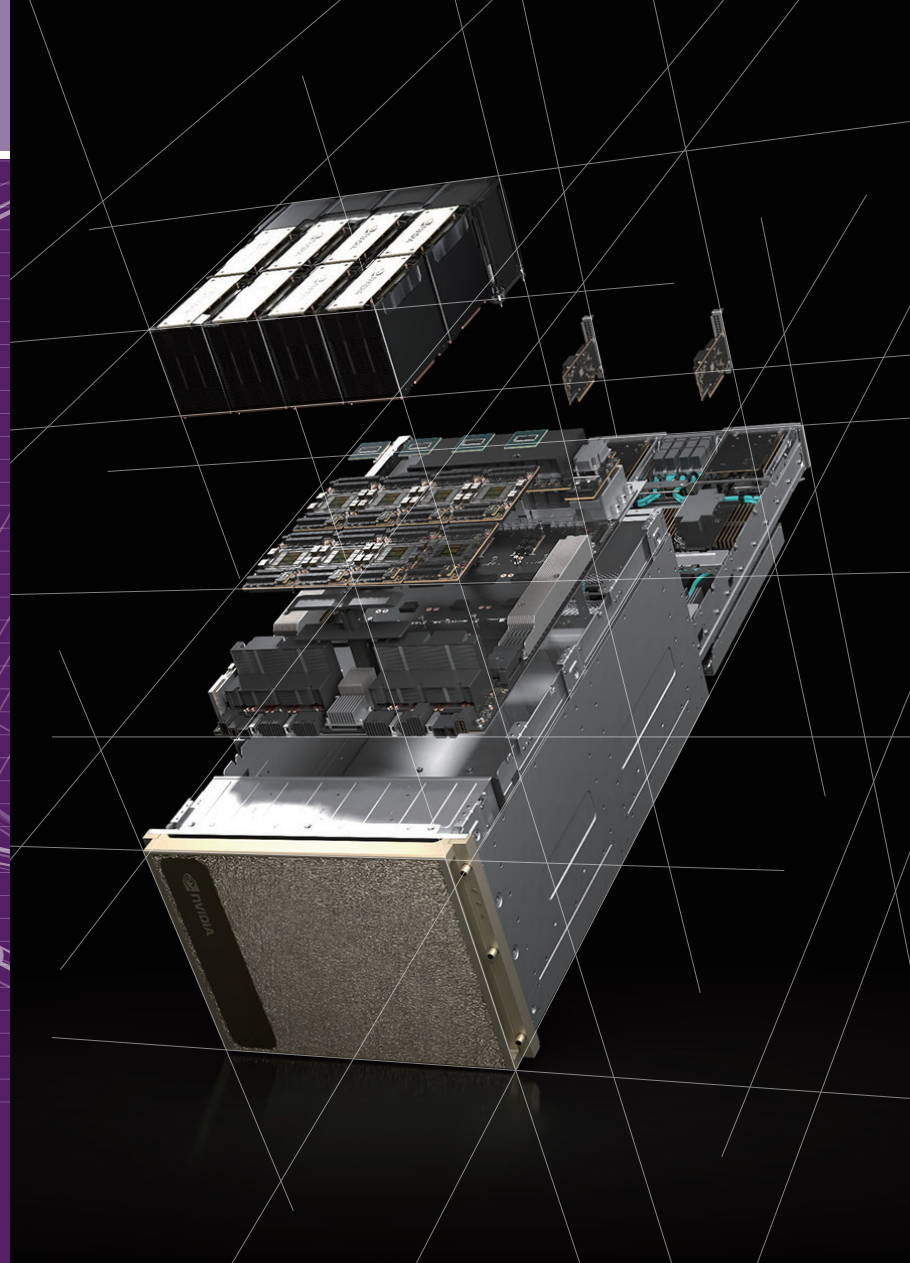
ネットワーク視点の

NVIDIA DGX™ B300 大解剖

GPUを最大限活かすNVIDIA ConnectX®-8の役割



MACNICA

目次

ネットワーク視点の

NVIDIA DGX™ B300 大解剖

GPUを最大限活かすNVIDIA ConnectX®-8の役割

NVIDIA DGX B300 概要 2

DGX B300 構成要素 3

- ConnectX-8が8基 搭載される理由 4
- ConnectX-8は帯域確保だけじゃない 5

DGX B300のシステム構成を考える 6

- InfiniBandネットワークの特長 7
- InfiniBand特有の機能 ..... RDMA 8
- InfiniBand特有の機能 ..... SHARP 9
- InfiniBand特有の機能 ..... SHIELD 10

