Mellanox AI Advantage

次世代AIが求める高効率インターコネクト

Feb 2018

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

ビッグデータ? … それはとてもとてもビッグなデータ

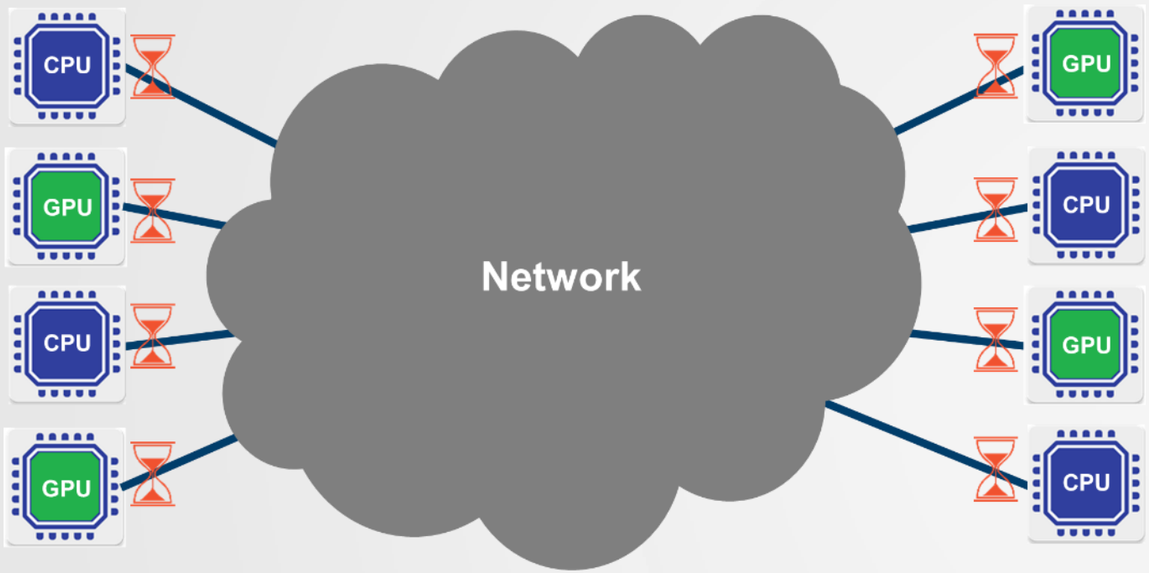
- 自動運転自動車から発生するデータは、およそ8時間の運転で40TBにも達します。（フルサービスの自動運転自動車の場合）

- The Pratt & Whitney PW1000G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

