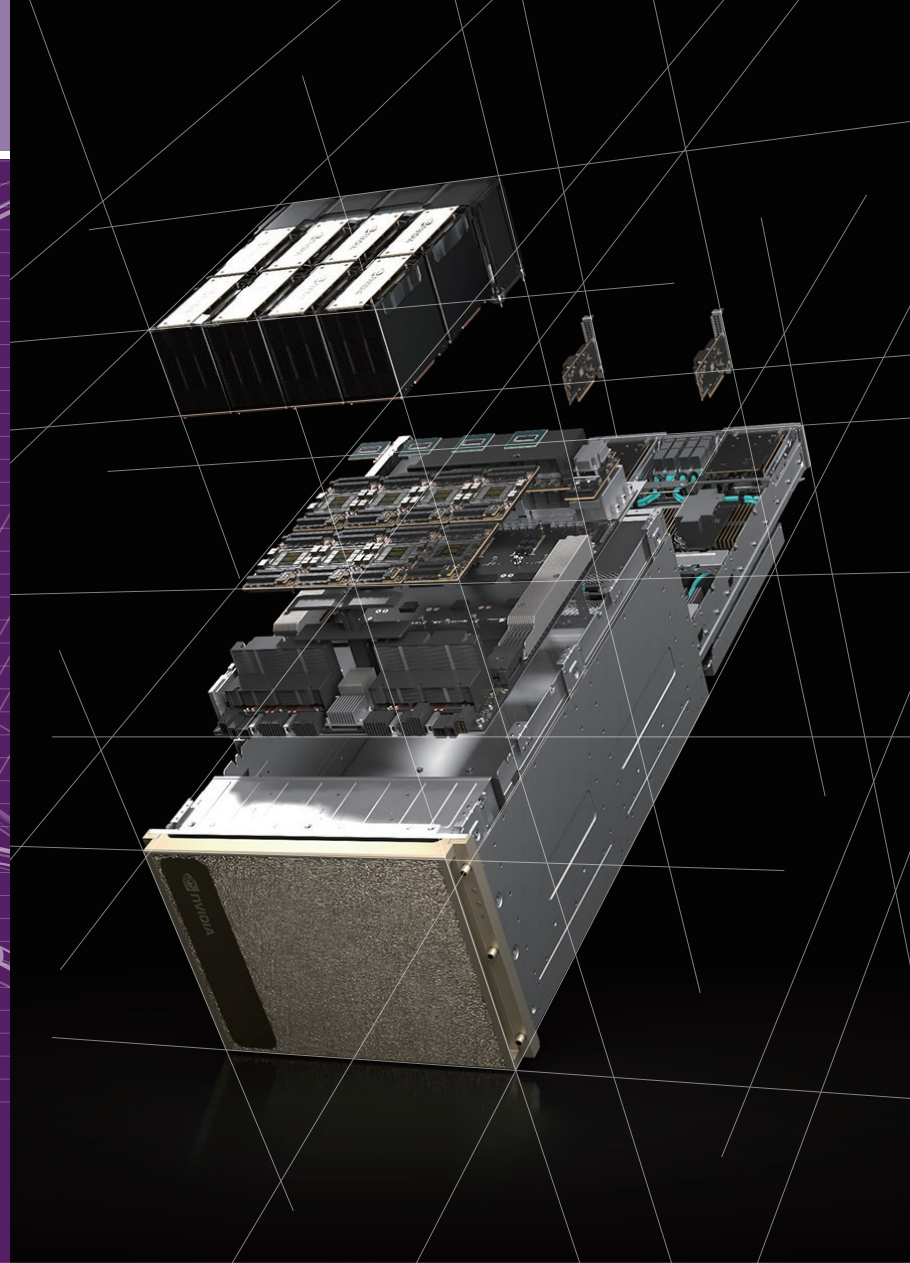
ネットワーク視点の

NVIDIA DGX™ H100 大解剖

GPUを最大限活かすNVIDIA ConnectX®-7の役割



MACNICA

目次

ネットワーク視点の

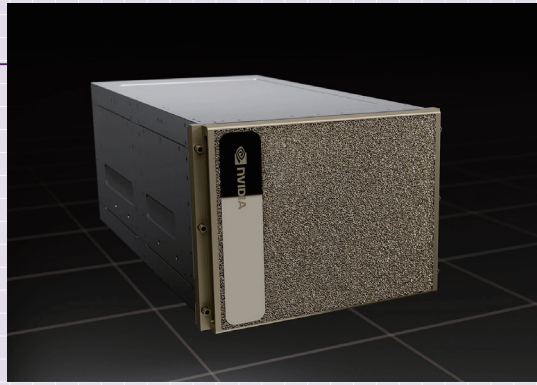
NVIDIA DGX™ H100 大解剖

GPUを最大限活かすNVIDIA ConnectX®-7の役割

NVIDIA DGX H100 概要

大規模で複雑なAIジョブの限界を突破

