April 2008

InfiniBand over Ethernetについて

10ギガビットEthernetにおけるIPCコンソリデーションの進化

1.0 はじめに

Fibre over Ethernet (FCoE) の背後で業界の流れは、Ehternet上のゼロコピー送受信とリモートDMA(RDMA)テクノロジーを使うサーバメッセージング(または、プロセス間通信、または、IPC)の最良な方法がなにかという問題を提起するいくつかの重要な優先順位を決めています。 IPCでは２つの競争するテクノロジー、インフィニバンドと(10GigBベースの)iWARPがあります。 FCoEの最初の成功後に同様のビジネスとテクニカルのロジックを適用するならば、InfiniBand over Ethernet (IBoE)が適当であると結論づけられます。 以下で、その理由を説明します。

1.1 SANコンソリデーション用 FCoEの魅力 – ビジネスの視点

Ethernetはどのサーバにも使われているのでサーバI/O統一用のテクノロジーになったとしても、それがデータセンターにおけるエンド・ツーエンドTCP,UDP,IPになるとは考えられません。 もしそうであるすれば、エンド・ツーエンドEthernetによるiSCSIはサーバからストレージ・ボックスまで世界に広まったでしょう。 しかし、そうなりませんでした。

それは、無視することのできない巨大なファイバーチャンネル・ストレージとソフトウェアへの投資があるためです。 それゆえ、Fibre Channel over Ethernet (FCoE)では、ITマネージャーが、サーバI/Oを10GigEへ統一し既存のファイバーチャンネル・ストレージへのシームレスな接続性を維持し、ファイバーチャンネルへのソフトウェア投資を保護することができるようになります。

1.2 SANコンソリデーション用 FCoEの魅力 – テクニカルな視点

iSCSIとFCoEの技術的な比較をしてみます。 iSCSIはLANとインターネット用に設計されたIPネットワークに基づいており、従来の損失のあるEthernetネットワークの問題を解決するためにTCPに依存しています。

リカバリーとフローコントロールをTCPに依存すると大きなオーバーヘッドが生じるため、いくつかのケースではネットワークの輻輳解決への反応が遅くなり(TCPは、ソフトウェアで使えるCPUサイクルに依存するので)、また、いくつかのケースでは高価で多くの電力を消費するスケーラブルしないI/Oアダプタ(TCPオフロードエンジン、または、TOE)となってあらわれます。 TOEへのLinuxコミュニティーの反対と、(VMware ESX, Citrix XenServeなどTOEを効果的に使えない)仮想化サーバ環境でのTOE価値観の不足により、TOEは限られた状況下での利用に縮小していきます。 しかし、、これだけではありません。

iSCSIとTOEは、Ethernetが信頼できる媒体でないという特有の仮定をしていました。つまり、損失性です。

つまり、TCPでは避けることができません。 信頼性のあるEthernetの出現とPer Priority Pauseのサポートによる、損失のない仮想レーンは、Ethernetによるストレージトラフィックで使うことのできるので、ストレージアプリケーションでのTCPの重要性は、さらに縮小します。様々なIEEEワーキングループで作業中のEthernetへのさらなる強化は、 レイヤー２Ethernet を使用する輻輳制御と管理を可能にし、TCPへの依存をさらに減らします。

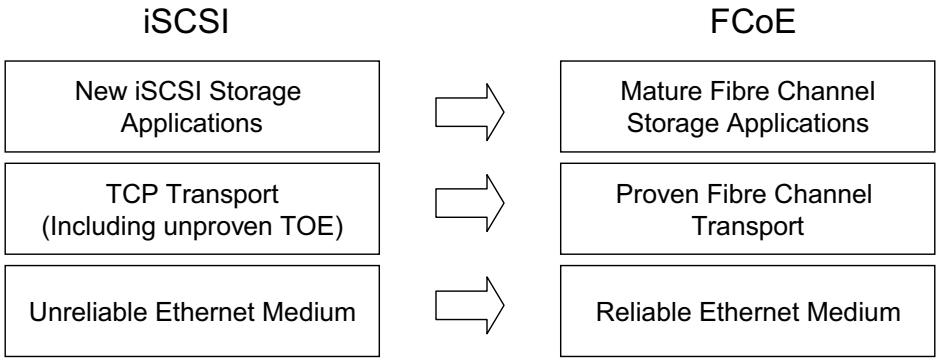

Figure 1: iSCSI versus FCoE

これらの強化は、FCoEのような信頼性のあるEthernetでファイバーチャンネルトランスポートを適用できることを裏付けします。そうすることにより、以下の利点が明らかになり、iSCSIを使用理由が減ります。

FCoEによる投資とスキルの保護

- ファイバーチャンネルトランスポートのストレージ用動作証明済み
- 簡潔で明瞭
- 高価な電力を多く必要とするTOEが不
- OSスタックの最小限の変更
- ストレージの管理スキルとツールの保護