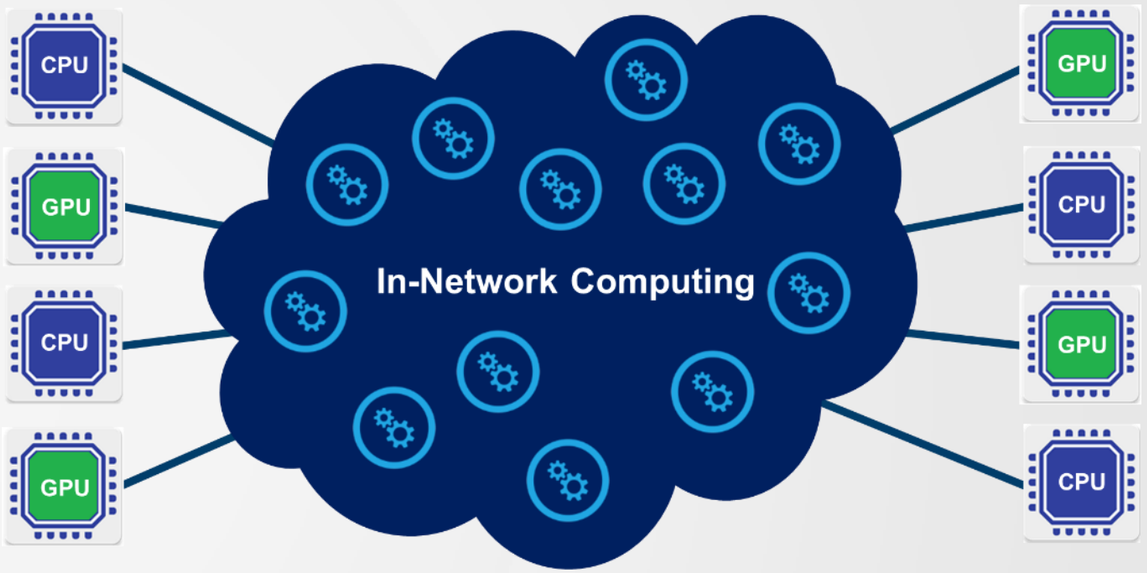
爆発的なデータの増加とIn-Network Computing

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

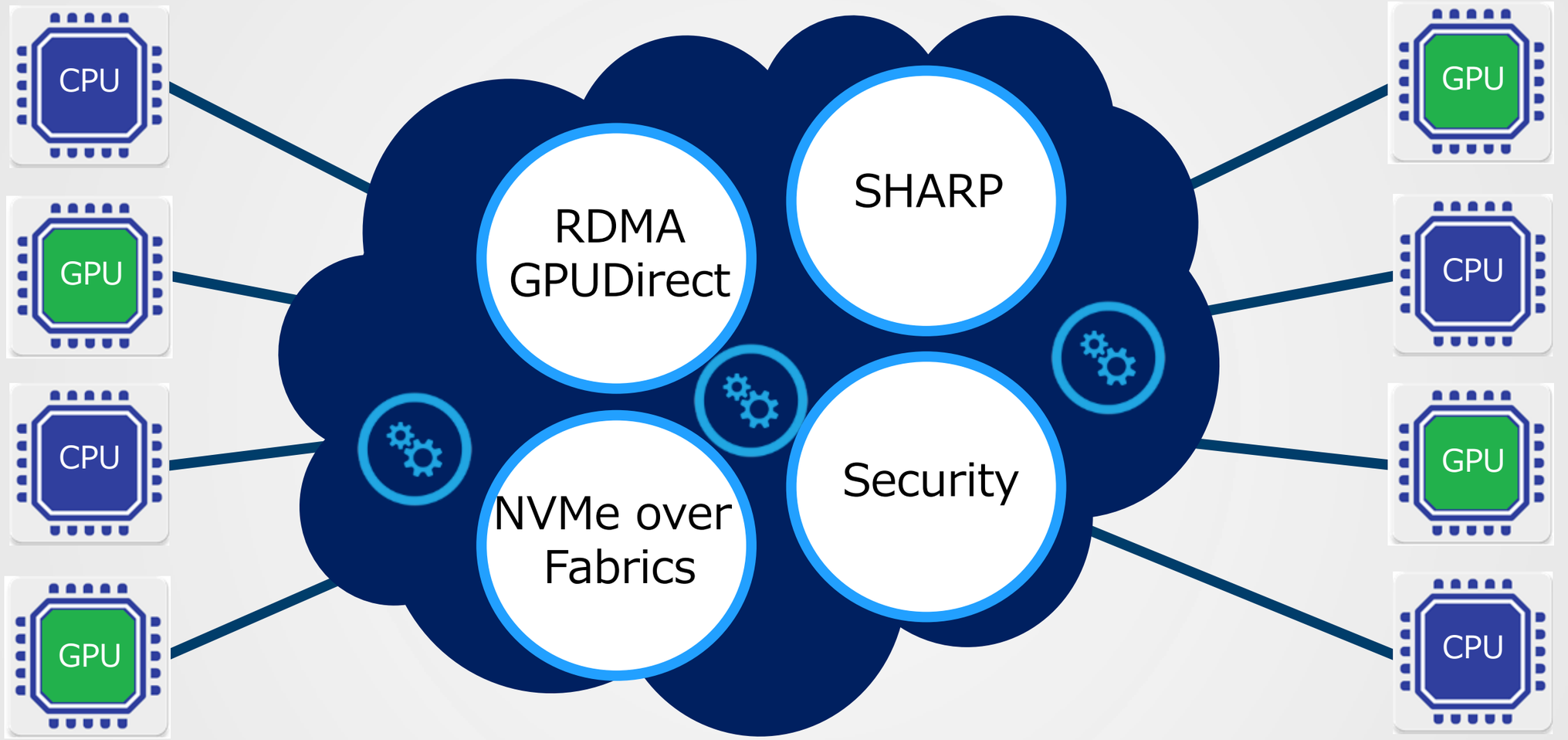
Data-Centric (Offload)



