HPC/AIにおけるビッグデータサイエンスを加速する NVIDIAのIOソリューションのご紹介

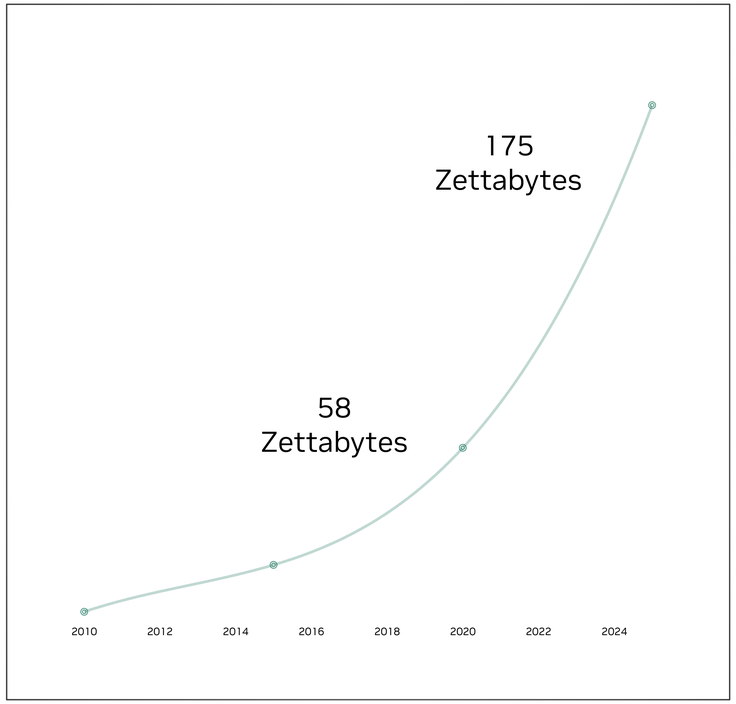
NVIDIA　Networking Product Marketing

HPC/AI Technical Computing Director

岩谷正樹

さらなるパフォーマンスとスケーラビリティを要求する次世代の AI

Exploding Data and Model Size

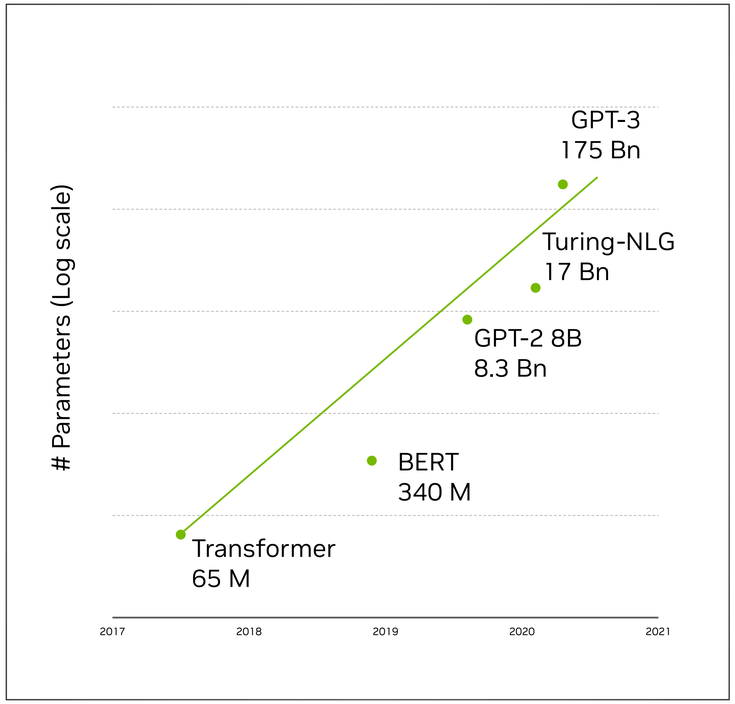


Big data growth

90% of the World's Data in last 2 Years

Growth in scientific data

Fueled by Accurate Sensors & Simulations



Exploding model size

Driving Superhuman Capabilities

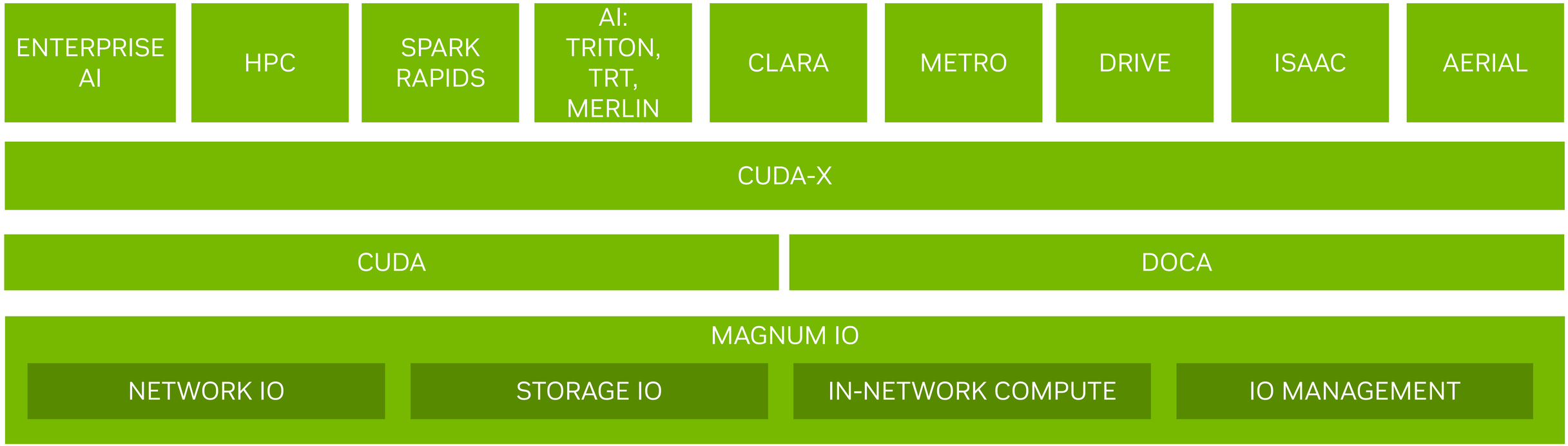
AIの急速に普及に伴い解析に使われるデータやﾊﾟﾗﾒﾀ加速度的に増加に伴い
計算資源だけでなくIOの高速化が必要

