超並列計算機RWC-1における光実装技術

西村 信治*

Optical Interconnection Subsystem for the RWC-1 Massively Parallel Computer

Shinji NISHIMURA*

*RWCP光インタコネクション日立研究室 (〒185-8601 東京都国分寺市東恋ヶ窪1-280 株式会社日立製作所中央研究所内)

*RWCP Optical Interconnection Hitachi Lab. (c/o Central Research Lab., Hitachi Ltd. 1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601)

1. はじめに

光インタコネクションは装置内・装置間において、大容量データを複数の光ファイバ・光送受信系を用いて並列転送する技術である。本技術は大容量データ転送を軽量光ファイバ・小型デバイスで実現し、しかも優れた耐EMI性能を有し、接続装置間のGNDの分離も可能となる。光インタコネクションの導入により装置内・装置間において大容量・柔軟なデータ接続が実現でき、通信ボトルネックの解消が可能になる^{1)~7)}。

システム構成から考えると光インタコネクションは装置内インタコネクション (Interboard Optical Interconnection, 別称PSI: Proprietary System Interconnect) と、装置間インタコネクション (SAN: System Area Network) に分類される。装置内インタコネクション (PSI) は同一システムのボード間等における低遅延・大容量データ転送技術を指す (Fig. 1, 2 参照)。PSIは接続される両者間の協調関係が密接な分、大容量性と共に転送の低遅延性が強く要求される。装

Fig.1 Interboard optical interconnection

Fig.2 Broadband connectivity technologies

置間インタコネクション (SAN) はクラスタ・サーバ間接続やディスクストレージ接続等における大容量データ転送を指す。SANは接続距離が長いので、低遅延性への要求は小さいが、同期信号間の伝送遅延時間差 (スキュー) がシステム性能上の問題となる。

超並列計算機は千台規模のプロセッサノード間を高速接続して計算処理を行う。プロセッサ単体の処理能力が飛躍的に向上した今日においては、プロセッサをつなぐネットワークの性能が最大の性能ボトルネックとなりつつある。われわれは、この超並列計算機内のノード間接続に光インタコネクションを採用することで、装置内ネットワークにおける通信ボトルネックの解消を図った。超並列計算機のノード間接続はPSIの領域に分類でき、数Gbps以上の大容量データ転送能力と数10mの接続距離が要求され、さらに低遅延なデータ転送、バースト信号等にも対応した高い透過性、高い通信信頼性等が要求される。特に信頼性に関して考えると、1000台規模のプロセッサノード間をそれぞれ転送容量1Gbpsのバスで接続し、1年以上の無障害運転を保証すると仮定した場合、ネットワーク全体に要求される信頼性は、ビットエラーレートに換算すると 10^{-20} 以下という高い信頼性に相当する。われわれは光インタコネクションに要求されるこの高性能・高信頼性を実証するため、実際に装置作成をして機能検証を行った。

2. 超並列計算機RWC-1⁸⁾

超並列計算機RWC-1は、RWCPつくば研究センタで開発中の独自アーキテクチャを持った計算機システムである。最大1024台のノードを直接網接続した構成を有し、1ノードは1つのプロセッサ (PE) とマスター・スレーブ2つの通信制御用スイッチングユニット (SU) から構成される (Fig. 3 参照)。各ノードは送受2系統ずつのI/Oポートを有し、それぞれ24ビットのデータと12ビットの制御データの並列同期信号から構成される。クロック周波数は100MHzであり、各ポートのスループットは2.4 Gbit/s (300 MByte/s) となる。われわれは電気ケーブル接続を前提に

Fig.3 Schematic structure in the RWC-1 massively parallel computer

Fig.4 Structure of MDCE network (3×8×8)

設計された本RWC-1のノード接続を光インタコネクションで置き換えることで、光インタコネクション技術の機能検証を実施した。

Fig.4には、RWC-1の3×8×8 (192ノード) 構成におけるネットワーク構成を示している。RWC-1はMDCE (Multi-dimensional Direct Cycles Ensemble extension)を採用し、低次元・小半径・小平均距離で高い通信パフォーマンスを実現している。本ネットワークは各ノード通信に高スループットを要求するため、特に大規模化に際しては光インタコネクションの採用が有望である。