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

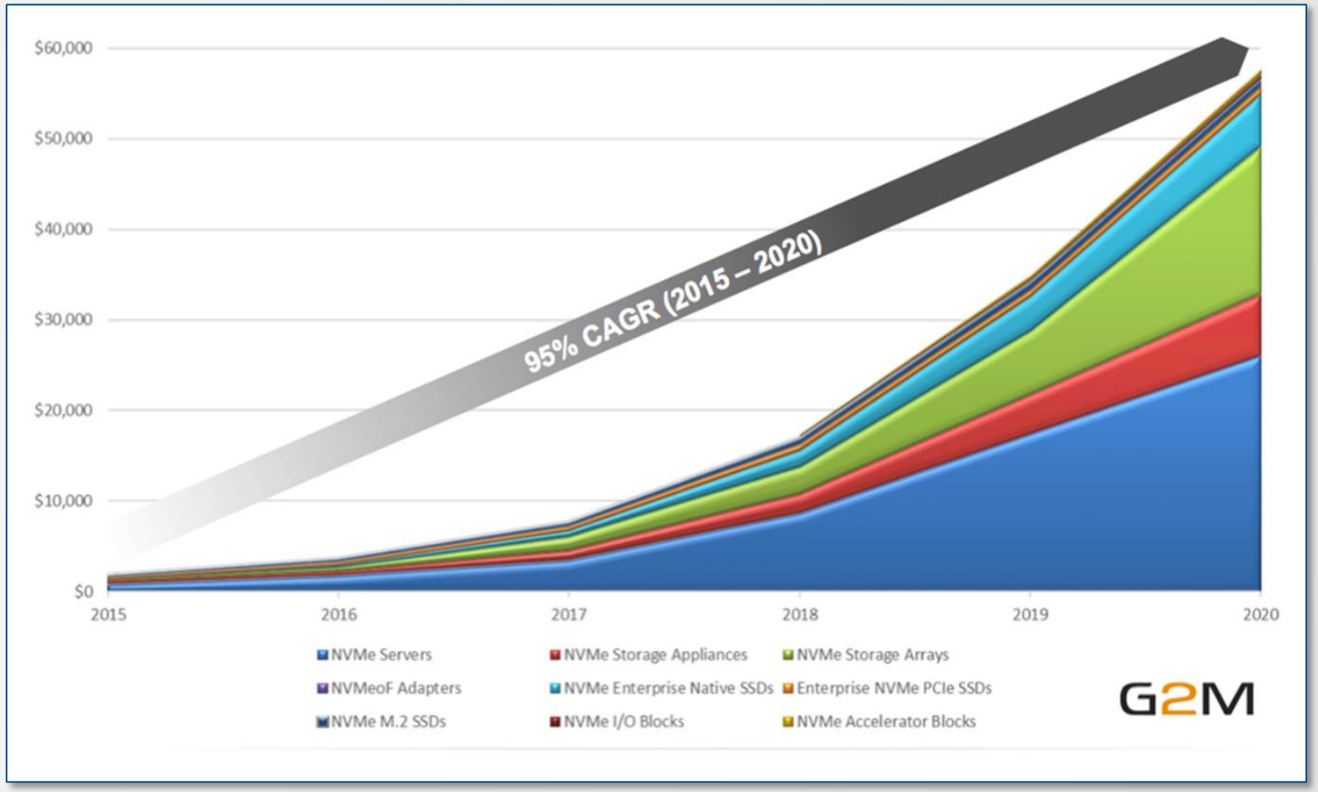


In-Network Computingにより、投資対効果を高めます

NVMe over Fabrics

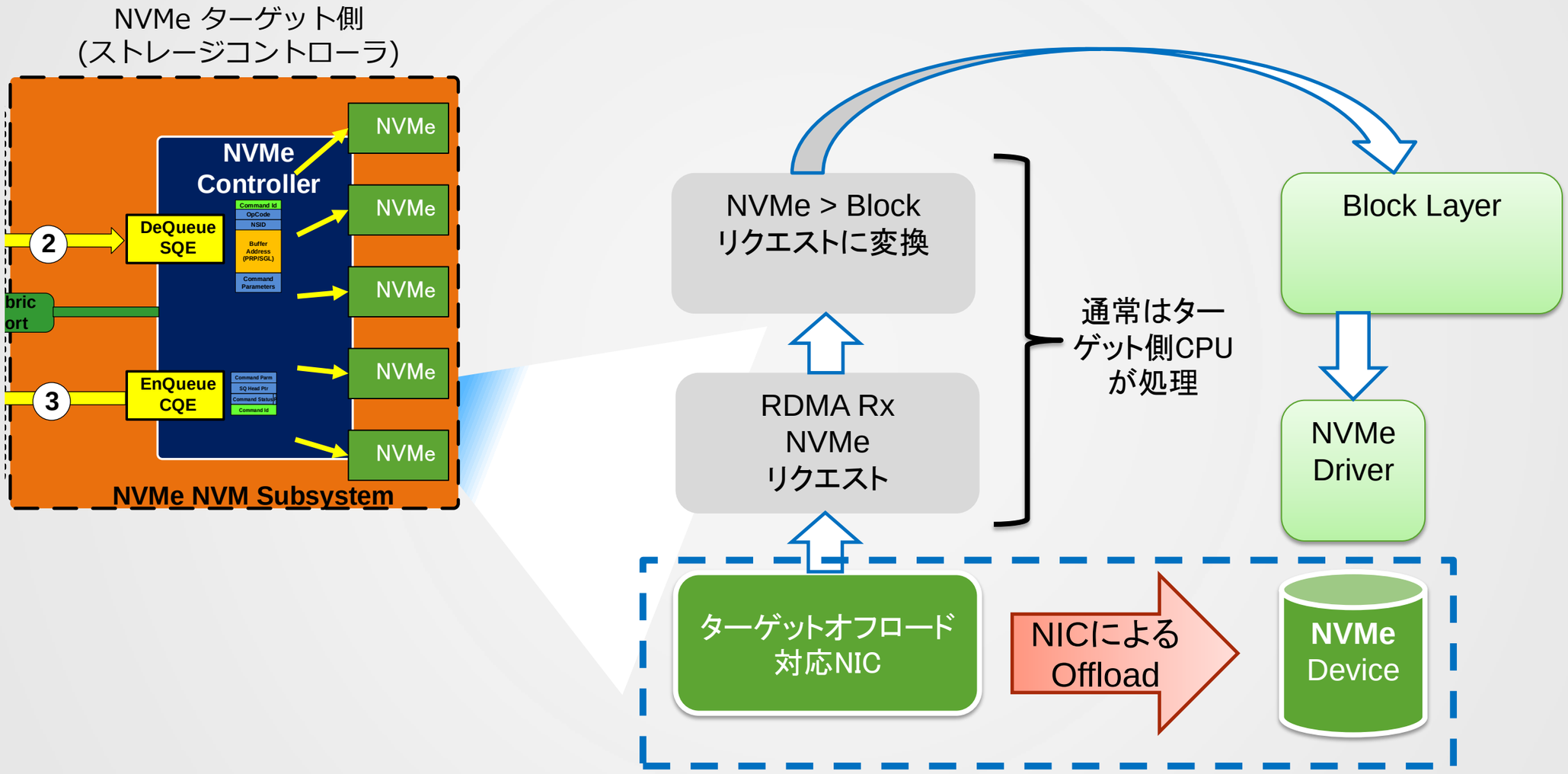
NVMeをとりまく市場動向

- 2020年までに 50% のエンタープライズサーバとストレージアプライアンスはNVMe をサポート
- 2020年までに 40% の all-flash arrayはNVMe ベースになる
- 2020年までに NVMe SSDの出荷量は25百万個に成長
- 2020年までに 740,000 NVMe over Fabrics アダプターが出荷される
- NVMe over Fabrics 市場の >75% がRDMA NIC になる



http://www.storagenewsletter.com/rubriques/market-reportsresearch/nvme-market-at-57-billion-by-2020-with-95-cagr-g2m-research/

