100ギガビットイーサネットについて

今後のトラフィック増大に伴い、近い将来導入の必要性が出てくる100ギガビットイーサネットについて、技術的な特徴を解説します。また、共同実証実験から得られたことについて報告します。

3.1 はじめに

本レポートでは、はじめに100ギガビットイーサネット(以下、GbE)の技術的な特徴を解説し、次に、インターネットマルチフィード株式会社、エヌ・ティ・ティ・コミュニケーションズ株式会社と共同で行った100GbE IX(ISP相互接続点)共同実証実験についての報告、最後に100GbEの光トランシーバ規格と今後の動向について解説します。

3.2 100GbEについて

100GbEは、2010年6月に標準化が完了してIEEE 802.3ba[*1]で規定されています。ここでは100GbEを理解する上で重要なポイントを説明します。また、10GbEで確立された技術を多く流用しているため、10GbEの詳細[*2]について理解していることが前提となります。

■ 100GbEの基本原理

10GbEの10倍の速度の信号をシリアル伝送することは技術的なハードルが高く、また、実装にかかるコストが問題になります。そこで100GbEでは10Gbpsや25Gpbsといった低速なデータ転送を並列に行い、100Gbpsの速度を実現しています(図-1)。この並列データ伝送を実現する技術が100GbEの特徴です。

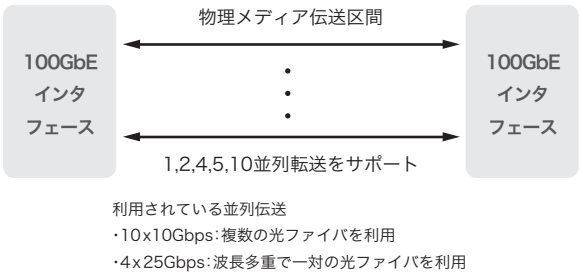

図-1 100GbEの基本原理

■ 100GbEとリンクアグリゲーションの比較

この並列データ伝送技術はMLD(Multi Lane Distribution)と呼ばれており、OSI参照モデルの物理層で実装されています(図-2)。同様の並列データ伝送を実現する仕組みに、複数インタフェースを束ねて仮想的に一つに見せる、IEEE 802.3adで規定されるリンクアグリゲーションがありますが、100GbEとは以下の点で異なります。

1. 100GbEは物理層で並列データ伝送を行うため、利用ユーザは並列データ伝送のための設定や挙動の詳細を気にする必要がありませんが、802.3adではリンクアグリゲーションの設定や挙動を気にする必要があります。
2. 100GbEの並列データ伝送は、イーサネットフレームを一定長に分割して伝送路に均等にデータ送信を行うため、特定の伝送路にデータが偏る問題は発生しませんが、802.3adではフレームの分散方法に問題が均等に分散しない場合があります[*3]。

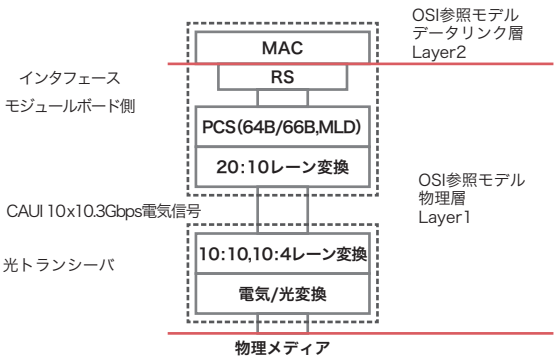

図-2 100GBASE-SR10/LR4/ER4概略図

*1 IEEE802.3baのドキュメント(http://standards.ieee.org/about/get/802/802.3.html)
*2 「10ギガビットEthernet教科書」石田修、瀬戸康一郎監修、IDGジャパン、2002年初版を一読しておくことをお勧めします。
*3 802.3adではフレーム長は考慮せず、フレーム単位での分散を行うためフレーム長に偏りがあると均等分散しない。また、フレーム分散のための実現方法が規格で定義されておらず、機器の実装に依存し、均等に分散しない場合がある。

■ MLDについて

RS（Reconciliation Sub-layer）では、データリンク層（MAC）からイーサネットフレームを受け取り、64ビット単位に区切り下位の層に渡します。PCS（Physical Coding Sub-layer）では、64ビットデータに物理メディア上のデータ転送用2ビットヘッダを付け、66ビットのブロックを構成する64B/66Bブロックを作成し、それを仮想レーン[*4]と呼ばれる並列データ伝送路に順番に送信することで、MLDを実現しています（図-３）。

