インターコネクトでAIを加速

～最新インターコネクト製品技術のご紹介

Mellanox – The Artificial Intelligence Interconnect Company

December 2017

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

会社概要

Ticker: MLNX NASDAQ®

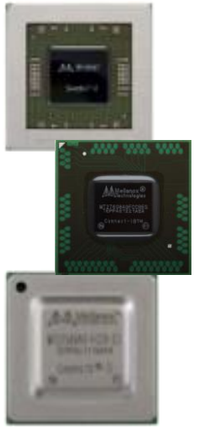
- **広帯域、低レイテンシーインターコネクトのリーディングカンパニー**
  - EDR 100Gb/s InfiniBand、10/25/40/50/56/100ギガビットEthernet
  - アプリケーションのデータ処理時間を大幅に削減
  - データセンターサービス基盤のROIを劇的に向上

- **会社概要**
  - 本社：ヨークナム（イスラエル）、サニーベール（米国）
  - 従業員数：全世界で2,900名（2017年1月末時点）

- **財務状況**
  - 2013年度売上 : $390.9M
  - 2014年度売上 : $463.6M
  - 2015年度売上 : $658.1M
  - 2016年度売上 : $857.5M
  - 2016年度税前利益率 : 21.0%
  - Cash + Investment : $328.4M（2016年12月末時点）

超高速ネットワーク市場をリードするエンドトゥエンド製品群

InfiniBand、Ethernetエンドトゥエンドソリューションを実現する製品群

IC	アダプタカード	NPU、マルチコア	スイッチ/ゲートウェイ	ソフトウェア	Metro / WAN	ケーブル/モジュール
						

AIとメラノックスの密接な関係

AIは幅広い分野での活用が期待されています

AIは、自然科学やビジネス、そして社会においてより重要な判断を
リアルタイムに行うために進化を続けています

ヘルスケア, さまざまなビジネス上の経営判断
知見の探求, セキュリティー, カスタマーサポートなどなど

より多くのデータ

+

より良いモデル

+

より高速な通信

GPUs

CPUs

FPGAs

Storage

More Data → Faster Interconnect → Better Insight → Competitive Advantage

ただのビッグデータ？… それはとてもとてもビッグなデータ

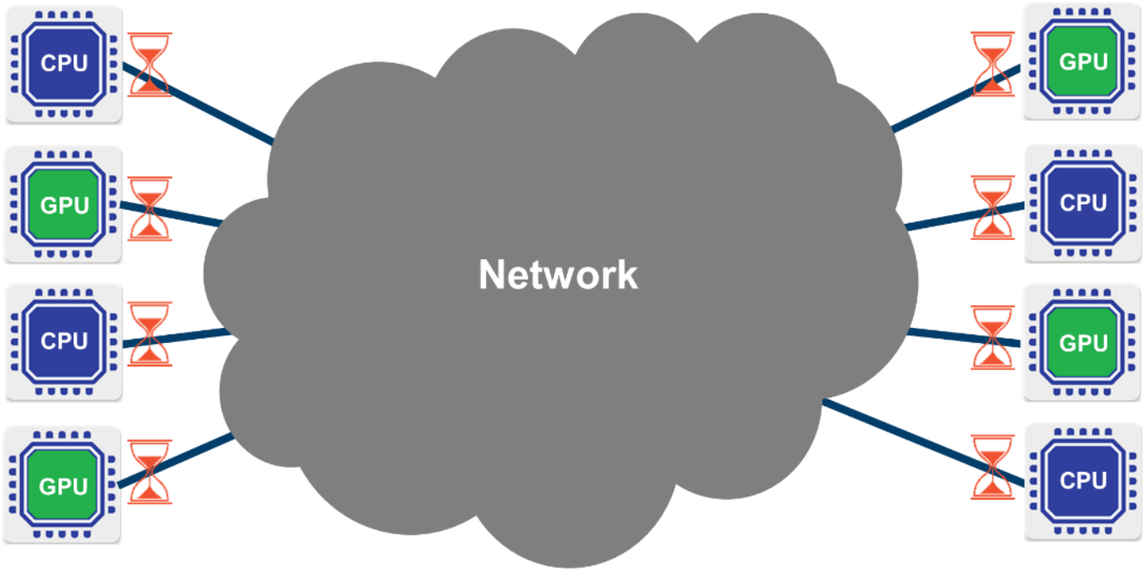
- 自動運転自動車から発生するデータは、およそ8時間の運転で40TBにも達します。
（フルサービスの自動運転自動車の場合）

- The Pratt & Whitney PC1000Gエンジンは、5000ものセンサーが内蔵されており秒間およそ10GBものデータを発生させます。これは、12時間のフライトでおよそ844TBにもなります。