Mellanoxが提供するNVMe over Fabrics Target Offload

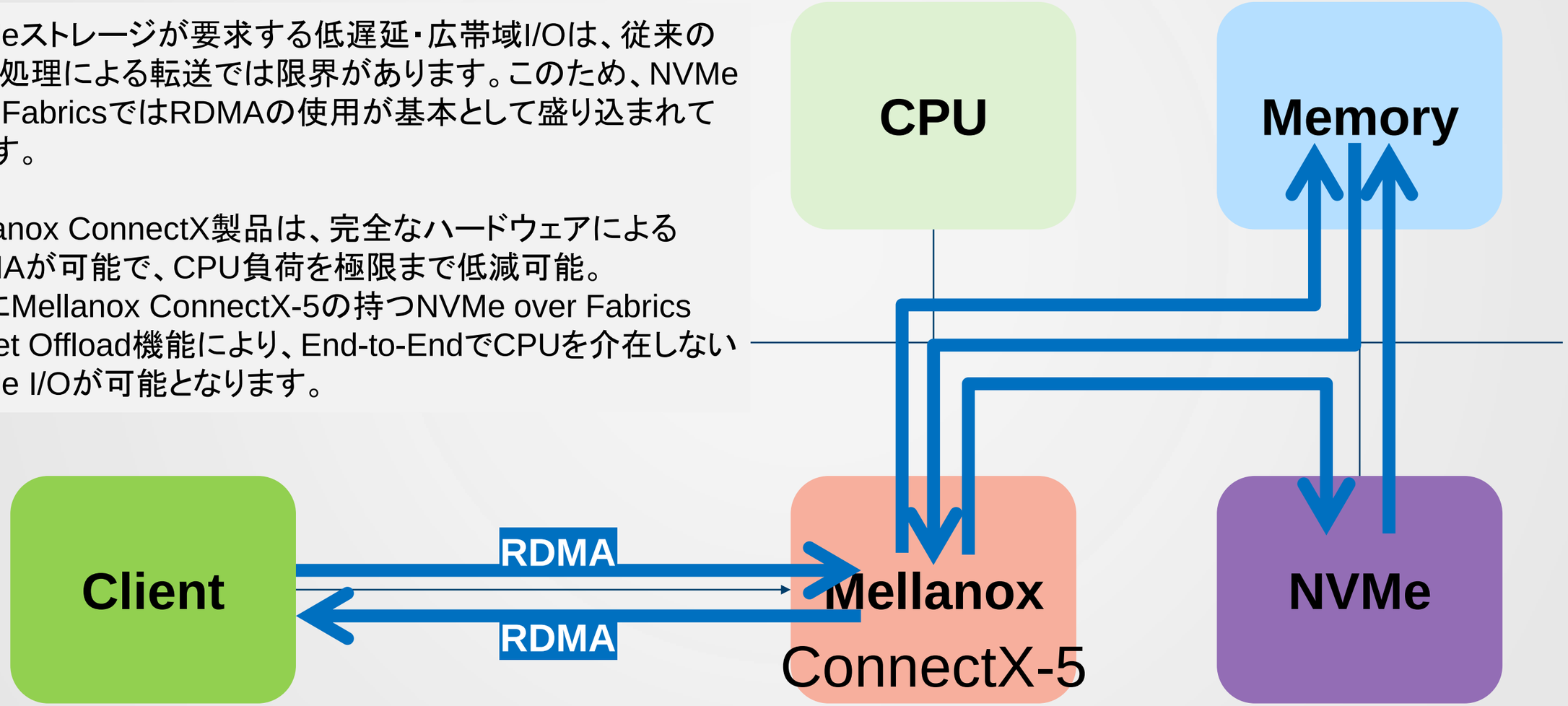


ConnectX-5/Bluefield はターゲット処理も含め、NICにオフロード

CPUを介在せずにストレージアクセスを実現

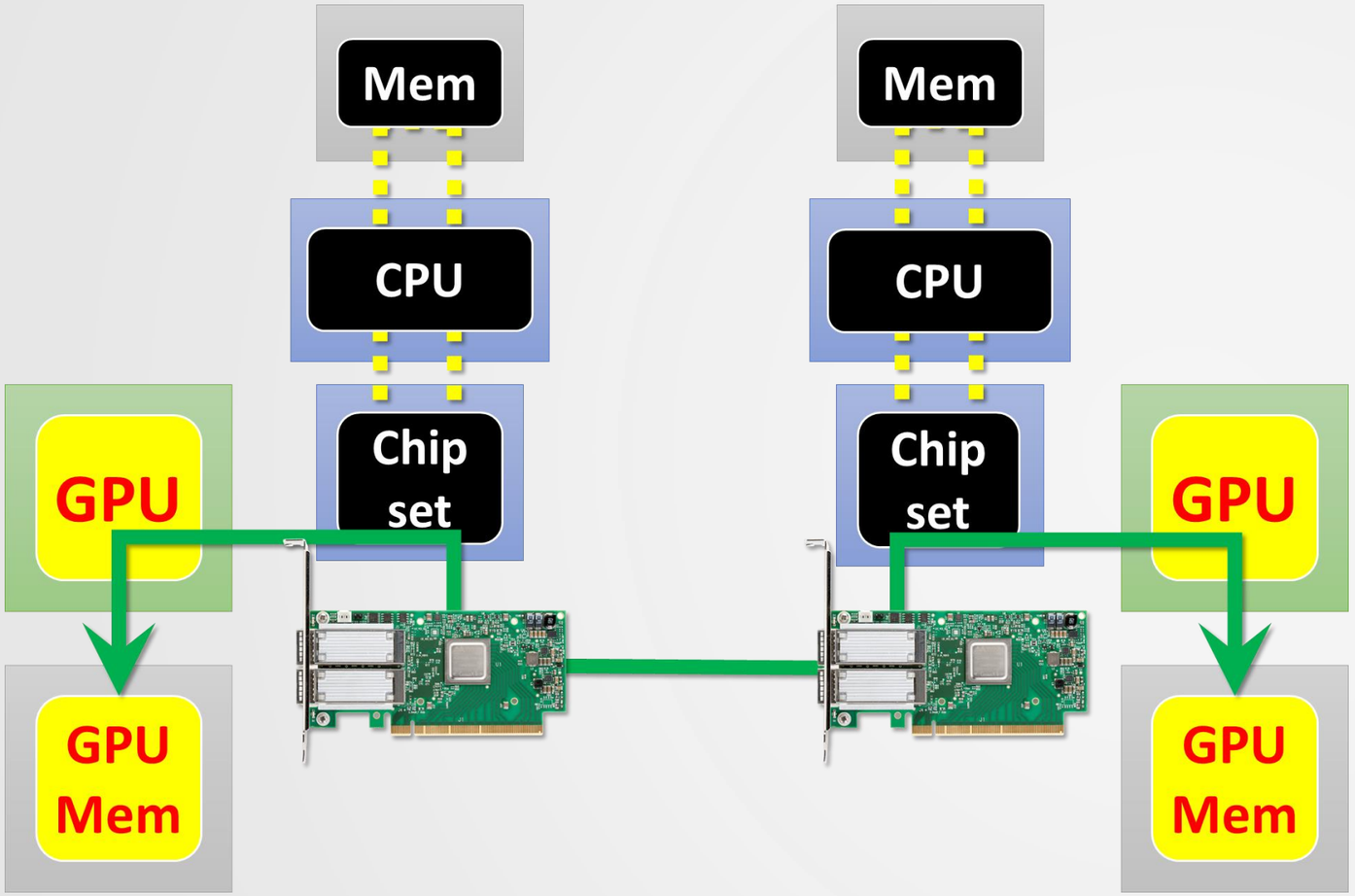
NVMeストレージが要求する低遅延・広帯域I/Oは、従来のCPU処理による転送では限界があります。このため、NVMe over FabricsではRDMAの使用が基本として盛り込まれています。

