筑波大学技術報告
26: 42-47 , 2006

Maestro2 スイッチボックスの開発

小野雅晃[1]
筑波大学システム情報工学等支援室（装置開発班）
〒305-8573 茨城県つくば市天王台 1-1-1

概要

クラスタ型コンピュータ向けネットワーク Maestro2 用のスイッチボックス(SB)を開発したので報告する。SB は８個のポートを備え、各ポートがパーソナルコンピュータ(PC)に挿されたネットワーク・インターフェース(NI)の通信ポートに接続される。SB はイーサネットのスイッチと同様に、ポート間をスイッチする。SB がスイッチするのはイーサネットなどの汎用プロトコルではなく、Maestro2 独自の通信プロトコルである。ポートの物理インターフェースは LVDS(Low Voltage Differential Signalling)を使用し、物理的な最大スループットは片方向 3.2Gbps(Giga Bit Per Second)である。

１．はじめに

システム情報工学研究科コンピュータサイエンス専攻の和田研究室では、クラスタ型のコンピュータに使用するために、高速、高機能のアドイン・カード及びスイッチから構成される通信システムを開発している。

6 年ほど前には、200Mbps の IEEE1394(i-LINK)を使用した第一世代の通信システム Maestro1 を開発した。

次に、Maestro2 として第 2 世代の通信システムを開発した。使用した通信インターフェースは 3.2Gbps の LVDS である。通信システムは NI と SB で構成される。昨年 NI について発表したので、今回は SB について発表する。

２．MAESTRO2 クラスタネットワークの構成

Maestro2 のネットワークを Maestro2 クラスタネットワークと呼ぶことにする。Maestro2 クラスタネットワークはイーサネットの様に PC を相互接続するネットワークである。構成要素は PC の PCI スロットに挿入される NI と複数の NI に接続され、通信パケットをルーティングする SB である。

SB には８個のポートがあり、それぞれのポートが NI のポートに接続される。それらのポートは２本のケーブルで接続される。ケーブルの物理インターフェースは 700MHz 動作の８本のデータラインと１本のクロックラインを持つ LVDS 信号である。

Maestro2 ネットワークの特徴としては、次の３点が挙げられる。

1. NI、SB 共に高性能なプロセッサを搭載していて、PC と処理を分担することが出来る。
2. 制御回路用 IC チップとして FPGA を搭載しているので、仕様変更を行うことが出来る。
3. イーサネットに比べて、プロトコルが軽量で、しかも実効スループットが大きい。

図 1 に Maestro2 ネットワークの構成図を示す。

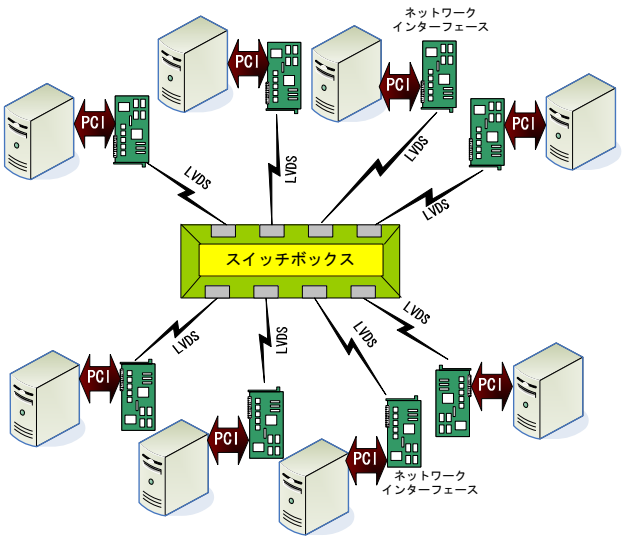

図 1　Maestro2 クラスタネットワークの構成図

３．SB の構成

SB は現在 265mm×240mm の 8 層基板で作られている。SB は LVDS 送信チップ、LVDS 受信チップ、PowerPC603e プロセッサ、32MByte の SDRAM (Synchronous Dynamic Random Access Memory)、Xilinx 社の FPGA(Field Programmable Gate Array)チップ、XPort で構成される。