メラノックスは、このような大量のデータ通信を支えるインターコネクトを提供します

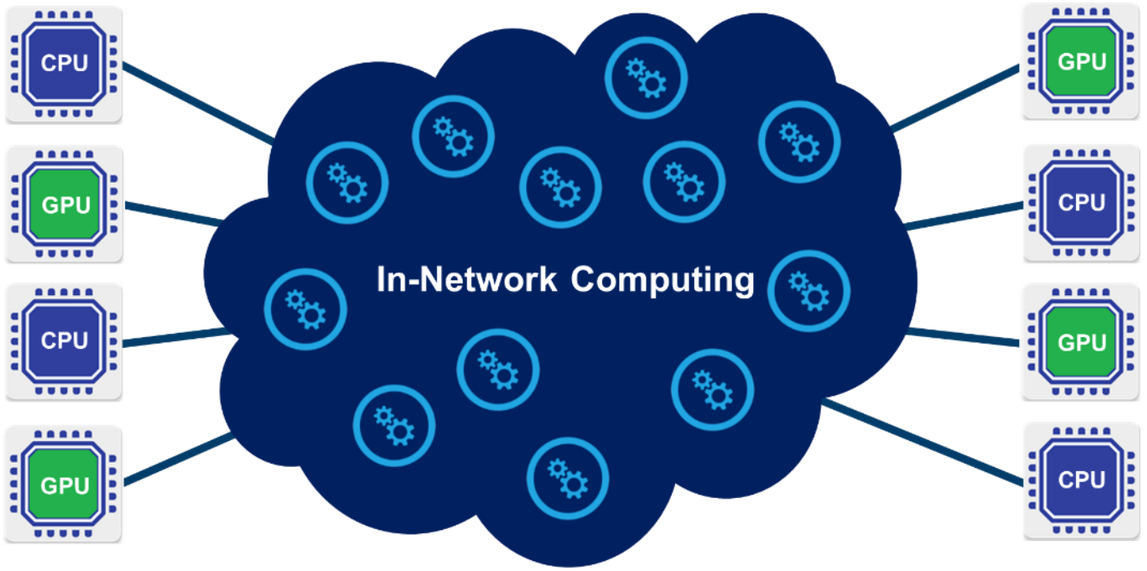
爆発的なデータの増加と求められるIn-Network Computing

CPU-Centric (Onload)



つねに大量のデータを「待つ」必要
パフォーマンスのボトルネック
数十usのノード間通信遅延

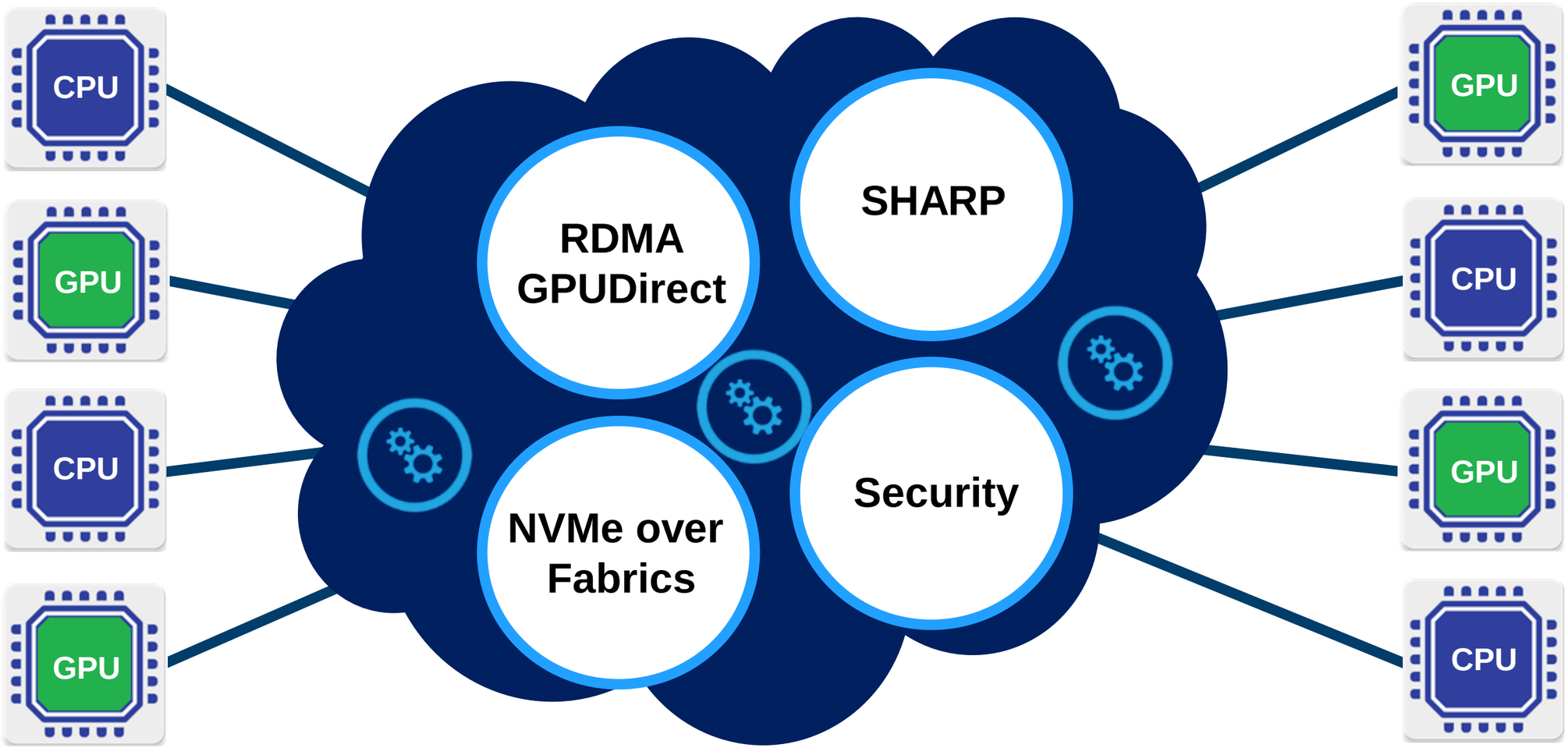
Data-Centric (Offload)



