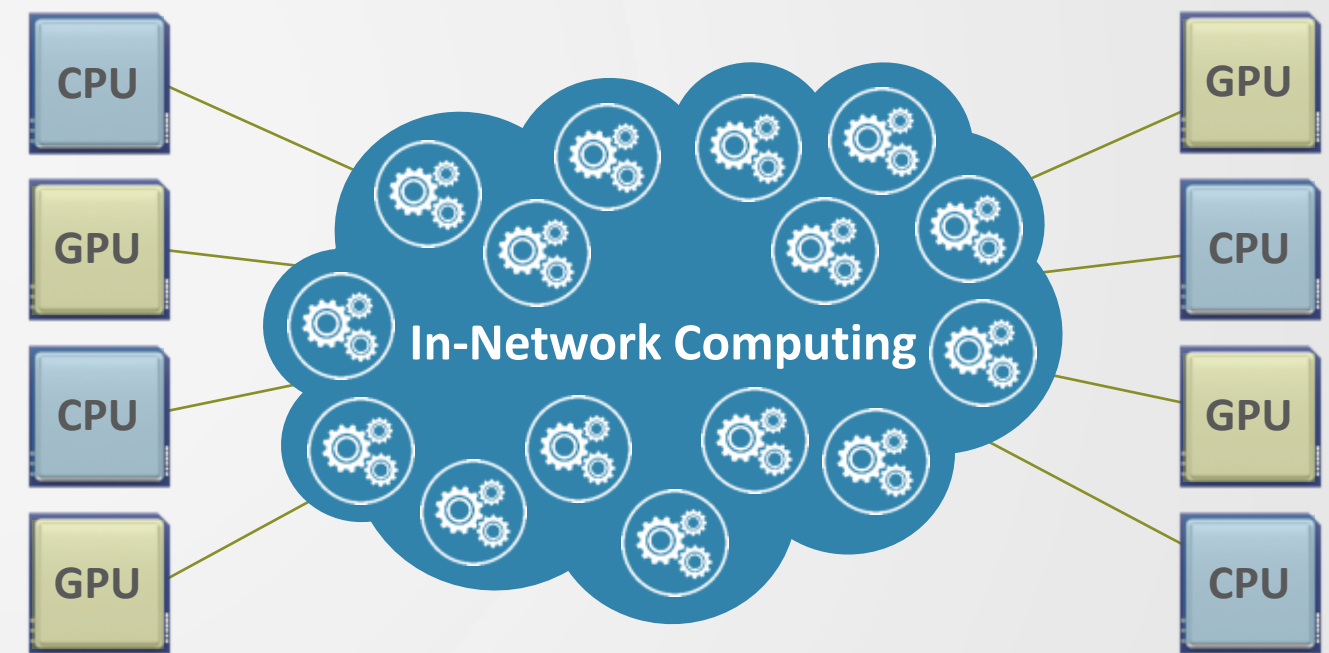
より高速なだけでなく、In-Network Computingによって、はじめてパフォーマンスとスケールが両立できます

CPU-Centric (Onload)

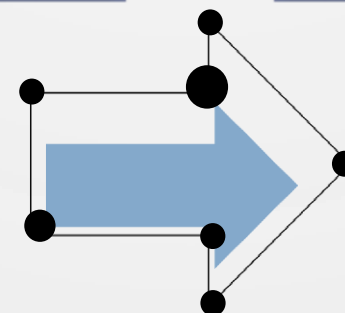


データ処理待ちが発生
パフォーマンスボトルネックに

Data-Centric (Offload)