図 2 に SB の写真を示す。真ん中の 2 つが FPGA と PowerPC プロセッサである。SDRAM は基板の後ろに搭載されているので見えない。LVDS 送信、受信チップは基板の端のコネクタのそばに搭載されている QFP(Quad Flat Package)チップで、LVDS 送信チップが 8 個、LVDS 受信チップが 8 個、合計 16 個搭載されている。基板の左上端の青い LAN ケーブルが接続されているコネクタが XPort である。

LVDS 受信チップはナショナルセミコンダクタ社の DS90CR481、LVDS 送信チップは同社の DS90CR484 を使用している。LVDS 送信チップは FPGA から動作周波数 100MHz、32 ビット幅のデータ信号を受け取り、動作周波数 700MHz、データ 8 ビ

[1] E-mail: ono@sie.tsukuba.ac.jp; Tel: 029-853-5195

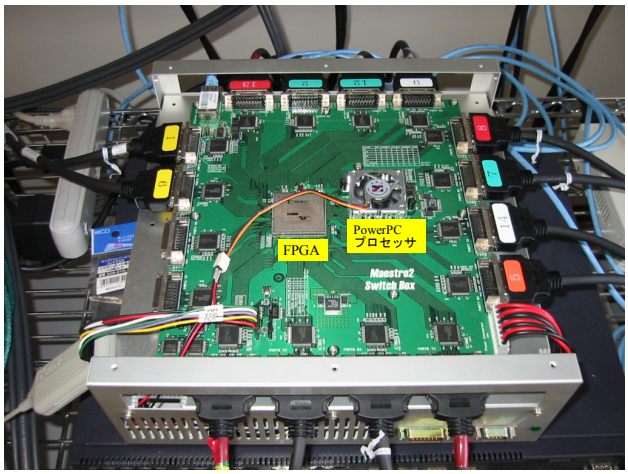

図 2　SB の写真

ット、クロック信号 1 ビットの LVDS 信号に変換する。その後、LVDS 信号はコネクタからケーブルを通り、相手方の NI に届く。LVDS 受信チップは、LVDS 送信チップの逆の変換を行う。

PowerPC プロセッサ(MPC603RRX300)は内部動作周波数が 300MHz、バスの動作周波数が 66MHz の PowerPC プロセッサを使用している。この PowerPC プロセッサはデータバスを 64 ビット持っているが、起動時に 32 ビット分だけ使用するモードに設定することが出来る。SB ではこのモードを使用して、PowerPC プロセッサのバス幅を 32 ビットとしている。

SDRAM の容量は 32MByte、動作周波数は 66MHz である。SDRAM チップはエルピーダ社の uPD45123163G5-A74 を使用している。この SDRAM チップは 128Mbit の容量を持ち、データバス幅は 16 ビットである。SB のデータバス幅は 32 ビットなのでこのチップを 2 個使用している。SDRAM の動作モードは PowerPC プロセッサがキャッシュ OFF 状態での SDRAM アクセスの無駄を省くために、シングル転送に設定されている。バースト転送が必要な場合には、シングル転送を 1 クロックごとに連続的に発生させる。命令ロードやキャッシュ ON 領域へのデータアクセスなどは 8 バーストのデータ転送が行われる。

FPGA は Xilinx 社の Virtex2 シリーズの内、XC2V3000-4BF957C を使用している。この FPGA は標準ゲート換算で 300 万ゲート相当、957 ピンの BGA(Ball Grid Array)パッケージである。FPGA には、SDRAM 制御回路、PowerPC のインターフェース回路を含め、すべての制御回路やスイッチ回路などが内蔵されている。

XPort は LANTRONIX 社の製品で、シリアル通信ポートをイーサネットに変換するモジュールである。FPGA のシリアル通信ポートをイーサネットに変換する。