Source for Big Data Growth chart: IDC – The Digitization of the World (May 2020)
storagenewsletter.com: 64.2ZB of Data Created or Replicated in 2020

NVIDIA

NVIDIA Magnum IO: データセンター IO を加速するNVIDIAのアーキテクチャ

AIおよびHPCコンピューティングを加速するためのテクノロジー、SDK、ライブラリ



データサイエンスを加速するMagnum IO 技術

IOアクセラレーション:データの移動、アクセス、管理

Network IO

NCCL

NVSHMEM

ASAP²

GPUDirect RDMA

HPC-X, UCX

DPDK

MOFED

Storage IO

GPUDirect Storage

NVMe SNAP

In-Network Compute

SHARP

Hardware Tag Matching

IO Management

Ethernet NetQ

UFM

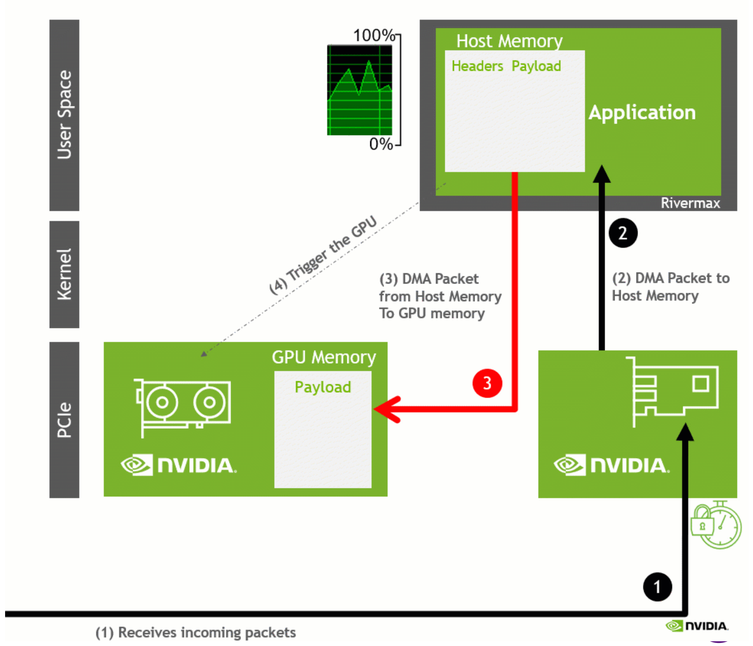
NVIDIA イノベーション | ピークパフォーマンス | どこでもRDMA | 柔軟な抽象化

GPUDirect™ RDMA による 10 倍のパフォーマンス向上

Magnum IOアーキテクチャは、DMAエンジンを活用して不要なコピーを排除し、IOを最適化

No GPUDirect

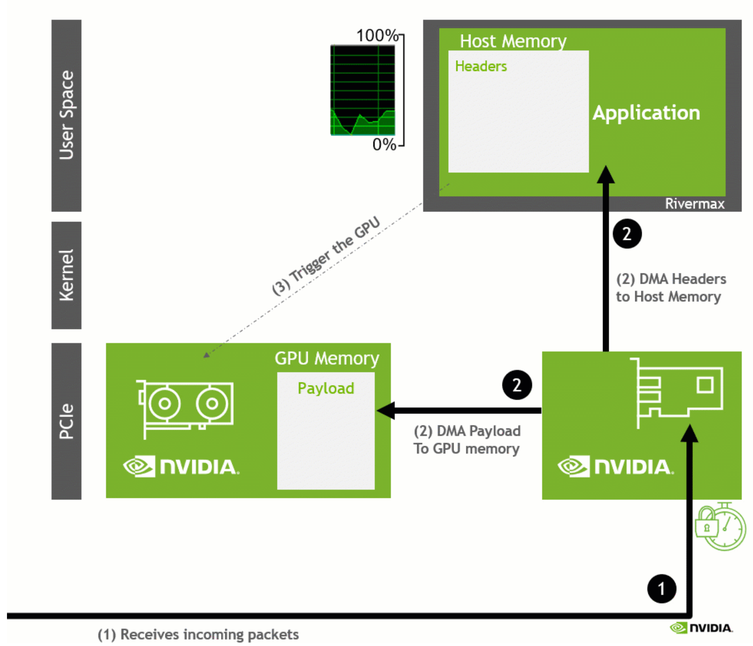
Network handled by CPU and CPU-Memory



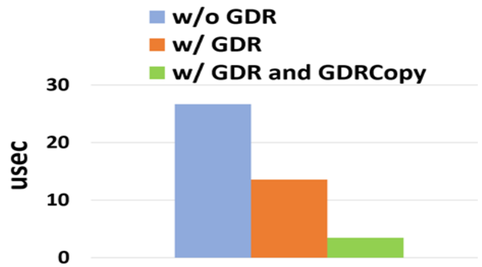
2	Full copy operations	0
2	PCIe transactions	1
↓	GPU utilization	↑
↑	CPU usage	↓
↑	Latency	↓