Mellanox ConnectX製品は、完全なハードウェアによるRDMAが可能で、CPU負荷を極限まで低減可能。
さらにMellanox ConnectX-5の持つNVMe over Fabrics Target Offload機能により、End-to-EndでCPUを介在しないNVMe I/Oが可能となります。



GPU Direct/RDMA

GPU Direct RDMAのご紹介

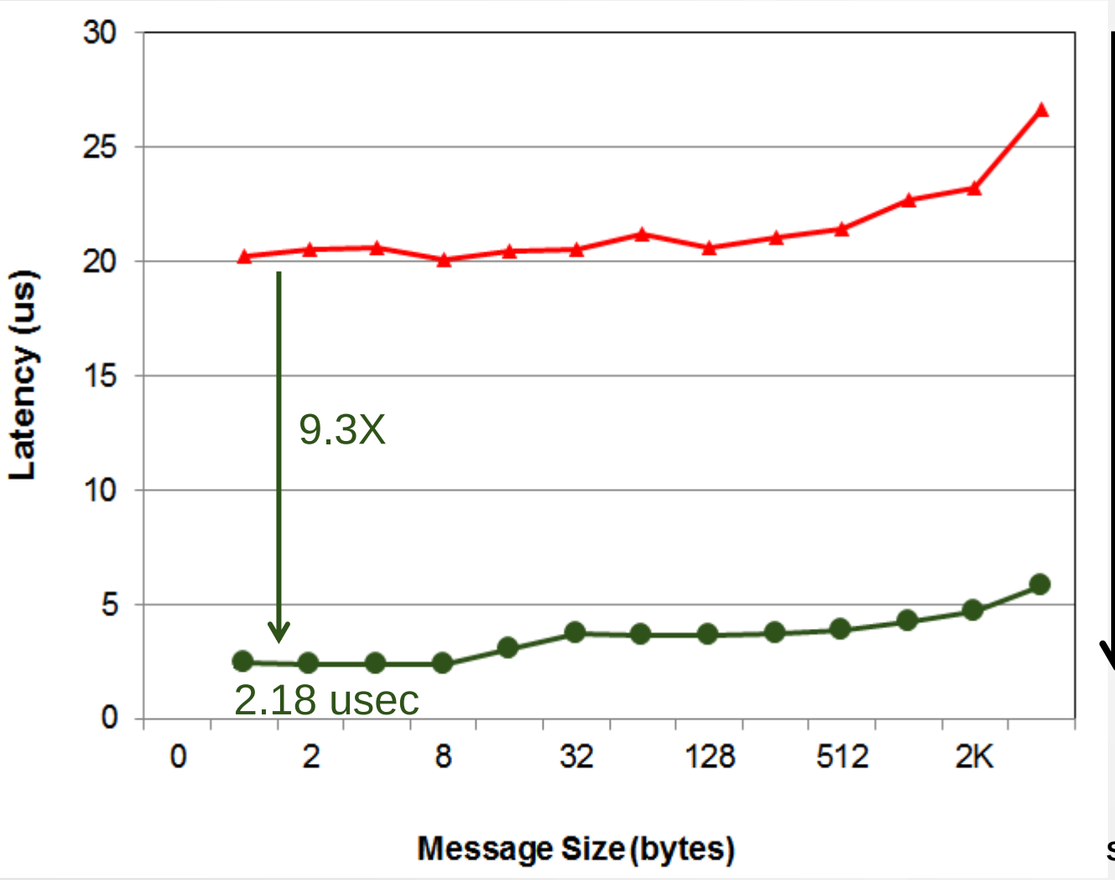


- ノードをまたいだGPU間のコミュニケーションに伴う遅延を劇的に削減
- Mellanox OFEDにてサポート
- Mellanoxアダプタとサードパーティデバイス間のP2P通信をサポート
- CPU負荷やシステムメモリへのコピーが不要
- InfiniBand および RoCEで使用可能
- Mellanoxは真のハードウェアRDMAにより、CPU消費を極限まで低減