４．FPGA 内部回路構成

FPGA 内部の回路は、PowerPC 制御回路、SDRAM 制御回路、内部 RAM、シリアルインターフェース、内部レジスタ、MLX、メッセージアナライザ、スイッチコントローラである。

メッセージアナライザは山際伸一氏[2]が作製し、MLXは青木圭一氏[3]が作製した。

FPGA 内部回路のブロック図を図３に示す。

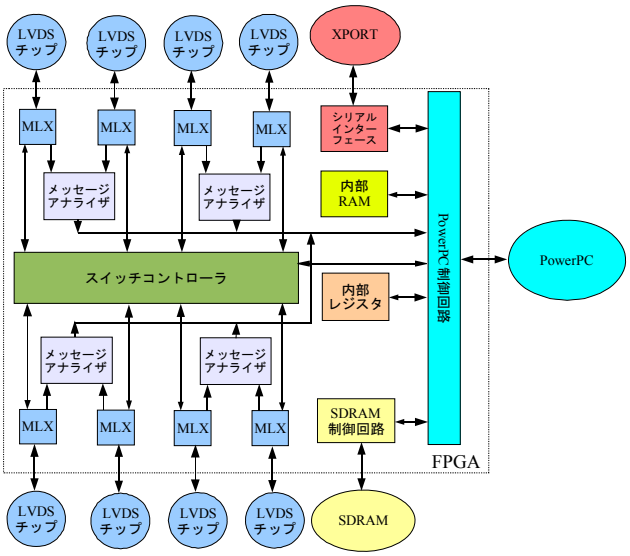

図 3　FPGA 内部回路のブロック図

これから、各回路について説明する。

4.1　PowerPC 制御回路

PowerPC 制御回路は、PowerPC プロセッサ(MPC603e,300MHz,バス動作周波数 66MHz)へのインターフェース回路である。

PowerPC のメモリおよびメモリ領域にマップされた I/O へのアクセス手順はアドレス転送とデータ転送に分けられる。アドレス転送は以前の転送が処理できなくても、次に 1 回は前倒しに発行できる。

PowerPC 制御回路はアドレス転送を保存し、/DBG（Data Bus Grant）信号をアサートして、データ転送を開始する。その後、PowerPC 制御回路は、データの書き込みや読み出しを終了する時点で/TA(Transfer Acknowledge)をアサートして、PowerPC プロセッサにデータ転送の完了を知らせる。

PowerPC 制御回路は FPGA 内の各回路をアドレスマップし、PowerPC からアクセスできるようにする。アドレスマップを表 1 に示す。

表 1　SB のアドレスマップ

デバイス	アドレス (上位 6 ビット)
内部 RAM	111111 (0xFC~)
SDRAM	000000 (0x00)
スイッチコントローラ	111010 (0xE8)
メッセージアナライザ	111000 (0xE0)
シリアルインターフェース	111001 (0xE4)
内部レジスタ	111011 (0xEC)

[2] Instituto de Engenharia de Sistemas e Computadores (INESC-ID), Portugal

[3]筑波大学　システム情報工学研究科

4.2　SDRAM 制御回路

SDRAM 制御回路は 32MByte の SDRAM を制御する回路である。PowerPC プロセッサと SDRAM のバスは同一であるため、SDRAM 制御回路はデータの入出力回路を持たない。SDRAM 制御回路は SDRAM の制御信号のみ駆動する。SDRAM の制御信号は /RAS(Row Address Strobe)、/CAS(Column Address Strobe)、/WE(Write Enable)、DQM(DQ Mask Enable)、/CS(Chip Select)、Address(0~11)である。各制御信号を適切に駆動することにより SDRAM へコマンドを与える。ただし、/CS は常時 LOW にアサートされている。

電源 ON 時には初期化手順を実行する。初期化手順はまず全バンクプリチャージを実行する。その後、モードレジスタを設定し、リフレシュを 2 回実行する。モードレジスタは SDRAM のモードを記憶しておくレジスタで、SB では CAS レイテンシ２、バースト長１、シーケンシャルモードに設定している。

