VIOPS06 WORKSHOP
Infiniband技術動向および導入事例

July 22nd, 2011

メラノックス テクノロジーズ ジャパン株式会社
日本地区担当ジェネラルマネージャ
津村英樹

Mellanox TECHNOLOGIES

会社概要

Ticker: MLNX

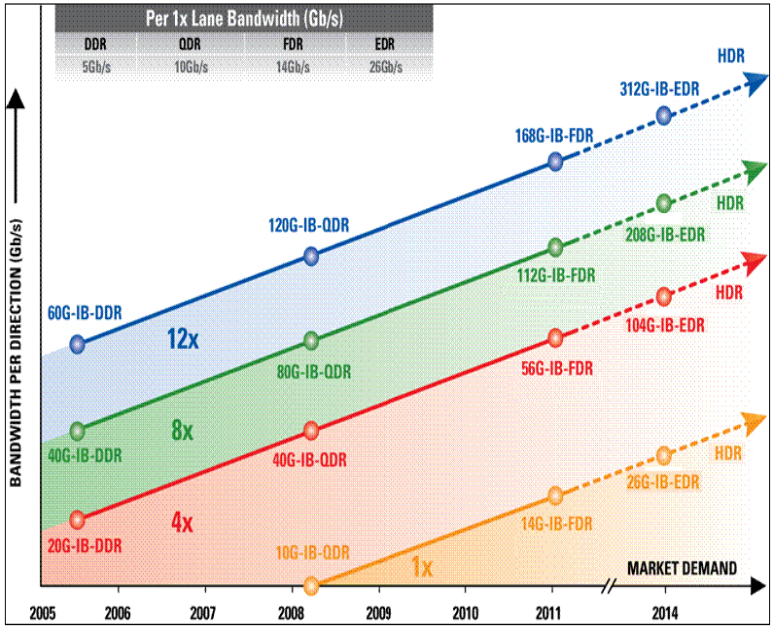
- HPC、金融、クラウド市場の要求を満たす最高性能のITシステムを実現するソリューションプロバイダー
  - インフィニバンドと10G/40GbEtherのEnd-Endソリューションを提供する高速ネットワーク市場のリーディングカンパニー
  - 2011年3月時点で、750万ポート以上を出荷
- 本社:
  - イスラエル ヨークナム, カリフォルニア州サニーベール
  - 675人の従業員; ワールドワイドのセールス＆サポート
- 順調な業績推移、優良な財務状況
  - 2010年度の記録的な売上; $154.6M
  - 2011年度第一四半期売上= $55.1M
  - 2011年度第二四半期見込~$61.5 – 62.0M
- Voltaire社との相互補完的な合併を完了

Recent Awards

インフィニバンド概要

物理層の仕様

- Infiniband Trade Associationで定義された標準規格
- 物理層は10Gbit EthernetやFibre Channelと同じIEEE802.3がベース
- 現行の規格では一方向あたり信号帯域幅2.5Gbps（QDR）、3.5Gbps（FDR）
- QDRまでは8B/10B、FDRからは64B/66Bの符号方式を採用
- 全二重通信のため双方向の合計帯域幅は2倍
- 複数の接続（レーン）の集束が可能（4x、12x）
- Node-Node ： 56Gbps (FDR 4x)
- Switch-Switch ： 168Gbps (FDR 12x)
- 1μs以下のアプリケーション・レーテンシー



インフィニバンド固有の様々な特長

MPI（メッセージ・パッシング・インタフェース）

- 高い伝送効率を実現するチャネルアーキテクチャ
  - PCIエクスプレスのロードストア・アーキテクチャはCPUが接続するI/Oデバイスと直接送受信するため、遅いI/Oデバイスの使用時や接続数が多くなるとオーバーヘッドが大きくなる。
  - チャネル・アーキテクチャではCPUは必要なデータをメモリに書き込んだらあとは別のタスクに移行できる。I/Oデバイスは自分の速度に合ったペースでメモリからデータを取得する。
  - I/Oデバイスの数が増えてもオーバーヘッドは小さいため、大規模システムにも対応可能。
  - メラノックス社のConnect-Xシリーズ Host Channel Adapter用コントローラ チップが提供する様々なオフロード機能により、さらに高効率、低レイテンシーな通信が実現できます。

© 2011 MELLANOX TECHNOLOGIES - MELLANOX CONFIDENTIAL -

メラノックスが提供する効果的なオフロード機能

RDMA（Remote Direct Memory Access）

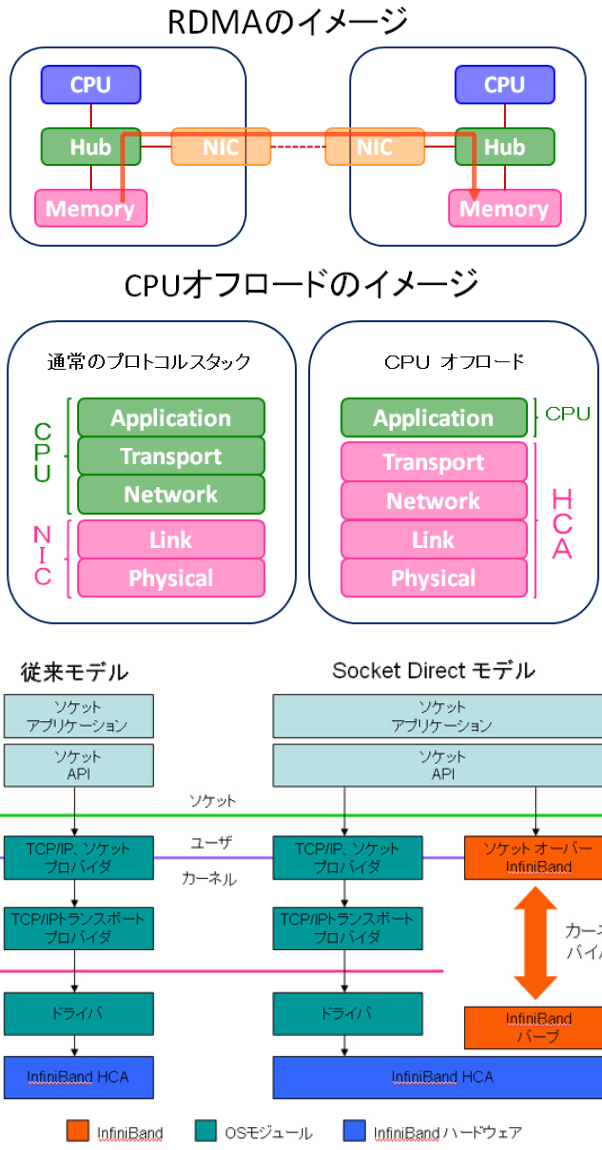
- 独立したシステムのメインメモリ間、システムとストレージ間のデータ転送を効率化するテクノロジ。メモリ間のデータ転送時にはホストCPUの割り込みがほとんど発生しないため、ホストCPUの　オーバーヘッド低下に
  加え、データ転送時のレイテンシ短縮にもつながる。
- 10Gbit Ethernetを利用する場合、システム構築や運用管理がInfiniBandよりも容易という利点を持つ。
  Ethernet上でRDMA機能を実現するRDMA over EthernetもMellanox社HCAではサポートしており、non-HPC市場での導入が進んでいる。

