論 文 の 内 容 の 要 旨

論文題目　大容量光接続ネットワークスイッチ設計・構築技術とその並列計算機システムへの応用

氏 名　　西 村 信 治

1.　新しい光計算機ネットワークの必要性

インターネットなどの情報通信の革命的な進歩の中で、計算機ネットワークを流れる信号の変調速度もギガヘルツ帯を超えつつあり、従来のネットワークで用いられてきた電気信号配線では、帯域、入出力ピン数、消費電力などが限界に来ている。この電気配線のボトルネックを解決し、大容量・長距離接続を容易に実現するには、光技術のより積極的な活用が強く期待されている。

光技術は基幹系からイーサネットまで、既に幅広く用いられている。特に基幹伝送系の光技術は、10 Gbit/s を超える帯域と数 km の接続が可能なデータ伝送を既に実現している。しかし、この基幹系光伝送技術を計算機ネットワークにそのまま適用した場合、信頼性確保のために複雑な温度制御回路などが必要となり、小型化と経済性の向上が困難である。そこで、次世代の高速で柔軟な計算機ネットワークには、高速・低遅延なデータ伝送を簡素な回路規模で実現する新たな光ネットワーク技術が必要になる。

2.　高速光ネットワークスイッチ

次世代（1～10 ギガビットクラス）の計算機光ネットワークの実現には、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、特に大きな課題となる。本研究では、このネットワークスイッチの構成方法として「光・電気混載ネットワークスイッチ」を採用した。本構成は、入出力に光インターコネクションを使用し、スイッチ処理に電気信号処理 LSI を用いるハイブリッド実装にてスイッチを実現する。光インターコネクションの使用により、信号の高速・長距離伝送が可能になり、さらに電気信号にて信号処理する事で高機能なスイッチ処理が実現できる。本研究では、この高速光・電気混載スイッチの実現に必要な設計・構築技術を開発し、実機でのテストを通じてその有効性を実証した。

3.　光・電気混載ネットワークスイッチの実現に必要な技術

3.1　並列光インターコネクション

高速光ネットワークには、大容量・長距離データ伝送を小型かつ低コストで実現できる光技術が必要である。本研究では、シリアル、並列、および WDM の各光技術を比較検討し、並列光インターコネクションを採用した。並列インターコネクションは、実現の容易な低速駆動のデバイスを並列駆動する構

造ゆえ、数 100m までの短い距離の接続を前提に同じ通信容量（～10 Gbit/s）で比較すると、シリアルデータ通信よりも高い経済性と信頼性を実現できる。

3.2 大容量 CMOS スイッチ LSI

大容量信号処理を実現するには、光インターコネクションの能力を最大限に生かせる新しい高速スイッチ LSI が必要である。光インターコネクションの使用により、スイッチ LSI の各入出力バッファを軽い入出力負荷を前提に設計でき（数 10m の長距離・大負荷の電気ケーブルを駆動する必要が無いため）、より高速な入出力回路を LSI に搭載できる。その結果、駆動能力ではバイポーラより劣る CMOS を積極的に使用する事が可能になり、CMOS 化のメリットである低消費電力、同一チップ上のメモリ搭載、さらに低コストが実現できる。本研究においては、並列光インターコネクションの使用を前提に、光インターコネクションに適した専用のスイッチ LSI 技術を ASIC テクノロジーを用いて設計開発した。

3.3 高速回路

光・電気混載スイッチにおいては、装置内の LSI と光デバイス間を接続する高速信号回路のタイミング設計と、並列チャネル間のスキュー管理が重要な設計事項になる。

特にタイミング設計においては、高速動作回路の短いクロック周期の中で十分な位相マージンを確保するための、精密な設計が必要である。本研究においては、出力側データチャネル間のスキュー、出力データ信号のジッタ、入力バッファの要求するセットアップ・ホールドタイム、クロックジッタ、プロセスや電圧のばらつき係数の 5 項目を積算し、位相マージンを求める計算方法を用いた（図 1）。個々のデバイスの実測評価結果に基づいて本 5 項目を積算することで、高精度なタイミング設計が実現できる。