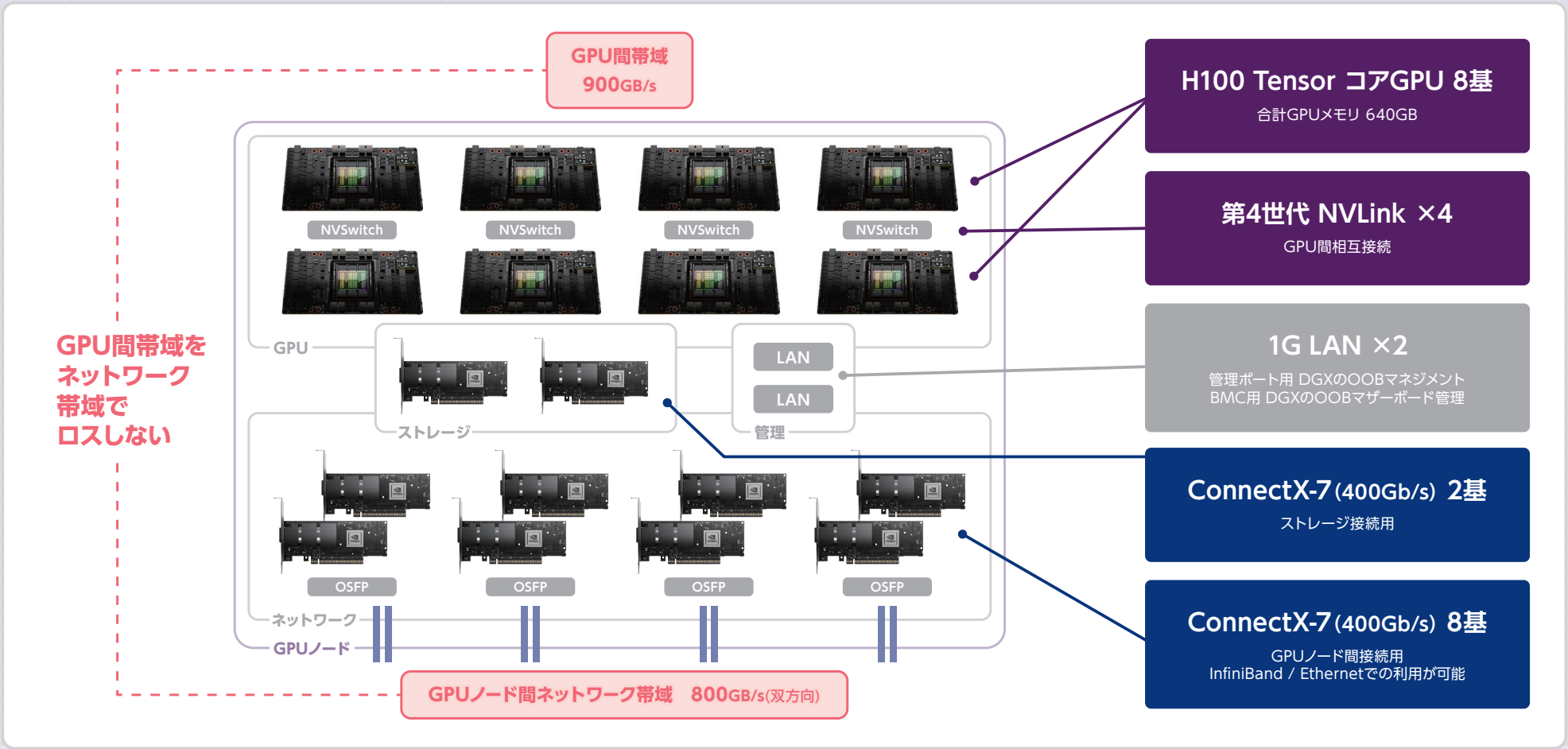
NVIDIA DGX™ H100は、第4世代のNVIDIA DGXシステムで、NVIDIA H100 Tensor コア GPUとNVLink Switch System、NVIDIA ConnectX®-7を組み合わせた世界最速のAIシステムです。次世代アーキテクチャーを採用することで前世代と比較して6倍のパフォーマンスと2倍の高速ネットワークを実現し、企業のAIインフラの中核を担うように設計されています。