CPU Offload

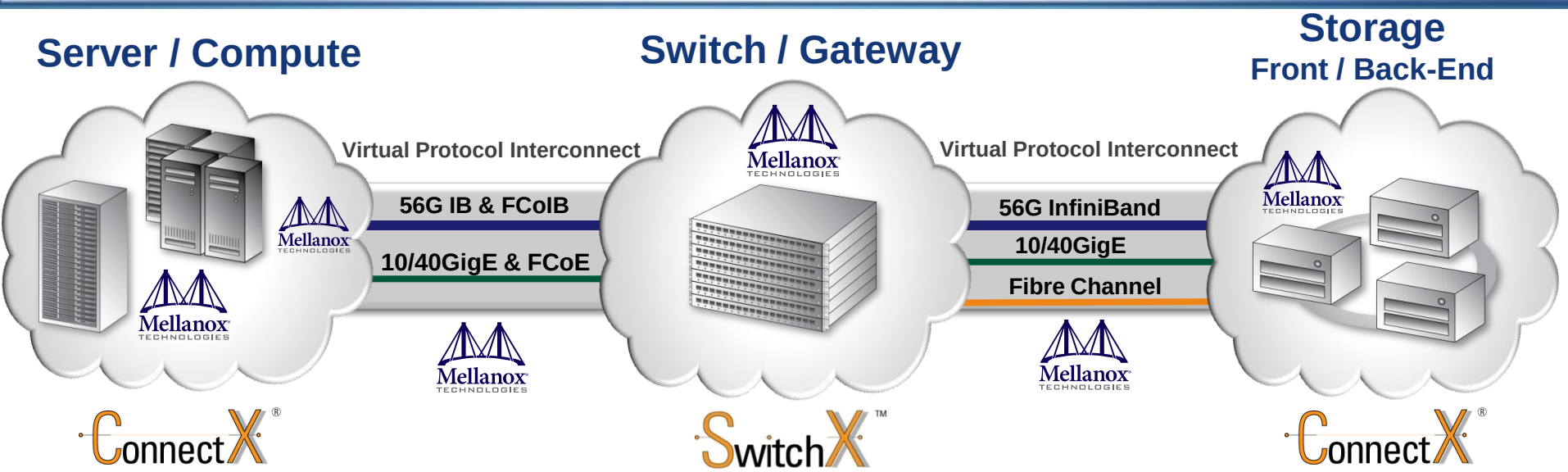
- TransportレイヤとNetworkレイヤをHCAのチップがハードウェアで処理することにより、CPUの負荷を軽減できる。

カーネル・バイパス

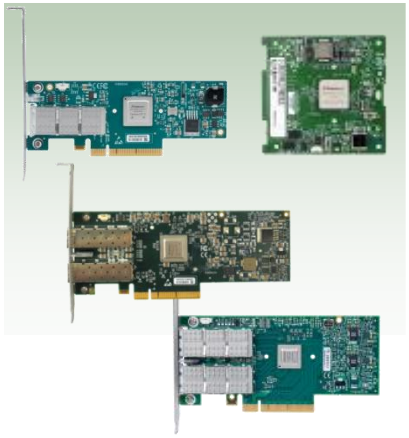
- InfiniBandで使用されるSDP（Sockets Direct Protocol）は、アプリケーションとその下にあるハードウェアを透過的に接続する基本的なソケットメカニズムを提供する。これによりInfiniBandを併用しながら従来のTCP/IPソフトウェアモデルを使い続けられる。
- ハードウェアがInfiniBandの場合には、複雑なトランスポート処理をInfiniBand HCA（Host Channel Adapter）がOSカーネルで直接行えるようになる。これにより、ソフトウェアからはTCP/IPに見せながら、設定の変更は殆ど無しにInfiniBandの利点を享受できる。（IP over IB）



最先端のE2Eソリューションを実現するラインナップ



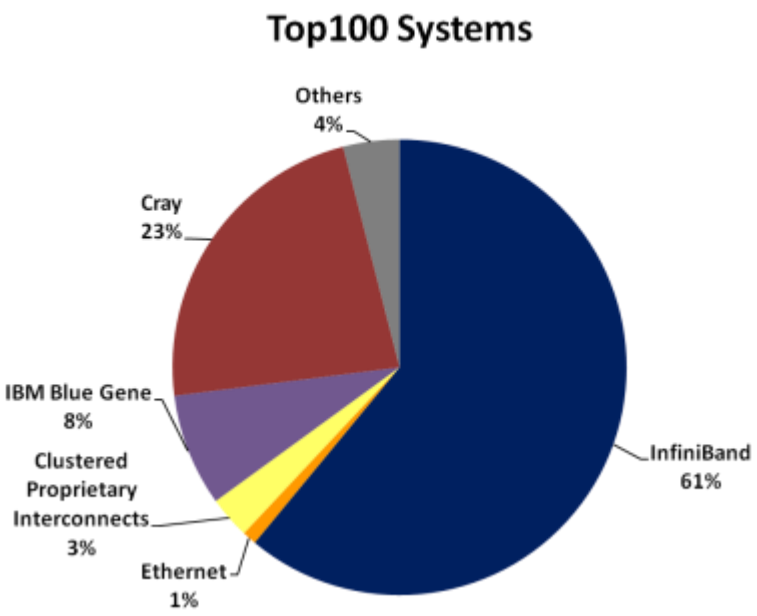
Industries Only End-to-End InfiniBand and Ethernet Portfolio

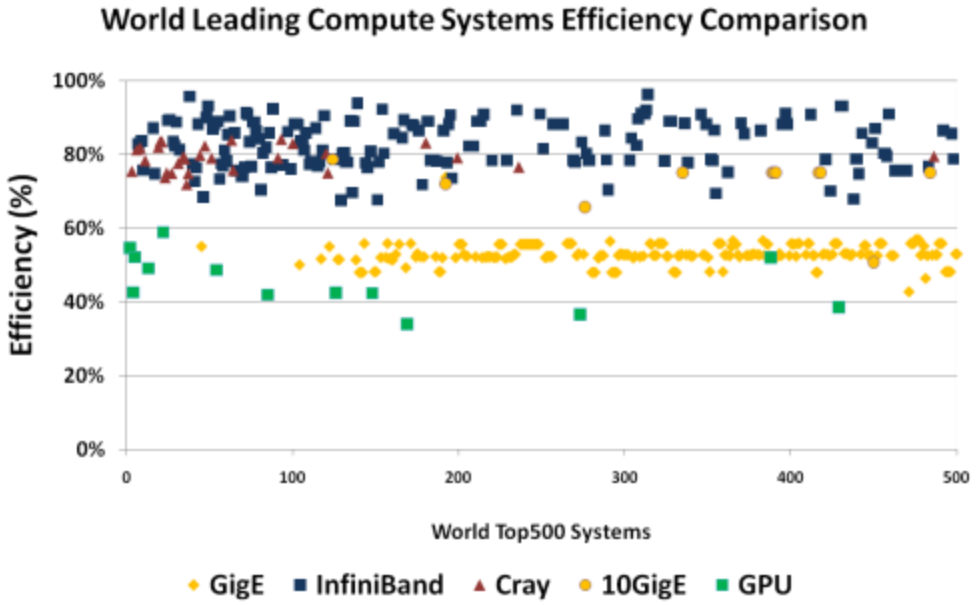
ICs	Adapter Cards	Host/Fabric Software	Switches/Gateways	Cables
				

メラノックスInfiniBand製品群の特長

- 最高のスループット
  - node to nodeの転送レートがQDRで40Gb/s、FDRで56Gb/s
  - QDRで約90M MPI messages / 秒
  - ゼロ・コピーでSend/receiveおよびRDMA operationsを実行
- 低レイテンシー
  - 1-1.3usec MPI end-to-end
  - 0.7-1us InfiniBand latency for RDMA operations
  - Lowest latency switch fabric – up to 45% lower vs other solutions
- 最小のCPUオーバーヘッド
  - フル・トランスポート・オフロードにより、ユーザ・アプリケーションに対するCPUの稼働率を最大化

Mellanox in the TOP500 - (June 2011)





- Mellanox InfiniBand builds the most powerful clusters
  - Connects 5 out of the TOP10, 61 systems in TOP100, 115 in TOP200, 147 in TOP300
- InfiniBand represents 42% of the TOP500
- Mellanox InfiniBand enables the highest utilization on the TOP500
  - **Up to 96% system utilization**
- Mellanox InfiniBand enables the most energy efficiency Petascale system
- Mellanox 10GbE is the highest ranked 10GbE system (#124)