GPUDirect

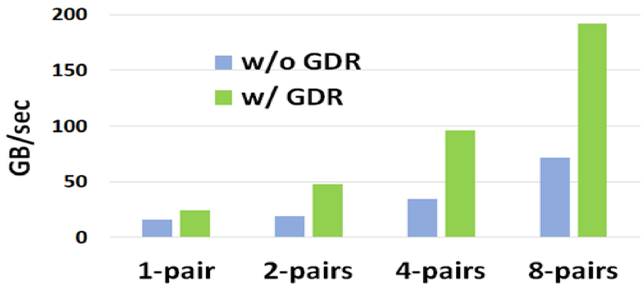
Network goes directly to GPU memory



***8x Latency** 改善.*

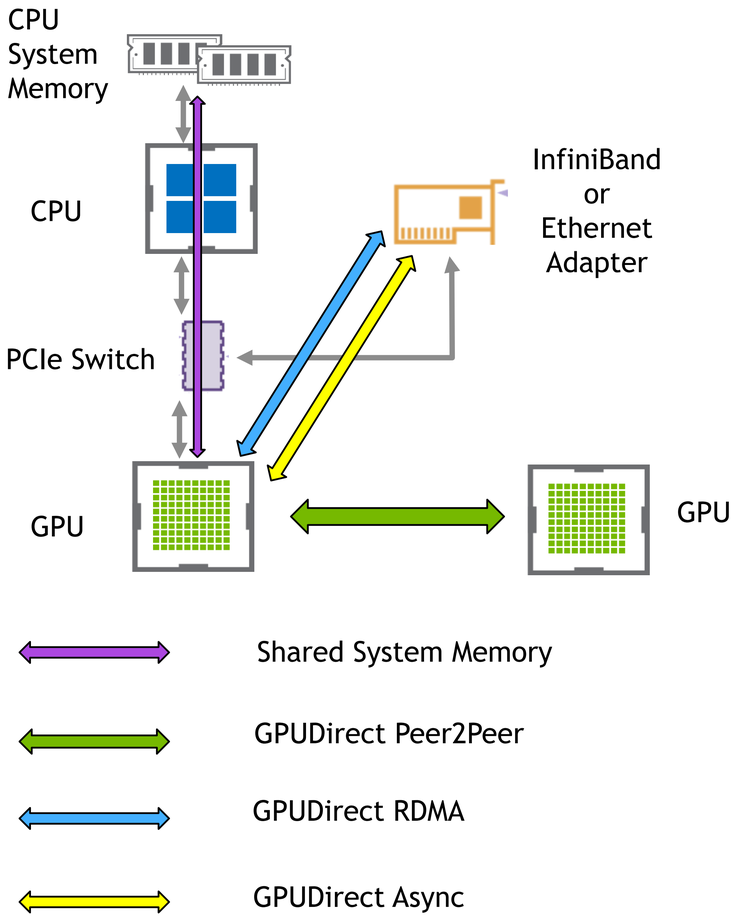


*最大**2.6**倍のノード間**MPI**帯域幅*



Magnum IO GPUDirect Async

HPCシステムのネットワークパフォーマンスの向上



NVIDIA GPUDirect

- Magnum IOによりデータ移動とGPUアクセスを強化
- ネットワークアダプタやストレージドライブからGPUメモリへの直接の読み書きが可能.
- 不要なメモリコピーを削減
- CPUのオーバーヘッドや遅延を低減し、パフォーマンスを大幅に向上.
- 共有システムメモリによりCPUとGPUの通信が可能だがマルチGPUによる加速コンピューティングはCPUをバイパスした方が高速.
- GPUDirect Peer2Peer は、PCIe または NVLink を介して GPU と GPU の直接通信を実現
- GPUDirect RDMAにより、周辺機器がGPUメモリに直接アクセスが可能.

NVIDIA GPUDirect Async

- GPUDirect AsyncではGPUはCPUプロキシを使用せずに通信開始することが可能.
- GPUDirect Async-Kernel Initiated（GDA-KI）では、GPUがNICに直接通信要求を出し、CPUを介さずにネットワーク通信を行うことが可能。
- GDA-KIは、より小さなメッセージサイズでも高いスループットレートを維持することが可能
- スケーリングが必要でかつ、ワークロードが大きくなるとメッセージサイズが縮小しより沢山のGPUを使う傾向があるアプリケーションでは、GDA-KIによりパフォーマンスが向上

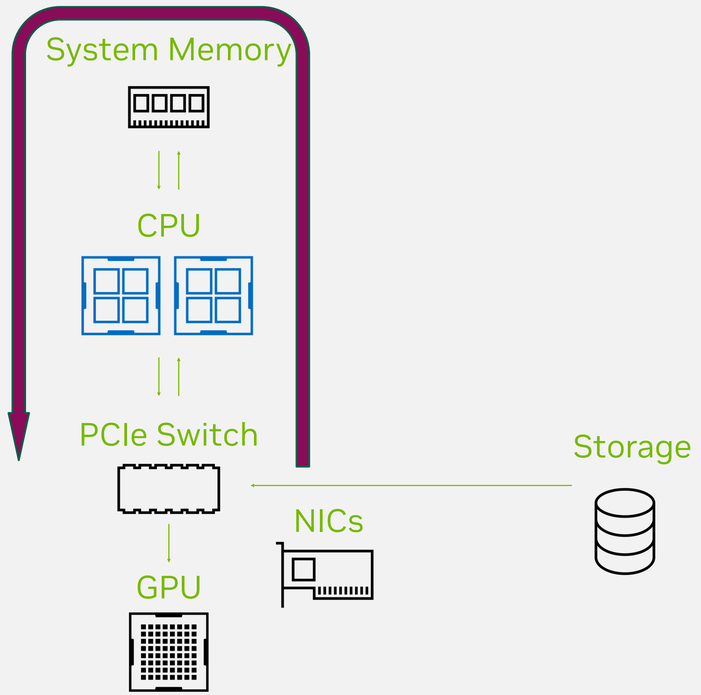
SM – Streaming Multiprocessor

Improving Network Performance of HPC Systems Using NVIDIA Magnum IO NVSHMEM and GPUDirect Async

