# InfiniBand ネットワークの IA サーバにおける実用性の検証

## Verification of Utility of InfiniBand Network for IA Servers

本　間　一　久

**要　約**　コンピュータを使った処理は，ビッグデータを扱うようになり大量のデータを高速に処理することが求められてきている．サーバについては，CPU の処理性能がムーアの法則に倣って向上している一方，IO の速度がボトルネックになる傾向がある．これらを解決する手段として InfiniBand がある．InfiniBand は長らく HPC（High Performance Computing）専用と考えられてきたが，IA（Intel Architecture）サーバでも，ハードウェアの進化と Windows Server 2012 のリリースにより，商用環境に取り入れる条件が整った．

本稿では，IA サーバにおける InfiniBand のネットワーク技術に焦点をあてる．最新の IA サーバを用意して，理論値に対してどの程度の性能が実測されるか確認するとともに，10Gb イーサネットとも比較し，InfiniBand の優位性を確認した．可用性の実機検証も行い，注意点を明確にすることで，最適な実装方法を確認することができた．InfiniBand の現状を明らかにし，IA サーバ環境での実用性を検証した結果を報告する．

**Abstract**　High-speed processing of large volume data is required for the computer processing as it handles the Big Data nowadays. As for a server, while the processing power of the CPU is improved by the Moore's law, there is a tendency that the rate of the IO becomes a bottleneck. InfiniBand can be used as a mean to solve these problems. InfiniBand has been considered exclusively for HPC (High Performance Computing) long time, but the evolution of hardware and the release of Windows Server 2012 enabled the installation of InfiniBand into the IA (Intel Architecture) server for commercial environment as a solution.

This paper focuses on the network technology of InfiniBand in the IA server. With brand-new model of IA servers, we measured and verified the performances by using InfiniBand HCAs, and also compared with 10Gb Ethernet to confirm the superiority of InfiniBand. Also an actual verification of availability was performed to clarify the considerations for the implementation to determine the optimal implementation. This paper reports the verification results to clarify the current status of InfiniBand and its usefulness in the IA server environment.

## 1. は　じ　め　に

コンピュータを使った処理は，ビッグデータを扱うようになり大量のデータを高速に処理することが求められてきている．サーバについては，CPU の処理性能がムーアの法則に倣って向上している一方，IO の速度がボトルネックになる傾向がある．既に HPC（High Performance Computing）の分野では，科学技術計算や生物構造の解析など，大量のデータを高速処理するために，解決手段として InfiniBand が採用されている．InfiniBand は，ネットワークとストレージ両方の機能を提供する仕組みが備わっており，サーバとストレージ間の接続や，データベースクラスタ構成で実装されている．日本ユニシスの大型汎用機 CS6200L シリーズ，中型汎用機 CS4200L シリーズはプロセッサと IO 間のインターコネクトに InfiniBand を

採用している．オラクル社の Exadata（データベースに特化したアプライアンス製品）でも，サーバとストレージ間の通信インターコネクトに InfiniBand を採用している．

InfiniBand は 2000 年に IBTA（InfiniBand Trade Association）により規格が発表されて以来，長らく HPC 専用と考えられてきた．IA（Intel Architecture）サーバの分野でも製品はリリースされてきたが，製品を提供するベンダも限られていたため，一般に広く用いられる環境は整っていなかった．しかし，2012 年に Intel 社が Qlogic 社から InfiniBand 事業を買収したことや，PCI バスなどのハードウェアのアーキテクチャーの進化，ならびに Microsoft 社の Windows Server 2012 のリリース等をきっかけに，IA サーバ分野でも商用環境で使用できる可能性が出てきた．

本稿では，IA サーバにおける InfiniBand のネットワーク技術に焦点をあてる．2 章にて，InfiniBand になぜ今注目すべきか，現状を明らかにする．3 章では，性能測定を行い理論値に対してどの程度の性能が測定されるか確認するとともに，10Gb イーサネットとも比較し，InfiniBand の優位性を確認した結果を報告する．また可用性の実機検証も行い，実装における注意点を明確にする．4 章では IA サーバ環境での最適な実装方法を報告する．

## 2. InfiniBand を取り巻く環境

本章では，IA サーバにおける InfiniBand を取り巻く環境を述べる．

### 2. 1 InfiniBand の特徴

InfiniBand の接続形態はポイント・ツー・ポイントの双方向シリアル接続である．複数の転送レートをサポートするほか，高速化のために PCI Express のように複数のレーンを束ねて利用することができる．束ねるレーン数に応じて 1X，4X，8X，12X が規定されている．InfiniBand ネットワークは，サーバの PCI Express の拡張スロットに搭載する HCA（Host Channel Adapter）により接続されるが，2014 年 8 月現在，4x の HCA がリリースされている．

InfiniBand の大きな特徴は，RDMA（Remote Direct Memory Access）の機能を備えていることである．RDMA はつながる 2 台のコンピュータのメモリ上のデータを，CPU を介さず転送する技術である．図 1 に示すとおり，RDMA を使用すると System Buffer への一時的なデータコピーを行わず，つながるコンピュータのメモリ（User Buffer）と直接データを送受信することができる．これによって，システムバス，CPU の負荷も低く抑えることができる．RDMA は HCA の機能として提供され，2014 年 8 月時点で，市場で販売されている HCA はこの機能を搭載している．