Paving The Road to Exascale Computing

Dawning (China)

TSUBAME (Japan)

NASA (USA)
>11K nodes

LANL (USA)

PetaScale

Mellanox Connected

CEA (France)

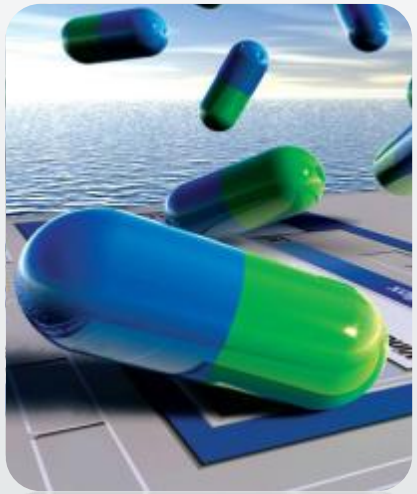
- Mellanox InfiniBand is the interconnect of choice for PetaScale computing
  - Accelerating 50% of the sustained PetaScale systems (5 systems out of 10)

様々な業界の代表企業への導入実績

30%+ of Fortune-100 & top global-High Performance Computers

6 of Top 10
Global Banks

9 of Top 10
Automotive Manufacturers



5 of Top 10
Pharmaceutical Companies

7 of Top 10
Oil and Gas Companies

コレクティブ・オフロード

- MPI Collective Operations
  - コレクティブ通信は初期の入力データを送信するためのブロードキャストや複数の発信源のデータを統合するリダクション(Reduction)、全体的な同期化のためのバリア(Barrier) のようなオペレーションのためにアプリケーションによって頻繁に使用されており、工学・科学アプリケーションの拡張性にきわめて大きな影響を与えます。
  - コレクティブ通信はシステム内の全てのプロセス/ノードを連結させるために全体的な通信操作を行います。従ってできるだけ効率よく、またできるだけ早く実行されなければなりません。コレクティブ通信の現在の実装では拡張性が無くシステムノイズに苦しみ、CPU処理 時間の多くを消費してしまいます。
- **メラノックス社のコレクティブ・オフロード機能**
  - Mellanox ConnectX®-2 アダプタと InfiniScale® IV スイッチは、それらアダプタやスイッチへコレクティブ通信をオフロードすることによってこの通信の拡張性の問題に対応します。
  - Mellanox のアダプタとスイッチは通信パターンをオフロードし、ノンブロッキングのコレクティブオペレーションとアプリケーション計算のオーバーラップを許容することによく適しております。通信の処理をこのHCA にオフロードすることは、システムノイズの影響を最小限に抑え、通信と計算のオーバーラップを飛躍的に増加するための機能を提供します。

CORE-Direct アダプタ機能	CORE-Direct スイッチ機能
ワイヤースピードで且つ信頼性のある方法でコレクティブオペレーションを対処し、インフライトコレクティブを制限無くサポートする	世界屈指のマルチキャストオペレーション・グループをサポートする
ローカルノード、リモートノードの通信に対処する	世界屈指の低いスイッチレイテンシー
浮動小数点演算を実行する	プログラマブルルーティング・アダプティブルーティング
異なるコレクティブトポロジーを柔軟にサポートする	
ブロッキング、ノンブロッキングコレクティブオペレーション(MPI-3)をサポートする	

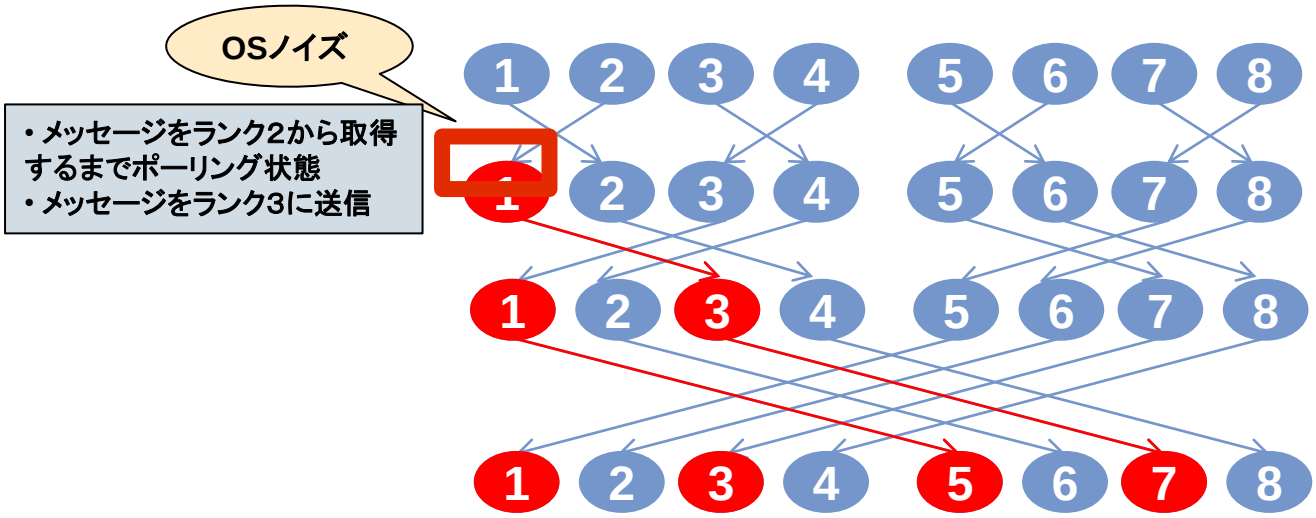
Collective Algorithms (Recursive Doubling)

システムノイズの軽減

- CPU は計算と通信の両方のタスクを実行する必要があり、一般的にはそれらのタスクを順次実行します。コレクティブオペレーションは CPU のランタイムに深刻なオーバーヘッドを与え、それによって OS ノイズは CPU タスクの全体の実行時間に対して追加の遅延を発生させます。
- MPI オペレーションによるアプリケーション性能への OS ノイズの影響を以下に図解します。この例では、ランク 1 (またはノード 1)上の OS ノイズがランク 2 から期待されるメッセージの到着を遅らせます。この遅延はランク 1 の実行時間を増加させるだけでなく間接的に影響を与えるランク 3，5，7 の実行時間も増加させます。それによってアプリケーション全体の実行時間に影響を与え、システム全体の性能を劣化させます。

メラノックス社のコレクティブ・オフロード機能

- コレクティブ通信のオペレーションをHCA にオフロードすることはそのCPUの通信管理の負荷を軽減させ、コレクティブ通信と計算のオーバーラッピング/並列処理の増加を促進します。このオーバーラップにおける更なる利点は CPU の代わりにその HCA に計算の一部を実行させることによって実現されます。ConnectX-2 はMPI_Reduce とMPI_Allreduceコレクティブオペレーションのデータ操作部分をHCA にオフロードすることを可能にした浮動小数点演算ユニットを含んでいます。

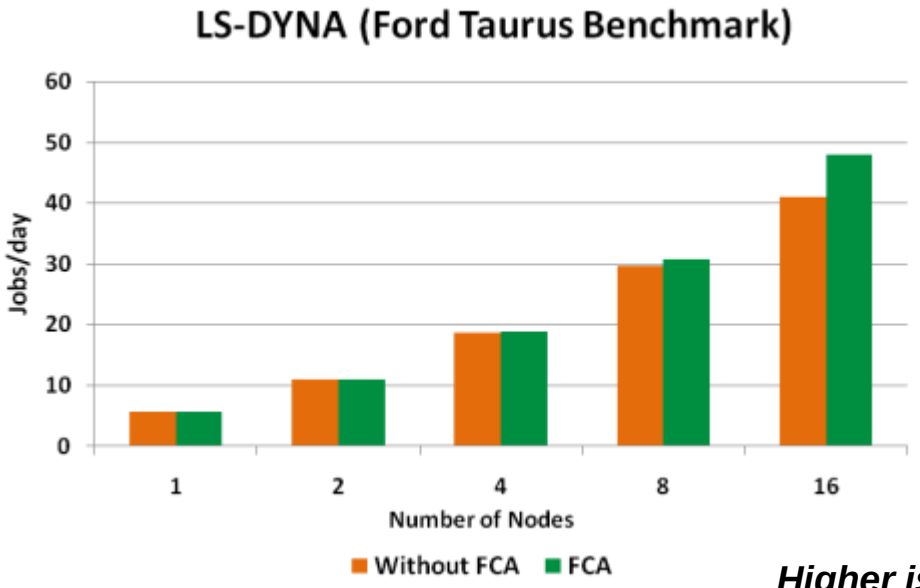


採用事例: LS-DYNA (自動車衝突シミュレーション)