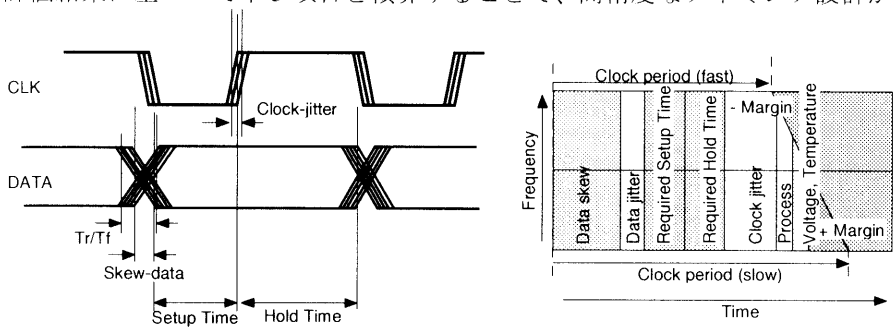

図 1: タイミングマージン（Margin）の積算方法

そして、タイミング設計の検証のため、実際のビットエラーレート測定の結果を基に、Q ファクタを測定する事で、理論上実現可能なビットエラーレートの最良値 BER_{margin} を近似的に導いた（式 1）。

$$BER_{margin} \cong \frac{e^{-(Q_m^2/2)}}{Q_m\sqrt{2\pi}} \quad \text{但し} \quad Q_m = \frac{\mu_1 - D - \mu_0}{(\sigma_1 + \sigma_0)} \qquad \text{式 1}$$

μ_0, μ_1 は立上・立下りエッジの平均時刻、σ_0、σ_1 は立上・立下りエッジのランダムジッタの分散値
D は次段が必要な位相マージン。

3.4 高速信号配線の高密度実装技術

本研究がターゲットとする総通信容量数 10 Gbit/s のスイッチをワンチップに実装するには、1 Gbit/s クラスの信号線を、差動信号なら 200～400 本程度 1 チップに集中配線する事になる。この際、信号の伝搬損失、クロストーク、同時スイッチングノイズが実現の障害となる。本研究においては、伝搬損失・クロストーク等に対して、実際に使用する基板材料を使用してその特性を測定し、そのデータに基づいて設計ルールを決定する手法を構築した。さらに、同時スイッチングノイズなどのノイズ対策として、電源レイアウトの最適化と、ノイズ源のフィルタ分離、電波吸収材の配置などを検討・実施した。

4．実例としてのシステム

光 RWC-1 と RHiNET の両並列計算機システムの開発を通じて、設計・構築技術の有効性を実証した。

4.1 光 RWC-1

交換機や汎用計算機向けの大規模データ伝送技術である光インターコネクションの新規応用として、RWC つくば研と共同で、光接続並列計算機(光 RWC-1)の設計開発を行った。光 RWC-1 では、超並列計算機 RWC-1 のデータ転送部の一部を光化する試みを実施した。本装置は 8 つの並列ノード間を通信容量 2.4 Gbit/s の光インターコネクションで接続した構成を有する。光 RWC-1 は大容量光データ伝送を実機搭載した計算機として世界初の試みとなる。本機は通信制御 LSI の制御方式や光接続系の構成を光インタ

ーコネクションに最適化する形で設計改良する事で、大容量・長距離なデータリンクを小型・低消費電力に実現できる構造になっている。4 ノードを搭載したプロセッサボード（図 2）は、日立製の 12 チャンネル光モジュールを 48 個用い、スループット 9.6 Gbit/s の光通信機能を有する。ベンチマークテストの結果、ファイバ(50 m)で接続した本装置と電気ケーブル(10 m)で接続した同種機 RWC-1（Elec. RWC-1）と同等の性能を実現した（図 3)。これは、低遅延な光インターコネクションとレイテンシーを隠蔽するアーキテクチャの導入により、処理能力の距離依存性を小さくした結果と言える。

