計算機アーキテクチャ　138－4
（2000．4．26）

光インタコネクト搭載ネットワークスイッチ

西村信治、工藤知宏*、西宏章*、原澤克嘉**、松平信洋**、坏重人**
RWC光インターコネクション日立研究室

東京都国分寺市東恋ヶ窪1-280　（株）日立製作所中央研究所内
Tel:0423-23-1111 (ext. 3560), Fax:0423-27-7740
E-mail:nisimura@crl.hitachi.co.jp
*RWCつくば研究センタ, RWCP並列分散システムアーキテクチャ研究室
**日立通信システム

要旨

分散環境に配置したパーソナルコンピュータを高速光ネットワークにて接続することにより高性能な並列計算機システムが構築できる。これを実証する為、大容量光インタコネクションと高速スイッチLSI（LVDS-I/O; 784 ピン-BGA; ワンチップCMOS-LSI）を組み合わせた大容量8×8スイッチシステムを（RHiNET-2/SW）開発した。RHiNET-2/SWは分散環境での高速並列計算処理に必要な、大容量（64 Gbit/s）パケットデータスイッチ機能を小型装置で実現する。光インタコネクトの使用により、ノード間接続距離は最大100mを実現している。

和文キーワード:光インタコネクション、並列計算機、ネットワークスイッチ

Network switch implemented with optical interconnection

Shinji Nishimura, Tomohiro Kudoh*, Hiroaki Nishi*,
Katsuyoshi Harasawa**, Nobuhiro Matsudaira**, and Shigeto Akutsu**

RWC Optical Interconnection Hitachi Laboratory
c/o Central Research Lab., Hitachi, Ltd.,
1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo
Tel:+81-423-23-1111 (ext. 3560), Fax:+81-423-27-7740
E-mail:nisimura@crl.hitachi.co.jp
* RWCP Distributed and Parallel System Architecture Tsukuba Laboratory, RWC Tsukuba research center
** Hitachi Communication Systems Inc.

Abstract

A network based parallel computing system can be constructed by optically interconnected PC network. We have developed a high-throughput, compact network switch (the RHiNET-2/SW) for a distributed parallel computing system. Optical interconnection modules and one-chip CMOS SW-LSI (LVDS-I/O, 784-pin BGA package) are integrated on a compact RHiNET-2/SW circuit board. The RHiNET-2/SW switch enables large-throughput (64 Gbit/s), packet data switching for a high-speed parallel processing in a distributed computing environment. Optical interconnection provides long-transmission-length (< 100 m) interconnection between the nodes.

Keywords:Optical interconnection, Parallel computer, Network switch

1　はじめに

パソコン（PC)やワークステーション(WS）の単体性能は近年著しく向上しており、ハイエンドの超並列計算機のノード性能に匹敵もしくは上回るようになっている。このため高速光ネットワークでPC・WS間を接続し並列計算システムを構成した場合、非常に高い計算機能力を得られる可能性がある。

我々は本可能性を実証するため分散並列計算機システムRHiNET-2 (RWCP high-performance network-2)の開発を進めている。RHiNET-2システムは、長距離を広帯域接続できる光インターコネクションをPCノード間接続に採用し、ビル内フロア内に分散配置された計算機類を自由に接続して高性能な並列計算処理を実現する事を目標としている（図1参照）[1-3]。高性能PCがオフィスや研究機関等に数多く配置されるようになった現状を考えると、一箇所に集められた計算機だけでなく、ビル・フロア内などに広く分散配置された計算機群を効率良く使用することで、「オフィスを丸ごとスパコンとして使う」ような形態の並列コンピュータシステムが実現可能になると考えている。

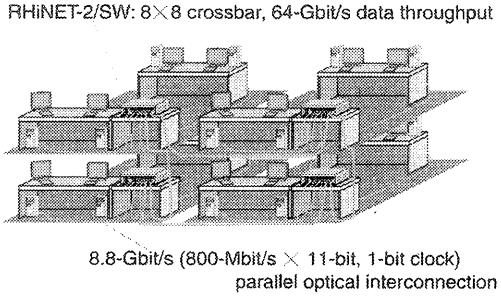

図1　RHiNETシステムの概念図

RHiNET-2システムにおいては、各PCはPCIバスベースのNIC（network interface card)を搭載し、8.8Gbps光インターコネクションを介して8×8の専用クロスバースイッチ（RHiNET-2/SW：図2参照）に接続される。本システムに用いた光インターコネクションは800Mbps×12ビットの並列同期伝送（内訳：データ11ビット、クロック1ビット）を実現し、接続距離は最大100mと電気ケーブルの10倍を有する。また搭載されたスイッチLSIには、大容量のSRAM(512kbyte)をオンチップ搭載し、メモリアクセス時間を最小化を図っている。本稿では、RHiNET-2システムの中心装置である8×8スイッチボード（RHiNET-2/SW：通信容量64Gbps）に関して、構造と動作試験結果に関して述べる。