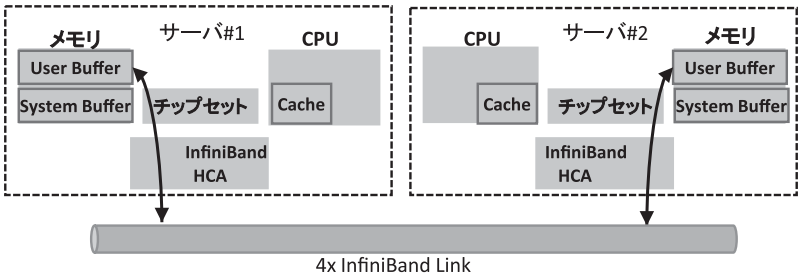


図 1　RDMA のデータ処理

### 2. 2　InfiniBand の規格

InfiniBand の規格は，業界団体である IBTA によって策定されている．表 1 のとおり，複数の規格がありそれぞれ帯域幅が異なる．2014 年 8 月現在，最新の規格は EDR（Enhanced Data Rate）である．EDR はスイッチ製品のリリースはされているが，まだ HCA はリリースされておらず，販売の主流は FDR（Fourteen Data Rate）である．EDR に続き，HCA（x4）の帯域幅が 500Gbps になるとみられている HDR（High Data Rate）の規格のリリースが，2017 年に計画されている．このようにこれまで SDR（Single Data Rate），DDR（Double Data Rate），QDR（Quad Data Rate），FDR と規格策定後製品がリリースされ，今後 HDR までは継続的に性能の向上が図られている．

InfiniBand は規格によりデータ通信符号化の方式が異なる．SDR，DDR，QDR は 8b/10b 変換（コンピュータ内部で扱うデータは 8bit だが，外部との通信の際は，誤り訂正符号を付加するために 10bit となる）を用い，FDR 以降は 64b/66b 変換を用いる．つまり，SDR，DDR，QDR の通信データに占める有効データの比率が 80％なのに対し，FDR，EDR は 97％と高く，ロスが少ない．FDR の実質的な帯域幅は，54.3Gbps と高性能を実現することができる．

**表 1　InfiniBand の規格と帯域**

| | SDR | DDR | QDR | FDR | EDR | HDR |
|---|---|---|---|---|---|---|
| リリース年 | 2000 年 | 2006 年 | 2009 年 | 2011 年 | 2014 年 | 2017 年（予定） |
| 1 レーンの帯域幅（Gbps） | 2.5 | 5 | 10 | 14 | 26 | 125（予定） |
| HCA（4X）の帯域幅（Gbps） | 10 | 20 | 40 | 56 | 104 | 500 |
| HCA の実質的な帯域幅（Gbps） | 8 | 16 | 32 | 54.3 | 100.8 | 484.8 |

※ EDR，HDR の HCA はまだリリースされておらず，実質的な帯域幅は 64b/66b 変換を用いた場合を仮定

### 2. 3　価格性能

InfiniBand 製品の性能向上と価格低下が進んでいる．図 2 に InfiniBand システムとイーサネットシステムの価格性能比較を示す．

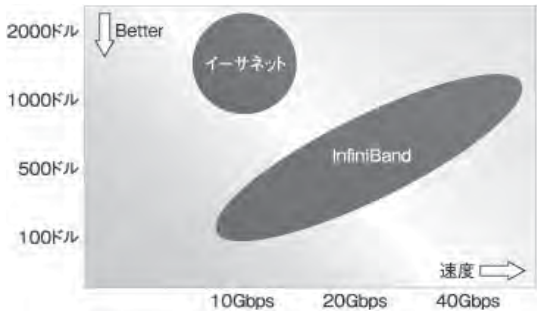


**図 2　InfiniBand システムとイーサネットシステムの価格性能比較**[1]

図2のとおり，InfiniBandを採用したシステムはイーサネットを採用したシステムと比べて価格性能で勝る．その要因としてインターフェースカード単体の価格性能があげられる．表2は，IBM System x3650M4サーバのオプションとして販売されている製品の比較である．40Gb InfiniBand HCAは10Gbイーサネット HBA（Host Bus Adapter）とほぼ同額であり，InfiniBand HCAの価格性能の良さがわかる．

**表2　InfiniBand HCAとイーサネット HBAの価格性能比較（IBM社）**

| | 価格（税込） | 帯域幅 |
|---|---|---|
| InfiniBand HCA<br>Mellanox ConnectX-2 デュアルポート QSFP QDR IB アダプタ | ¥116,000 | 40Gbps |
| イーサネット HBA<br>NetXtreme Ⅱデュアルポート 10GBase-T アダプタ | ¥112,000 | 10Gbps |
| イーサネット HBA<br>インテル X540-T2 デュアルポート 10GBaseT アダプタ | ¥105,000 | 10Gbps |

各システムでは，HCAまたはHBAを2枚，4枚と複数枚で構成することから，その差は広がる傾向にある．またInfiniBandでは，安価な銅線ケーブルを使用することができるため，光ファイバケーブルの採用が多い10Gb以上のイーサネットよりも，ポートあたりの単価を下げられることも要因としてあげられる．

### 2. 4　PCI Express

InfiniBandに注目する理由の一つに，PCI Expressのアーキテクチャーの進化によるサーバのIO性能の向上があげられる．InfiniBandのHCAはサーバのPCI Expressの拡張スロットに搭載する．FDRのHCAは，PCI Express3.0の仕様だが，PCI Express3.0は，2012年春に各社からリリースされたIAサーバで初めて採用された．

PCI Expressには1.1，2.0，3.0の規格があり，1レーンあたりの帯域幅が異なる．また，サーバの拡張スロットは，各メーカ，各機種のスロットごとに，レーン数および帯域幅が異なる．HCA搭載時は，搭載するサーバの拡張スロットの仕様を確認し，PCI Expressのレーン数と1レーンあたりの帯域幅に注意して，PCI Expressの帯域幅がボトルネックとならないよう，適切なスロットを選択する．

