単方向トーラスネットワークを用いたPCクラスタ

江草俊文*・小畑正貴**

*岡山理科大学工学研究科博士課程システム科学専攻

**岡山理科大学工学部情報工学科

(1997年10月6日　受理)

1．はじめに

現在，科学技術計算はもちろん，データベース，マルチメディアなど，さまざまな応用分野でより高い計算性能が求められている。その結果，さまざまな並列計算機が研究・開発されており，性能を重視した専用並列計算機や，コストを重視し既存のワークステーション（WS）やパーソナルコンピューター（PC）を要素プロセッサと見立てEthernetなどのネットワークを付加することで並列計算機を構成するものなどがある。

多数のWSやPCをネットワーク接続して並列計算機を構成するWS/PCクラスタシステム[1)2)3)]は，量産による価格性能比の優れた要素プロセッサを利用でき，また最新プロセッサに素早く対応できるなどの利点をもっている。

WS/PCクラスタの構成として以下のような形態がある。

(1) 低速LANによる接続（Ethernet）

(2) 高速LANによる接続（ATMなど）

(3) 専用スイッチによる接続

順に，性能は高くなるが，価格も高価になる。

WS/PCクラスタでは，多くの場合それぞれのノードにハードディスクがあるので，各ノードがプロセッサとメモリのみで構成される並列計算機とは異なる応用が考えられる。入出力システム(特に分散ファイルシステム)，データベース処理，マルチメディアサーバなどである。また，高速な仮想記憶をサポートすることもできる。

また，一般ユーザの並列処理の利用を見てみると，並列処理に興味をもち必要性を感じながらも利用に至っていない計算機ユーザは多い。その理由としては，多くのWS/PCクラスタシステムではMPIなどを用いたメッセージパッシングによるプログラミング環境しか提供されないので比較的プログラミングが困難であること，利用環境が身近にないことなどがあげられる。

本研究では，低価格PCを単位プロセッサとし，単方向トーラスネットワーク[4)]を採用することで(1)程度の接続コストと(2)程度の性能を持つPCクラスタシステムの構築を行なう。また，通信ハードウェアにFPGAを利用することで，通信制御機能の変更を可能とし，メッ

セージパッシングライブラリ，分散共有メモリの実装や，さまざまな応用問題に対する実験を行なう予定である。

2章では，単方向トーラスネットワークの概要，ルーティングアルゴリズム，およびデッドロック回避法について解説し，3章では，PCクラスタシステムの全体構成，専用ネットワーク基板の構成について述べる。

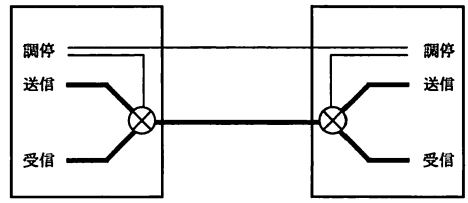
2．単方向トーラスネットワーク

2．1　単方向トーラスネットワークの概要

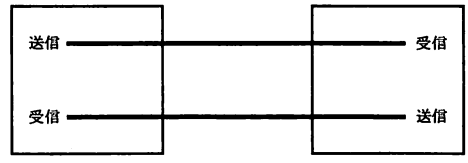
従来，並列計算機ではプロセッサ間の通信に双方向のネットワークを用いていた。双方向の通信路を用いた方が，単方向のものよりさまざまの面で柔軟な対応ができそうである。しかし，構成さえ十分に検討されていれば単方向のネットワークの場合でも，ルーティング自体は行なうことができる。しかし，単方向化することで経路長が伸びてしまうことが予想される。そこで，単方向のネットワークをうまく組み合わせることで経路長が長くなることを避ければ，パケットの密度が高い場合について，双方向のネットワークを用いた場合と比較して大差のない性能を得られる可能性がある。