ビジネスとテクニカルの観点から、FCoEは10GigE上SANコンソリデーションの進化段階にあります。

1.3LANとIPC のコンソリデーション

10GigEとLANが連携し、FCoEととも、データセンターの２つの重要なトラフィックタイプを扱い、FC SANとEthernet LANはサーバの10GigEアダプタ上で統一することができます。

トラフィックタイプの3番目のカテゴリーには、クラスタ、グリッド、ユーティリティ・コンピューティングを支援するIPCトラフィック、 サービス指向インフラの増加するコンポーネントをサーバノード間の低レイテンシにより少ないサーバ(より高い使用率で)で行うような変換、 (例えば、アルゴリズム・トレーディングなど)トランザクションがミリ秒遅れるたびに数百万ドルの損失になる“time is money”があります。

1.4 IPCコンソリデーション IBoE 対 iWARP – ビジネスの視点

まさにSANのファイバーチャンネルテクノロジでの莫大な投資と成熟があるように、IPC用インフィニバンドにも同様の投資と成熟があります。

Ethernetはどのサーバにも使われているのでサーバI/O統一用のテクノロジーになったとしても、TCPベースのiWARPのみに基づいたIPCコンソリデーションとして想定することはできません。FCoEがEthernetフレームでFCデータをカプセル化するように(そして、慣れ親しんだSANソフトウェア、インターフェース、管理性を保てるように)、InfiniBand over Ethernet または IBoEは、EthernetフレームでIBデータをカプセル化し、使い慣れたIPCソフトウェア、インターフェース、管理性を保ちます。これによりITマネージャーがサーバのI/Oを10GigEへ統合し、IPCとクラスタアプリケーションのシームレスな相互運用とソフトウェアへの投資(例えば、金融、クラスタデータベース、商用や学術、ハイパフォーマンスコンピューティングアプリケーション)を保護します。

そのようなアプリケーションはすでに、インフィニバンド上のOpenFabrics (www.openfabrics.org) IPC プロトコルスタック(または、ポピュラーなLinuxとWindowsで利用できる)は、ゼロコピーsend/recevとIBoEを使ったRDMA Ethernetでシームレスに展開できます。

IBoEは、Mellanox社のConnectX アダプタでサポートされています。

1.5 IPC コンソリデーション用 IBoE – テクニカルな視点

iWARPとIBoEの技術的な比較をしてみます。

iWARPはLANとインターネット用に設計されたIPネットワークに基づいており、従来の損失のあるEthernetネットワークの問題を解決するためにTCPに依存しています。リカバリーとフローコントロールをTCPに依存すると大きなオーバーヘッドが生じるため、前述のiSCSIとFCoEの比較と同じになります。

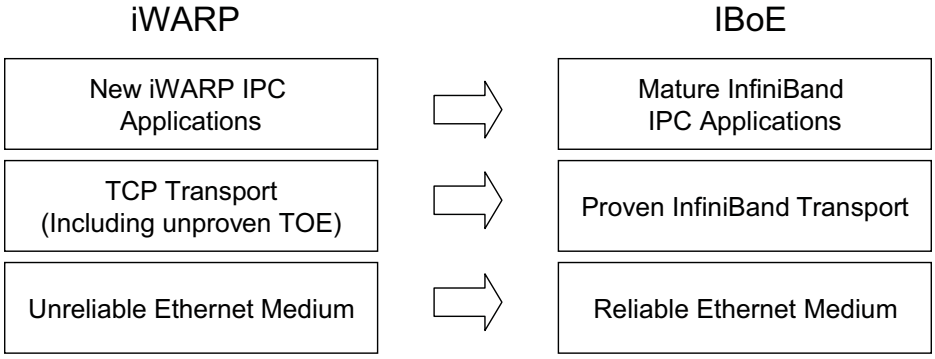

Figure 2: iWARP versus IBoE

信頼性のあるEthernetの出現と、Per Priority Pauseのサポートにより、レイヤー 2 Ethernet を使う輻輳管理と制御、実証された効率的なインフィニバンドトランスポートは、信頼性のあるEthernet上のストレージ用FC transportのように、信頼性のあるEthernet上のIPCで最適な選択になります。

これにより、以下のことが明らかになり、TOEによるiWARPを使用する理由がなくなります。

IBoE投資とスキルの保護

- 実績のあるIPC/サーバ間通信
- 簡潔で明瞭
- 高価な電力を多く必要とするTOEが不要
- OSスタックの最小限の変更 IPC/サーバの管理スキルとツールの保護

SANコンソリデーションのFCoEのように、IBoEは、ビジネスとテクニカルの観点から、10GigE上のIPCコンソリデーションの進化段階にあります。