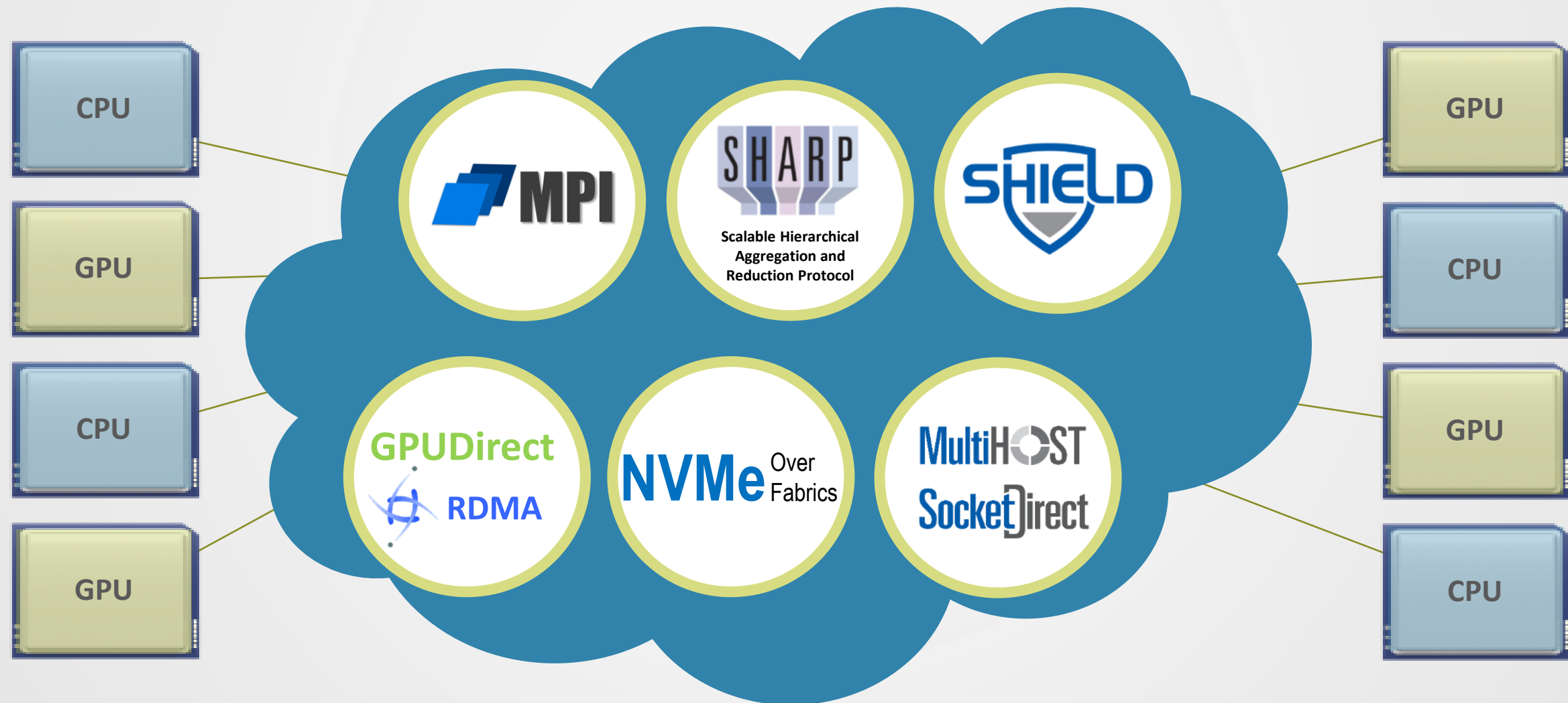


データ移動中でも解析を実行
高パフォーマンスと高スケールの実現



In-Network Computingが実現する データを中心とするデータセンター

高速なデータ通信とIn-Network Computingが実現する
高速で高スケールなデータセンター



Mellanox In-Network Computing and Acceleration Engines

HDR製品のご紹介

Highest-Performance 200Gb/s Interconnect Solutions

Adapters

ConnectX[®] 6

200Gb/s Adapter, 0.6us latency
215 million messages per second
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)



Switch

Quantum
The smartest switch, became smarter

40 HDR (200Gb/s) InfiniBand Ports
80 HDR100 InfiniBand Ports
Throughput of 16Tb/s, <90ns Latency



Switch

Spectrum[™] 2

16 400GbE, 32 200GbE, 128 25/50GbE Ports
(10 / 25 / 40 / 50 / 100 / 200 GbE)
Throughput of 6.4Tb/s



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 – エクサスケールへの布石

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
- Media – Video packets hardware reconstruction

ConnectX[®]·6



HDR InfiniBand スイッチ: QM8700, 1U Series



40 QSFP56 ports (50G PAM4 per lane)

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

Superior performance

- 90ns latency
- 390M packets per sec (64B)
- 16Tb/s aggregate bandwidth

Superior resiliency

- 22" depth
- 6 fans (5+1), hot swappable
- 2 power supplies (1+1), hot swappable



HDR InfiniBand スイッチ: CS8500, モジュラ型

800 QSFP56 ports

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

Superior performance

- 300ns latency
- 320Tb/s aggregate bandwidth
- Max power : 800p up to 22KW
- LCD Tablet IO panel

Water-cooled solution

- Liquid – Liquid 4U CDU
- Liquid – Air 42U (350mm wide) stand alone HEX
- 0C – 35C (air) or 40C (water) operating air range