- LS-DYNAは、複雑な実世界での問題をシミュレーションできる構造解析、流体解析用の汎用シミュレーション・ソフトウェアです。
  - LS-DYNA は自動車や飛行機などの製造業における様々なシミュレーションで使用されています。

- 導入効果: FCAはLS-DYNAの性能を約20%向上することに成功
  - 16ノード, 192コア
  - アプリケーション性能はクラスタサイズが大きくなるほど向上する

LSTC
Livermore Software Technology Corp.



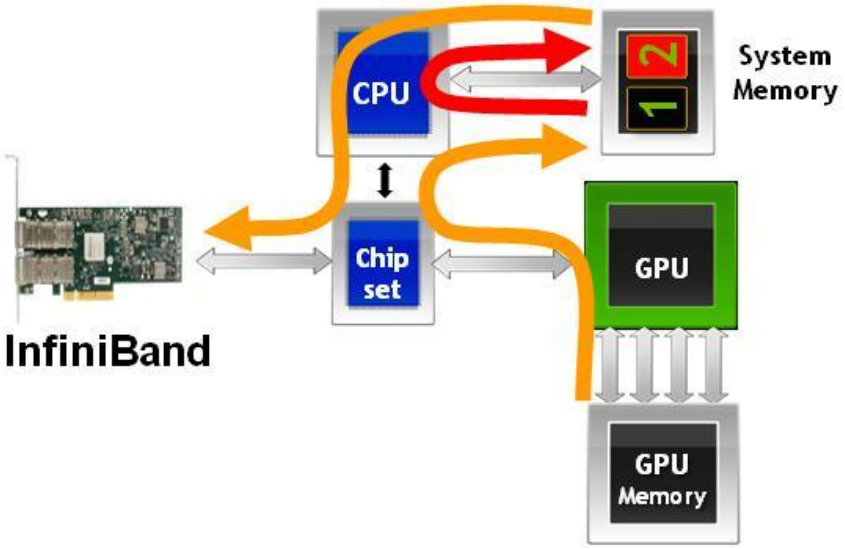
Higher is better

*Acknowledgment: HPC Advisory Council for providing the performance results

GPU Direct の概要

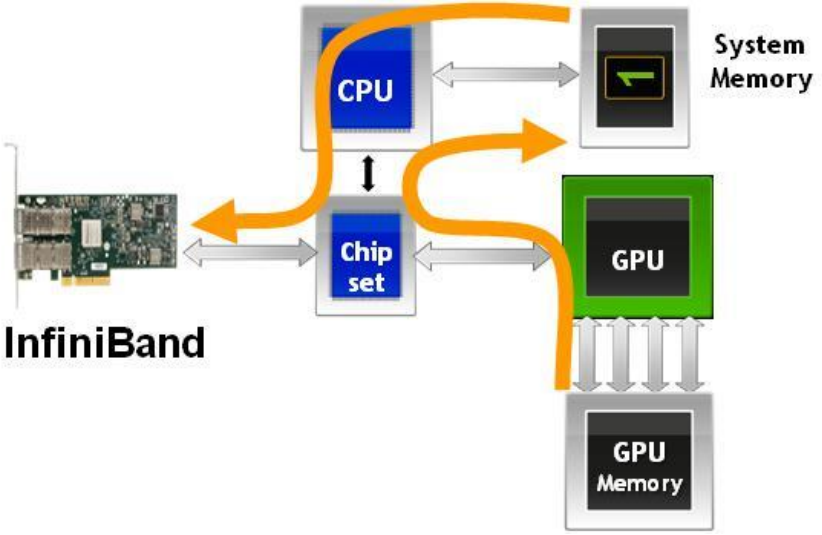
- Mellanox InfiniBandとNVIDIA GPUの最適な組合せ
  - **InfiniBandとGPU間のメモリコピーのステップを削減、GPU-to-GPU の通信時間を 30 % 短縮**

No GPUDirect



1. GPU が GPU 専用のホストメモリにデータをライトする。
2. ホスト CPU はそのデータを GPU 専用のホストメモリからインフィニバンドデバイスが RDMA 通信で利用できるホストメモリへコピーする。
3. インフィニバンドデバイスはそのデータをリードし、リモートノードへデータを転送する。

GPUDirect

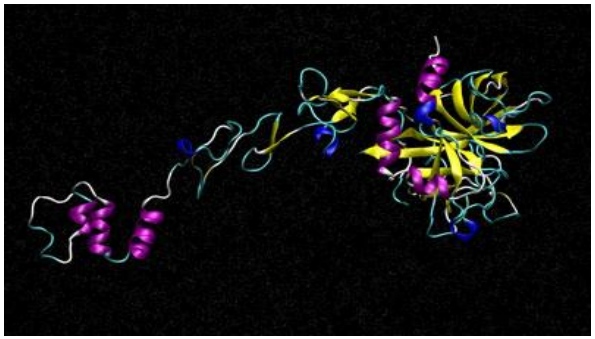


NVIDIA GPU Direct は NVIDIA GPU と Mellanox インフィニバンドアダプタの間で新しいインタフェースを提供し、双方のデバイスで同じシステムメモリを共有することを可能にします。

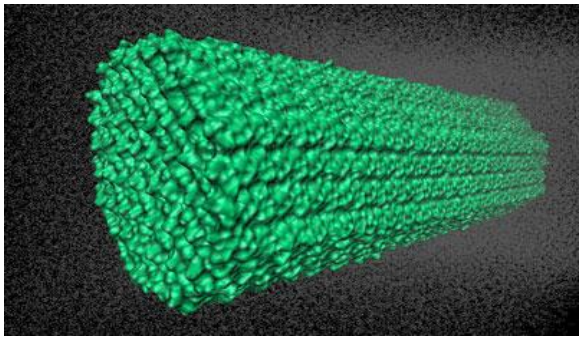
1. GPU が GPU 専用のホストメモリにデータをライトする。
2. インフィニバンドデバイスはそのデータをリードし、リモートノードへデータを転送する。

性能向上例1 - Amber

- Amber　：　分子動力学法向けに開発されたソフトウェアパッケージ
  - 二分子反応に関して最も広く利用されているプログラムの一つ
  - 広範なユーザーベース
  - 原子の動きを計算する分子動力学シミュレーション
- **ベンチマーク内容**
  - FactorIX - 90,906原子から構成される血液凝固第Ⅸ因子のシミュレーション
  - Cellulose - 408,609原子から構成されるセルロース繊維のシミュレーション
  - PMEMD（粒子メッシュエーヴァルト分子動力学）シミュレーションプログラムを使用
- **ベンチマーク システム – 8ノード PCクラスタ**
  - Mellanox ConnectX-2 HCA と InfiniBand スイッチ
  - 各ノードはNVIDIA Fermi C2050 GPUを一つ搭載