**表3　PCI Expressの規格と対応するInfiniBandの規格**

| | | PCI Express 1.1 | | PCI Express 2.0 | PCI Express 3.0 | | |
|---|---|---|---|---|---|---|---|
| | | 2.5 Gbps<br>（1レーンあたり） | | 5 Gbps<br>（1レーンあたり） | 8 Gbps<br>（1レーンあたり） | | |
| | | × 4 | × 8 | × 8 | × 4 | × 8 | × 16 |
| SDR | 10Gbps | ○ | ○ | ○ | ○ | ○ | ○ |
| DDR | 20Gbps | × | ○ | ○ | ○ | ○ | ○ |
| QDR | 40Gbps | × | × | ○ | × | ○ | ○ |
| FDR | 56Gbps | × | × | × | × | ○ | ○ |
| EDR | 104Gbps | × | × | × | × | × | ○ |

※EDRのHCAはまだリリースされておらず，PCI Expressの規格はPCI Express 3.0と仮定している

表 3 は PCI Express の規格に対応する InfiniBand の規格を，ボトルネックが生じない場合を○，生じる場合を×で表したマトリックスである．表 3 の太枠では，FDR の HCA が PCI Express3.0 × 4 のバスにてボトルネックが生じることになる．


**図 3　PCI Express のボトルネック**

図 3 のように，FDR 対応の HCA を PCI Express3.0×4 のスロットに搭載すると，FDR のネットワーク接続ポートの帯域幅 56Gbps に対して，PCI Express の帯域幅は 32Gbps であるため，サーバ間通信では PCI Express がボトルネックとなってしまう．

**表 4　最新の IA サーバが搭載する PCI Express のスロット**[2], [3], [4]

| | Unisys rE5000 RS220HM2 | HP ProLiant DL380p Gen8 | IBM System x 3650M4 |
|---|---|---|---|
| PCI Express 3.0 × 16 スロット | − | 1 スロット | − |
| PCI Express 3.0 × 8 スロット | 2 スロット | 1 スロット | 3 スロット |
| PCI Express 3.0 × 4 スロット | 2 スロット | − | − |
| PCI Express 2.0 × 8 スロット | − | 1 スロット | − |
| PCI Express 2.0 × 4 スロット | − | − | − |
| PCI Express 2.0 × 1 スロット | − | − | − |

表 4 に各メーカの 2014 年 8 月現在最新の IA サーバの PCI Express スロットの数と規格を示す．サーバのメーカ，機種によって異なることがわかる．例えば，Unisys の rE5000 RS220HM2 に FDR の HCA を搭載する場合は，PCI Express3.0×8 スロットに搭載する．PCI Express3.0 × 4 のスロットはボトルネックとなるため使用しないようにする．

### 2. 5　サポート OS

2012 年 9 月に Microsoft 社より Windows Server 2012（以下 Windows2012 と示す）がリリースされ，IA サーバに搭載される OS で初めて InfiniBand が標準サポートされた．Windows 2012 は，標準で HCA のドライバを持つほか，従来から WindowsOS で利用されてきた SMB（Server Message Block）が実装されている．Windows2012 で実装される SMB のバージョンは 3.0 である．この SMB3.0 で初めて実装された SMB Direct により，OS 上で RDMA 機能を持つネットワークアダプタがサポートされるようになった．表 5 のとおり SMB Direct は，SMB3.0 で初めて実装された．

表 5　SMB Direct の対応

| | Windows2008 | Windows2008 R2 | Windows2012 | Windows2012 R2 |
|---|---|---|---|---|
| SMB のバージョン | 2.0 | 2.1 | 3.0 | 3.2 |
| SMB Direct（RDMA 機能） | – | – | ○ | ○ |

## 3.　実用性の検証

本章では，実装方法，性能，可用性の三つの観点で，実機検証を行った結果を報告する．

### 3. 1　Windows2012 への実装

ここでは Windows2012 で，InfiniBand を効率的かつ効果的に機能させるための方法を述べる．

#### 3. 1. 1　容易なドライバの適用とネットワーク設定

Windows2012 は，HCA のドライバ WinOF（Open Fabric Enterprise Distribution for Windows）を標準で搭載している．HCA は Windows2012 により自動認識され，このドライバが自動で適用され RDMA に対応する．ただし，イーサネットなどの一般的なドライバと同様，OS のリリース時期により最新のドライバが適用されないため，HCA を提供しているメーカがリリースする最新版のドライバへ更新が必要となる．メーカ提供の最新ドライバには，標準ドライバには含まれない，障害切り分け時に必要なコマンドツールが含まれる．また，機能拡張と，不具合の修正が含まれる．ドライバの適用はイーサネット HBA と手順は変わらず，インストーラーを実行することで容易に行える．

次に InfiniBand ネットワークでは，HCA に IP アドレスの設定が必要だが，設定のための操作方法はイーサネットと同じで，Windows2012 のネットワークと共有センターで行う．日頃 IA サーバを扱っていれば，特別な操作を覚える必要はない．

#### 3. 1. 2　サブネットマネージャの必要性

InfiniBand ネットワーク内では，サブネットマネージャが最低一つ起動している必要がある．サブネットマネージャは，サブネットを管理および制御する．サブネット全体の物理トポロジーを検出し，サーバ間のすべてのパスに関して最短経路を計算したり，サブネット内の構成変更を監視したりしている．サブネットマネージャはサーバ，またはスイッチ上で起動する．スイッチを使用しない構成の場合，Windows2012 では HCA ドライバの適用後に，Program Files にインストールされている opensm.exe を手動で起動し，サービスに登録して使用する必要がある．これに対して，スイッチを使用する構成の場合，スイッチの管理コンソール上でコマンド操作によりサブネットマネージャを起動する．

サブネット上に複数のサブネットマネージャが存在する場合，最も優先度の高い一つのサブネットマネージャが Master，他のサブネットマネージャは Standby となる．Master に問題が発生した時には，Standby の中で優先度が高く設定されているサブネットマネージャが役割を自動的に引き継ぐ．