SDRAM は定期的にリフレッシュをしないとデータが消えてしまう。SB では 15.6usec ごとに/RAS と /CAS を LOW レベルにアサートして、リフレッシュを実行している。

PowerPC プロセッサの SDRAM アクセス手順は、最初に行アドレスを与えて/RAS を LOW にアサートし、次に列アドレスを与えて/CAS をアサートする。書き込みの場合は同時に/WE をアサートする。キャッシュ OFF の領域にアクセスする場合には 1 回で終わるが、キャッシュ ON の領域にアクセスする場合は、8 バースト転送となるのでアドレスを変更しながら 8 回繰り返す。最後に/RAS と/WE を LOW にアサートしてプリチャージを行う。

4.3　内部 RAM

内部 RAM は FPGA に内蔵された BlockRAM を使用している。容量は 32Kbyte、データ幅は 32 ビットである。バーストアクセスに対応し、PowerPC プロセッサの最大バースト長 8 ワードに対応している。

内部 RAM には、最初に起動するブート用ソフトウェアを入れておく。現在 SB ではシリアルインターフェースからソフトウェアをロードするダウンローダーを BlockRAM に書き込んである。

4.4　シリアルインターフェース

シリアルインターフェースはパーソナルコンピュータと SB を結ぶインターフェースである。シリアルインターフェースのブロック図を図 4 に示す。

TXD、RXD などのシリアルの信号線は XPort に接続される。XPort によってシリアル信号がイーサネットに変換され、PC から IP アドレスを指定して通信すれば、PC は SB のシリアルインターフェースと通信出来る。

シリアルデータの送信手順は、SB の PowerPC プロセッサがアドレス 0xE4000000 に 8 ビットのデータを書き込むと、そのデータは送信 FIFO へ送られる。送信 FIFO から出力されたデータは、パラレルーシリアル変換され TXD へ直列に出力される。

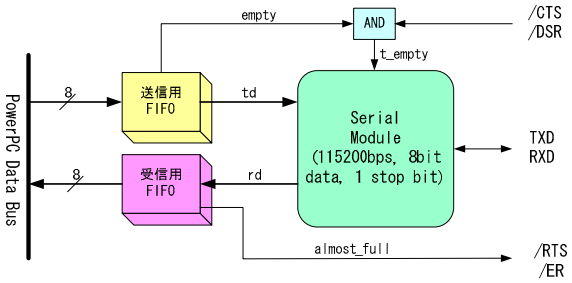

図 4　シリアルインターフェース　ブロック図

シリアルデータの受信手順は、RXD からシリアルデータが来ると、シリアルーパラレル変換されて受信 FIFO にパラレルデータが入力される。これにより、ステータスレジスタのエンプティフラグが 0 になる。同時に PowerPC プロセッサはステータスレジスタをポーリングまたは、割り込みによって監視してエンプティフラグが 0 になるのを待つ。PowerPC プロセッサはエンプティフラグが 0 になっていることを確認すると、受信準備が整ったと判断し、0xE4000000 からデータを読み込む。

このシリアルインターフェースの仕様は、調歩同期、転送速度 115200bps、データ 8 ビット、1 ストップビットである。

現在シリアルインターフェースは SB に PowerPC プロセッサのプログラムのダウンロード用やモニタソフトウェアとの通信用に使用されている。

4.5　内部レジスタ

内部レジスタには、いろいろな設定レジスタ、ステータスレジスタなどが実装されている。例えば、MLX のリセットや各ポートのバッファの使用状況などである。ここに各ポートから PowerPC プロセッサにデータを渡すためのプロセッサ FIFO のステータスやデータ読み出しポートがある。

4.6　MLX

MLX はケーブルが接続されている先の NI にデータを送り出すためにカプセル化する。データは 32 バイト単位の 1 つ又は複数のパケットにまとめられ、ヘッダを付けられて送り出される。