また，双方向のネットワークでは図1に示すように，1組の通信路を送受信で共有し調停を行なうことで双方向の通信を実現する方法と，2組の通信路を設けることで双方向の通信を実現する方法がある。前者の場合では通信路は1組で済むが，その切替えのための回路と調停のための回路が必要になる。また，調停のためのバスフェーズが必要となり，それらを機能の一部として実装しなければならない。後者では，調停回路は不用であるが，通信路を2組持たねばならない。

これに対して，単方向の通信路だけでネットワークが構成できれば調停回路は不用にな



a) 通信路を共有する場合



b) 通信路を共有しない場合

図1　双方向通信の実装

り，通信路も1組で済む。したがって，ネットワークを構成するための配線量や，ルーターの規模の縮小が期待できる。

単方向の通信しか行なえないメディアを用いてネットワークを構成する場合を考えてみる。例えば，光ファイバーを用いた場合，双方向バスを構成するためには，2本1組で用いるほかないが，単方向ならば1組で済むため，高速なネットワークを低コストで実現できる可能性がある。

そこで，

●単方向化によってハードウェアを簡略化する。

●ネットワーク構成の工夫で直径を大きくしない。

●ルーティングや，デッドロックの回避がハードウェアで容易に実現できる。

の3点を最重要の条件として，次に示すような単方向2次元トーラスネットワーク(図2)を考えた。

●ネットワークの外観は2次元トーラスと同様である。

●プロセッサ間は単方向のネットワークで接続。

●列・行の番号が偶数か奇数かで通信可能な方向が異なる。

ようなネットワークである。

このネットワークは，単方向のネットワークを採用することでハードウェア量を減らし，通信可能な方向を互い違いに配置することで，直径が極力大きくならないように配慮した。

$N \times N$ ノードで考えた場合，双方向ネットワークでは，ネットワークの直径は N である。次に，単方向で南北・東西のすべてのネットワークの通信可能な方向を同じ向きにすると，ネットワークの直径は $2N$ となる。それに対して，互い違いに配置することで $N+2$ となり，双方向の場合とほぼ同じになる。また，平均距離を考えた場合でも，双方向の場合 $N/$

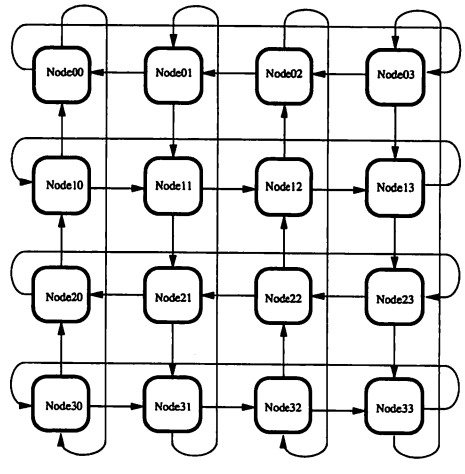


図2　単方向2次元トーラスネットワーク

2に対して，単方向は$N/2+1$となる。

実際，次に述べるルーティング法，デッドロック回避法などを用いて，ソフトウェアシミュレーションした結果，単方向2次元トーラスネットワークでは，近接ノードに対する通信性能は双方向のものと比較して劣っているものの，ランダム通信性能に関しては近接通信の場合と比較して，双方向2次元トーラスネットワークに近い通信性能が期待でき，特にパケットの密度が高くなる場合は，ほぼ同等となることが確かめられている[4]。

2.2 ルーティング

ルーティングアルゴリズムを決定するに当たって，次の点について考慮した。

(1) デッドロックの回避も考慮して全体のハードウェア量が少ない。

(2) ルーティングのアルゴリズムがハードウェアで実現しやすい。

(3) 最短距離を通るようにする。

この条件を満たすために，送信ノードを起点に，受信ノードを通過して送信ノードに至る最小の矩形経路をネットワークの通信可能な方向に逆らわないように決定し，その経路に沿ってルーティングを行なえばよい。