### 3. 1. 3　SMB マルチチャネルと SMB Direct の確認

SMB マルチチャネル（以下マルチチャネルと示す）は，同時に複数の経路を使用することにより，ネットワーク帯域を拡張し性能を向上させる機能と，ロードバランスや冗長構成を実現し可用性を向上させる機能を提供する．このマルチチャネルの機能と 2. 5 節で述べた SMB Direct は，Windows2012 では初期設定として有効になっているが，その設定を GUI で確認することはできないため，PowerShell で確認する．コマンドを入力することにより簡単に確認できるので，事前に確認することを勧める．無効になっている場合は，PowerShell により有効にすることができる．

## 3. 2　性能測定

Windows2012 をインストールした環境にて，FDR，QDR の性能測定を実施した．理論値から，InfiniBand はイーサネットの性能を上回ると想定し，比較のため 10Gb イーサネットの性能測定も実施した．測定ツールは Iometer を使用し，測定内容はシーケンシャルリードとした．また，性能測定を行うにあたって想定される懸念点を払拭するため，事前にストレージのボトルネックと PCI Express のボトルネックの有無を確認した．

### 3. 2. 1　ストレージのボトルネック

InfiniBand の測定では，ストレージとの IO アクセスがボトルネックになる可能性がある．このため，RAM ディスクを使用することでこの懸念点が払拭できるかどうか予め確認した．

1）　ストレージの性能測定

図 4 に示すとおり，サーバにファイバーチャネル 8Gbps で直接接続したストレージに対して，サーバが直接アクセスした時（以後ローカルアクセスとする）と，クライアントから InfiniBand ネットワークを経由してアクセスした時（以後ネットワークアクセスとする）の性能を測定し比較した．

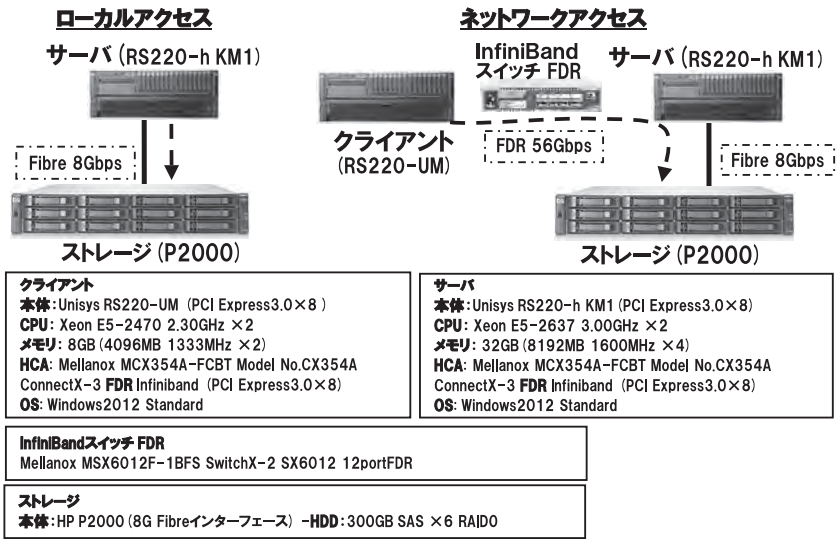


図 4　ローカルアクセスとネットワークアクセスの測定構成

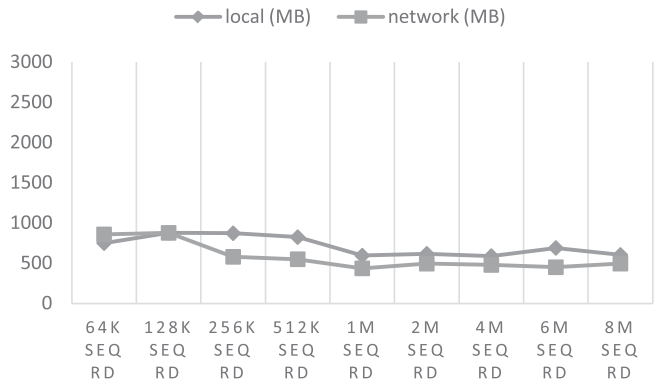


**図5　ローカルアクセスとネットワークアクセスの測定結果**

図5に示すとおり，ローカルアクセスとネットワークアクセスでは，想定どおり性能差が出なかった．ピーク値で比較すると，128Kのシーケンシャルリードで，ローカルアクセス，ネットワークアクセスともに877MB/s（6.9Gbps）と同じ値を計測した．これはInfiniBandを介したネットワークアクセスにおいて，帯域が8Gbpsであるサーバとストレージ間のファイバーチャネルがボトルネックになったことを示している．

2）　RAMディスクの性能測定

図6のとおり，サーバにRAMディスクを設定してローカルアクセス時の性能測定を実施した．検証環境のサーバに搭載したメモリの周波数は1600Mhzに対応するが，サーバの仕様により1333MHzで動作する．1333MHzで動作するメモリの1チャネルの転送速度は10.67GB/s（85.36Gbps）である．

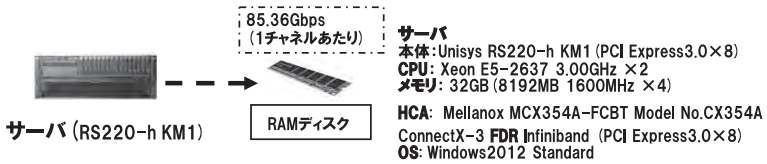


**図6　RAMディスクの測定**

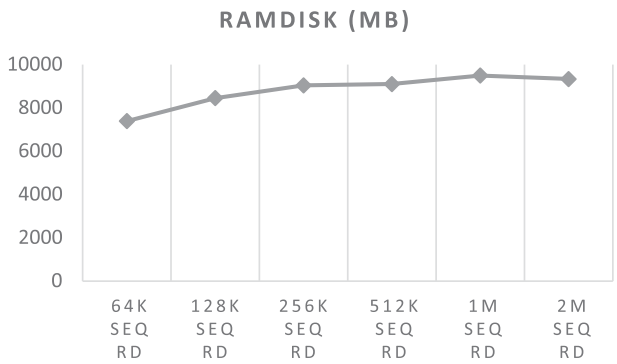


**図7　RAMディスクの測定結果**

図 7 のとおり，RAM ディスクの性能は，最も低い値の 64K シーケンシャルリード時でも，7387MB/s（57.7Gbps）と InfiniBand FDR の理論値 56Gbps を越えており，InfiniBand の測定に十分な構成を確保した．

