計算機アーキテクチャ 136－4
（2000．1．7）

PCクラスタのためのネットワークカードの設計

西村公行† 小畑正貴††

安価なPCクラスタを構築するためにPCIバス上で動作する単方向2次元トーラス網用ネットワークカードを設計している。その特徴は高速化を図るために、物理面では液晶パネル用LVDS（Low Voltage Differential Signal）インターフェースを搭載している。加えてルーティングアルゴリズムには少ないリソースで構成できるMSN（Manhattan Street Network）を並列計算機のネットワークに応用できるように改良されたMSN/P（MSN with Proxy Send/Receive）を用いる。本稿では、我々が開発しているネットワークカードについて述べる。

Design of a Network Card for a PC Cluster

KIMIYUKI NISIMURA† and MASAKI KOHATA††

In this paper, we describe a new network card on a PCI bus, for a unidirectional two-dimensional torus network, in order to build a low price PC cluster. One feature is the use of a Low Voltage Differential Signal Interface (LVDS) panel for high speed data transmission. Also, we use an improved MSN with Proxy Send/Receive, an algorithm with few resources, as the routing algorithm.

1. はじめに

近年、さまざまな分野で計算機の高速化の要求に伴い、プロセッサの集積度や動作周波数による高速化が行なわれている。これと共にデータやプログラムを細分化し、複数のプロセッサで処理する並列処理による高速化も検討され、開発されている。このような並列計算を行なうシステムとして量産品のコンピュータを多数使用するWS/PCクラスタ[3) 6)]があり、Gigabit Ethernetやmyrinet[1)]のような高速ネットワークを用いたクラスタシステムの構築が注目されている。しかしながら汎用の高速ネットワークを利用するのは現時点では高価であり、100BaseTでは通信速度が遅い。そこで安価なPCクラスタを構築するためにPCIバス上で動作するネットワークカードを開発する。

その主な特徴として各PCに対してデータの受渡しを高速にするため物理面には液晶パネル用LVDS（Low Voltage Differential Signal）インターフェースを搭載した。またルーティングアルゴリズムにはMSN/P（Manhattan Street Network with Proxy Send/Receive）[2)]を用いた。MSN（Manhattan Street Network）とはローカルエリア・メトロポリタンエリアのパケット通信を行なうために設計された単方向の通信メディアをトーラス状に接続し，通信可能な方向を互い違いにすることで単方向の欠点をカバーしたネットワークである。このMSNを並列計算機のネットワークへ応用するためルーティング法における問題点を代理送受信ノードの概念を導入することにより改善したネットワークがMSN/P[5)]である。単方向のトーラスネットワークを利用することで双方向よりも少ないリソースで構成することができる[4)]。

2. LVDSインターフェース

2.1 LVDSの概要

マルチメディアやネットワーキングなどの分野において、高速のデータ転送レートが求められている現在では、今までの標準規格では、これらの要求に対応できない状態が生じてきている。このため、機器の処理能力を上げるために、従来のシリアル・インターフェースを同期式のパラレル・インターフェースにし、なおかつ伝送幅を広げたり、光ファイバの採用をしてきたがEMI対策の複雑化、線材の増加、消費電力の増加、コストの増加など課題が多い。これらの物理層ボトルネックの問題を解決する方法としてLVDS（低電圧差動信号）が提唱され定義された。このLVDS方式を搭載したものがLVDSインターフェース[10)]である。

† 岡山理科大学大学院工学研究科
The Graduate School of Engineering, Okayama University of Science
†† 岡山理科大学工学部情報工学科
Department of Information and Computer Engineering, Faculty of Engineering, Okayama University of Science

2.2 LVDS 信号の特徴

データ伝送線を 2 本用い、その差分を取ってデータを受信することにより同相ノイズを除去するので外的要因に対して強い。これにより低電圧でも正確にデータ伝送が可能となる。それに伴い信号を小振幅にすることで終端抵抗に流れる電流が小さくなり消費電力が少なくなる。また信号ラインに流れる電流の向きがツイスト・ペア線上でそれぞれ逆向きになるツイナックス・ケーブルを用いた場合の差動信号は電流により生成される磁界が互いにキャンセルする効果があり、互いの信号で余分な EMI の発生が小さくなり長距離伝送が可能となる。