主な仕様と性能

▶ 仕 様	**H100 Tensor コアGPU 8基** 合計GPUメモリ 640GB	**第4世代 NVLink** GPU間相互接続	**ConnectX-7 10基** GPUノード間接続用+ストレージ接続用
▶ 性 能	**性能6倍(32PFLOPS)** FP8 精度で前世代の 6倍以上	**GPU間帯域 900GB/s** 双方向帯域幅	**ピークネットワーク帯域 1TB/s** 双方向帯域幅

DGX H100 構成要素

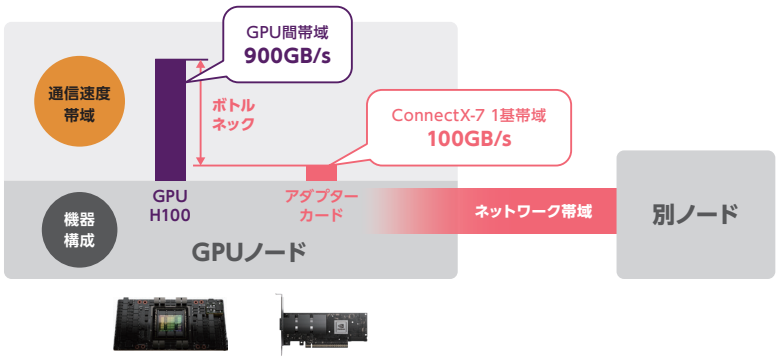


ConnectX-7が8基 搭載される理由

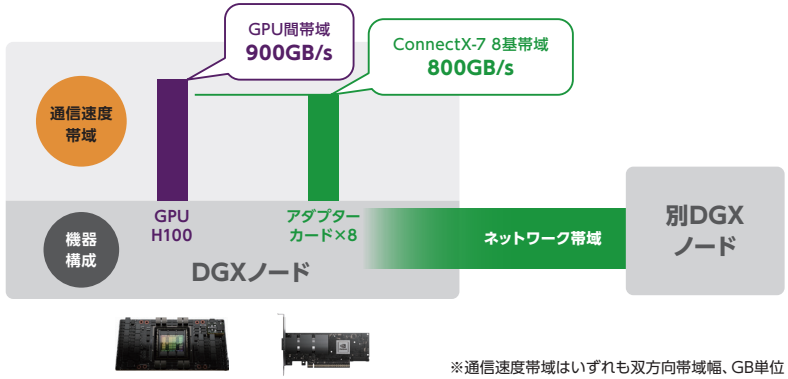
ネットワーク帯域を確保し、ノード間のGPU演算速度を維持

GPUノード内(筐体内)の通信は高速GPU同士での処理となり、転送速度は理論上900GB/sとなります。一般的な構成でネットワークインターフェースが1ポートの場合、別ノード間通信でネットワークがボトルネックとなり通信速度を損失します。一方、DGX H100では ConnectX-7 8基により 800GB/sの帯域を確保することができ、ネットワークにおけるボトルネックを解消します。GPUの演算速度を落とさずネットワーク構築が可能なDGX H100は、クラスター性能が非常に重視される大規模なAI分散処理に、最適なシステムといえます。