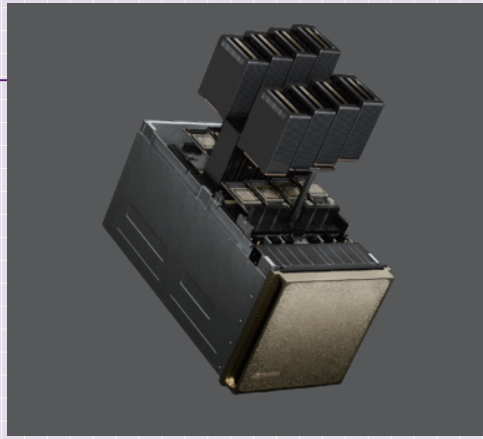
DGX B300 システム構成例(5ノード構成) 11

関連製品紹介 12

問い合わせ 13

NVIDIA DGX B300 概要

大規模で複雑なAIジョブの限界を突破

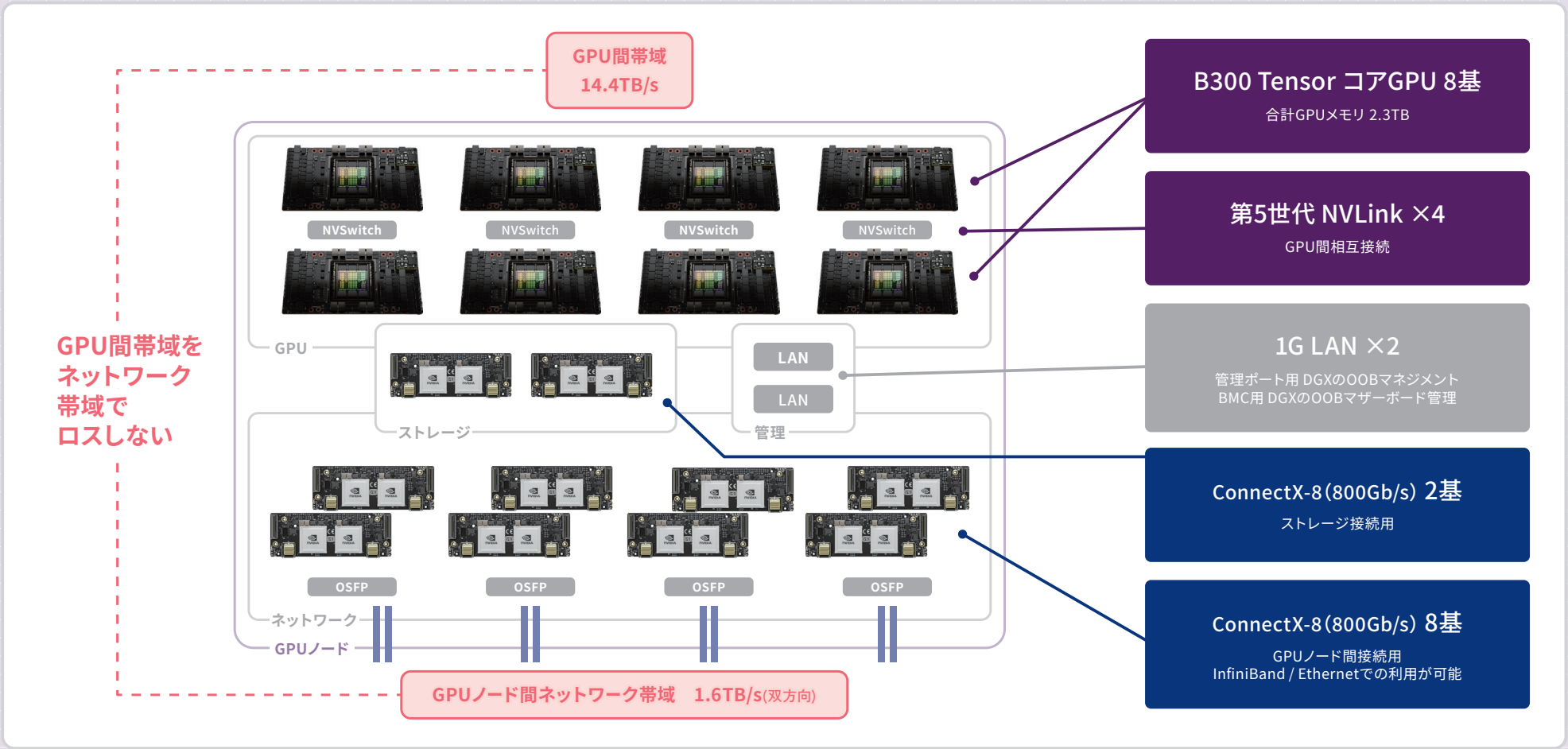


NVIDIA DGX™ B300は、第5世代のNVIDIA DGXシステムで、NVIDIA B300 Tensor コア GPUとNVLink Switch System、NVIDIA ConnectX®-7を組み合わせた世界最速のAIシステムです。次世代アーキテクチャーを採用することで前世代と比較して2倍のパフォーマンスと2倍の高速ネットワークを実現し、企業のAIインフラの中核を担うように設計されています。

主な仕様と性能

▶ 仕 様	B300 Tensor コアGPU 8基 合計GPUメモリ 2.3TB	第5世代 NVLink GPU間相互接続	ConnectX-8 10基 GPUノード間接続用+ストレージ接続用