そのような経路は，次に示す単純なルールで決定できる。

(1) パケットが目的ノードに近付ける方に進む。南北方向，東西方向のどちらに送信しても近付ける場合は方向転換しない方向に進む。

(2) 近付ける方向がなければ目的ノードと現在のノードのノード番号の差（相対ノード番号）が南北方向，東西方向ともに奇数もしくは，偶数なら方向転換をしない方向

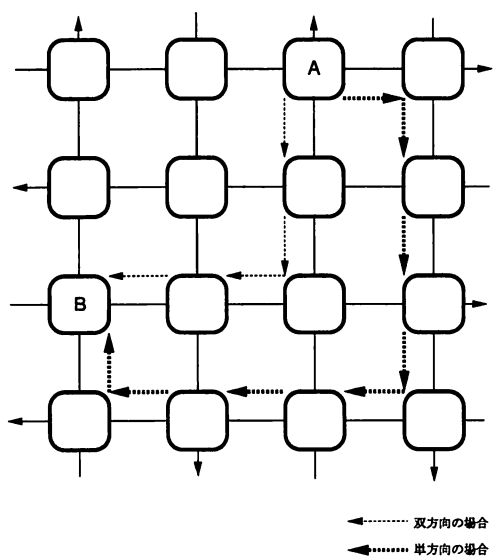


図3 ルーティングの例

に進む。

(3) 南北方向，東西方向の相対ノード番号が，奇数になるように移動する。

また，パケットが発生したノードにおいては，「方向転換しない方向」が存在しないが，その場合はどちらか任意の方向を選択すればよい。

このようにルーティングすることによって，ネットワーク上でのパケットの通信方向の変化が最小になり，次に述べるデッドロックの回避法と合わせて，仮想チャネル用のバッファ量を削減できる。また，常に経路長が最短になるような経路が選択される。

図3は実際のルーティングの例である。単方向にしたことによって，もっとも経路が長くなってしまうケースである。この例では，距離が4伸びているが平均的には1伸びるだけである。

また，上記のルールを実現するためには，常に宛先ノードとの相対座標を知っておく必要があるが，それ以外の判定は相対座標の最下位ビットのみで判定できるためハードウェアでの実現は困難ではない。

2.3 デッドロックの回避

デッドロックの回避については，ネットワークの物理的な循環構造を仮想チャネルを用いて論理的に断ち切る方法を採用する[5]。この方法は，仮想チャネルの構成に必要になるバッファの量を最小に押えることができる。また，仮想チャネルは，東西南北それぞれのネットワークで，独立して持たせる。

具体的には，次に示す条件でチャネル変更を行なう（図4）。

Case 1. ラウンドトリップループを用いる場合にはチャネル0からチャネル1に変更する。

Case 2. 方向転換をする場合は，必ずチャネル0に変更する。

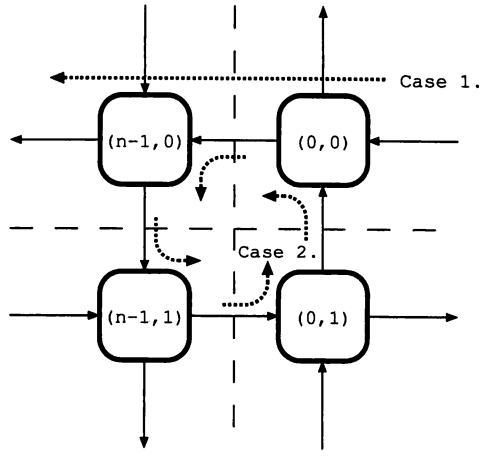


図4　チャンネル変更によるデッドロックの回避

このように仮想チャネルを変更すれば論理的な循環構造がなくなるので，デッドロックしない。

上記のルーティング法にしたがった場合に必要になる仮想チャネルは南北方向，東西方向のネットワークに対して，独立して2チャネルずつあれば良い。

3．PCクラスタ

3.1 全体構成

本研究では，ノード数が数個から数十個程度のPCクラスタシステムを構築する。また，各ノードには，安価に入手できるIBM PC/AT互換機を用い，64Mバイト程度のメモリと数Gバイトのハードディスクを接続し，さまざまな応用問題に対応できるようにする。