同期伝送である LVDS インターフェース・チップにおいて周波数の変化に対応してデータ伝送されるので比較的容易に高速化を実現することができる。

2.3 LVDS インターフェース・チップの構造と性能

本ルータカードに搭載した LVDS インターフェース・チップは THine Microsystems の THC63LVDM83A（送信）と THC63LVDF84A（受信）であり、内部には CMOS/TTL のパラレル信号を LVDS 信号に変換する変換器（受信側は LVDS のシリアル信号をパラレル信号に変換）と入力信号と周波数や位相のズレが発生しないようにするため PLL 回路が組み込まれている。

送信側はコントロール回路から送られてきた 28bit の CMOS/TTL 信号をクロックの立ち下がりサイクルごとにサンプリングし、内蔵の PLL 回路で 4 チャネルの LVDS シリアル信号に変換する。1 チャネルあたりの 7bit のデータがサイクルごとに出力される。最大周波数は 70MHz であり、1 チャネルあたりのデータ伝送レートは 490Mbps になり、4 チャネルで 1960Mbps になる。

一方、受信側は送られてきた 4 チャネルの LVDS シリアル・データとクロックに対してシリアル - パラレル変換を行なう。受信側も PLL 回路を内蔵しており、クロックの立ち下がりで、入力された 7 bit の信号を検出するためのストローブ信号を自動的に発生させ、パラレル・データに変換する。

3. MSN with Proxy Send/Receive

3.1 MSN with Proxy Send/Receive の概要

並列計算機のネットワークでは，デッドロック対策が必須である。また，メッセージの順序を保証することで，通信プロトコルの単純化が可能となる。また，受信メッセージの再構築が不要となる。

そこで我々は，従来の MSN のルーティングに代理送受信の概念を採り入れることでデッドロックフリー，メッセージの順序の保証、効率的なブロードキャスト，容易な中継ノードでのルーティング制御が実現できる Manhattan Street Network with Proxy Send/Receive（MSN/P）を提案した。

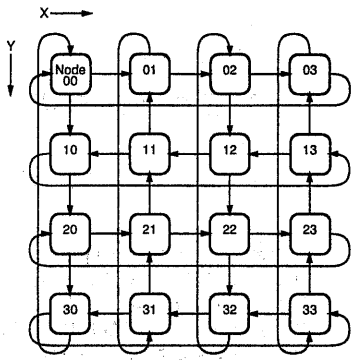

図 1 MSN with Proxy Send/Receive のトポロジ

MSN/P のトポロジを図 1 に示す。MSN/P でのルーティングは，X リンクを優先して使い，その後 Y リンクを用いて受信ノードへと転送する次元順ルーティングを行なう。ただし，MSN/P では通信可能なリンクの向きが強く制限されているので，単純に X リンクを用いると受信ノードから遠ざかってしまう場合がある。そのため，代理送信ノード・代理受信ノードという概念を導入することで，出来るだけ最短に近い経路を選択できるようにする。

(1) 必要があれば，代理送信ノードへ送信する。
(2) X リンクを用いて，受信ノードに近付く。
(3) Y リンクを用いて，受信ノードに近付く。
(4) 必要があれば，代理受信ノードから受信ノードへ送信する。

代理送信・代理受信ノードを用いることで，その間のルーティングは，単純な X リンク優先になり，かつ，大抵の場合は最短経路が選択される。

MSN/P では，まず，送信開始時にルーティング情報の生成を行なう．ルーティング情報は，代理送信・受信のフラグおよび X・Y の相対アドレスで構成される。代理送信・受信のフラグは，真なら代理送信・受信が必要であることを示す。相対アドレスは，送信ノードと受信ノードの絶対アドレスの差の絶対値である。しかし，代理送信ノードが必要な場合は送信ノードの変わりに代理送信ノードの絶対アドレスを用い，代理受信ノードが必要な場合は受信ノードの変わりに代理受信ノードの絶対アドレスを用いて相対アドレスを求めることとする。