一般的な構成（ネットワークインターフェース1ポートの場合）



DGX H100構成（ネットワークインターフェース8ポートの場合）



※通信速度帯域はいずれも双方向帯域幅、GB単位

ConnectX-7は帯域確保だけじゃない

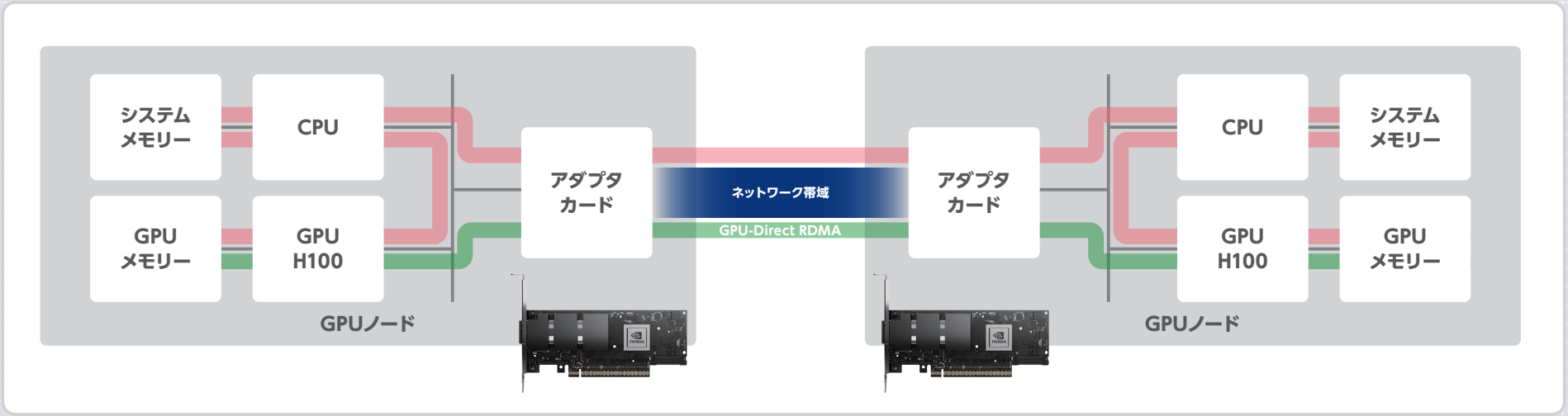
GPU-Direct RDMAがノード間通信高速化に貢献

GPU-Direct RDMAとは?

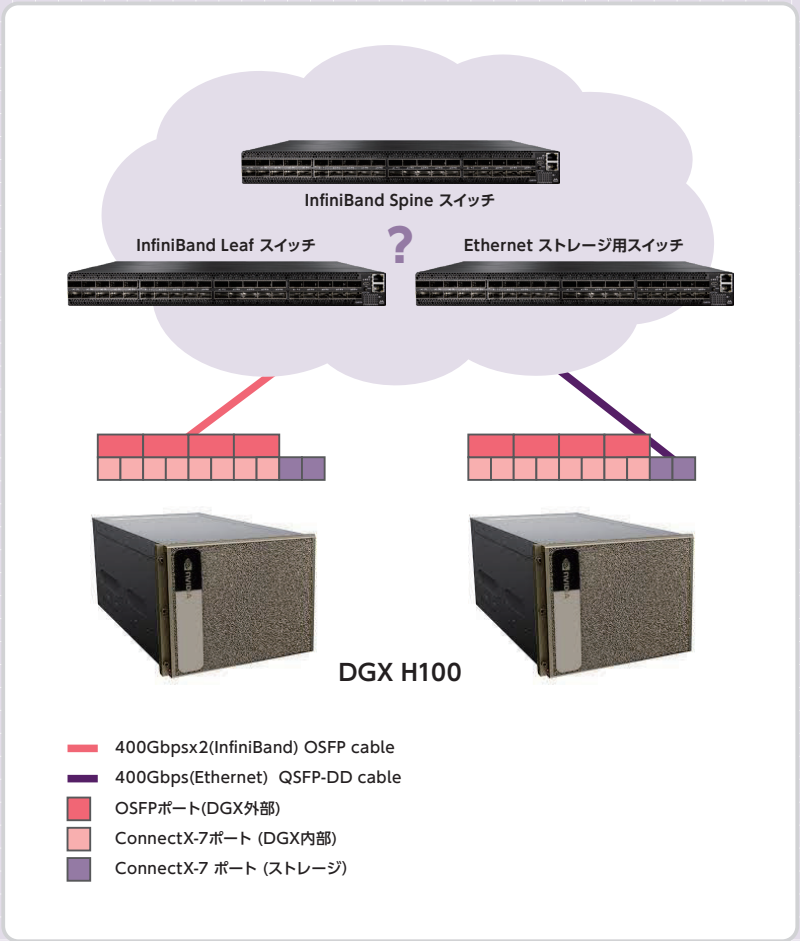
GPUノード間通信でホストCPUを介さずGPUメモリー同士が直接データ転送をする技術です。
ConnectX-7のRDMA機能とCUDA TOOL KITによりGPU-Direct RDMAを実現し、ノードを跨いだ
GPUメモリー間データ転送をハードウェアへオフロードが可能となります。