■ 仮想レーンの多重化と分離の仕組み

仮想レーンは途中で段階的に集約することも可能で、その場合はレーン間のデータをビット単位で多重化します（図-4）。多重化して集約することができると必要に応じてレーン数を変更して、様々な物理メディアに適応した伝送が可能になります[*5]。しかし、伝送路の途中でビット単位でレーンの多重化が行われると、送信側と受信側のレーン間の対応関係が成り立たなくなります（図-4のレーン5を参照）。そこでアライメントマーカー（Alignment Marker）という仕組みを利用して受信側レーンと送信側レーンの対応関係を受信側に伝えます。

■ アライメントマーカーについて

64B/66Bブロックのデータを16383個送信するごとに、一時的にデータブロックの送信を中断して、1アライメントマーカーを全レーンに同じタイミングで送信します（図-5）。例えば、送信側物理レーン5のアライメントマーカーには、レーン5の識別情報が含まれています。これを受信側物理レーン1で受信した場合は、識別情報をみることで、仮想（送信側物理）レーン5に対応したデータであると解釈します。

■ 仮想レーンのビットエラー監視の仕組みについて

複数レーンの並列伝送では、レーンごとに異なる物理メディア上でデータ伝送される可能性があります。そのため、各レーンの回線品質を監視するためのBIP（Bit Interleaved Parity）機能が実装されました（図-6）。BIP機能は、各レーンで前のアライメントマーカーを含む64B/66Bブロック16384個のビットパリティを計算して、その値を送信し