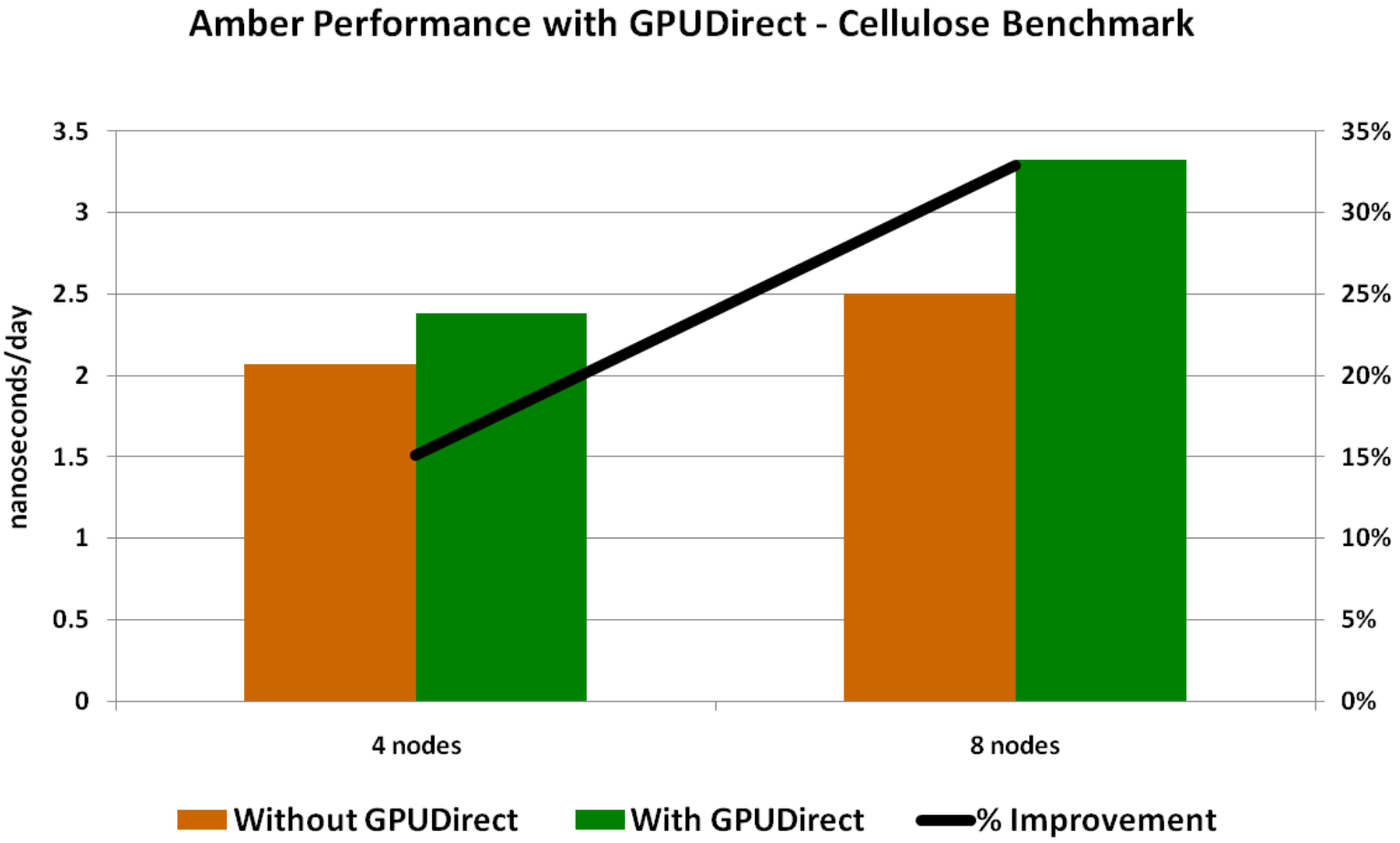


FactorIX



Cellulose

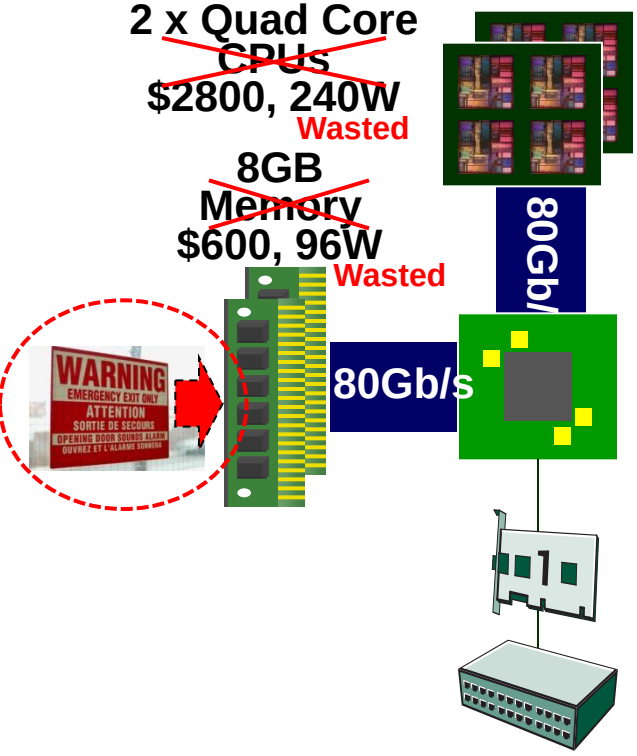
Amber Performance with GPUDirect



- GPUDirect によって33%の性能向上
- クラスタサイズの増加に応じてさらに性能向上

データセンタ内IT基盤のボトルネックの推移

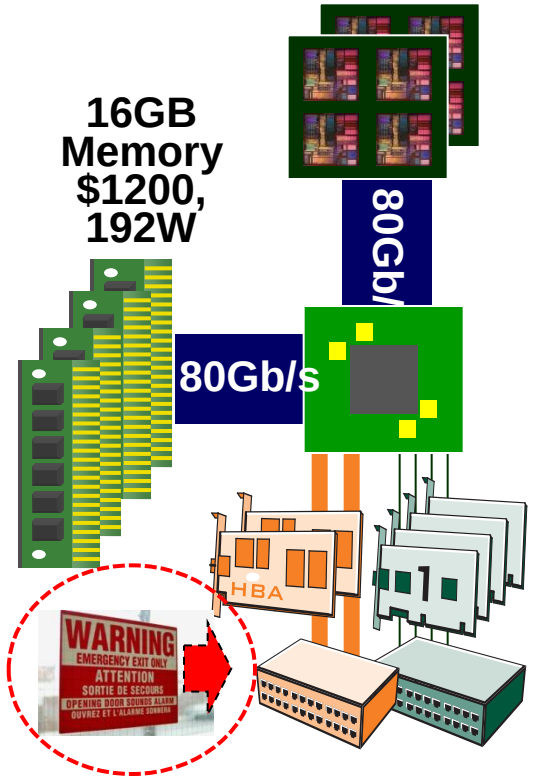
過去



1GigE, 50µs, $100, 4W

- 強力な物理サーバのスペックをアプリケーションが使用しきれずCPUとMemoryはアイドル状態
- 電力、性能、投資の無駄

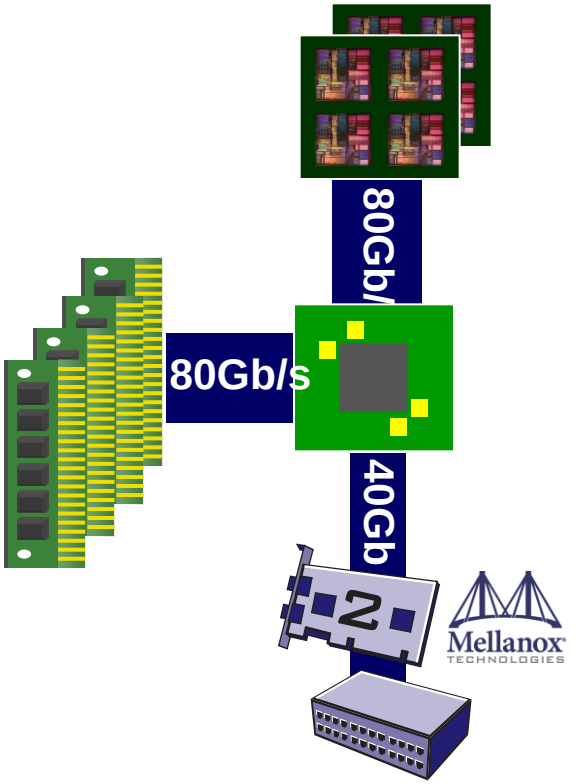
現在



12Gb/s, 50µs, $2400, 28W

- サーバ仮想化により、複数の仮想サーバがCPUやメモリを共有することで効率化
- I/Oがボトルネックに配線など、I/O接続機器と複雑性の増加＝インターコネクト投資増加