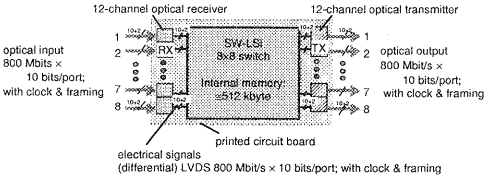

図2　RHiNET-2/SWの構造

2 RHiNET-2における光インターコネクション

RHiNET-2は並列コンピュータシステムである。この為、通信容量の拡大ばかりでなくノード間通信のレーテンシーを最小限にする必要がある。また、通常のPCをノードしとして使用しているため、光素子の消費電力や物理的サイズも小さいものが要求される。

RHiNET-2においては、リボンファイバを用いた並列同期転送を採用した[4-7]。用いた光モジュールは日立製作所製のインタコネクトモジュール[MDS4212A/MDR4212A]である[4-6]。本モジュールは12チャネルのレーザとホトダイオードを並列駆動する事により、小型パッケージ（3.9cc）で8.8Gbpsの大容量光データ接続を可能にしている（800Mbit/s×データ11チャネル+クロック1チャネル）。発振波長1310nmの端面発光レーザアレイを使用し、12芯シングルモードリボンファイバ（MPO [multi-path push-on]コネクタ付）と組合せる事により、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最小化している。電気インタフェースはLVDSとP-ECLインタフェースを搭載し、コンピュータ内部回路と容易に接続できる。ファイバ接続距離は最大100mを実現しており（制限要因はファイバスキュー）、電気ケーブル（限界距離: 約10m）よりも自由度の高いノード間接続が可能である。また、搬送クロックを並送する並列同期伝送の採用により、受信側でフレーム同期やクロック生成等の複雑な信号処理を必要とせず、同期系で構成されるコンピュータシステム同士を接続するのに適したコンパクトで低遅延なデータ転送を実現している。

1Gbpsクラスの伝送速度を有する標準技術Gigabit Ethernet[8, 9]・Fiber Channelにおいては、送受各々において1本のファイバと1つの波長信号を用いたシ

リアルデータ転送を採用している。しかし、シリアルデータ転送にて、より高速なデータ転送レートを実現する為には厳密な温度管理が必要となり、結果として送受信モジュールの大型化と消費電力の増大をもたらす。この為、現在のデバイス技術では、厳密な温度管理なしに実現できるデータレートは、最大2.5 Gbps前後と考えられている。それ故、2.5 Gbpsを超える伝送レートを、装置間接続等の近距離において小型装置サイズで実現するには、複数のデータ転送系を用いた並列データ転送がもっとも有利であると考える。

3　RHiNET-2/SWシステムの構造

(1)　スイッチLSI

RHiNET-2/SWに使用したスイッチLSIは8入力8出力のクロスバスイッチ機能を搭載している。各ポートは10ビットのパケットデータと1ビットのフレーム同期信号、1ビットのクロック信号の12ビット構成を有し、8Gbps/portの通信容量を実現できる。前節の高速動作とCMOS化に対する要求を満たすため、全ての高速電気インタフェースを高速CMOS-LVDS (low-voltage differential signaling)で統一した。I/Oピンは800Mbit/sの差動信号線を192組（384本）入出力する構成とし、単一チップで64Gbpsの大容量スループットを実現している（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）。0.18 μ m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。LSIパッケージは784pin BGA[ball grid array]である。

図3にLSIの内部ブロック図、図4はスイッチLSIのダイのレイアウトを示す。800Mbps×10bitsの入力信号は、入力端で1:8DEMUXにて100Mbps×80bitsに展開された後、ECCデコーディングとエラスティックバッファを通過する。搬送波クロックに同期した入力信号はエラスティックバッファ内で内部クロック（BASECLK: 100MHz）に再同期される。エラスティックバッファからの入力信号（80bit）はスイッチコア部（SW-core）にて行路切換処理される。スイッチコアからの出力信号はECCエンコーディング処理を経て、8:1MUXにて800Mbps×10bitsの信号に多重化され、フレーム同期信号(AO)と搬送波クロック(OUTCLK)を付加されてLSIから出力される。LVDS入出力バッファ部は800Mbpsを実現するため、高速特性を要求される。