GPUDirect Storage

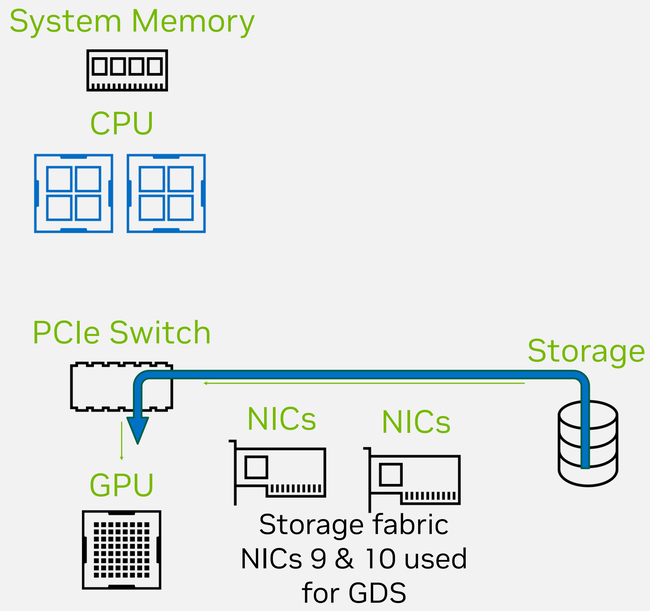
ローカルまたはリモートで保存されたデータからGPUへのデータ転送の高速化

Without GPUDirect Storage



Low Bandwidth | High Latency | Limited Capacity

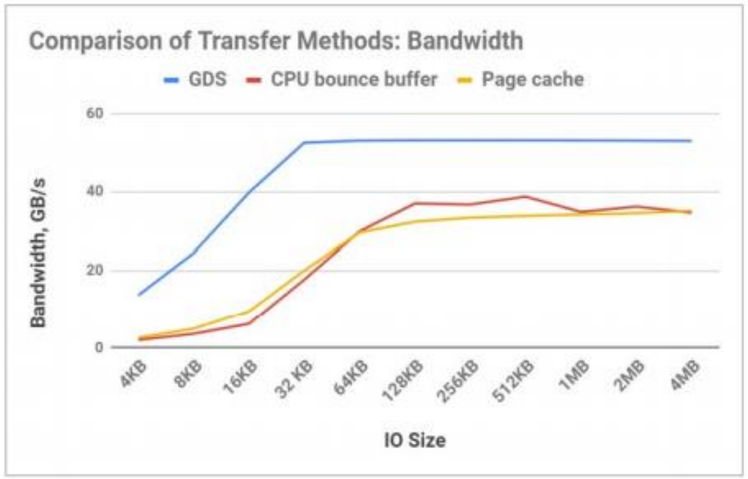
Without GPUDirect Storage



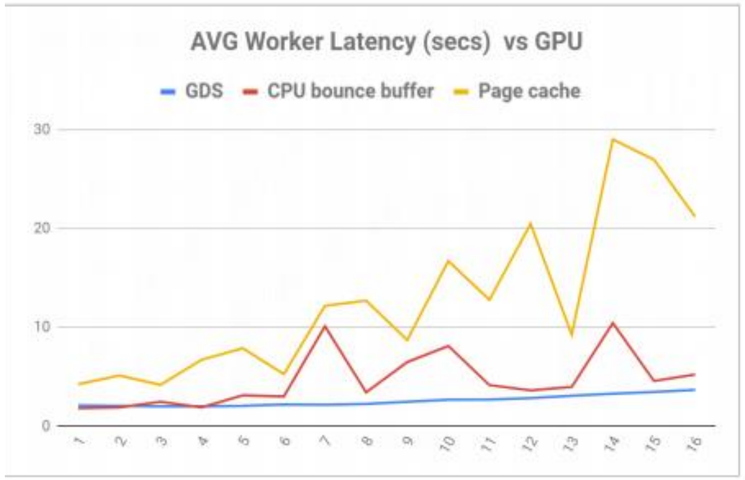
Higher Bandwidth | Lower Latency
O(PB) capacity | CUDA programming model

GPUDirect Storageの効果

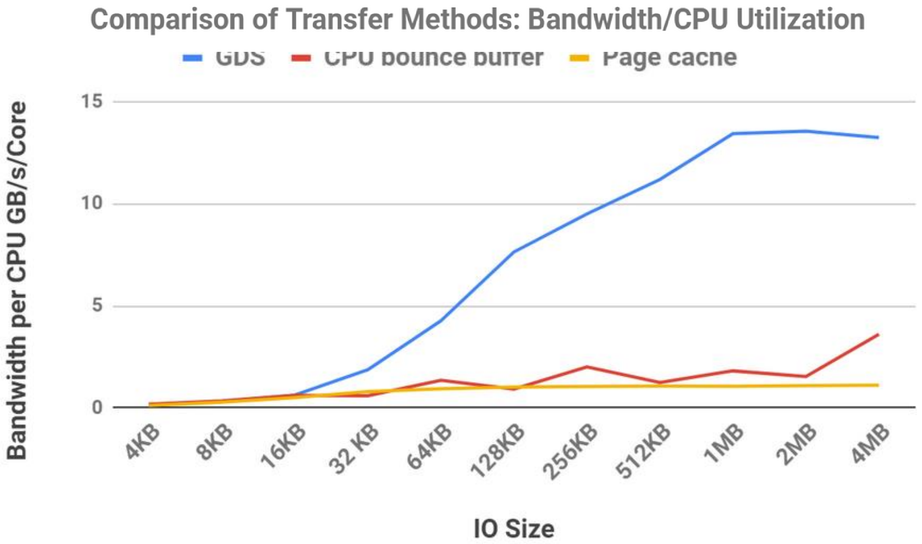
ストレージとGPUメモリ間のダイレクトパス



より高い帯域幅：ダイレクトパスにより
スループットを向上



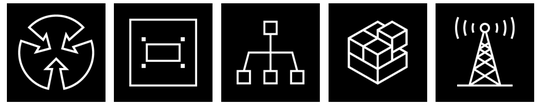
低レイテンシー :GPUDirect Storage は
ほぼ横ばい非常に低いのレイテンシー
での運用が可能



転送サイズが大きくなってもCPUの負担が少ない

BlueField Data Processing Unit

ソフトウェア定義ネットワーク



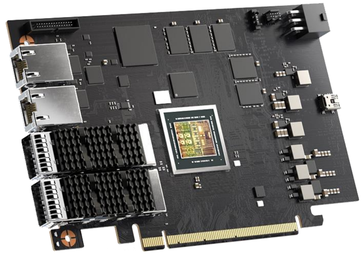
ソフトウェア定義のセキュリティ

ソフトウェア定義ストレージ

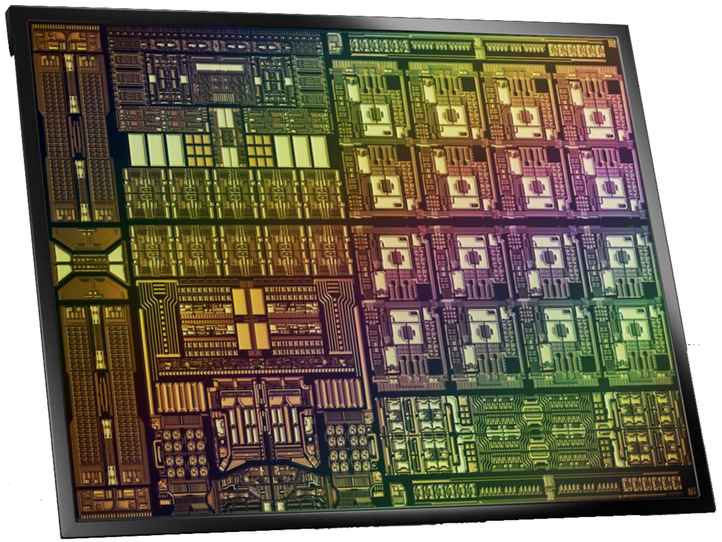
Infrastructure Services

Data Center on a Chip

- 16 Arm 64-Bit Cores
- 16 Core / 256 Threads Datapath Accelerator
- ConnectX InfiniBand / Ethernet
- DDR memory interface
- PCIe switch