### 3. 2. 2　PCI Express のボトルネック

HCA をインストールする PCI Express のスロットによって，ボトルネックが発生するかどうかを確認した．図 8 の InfiniBand ネットワーク構成で，サーバ側は 56Gbps の帯域を持つ FDR の HCA を 32Gbps と帯域幅の少ない PCI Express3.0×4 スロットに搭載した．

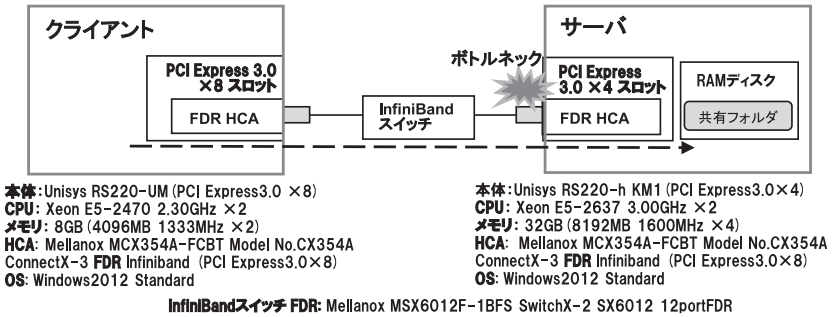


図 8　PCI Express のボトルネック測定構成

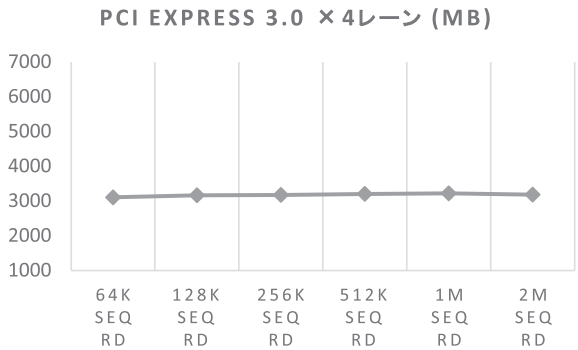


図 9　PCI Express のボトルネック測定結果

結果，図 9 のとおり，FDR の理論値 56Gbps に対して，ピークでも 3221MB/s（25.2Gbps）と PCI Express × 4 の仕様である 32Gbps に抑えられ，PCI スロットがボトルネックになることを確認した．

### 3. 2. 3　FDR，QDR の性能測定

FDR，QDR の測定は，それぞれの規格にあったクライアント，サーバならびにスイッチを用意した．図 10，図 11 はそれぞれ FDR と QDR の測定構成を示す．