効果

- レイテンシーの大幅な改善
- スループット性能の向上
- CPU使用率の低減



DGX H100のシステム構成を考える



複数ノードのDGX H100を構成するときに考慮すべきポイント

■コンピューターネットワーク

・InfiniBandのLeaf – Spineのノンブロッキング構成を組む
…多くのHPCシステムで採用されるアーキテクチャー
…AIの多岐に渡るワークロードに対応可能、DGX H100のような
ハイパフォーマンスコンピューティングなサーバーに最も適した構成

・スイッチはInfiniBandを使用
…SHARP・SHIELDと呼ばれるIn-Networkingコンピューティングをサポート
…高性能なネットワークファブリックを構成可能

■ストレージネットワーク

・DGX H100内ConnectX-7をEthernetスイッチに接続し、冗長構成を組む
…DGX H100はノードあたり30TBの内部ストレージを保有
…NVIDIA提供Magnum IOの利用で、高速ストレージの拡張が可能
…ストレージの高速化とスケーラビリティ にも対応した構成を実現

■物理結線

・OSFPのOpticalモジュールおよびケーブルを利用して構成
…DGX H100はコンピューター用インターフェースとしてOSFP(800Gbps) 4ポートをサポート
…スイッチおよびDGX H100の内部は400Gbpsで接続
…ストレージ用インターフェースはQSFP-DDをサポート

InfiniBandネットワークの特長

主な特長

高帯域なネットワークリンク	▶ 最新の世代はNDR(400Gb/s) のネットワークリンクを提供
超低遅延なパケット転送	▶ 大規模環境でも、End to Endで1us未満の低遅延ネットワークが可能
ハードウェアベースの転送プロトコル	▶ 転送の高速化やCPUリソースの削減に寄与
信頼性の高いネットワーク	▶ クレジットベースのフロー制御でロスレスファブリックを実現