ヘッダに書かれている送信パケット数を全部送り終わっても、後からバッファに溜まったパケットがあればコンティニューコマンドを送り、データ転送を切れ目なく継続することが出来る。

昨年発表した NI の技術報告書ではエラー訂正もハードウェアで実装されていたが、今回からエラー検出のみハードウェアで実装し、エラー訂正は上のレイヤーに任せることになった。

MLX は OSI 参照モデルで言うとデータリンク層に相当する。MLX の入力は 64 ビット 66MHz で、出力は 32 ビット 100MHz である。MLX の最大ビットレートは 3.2Gbps である。

4.7　メッセージアナライザ

メッセージアナライザは MLX から通信データを取得し、メッセージのヘッダだけを抽出し、FIFO に

格納する。メッセージアナライザは 2 つの MLX に 1 つずつ、全部で 4 つ搭載されている。PowerPC プロセッサは 4 つのメッセージアナライザをポーリングし、ヘッダがないかどうかを探している。ヘッダがあると PowerPC プロセッサはヘッダを解析して、その通信データを適切に処理する。

4.8 スイッチコントローラ

スイッチコントローラは各ポートの MLX 出力から入ってきた通信パケットを各ポートの MLX 入力にスイッチする。

スイッチコントローラはコマンド発行ユニット、リソーステーブル、スイッチユニットの 3 つのユニットで構成される。スイッチコントローラは、現在 8 つのスイッチユニットを持っているが、パラメータを変更するだけでスイッチユニットの個数を変更することが出来る。

PowerPC プロセッサはメッセージアナライザからのヘッダ情報を元にスイッチコマンドを作成し、コマンド FIFO に書き込む。コマンド FIFO に書き込まれたスイッチコマンドはコマンド発行ユニットに送られる。コマンド発行ユニットはリソーステーブルやコマンド発行ユニットにすでにエントリされているスイッチコマンドとの優先順位を判定し、発行可能なスイッチコマンドをスイッチコマンドバスに発行する。スイッチユニットはスイッチコマンドバスを常に監視して、自分に対するスイッチコマンドならばコマンドで指定された処理を行う。スイッチコントローラのスイッチコマンドバス構成を図 5 に示す。

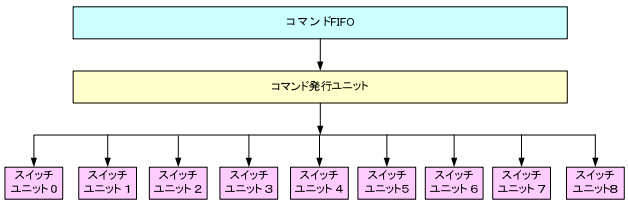

図 5 スイッチコマンドバス構成図

スイッチコマンドは 64 ビット幅のフィールドを持っている。構成要素はスイッチコマンドの機能、書き込みポートベクタ、読み出しポート番号、転送パケット数である。スイッチコマンドの機能にはパケット転送、パケット消去、パケット生成の 3 つの機能がある。パケット転送は MLX に入ってきたデータを他の MLX に転送する機能である。パケット消去は MLX に入ってきたパケットをどこにも転送せずに消去する。パケット生成は PowerPC プロセッサがパケットを生成し、コマンド FIFO に書き込むことで MLX に出力する機能である。書き込みポートベクタは書き込むスイッチユニットに割り当てられているビットを持っている。そのビットを 1 にすると指定されたポートに書き込むことが出来る。ビットベクタなので、すべてのビットを 1 にするとブロードキャストが出来る。読み出しポート番号はパケットを読み出すポートの番号を指定する。