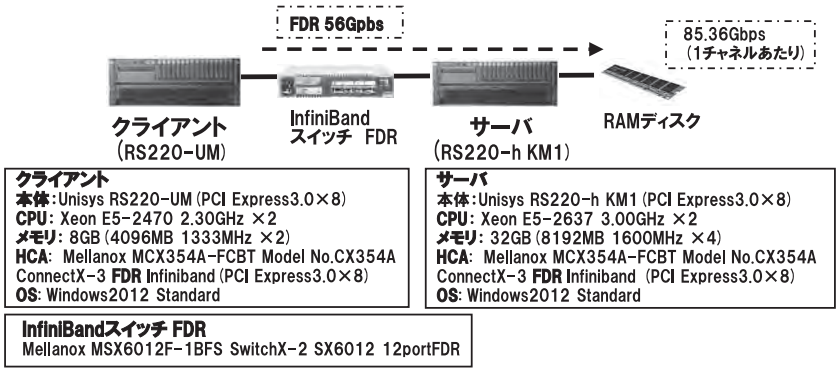


図 10　InfiniBand FDR 測定構成

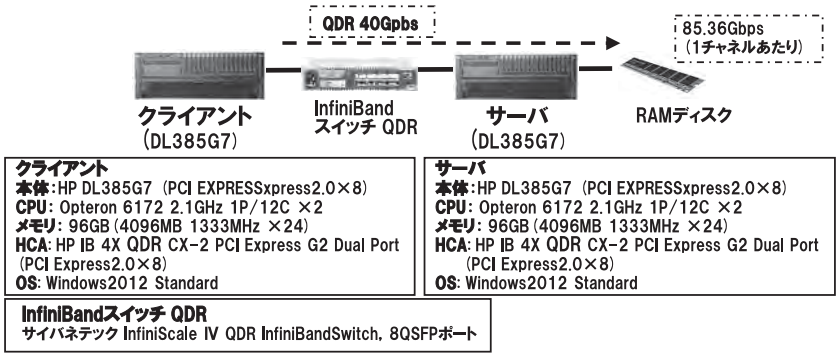


図 11　InfiniBand QDR 測定構成

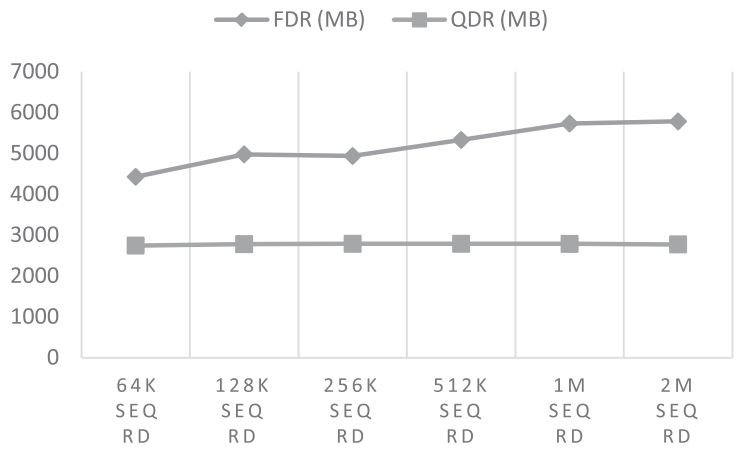


図 12　FDR，QDR の測定結果

図 12 に測定結果を示す．FDR は，ピーク値が 2M のシーケンシャルリードで 5784MB/s（45.2Gbps）と QDR のピーク値 2789MB/s（21.8Gbps）の約 2 倍の性能を計測した．FDR は，符号化後の理論値 54.3Gbps に対して，実測値は 17％低い値となった．また QDR は，符号化後の理論値 32Gbps に対して実測値は 32％低い値となり，FDR と QDR の性能効率で 15％の差がでた．この 15％の差の要因として，QDR HCA の 40Gbps の帯域と搭載した PCI Express2.0

× 8 の帯域が同じであったため，そこにボトルネックがあったと考えられる．

### 3. 2. 4　10Gb イーサネットの性能測定

InfiniBand の性能測定と同様に，サーバの RAM ディスク上の共有フォルダを，クライアント側でマウントし，Iometer によるシーケンシャルリードを実施した．図 13 に 10Gb イーサネットの測定構成を示す．

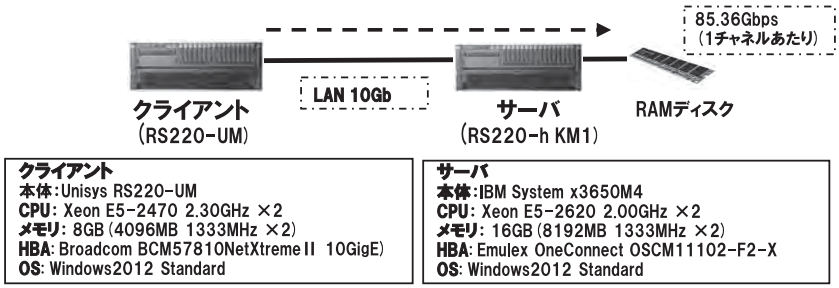


**図 13　10Gb イーサネットの測定構成**

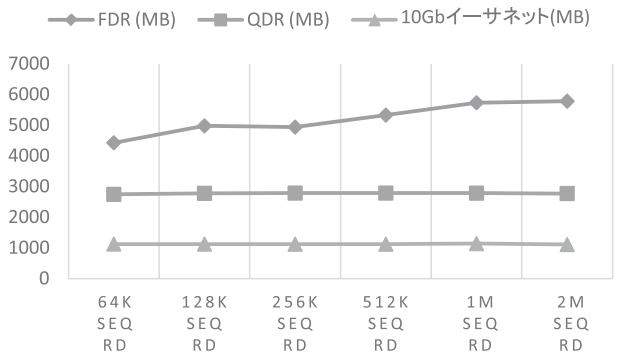


**図 14　InfiniBand と 10Gb イーサネットの測定結果**

結果は図 14 のとおり，FDR のピーク値 5784MB/s（45.2Gbps）は，10Gb イーサネットのピーク値，1142MB/s（8.9Gbps）の約 5 倍の性能を示した．QDR でも，ピーク値 2789MB/s（21.8Gbps）は，10Gb イーサネットの 1142MB/s（8.9Gbps）の約 2.5 倍の性能となった．

## 3. 3　可用性の確認

### 3. 3. 1　マルチチャネルによる HCA の冗長化

ここでは，マルチチャネルの冗長化の検証結果を報告する．3. 1. 1 項で述べたとおり，ネットワークの設定方法はイーサネットと同じだが，冗長化の方式は異なる．イーサネットの冗長化の方式はハードウェアに近いドライバのレイヤーで実現する．InfiniBand では，2014 年 8 月時点でリリースされている WinOF ドライバにて冗長化をサポートしておらず，ドライバのレイヤーで実現できない．代わりに上位のアプリケーションレイヤーで動作するマルチチャネルによって実現する．

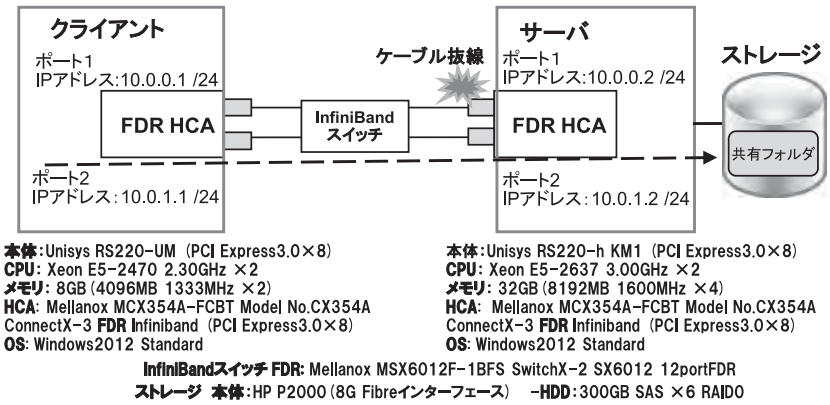


図 15　マルチチャネル冗長化

図 15 は，マルチチャネル冗長化の動作を確認した時の構成である．確認方法は，クライアント側でマウントしたサーバの共有フォルダ（外付けストレージ領域）に対して連続した IO 負荷をかけ，サーバの HCA ポート 1 のケーブルを抜いても，ファイルアクセスが継続されるかどうかをパフォーマンスモニタにて確認する．