InfiniBand特有の機能 ▶ 更なる高性能ネットワークを実現

RDMA
ロスレスなデータ転送を実現

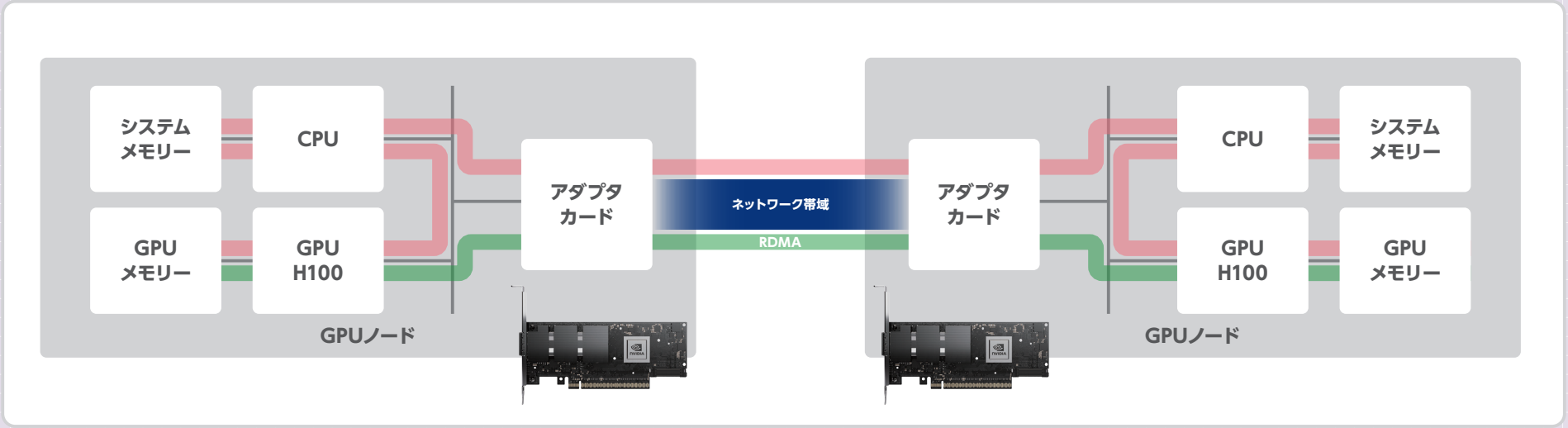
SHARP
AI/深層学習環境の分散処理を効率的に実行

SHIELD
大規模なInfiniBandネットワークを堅牢に

InfiniBand特有の機能…RDMA

RDMA(Remote Direct Memory Access)
InfiniBand HCAで実装可能なCPUオフロード機能

- ノード間データ転送をメモリ to メモリで完結
- CPUを介したデータ転送処理のボトルネックを解消し、CPUリソースの削減に寄与
- クレジットベースの転送で、ロス レス ファブリックを実現



InfiniBand特有の機能…SHARP

SHARP(Scalable Hierarchical Aggregation and Reduction Protocol)
InfiniBandスイッチに実装されるMPI・SHMEM通信のオフロード機能

大規模集合演算環境の課題

・ノード数を増加させても処理性能が頭打ちに
-集合演算によるCPU負荷高騰により通信遅延も発生

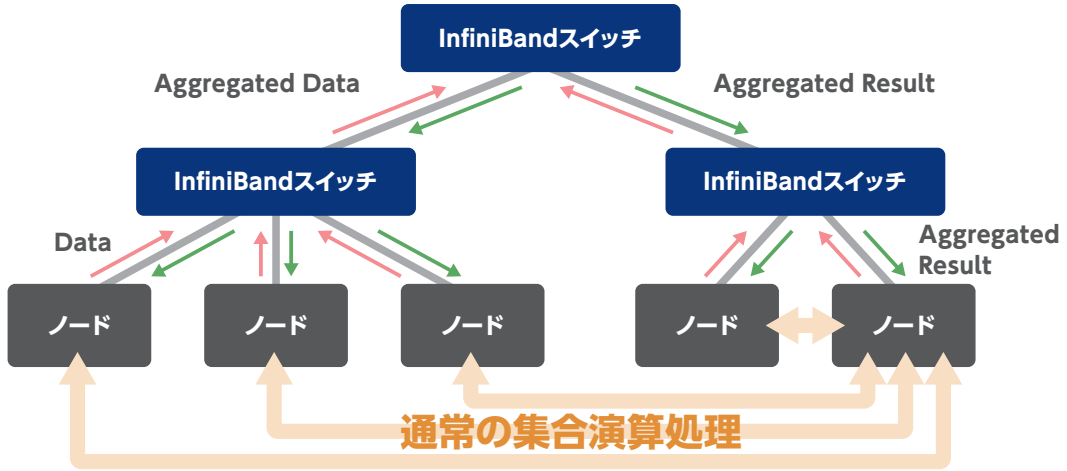
・演算処理がネットワーク負荷を増加させる
-ノード任せの処理ではネットワークを効率的に使用できず、
ネットワークのボトルネックを生む原因になり得る