今後



10-40Gb/s, 1µs, $900, 10W

- InfiniBandで統合化されたI/O
- ケーブリング、複雑性の解消
- 消費電力やコストの大幅な削減
- 低価格&高性能なマルチプロトコル仮想環境向Converged Network

インフィニバンド導入事例

Salesforce.com 社の削減例

xsigo systems

salesforce.com

12ブレードシャーシによる統合で、
約1億円のコスト削減を実現

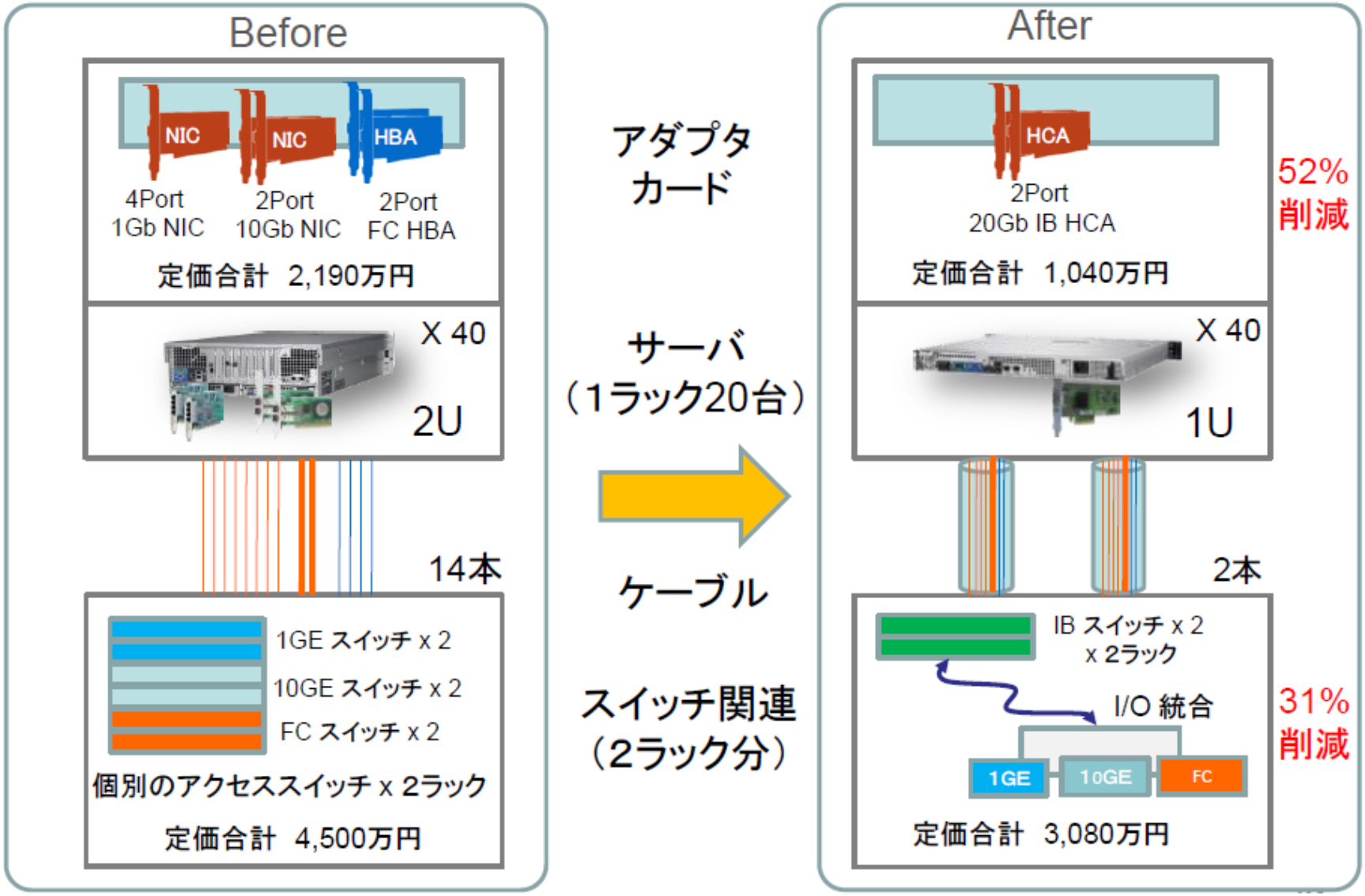
Xsigo 導入前

98 ケーブル

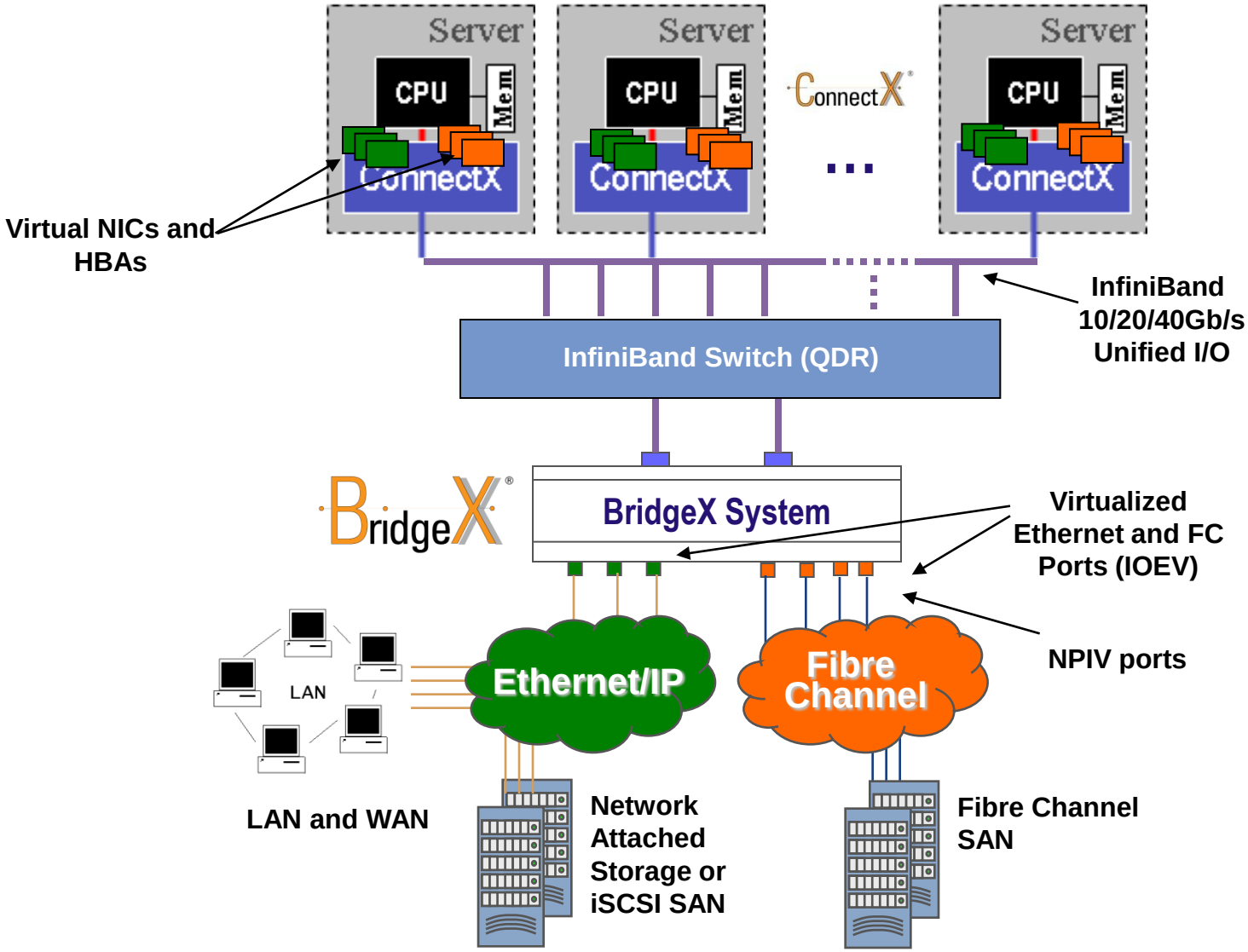
Xsigo 導入後

6 ケーブル

IO統合によるROI（サーバ40台のケース）

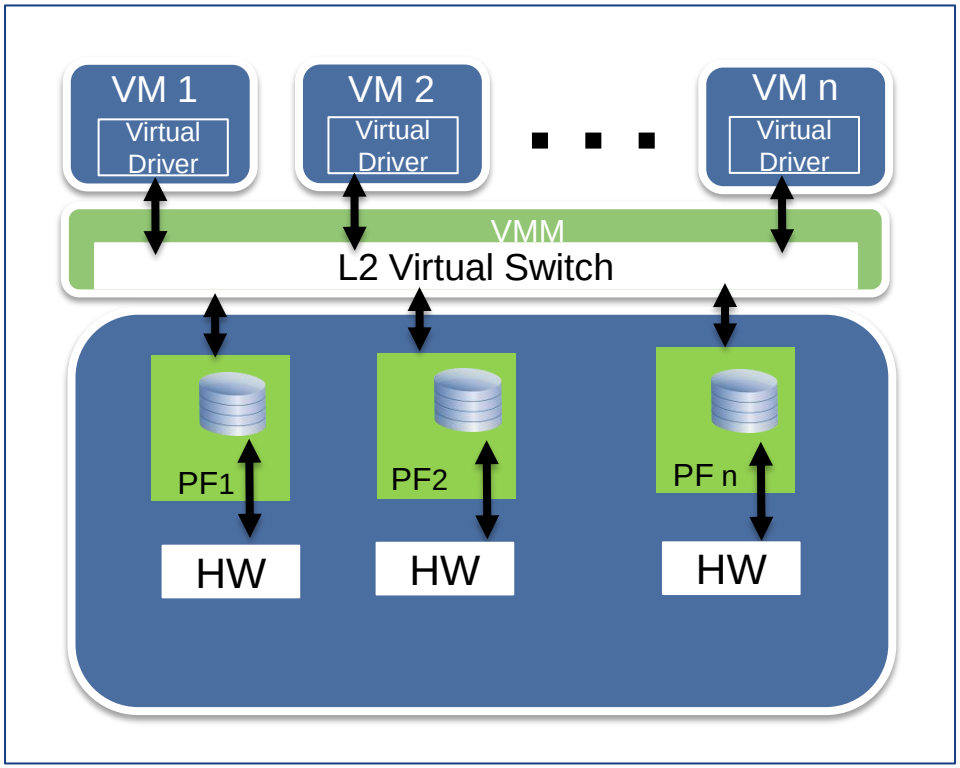


Mellanox BridgeX によるIO統合イメージ



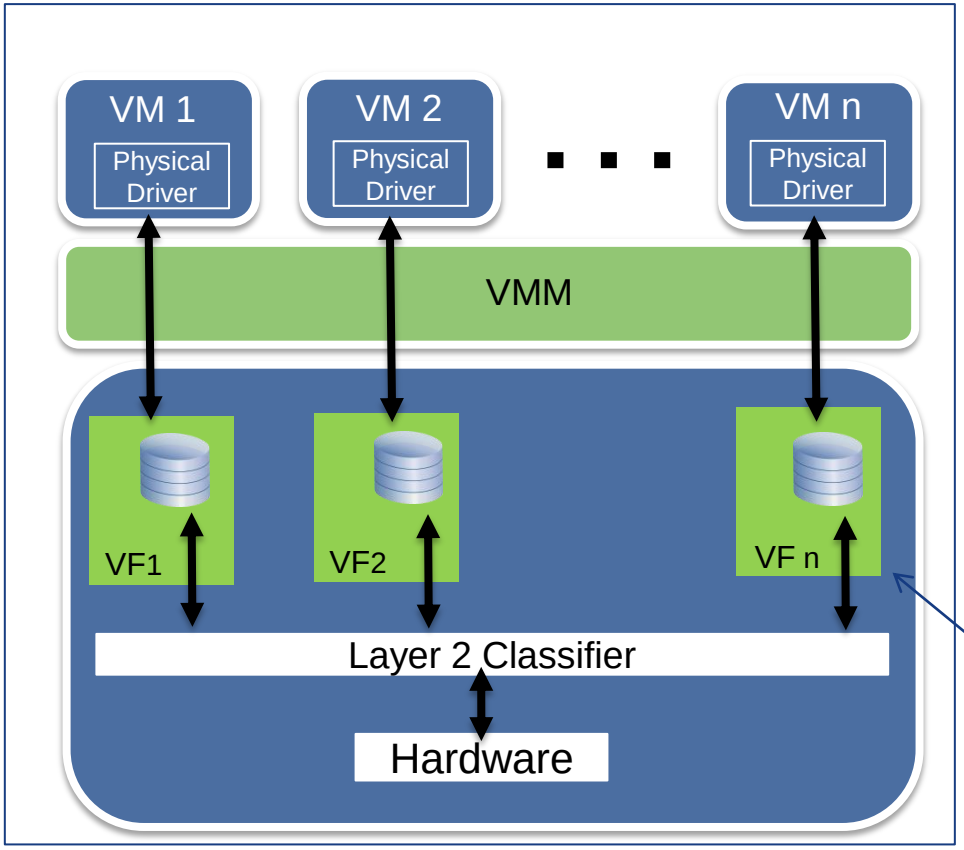
Para-Virtualizationとは？

- シームレスにインフィニバンドを導入できる簡易型ソリューション
  - バーチャルマシンモニタ(VMM)がダイナミックにソフトウェアをエミュレーションすることにより、HCAなどのハードウェアリソースを複数の仮想マシンが共有できる機能
    - Mellanox Para-Virtualization DriverがVMごとにVNICを設定しHypervisorのvSwitchと連携、vNICとその先にあるVMのMACアドレスのみ学習、それ以外のアドレス管理はBridge-Xの外側に接続されるEthernet Switchが管理。Ethernetの代替ネットワークとして簡単・低価格にIO統合を実現



SR-IOVとは？

- PF とVFを組み合せることで、一つのIOデバイスを各仮想マシンに対して 複数の独立した物理デバイスのように利用できる環境を実現
- VMMのオーバーヘッドを解消し、低いCPU利用率でIOの高速化を実現できるため、仮想マシンの統合率をさらに向上