データの流れに応じた解析が可能に
数usのノード間通信遅延

In-Network Computingにより、効率的なデータ解析を実現

メラノックスの技術がマシンラーニングを加速します

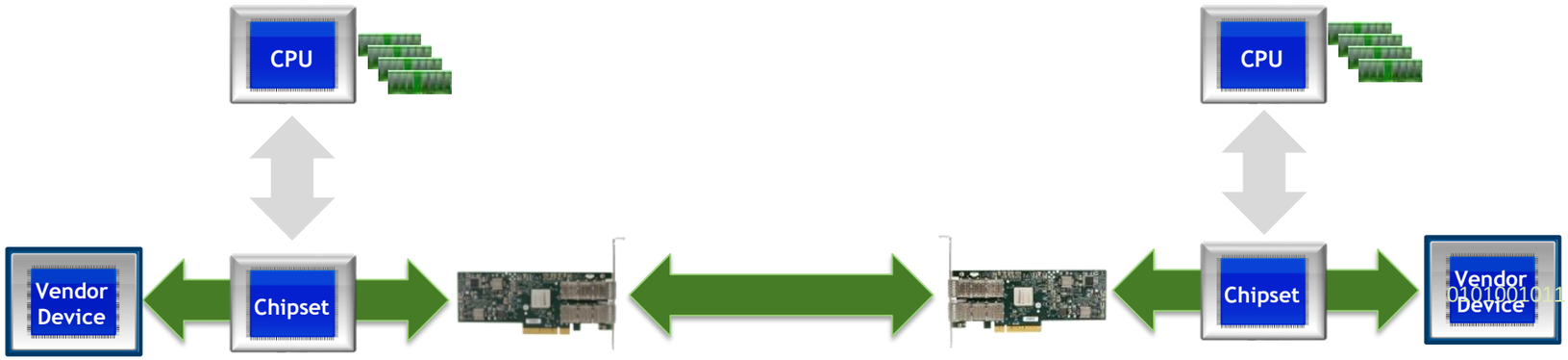


In-Network Computingにより、投資対効果を高めます

GPUDirect® RDMA and ASYNC技術のご紹介

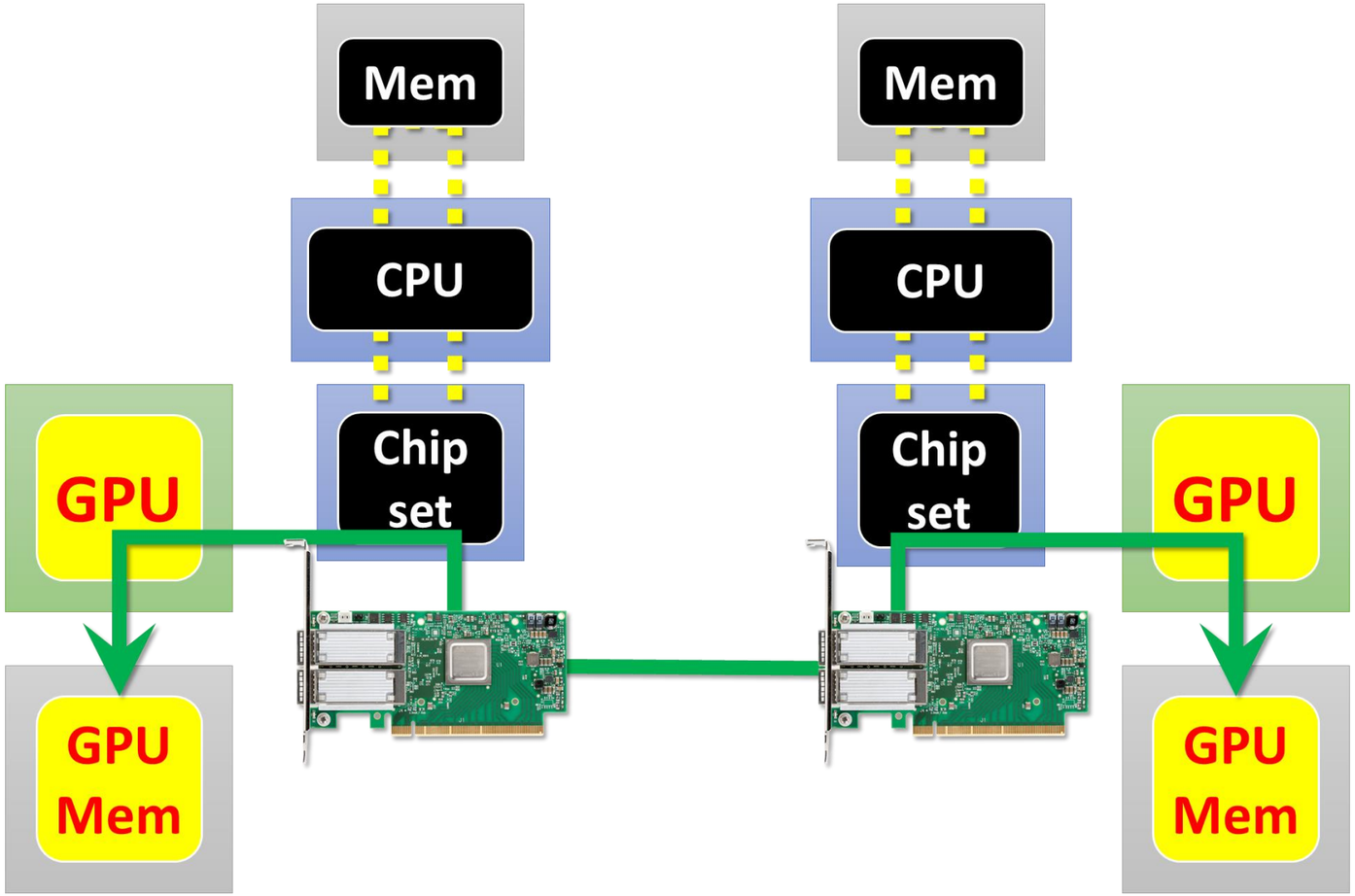
GPUDirect™ とは

- ノードをまたいだGPU間のコミュニケーションに伴う遅延を劇的に削減
- Mellanox OFEDにてサポート
- Mellanoxアダプタとサードパーティデバイス間のP2P通信をサポート
- CPU負荷や、システムメモリへのコピーが不要
- InfiniBand および RoCEで使用可能