3.2 ソフトウェアシミュレーションによる評価

単方向と双方向のトーラスネットワークの通信性能を比較するために，通信シミュレーターで作成した実験の結果を示す。シミュレーションの対象は，MSN,MSN/P と

Bi-dir 双方向 2 次元トーラスネットワーク．e-cube ルーティングを用いる．半二重で通信を行なう。

Simple 通信可能な方向がすべて同じ単方向トーラ

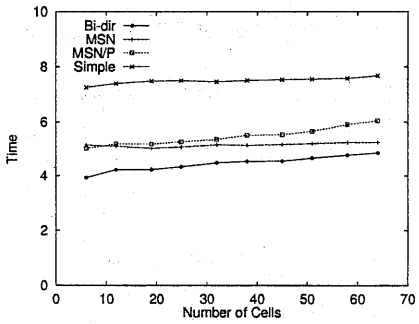

図 2　平均通信時間

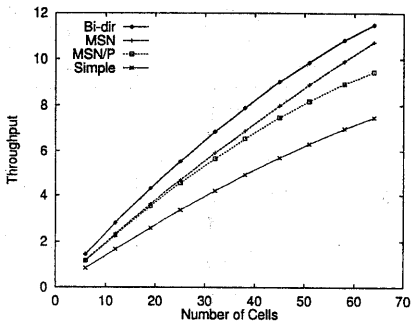

図 3　スループット

スネットワーク．次元順ルーティングを用いる。を含む 4 つのルーティングで行なう。

シミュレーションの条件として，

(1) 各ノードは，異る方向の送受信を同時に行なえる。

(2) 簡単のためセル単位で Store&Forward を用いる。セルとは，シミュレーションに用いるルーター上でルーティングに必要な情報を持ったヘッダと，固定長のデータを持っている。

(3) 1 セルの通信に必要な時間を 1 セル時間とする。「1 セルの通信に必要な時間」とは，受信バッファにあるセルを次ノードの受信バッファに格納するまでの時間である。したがって，セル中のデータ長はシミュレーションに影響しない。

(4) ノード構成は，図 4 に示したものを用いた。
MSN　X，Y 共に，深さ 4 の FIFO
MSN/P　X，Y の仮想チャネル 0 を深さ 2 の FIFO
Simple　X，Y のすべての仮想チャネルを深さ 2 の FIFO

(5) 時間 0 で複数のセルを同時に発生させ，すべてのセルが受信ノードに到達するまでの時間を“通信時間”とする．ただし，ルーティングの途中でセルを生成しない。

(6) ネットワーク上に常に同数のセルが存在するようにし，対象とするすべてのネットワークが定常状態になった時間 600 から，時間 1000 の 400 単位時間に到達した平均のセル数を“スループット”とした。

不規則で局所性をもたない通信において 8×8 ノードでの平均通信時間とスループットを図 2 と図 3 に示す。平均通信時間では，Simple が悪く，Bi-dir がよい．MSN，MSN/P はその中間となる。これは，MSN が通信可能な方向を互い違いに配置したために，Bi-dir と比較して Simple ほどネットワークの直径が大きくなっていないことを示している。また，MSN では動的ルーティングを行なっているが，動的ルーティングを行なわない MSN/P と大差のない結果が得られた。

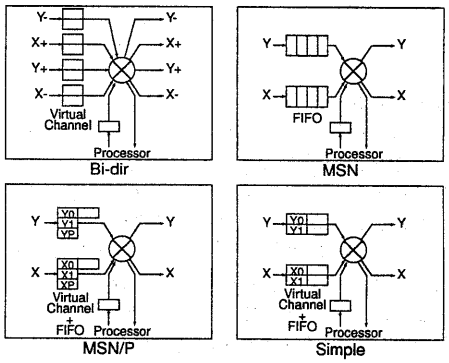

図 4　ノード構成

また，MSN 以外はセル密度があがるにつれて通信時間も伸びているが，MSN はほぼ一定である。

スループットも，平均通信時間と同じ傾向を示している。Bi-dir が有利だが，MSN，MSN/P ともに 2 割程度の悪化に留まる。MSN，MSN/P はほぼ同等のスループットが得られる。これは MSN/P はほぼ MSN と同様の性能であることを示している [5]。