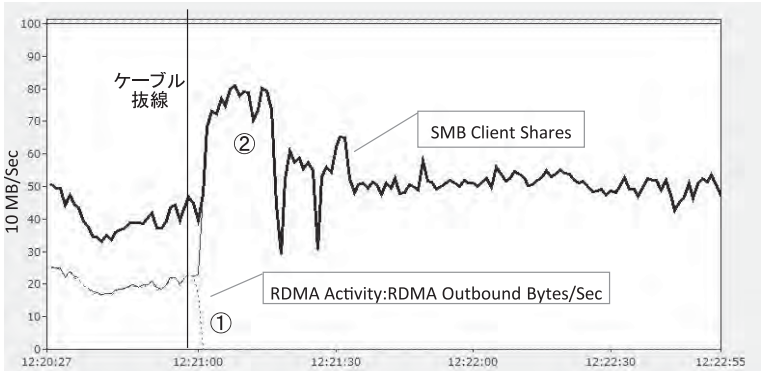


図 16　マルチチャネルパフォーマンスモニタ出力

図 16 は検証時のパフォーマンスモニタの出力である．図中の SMB Client Shares は，システムの Total のパフォーマンスを示す．RDMA Activity は，HCA のポート 1，ポート 2 それぞれのパフォーマンスを示す．HCA のポート 1 経由，ポート 2 経由の二つのパスが疎通可能なうちは，マルチチャネルの機能でロードバランスされているため，それぞれのポートの RDMA Activity の値の和が，SMB Client Shares の値になっている．

①の点線はサーバの HCA ポート 1 のケーブルを抜いたために，ポート 1 を使っていたパスのアクセスが停止したことを示す．②は，二つのパスのうちポート 1 のパスで疎通が無くなっても，ポート 2 のパスだけでアクセスが継続されたことを示している．また，切断されたパスにあったデータ量が残されたパスに追加され，システムとしての性能が維持されたことがわかる．

ポート 1 の抜線後，ケーブルを接続し直したところ，物理的にリンクアップはするがポート

1 のパスを使用したファイルアクセスは再開されなかった．アプリケーションを再実行することでポート 1 を再度利用して，ファイルアクセスができるようになった．一度抜線したポートを再度利用するには，アプリケーションを実行し直す必要があるため，注意が必要である．

### 3. 3. 2　クラスタの動作

InfiniBand の実用性を見極める上で，Windows2012 のクラスタ構成の検証を行った．InfiniBand を用いたクラスタ構成の構築作業は，イーサネットを使用したクラスタ構成の設定知識があれば行うことができた．イーサネットを使用したクラスタ構成の設定と比べて，特別考慮すべき点はない．

クラスタのフェールオーバー機能の確認のため，透過フェールオーバー環境を構築した．透過フェールオーバーは，Windows2012 の SMB3.0 で追加された機能で，ファイルサーバをクラスタ化し，フェールオーバー時に一時的に処理が停止しても，自動的にクライアントから継続してファイルアクセスを可能にする機能である．

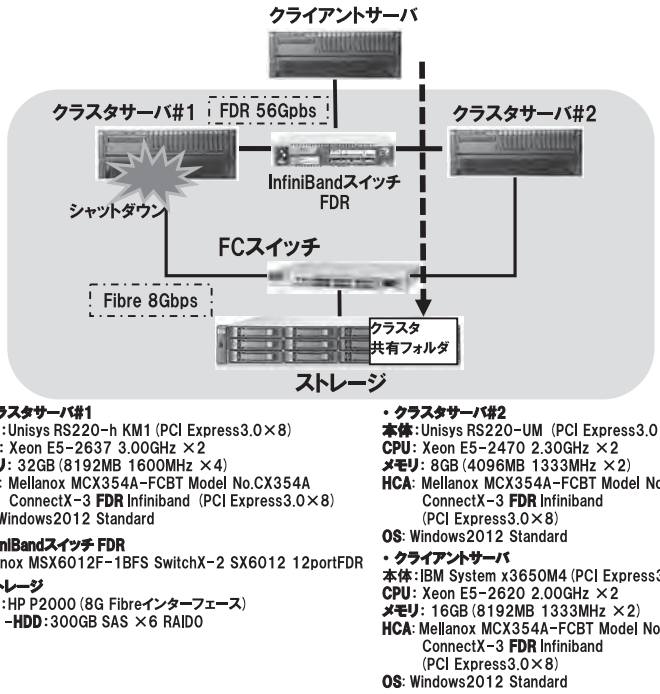


図 17　透過フェールオーバー構成

図 17 に検証環境を示す．クラスタのフェールオーバー機能の確認方法は，クライアントサーバでマウントしたクラスタの共有フォルダに対して連続した IO 負荷をかけ，クラスタサーバ #1 をシャットダウンしても，ファイルアクセスが継続されるかどうかをパフォーマンスモニタにて確認した．

