Mellanox社とInfiniband概要

2010年9月16日

Mellanox Technologies Inc.
津村 英樹

Mellanox
TECHNOLOGIES

Mellanox
TECHNOLOGIES
~10 Years~
Efficiency, Scalability, Performance

会社概要

ティッカー: MLNX NASDAQ® TEL-AVIV STOCK EXCHANGE Mellanox TECHNOLOGIES

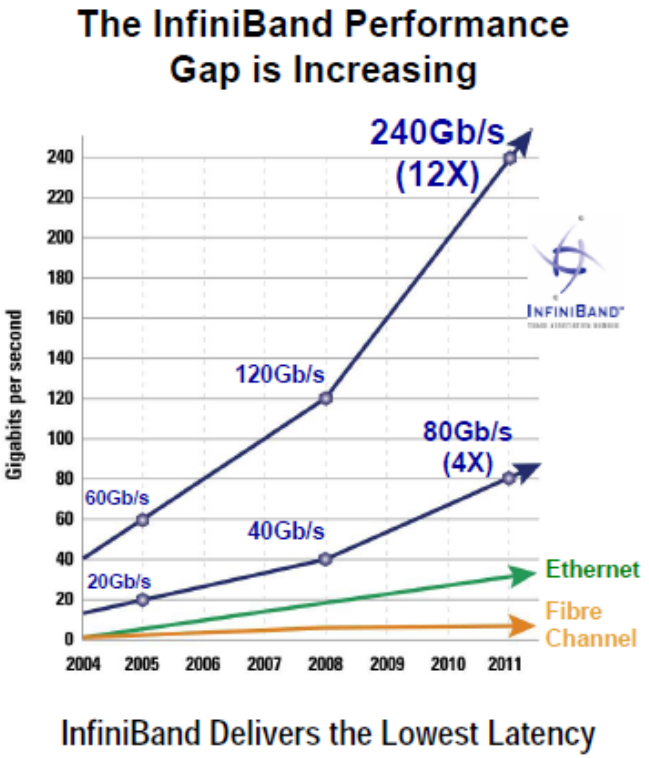
- **HPC、金融、クラウド市場の要求を満たす最高性能のITシステムを実現するソリューションプロバイダー**
  - Chip、HCA、スイッチなどEnd-Endソリューションを提供するインフィニバンド/10GbEther市場のリーディングカンパニー
  - 2010年6月時点で、620万ポート以上を出荷
- **本社:**
  - イスラエル ヨークナム, カリフォルニア州サニーベール
  - 400人以上の従業員; ワールドワイドのセールス＆サポート
- **揺るぎのない財務状況**
  - 2010年Q2の記録的な売上高； $40.0M
  - 2009年会計年度の記録的な売上高 = $116.0M
  - $233.9Mのキャッシュ/借入れなし

最近の受賞

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、8B/10B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　：　40Gbps (QDR 4x)
  - Switch-Switch　：　120Gbps (QDR 12x)
  - 1μsのアプリケーション・レーテンシー
- **高い伝送効率を実現するチャネルアーキテクチャ**
  - MPI（メッセージ・パッシング・インタフェース）
  - RDMAとTransport Offload
  - Kernel Bypass

＜メッセージ・パッシング：補足説明＞
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。



補足説明

RDMA（Remote Direct Memory Access）

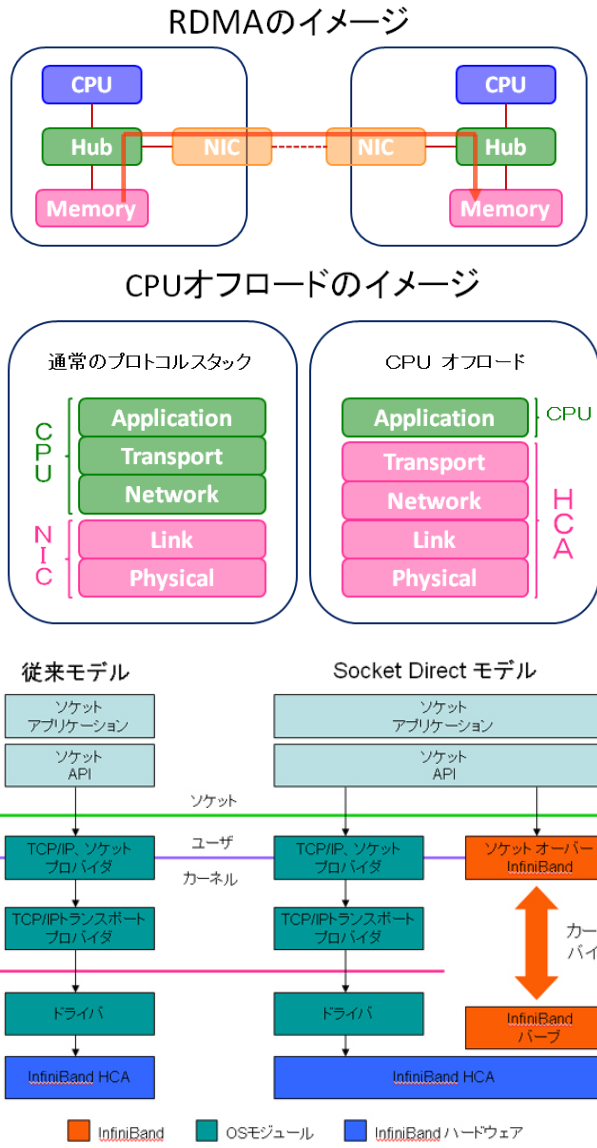
- 独立したシステムのメインメモリ間、システムとストレージ間の　データ転送を効率化するテクノロジ。メモリ間のデータ転送時にはホストCPUの割り込みがほとんど発生しないため、ホストCPUの　オーバーヘッド低下に加え、データ転送時のレイテンシ短縮にもつながる。
- 10Gbit Ethernetを利用する場合、システム構築や運用管理がInfiniBandよりも容易という利点を持つ。Ethernet上でRDMA機能を実現するRDMA over EthernetもMellanox社HCAではサポートしており、non-HPC市場での導入が進んでいる。