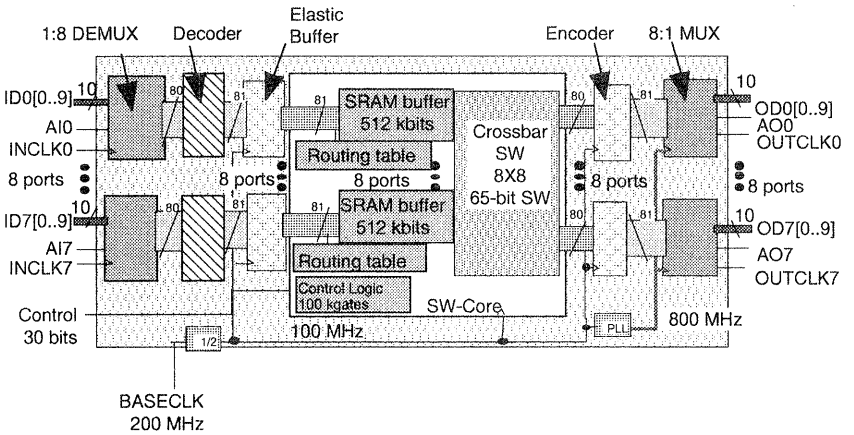

図3　LSIの内部ブロック図

Routing Table (8 ports): 2 bit * 4096- for single cast+ 4 bit * 64- for multicast

ID0[0..9], ID1[0..9], ..., ID7[0..9]: input data (LVDS); 10 bits*800 Mbit/s

OD0[0..9], OD1[0..9], ..., OD7[0..9]: output data (LVDS); 10 bits*800 Mbit/s

INCLK 0..7: input clock (LVDS); 400 MHz,

OUTCLK 0..7: output clock (LVDS); 800 MHz

AI0..7, AO0..7: framing (LVDS) 800 Mbit/s

RHiNET-2を含めた高速計算機システムのノード内の計算処理は電気信号にて処理される。このため、各ノードのデータ入出力部においては光・電気信号変換処理が必要となる（光信号のままでのスイッチング技術[11,12]も様々開発されているが、通信路の設定に必要な時間が最低でもマイクロ秒オーダの時間を要し、LAN装置には使いにくい）。この為、高速な光I/Oを有するネットワークスイッチには、高速電気・光インターフェースが必要となる。また同時に、メモリアクセスのボトルネックを回避するためにロジック回路と大容量メモリのオンチップ集積が必須となる事から、構成回路のCMOS化が求められる。

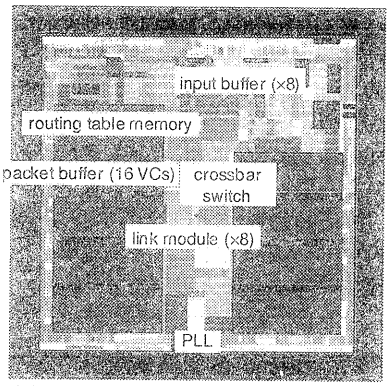

図4　LSIダイ内部のブロック配置図

(2) .RHiNET-2/SWのボードレイアウト

RHiNET-2/SWにおいては、ボードの中心にCMOSスイッチLSIが搭載され、その直近に光インタコネクションモジュール送受8対を高密度実装して8×8のクロスバースイッチを実現している（図5, 6参照）[12]。各ポートは12bit並列データを速度800 Mbpsで並列同期入/出力できる(ビット内訳：パケットデータ信号10 bit、搬送クロック1 bit、フレーム信号1 bit、総データ容量：8 Gbps/port)。ボードサイズは280×366mmである。

LSIは高速デバイス専用のソケットに搭載されている。このLSIソケットは、帯域（DC-6GHz）、path-inductance <1 nH、capacitance[singal-to-signal] < 1pFと優れた高周波特性を有する。スイッチLSIの制御用にRS-232Cインタフェースを持つH8マイコンのドータボードを搭載しており、ルーティングテーブル等の書換はRS-232C, H8ボードを介して外付けのPCから行う。周波数200MHzの内部クロック源として水晶発振器を1つ搭載している。LSIのピン配置およびボード内配線のレイアウトは、前述のテスト基板による実測結果を踏まえ、プリント基板配線の最短化（最長150mm）と交差数0の制約条件の元に最適化設計している。