応用としては，現在のもっとも一般的な利用法である科学技術計算以外に各ノードのプロセッサに大容量のハードディスクを持つことを利用した用途も考えられる。よって，本PCクラスタの応用範囲としては，

(1) 科学技術計算

(2) 入出力システム（分散ファイルシステムなど）

(3) データベース

(4) マルチメディアサーバー

などである。特に，(2)，(3)，(4)に関しては，ノードプロセッサに大容量のハードディスクを持つことで非常に有利になると考えられる。

各ノード間は，専用インターフェースによる単方向トーラスネットワークと，Ethernet (100BASE-TX) で接続し，どちらか一方，もしくは両方のネットワークを用いたソフトウェアシステムを構築する。

また，各ノードではUNIX Like OSであるLinuxをベースとし，

- MPI，PVMなどによるメッセージパッシングモデル
- 分散共有メモリモデル

でのプログラミングができる環境を構築する。

OSとしてLinuxを選択した理由は，

- 安価に入手できる。

（表1） PCクラスタのハードウェア

ノードプロセッサ	IBM PC/AT互換機
ノード数	数個から数十個
メモリ	64Mバイト程度
ハードディスク	数Gバイト
ネットワーク	専用ネットワーク（PCI） 100BASE-TX

- Kernelも含めてすべてのソースコードが入手可能である。
- AT互換機で安定した動作が望める。
- ドキュメントが充実している。
- OSの再起動なしにデバイスドライバをロード・アンロードできるためドライバソフトの開発が短時間で行なえそうである。

などである。

単方向トーラスネットワークを構成するための専用インターフェースは，各ノードとPCIバスで接続する。PCIバスは，32ビット，64ビットのプロセッサアーキテクチャに依存しない高性能バス規格として普及しており，AT互換機はもちろん，ほとんどすべてのPCや多くのWSで採用されている。

3-2　専用ネットワーク基板

単方向トーラスネットワークを実装するためには，東西南北にネットワークを出さなければならないので，EthernetやATMなどのネットワークインターフェースを流用することが難しい。そこで，専用のネットワークインターフェースを開発する。ただし，PCクラスタの最大の長所であるコストパフォーマンスの良さを失わないように，できるだけ低コストであることが大前提である。

専用ネットワーク基板は，以下のことを前提として開発を進める。

(1) ノードプロセッサの配置の柔軟性をできるだけ犠牲にしないよう，数メートルから数十メートルのケーブルが使えるようにする。

(2) 単方向トーラスネットワークを構成するためには，入出力を各2チャネルと，フロー制御線が必要。

(3) 完全なハードウェアによるルーディング。

(4) 出力方向が異なるセルの同時ルーティング（最大3入力3同時出力）。

(5) PCIのバンド幅を十分に使えるようにする。

(6) 一枚のPCIカード上に実装できるようにする。

試作基板を作るに当たって，まず，(1)について。単方向トーラスを構成するためには，東西南北に通信路を持たねばならない。しかし，高速なパラレルを用いて長距離の通信を安定して行なうとは困難であるため，高速シリアルリンクで実装する。近年のシリアル通信技術の向上で，高速シリアル制御回路が1チップに収まり，簡単に扱えるようになったためである。試作基板では，AMD社のTAXIchip[7]を用いる。TAXIは1チップで最大175Mビット/秒のシリアル通信を行なうことができ，外部からはクロックを与えるだけで動作するので使用は容易であり，基板面積の縮小もできる。ケーブルはEthernet(100BASE-T)のケーブルを流用することにした。

(2)については，ケーブルの都合などで単方向のシリアルリンクを1チャネルと，フロー

コントロールとして2ビット分のパラレルを1ケーブルにまとめ，入力用2チャネル，出力用2チャネルを用意する。

(3)，(4)については，回路の書換えが容易なFPGAを用いて実装し，転送方式やルーティング方式を変えた実験ができるようにする。試作基板で用いたFPGA XC4010Eは1万ゲート相当で，内部に最大12,800ビットのRAMを持たせることができる。各通信路について2チャネルの仮想チャネルを持たせる必要があるため，ネットワークインターフェース上に仮想チャネル用のメモリが必要となるが，FPGAのRAMを用いることで部品点数を減らすことができる。