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。

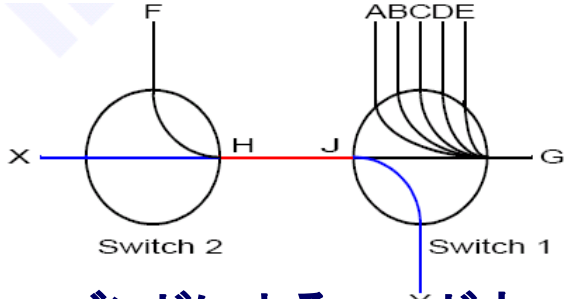
カーネル・バイパス

- 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 TCA（Target 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に見せながらも、同時にInfiniBandの利点を享受できる。

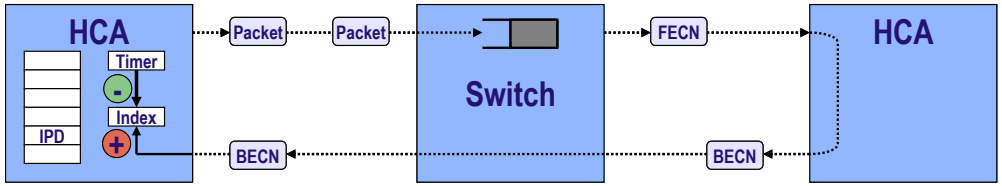


ハードウェアによるコンジェッション・コントロール

- **コンジェッション・スポット → 破壊的なスループット・ロス**
  - 今日の要求に古い技術では適応できない

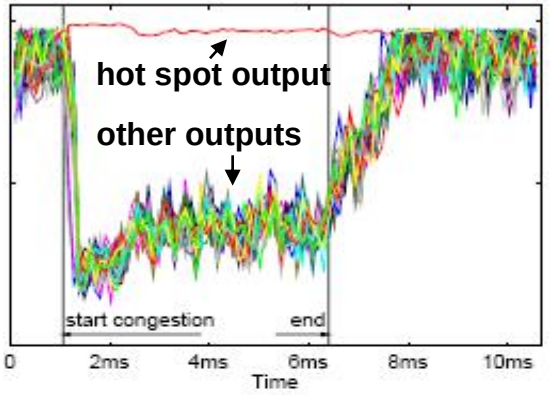


- **インフィニバンドによるハードウェアベースのコンジェッション・コントロール**
  - 情報の流れる経路を事前に規定する必要なし
  - ホットスポットの自動検知
  - データトラフィックの調整
  - 帯域変動などの副作用なし
  - コンジェッション通知
- **最大の実効帯域を確保**

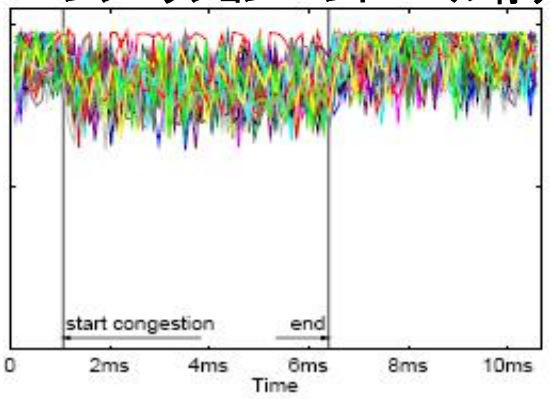


シミュレーション結果
32-port 3 stage fat-tree network
High input load, large hot spot degree