Deep Learning性能を効果的に加速

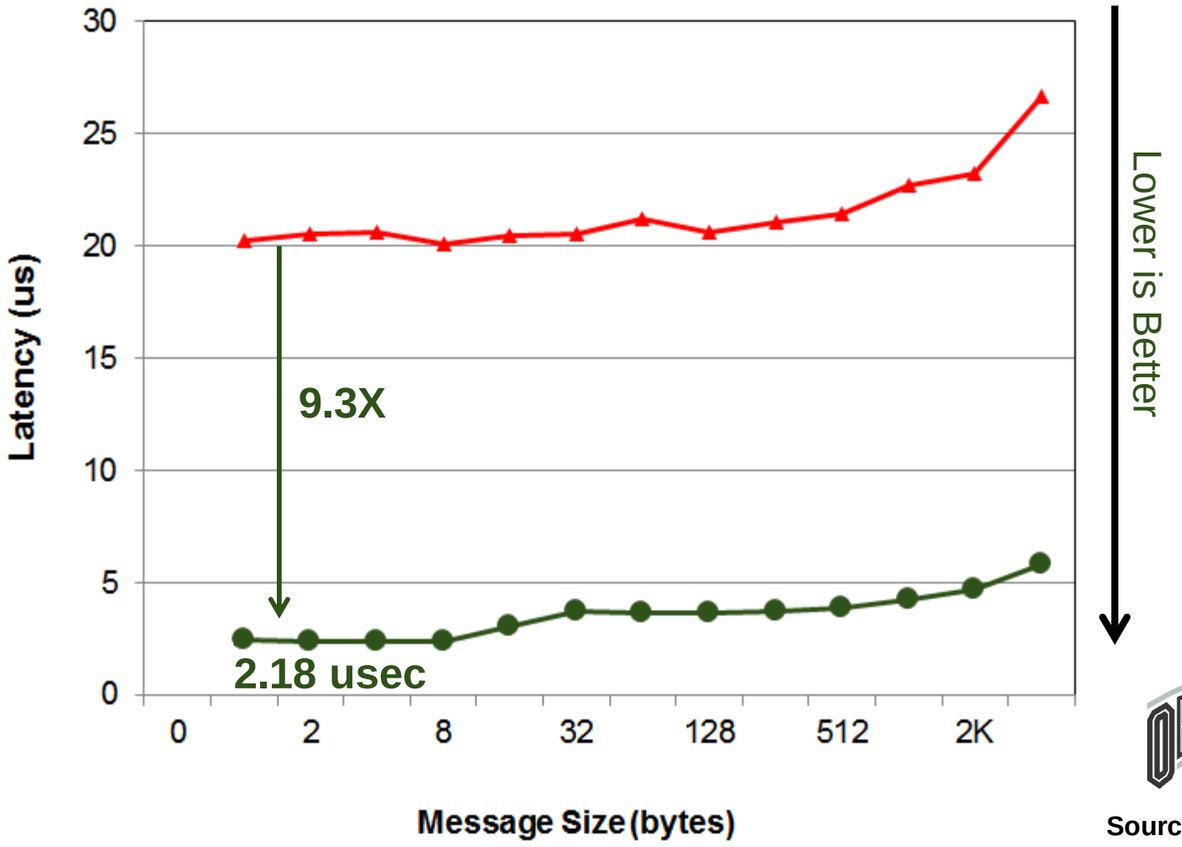
GPUDirect™ RDMA の概念



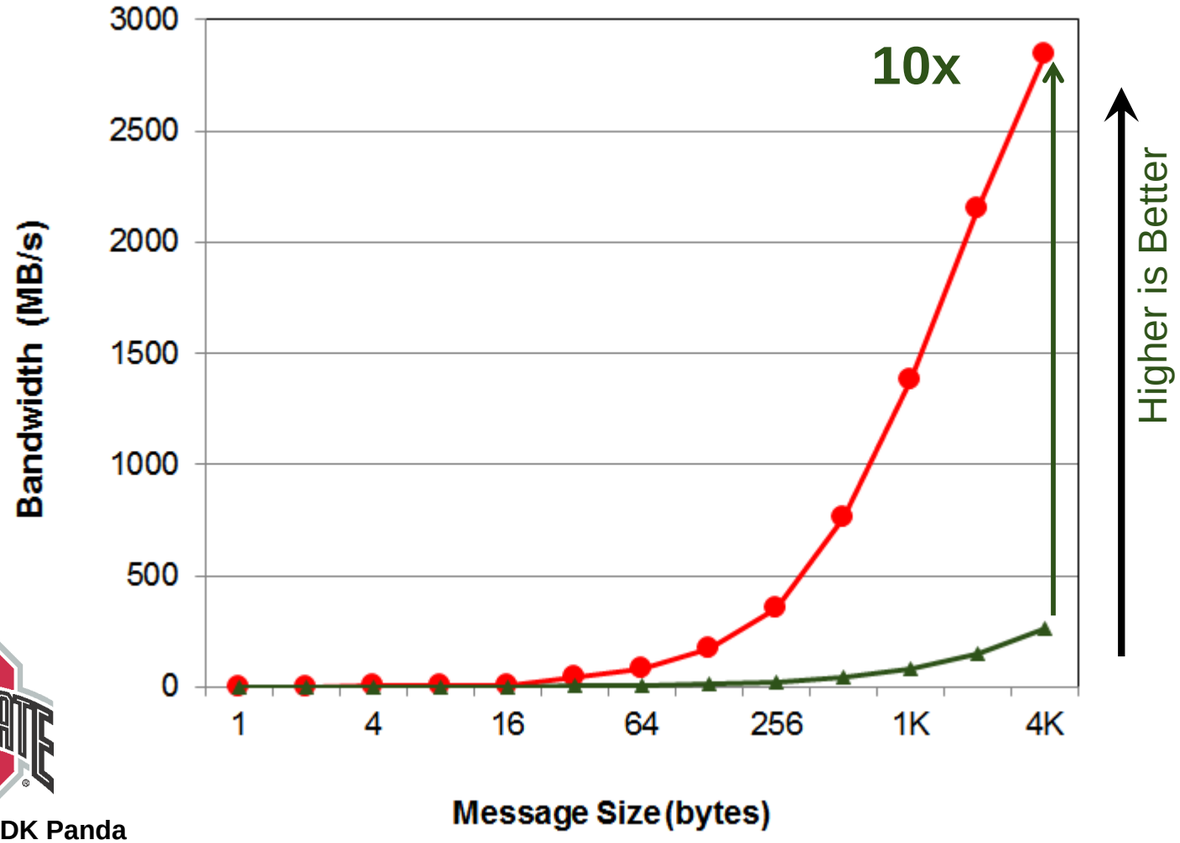
GPU間で直接的な通信を可能に

GPUDirect™ RDMA による効果

GPU-GPU Internode Latency



GPU-GPU Internode Bandwidth



Source: Prof. DK Panda

9.3X Better Latency