業界最先端のMellanox HDR InfiniBand



HDR InfiniBand は最もスケーラブル

1.7X Better

80-Ports Top of Rack Switch

2.8X Better

3200 ノードが2Tierのネットワークで実現

4.6X Better

128K ノードが3Tireのネットワークで実現

400-Nodes



およそ4割のスイッチと5割のケーブルを削減

InfiniBand: 15 ToR Switches; Other solution: 24 ToR Switches

1600-Nodes



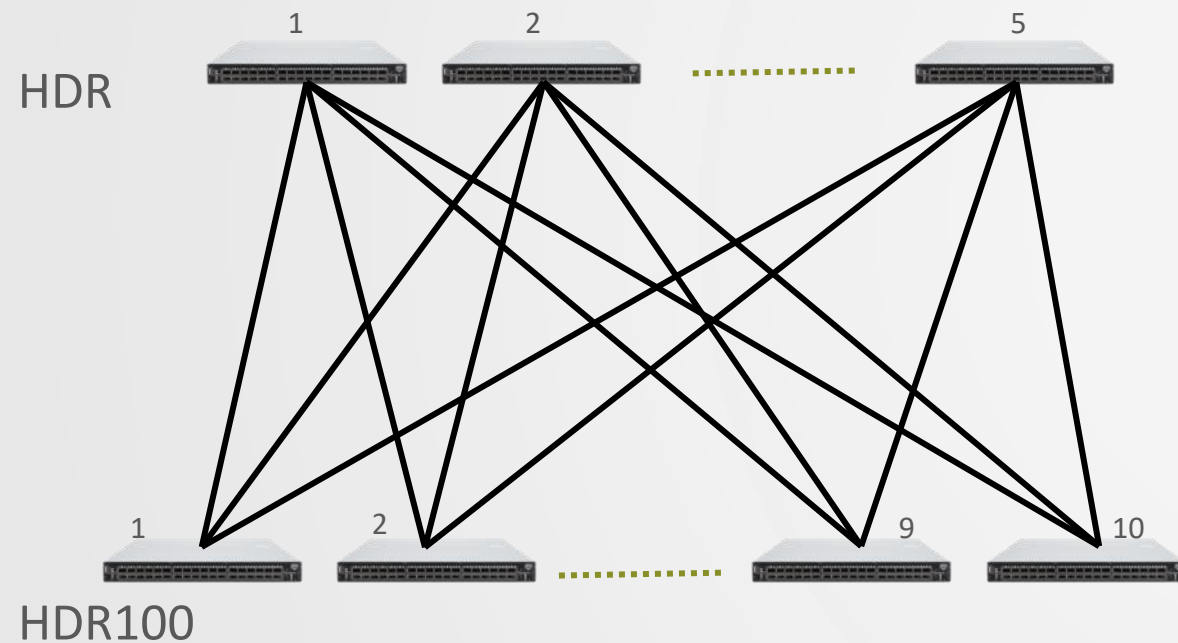
ケーブルは1/4に、電力も半減

InfiniBand: 1 Modular Switch; Other solution : 64 ToR + 2 Modular

優れたTCOで最高のスケールを実現

スイッチ台数を4割近く削減し、ケーブル本数も半減

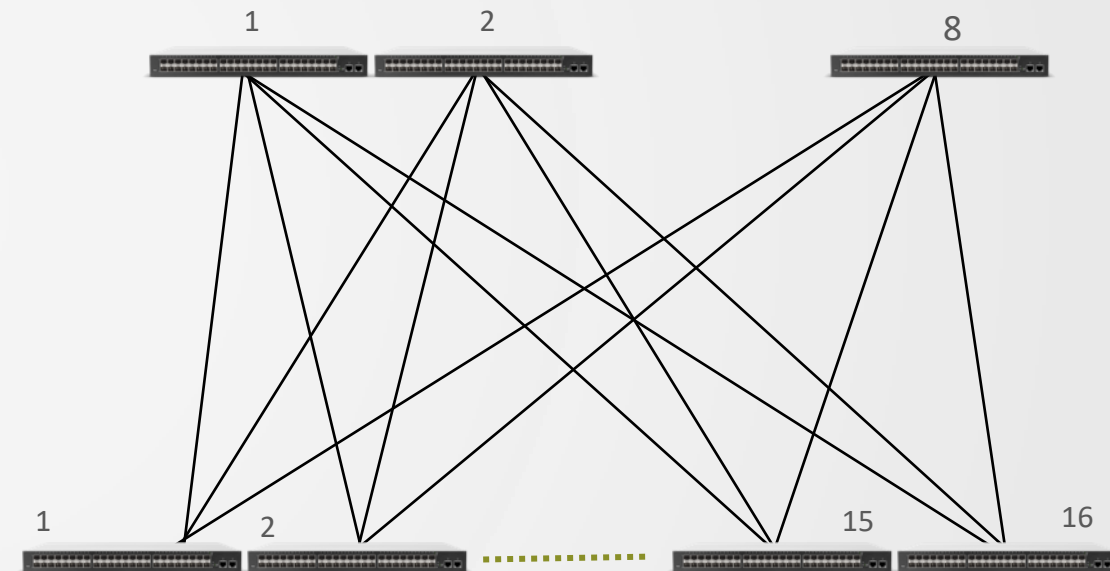
400-Nodes 100G with HDR100 InfiniBand



15 Switches, 400 Cables

Mellanox Quantum | ConnectX-6

384-Nodes 100G with other solution.