コンジェッション・コントロール無し



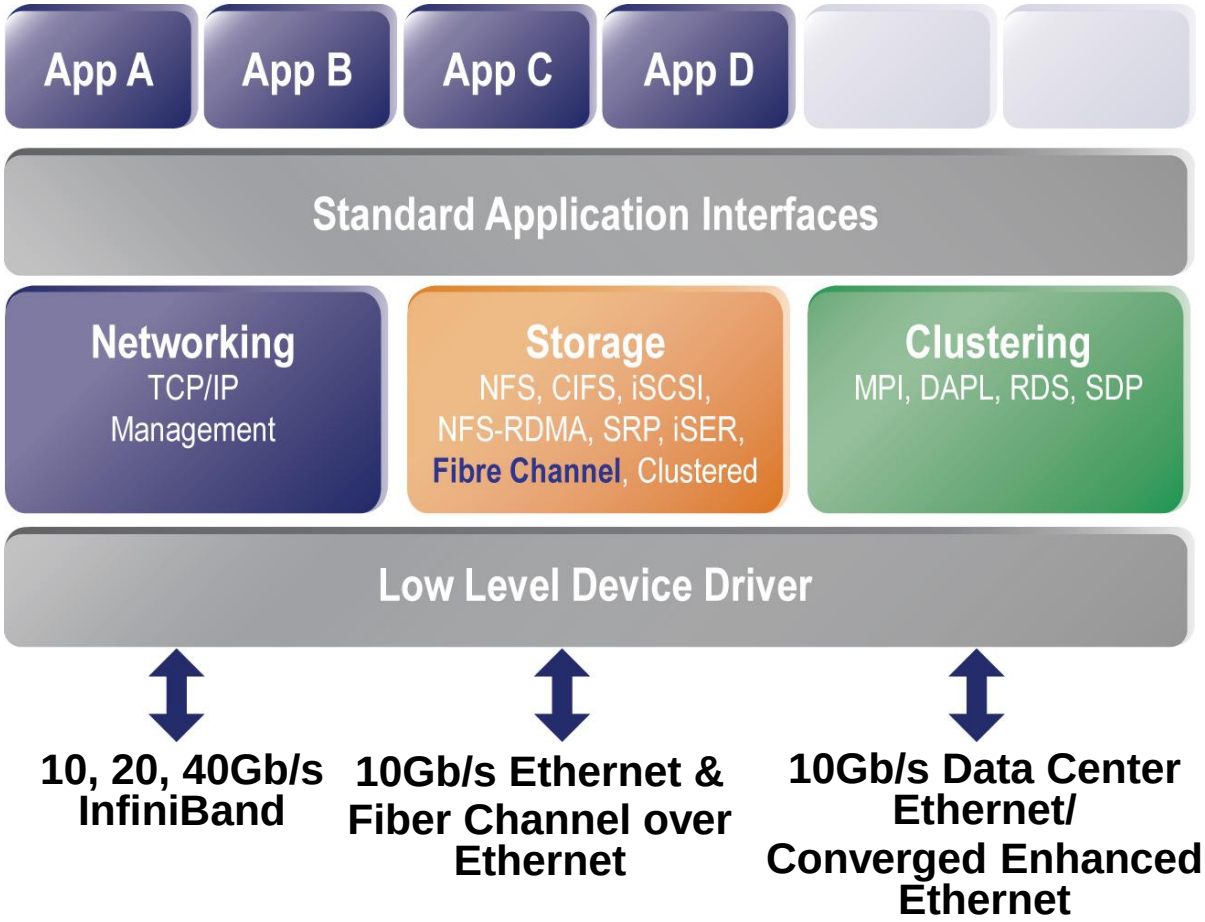
コンジェッション・コントロール有り



"Solving Hot Spot Contention Using InfiniBand Architecture Congestion Control*IBM Research; IBM Systems and Technology Group; Technical University of Valencia, Spain*