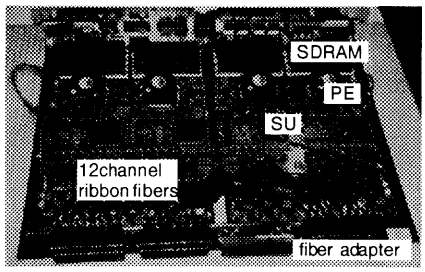

図 2: 光 RWC-1 のマザーボード

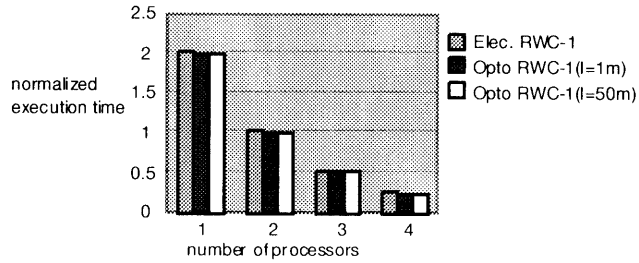

図 3: ベンチマークテストの結果

4.2 RHiNET (RWCP high-performance network)

PC クラスタの内部ネットワークに光インターコネクション技術を応用する事により、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（LASN）が実現できる。この事を実証するため、LASN でノード間を接続した並列分散計算機システム RHiNET-2（図 4）を RWC つくば研と共同で開発した。本研究では、RHiNET-2 に使用する光インターコネクションを搭載した高速ネットワークスイッチ RHiNET-2/SW（図 5）における、光インターコネクションと高速スイッチ LSI の高速回路やボード実装を含む、ハードウェアの方式・回路および実装設計を行った。

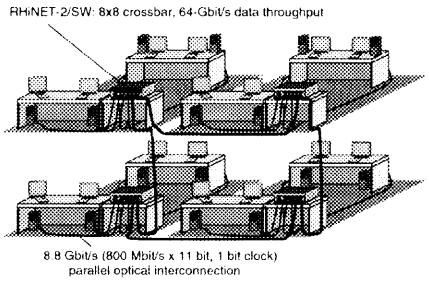

図 4: RHiNET-2 システムの概念図

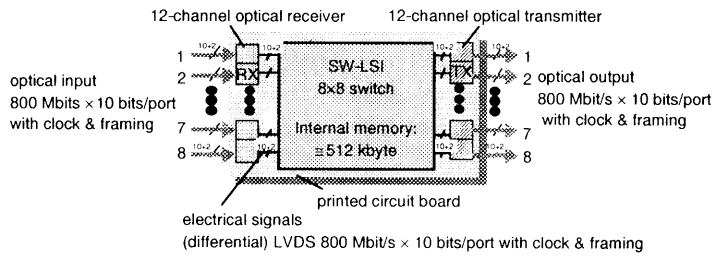

図 5: ネットワークスイッチ RHiNET-2/SW（総通信容量: 64 Gbit/s）の概念図

RHiNET-2/SW は、大容量スイッチ LSI と高速光モジュールをワンボードに実装し、各ポートは大容量 8 Gbit/s のスループットを有する 8×8 スイッチシステムである。

実現したハードウェアの特徴は以下の通りである。

(1) 大容量並列光インターコネクション

RHiNET-2/SW にて使用した日立製・光送受信モジュールは 12 チャネルのレーザーとホトダイオードを並列駆動する事により、8.8 Gbit/s の大容量光データ接続を実現する（800 Mbit/s×データ 11 チャネル+クロック 1 チャネル)。

(2) 高速 CMOS-スイッチ LSI

RHiNET-2/SW 用に開発したスイッチ LSI は、8 入力 8 出力のクロスバースイッチ機能を搭載し、各ポートは 8 Gbit/s の高速通信容量を実現する（図 6)。チップ全体では 64 Gbit/s の大容量スループットを実現し、計算機ネットワーク用の CMOS スイッチ LSI としては、世界最高クラスの通信容量をもつ。全ての高速電気インタフェースは高速 2.5V-LVDS（low voltage differential signaling） で統一し、高速（800 Mbit/s）差動入出力ピンを、192 組（384 本）入出力する構成を持つ。0.18 μm プロセスの CMOS-ASIC を使用し、大容量 SRAM（512 kbyte）をオンチップ搭載した。

(3) 800 Mbit/s の高速クロック回路の高密度実装

マザーボード（図 7）には、8×8 スイッチ LSI（784 ピン）1 個と 800 Mbit/s×12 チャネル光送受信モジュール計 8 対を高密度実装した（高速 800 Mbit/s-LVDS 信号線の総数は、384 本)。高速回路の実現のために、スイッチ LSI と光モジュール間を接続する回路に対して、高精度なタイミング設計とスキュー管理を実施した。また高速回路の高密度実装を実現するために、まずテスト基板にて配線のロス・反射・