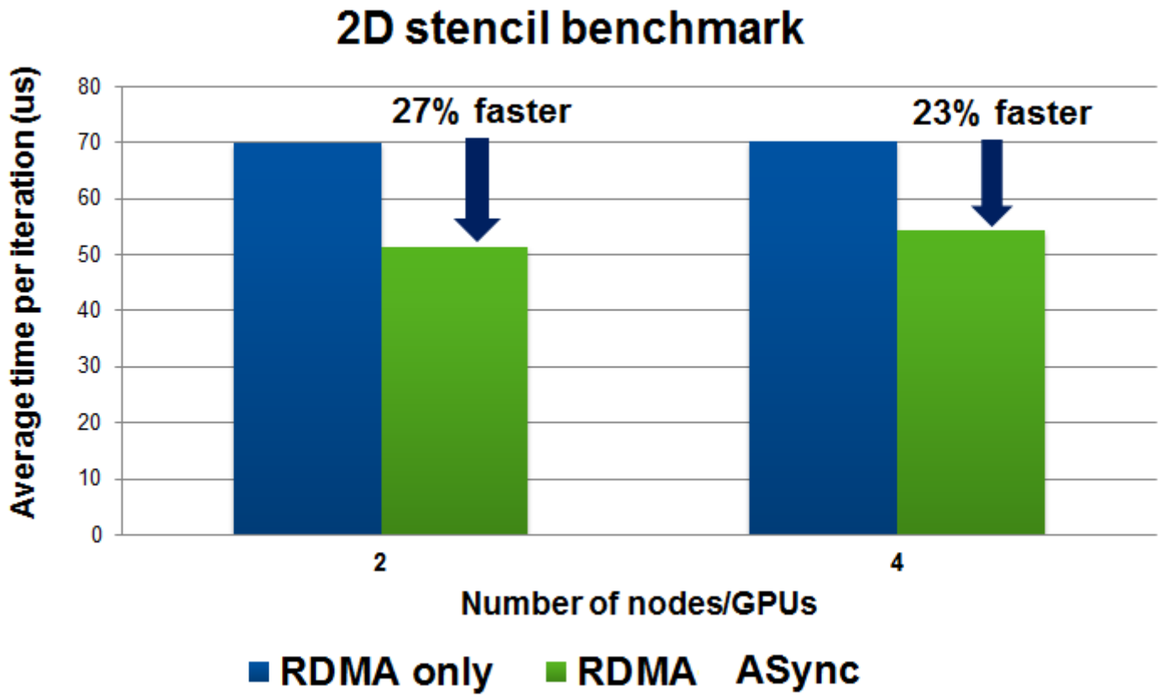
10X Better Throughput

GPUDirect ASYNC

- GPUDirect RDMA (3.0) – GPUとメラノックスデバイス間で直接のデータ通信を可能に
  - コントロールパスは従来通りCPUで処理
    - CPUが、通信タスクをGPU上に準備し、キューする
    - GPUが、Mellanox に対して通信タスクの開始をトリガー
    - Mellanox が、GPUメモリを直接アクセスして通信を開始