将来性を考慮した理想的なアーキテクチャ

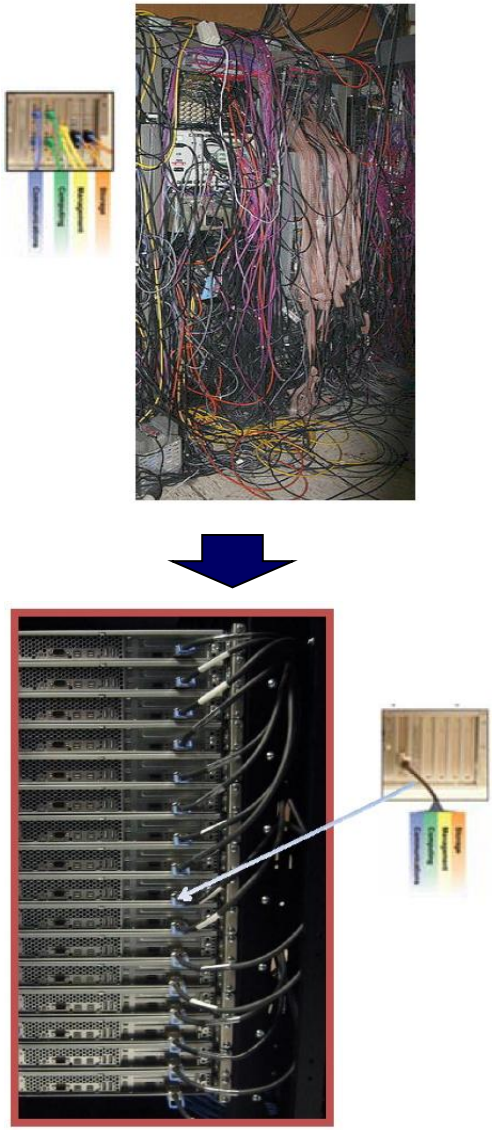
完全なアプリケーションの透過性



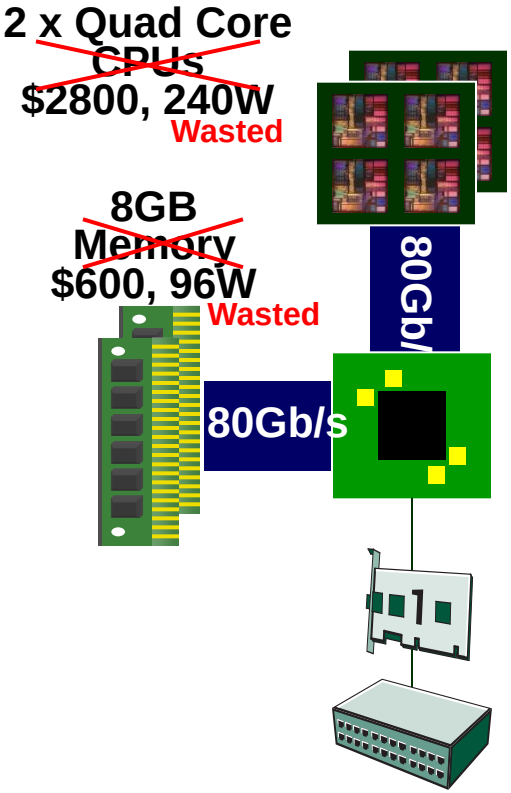
全てのインターコネクトに対する統合化されたソフトウェア・スタック

インフィニバンドによる問題解決

- **データセンター環境の問題点**
  - サーバ増加による消費電力の急増
  - サーバ仮想化の普及に伴うI/Oボトルネック
  - ケーブル量増加に伴う配線の複雑化
- **アプリケーション性能の向上**
  - ストレージサーバやファイルシステムの応答速度
  - データベースの高速化
  - アプリケーション・レイテンシの改善
- **高まる次世代ネートワークへの期待**
  - シンプルで高性能なConverged Network
  - 高帯域と低レイテンシを併せ持つNetwork
  - 低消費電力で低価格なNetwork