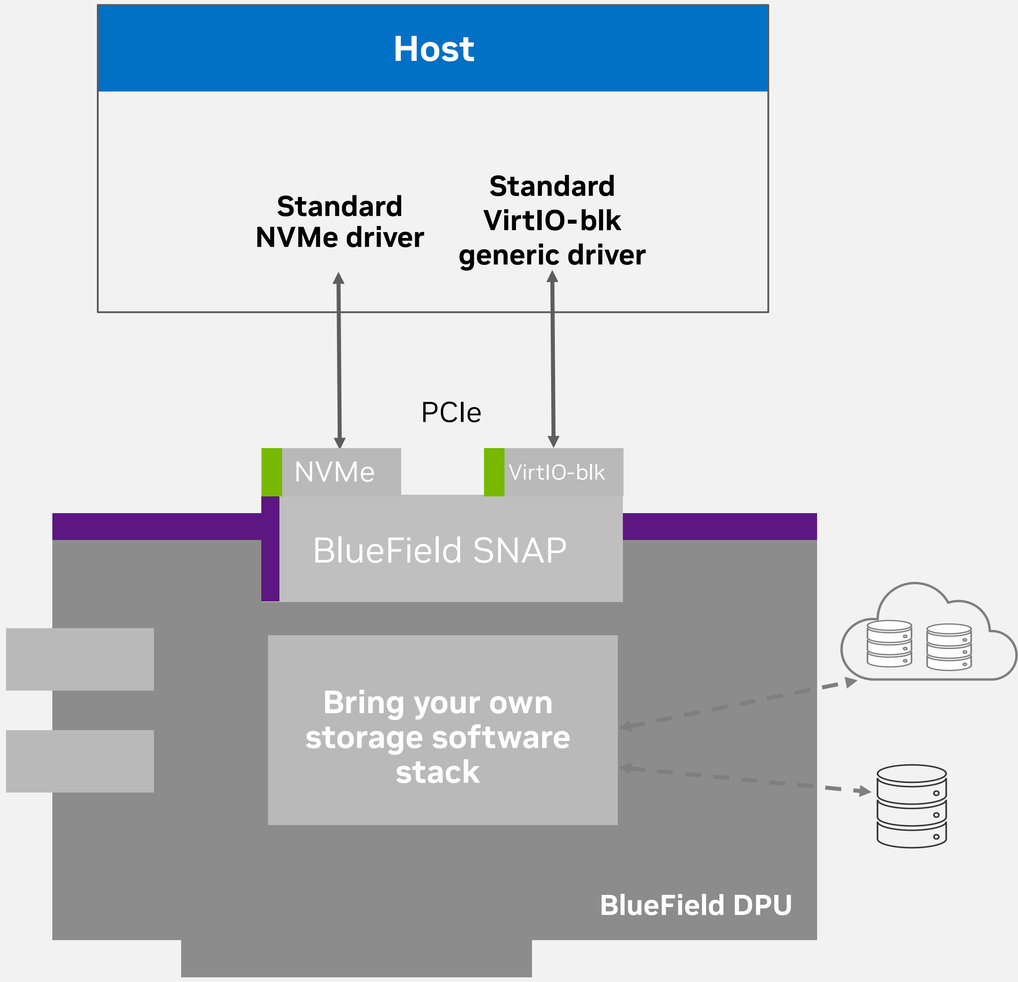
BlueField Infrastructure Compute Platform



BlueField SNAP

ソフトウェア定義のネットワークアクセラレーション処理

- BlueField SNAPにより、NVMeストレージのハードウェアアクセラレーションによる仮想化が可能
- ネットワークストレージを、ローカルのNVMe SSDとして表示
- BlueField SNAPは、特別なドライバを必要とせず、PCIe上のNVMeまたはVirtIO-blkストレージインターフェースを公開
- ベアメタルおよびHypervisor/VMsのユースケースに対応
- VMに数百のインターフェイスを提供するための拡張性
- ライブマイグレーションに対応.
- ローカルストレージのパフォーマンスメリットとネットワークストレージの効率性.