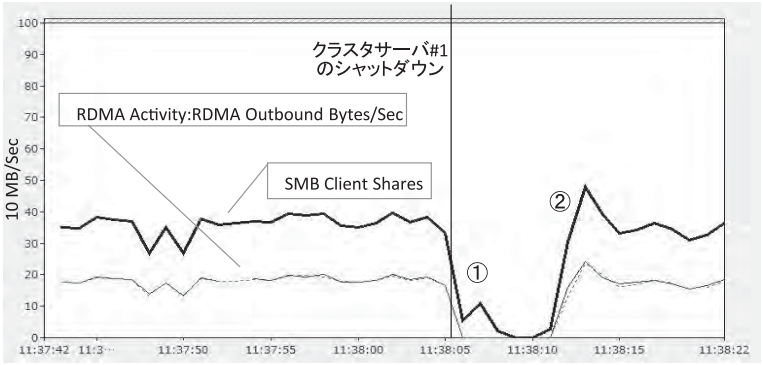


**図 18　透過フェールオーバーパフォーマンスモニタ出力**

図 18 はクライアントサーバ上で確認した，透過フェールオーバー時のパフォーマンスモニタの出力である．図中の SMB Client Shares はシステムの Total のパフォーマンスを示す．RDMA Activity は，HCA のポートのパフォーマンスを示す．①でクラスタサーバ #1 のシャットダウンを実施している．クラスタサーバ #2 にフェールオーバーする①から②までの約 5 秒間ネットワーク遅延が発生しているが，ファイルアクセスは継続された．

## 4.　InfiniBand の実用性の確認

前章の実機検証にて，InfiniBand は HPC 分野だけでなく，IA サーバでも実用性があることを確認した．PCI バスの規格の進化と，Windows2012 で新しく加えられた SMB Direct による RDMA への対応で，InfiniBand の利点を活かせるようになった．実装については，特別な設定を行うことなく，イーサネットを設定できる知識があれば，InfiniBand も同様に設定できることが確認できた．性能面では，IA サーバで普及している 10Gb イーサネットの性能の約 5 倍を記録した．可用性については，Windows2012 のマルチチャネルにより HCA が冗長化でき，またクラスタの透過フェールオーバーにより，サーバの冗長性を確認することができた．IA サーバで InfiniBand を取り入れる条件は揃ったといえる．

注意点として，HCA 搭載時には，PCI Express がボトルネックにならないよう，対応したスロットを使用することが求められる．また Windows2012 上で SMB Direct を有効にしていないと，RDMA が機能せず性能が落ちてしまう．3. 3. 1 項で述べたとおり，マルチチャネルの冗長構成では，ケーブルの再接続後にアプリケーションを実行し直す必要があった．アプリケーション側で自動または手動でリカバリする仕組みを備える必要がある．

ストレージが必要なシステムでは，3. 2. 1 項の検証のとおり，サーバとストレージ間の帯域速度が問題となる．InfiniBand の帯域を活かすにはストレージへ直接アクセスするよりも，より高速なメモリを介してストレージへアクセスしたほうが良い．そのためには，メモリ容量を増やし利用頻度の高いデータをメモリに展開することで，システム全体の通信性能を上げることができる．サーバ間の大量のデータ通信には InfiniBand を使い，外部とのデータ通信や運用・監視にはイーサネットを使うなど目的に合った使い分けが重要である．

検証結果から提案する構成として，大量のオンライントランザクション処理が発生する大規模データベースシステムに InfiniBand は最適と考える．データベースサーバとストレージとの間に，頻繁に使用するデータをキャッシュするキャッシュサーバを配置する．このデータ

ベースサーバとキャッシュサーバ間のインターコネクトに InfiniBand を使用することで，データベースへのトランザクション性能を向上させることができる．またキャッシュサーバの内蔵ストレージに SSD（Solid State Drive）を採用することで，外部ストレージのボトルネックを軽減し，高速データベースシステムが構築できる．

InfiniBand の規格は，EDR に続き HDR のリリースが計画されており，今後もアーキテクチャーの進化が期待できる．イーサネットに比べ価格性能が優れている利点もあり，今後は IA サーバでも，InfiniBand がファイバーチャネルやイーサネットと同様に，商用環境で使われるようになっていくと考える．

## 5. お わ り に

本稿では InfiniBand の HCA の性能と機能の検証に注力したが，InfiniBand スイッチや，InfiniBand 接続のフラッシュストレージの性能測定，また InfiniBand を採用した SQL サーバの構成についても今後考察を深めたい．また，マルチチャネルの可用性の検証では，障害復旧時にアプリケーション側で考慮が必要だった．今後 HCA ベンダのドライバまたは Windows の機能拡張に期待したい．

今後も培ってきたノウハウを活かし，システム全体の最適化を考え，顧客の期待に応えるソリューションを提案していく所存である．

**参考文献**
［1］ 松本直人,「InfiniBand で変わるデータセンター内通信」, @IT, アイティメディア(株), 2011 年 2 月 15 日，http://www.atmarkit.co.jp/fnetwork/tokusyuu/51ib01/01.html
［2］「HA8000 シリーズハードウェアアーキテクチャーガイド（2013 年 9 月～モデル）」, 株式会社日立製作所，2013 年 9 月，3 ページ，29 ページ
［3］「HP ProLiant DL380p Generation8 システム構成図」，日本ヒューレット・パッカード株式会社，2014 年 8 月 7 日，65 ページ，66 ページ
［4］「IBM System x3650M4 System Guide」，日本アイ・ビー・エム株式会社，2014 年 8 月 26 日版，18 ページ，29 ページ，30 ページ
［5］「Database Acceleration Solution for HP ProLiant 技術解説」，日本ヒューレット・パッカード株式会社，2010 年 9 月
［6］「RDMA プロトコル：ネットワーク性能の向上 技術概要」，日本ヒューレット・パッカード株式会社，2006 年 1 月
［7］「Windows Server 2012 ファイルサーバに SMB 3.0 の新機能」，日本マイクロソフト株式会社，http://support.microsoft.com/kb/2709568/ja
［8］「Infiniband hack-a-thon #2」
http://www.slideshare.net/detteiu/infiniband-hackathon-2-windows
※上記参考文献内の URL は 2014 年 8 月 31 日時点での存在を確認．

**執筆者紹介** 本 間 一 久（Honma Kazuhisa）
2001 年ユニアデックス(株)入社．カスタマーサービスエンジニアとしてオープンシステムを担当．2011 年よりオープンサーバのプロダクト主管業務へ従事．