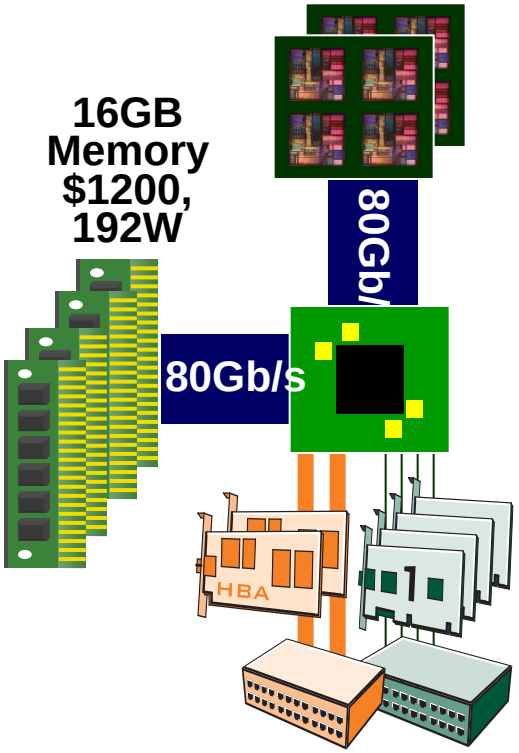


ConnectX バーチャルプロトコルインタフェース



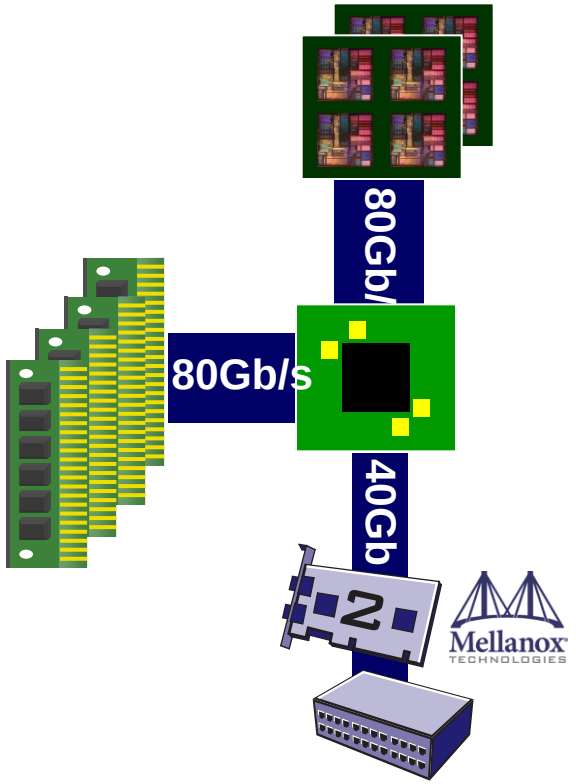
1GigE, 50µs, $100, 4W

- I/O性能の不足
  CPUとMemoryはアイドル状態
- 電力と性能の無駄



12Gb/s, 50µs, $2400, 28W

- 仮想化によるI/Oの増強
- 消費電力、ケーブリング、複雑性の増加 ＝ インターコネクト投資増加

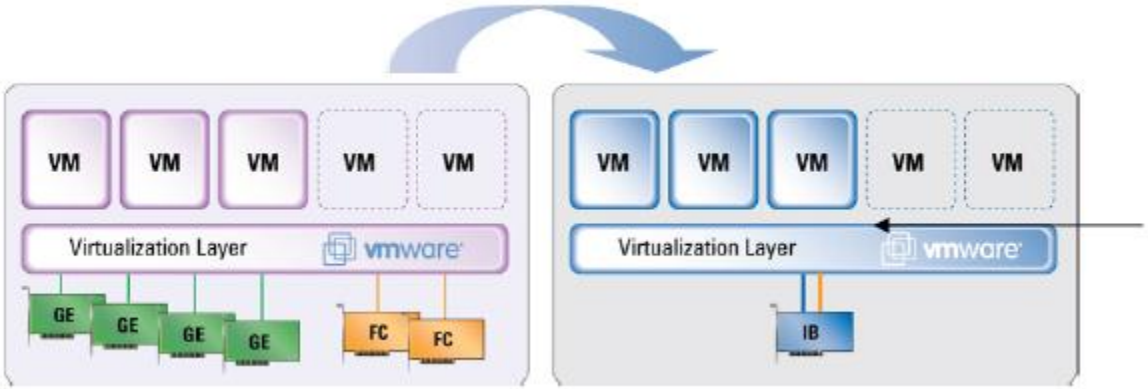


10-40Gb/s, 1µs, $900, 10W

- 統合化されたI/O
- 消費電力やコストの最適化
- ケーブリング、複雑性の解消

仮想化とUnified I/O

- **仮想化環境における統合率の向上**
  - 物理サーバ1台あたり30台以上の仮想サーバが搭載可能に
  - 統合率向上を阻むI/Oボトルネックの解消
  - 管理効率を向上するネットワークの統合
- **インフィニバンドを活用したUnified I/O Fabric**
  - 40Gbpsの高いバンド幅＋圧倒的に低いレイテンシ
  - 仮想環境の変更の必要がない透過的なソリューション
  - EthernetやFibre Channelとの接続実績