24 Switches, 768 Cables

優れたTCOで最高のスケールを実現

専有面積は1/3, ケーブルも1/4, 電力も半減

1536-Nodes 100G with HDR100 InfiniBand

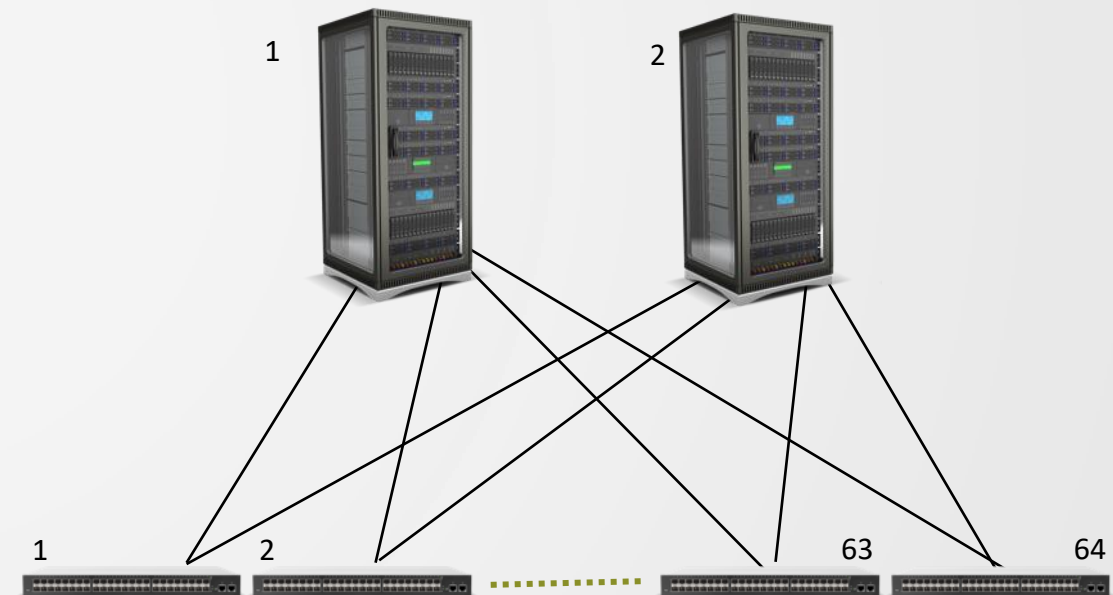


Mellanox
Quantum

ConnectX[®] 6

768 Cables, 22KW Power

1536-Nodes 100G with other solution



3072 Cables, 49KW Power

リアルタイムなネットワークの可視化

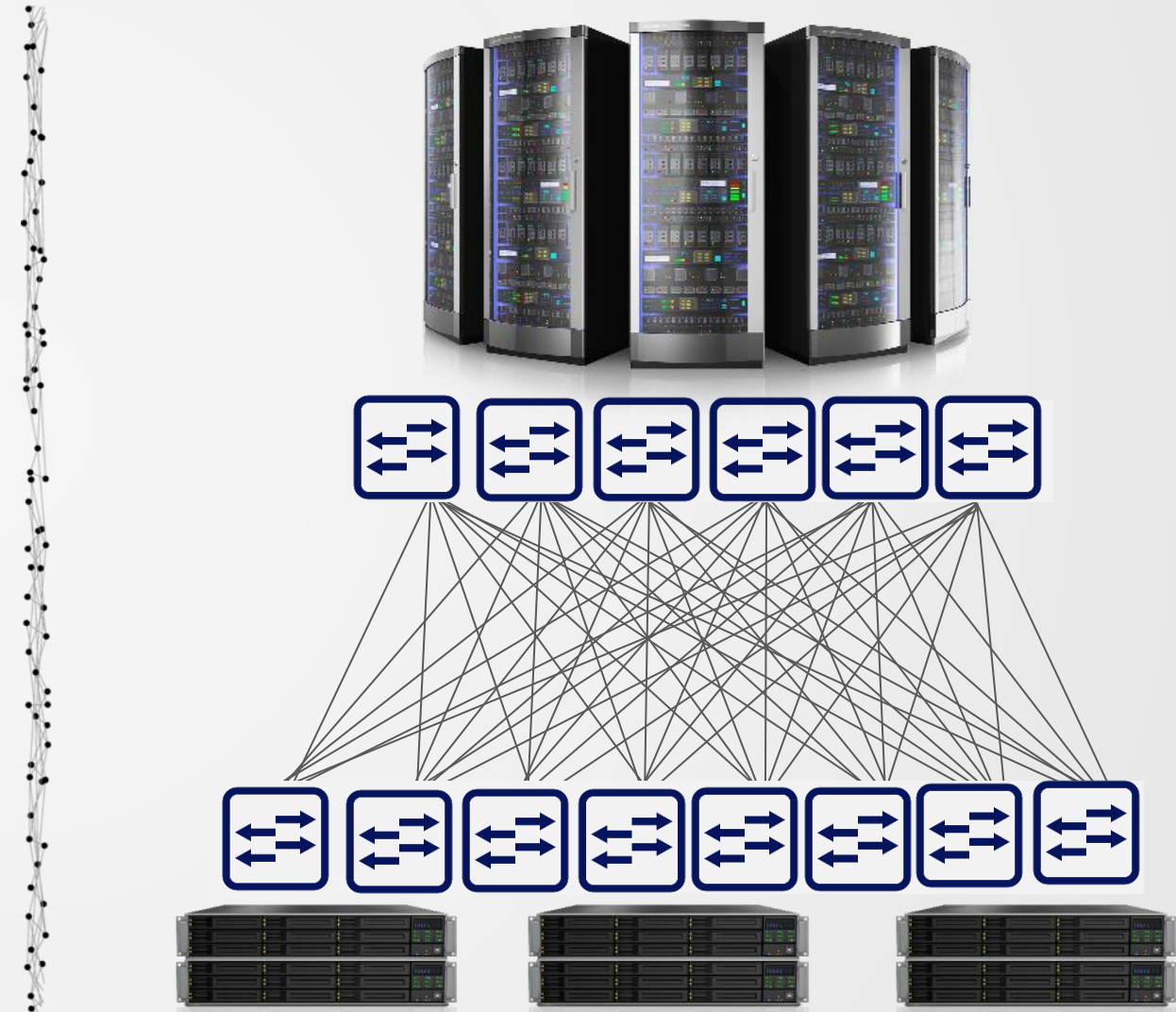
テレメトリやデータ収集のための、ハードウェア内蔵センサ

トラブル解析のための高度なモニター

- 8つのミラーエージェント（輻輳、バッファ使用率、遅延でトリガ可能）