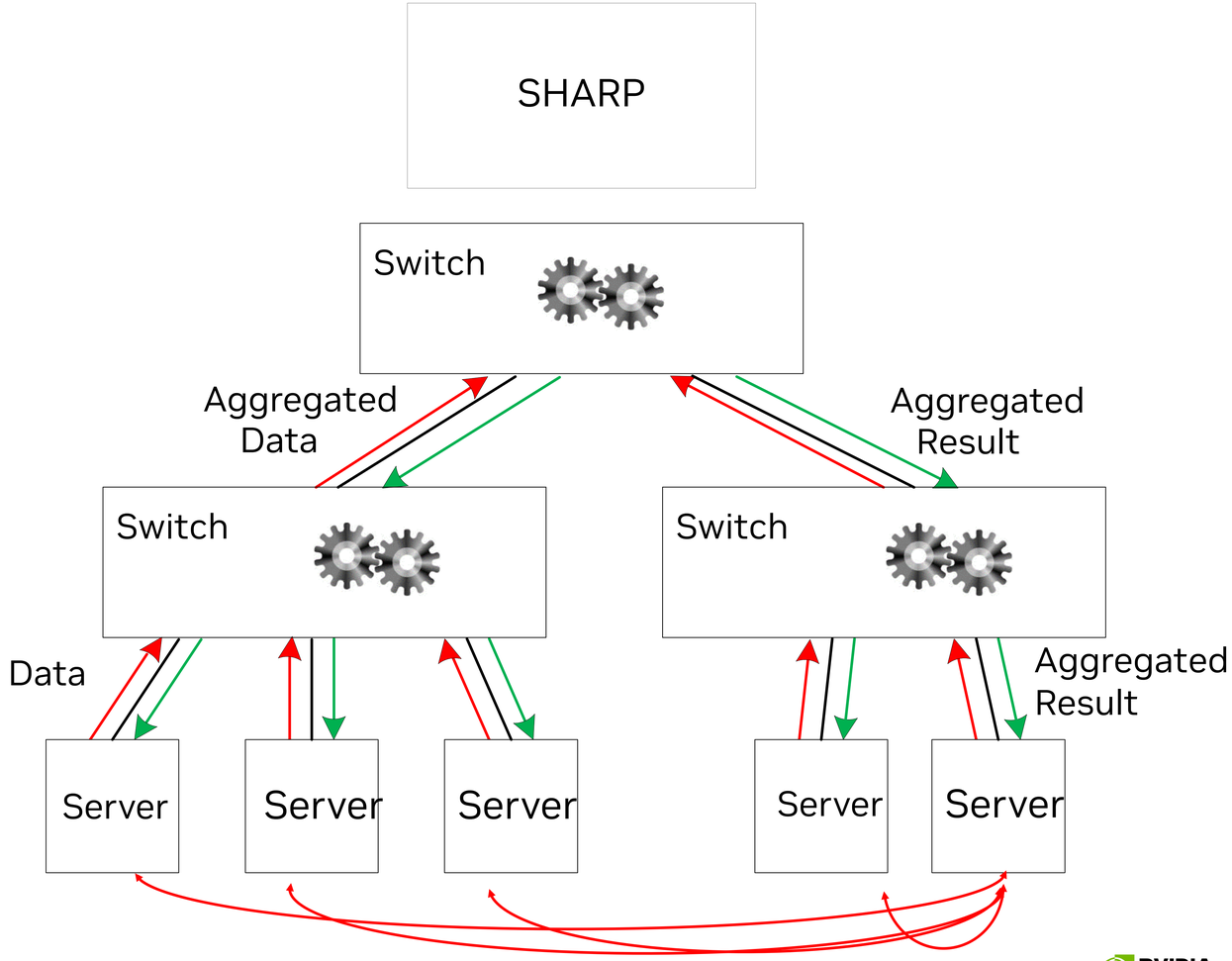
第3世代SHARP V3

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

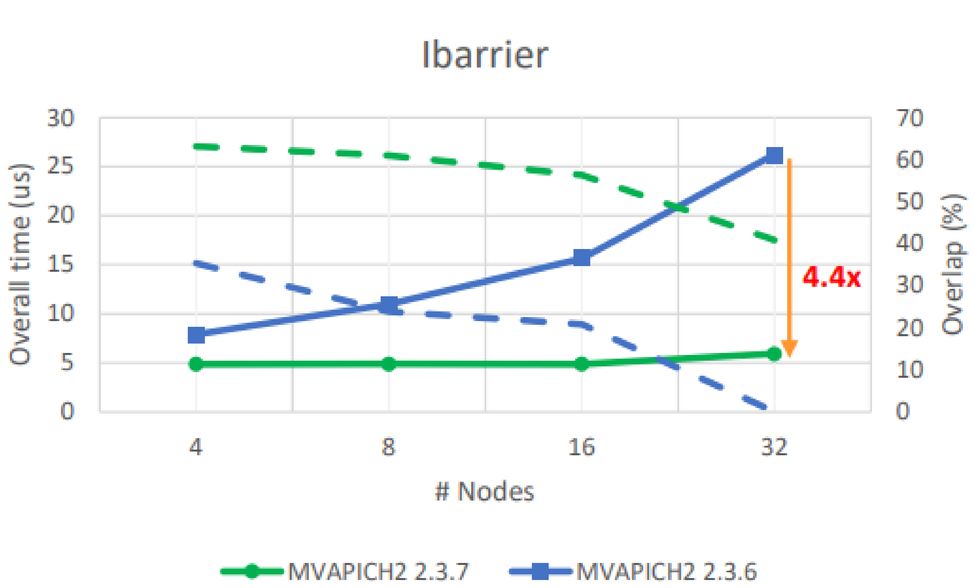
第３世代 SHARP V3エンハンスポイント

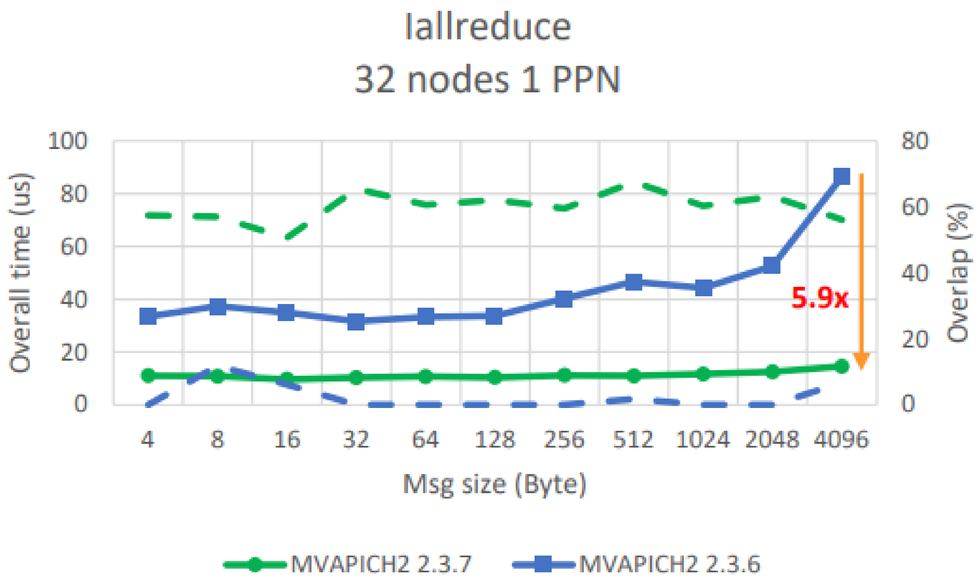
- 論理ツリー数が２から 64 になることで同時に大量の AI 集合演算が可能

- 従来のSHARP V2の機能はそのまま利用可能
  - 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（8は無)：8/16/32/64 ビット
    - Bfloat16