インフィニバンドの効果

- **I/Oコストの削減**
- **I/O消費電力の削減**
- **SAN性能が4GbpsFCの4倍**
- **LAN性能が10GbEの3-5倍**

効率的なコンピューティングのための接続性ソリューション

クラウド コンピューティング

ハイパフォーマンス
コンピューティング

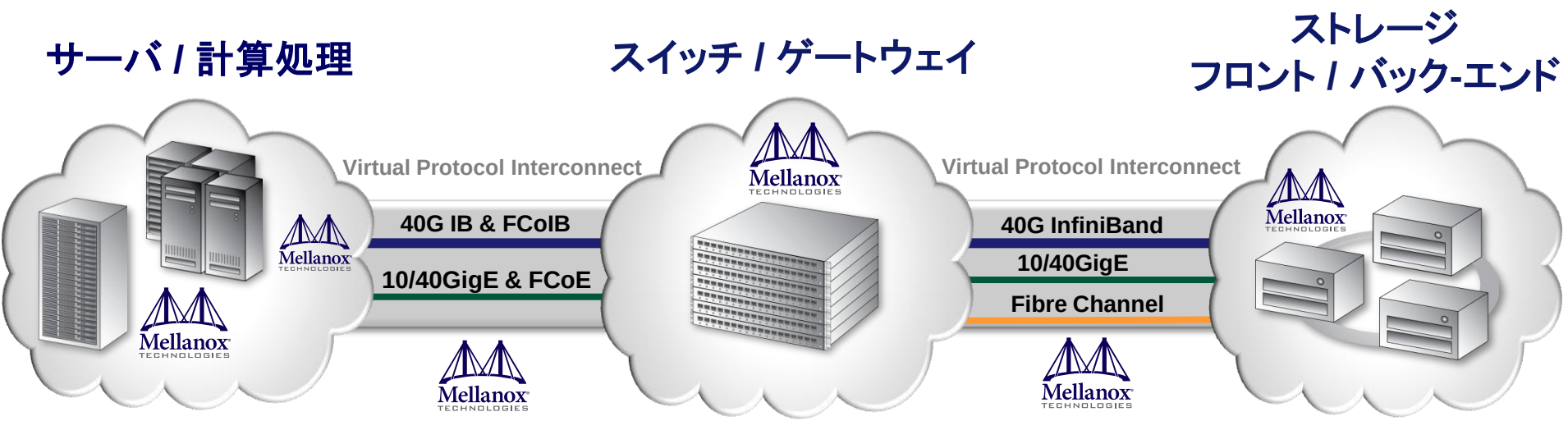
企業とストレージ

クラス最高のネットワーク帯域幅、アプリケーション性能、
そして応答時間をユーザに提供

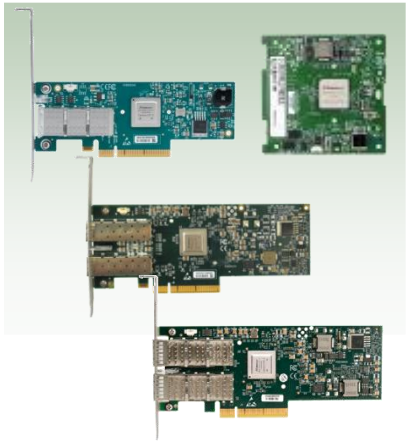
Mellanox製品概要

2010年 9月16日

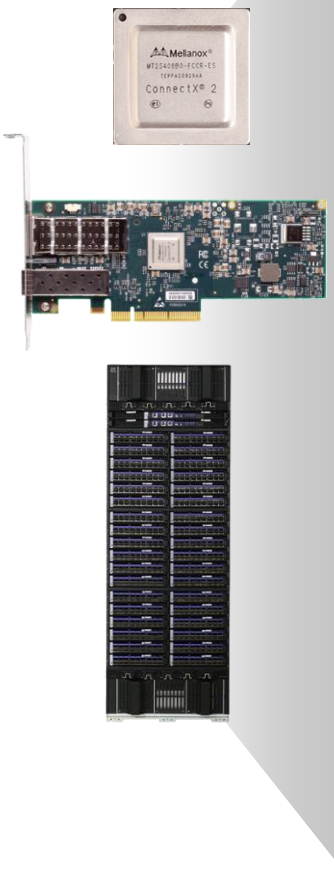
エンド･ツゥ･エンドの接続性ソリューション プロバイダー



Mellanox 相互接続ネットーワーキングソリューション

IC	アダプタ カード	ホスト/ ファブリック ソフトウェア	スイッチ / ゲートウェイ	ケーブル
				

Mellanoxがつなぐデータセンターエコシステム



ハードウェア OEM

サーバ

HP, IBM, Sun, ORACLE, DELL, NEC, FUJITSU, 曙光 DAWNING, SUPERMICRO www.supermicro.com, BULL

ストレージ

NetApp, HP, IBM, Sun, ORACLE, DataDirect NETWORKS, LSI, ISILON SYSTEMS

組み込み用途

GE FANUC, DALSA, rohati, Computer Systems MERCURY, SeaChange

ソフトウェア パートナー

Microsoft, redhat, Novell, SOLARIS, AIX L, HP-UX 11i, vmware, ORACLE, IBM DB2, REUTERS, ANSYS, SYNOPSYS, lustre, Autodesk

エンドユーザ

エンタープライズデータセンター

cme Chicago Mercantile Exchange, JPMorgan, Burlington Coat Factory, MTV MUSIC TELEVISION, HBO IT'S NOT TV IT'S HBO!, VIACOM, KELLY SERVICES, BT, YAHOO!, Alexa The Web Information Company