- GPUDirect ASYNC (GPUDirect 4.0)
  - コントロールパスも含めてMellanoxがGPUに対して直接アクセス

GPU Clustersに対して最大限の性能を発揮



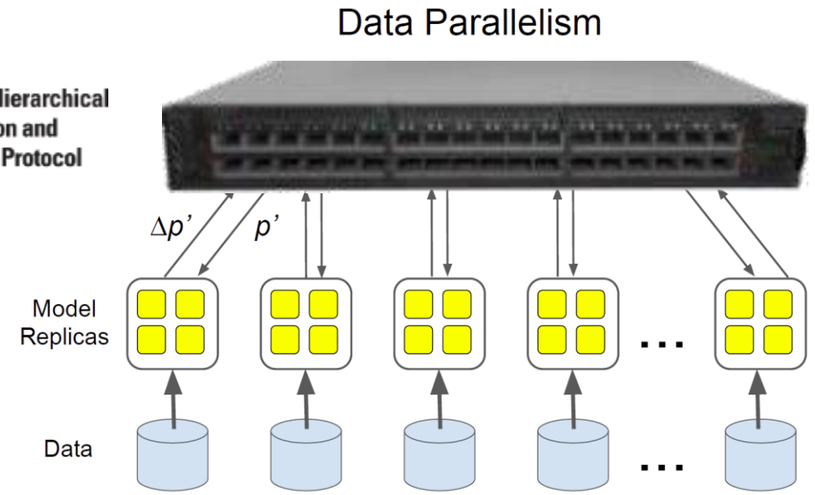
Mellanox SHARP

Scalable Hierarchical Aggregation Reduction Protocol

Mellanox SHARP によりAI性能を加速

AIにおいては、パラメータサーバのCPUがすぐにボトルネックに
(だいたい 4 nodes から)