MVAPICH を用いたSHARP 性能

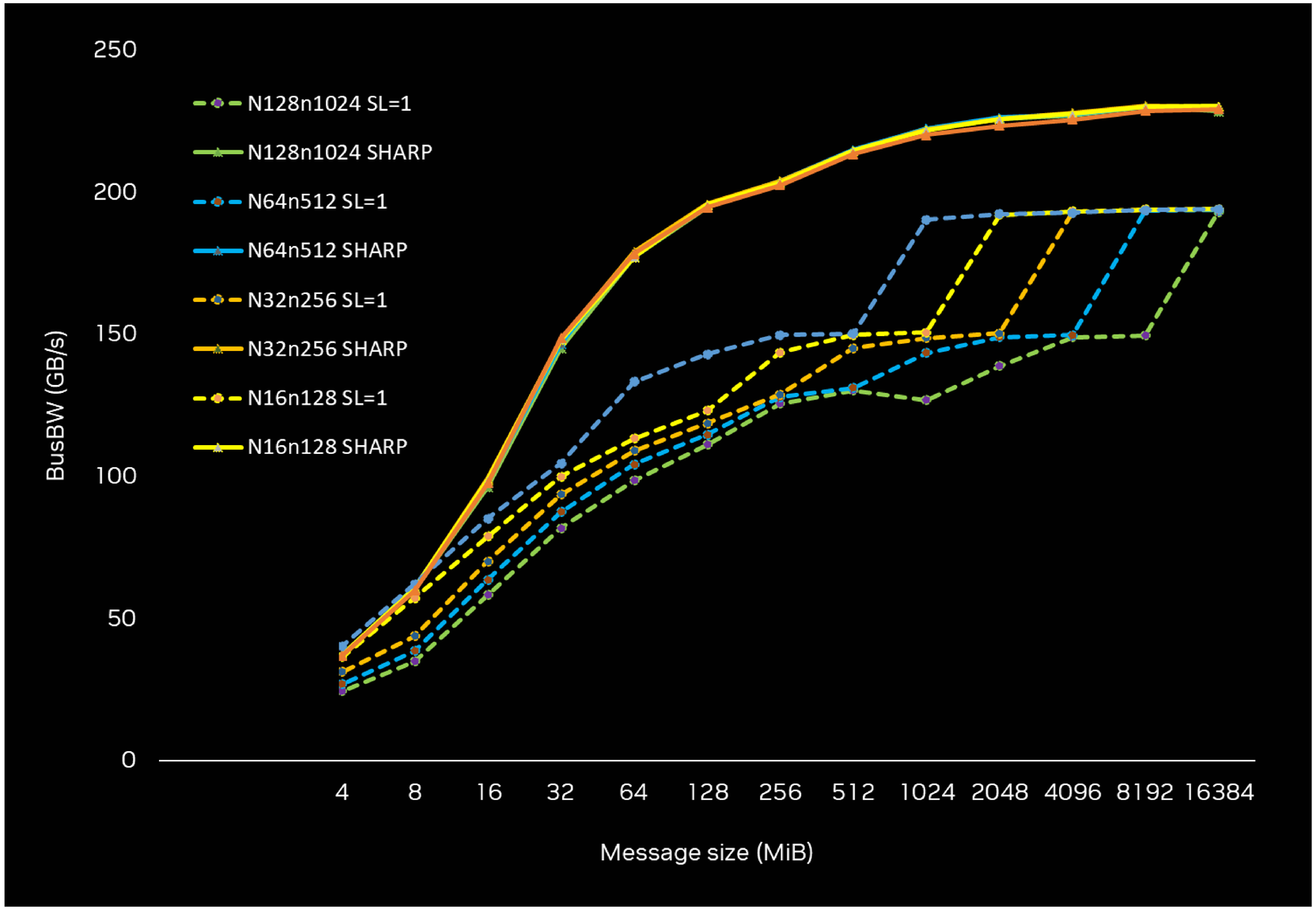




- With SHARP:
  - 全体の時間から見てフラットなスケーリング
  - 計算と通信の重なりが多い

Platform: Dual-socket Intel(R) Xeon(R) Platinum 8280 CPU @ 2.70GHz nodes equipped with Mellanox InfiniBand, HDR-100 Interconnect

NCCL All Reduce Performance with NVIDIA InfiniBand



InfiniBand SHARPは、最大100%の優位性を持つ大規模な帯域幅を維持

Magnum IO Resources

Storage IO
GPU Direct Storage

IO Management
NetQ
UFM
Nsight Cluster Debug and Profiling

Network IO
GPU Direct RDMA, Peer-to-Peer in CUDA Toolkit (v6-11.5.1)
NCCL, HPC-X UCX, MPI in HPC SDK
NCCL
ASAP^2

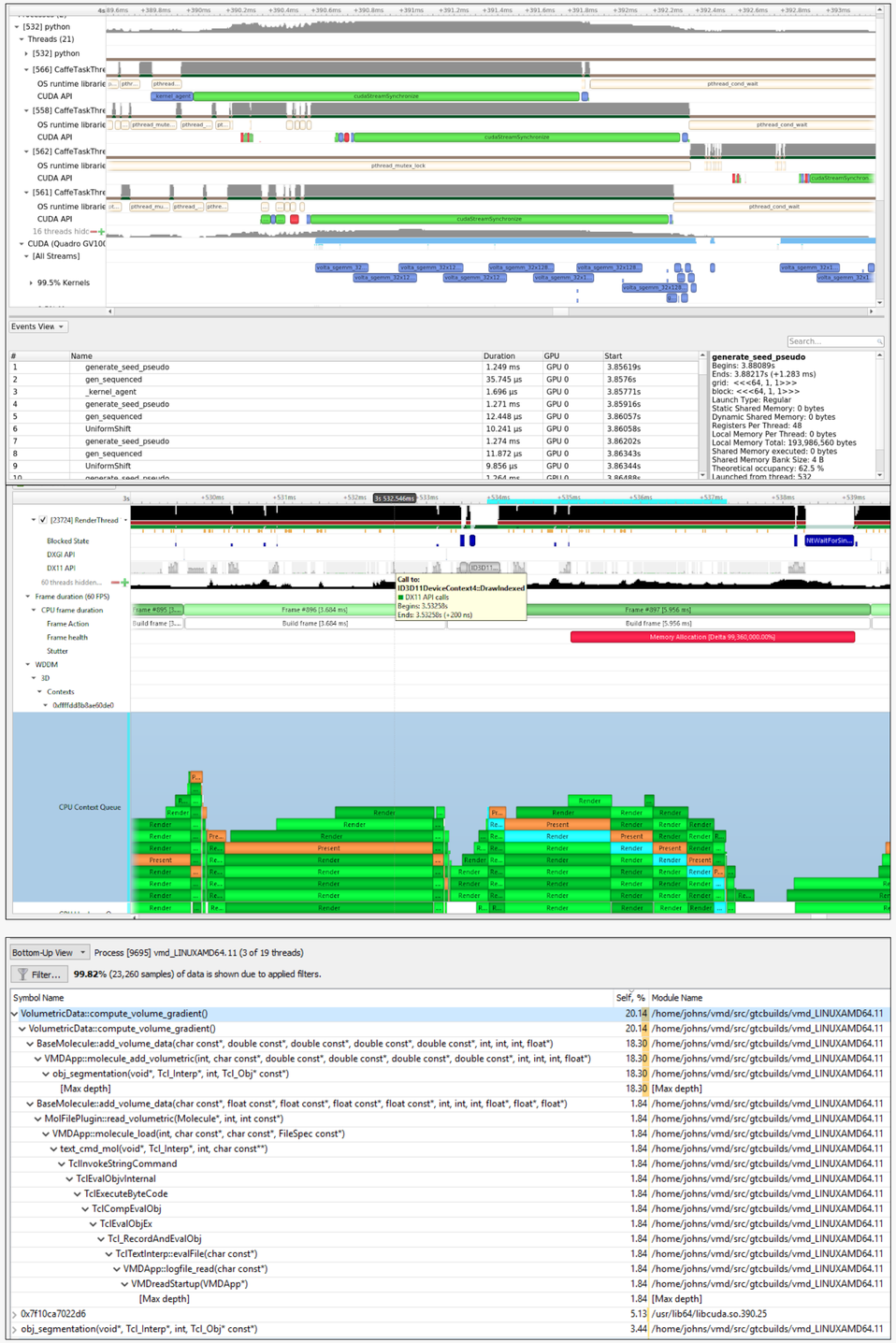
NVIDIA Deep Learning Institute Courses
The Fundamentals of RDMA Programming
Working with Mellanox OFED in InfiniBand Environments