ハイパフォーマンス コンピューティング

HONDA, Chevron, CITGO, TOYOTA, NCSA, DREAMWORKS ANIMATION SKG, EXXON, AIRBUS, Ford

組み込み用途

COX COMMUNICATIONS, CABLEVISION, comcast, TimeWarner

HPC とデータセンターに不可欠なMellanox

スケール-アウト クラスタリング

データウェアハウジング

金融サービス

クラウド コンピューティング

Web 2.0

- **増大するネットワーキングI/Oの要求**
  - マルチコア CPU、GPU、仮想化の市場普及に伴い増大
  - LAN、SAN、IPC I/O 統合
- **サーバとストレージの効率性と拡張性**
  - それぞれのサーバの使用を最大化、性能保証、SLAの満足
  - ユーザとデータボリュームの指数関数的増大に弾力的に対応するストレージの拡張性
- **広帯域で低レイテンシI/Oは、ROIに重要**
  - データスループット、アプリケーション応答、そして、ジョブの実行時間を加速し、利用を増大

Mellanox Network Connectivity Benefits for IT*

* Based on end-users testimonies

インフラの削減 60%

エネルギー費用の削減 65%

性能の増大 10X

一枚のアダプターで実現:
Virtual Protocol Interconnect (VPI)

- **広範なOS / 仮想化をサポート**
  - 強力なソフトウェアエコシステムの礎
- **統合 / 豊富な接続性オプションと機能**
  - InfiniBand/FCoIB あるいはEthernet/FCoEを介してのコスト効率に優れた統合
- **性能**
  - アプリケーションの加速化、PCIe 2.0、低レイテンシ、広帯域

オンデマンドの性能とサービス

業界を牽引する InfiniBand アダプタ

- **InfiniBandの市場とパフォーマンスを主導**
  - 40Gb/sアダプタを、最初に市場投入
    - 2011年には、100Gb/s アダプタへのロードマップ
  - 重要！次世代の処理効率の改善
    - GPU-Direct™
      - GPUベースのクラスタのアプリケーション性能を改善
    - CORE-Direct™
      - ハードウェアで、アプリケーション通信をオフロード
  - 40Gb/s InfiniBand の好調な採用動向
  - HPC Top500 システムが100% IBを採用

(Worldwide Tier-1 Server OEM Availability)

BULL 曙光 DAWNING DELL FUJITSU hp invent IBM Sun microsystems PLATFORMS

ITBRAND MARKET LEADER InfiniBand Adapters November 2009

HPC wire Editors' Choice Awards 2009

市場の広がりとリーダーシップ:
10/40 ギガビットイーサネットアダプタ

- **業界初の**
  - デュアルポート PCIe 2.0 10GigE アダプタ
  - ハードウェア・オフロード機能つき、FCoE対応10/40GigE
- **イーサネット業界で、最も低いレイテンシ**
  - 1.3µs のエンド-ツゥ-エンド・レイテンシ
  - より効率的なサーバ利用と高速なアプリケーション処理を可能に
- **驚異的なエコシステムサポートの推進力**
  - 複数のTier-1 OEMでの採用実績
    - ～10% のワールドワイド市場のシェアと成長
    - サーバ、マザーボード上のLAN (LOM)、およびストレージシステム
  - VMware vSphere 4 Ready
  - Citrix XenServer 4.1 in-the-box support
  - Windows Server 2003 & 2008, RedHat 5, SLES 11

CITRIX® Certified for Windows Server 2008 Certified for Microsoft® Windows Server® 2003 Novell. redhat. vmware®

