3

光インタコネクションを応用した計算機システム

西村 信治　新情報処理開発機構 光インターコネクション日立研究室　nisimura@crl.hitachi.co.jp

◆光接続した並列計算機システム◆

計算機内部ネットワークに光インタコネクション技術を応用することにより，高速で低遅延なデータ伝送を長距離で実現でき，計算機ネットワークの広帯域性とレイアウトの柔軟性を格段に向上できる．従来，光技術は，その広帯域性を利用し長距離基幹伝送系やEthernetなどのLAN（Local Area Network）に使用されてきた．基幹伝送系やLANにおける光伝送は，光信号伝送の広帯域・低損失な性質を生かすことで，GbpsからTbpsクラスの広バンド幅と数kmの長距離伝送を実現している．しかし，これら従来の光技術を並列コンピュータの内部ノード間接続にそのまま適用した場合，データ伝送の信頼性が低く（エラーレートに換算して10^{-12}程度），上位レイヤのプロトコル処理にてデータの破棄・再送などを実施して通信の信頼性を確保する必要がある．このため，高性能並列計算機内部のネットワークとしては，この重いプロトコルがネックとなる．このため，PCクラスタを始めとした並列計算機内部のノード間接続においては，Myrinet[1]などのSAN（system area network）が用いられている．これらの従来のSANは，電気ケーブルを用いた広バンド幅でシンプルなプロトコルを有するネットワークであり，低遅延で高速なネットワークであるが，電気配線の延長距離やトポロジにおける制約が厳しい．システムの大規模化には，電気ケーブルからのノイズの影響などが問題となる．このため，次世代の並列計算機内のネットワークには，LANの長距離伝送性能と，SANの高速・低遅延スイッチ機能を合わせ持つ新たなネットワーク技術が必要になる．並列光インタコネクション[2),3)]の採用により，このLANとSANの性能を合わせ持つネットワークが実現できる（**図-1**参照）．

10Gbpsクラスのネットワークを並列光インタコネクションを用いて実現する場合，大容量データを集中処理するネットワークスイッチの実現が大きな課題となる．将来的には，このスイッチは入力光信号を光信号のまま行路切換処理する「全光スイッチ」が有望と考えられている．しかし現在の全光スイッチは，パケット中のルーティング情報に従って出力先を決めるアドレス行路切換機能や衝突回避のためのバッファリング機能などの，実際の計算機ネットワークに必須な機能の実現が難しい．このため，現実の光ネットワークスイッチは，「光・電気信号混在スイッチ」の構成が現実的である（**図-2**）．光・電気信号混在スイッチは入力光信号を電気信号に変換し，行路切替え後に再び光信号に再変換し出力する．この大容量スイッチの実現には，以下の技術が必要となる．

1）大容量並列光インタコネクションモジュール: 大容量電気信号を低遅延に光信号に変換しデータ伝送する光インタコネクションモジュール[2),3)]（光インタコネクションモジュールには，複数のレーザ・ホトダイオードとそれらを駆動するIC回路を集積実装している）．

2）高速スイッチLSI: 高速電気信号の行路切替を実現するスイッチLSI．スイッチLSI自体には，ギガビットクラスの大容量・高速信号の入出力動作が必要とされる[4)]．

