並列計算機 Ships1 のノード間結合装置の構築

†加藤 渉, †松原 裕人, †三浦 康之, †大谷 真, †渡辺 重佳, †高野 誠一
湘南工科大学

■ 1. はじめに

近年、PC の高性能・低価格化が進んでいる。また、計算機によって処理すべき分野が拡大、大規模な計算を必要とするようになり処理速度の飛躍的な向上が必要なる。そこで、複数の計算機を統合化し高い処理速度・信頼性を得ることを目的とした並列計算機が実用化されている[1]- [3]。しかし、グリッドなどの疎密結合並列計算機や密結合計算機はコストが高くなる傾向にある。

湘南工科大学情報工学科では、Ships1(SHonan Institute of technology Parallel System 1)という安価な CPU を複数用いたコストパフォーマンスの高い並列計算機の研究・開発している。Ships1 は、並列計算機を構築するために高いスループットを持っている Gigabit LAN を使用している。しかし、オーバーヘッドが大きい他、バッファのコピーによる遅延などにより細かいデータ転送を必要とする並列計算機には不向きな場合がある。そこで、専用の通信機構として LVDS を用いて小規模並列計算機向けにこの問題を解決する。

Ships1 では、Gigabit LAN と LVDS の両方を使用し、それぞれの特性を生かしたデータ転送を目指す。

本稿では Gigabit LAN に代わりノード間の通信を行うための装置として、LVDS コネクタ付 FPGA ボードの開発状況を報告する。

■ 2. クラスタ型並列計算機

2.1 並列計算機 Ships1 の構成

Ships1[3][4]は、16 台の PC とネットワークインターフェイス(Gigabit LAN)、Gigabit スイッチで構成される。本研究では市販されている LVDS 付き FPGA ボードを使用する。LVDS を直接網でリング型ネットワークに構成することにより、専用のスイッチは使用せずに小規模限定であるものの、低コストかつ高速な通信網の構築が可能になる。

■ 3. FPGA 回路の設計

3.1 FPGA ボードの仕様

図 1 に本研究で使用する FPGA ボードの内部構成を示す。本研究で使用する FPGA ボードは、PCI 制御用の Spartan2 と LVDS 制御用の Virtex2 の 2 つの FPGA が搭載されている。

また、ボードコネクタが 2 つあり、他の回路と接続することができる。LVDS 通信は 1 対 1 の通信のため、1 台で 2 つのノードを結ぶ通信を行う場合にはボードを 2 枚用意する必要がある。同 PC での 2 枚ボード間通信はボードコネクタを経由して実現することが可能である。

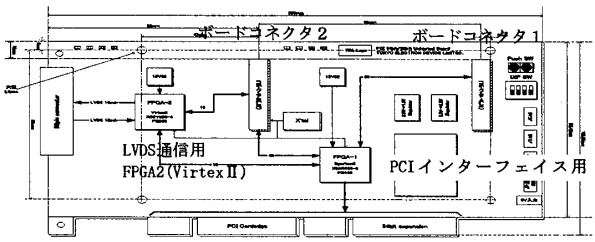

図 1 FPGA ボードの内部構成

3.2 FPGA ボードの開発状況

PCI を制御する Spartan2 には、Xilinx 社の PCI LogicCore および東京エレクトロンデバイス社の基本回路を用いていることにより、DMA 制御方式を用いた PCI バスから Dual Port RAM へのデータ転送を可能にした。現在 LVDS 間通信を行うため、Dual Port RAM に送られる 32bit のデータとクロックを Virtex2 へ転送する。図 2 に、Spartan2 の構成を示す。

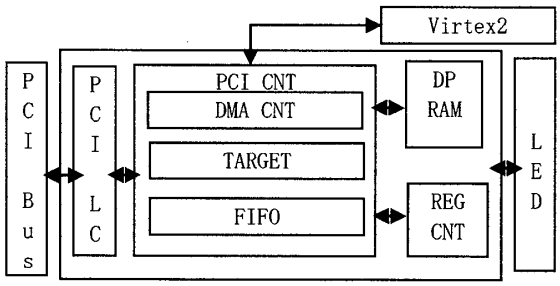

図 2 Spartan2 の回路

LVDS を制御する Virtex2 では、今回使用する LVDS ケーブルが 8bit の双方向全二重の構成のため、データ転送を行うために Spartan2 から転送されてきた 32bit のデータを 8bit へ変換する必要がある。DCM 回路を用いてクロックを早めることにより、送信時には serdes_8b_7to1 を使用し 32bit から 8bit×4、受信時には obufds_lvds_33 を使用し 8bit×4 から 32bit へ変換する。FPGA 内部では、32bit のデートとして扱うことができるためトラフィックの増大に対応する。データの送受信は Virtex2 で使用可能な lvds_33 モジュールを使用

Development of Inter-Node Connection of Parallel Computer - Ships1
† Wataru Kato, Hiroto Matubara, Yasuyuki Miura, Makoto Oya, Shigeyoshi Watanabe, Seiichi Takano
Shonan Institute of Technology