4. 通信方式

4.1 転送方式

プロセッサ間のデータ転送はパケット（図 5）を単位として行ない、そのパケットのヘッダには転送先ノード，転送元ノード，パケット長などが格納されている。転送方式として

- Store & Foward 方式
- Wormhole 方式
- Virtual Cut Through 方式

などが挙げられる。ハードウエア上の制約から、Virtual Cut Through 方式ではすべてのパケットを格納できるバッファを用意することが困難であり、バッファ数の少なくてすむ Wormhole 方式では先頭のフリットがストップした場合，その後に続くフリットが，経路上のルータのバッファをすべて占有してしまう。単

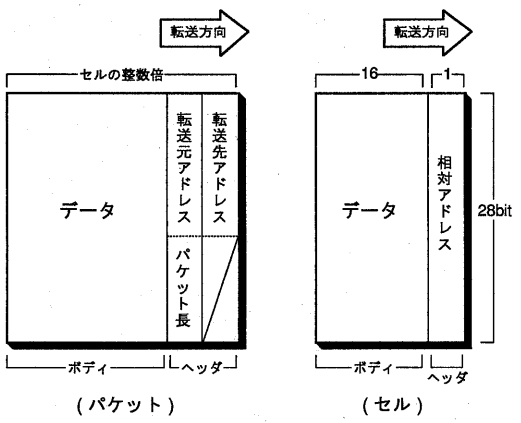

図 5　パケットとセルの構成

方向 2 次元トーラスネットワークにおいてはデータの転送の方向制限という制約によりデータがつまりやすく，転送性能の低下につながる [7]。

フリットより大きくルータに格納できる程度のバッファの大きさであるデータの単位をセル（図 5）としてデータ転送を行なう。また、これにより Wormhole 方式のようにパケットの最初から最後までをブロックすることがない。その変わりに，セルにはパケットのようなデータ転送の情報を必要とするためセルのヘッダにはパケットのヘッダから生成された転送情報を付加する。また，セルは固定長とし，パケットはセルのサイズの整数倍の可変長とする。

4.2　スイッチ構成

デッドロック回避のため仮想チャネルを用意する[6][7]。X 方向においては、通常チャネル、ラップアラウンドチャネル、代理受信チャネルが必要である。Y 方向では、通常チャネル、ラップアラウンドチャネル、代理送信チャネルが必要である。すべてのチャネルを通る場合、(Y 代理送信－X 通常－X ラップアラウンド－Y 通常－Y ラップアラウンド－X 代理受信）の順となる。 FPGA 内のデータパスを図 6 に示す。必要なバッファは HOST からの入力、HOST への出力、X 入力バッファ3 チャネル、Y 入力バッファ3 チャネルとなる。入出力間のスイッチは以下のようになる。X 出力に対しては、X 代理受信が除かれる（代理受信は連続しない）ので 6 入力となる。Y 出力に対しては、Y 代理送信と X 代理受信が除かれる（代理送信は連続せず、代理受信は HOST に出力する）ので 5 入力となる。HOST に対しては、Y 代理送信が除かれるので 5 入力となる。

4.3　フロー制御と誤り制御

制御には、図に示す 2 ビットの制御信号を用いる (ACK および NACK)。また、セルの先頭ワードには仮想チャネル番号が含まれている。通信はセル単位のストアアンドフォワードで、フロー制御と誤り制御も