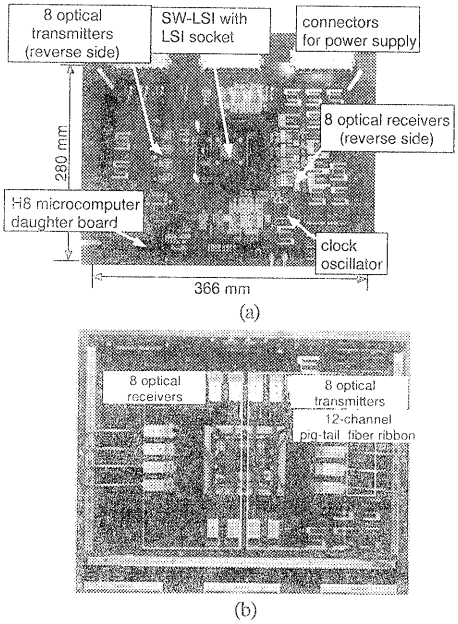

図5　RHiNET-2/SW, マザーボードの部品レイアウト (a): LSI搭載側、(b): 光モジュール搭載サイド

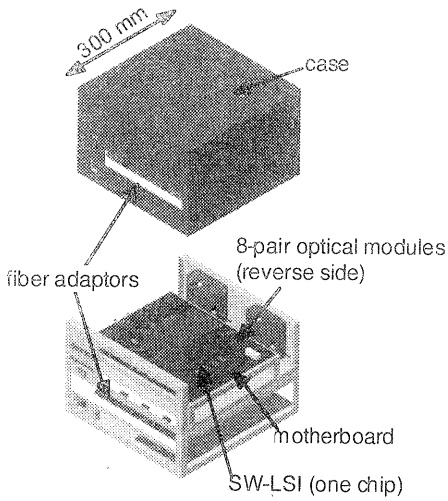

図6　RHiNET-2/SWの概観図

4　RHiNET-2/SWシステムの評価結果

RHiNET-2/SWの各入力ポートに800 Mbps×10bitのPRWS信号を入力し、出力信号のアイパターンの観測し、ビットエラーレート（BER）を測定した。図7に評価系の構成を示す。800Mbps×12チャネル（クロックCLK, フレーム信号AI, データDI[9..0]）の同期信号をデータジェネレータ（HP80000)にて生成し、光インターコネクションモジュール[MDS4212A]にて光信号に変換したのち、50mのリボンファイバを介してRHiNET-2/SWの入力ポートに入力した。RHiNET-2/SWの出力ポートからの光信号は、同じく50mのリボンファイバを通過の後、光インターコネクション受信モジュール[MDR4212A]にて電気信号に変換され、ロジックアナライザ（HP81200-LA）とオシロスコープ（HP54120）にて観測した。

ロジックアナライザの入力段の信号のアイパターンを見ると（図8参照）Tr/Tf：150 ps、ジッタ< 100 psの良好な波形を得られた。またBERはデータレート800 Mbpsの条件で10^{-11}以下の高信頼性を実現した。本評価結果はRHiNET-2/SWのデータI/O系の信頼性は、並列計算機システムを構成する上で要求されるものを十分満たしていると考える。