SHARP Scalable Hierarchical Aggregation and Reduction Protocol



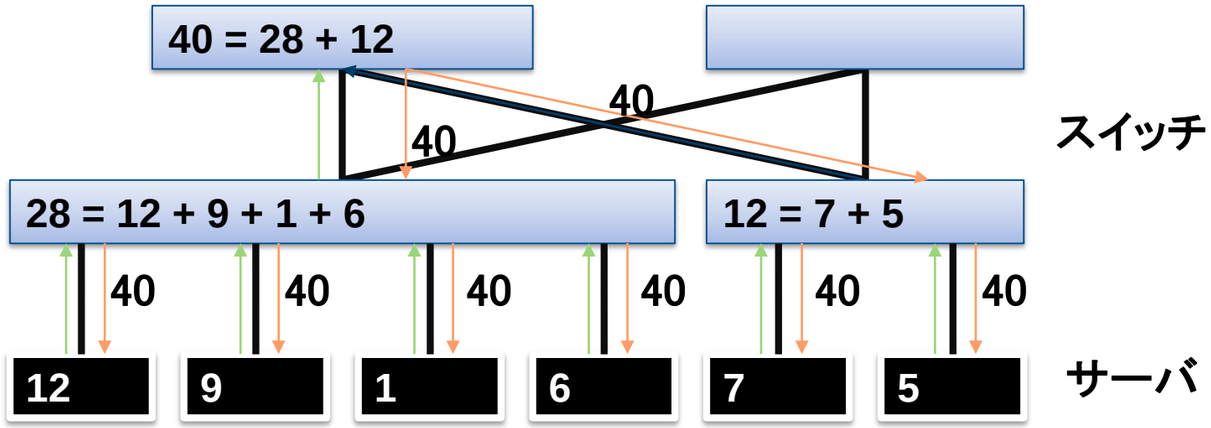
SHARPによってMellanox スイッチが
勾配平均値を処理することで
パラメータサーバ不要でオーバヘッドを削減

AI/DLアプリケーションへの最適化 – コレクティブオフロード –

- 大規模分散処理における課題
  - ノード間通信の量が多い。その処理をホストプロセッサで処理している
    - Machine Learning における勾配平均値の計算やリデュース処理などなど
    - ノードが増えるほど、これらの処理負担が増大し、思ったほどスケールしなくなる。

- Mellanox のソリューション
  - SHARP ※
    - コレクティブオフローディングメカニズムの提供
      - Barrier, Reduce, All-Reduce, Broadcast
      - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
      - Integer and Floating-Point, 16 / 32 / 64 / 128 bit