4.8.1 コマンド発行ユニット

コマンド発行ユニットの構成を図 6 に示す。

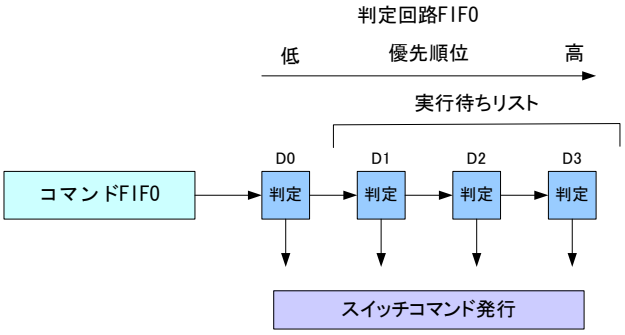

図 6 コマンド発行ユニットの構成図

図 6 に示すように、コマンド発行ユニットはスイッチコマンドの発行を判定する判定回路 FIFO で構成される。判定回路 FIFO の個数は VHDL の constant 文で指定されるパラメータによって変更できる。コマンド発行ユニットは容易にコマンド発行の効率を変更できるように設計されている。

コマンド FIFO から入力されたスイッチコマンドは D0 の判定回路に入力され、リソーステーブルの使用中の読み出しポート、書き込みポートの情報と比較される。D0 のスイッチコマンドと使用リソースが競合しない場合に、スイッチコマンドが各スイッチコントローラに発行される。スイッチコマンドが発行されれば D0 のエントリは削除される。

リソースが競合した場合には、スイッチコマンドは発行されずに実行待ちリスト D1 にシフトされる。これで D1 が有効となる。有効となった D1 はリソーステーブルの使用中リソースと比較される。スイッチコマンドが発行できる場合にはスイッチコマンドが発行され、発行されたエントリは削除される。

スイッチコマンドが発行されないで D2 のエントリが空の場合は、次のクロックで D1 のスイッチコマンドは D2 にシフトされる。このように、前のエントリが空の場合は順々に前のエントリにシフトされる。

実行待ちリストに入力されているスイッチコマンドは各クロックで発行可能かどうかチェックされ、可能であれば優先順位の順に発行される。各スイッチコマンドの優先順位は、図 6 のように D0 が最低で D3 が最高となる。

複数のエントリにスイッチコマンドが入っている場合は、リソーステーブルの使用中リソース及び優先順位の高いエントリのスイッチコマンドとの依存関係をチェックし、発行可能なスイッチコマンドを発行する。以前に発行されたスイッチコマンドを後で発行されたスイッチコマンドが追い越すことが出来るアウト・オブ・オーダー発行をサポートしている。

スイッチコマンドがパケット生成の場合は D0 にとどめ、実行待ちリストには入力しない。スイッチコマンドが発行できない場合は D0 で発行できるようになるまで待つ。これはスイッチコマンドの次からのデータが生成データであるため、実行待ちリストに入れないための処置である。つまり、スイッチコマンドがパケット生成の場合はここでブロックされる。

4.8.2　リソーステーブル

リソーステーブルは現在使用中の読み出しポート、書き込みポートを記憶する。読み出しポートは 8 ビット幅の使用中ビットを用意し、ポート番号のオフセットのビットが立っていたら使用中を示す。書き込み中スイッチユニットも同様に 9 ビット幅の使用中ビットを用意し、スイッチユニット番号のオフセットのビットが立っていたら該当するスイッチユニットは使用中を示す。

4.8.3　スイッチユニット

スイッチユニットは通信パケットのスイッチを受け持つユニットである。スイッチユニットはパラメータによって 9 個、5 個、4 個、2 個に変更できる。9 個の場合はクロスバスイッチと同等になる。8 個の MLX の接続されたスイッチユニットと PowerPC プロセッサにスイッチするためのスイッチユニットが 1 個の構成となる。PowerPC プロセッサにスイッチするためのスイッチユニットは PowerPC プロセッサが読めるようにプロセッサ FIFO にスイッチする。9 個のスイッチユニットの場合のスイッチユニット構成図を図 7 に示す。