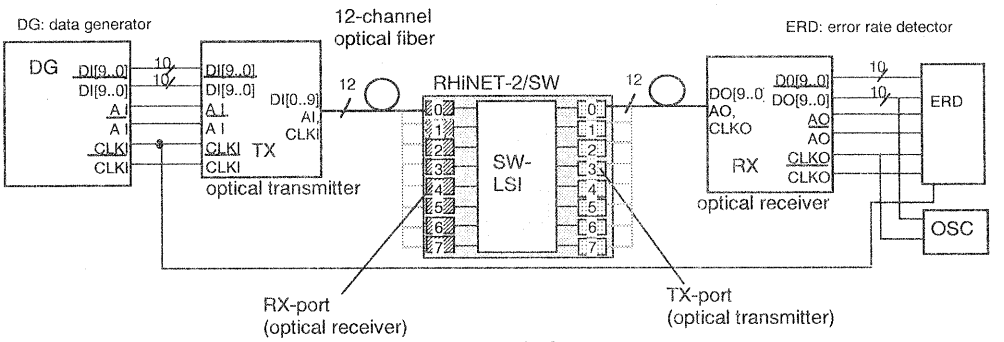

図7　測定系

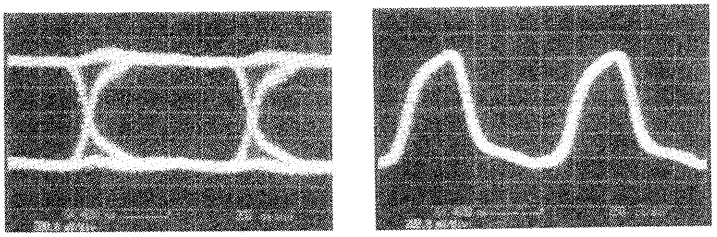

図8　アイパターンの測定結果
0データビット[D0], クロック[CLK]
250 ps/div, 200 mV/div

5 まとめ

RHiNET-2システムはオフィスLAN等の分散配置された計算機環境において、高性能な並列計算処理を実現できる。8.8Gbps大容量光インタコネクション8対と64Gbps高速スイッチLSI（LVDS-I/O; 784ピン-BGA; ワンチップCMOS-LSI）を組み合わせた大容量8×8スイッチボード（RHiNET-2/SW）開発した。本装置は分散環境での高速並列計算処理に必要な、大容量（64 Gbps）パケットデータスイッチングを、小型装置で実現可能とする。並列同期光インターコネクションを採用することにより、小型装置規模・低遅延でスキューの小さいデータ転送を実現している。高速動作素子をボード上に高密度実装するため、MWB™を用いた高速・高密度配線技術を開発し、RHiNET-2/SWボードに応用した。全ての電気インタフェースはCMOS-LVDSに統一し、高速動作・低コスト化を容易にしている。本プロトタイプシステムRHiNET-2/SWにおいてBER測定による信頼性評価試験を実施した結果、データレート800Mbps動作時においてBER 10^{-11} 以下の高信頼性を確認した。RHiNET-2/SWネットワークシステムを用いることにより分散配置環境での高性能並列計算機能が実現可能である。

参考文献

[1] T. Kudoh, J. Yamamoto, F. Sudoh, H. Amamo, Y. Ishikawa, and M. Sato: "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).

[2] S. Nishimura, H. Inoue, H. Matsuoka, and T. Yokota: "Optical interconnection subsystem used in the RWC-1 massively parallel computer", IEEE Journal of selected topics on quantum electronics, vol. 5, pp. 360-367, (1999).

[3] T. Yoshikawa and H. Matsuoka: "Calibration-free parallel optical-interconnection subsystem implemented by a GByte/s-array optical tranceiver and a one-ship link LSIs (Invited)", Technical Digest of OC'98, pp. 524-527, Brugge, Belgium, (1998/6).

[4] A. Takai, T. Kato, S. Yamashita, S. Hanatani, Y. Motegi, K. Ito, H. Abe, and H. Kodera: "200-Mb/s/ch 100 m Optical Subsystem Interconnections Using 8-Channel 1.3-μm Laser Diode Arrays and Single-Mode Fiber Arrays", J. of Lightwave Technology, vol. 12, pp. 260-270, (1994).

[5] 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: "800 Mbit/s/ch×12ch 光インタコネクト受信モジュール"、1999電子情報通信学会ソサエティ大会、SC-4-2, (1999).

[6] http://www.hitachi.co.jp/Prod/tcd/hikari/tjn00634.htm

[7] 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回光インターコネクト情報処理研究会, OIP99-8, (1999).

[8] HIPPI-6400 working drafts, T11.1 maintenance drafts of ANSI NCITS

[9] IEEE802.3 Higher Speed Study Group http://grouper.ieee.org/groups/802/3/

[10] Keishi Habara, Tohru Matsunaga, and Ken-ichi Yukimatsu: "Large-Scale WDM Star-Based Photonic ATM Switches", Journal of Lightwave Technology, Vol.16, No.12, pp.2191-2201, (1998/12).

[11] S. Kitajima, H. Takano, and M. Kobayashi: "Traveling type optical cell buffer with small variation of output cell level", IEICE Trans. of Communications, vol. E82-B, pp. 281-287, (1999).

[12] S. Nishimura, K. Harasawa, N. Matsudaira, S. Akutsu, S. Sasaki, T. Kudoh: "Network switch using optical interconnection for high performance distributed parallel computer using PCs", Proceedings of the Sixth International Conference on Parallel Interconnects , pp. 5-12, Anchorage, (1999/10).