▶ 性 能	性能2倍(72PFLOPS) FP8 精度で前世代の 2倍以上	GPU間帯域 14.4TB/s 双方向帯域幅	ピークネットワーク帯域 2TB/s 双方向帯域幅

DGX B300 構成要素

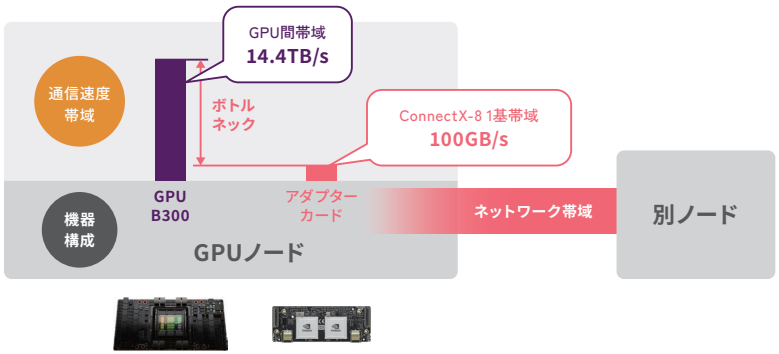


ConnectX-8が8基 搭載される理由

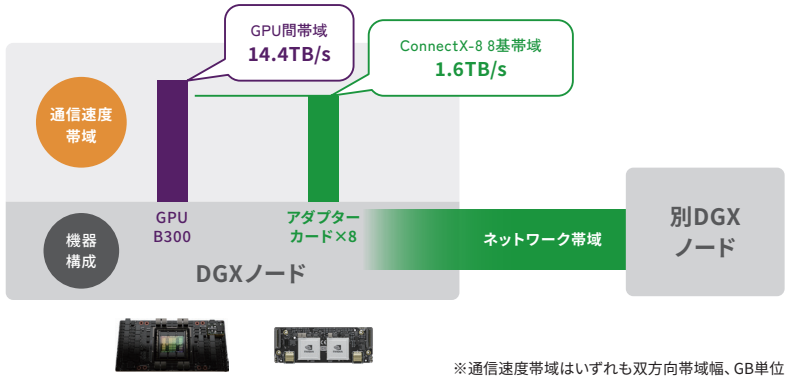
ネットワーク帯域を確保し、ノード間のGPU演算速度を維持

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14.4TB/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 B300では ConnectX-8 8基により1.6TB/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 B300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 B300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-8は帯域確保だけじゃない