▼

SHARPが解決の糸口

・MPIやSHMEM通信を使用した集合演算処理を、スイッチ上で処理

・HPCや分散学習を行うアプリケーションで効果を発揮



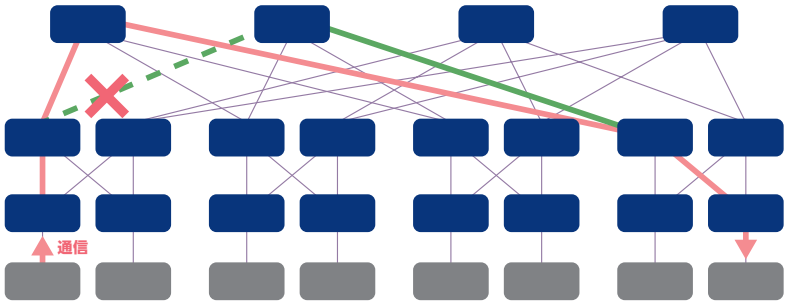
InfiniBand特有の機能…SHIELD

SHIELD(Self-Healing Interconnect Enhancement for InteLligent Datacenters)
InfiniBandネットワークの障害から高速に自己回復するための機能

- 専用のリニア・フォワーディング・テーブル(LFT)で経路管理しサブネットマネージャー (SM)の介入なしに通信経路を選択
- SMによる回復よりも最大5000倍の回復速度を提供
  ※SMによる経路管理をしている数千、数万ノードのクラスタ環境だと回復に30秒程度かかるケースもある
- 障害後もシームレスな計算を維持し、堅牢で回復力のあるネットワークを実現