3）高速デバイスの回路・実装技術: 高速デバイスを，低ノイズかつコンパクトにボード搭載する実装技術[4)]．

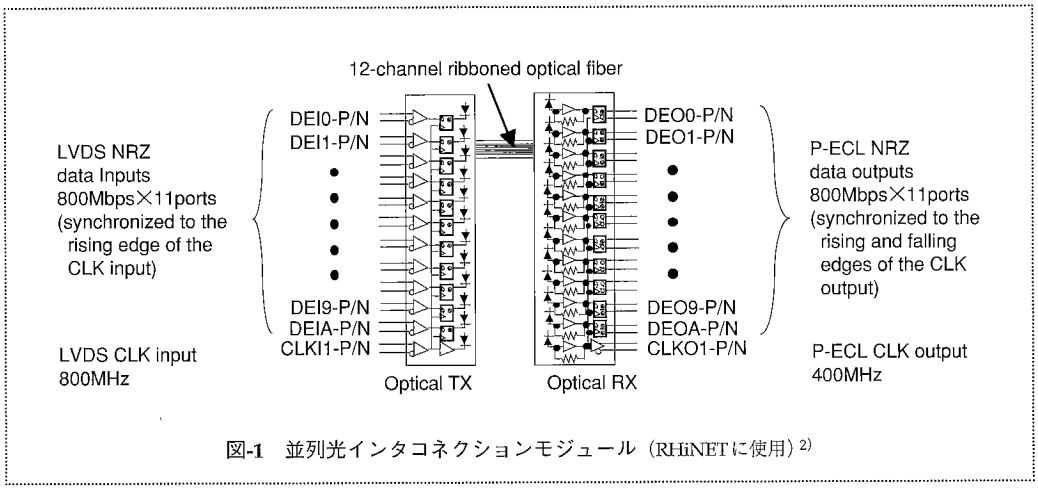

図-1　並列光インタコネクションモジュール（RHiNETに使用）[2]

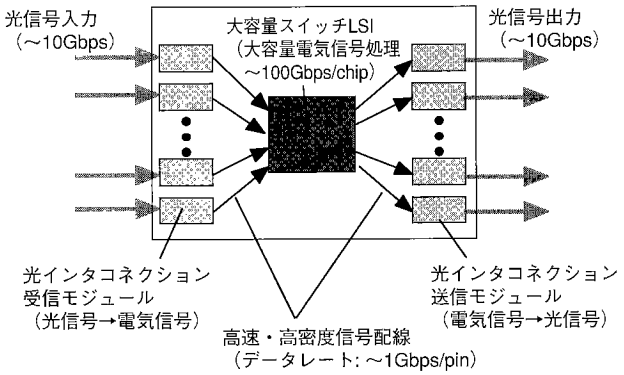

図-2　光・電気信号混在スイッチ

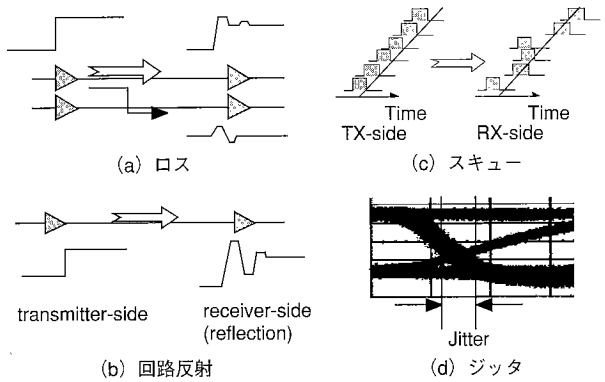

図-3　高速デバイスの回路・実装における課題

◆高速デバイスの回路・実装における課題◆

10Gbpsクラスの高速信号デバイスを用いる場合，電気信号配線の信号速度も1ピン当たり1Gbpsクラスに高速になる．このため，高速光ネットワークを実際に開発する場合，光モジュール・高速スイッチLSIと併せて回路・実装技術が必要となる．この際問題となるのが，電気信号配線のロス・スキュー・信号反射・ジッタである（図-3参照）[4]．

■ロス・信号反射

高速動作回路のインタフェースはLVDS（low-voltage differential signaling）などの小振幅差動信号が一般的であり，ボード内配線内での伝搬損失（ロス）や，回路反射による信号乱れの影響を受けやすい（LVDSの信号振幅は200mVから400mVと小さい）．このため，光モジュールと高速スイッチ間のボード内配線距離は数10cmと制限されるとともに，高精度な回路インピーダンス制御が必要になる．

■スキュー・ジッタ

並列信号を同期伝送する際，並列回路間の特性ばらつきが原因で受信信号の到着時間がばらつく．これをスキューというが，高速信号の同期伝送においてこのスキューを小さく抑えることが，回路実現上の大きな課題となる．たとえば，800Mbpsの高速信号の周期は1,250psと短く，故に許容できるスキューも約250psと小さくなる．このため，LSI，光モジュールの入出力バッファの高精度な特性制御が必要となる．また，高速デバイスが発生する電磁波ノイズも，信号のジッタ成分を増大して通信エラーの原因となるため，回路実装上の大きな課題となる．