In-Network Computing
SHARP

Magnum IO Management

NVIDIA Nsights Graphics

- マルチノードのプロファイリングとデバッグ
- Nsightは、以下の方法で生成されたトラフィックの可視性を向上
  - GPUDirect RDMA and GPUDirect Storage
- NsightはMagnum IOと組み合わせることによりより利便性を向上
  （可視化による管理の向上）



UFM Platforms 製品ラインナップ

監視・モニトリング

UFM Telemetry
Real-Time Monitoring

監視・モニトリング
オーケストレーション
ネットワーク設定変更

UFM Enterprise
Management, Monitoring & Orchestration

(UFM Enterprise includes UFM Telemetry)

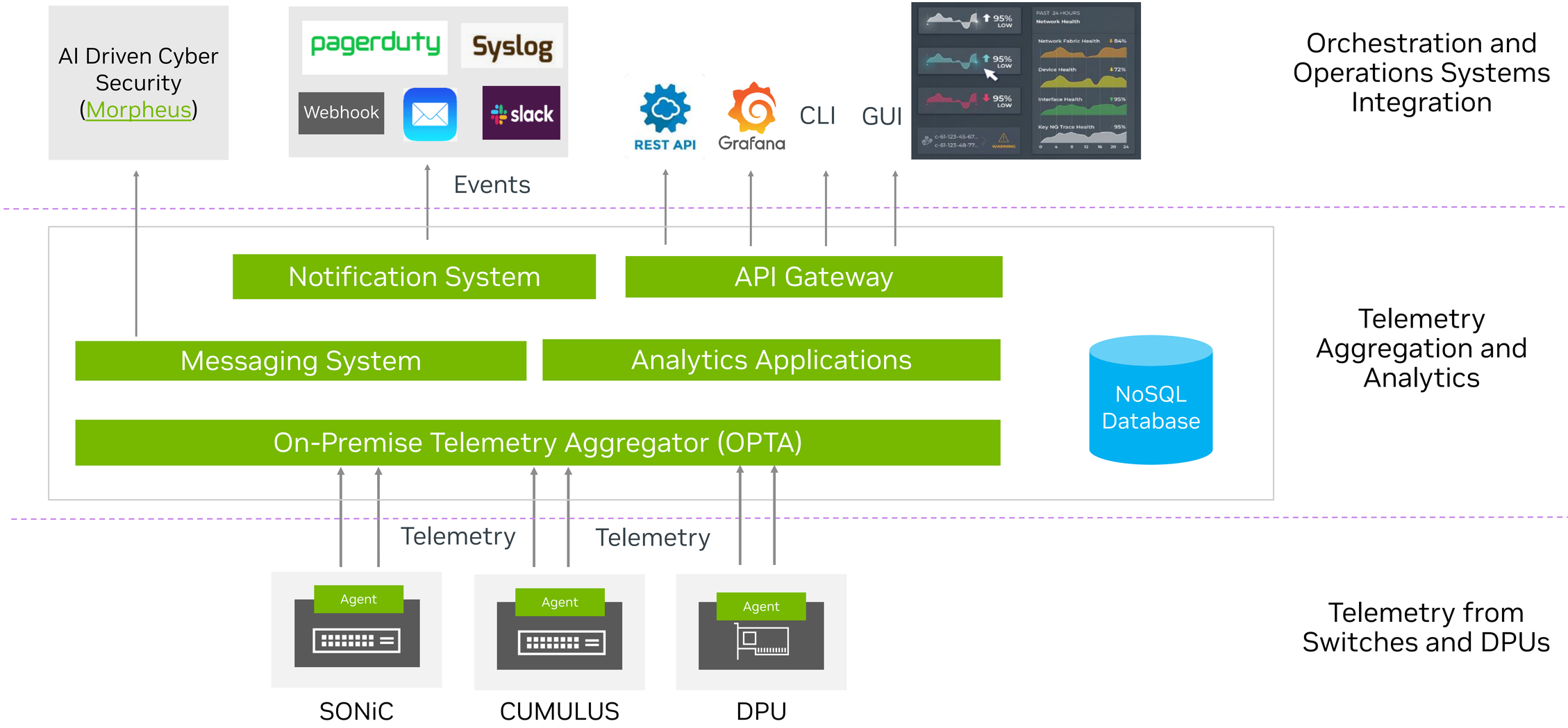
監視・モニトリング
オーケストレーション
セキュリティ設定変更
AIによるネット解析

UFM Cyber-AI
Cyber Intelligence and Analytics

(UFM Cyber-AI includes UFM Enterprise)

NVIDIA NetQ – Transforming Network Operations

スケールアウトアーキテクチャによるオンプレミスとSaaSの展開オプション




NVIDIA