- ヒストグラムを使用してキュー深度を計測（64ns単位）

リアルタイムネットワーク解析

- バッファスナップショット
- 輻輳通知とバッファステータス



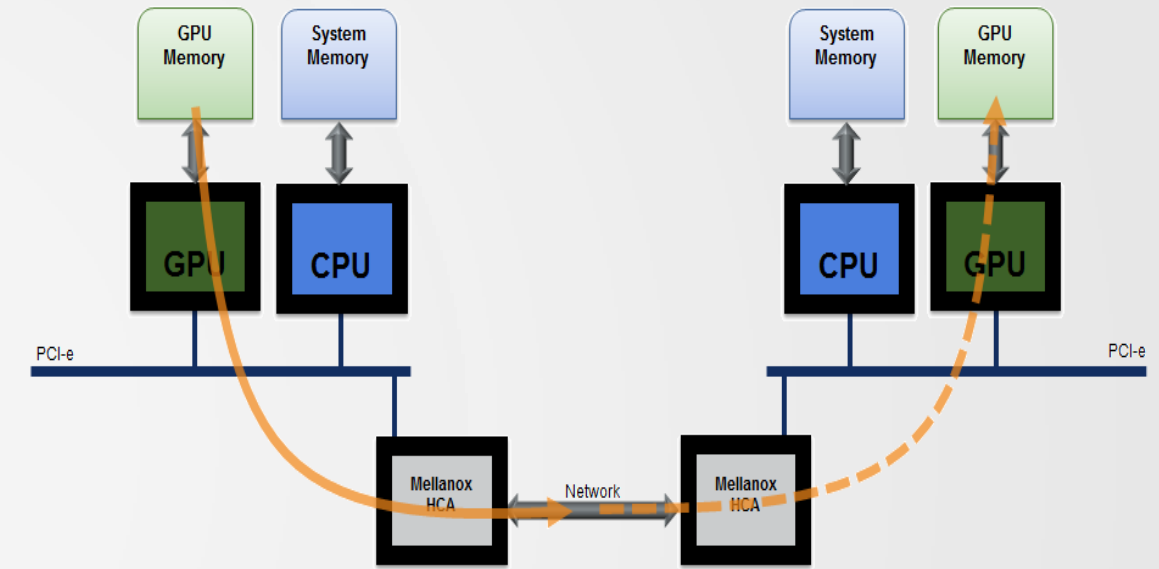
Mellanox Advantages



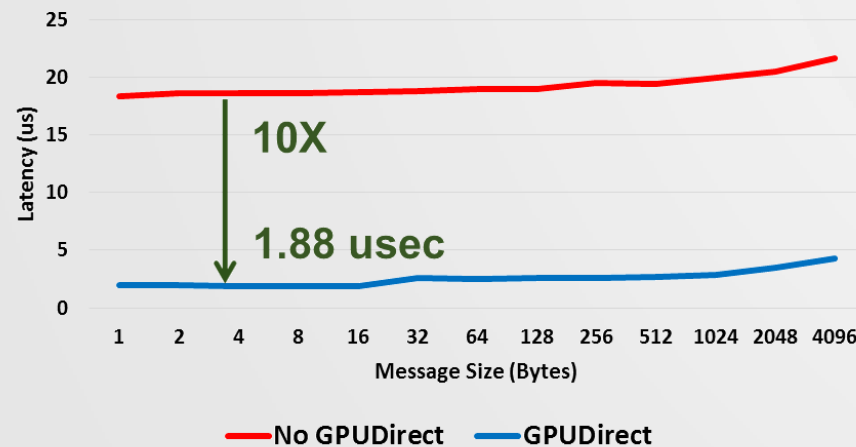
GPUDirect™ RDMAで10倍の性能を実現

- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs

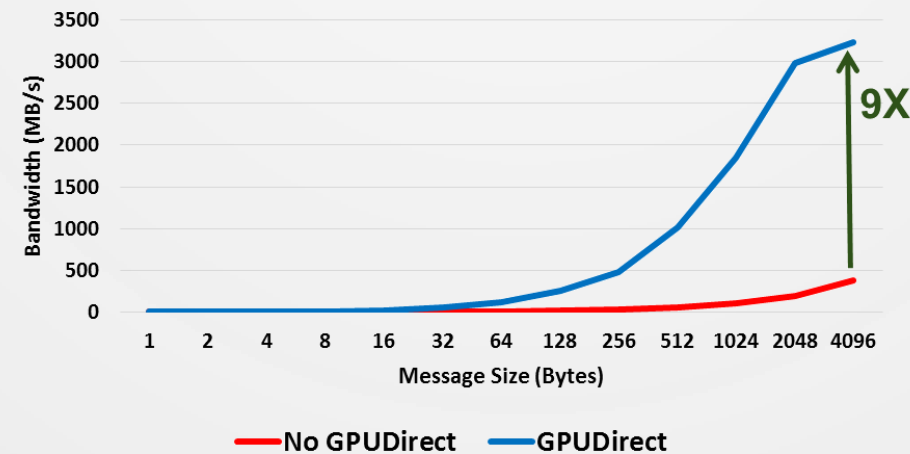
GPUDirect™ RDMA



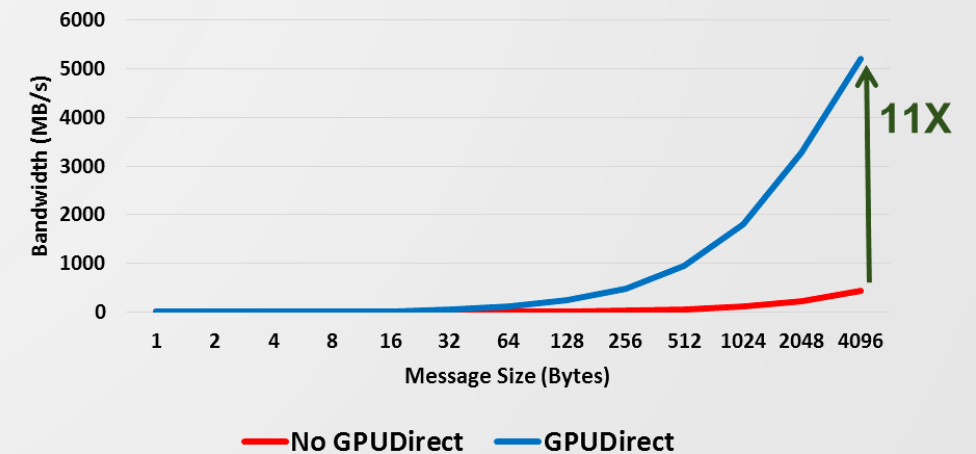
GPU-InfiniBand-GPU Latency