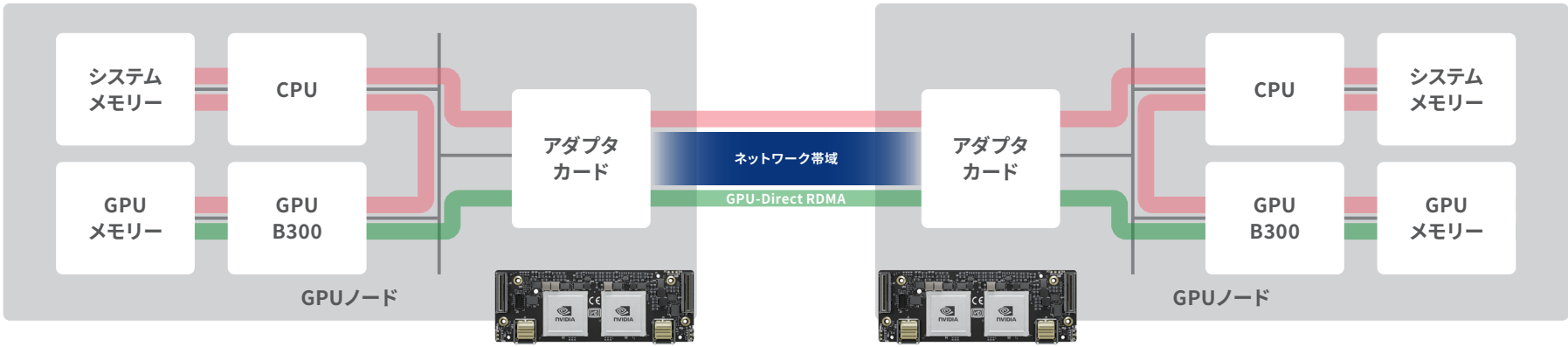
GPU-Direct RDMAがノード間通信高速化に貢献

GPU-Direct RDMAとは？

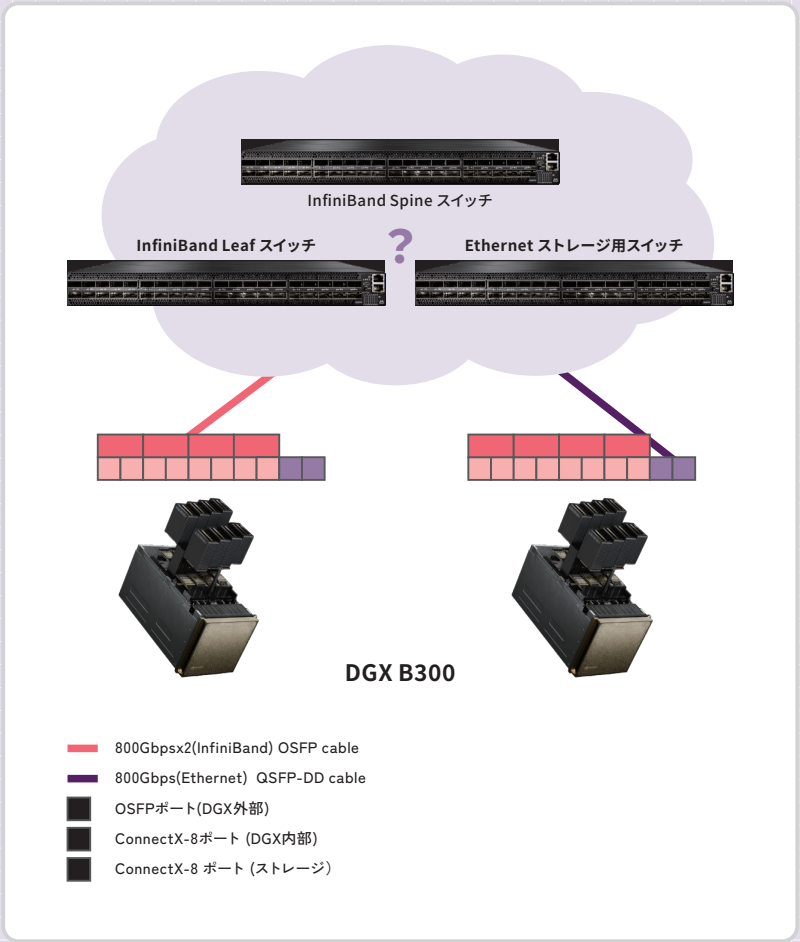
GPUノード間通信でホストCPUを介さずGPUメモリー同士が直接データ転送をする技術です。ConnectX-8のRDMA機能とCUDA TOOL KITによりGPU-Direct RDMAを実現し、ノードを跨いだGPUメモリー間データ転送をハードウェアへオフロードが可能となります。