(5)，(6)については，複雑なPCIのバスプロトコルをFPGAで直接扱うのは難しいので，適当なBUS bridgeを利用することにする。今回試作基板で用いたPLX 9060は，1チップで2チャネルのDMA，双方向のFIFOを用いて0ウェイトのバースト転送への対応，バスマスターとしては132Mバイト/秒の転送ができる，ローカルバスはPCIより単純なi860のバスプロトコルであるなどの利点がある。

以上のことを考慮して現在製作中の，試作基板の構成を図5に，FPGAの内部構成を図6に示す。

FPGAには，高速シリアルリンクからの入力・出力，PCI（プロセッサ）への入出力，フローコントロール用のパラレルの入力・出力が接続される。内部では，シリアルリンク・PCIからのデータをバッファリングするためのメモリ，ルーティングの制御回路，宛先ノードを格納するためのアドレスレジスタ，クロスバスイッチ，PCI bus bridgeの制御回路，TAXIの制御回路，フロー制御回路などのすべてを入れる予定である。図6では直接ルーティングに関係する部分のみ示した。

TAXIと接続されたメモリは，入力・出力を同時に行なえるFIFOを構成するために，XC4000EのRAM機能を用いてダブルバッファ構成することによって実装する。また，転送方式としてはセル単位のストア&フォーワードを前提としているので，1セルは宛先アドレスなどを含んだヘッダと16バイト程度のデータにすることで必要なすべてのバッファをFPGA内部に実装できる。また，宛先ノードは次ノードとの相対座標で送信するので，ルーティングコントローラではルーティング結果に基づいて次ノードに送信すべき宛先アドレスを計算しなければならない。

入出力ポートは，3入力3出力のクロスバスイッチで接続するので，出入力，出力ポートが競合しなければ同時にルーティングすることができる。

作成中の試作基板を図7に示す。

4．ま と め

単方向トーラスネットワークの概要，ルーティング，デッドロック回避法について述べ，それを用いることでコストを削減したPCクラスタシステムの全体構成について述べた。ま