GPU-Mellanox-GPU Throughput (Uni-Dir)



GPU-Mellanox-GPU Throughput (Bi-Dir)



ノード間演算最適化 - SHARP

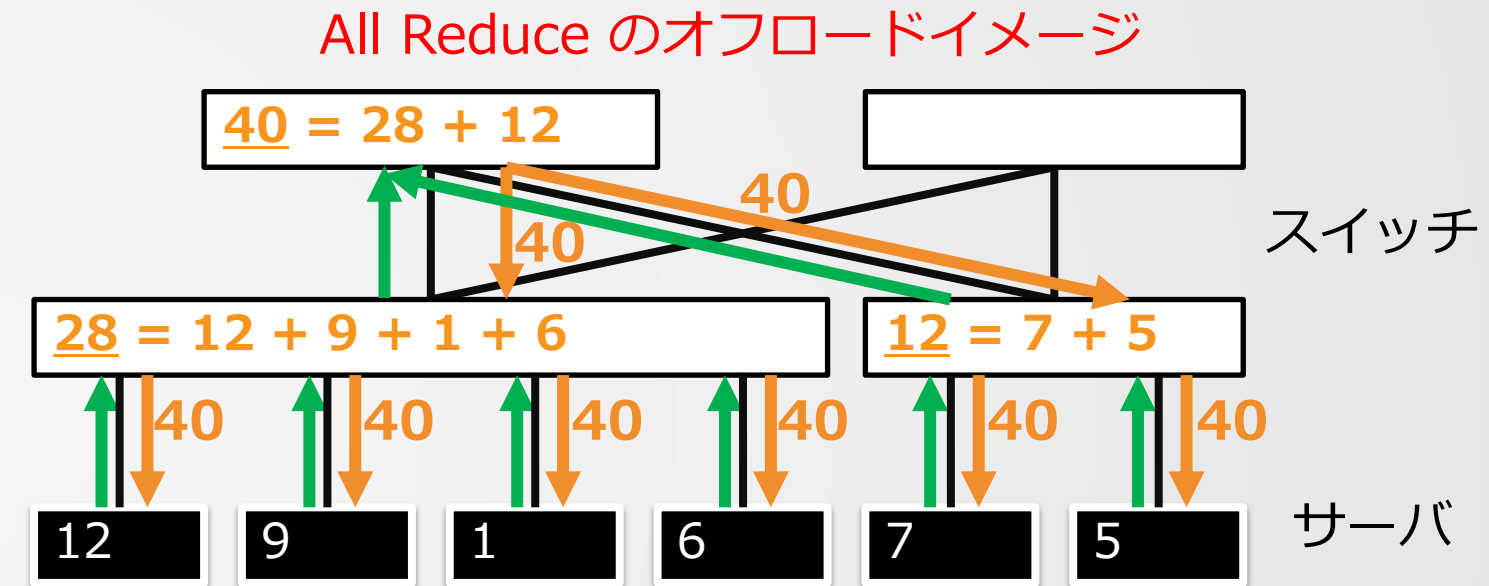
■ ノードが増えるとノード間演算 (コレクティブ処理) が爆発的に増える

- 100ノードならそれぞれのノードが他の99ノードとの演算を必要とする可能性
 - 機械学習における勾配平均値の計算やリデュース処理などなど、AIではこの様な演算を多用することが問題に

■ Mellanox SHARP ※で効率化

- コレクティブ処理をネットワークで集約
- スケールが増えてもノード間コミュニケーションに伴う効率低下が少ない
- 本機能を使用するためのソフトウェアは無償にて提供 (HPC-X)