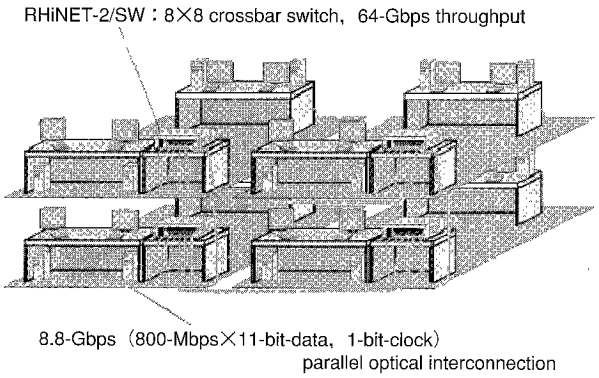

図-4　RHiNETの概念図

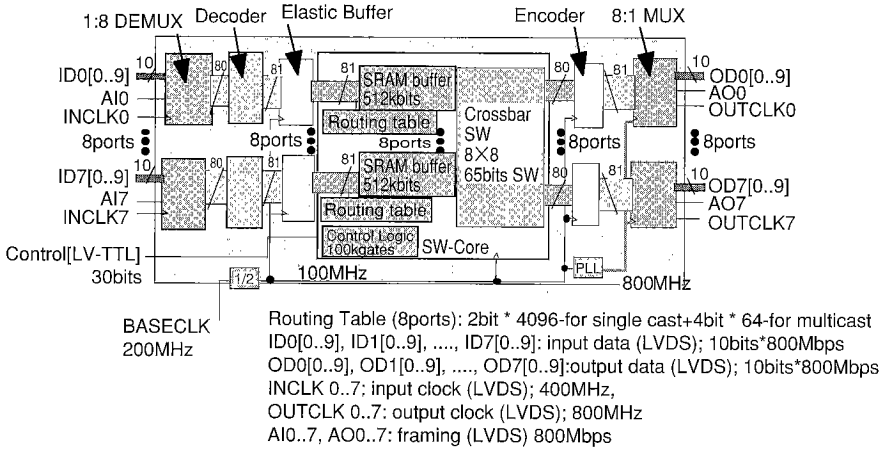

図-5　スイッチLSIの内部ブロック図

◆高速ネットワークスイッチ◆

光・電気信号混在スイッチに必要な前述の3技術を，我々の開発した並列計算機システム（RHiNET: RWCP high-performance network）用ネットワークスイッチ（RHiNET-2/SW）を具体例として説明する[4),5)]．RHiNETシステムはPCノードと，ネットワークスイッチおよびそれらを接続する光インタコネクションから構成される．RHiNET-2は高速スイッチ（RHiNET-2/SW）と各PC間を800Mbps×12チャネルの並列光インタコネクションで接続する（図-4参照）．

■大容量並列光インタコネクションモジュール[2)]

RHiNET-2にて使用した光送受信モジュールは12チャネルのレーザとホトダイオードを並列駆動することにより，8.8Gbpsの大容量光データ接続を実現している（800Mbps×データ11チャネル＋クロック1チャネル：図-1参照）．発振波長1.31μmの端面発光レーザレイを使用し，12芯シングルモードリボンファイバと組み合わせて使用することにより，伝送中の波形劣化を抑えて，並列同期信号のチャネル間スキューを100ps以下に抑えている．電気インタフェースはCMOSの高速動作I/Oでは一般的なLVDSとP-ECL（pseudo-emitter coupled logic）インタフェースを搭載し，他のコンピュータ内部回路との接続を容易にしている．最大伝送距離は100mである（リボンファイバのスキューで制限される）．

■高速スイッチLSI[4), 5)]

RHiNET-2/SWに搭載したスイッチLSIは，単一チップで8Gbps×8ポート（合計64Gbps）の大容量スループットを実現する．これはCMOSのネットワークスイッチLSIとしては世界トップクラスの性能である．**図-5**にRHiNET-2/SWに搭載したスイッチLSIの内部ブロック図を示す．I/Oピンは800MbpsのLVDS差動信号線を192組（384本）である．0.18μmプロセスのCMOS-ASICの使用により，大容量SRAM（512kbyte）のオンチップ搭載を実現している．大容量オンチップメモリは高速メモリアクセスを実現し，低遅延スイッチ動作を可能にしている．CMOSの使用は，大容量オンチップメモリの実現，スイッチLSIの低コスト化，他のPC内部回路との互換性向上などの面で大きなメリットを有する．