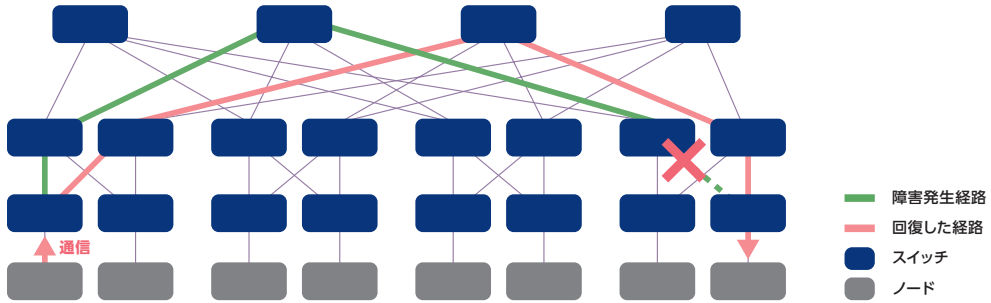
通信起点から見て障害発生ポートの代替ポートが存在する場合

代替ポートを使用した経路へ即座に切り替わり、ネットワークが回復する。



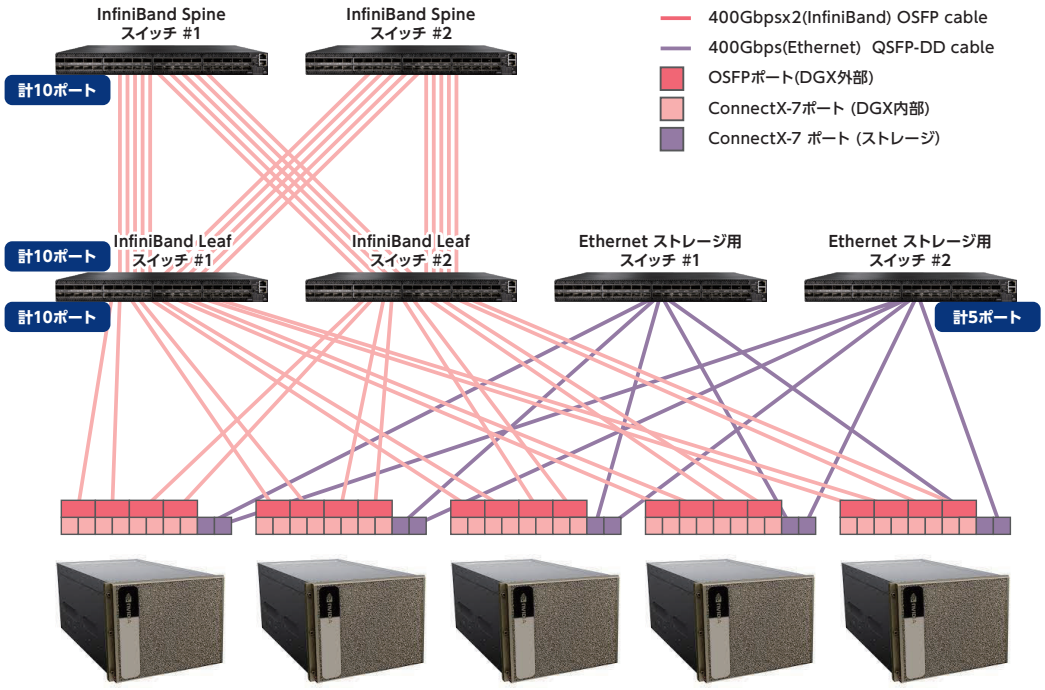
通信起点から見て障害発生ポートの代替ポートが存在しない場合

アダプティブ・ルーティングの通知機能も駆使し、経路の再計算を高速に行うことでネットワークを回復する。この時の回復までの時間は1us程度。



DGX H100 システム構成例(5ノード構成)

システム構成例

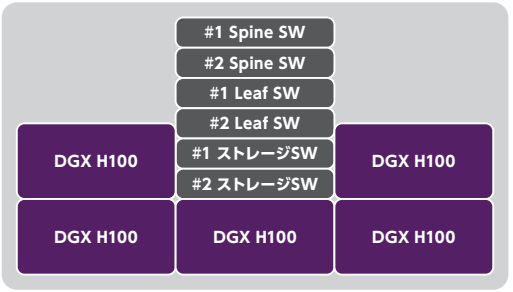


DGX H100ノード

構成例のBOMリスト

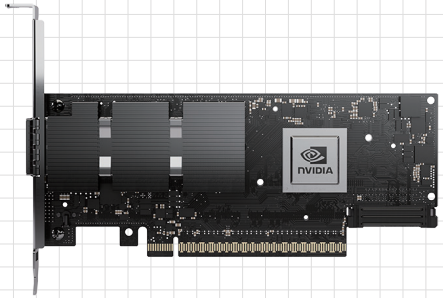
カテゴリ	型番	数量	説明
GPUサーバー	DGX H100	5	
InfiniBand スイッチ	QM9700	4	Quantum-2-based NDR InfiniBand switch, 32 800Gbps OSFP ports (64 400Gb/s ports)
Ethernet スイッチ	MSN4700	2	Spectrum-3 based Ethernet switch, 32 400Gbps QSFP-DD ports
トランシーバー (InfiniBand)	MMA4Z00	60	OSFP twin-port 850nm transceiver コンピュート向けDGX – スイッチ間 およびスイッチ間用
MPOケーブル	MFP7E10-Nxxx	60	passive fiber cable, MMF, MPO to MPOコンピュート向けDGX – スイッチ間 およびスイッチ間用
AOCケーブル	C-DQ8FNM0xx	10	400GbE PAM4 QSFP-DDストレージ向けDGX – スイッチ間用

ラックデザイン例



関連製品紹介

DGX H100をご利用いただく場合、こちらのNVIDIA関連製品をおすすめします。



InfiniBandアダプター

■NVIDIA ConnectX-7

- 16 Core / 256 Threads Datapath Accelerator
- Full Transport Offload and Telemetry
- Hardware-Based RDMA/GPUDirect
- MPI Tag Matching and All to All

InfiniBandスイッチ

■NVIDIA Quantum-2 QM9700

- 64-Ports of 400 Gbps
- 128-Ports of 200 Gbps
- 66.5B packets per sec (64B)
- 51.2Tb/s aggregate bandwidth

Ethernetスイッチ

■NVIDIA Spectrum SN4700

- 32-Ports of 400 Gbps
- 64-Ports of 200 Gbps
- 8.4B packets per sec
- 12.8Tb/s Max throughput

お問い合わせ

株式会社マクニカ　クラビス カンパニー　NVIDIA製品担当　｜　〒222-8561 横浜市港北区新横浜1-6-3 マクニカ第1ビル

045-470-9825　clvinfo@macnica.co.jp　https://www.macnica.co.jp/business/semiconductor/support/contact/