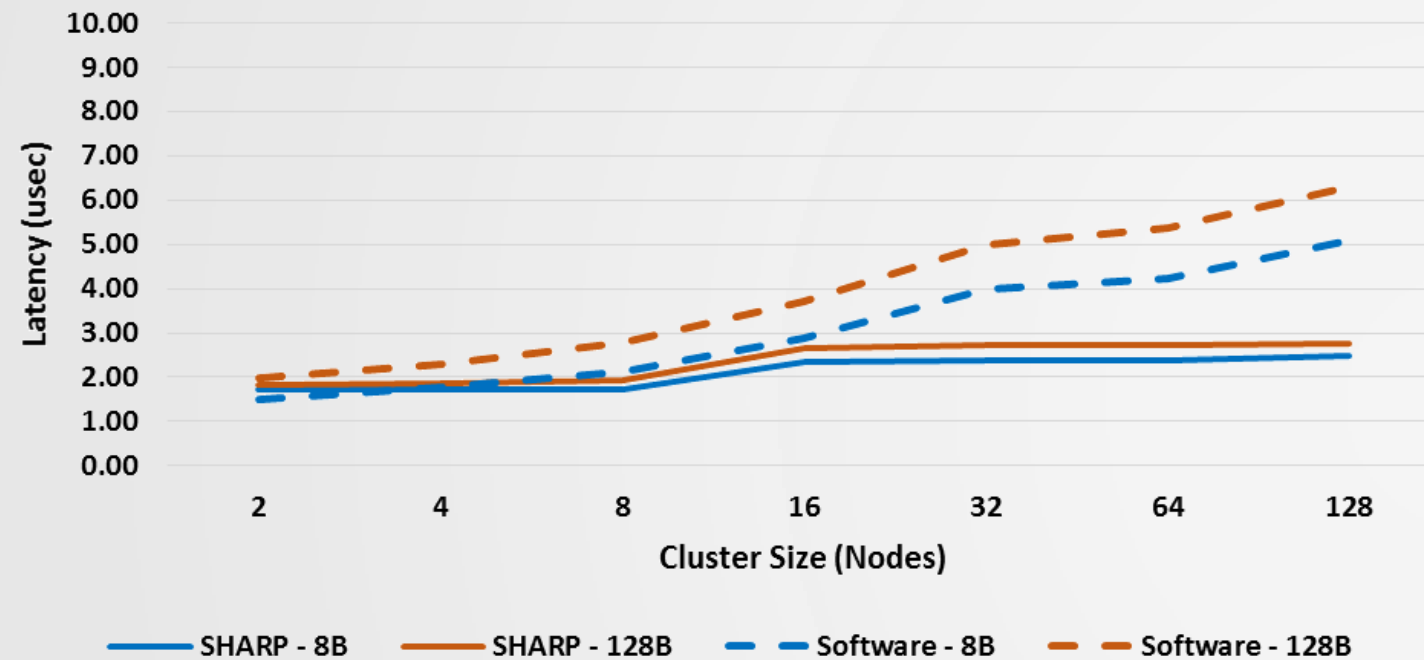
Barrier, Reduce, All-Reduce, Broadcast
Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
Integer and Floating-Point, 16 / 32 / 64 bit



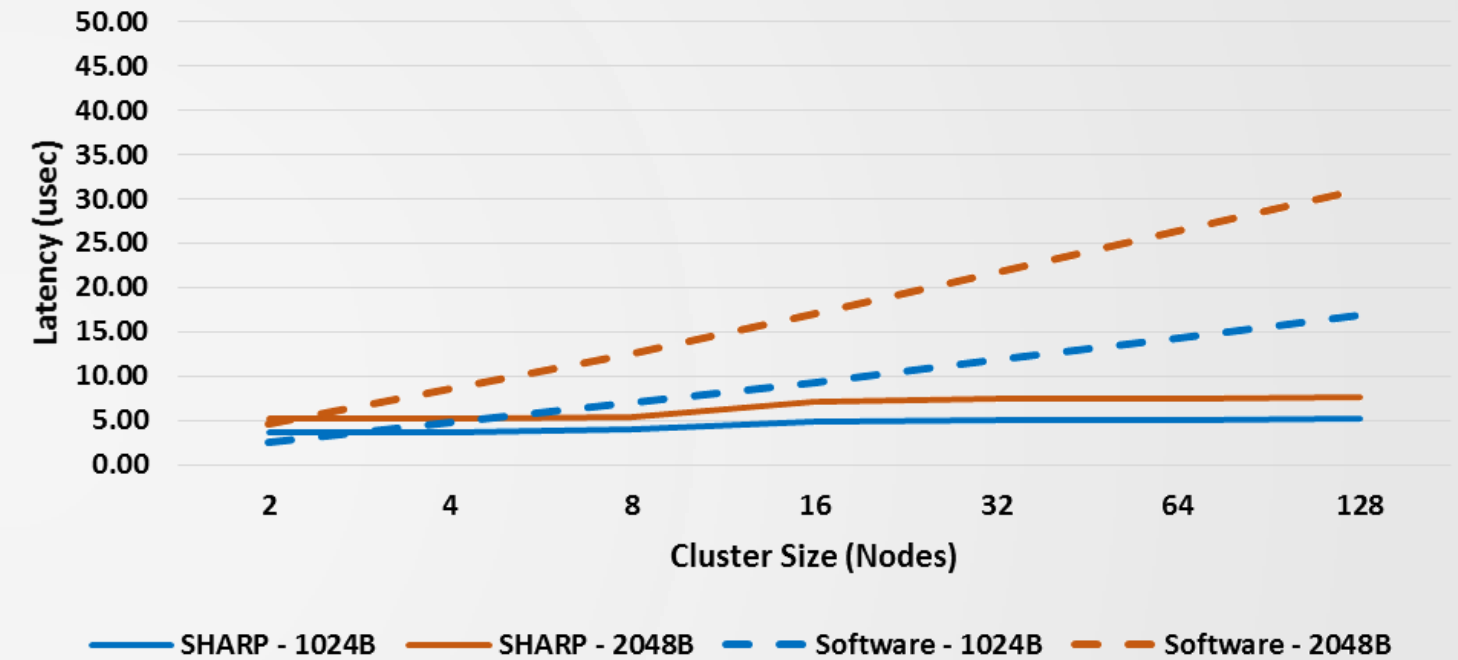
※Scalable Hierarchical Aggregation and Reduction Protocol