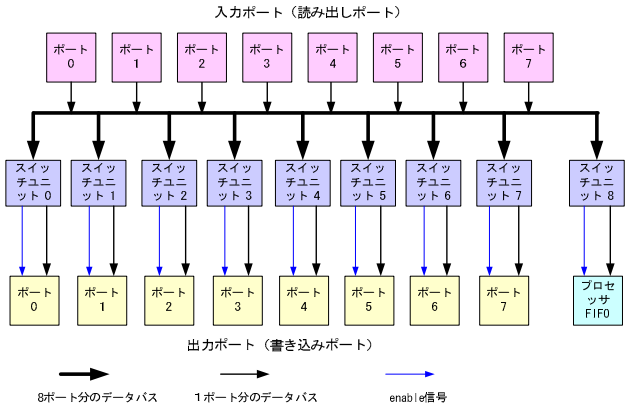

図 7　スイッチユニット構成図（9 個のスイッチユニットの場合）

図 7 の入力ポート、出力ポートは MLX の入出力ポートに接続されている。

出力ポート 1 つごとにスイッチユニット 1 つが割り当てられているので、スイッチユニットが担当する出力ポートは 1 つとなる。

各スイッチユニットは、コマンド発行ユニットが発行したスイッチコマンドを常時監視し、自分へのスイッチコマンドを認識すると動作を始める。

スイッチコマンドで指定された入力ポートに接続された MLX にパケットが到着すると、MLX の受信パケット容量が 0 ではなくなる。スイッチユニットはそれを受け、さらに送信先の MLX の送信パケット容量が 0 でないことを確かめた後に入力ポートのデータを出力ポートに送る。スイッチコマンドで複数のスイッチユニットを起動した場合には、起動したスイッチユニットすべてについて、出力する MLX の送信パケット容量が 0 でないことを調べる。

図 8 にスイッチユニットが 5 個の場合のスイッチユニット構成図を示す。出力ポート 2 個ごとに 1 つ

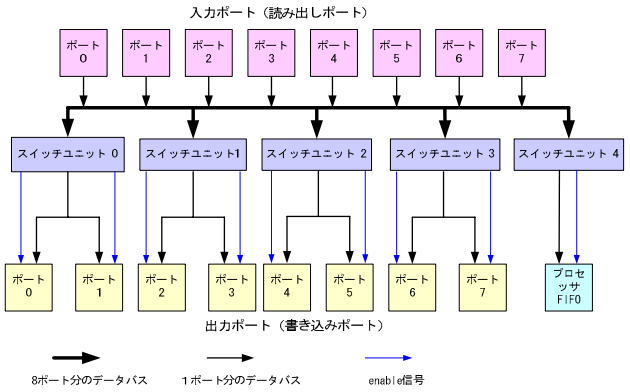

図 8　スイッチユニット構成図（5 個のスイッチユニットの場合）

のスイッチユニットが割り当てられる。例えば、ポート 0 またはポート 1 へパケットを出力するスイッチコマンドの場合は、スイッチユニット 0 が動作する。ポート 0 にパケットを出力するためにスイッチユニット 0 が動作中の場合には、コマンド発行ユニットはポート 1 にパケットを出力するスイッチコマンドを発行できない。つまり、コマンド発行ユニットもパラメータによって構造を変える。リソーステーブルも同様の構成である。

パラメータによって構造を変化することの利点としては、FPGA の使用リソースを簡単に変更できることが挙げられる。当初、スイッチコントローラを変更する場合、どのくらい FPGA のリソースを消費するか予想が立たなかったので、パラメータで簡単に使用リソースを変更できるようにした。また、パラメータを変更すればスイッチユニットの数による性能の差が簡単に測定できる。

5．基板実装

SB には少なくとも 100MHz で 256 本の同時スイッチング出力がある。このような高速信号を多数使用しているプリント基板では、グランドバウンズやクロストークに注意しなければならない。

1 枚目の SB のプリント基板を基板設計業者に依頼して作成した時点では、グランドバウンズやクロストークを軽視し、伝送線路や電源インピーダンスに対する配慮が足りなかった。