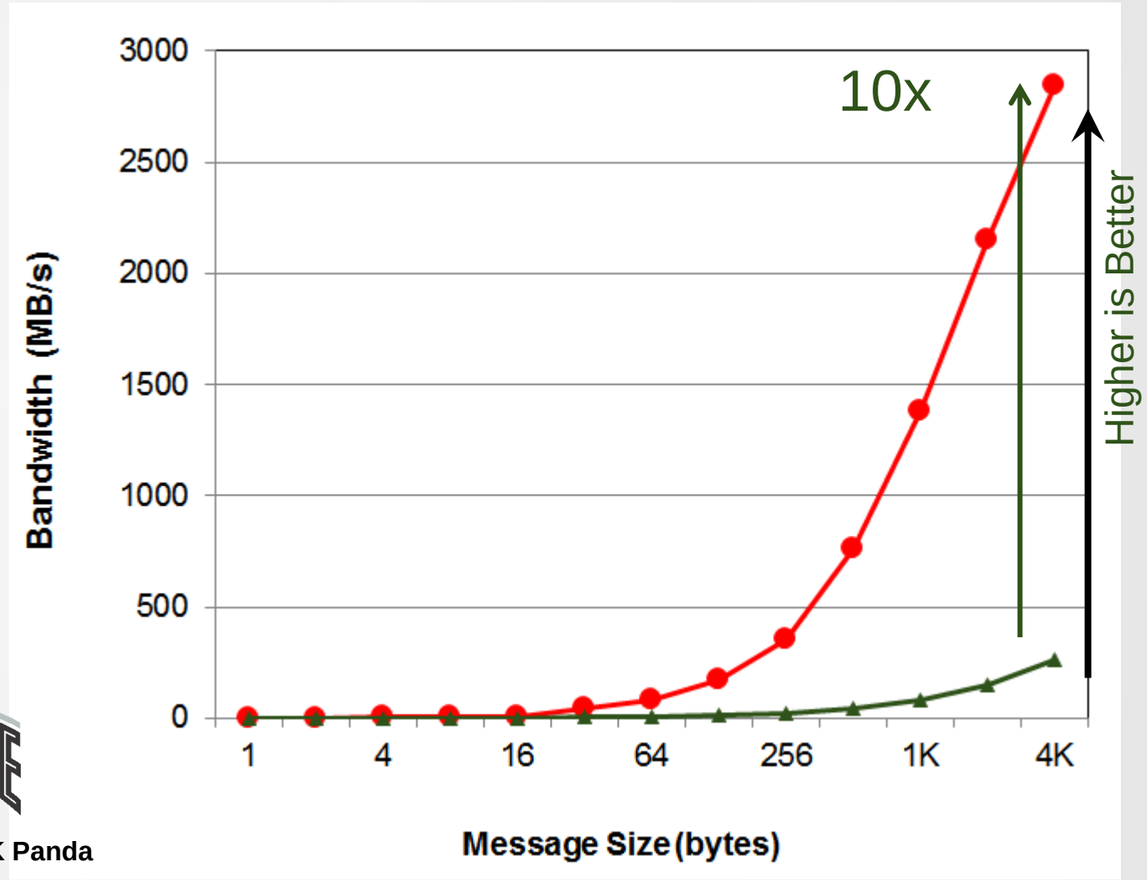
GPU間で直接的な通信を可能に

GPU Directの効果（GPU間通信）

GPU-GPU Internode Latency



GPU-GPU Internode Bandwidth



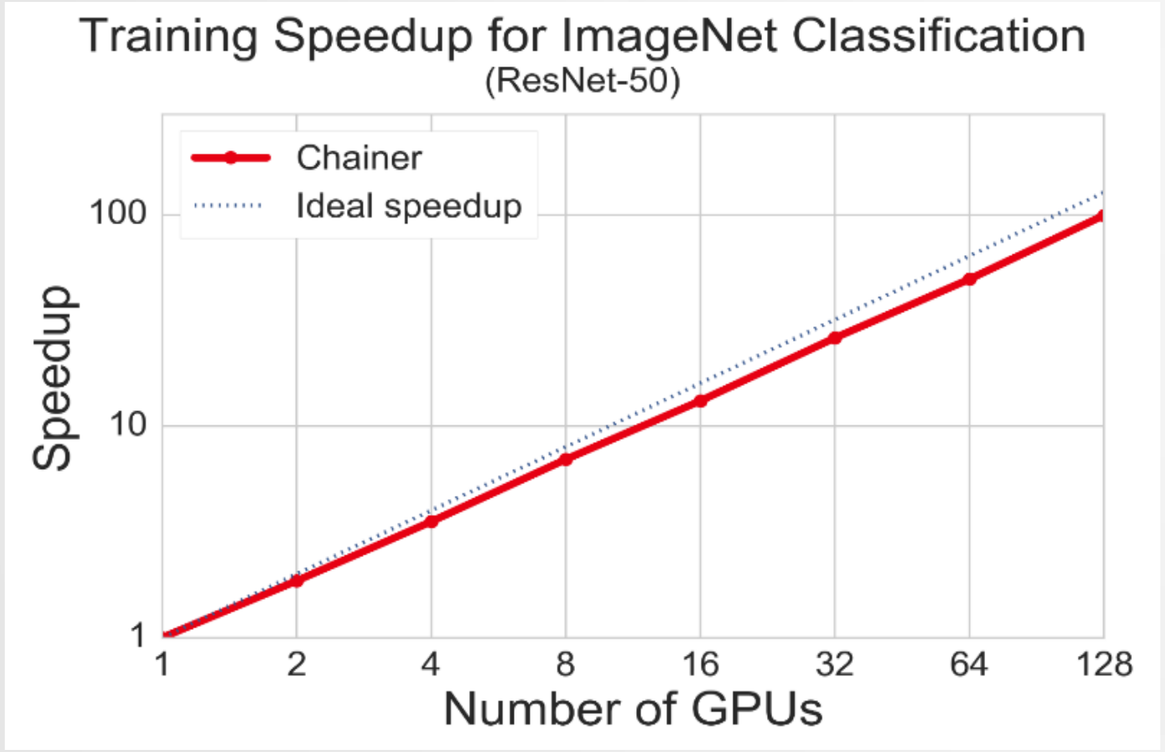
Source: Prof. DK Panda

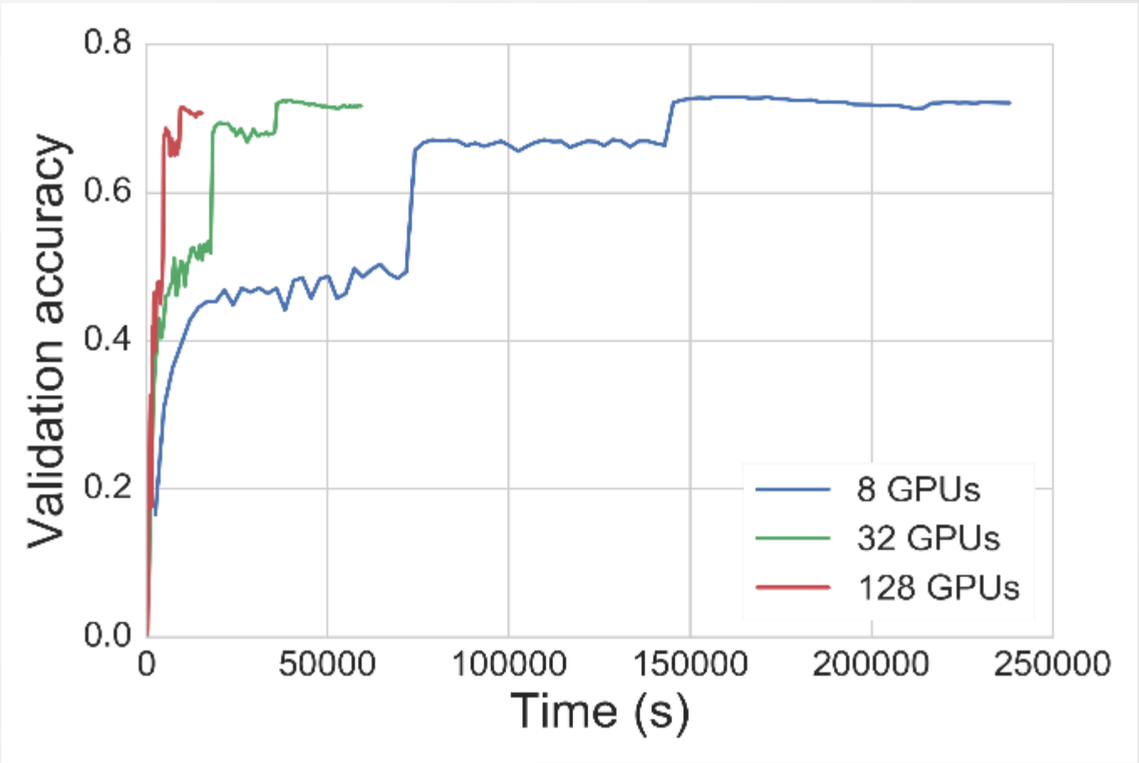
9.3X Better Latency

10X Better Throughput

GPU Directの効果（Chainer MN）

- ChainerMN depends on MPI for inter-node communication
- NVIDIA® NCCL library is then used for intra-node communication between GPUs
- Leveraging InfiniBand results in near linear performance
- Mellanox InfiniBand allows ChainerMN to achieve ~72% accuracy.





Source: http://chainer.org/general/2017/02/08/Performance-of-Distributed-Deep-Learning-Using-ChainerMN.html