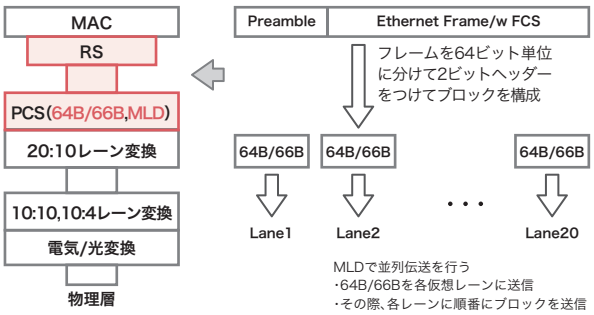

図-3 RS+PCS（64B/66B, MLD）

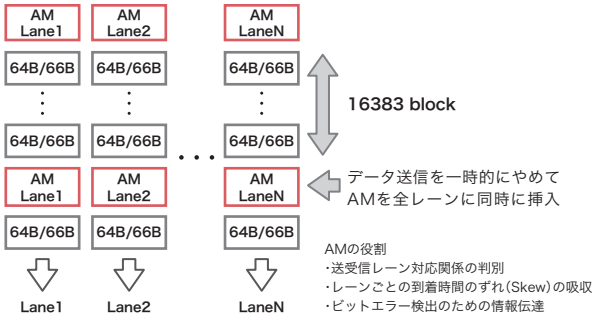

図-5 アライメントマーカー（AM）

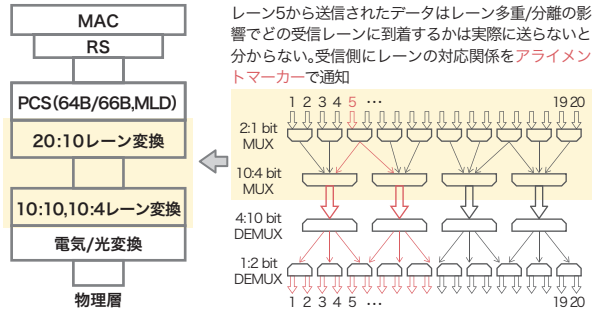

図-4 レーン多重/分離による送受信レーンの関係

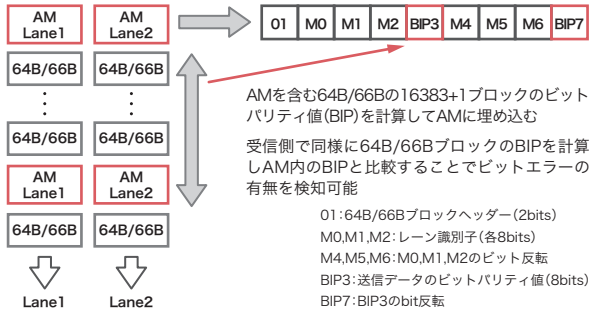

図-6 アライメントマーカーとBIPの関係

＊4　100GbEでは20レーンを利用している。

＊5　現在のIEEE規格では、インタフェースボードと光トランシーバモジュール間のデータ伝送は10レーンで行うように規定しているため、物理メディア部分で利用可能なレーン数は1、2、4、5、10になります。実際の100GbEでは物理メディア部分が4レーンの100GBASE-LR4/ER4、10レーンの100GBASE-SR10、10x10 MSAが利用されています。

ます。受信側では同様の方法で受信したデータのビットパリティを計算して、アライメントマーカーのBIP値と比較し、16384ブロックの間にビットエラーが発生していたかを確認します。BIP機能があることで、伝送路でビットエラーが発生した場合に、並列伝送路すべてに影響している問題なのか、一部のレーンを運ぶ伝送路だけの障害なのかという切り分けが可能になります。

■ まとめ

ここでは100GbEで並列データ伝送を実現するため重要な機能を解説してきました。詳細について興味のある方は802.3baのドキュメントを確認することをお勧めします。

3.3 100GbE IX共同実証実験

6月1日に3社共同で100GbE IX(ISP相互接続点)共同実証実験に関するプレスリリースを発表しました[*6]。また、7月15日のJANOG28 Meetingにおいて共同実証実験の内容の一部を公開しました[*7]。本レポートでは共同実証実験の内容について公開されている範囲で簡単に触れます。

本実験の目的は、100GbEの導入に備えた安定性の確認と運用上の問題点の検証です。また、IX接続する際のマルチベンダー機器環境での相互接続の問題点を確認しました。実験から得られた重要な点は、以下のとおりです。

1. マルチベンダー機器環境での相互接続に大きな問題はなかった
2. 運用面では従来の10GbEまでと異なる点があるので注意する必要がある
   - 100GBASE-LR4の送受信の光レベルは波長多重されたもので、光のパワーが強いことに注意
   - 100GBASE-LR4の各波長の光レベル測定したい場合は専用のパワーメータを利用する、または、サポートしていればCLIを利用するが必要ある
   - 各レーンの品質監視のためのBIPカウンターを参照するための機能を実装してないものが多い
   - 100GBASE-LR4は僅かな汚れに対して非常にシビアでエラーが発生しやすく、クリーナーでCFP及びファイバー端面を綺麗に掃除する必要があった

3.4 100GbEの光トランシーバについて

■ CFP MSAと10x10 MSAの規格

現在、100GbEで利用可能な光トランシーバには、IEEEに標準準拠しCFP MSA[*8]で規定されたCFPモジュールとIEEEに準拠しない独自規格の10x10 MSA[*9]準拠の二種類のトランシーバが存在しています。図7は光トランシーバの構造比較です。

100GBASE-LR4/ER4 CFPの光トランシーバは、4:10レーン変換のGearBoxチップと4x25.8Gbpsの高速レーザ素子のコストが高く普及を妨げるという問題がありました。そこで、10x10 MSAでは既に量産されている10GbE技術を流用して、10x10.3Gbpsの電気信号をそのまま10x10.3Gbpsの光信号に変換し、部品コストを低減する独自仕様を作成しています。しかし、10x10 MSAはIEEE標準準拠ではないため、一部ベンダー機器のみでのサポートとなり注意が必要です。

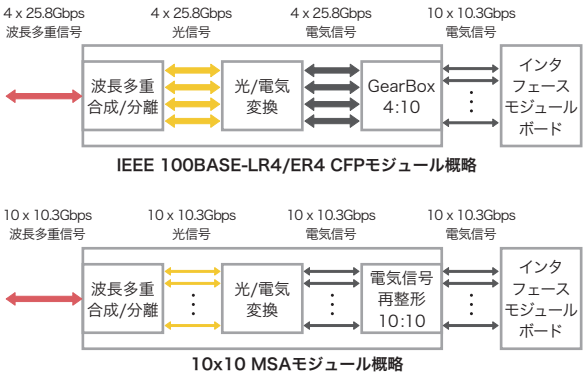

図-7 100GBASE-LR4/ER4 CFPと10x10 MSAの比較)