■高速デバイスの回路・実装技術

RHiNET-2/SWは，1つのCMOSスイッチLSIと送受8対の8.8Gbps光インタコネクションを単一ボード上に高密度実装し，8入力8出力のクロスバースイッチ機能（通信容量: 64Gbps）を実現する（**図-6**参照）．ボードの中心にCMOSスイッチLSIを搭載し，その直近に送受8対の光インタコネクションモジュールを高密度実装して8×8のクロスバースイッチを実現している．8つの各ポートは800Mbps×10bitの並列データ（残りの2チャネルはクロックとフレーミング信号）を並列同期入／出力する（総データ容量: 8Gbps/port）．信号入出力は1ピン当たり800Mbpsの高速動作をLVDSを用いて実現している．本高速動作を実現するため，実装ボードの配線レイアウト・各デバイスの実装は，別途開発したテストボードを用いて信号のロス・スキューを実測し，その結果に基づいた最適設計を実施した[5)]．

◆ノード間接続に適した光インタコネクションの将来◆

並列リボンファイバを使用した並列光インタコネクションは近年格段に進歩し，2.5Gbps×12bit（総容量30Gbps）のモジュールも10Gbit-Ethernet[6)]のソリューションとして提案されている（接続距離300m以下．数mの至近距離は電気ケーブルの方がコスト的に有利であり，300mから長距離は，ファイバコストからみてシリアルファイバが主流と考えられている）．さらに並列リボンファイバを使用せずシリアルファイバと波長多重を用いた方式なども10Gbit-Ethernet技術の1つとして開発されつつあり，合分波器の低コスト化が進めば，有力なソリューションの1つとなる．今後の計算機システムにおけるノード間接続方法は，接続距離・回路規模・遅延時間・コストなどの複合的要素を踏まえ，さまざまなソリューションを組み合わせることが可能になる．将来的には，10Gbpsの伝送系を並列使用することで100Gbpsクラスの光データ接続も実現可能である．光インタコネクト技術を積極的に応用していくことにより，今後も一層，計算機ネットワークは高速化は加速していくと考える．

参考文献
1) http://www.myrinet.com
2) 刀祢平高一朗, 三浦　篤, 高井厚志, 上野　聡, 内田勝己, 豊中隆司, 斉藤勝美: 800Mbit/s/ch×12ch光インタコネクト受信モジュール, 電子情報通信学会ソサエティ大会, SC-4-2 (1999).
3) 三好一徳: 並列光インタコネクション用622Mbit/s×12ch送受信モジュール, 電子情報通信学会第1回光インタコネクション情報処理研究会, OIP99-8 (1999).
4) Nishimura, S., Harasawa, K., Matsudaira, N., Akutsu, S., Kudoh, T. Nishi, H. and Amano, H.: RHiNET-2/SW: A Large-throughput, Compact Network-switch using 8-Gbps Optical Interconnection, New Generation Computing, Vol.18, pp.188-197 (2000).
5) Kudoh, T., Yamamoto, J., Sudoh, F., Amamo, H., Ishikawa, Y. and Sato, M.: Memory Based Light Weight Communication Architecture for Local Area Distributed Computing, Innovative Architecture for Future Generation High-performance Processors and Systems, IEEE Computer Society Press, pp.133-139 (1997).
6) IEEE802.3 Higher Speed Study Group, http://grouper.ieee.org/groups/802/3/

(平成12年8月1日受付)

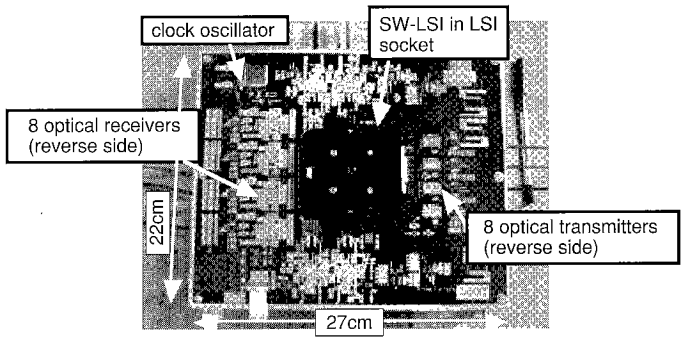

図-6　マザーボードの外観