Two new function types that I/O adapters must support

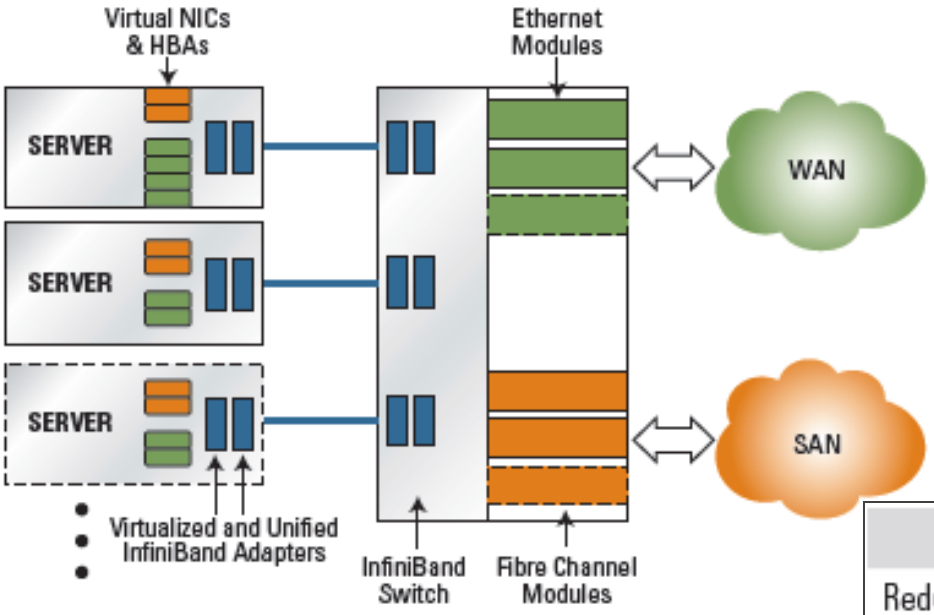
PFs: Physical Functions:
Full PCIe functions that include the SR-IOV Extended Capability

VFs: Virtual Functions:
Lightweight' PCIe functions that contain the resources necessary for data movement but have a carefully minimized set of configuration resources

vSwitchを介さずに物理IOデバイスのVFが直接各仮想マシンにアロケーションされることにより、インフィニバンドの40Gbpsのバンド幅と1μsの低レイテンシーをフルに活用することが可能に！

Virtual Functions

Case Study: Travel/Transportation



Web 2.0

NAVITAIRE

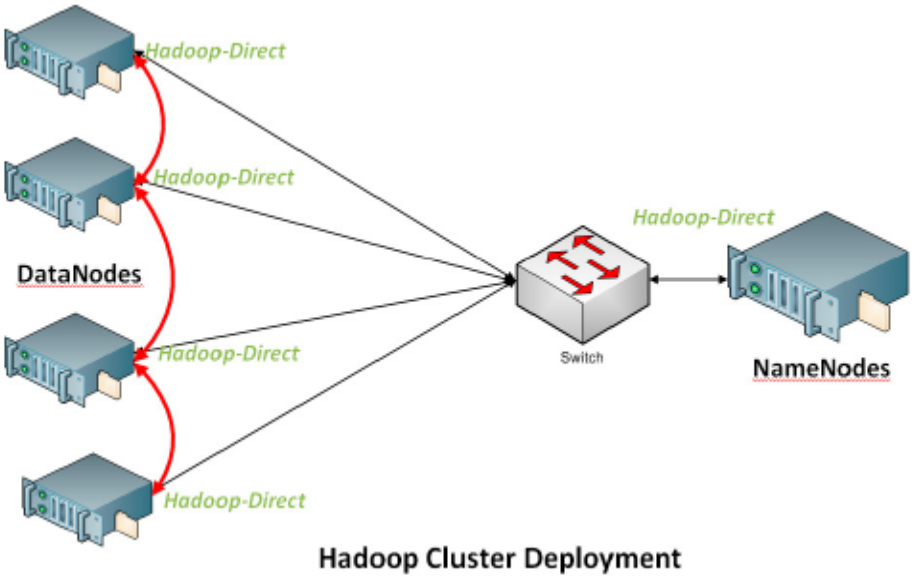
Managed hosting server provisioning time down from 216 hours to 12 hours improving SLA

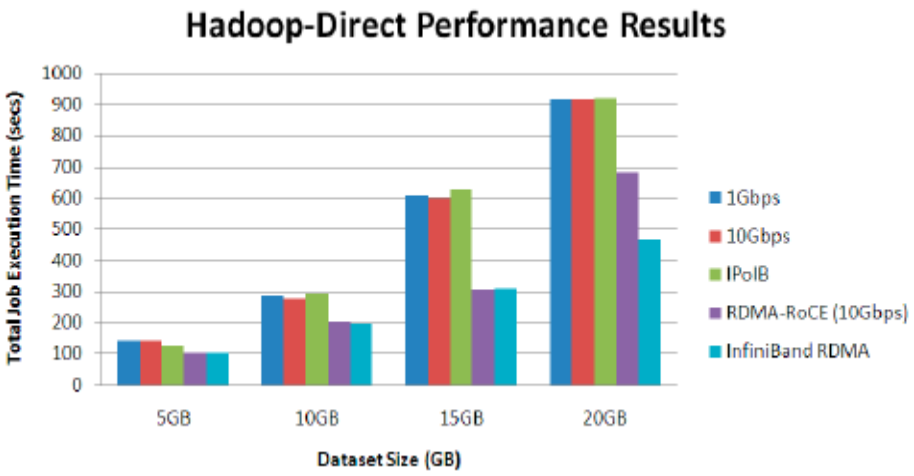
Lower Cap Ex, Op Ex through I/O consolidation

Objective	Benefit Achieved
Reduce the number of physical servers	The number of physical servers dropped from nearly 2,000 to only 150
Decrease server provisioning time	Average server provisioning time down from 216 hours to 12 hours
Simplify cabling complexity	Cabling complexity simplified by 70 percent
Lower I/O capital expenditure	Capital expenditures down 50 percent
Reduce I/O power consumption	Power consumption reduced 30–40 percent
Achieve ROI for the new I/O infrastructure	Immediate ROI realized through capital, labor and power savings

その他のインフィニバンド使用例 (Hadoop Direct)

- Hadoop Direct





- ソフトウェアプラグインとなる Mellanox Hadoop-Direct は、Hadoop ネットワークを加速し、アプリケーションに注力するデータ解析用の Hadoop クラスタの拡張性を改善します。効率の良い Merge-sort アルゴリズムとこの RDMA を使用した斬新なデータ移動のプロトコルは、Mellanox の Infiniband と 10GigE RoCE (RDMA over Converged Ethernet) アダプタカードを搭載した Hadoop クラスタに対して、Hadoop フレームワークを加速させ、サーバー間で効果的にデータを移動させることを可能にします。これは優れたロスレスファブリックを実現する Infiniband と RoCE を活用し、クラスタ観点から Hadoop の帯域と演算の問題に取り組んだ初めてのネットワークキングソリューションとなります。Hadoop-Direct は Hadoop ユーザーに対して透過的であり、現在のアプリケーションを変更することなくそのまま実行することができます。
- Hadoop-Direct はデータ処理のスループットを2倍以上に向上させ、また全体のジョブ実行時間を Hadoop ノードごとに半分に減少させます。より短いジョブ実行時間によるノードあたりの CPU 効率の改善によって、クラスタはより少ないサーバー数で要求された性能を満たすことができ、その結果データセンターの消費電力を直接的に削減しデータセンターの環境への取り組みに貢献いたします。

まとめ

- インフィニバンドは最高速ネットワーク
  - 56Gbpsの転送レート＋1μs以下の低レイテンシー
  - RDMA、Transport Off-load、Kernel BypassなどによるCPU負荷の劇的な軽減
  - 様々なオフロード機能：Core Direct/FCA、GPU Direct、Hadoop Direct
- HPC向けだけでは無い、幅広い規格
  - ノード間　：　IPC/MPI、RDS、SDP、IPoIB、EoIB、RoCE
  - ファイルアクセス　：　SRP、iSER、Luster、FCoIB、FCoE
  - 仮想化対応　：　SR-IOV、Para-Virtualization、vmDirectPathIO
- 拡大する利用用途
  - HPC:　大規模科学計算の分野から構造解析、画像処理など小規模クラスタへ
  - 用途：　DBクラスタ、Hadoop、GemFire、IN-bound/Out-bound Storage System
  - 市場：　金融、クラウドコンピューティング、Web2.0、グラフィック