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 B300のシステム構成を考える



複数ノードのDGX B300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 B300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 B300内ConnectX-8をEthernetスイッチに接続し、冗長構成を組む
…DGX B300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 B300はコンピューター用インターフェースとしてOSFP(1.6Tbps) 4ポートをサポート
…スイッチおよびDGX B300の内部は800Gbps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XDR(800bps) 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

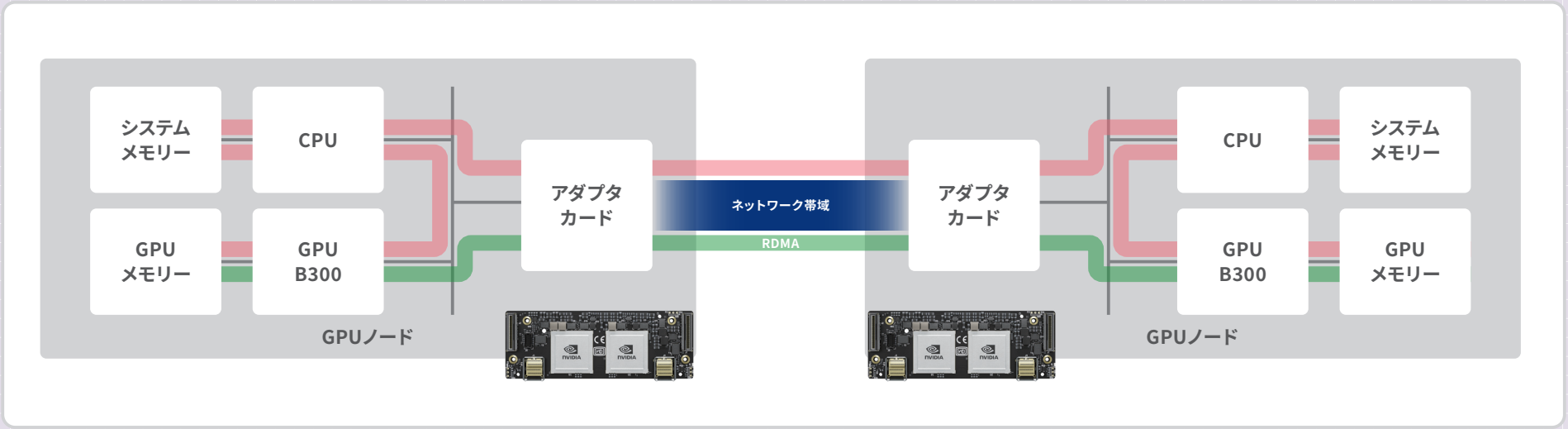
SHARP
AI/深層学習環境の分散処理を効率的に実行

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

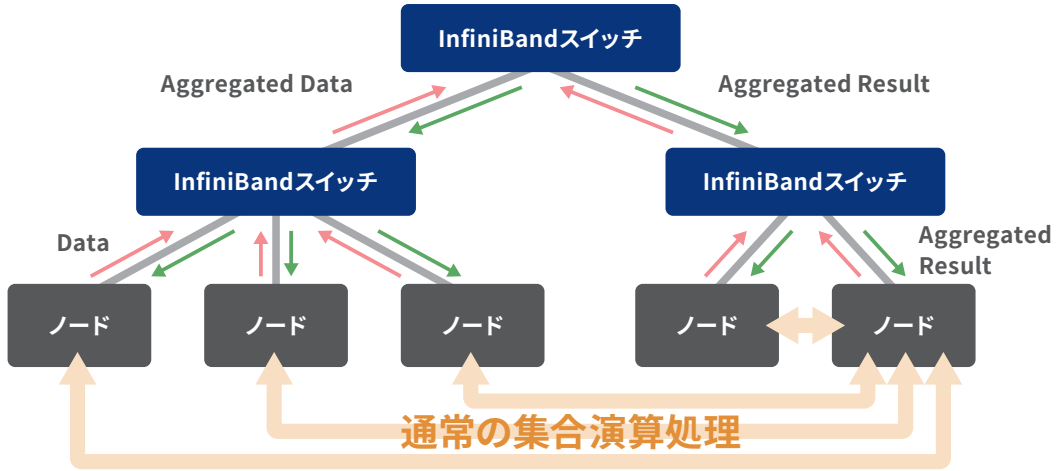
・演算処理がネットワーク負荷を増加させる
-ノード任せの処理ではネットワークを効率的に使用できず、ネットワークのボトルネックを生む原因になり得る