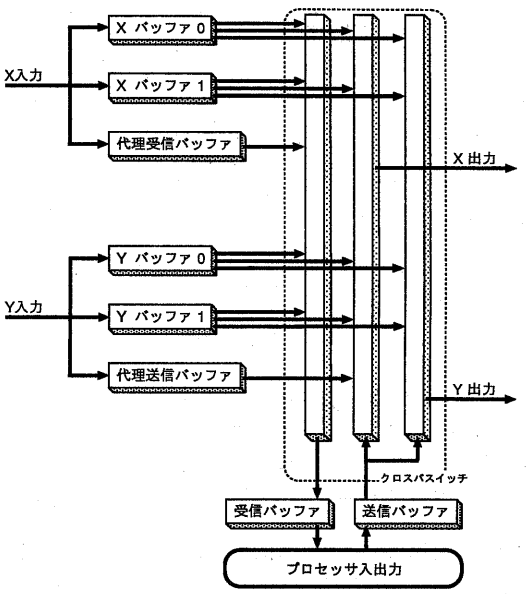

図 6　ルータのデータパス

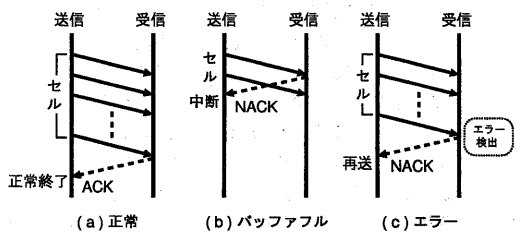

図 7　フロー制御と誤り制御

セル単位で行う（図 7）。

まず送信側はセルを送り始める。受信側でチャネルが開いていれば転送を続行する（図 7 (a)）。エラーが無かった場合、ACK を返す。チャネルが開いていなければ、受信側はすぐに NACK を返す。送信側は送信を中止し別のチャネルを使う通信に切り替える（図 7 (b)）。

誤り検出にはパリティを用いる。誤りがあった場合、受信側は NACK を返し、送信側は再送を行う（図 7 (c)）。

4.4　ルーティング

各セルのヘッダには（代理送信ビット：ps、代理受信ビット：pr、チャネル番号：ch、X方向距離：dx、Y方向距離：dy）が含まれる。

各方向からの入力セルに対する処理は図 8 のようになる。ここで、XEDGE と YEDGE はそれぞれ各方向における端ノード（ラップアラウンドリンクを使用するノード）において 1 とする。通常の次元順ルーティングに対して、ホストからの送信時における代理

```
from HOST:
 if(ps==1){ch=2; out(Y);}
 else if(dx!=0){dx--; ch=0; out(X);}
 else{dy--; ch=0; out(Y);}
from X:
 if(dx==0 && dy==0){out(HOST);}
 else if(dx!=0){dx--; ch=ch|XEDGE; out(X);}
 else {dy--; ch=0; out(Y);}
from Y:
 if(ps==1){dx--; ch=0; out(X);}
 else if(dy==0 && pr==1){ch=2; out(X);}
 else if(dy==0){out(HOST);}
 else {dy--; ch=ch|YEDGE; out(Y);}
```

図 8 ルーティングアルゴリズム

送信処理と、Yからの受信時における代理受信処理を付加したものであるが、比較的単純であり実現が容易である。

5. ルータカードの実装

図 9 に PCI バス・インターフェース[9] と FPGA (Field ProgrammableGate Array) [8] 及び VHDL インターフェース[10] を実装した基板の構造を示す。

PCI インターフェースには既製の PCI バスブリッジ IC である PCI9052 (PLX TECHNOLOGY) を利用し、それ以外の回路は FPGA で作成するいう方法をとっている。プロセッサから出されたデータは, PCI バスから PCI インターフェースを経由し, ルータである FPGA と送受信する。FPGA は XILINX の XCS40XL-PQ240C を使用している。このデバイスの概要を以下に示す。

Specifications of XILINX XCS40XL

Logic Cells	1862
Max System Gates	40,000
Typical Gate Range	13,000 - 40,000
CLB Matrix	28 × 28
Max. Available User I/O	205