*6 IIJプレスリリース「世界初の超高速100GbE IX(ISP相互接続点)共同実証実験に成功」(http://www.iij.ad.jp/news/pressrelease/2011/0601-02.html)
*7 JANOG28「IXと100Gbit Ethernet」(http://www.janog.gr.jp/meeting/janog28/program/100G.html)
*8 C Form-factor Pluggable Multi Source Agreementの略。IEEEでは100GbEの光トランシーバモジュールの形状や機能の仕様の詳細については規定していないため、光トランシーバモジュールのベンダーが集まり、共通仕様を定義して、それに基づき製品を作っている。(http://www.cfp-msa.org/)
*9 テンバイテン Multi Source Agreement。機器ベンダーとその利用ユーザが集まって、低コストの作成可能な独自の光トランシーバの規格を策定している。(http://www.10x10msa.org/)

■ 利用可能な物理メディアと到達距離について

100GbEで利用可能なインタフェース規格を表-1に示します。ここで赤く表示しているものは、共同実証実験時点で利用可能であった光トランシーバです。利用可能な主な物理メディアには、銅線(Copper)ケーブル、多芯マルチモードファイバー(MMF)、シングルモードファイバー(SMF)の3種類があります。しかし、実際のコアネットワーク機器間の接続では、既存のシングルモードファイバーをそのまま利用可能な100GBASE-LR4/ER4、10x10 MSAしか選択肢がありません。また、現在利用可能な光トランシーバは到達距離があまり伸ばせず、100GbEとダークファイバーを利用したメトロエリアネットワークの構築が困難であり、キャリアクラスの伝送装置がなければ中長距離伝送が難しいのが現状です。今後、中距離伝送が可能な100GBASE-ER4(30km)と10x10-40kmの光モジュールが利用可能になるはずですが、現状のCFPの価格を考慮するとコスト的に利用可能かは不透明です。

	IEEE 100G	10x10 MSA 100G
Backplane 1m	100G BASE-KR4(策定中)	N/A
Copperケーブル 5m	100G BASE-CR4(策定中)	N/A
Copperケーブル 7m	100G BASE-CR10	N/A
MMF(OM3)100m	100G BASE-SR10 100G BASE-SR4(策定中)	N/A
SMF 2km	100G BASE-FR4(策定中、WDM)	10x10-2km(WDM)
SMF 10km	100G BASE-LR4(WDM)	10x10-10km(WDM)
SMF 30km/40km	100G BASE-ER4(WDM)	10x10-40km(WDM)

表-1 IEEE/10×10 MSAで定義されているインタフェース一覧

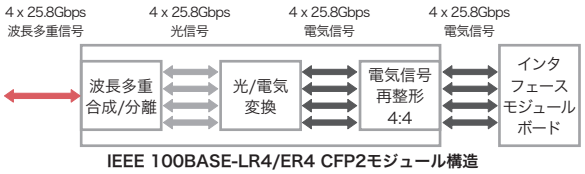

図-8 100GBASE-LR4/ER4 CFP2

■ 今後の光トランシーバ規格について

現在のCFPタイプの光トランシーバは、78x13.6x144(mm)と非常に大きくインタフェースモジュールに搭載可能なポート密度を上げることができませんが、大幅に小型化したCFP2、CFP4というモジュールが近い将来出てくる予定です。CFP2以降になるとインタフェースモジュールボード間の電気信号の速度が4x25.8Gbpsに変更されます(図-8)。これにより、10x10 MSAと同様にCFP上のGearBoxチップが不要で、CFPより安価になることが期待されます。

3.5 おわりに

現状では、100GbEに関して日本語でのまとまった情報が殆どありませんが、本レポートが100GbEを理解するための手助けとなれば幸いです。現時点で100GbEは非常に高価であり、また対応製品も少ないため、一般的な普及にはまだ時間がかかると思われます。また、IIJでは今後のトラフィック増大に備えて100GbE技術の導入検討を積極的に進める予定です。

執筆者:
大内 宗徳 (おおうち むねのり)
IIJ サービス本部 ネットワークサービス部 技術開発課。IIJ入社後、一貫してIIJバックボーンで利用する機器のテスト、インターネットに関する新技術の調査、研究開発に従事。