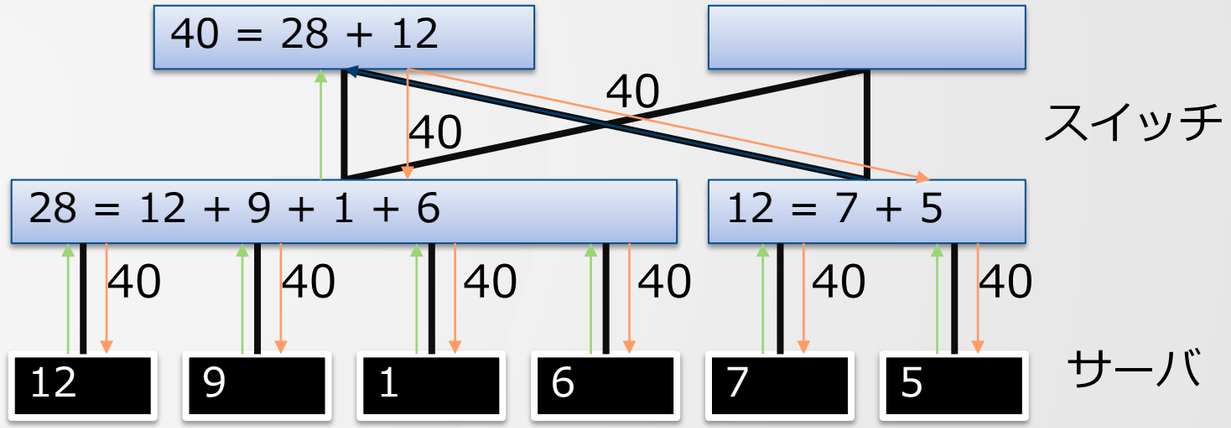
Mellanox SHARP

AI/DL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 bit

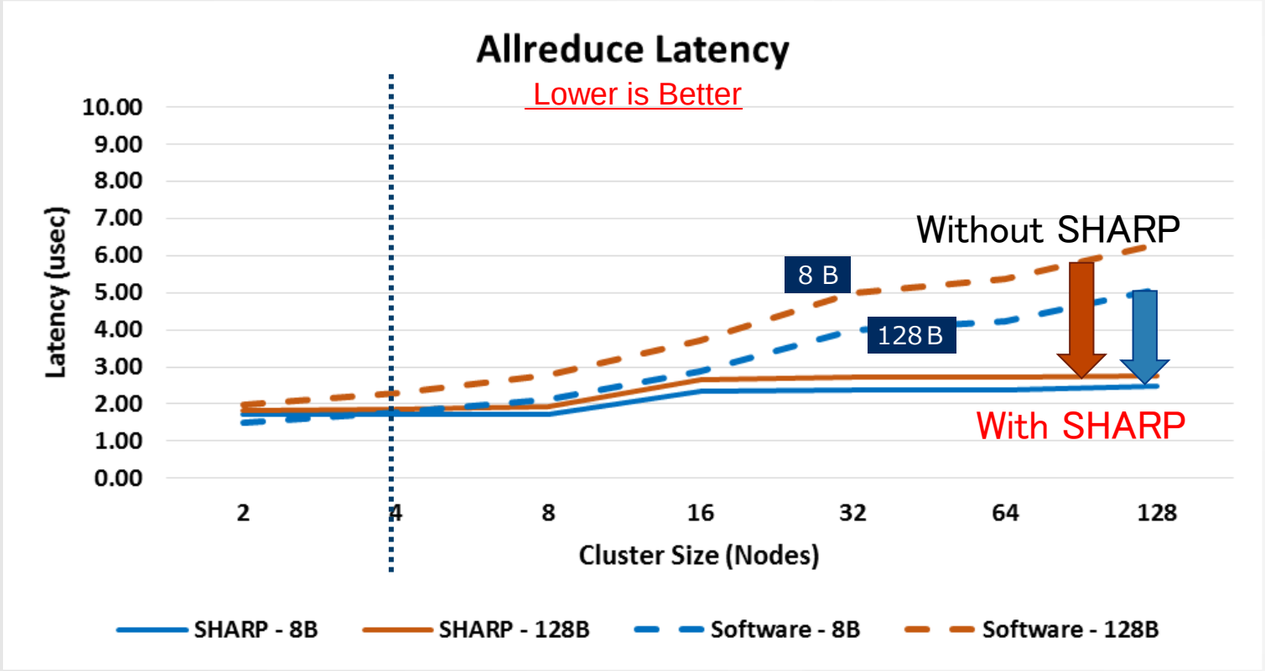
All Reduce のオフロードイメージ

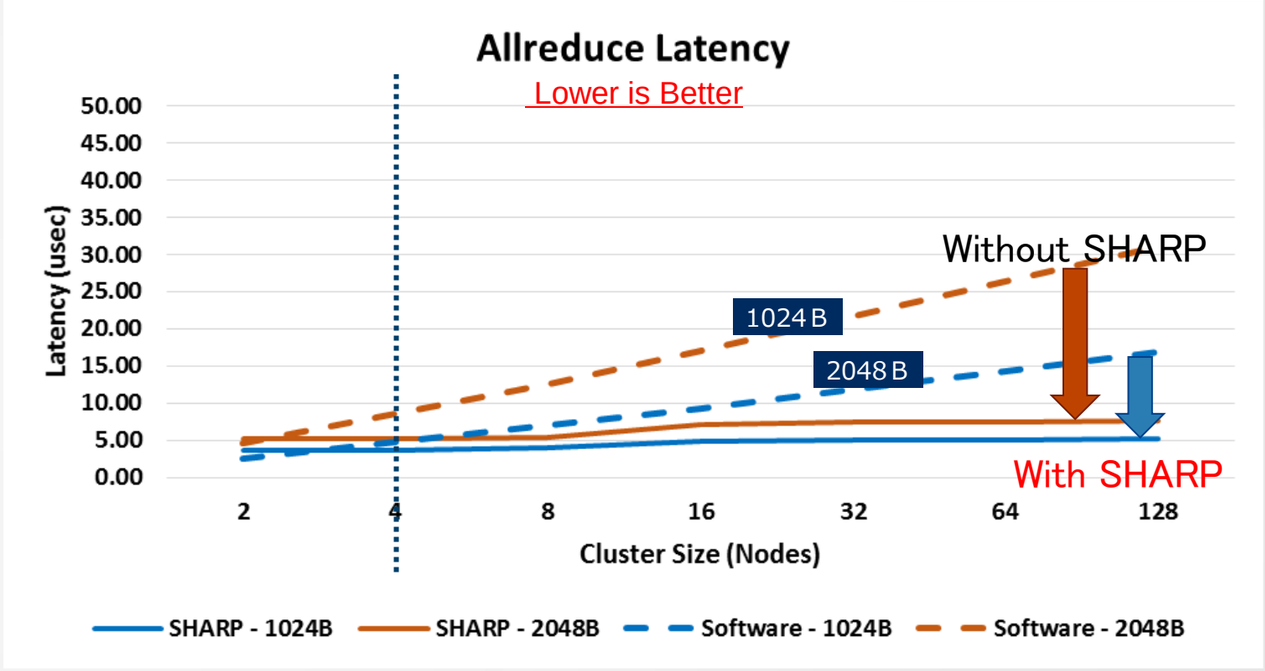


※Scalable Hierarchical Aggregation and Reduction Protocol

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
おおむね4nodes以上にクラスタにおいて、MPI遅延削減に効果

MellanoxでAI/DL環境の最大限の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

次世代製品のご案内

世界初のHDR（200G）
InfiniBandスイッチ

世界初のHDR（200G）
IB HCA

Thank You