LVDS インターフェースには液晶パネル用 LVDS リンクである THine Microsystems の THC63LVDM83A (送信) と THC63LVDF84A (受信) を用いている。ノード間のデータの転送はこの LVDS インターフェースによってシリアルで行なう。また送信と受信は各 2 系統あり、これにより単方向 2 次元トーラスネットワークの構築を可能とする。この高速差動信号を確実に伝送するためにコネクタとケーブルは住友スリーエムのツイナックスケーブル・ハーネスと汎用コネクタ MDR を利用した。この代表的な伝送性能を以下に示す。

伝送性能

特性インピーダンス	100ohms
チャネルスキュー	300ps. Max at 10m
エラーフリーデータレート	1Gbps at 10m

また PCI インターフェースとは逆向きに通信するためにハンドシェーク用のチップである差動ライン・ドライバとレシーバを用いている。

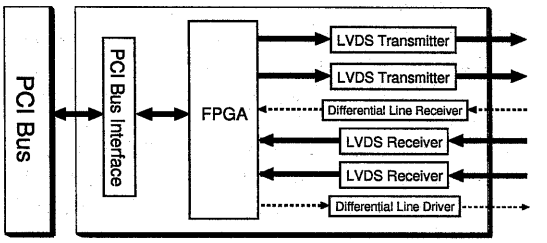

図 9 ルータカードの構成

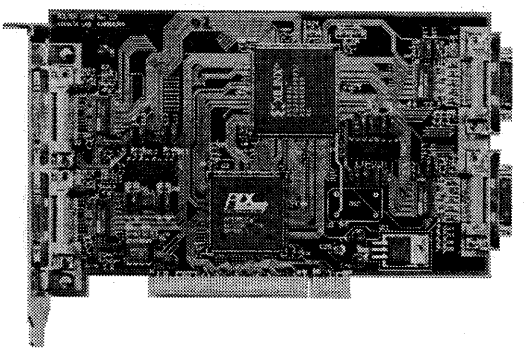

図 10 ルータカードの写真

図 10 が設計した基板の写真である。

6. おわりに

単方向トーラスである MSN/P を利用することによってルーティングアルゴリズムが単純になるという利点に加え、双方向より基板密度が低く済む。詳しく説明すると双方向の場合には設計した PCI のショート・サイズの基板にコネクタを 8 つ実装しなければならず、物理的に困難である。さらに FPGA のピン数が PQFP (Plastic Quad Flat Pack) の 240 ピンでは実装できない。ピン数を増やすために BGA (Ball Grid Array) の 256 ピンにしなければならず、コストが高くなってしまう。一方、単方向の場合にはそれらの問題は解消できる。

現在、MSN/P のルーティングアルゴリズムをルータカードに実装している。また、このルータカードの実用面での有効性を評価していく予定である。

参考文献

1) http://www.myri.com/
2) N.F.Maxemchuk. The manhattan street network. In *Proc. GLOBECOM'85*, pp. 255–261, Dec 1986.
3) 天野英晴. 並列コンピュータ. 昭晃堂, 1996.
4) 林匡哉, 堀田真貴, 大津金光, 吉永努, 馬場敬信. HDL 設計に基づく並列計算機ルータの評価. 情報処理学会研究報告 98-ARC-130, Vol. 98, No. 70,

pp. 79–84, 1998.
5) 江草俊文, 小畑正貴. Manhattan Street Network における代理送受信を用いたルーティング手法. 情報処理学会論文誌,Vol. 40, No. 5, pp. 1968–1976, 1999.
6) 田中良夫, 松田元彦, 安藤誠, 久保田和人, 佐藤三久. COMPas: Pentium Pro を用いた SMP クラスとその評価. IPSJ Symposium Series, Vol. 98, No. 7, pp. 343–350, 1998.
7) 可児純一, 江草俊文, 小畑正貴. PC クラスタのための単方向 2 次元トーラス網用ルータ. 電気・情報関連学会中国支部第 48 回連合大会講演論文集, p. 449, Oct 1997.
8) "Spartan and SpartanXL Families Data Sheet," XILINX,1999.
9) "PCI9052 Data Sheet," PLX TECHNOLOGY,1997.
10) "LVDS 24Bit COLOR HOST–LCD PANEL INTERFACE DATA Sheet," THine Microsystems,1998