3. 光インタコネクションモジュール¹⁾

本装置においては12チャンネルの光インタコネクションモジュールを使用してノード間接続を実現した (Fig.5 参

Fig.5 12-channel optical transmitter and receiver module

照)。各送受信モジュールは、モノリシック集積された12チャンネルのレーザ・PDアレイと駆動用ICを小型3.9ccのパッケージに一体集積した構造を有する。各チャンネルは200Mbit/sのデータレートで駆動でき、1モジュールで2.4Gbit/sの大容量スループットを実現できる。電気的なインタフェースは単一電源3.3V CMOSロジックで動作し、チャンネル間伝送スキューは±1ns以下を実現している。本モジュールはMUX/DEMUXを用いず、12チャンネルの電気信号を直接12チャンネル光信号に変換して送る。このため、低遅延性と高透過性が実現でき、計算機内の光インタコネクションデバイスとして最適な構成となっている。

4. 通信機能試験

われわれはまず光インタコネクションの計算機ノード間におけるデータ転送機能を実証するため、超並列計算機1ノード分の通信機能に特化したテストベッドを作成した (Fig.6, 7 参照)^{9)~11)}。本テストベッドはPEに相当する機能を外部接続ワークステーションで代替し、RWC-1と同等の通信機能を実現している (基板の大きさは51cm×46cm)。各2系統の送受信ポートはそれぞれ4つの光モジュールにて構成され、100MHz・24ビットのチャンネル構成を有する (スループットは300MByte/s)。各光モジュールからの並列光信号は12チャンネルのリボンファイバにて転送され、接続距離は20mを実現する。本装置においては信号の多重分離 (MUX/DEMUX) を行わず、DC直結の並列同期伝送方式を採用している。これにより信号処理の時間・回路のオーバーヘッドを最小に抑え、簡素な回路構成と低遅延・高データ透過性を有する構成となっている。本装置のSUの電気インタフェースは差動ECL、光モジュールはCMOSロジックを有するため、ECL-CMOSのレベル変換LSIを使用した。

Fig.8には本テストベッドを用いた機能試験内容を示している。信号は送信ポートo0もしくはo1より出力され、受信ポートi0もしくはi1にて受信される。本試験においてはパケットデータ形式の信号を用い、ノード間通信に必要な Add, Drop, Throughの3機能の動作をまず確認した。

Fig.6 Board layout of the testbed board

Fig.7 Photograph of the testbed board

Fig.8 Setup for the data transmission test

本システムにおける並列同期データ転送の障害要因は、帯域や受信感度よりも、信号間のスキューが大きなファクタを占めている。スキューのファクタは、(1)リボンファイバのスキュー、(2)レーザの発振遅延ばらつきによるスキュー、(3)受信系におけるレベル変動によるスキュー、(4)LSI内部回路・パッケージばらつき・プリント配線板内の配線ばらつきなど、さまざまな要因が考えられる。これらのスキューを補償するため、本システムはSUの受信端でスリップバッファ機構を用いた信号のリタイミングを行う。その機構は並列同期信号間のスキューを $\pm 1\text{ns}$ 以下に抑えることを要求するため、レーザアレイの高効率化・低ばらつき化を実現し、レーザ送信端にラッチによるリタイミング機構を採用するなどしてスキューの極小化を図った。その結果、SUの受信ポートにおける観測によると、並列同期信号間のスキューは $+0.06\text{ns}$ から -0.82ns の間に抑えられている (Fig.9 参照)。

Fig.10にSUの送受信ポートで観測したクロックストロープデータ (ビット0) の各信号波形を示す。受信端においても十分な信号振幅を有している。全信号はクロックの立ち上がりエッジに同期して転送される。OE/EO変換に伴う遅延時間は合わせて 15ns であり、非常に低遅延な動作を実現している。

次に本テストベッドを用いて、連続的にデータ転送を実行し、受信エラーを観測することにより通信信頼性の確認をした。試験のセットアップは同一装置のポートo1とi1を、24ビット10ワードの疑似ランダムパケットデータを周回転送する構成を採用した (PRBS= $2^{21}-1$ に相当する)。エラー判定はCRCを採用した。そして、80時間に渡るエラーフリー連続動作を達成し、高い信頼性を実証した (これはビットエラーレートに換算すると、 2.17×10^{-15} に相当する)。

Fig.11はビットエラーレートと平均障害発生時間 (MTBF) の間の相関を示している。各線は1システムを構成する信号経路が、それぞれ1から1000の場合を示している (1信号経路はそれぞれ 2.4Gbit/s のスループットを仮定)。1信号経路のシステムは電話等の長距離通信系を仮定し、1000本規模のシステムは超並列計算機等を考えている。Fig.11からもわかるように、超並列計算機等の大規模