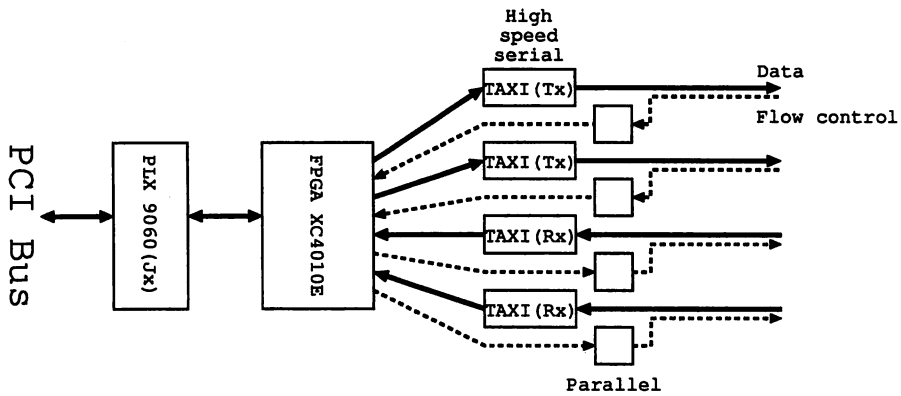


図5　ネットワーク試作基板の構成図

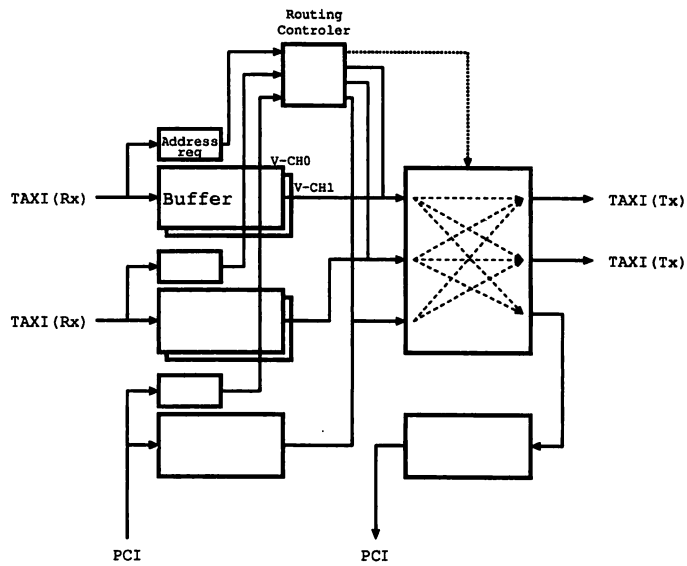


図6　FPGAの内部構成

た，単方向トーラスネットワークを構成するための専用ネットワークの試作基板の構成について述べた。

参考文献

1）http：//www. yamato. ibm. co. jp/rs6000sp/0030. html.
2）http：//www. think. com/Prod. web/index. html#gww 3）http：//now. cs. berkeley. edu/
4）江草, 小畑："単方向2次元トーラスネットワークの構成と，シミュレーションによる評価", 情研技報, Vol. 96, No. **106**, pp. 13-18 (1996).
5）天野 英晴：並列コンピューター，昭晃堂，1996.
6）PLX Technorogy："PCI Bus Interface and Clock Distribution Chips Product Catalog", 1996.
7）Advanced Micro Devices："Am7968/Am7969 TAXIchip Handbook", 1992.

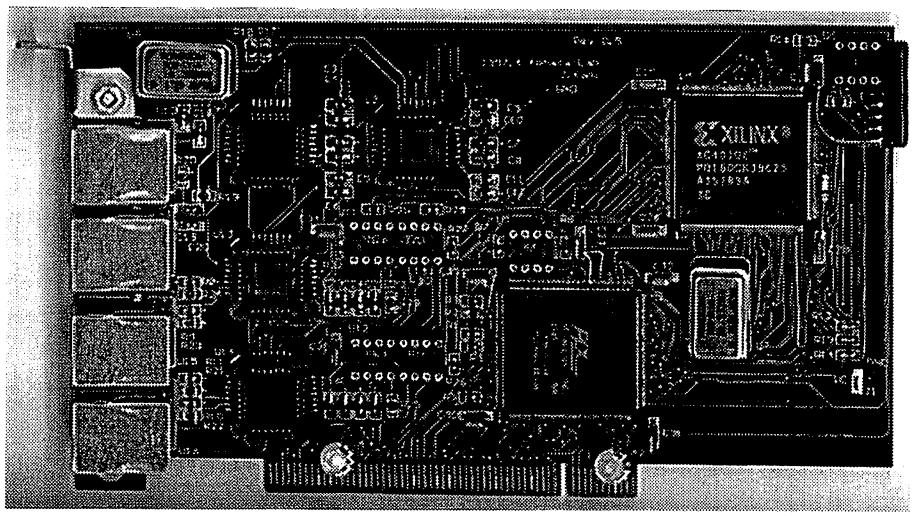


図7　試作基板の写真

Design of PC Cluster Using a One-way 2D Torus Network

Toshifumi EGUSA* and Masaki KOHATA**

*Graduate School of Engineering,

**Department of Information & Computer Engineering

Faculty of Engineering,

Okayama University of Science

Ridai-cho 1-1, Okayama 700-0005, Japan

(Received October 6, 1997)

This paper describes a design of PC cluster using the one-way 2D torus network whose each lhnk is one directional. The link directions are same on each row and column and each direction of lows and columns is reverse to their nearest neighbors. Routing algorithm and escape method of deadlok is introduced. Design of an experimental network interface card for PC cluster are also described.