SHARPの優れた性能(AllReduce, 128 Nodes)

Allreduce Latency



Allreduce Latency

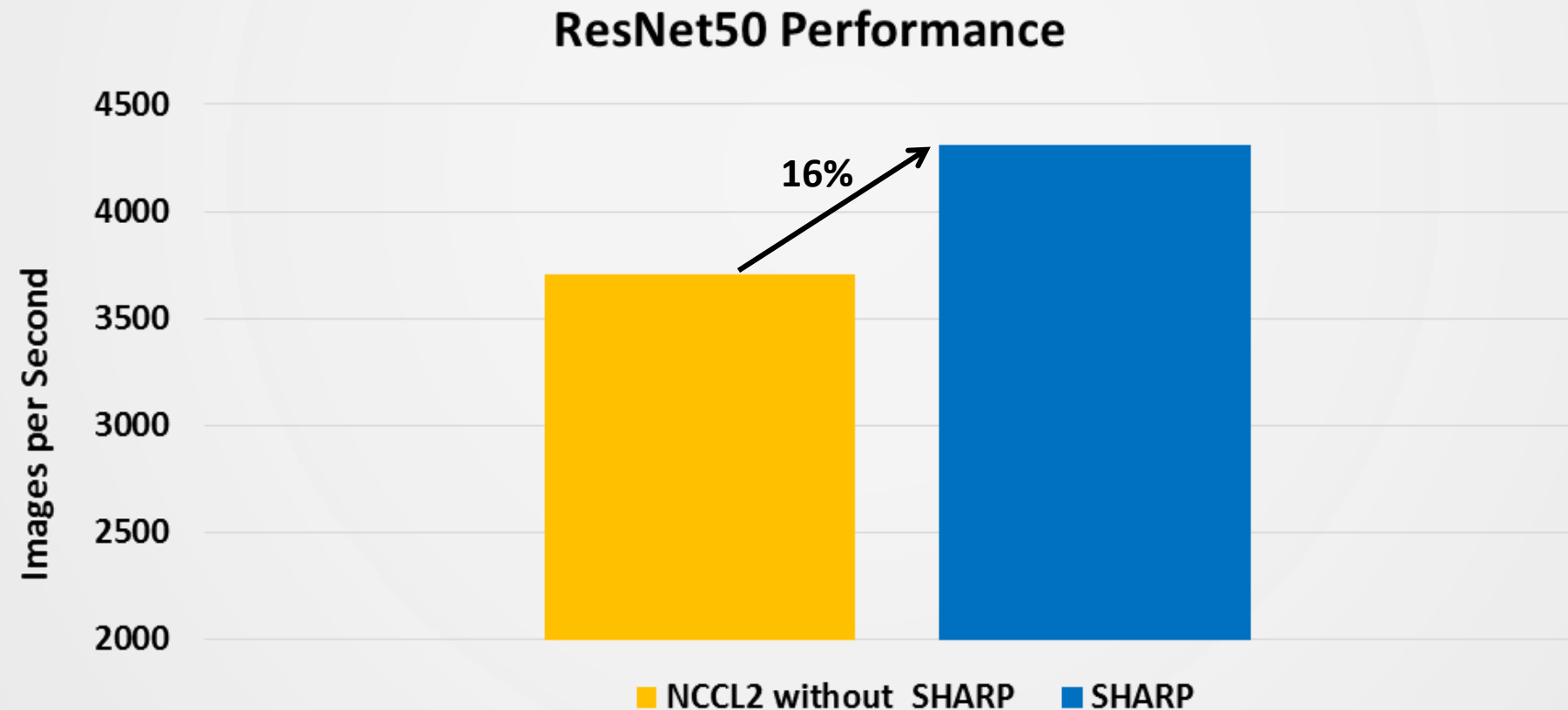


SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

AIの性能向上に寄与するSHARP (2.0)

HDRではAI/DL向けに最適化されたSHARP2.0へ進化

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



8 Nodes, 22 GPUs, HDR InfiniBand

Proven Advantages

- 最高のROIを実現
- 高スケールで柔軟性の高い高速なネットワークをエンドツーエンドで実現
- 莫大なエコシステムに裏付けされた標準技術(InfiniBand/Ethernet)
- x86, ARM, GPU, FPGAなど、アーキテクチャを問わずに対応
- オフロード機能やIn-Network Computingへの対応
- 柔軟なトポロジー: Fat Tree, Mesh, 3D Torus, Dragonfly+などなど
- I/O集約を実現: 計算ノード,ストレージ,単一ファブリックの管理
- 過去から将来にわたる互換性の維持



The Future Depends On Smart Interconnect



Thank You