Fig.9 Measured skew in the data signals (input port i0)

Fig.10 Waveform of the signals (left: transmitter side, right: receiver side)

Fig.11 Reliability of optical interconnection

Fig.12 Layout of the processor-board

Fig.13 Photograph of the processor-board

システムにおいて1年以上のエラーフリー動作 (MTBF=1 year) を保証する際、システム全体にはビットエラーレート 1×10^{-20} 以下という高信頼性が要求される。本図によると、今回実現した 2.17×10^{-15} 以下というビットエラーレートは、信号経路が10程度のシステムにおいてMTBFが約1日に相当する。これはPCクラスシステム等の小規模システムにおいては十分な性能に達成していると言える。

5. プロセッサボード

前述の通信機能テストベッドの結果を踏まえ、光インタコネクト技術を搭載したプロセッサボードを開発した (Fig.12, 13参照)。本ボードは430mm×540mmの単一ボード上に各4ノードの計算機機能を搭載し、信号入出力は送受各4ポートずつを有する (スループットは各300MByte/s、ボード全体のスループットは1.2GByte/s)。

(a) Previous type

(b) Newer designed type

Fig.14 Photograph of the SU

9.6Gbit/s)。1 ボードに12チャンネルの光インタコネクモジュールを送受各24個搭載し、光信号のボード外出力はすべてボード端の16連MPOファイバアダプタを介して行われる（本試作機は、16ノード：プロセッサボード4枚で構成される）。

本プロセッサボードにおいては、通信機能LSI（SU）の電気インタフェースを差動ECLからCMOSロジックに変更し、PE、SU、光モジュールのインタフェースをすべて3.3V CMOSロジックで統一した。これにより、前述の通信機能テストベッドで使用したECL-CMOSレベル変換LSI等の回路オーバーヘッドが不要となった。従来のSUは電気ケーブルによるノード間接続を前提に設計された。このため、電気ケーブルを用いて接続距離10m、クロックスピード100MHzの達成を求められ、線路の駆動能力の観点から差動ECLのロジックを採用した。これに対し、本プロセッサボードのSUは光インタコネクションの使用を前提に設計され、通信LSIの電気回路は同一プリント板内の短距離回路に限定される（それ以上の長い距離は光信号で送受信される）。このため、低消費電力かつロジックLSIとも好相性のシングルエンドのCMOSロジックの採用が可能となった。Fig.14に、従来のSUと今回設計改良したSUの外観写真を示す。ピン数は409ピンから304ピンに削減でき、パッケージ自体も57mm角のPGAパッケージから39mm角のQFPに大幅に小型化された。

このように光インタコネクションを採用することにより、高速・大容量な電気信号を信号処理LSIのピンの極近傍で光信号に変換可能となり、電気系のみで構成した装置間・装置内のデータスループットの拡大が可能となる。また、大負荷な長距離電気ケーブルを駆動する必要がなくなり、LSIのI/Oバッファの負荷や信号ピン数を大幅に削減でき、LSIのピンボトルネックやI/Oボトルネックの解消も可能となる。また、光ファイバの使用により接続距離も

24-ch Electrical Cable: 10 m

(a)

24-ch ribboned fiber cable: 20 m

(b)

Fig.15 Photograph of the 24-channel cables (copper cable: [10m: 10kg] and optical ribboned fiber: [20m: 1kg])

100m以上と大幅に拡大でき、信号ケーブル自体の小型軽量化も実現できる（Fig.15参照）。

6. おわりに

光インタコネクション技術を超並列計算機RWC-1のノード間接続に適用し、テストベッドを使用して機能検証を実施した（スループット600MByte/s、接続長20m、80時間連続エラーフリー）。そして、本試験結果を踏まえ16ノード構成の光インタコネクション接続並列プロセッサボードを開発した。本ボードは1ボード当たり1.2GByte/sの大容量スループットを有する。本システムにより、光インタコネクションを計算機システムに適用することによる通信系ボトルネック解消への効果を実証できるものとする。

（1998.4.6-受理 1998.4.30-再受理）

文 献

- 1) A.Takai, T.Kato, S.Yamashita, S.Hanatani, Y.Motegi, K.Ito, H.Abe, H.Kodera: "200-Mb/s/ch 100-m Optical Subsystem Interconnections Using 8-Channel 1.3- μ m Laser Diode Arrays and Single-Mode Fiber Arrays", J.of Lightwave Technology, Vol.12, pp.260-270, 1994.
- 2) J.W.Goodman, F.I.Leonberger, Sun-Yuan Athale, R.A.Kung:

- “Optical interconnects for VLSI system”, Proceedings of the IEEE, Vol.72, pp.159-174, July, 1984.
- 3) D.A.B.Miller, H.W.Ozaktas: “Limit to the Bit-rate Capacity of Electrical Interconnection from the Aspect Ratio of the System Architecture”, Journal of Parallel and Distributed Computing, Vol.41, pp.42-52, 1997.
 - 4) A.V.Krishnamoorthy, D.A.B.Miller: “Firehose architectures for free-space optically interconnected VLSI circuits”, Journal of Parallel and Distributed Computing, Vol.41, pp.109-114, 1997.
 - 5) T.K.Woodward, A.V.Krishnamoorthy, A.L.Lentine, L.M.F. Chirovsky: “Optical receivers for optoelectronics VLSI”, IEEE J.Special Topics in Quantum Electronics, Vol.2, pp.106-116, 1996.
 - 6) S.E.Greer: “An Extended Eutectic Solder Bump for FCBO”, Proceeding of ECTC '96, pp.546-551, 1996.
 - 7) D.L.Smith, A.S.Alimonda: “New Flip-Chip Technology for High-Density Packaging”, Proceedings of ECTC '96, pp.1069-1073, 1996.
 - 8) T.Yokota, H.Matsuoka, K.Okamoto, H.Hirono, A.Hori, S.Sakai: “Prototype Router for the Massively Parallel Computer RWC-1”, Proc. of ICCD '95, pp.279-284, 1995.
 - 9) Y.Yoshikawa, S.Araki, K.Miyoshi, N.Henmi, Y.Suemura, T.Nagahori, H.Matsuoka: “GByte-Class Skewless Parallel-Optical-Data-Link Subsystem”, Proc.of LEOS '96, WI6, pp.63-64, 1996.
 - 10) S.Nishimura, H.Inoue, S.Hanatani, H.Matsuoka, T.Yokota: “Optical Interconnections for the Massively Parallel Computer”, IEEE Photonics Technology Letters, Vol.9, pp.1029-1031, 1997.
 - 11) S.Nishimura, H.Inoue, S.Hanatani, H.Matsuoka, T.Yokota: “Error-free optical interconnection for the massively parallel computer”, IEEE Photonics Technology Letters, Vol.10, No.1, pp.147-149, 1997.

用語解説

MUX/DEMUX: Multiplex and demultiplex

並列同期信号を時間軸上で多重・分離する信号操作。シ

リアル・パラレル変換もこの一種である。

Add, Drop, Throughの3機能: 交換機能の基本となる機能。

各ノードからネットワーク上に信号を送出する機能 (Add機能), ネットワークに流れる信号から特定信号をノードに取り込む機能 (Drop機能), ネットワーク上の信号をノードに取り込み・再送する機能 (Through機能) に分類される。

スリッパバッファ機構: データのI/O時に同期のリタイミングを行う機構の一種。受信信号を書き込み3つ程度の高速度レジスタバッファにクロックに同期させて順次書き込み, 読み出しクロックに同期させて再度レジスタより順次読み出すことで, 書き込み・読み出しクロック間の位相差を補正する機構。

PGAパッケージ: Pin Grid Array Package

LSIのピンパッケージの一種。本記事のモジュールにおいてはピン間隔2.54mmの格子状のピン配列を有し, 基板貫通型の実装構造となる。

QFP: Quad Flat L-leaded Package

面実装型のLSIパッケージの一種。100ピン以上の多ピン構造を採用する場合, 本パッケージはBGA(Ball Grid Array package)が採用される場合が多い。

LSIのピンボトルネック: LSIのI/Oピン数が設計上必要な数だけ確保できず, 設計上の制約条件となる場合を指す。例えば, システム設計上必要なI/O信号が100あっても, LSIのI/Oピンが80程度しか確保できない場合をLSIのピンが設計上ボトルネックであるという。

I/Oボトルネック: LSIやシステムのI/O部の性能が, システム設計上の要求仕様を満たさない場合を指す。信号帯域, LSIのピン数, 接続距離, ケーブル体積, コネクタの実装等の様々な要因が考えられる。

西村 信治 (にしむら しんじ)

昭和41年生まれ。平成3年, 東京大学大学院工学系研究科電子工学専攻修士課程修了。現在, 株式会社日立製作所中央研究所において, 光インタコネクタサブシステムの開発に従事。