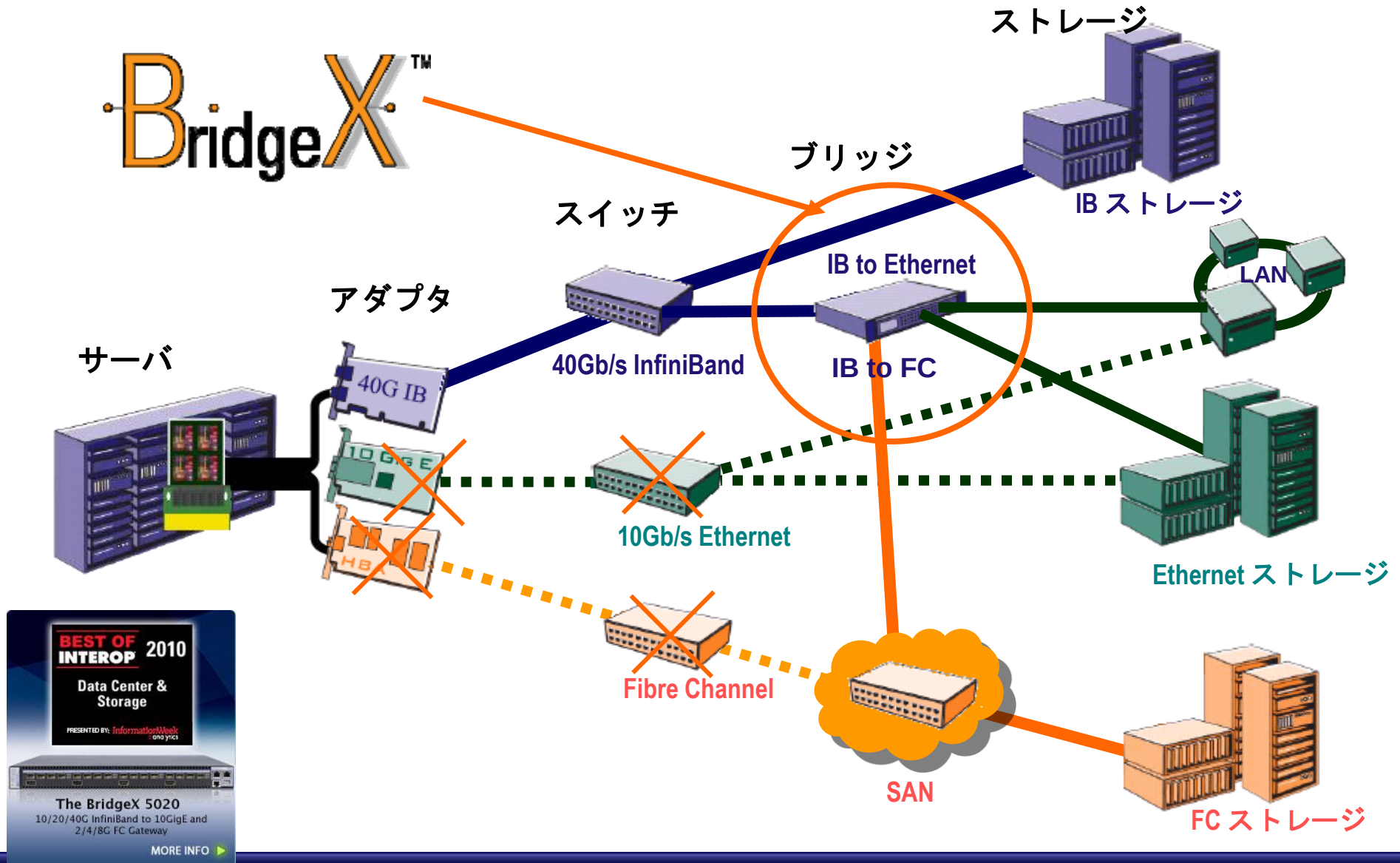
パフォーマンスを牽引するスイッチとゲートウェイ

- **InfiniBand 市場とパフォーマンスのリーダー**
  - 51.8TBで、業界最高密度のスイッチ
    - 包括的なファブリック管理ソフトウェアと共に
  - 世界で最も低いポート間レイテンシ
    - 競合よりも、25 - 50% 低い
  - 競合よりも最大4倍高い拡張性
  - 先進的なネットワーキング機能
    - 輻輳管理、QoS、ダイナミックルーティング
- **最高性能を誇る統合 I/O ゲートウェイ**
  - 最適な拡張性、統合、省電力
  - より低いスペースと電力そして、増加したアプリケーション性能
    - 3.5倍のデータ容量、7倍の性能

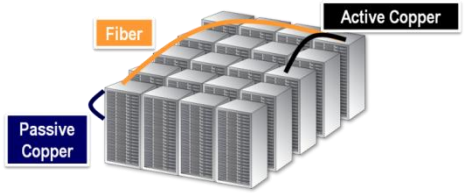
FabricIT Management Suite

- **Management (CLI, WebUI) ユニファイドアクセス**
  - RS232 コンソール (CLI のみ)
  - 10/100 管理ポート
  - IPoIB インバンド インターフェイス
- **FabricIT Chassis Manager (SCM)**
  - シャーシ管理: センサー読み込み、警告、ファームウェア更新、カウンター読み込み
- **FabricIT Fabric Manager (EFM)**
  - SM、診断、適応型ルーティング&輻輳管理、クラスタ診断
  - アップグレード可能な発注オプション（ライセンス）

DCネットワーク – IBゲートウェイを介して集束



カッパー（銅線）と光 ケーブル・ソリューション



- **完璧な製品ラインナップ**
  - パッシブと、最大12mのアクティブ カッパーケーブル
  - 最大300mのアッセンブル済みの光ケーブル
- **全数検証**
  - IBTAのメカニカル及び、エレクトロニクス基準を上回る品質
  - MellanoxのスイッチとHCAで検証済み
- **QSFP To SFP+ アダプタ (QSA)**
  - QSFPからSFP+変換に挑む、世界初のソリューション
  - 10GigE、40GigE、そして40Gb/s InfiniBand間のスムーズで費用対効果に優れた接続
  - 将来を見据え、今日の40Gb/s エコシステムを強化

Thank You