する。データが正しく転送されているかを確認するために、Virtex2に接続されたLEDにより確認を行う。図3にVirtex2の構成を示す。

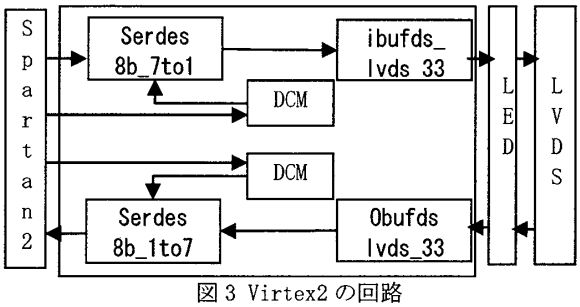

図3 Virtex2の回路

■ 4. デバイスドライバ

Ships1では通信ノード間のアプリケーションバッファでのZero-Copy転送を目指す。そのため、raw I/Oでノード間のメモリの直接転送を行う。アプリケーション-デバイスドライバ間ではユーザ空間のメモリをカーネル空間にマッピングする。デバイスドライバ- 半密結合装置間はDMA転送を行う。

Ships1では、CPUの使用率の低減と通信速度の向上するような処理方式を目指す。

送信時の処理の流れを下記に示す。

1) アプリケーションがシステムコールを出す。
2) 呼び出されたデバイスドライバは送信用キューを調べる。
3) 物理メモリアドレスを取得しDMAメモリとしてマッピングする。
4) DMA転送の情報を送信用キューのI/Oポートアドレスに追加する。
5) ノード間結合装置は、送信用キューのI/Oポートアドレスに送られてきた情報を別の場所にあるキューに保存する。キューに登録されたDMA転送が終わるとIRQを出す。

受信時の処理の流れを下記に示す。

1) アプリケーションがシステムコールを出す。
2) デバイスドライバは受信用キューを調べる。
3) 物理メモリアドレスを取得しDMAメモリとしてマッピングし、受信用キューにDMA転送の情報を登録する。
4) 登録後に排他的待ち列に自分のプロセスを追加しスリープする。
5) ノード間結合装置は、データを受信し始めると受信用キューからDMA転送の情報を取り出し、DMA転送を行う。
6) DMA転送が終了したらIRQを出しデバイスドライバを起こす。
7) デバイスドライバは受信したデータサイズをアプリケーションにリターンする。

■ 5 デバイステスト

デバイスが正しく動作しているか確認を行うため、プログラムを作成しテストした。テスト内容は、DMA転送を使用せず、システムとI/Oポートを経由したFPGAボード間の速度計測となる。結果を表1に示す。図4にGigabit Ethernetのラウンドトリップタイムを参考に示す。

データサイズ(byte)	時間(ms)
8	0.0000360012
100	0.0000619888
200	0.0000910759
400	0.0001499653
600	0.0002069473
800	0.0002639294
1000	0.0003221035

表1 システムとFPGAボード間との時間測定

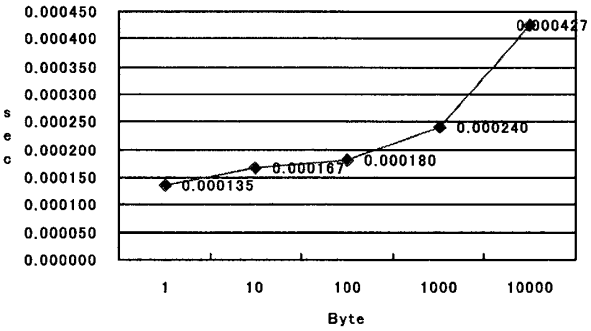

図4 Gigabit Ethernetのラウンドトリップタイム

データの確認にはVirtex2のLEDを使用したFPGAボードでの確認および、Linuxのlogを参照し確認を行った。データサイズが小さい場合、DMAコントローラを通すDMA転送より、I/Oポートを通した転送を行ったほうが早いと予測できる。DMA転送と比較しデータサイズによってDMA転送とI/Oポートを通した転送とを切り替える必要がある。

■ 今後の予定

今回作成したドライバで、DMA転送を行うことができなかったため作成する。また、現在LVDSを制御するVirtex2は、送信用回路と受信用回路とが別々に作成したため、それを統合し送受信用回路に作り変える必要がある。これらを実装したのち、LVDSを使用したラウンドトリップタイムを測定、Gigabit Ethernetを使用したTCPソケット通信との比較を行う。

■ 参考文献

[1] TOP500 Supercomputer Site
http://www.top500.org/

[2] N. J. Boden, D. Cohen, R. E. Felderman, A. E.Kulawik, C. L. Seitz, J. N. Seizovic and WenKing Su. "Myrinet { A Gigabit-per-SecondLocal-Area Network". IEEE MICRO, Vol. 15,No. 1, pp. 29{36, February 1995

[3] 松尾成志, 岡本恵介, 大谷 真, 中小規模並列コンピュータShips1の開発

[4] 松原 裕人, 和田 卓, 大谷 真, Ships1におけるノード間接続装置の研究