クロストークなどの物理データを実測評価した。そして、その結果に基づいて設計ルールを構築し、その設計ルールに乗っ取ってボードの実装・レイアウト設計を行った。

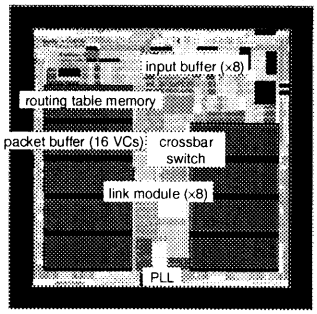

図 6: スイッチ LSI のフロアプラン

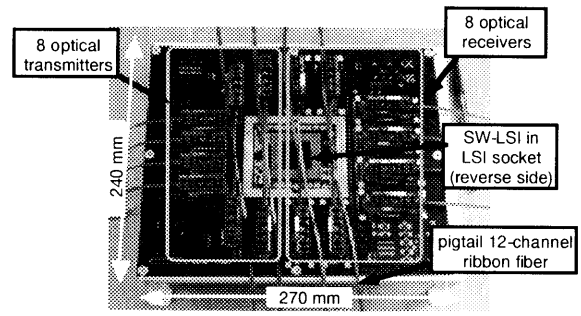

図7: マザーボードの外観図

(4) エラーレート測定

信頼性評価として、RHiNET-2/SW のスイッチシステム全体を通過後の信号のエラーレートを測定した。測定結果によると、エラーレートカーブ上で十分に開口系を確認でき、10^{-11} 以下のビットエラーレートを十分な位相マージン（890 ps）をもって確認した（図 8)。そして、ノイズ成分がガウシアン近似できる事を前提にビットエラーレートを直線近似すると、ビットエラーレート 10^{-20} の実現に際し 790 ps の位相マージンを持つことが推定できた。さらに、本実験結果をQファクタに換算して評価する事で、実現可能な最良のビットエラーレート BER_{margin} を計算した結果、約 1.0×10^{-30} の値を得（数 10 年のエラーなしに相当)、本近似計算では考慮していない経時劣化や装置故障の影響を考慮しても、実用上、十分な動作上の余裕を持つことが判った。

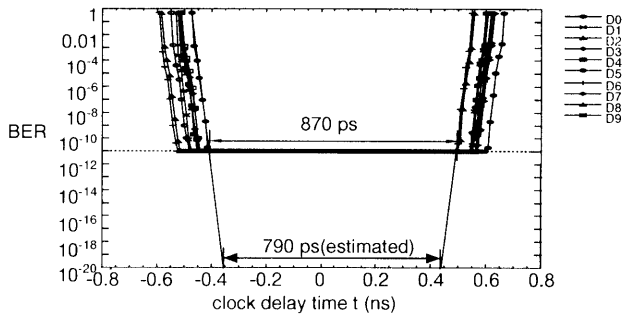

図 8: RHiNET-2/SW のエラーレート測定結果

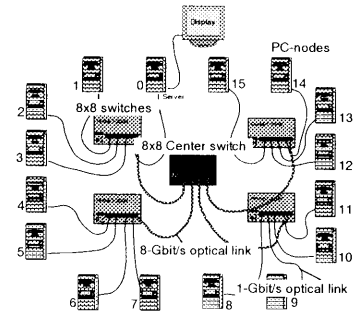

図 9: 16 ノードシステム

(5) 計算機システムとしての試験

16 台の PC と 5 台の RHiNET-2/SW を接続して 16 ノードクラスタシステムを構成し、並列計算機システムとしての動作試験を行った。本システムでは、16 台の PC と 5 台のスイッチを用いて、図 9に示すようなツリー型のネットワークを構成した。PC ノードに搭載した NIC(Network Interface Card)に PCI バスを用いたため、PC ノードと SW 間は 1 Gbit/s で接続し、SW-SW 間のみ 8 Gbit/s で接続している。計算機システムの実機使用時の機能を検証するため、本システムにて並列計算処理プログラム（レイトレーシング）を実行した。その結果、学会期間中の 4 日間に渡って安定した動作を確認でき、本スイッチが並列計算機システムにおいて安定して使用可能である事を、実機で確認した。

以上の 2 システムを用いた実機実証により、本研究の成果となる高速ネットワークスイッチの設計・構築技術の有効性が実証できた。

これらの 2 つのシステムにおいて実証できた光インターコネクションの技術的メリットは、以下の 5 項目が挙げられる、(1) 高速データ入出力、（2）長距離接続、（3）装置の小型化と実装の容易化、（4）実機使用に耐える高信頼性、（5）並列システム内部で使用可能な低遅延データ接続。

5． 高速ネットワークスイッチにおける光インターコネクションの将来

今後 100 ギガを超える大容量通信などでは、今以上に並列光インターコネクションが重要な技術となる（OEIC 技術や WDM を取入れる事も重要)。この将来の高速光ネットワークを、システム開発していく上でも、本研究の成果であるタイミング設計、スキュー管理、ノイズ対策、テスト手法などのネットワークスイッチの設計・構築そして評価技術は、有効な技術になると確信している。