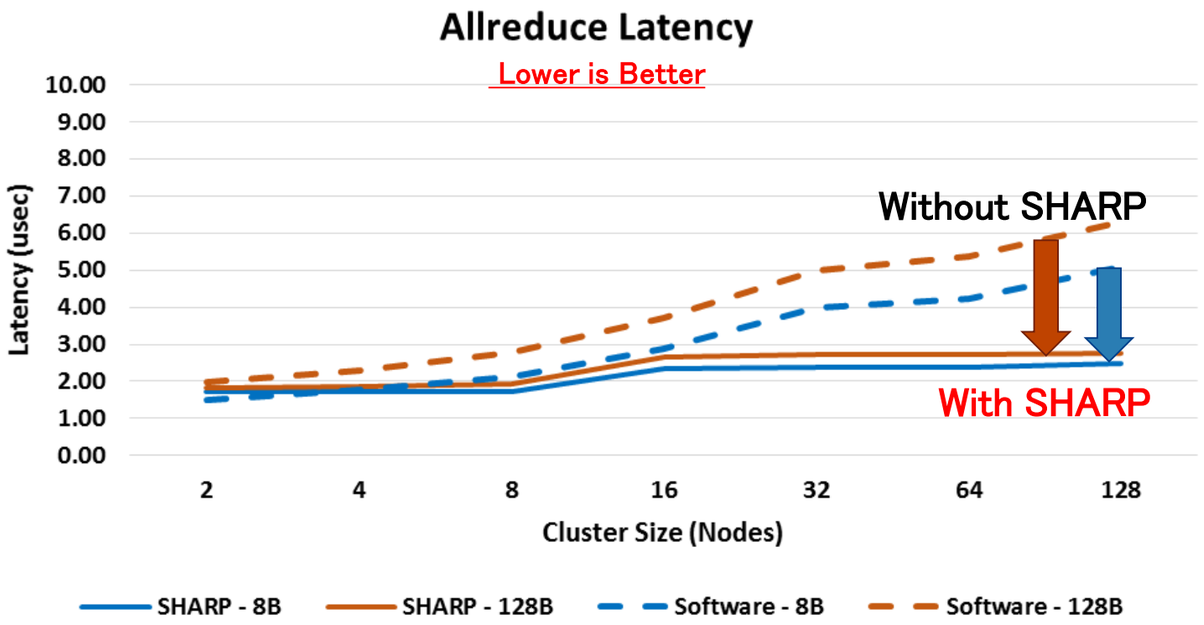
All Reduce のオフロードイメージ

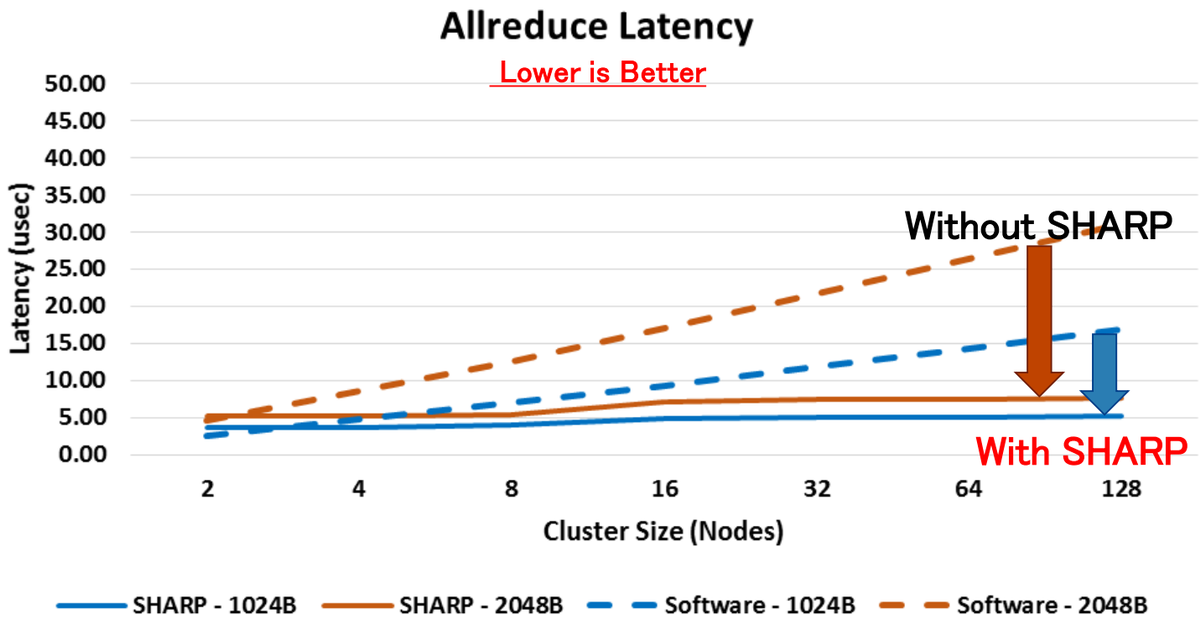


※Scalable Hierarchical Aggregation and Reduction Protocol

SHARP使用によるMPI通信の最適化

All Reduce 処理における遅延性能

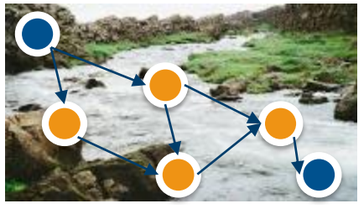




SHARPを使用することで、75%の遅延削減を実現

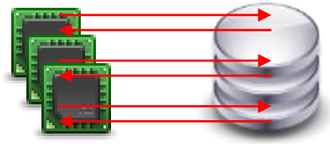
In-Network Computing は将来技術ではなく、今使える技術

In Network Computing



リアルタイム
(ストリーム)処理

リアル
タイム
応答



高トランザクション処理

最適化された
分析処理

ストレージ内、ネットワークでの移動時も処理を行うことで、スケール・性能を同時に向上

Mellanoxにより、AIシステムにおいて最大限のROIを実現

60% 高い投資対効果
2.5x の性能 及び95% のスケールアウト効率
CapExおよびOpExの 50%削減
ディープラーニングおいて最大のパフォーマンス、スケーラビリティー、生産性を実現

Caffe
IBM
TensorFlow
Microsoft Cognitive Toolkit
NVIDIA
Chainer
PaddlePaddle
Tencent 腾讯

Mellanoxは、AIの性能を加速させます

次世代製品のご紹介

High can Accelerate

InfiniBand H·D·R 200

- **Higher** performance
- **Higher** efficiency
- **Higher** scalability
- **Higher** speed

世界初のHDR（200G）InfiniBandスイッチ

世界初のHDR（200G）IB HCA

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™