その結果として、SB では FPGA で受け取った信号が誤るようになった。特に、FPGA と LVDS チップの距離が遠いポートが誤る率が高かった。この状況を改善しようと FPGA から LVDS の伝送クロックを 100MHz から 66MHz へ落としてみた。そうするとデータ誤りは改善するが、距離が遠いポートがデータ誤りを起こしてしまう。距離が近くデータ誤りを起こさないと思えるポートでも、0xFF7FFFFF と 0x00800000 のような 1 ビットのみ他のビットと異なるデータとそれを反転したデータの繰り返しではデータ誤りが生じた。

以上の結果からプリント基板の電源インピーダンス及び伝送線路特性が悪いと判断し、伝送線路シミュレーションをしてくれるスキルの高い基板設計業者に基板の再作製を依頼した。その結果、基板は当初 6 層基板だったがインピーダンスマッチングやク

ロストークを考慮し 8 層基板になった。FPGA 直下に 0603 と呼ばれる 0.6mm×0.3mm の大きさの極小コンデンサを取り付け、電源インピーダンスを改善した。さらに、伝送線路シミュレーションを行い、最適な波形になるように配線パターンを決定した。以上の対策によって再作製した基板は 100MHz で正常動作するようになった。

最初から動作する基板を作製するには、伝送線路シミュレーションや、グランドバウンズ対策のための電源インピーダンスの最適化などの対策を十分に検討することが不可欠である。

６．まとめ

本報告ではクラスタ型のコンピュータに使用するスイッチボックス(SB)の開発について述べた。

SB は NI からのパケットを他の NI にスイッチする機器である。SB は PowerPC プロセッサを搭載し、インテリジェントな処理をすることが出来る。また、ヘッダ処理をサポートするハードウェアを持ち、スイッチユニットの数を容易に変更できるように作られている。SB の 1 ポートあたりの理論的な最大スループットは 3.2Gbps である。

現在の Maestro2 クラスタネットワークは、ハードウェアのデバックがほぼ終わり、姫野ベンチマーク、Gauss-Jordan ベンチマークなどのアプリケーションを用いて評価を進めている。

謝辞

Maestro2 システムの開発補助をさせていただいたシステム情報工学研究科コンピュータサイエンス専攻の和田耕一教授に深く感謝いたします。また、共同制作者である INESC-ID の山際伸一氏とシステム情報工学研究科の青木圭一氏に深く感謝いたします。

参考文献

[1] 小野雅晃, Maestro2 ネットワークインターフェースの開発,第 4 回筑波大学技術職員技術発表会ポスターセッション(2005)

[2] 小野雅晃,Xilinx 社製 FPGA を搭載した PCI ボードのシミュレーション,平成 15 年度高エネルギー加速器研究機構技術研究会報告集ポスターセッション(2004)

[3] Shinichi Yamagiwa, Keiichi Aoki, Masaaki Ono, Tetsuya Sakurai, Koichi Wada, and Luis Miguel Campos. Maestro2: A new challenge for high performance cluster network. In The 6th World Multiconference on Systemics, Cybernetics and Informatics, volume XI, Computer Science II, pp.382-387, 2002

[4] Keiichi Aoki, Shinichi Yamagiwa, Masaaki Ono, Koichi Wada, Luis Miguel Campos. An architecture of high performance cluster network : Maestro2. In 2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing(PacRim03), 2003.

[5] Shinichi Yamagiwa, Kevin Ferreira, Luis Miguel Campos, Keiichi Aoki, Masaaki Ono, Koichi Wada, Munehiro Fukuda, Loenel Sousa. On the Performance of Maestro2 High Performance Network Equipment, Using New Improvement Techniques. In 23rd IEEE International Performance Computing and Communications Conference(IPCCC 2004), 2004.

[6] Keiichi Aoki, Shinichi Yamagiwa, Kevin Ferreira, Luis Miguel Campos, Masaaki Ono, Koichi Wada, Leonel Sousa. Maestro2: High Speed Network Technology for High Performance Computing. In 2004 IEEE International Conference on Communications (ICC 2004), 2004.