▼

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



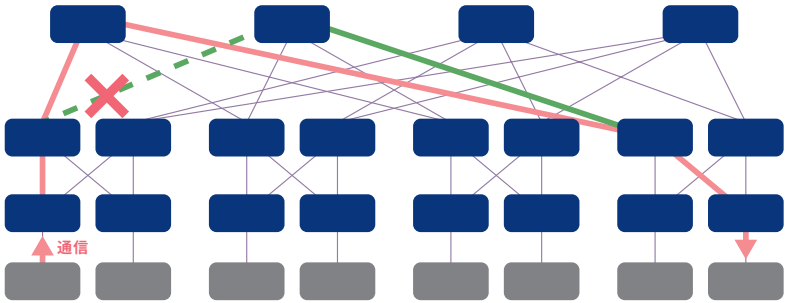
InfiniBand特有の機能...SHIELD

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

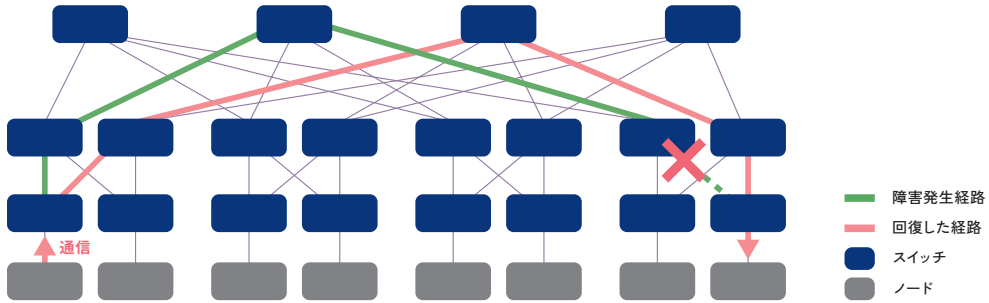
通信起点から見て障害発生ポートの代替ポートが存在する場合

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。



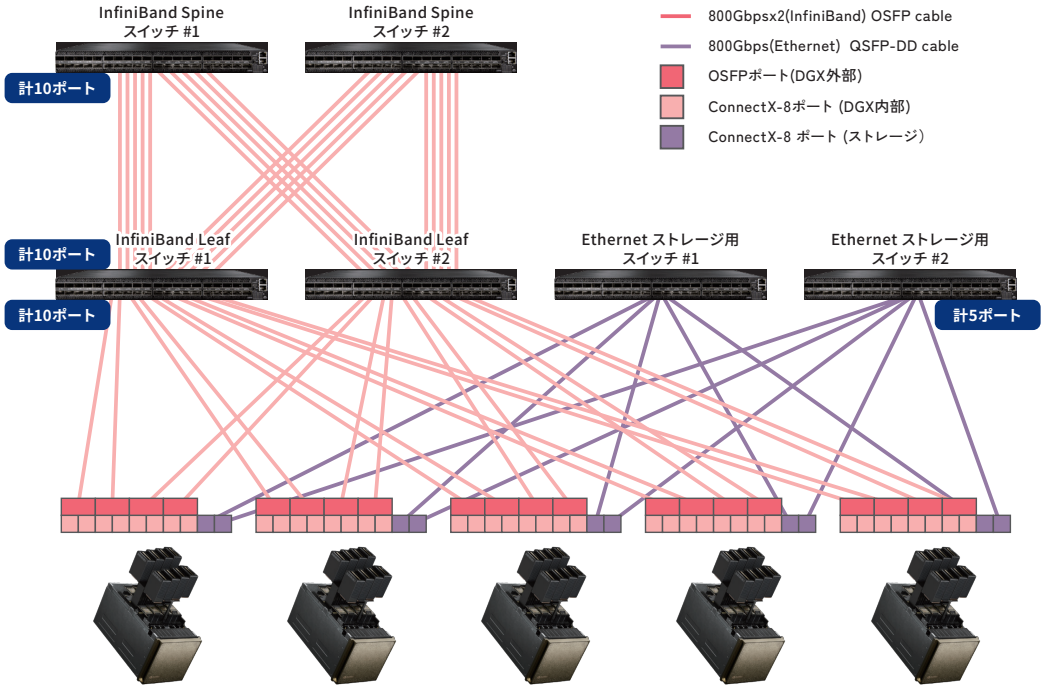
通信起点から見て障害発生ポートの代替ポートが存在しない場合

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 B300 システム構成例(5ノード構成)

システム構成例

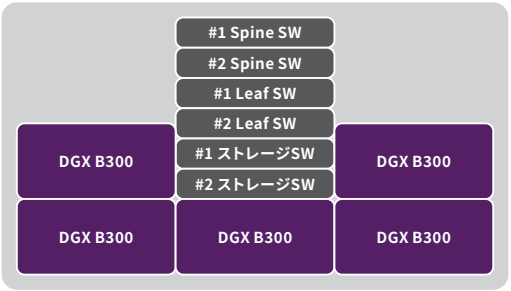


DGX B300ノード

構成例のBOMリスト

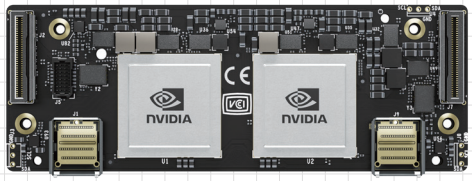
カテゴリ	型番	数量	説明
GPUサーバー	DGX B300	5	
InfiniBand スイッチ	QM9700	4	Quantum-2-based NDR InfiniBand switch, 32 800Gbps OSFP ports (64 400Gb/s ports)
Ethernet スイッチ	MSN5600	2	Spectrum-3 based Ethernet switch, 64 800Gbps OSFP ports
トランシーバー (InfiniBand)	MMS4A00	60	OSFP twin-port 850nm transceiver コンピュート向けDGX – スイッチ間 およびスイッチ間用
MPOケーブル	MPF7E30-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	MCP7Y10-N00x	10	800GbE 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 B300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-8

- 16 Core / 512 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 SN5600

- 64-ports of 800 Gbps
- 33.3 packets per sec
- 51.2Tb/s Max throughput

MACNICA

・本資料に記載されている会社名、商品、サービス名等は各社の登録商標または商標です。なお、本資料中では、「™」、「®」は明記しておりません。・本資料は、出典元が記載されている資料、画像等を除き、弊社が著作権を有しています。・著作権法上認められた「私的利用のための複製」や「引用」などの場合を除き、本資料の全部または一部について、無断で複製・転用等することを禁じます。・本資料は作成日現在における情報を元に作成されておりますが、その正確性、完全性を保証するものではありません。

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

045-470-98　clvinfo@macnica　https://www.macnica.co.jp/business/semiconducto

©Macnica, Inc.　● 本ホワイトペーパーに掲載されております社名および製品名は、各社の商標および登録商標です。