

並列推論マシンPIM/pのネットワーク

5N-3

久門耕一*, 服部彰*, 後藤厚宏**

(kumon@flab.fujitsu.junet, hattori@ayumi.stars.flab.fujitsu.junet, goto@icot.junet)

*富士通株式会社 **新世代コンピュータ技術開発機構

1. 始めに

われわれは、第五世代コンピュータプロジェクトに関する研究の一環として、並列推論マシンの研究を行なっている。

本論文では、PIM/pのクラスタ間を接続するネットワークの概要について述べる。

PIM/pは、100台規模のCPUを8CPUを一つのクラスタとして共有バスにより接続し [1][2]、クラスタ間を本節に述べるクラスタ間ネットワークにより相互に接続している。以下に、PIM/pのネットワークの構造について述べる。

2. ネットワーク構造

ネットワークの性能は、ほぼ結合線路の数に比例するが、実際に実現する場合にはネットワークを構成するスイッチや、伝送路の実現に関しての多くの制限から、必ずしも理想的な構成が取れない。本ネットワークでは、ハイパーキューブ構成を選択した。

2.1. ネットワークの実装方法

ネットワーク用のきょう体を特別に設けずに、クラスタのなかにクラスタ内に1ないし2枚のボード上に実装する。使用する回路テクノロジーとしてはCPUと同程度の規模のCMOS VLSIを用いる。

ネットワークの構造として、ノード上にCPUを接続した構造を採用し、トポロジとしてはハイパーキューブ構成を選んだ。ノード上にCPUが接続された構造のネットワークは、一般にノード間の距離の平均値が小さく、またネットワークスイッチの分散配置に適する。

2.2. ネットワークの規模

基本構造で128台(16クラスタ)を考えているので4次のハイパーキューブとなる。しかし、拡張性を確保するために、1次大きい5次のハイパーキューブ構

造を取る。ネットワーク伝送路は1バイト巾として、ネットワークスイッチのピンネックを回避し、実装を簡単にする。

2.3. ネットワークスループット

クラスタからネットワークに出力されるデータのスループットは40MB/s程度と見積もられている。クラスタ間通信をランダムだと考えると、ハイパーキューブではクラスタ間アークを通過するデータのスループットはクラスタから出るデータスループットの1/2となるので、クラスタ間接続路は20MB/sのスループットであるが余裕を見込んで40MB/sとした。これを、1バイトの通信路で実現するためには、40MHzのデータ転送レートが必要である。現在われわれが用いようとしているテクノロジーでは、20MHzのクロックが上限と考えられるので、クラスタ間の通信路を2重に張って通信容量を2倍にする。

2.4. ネットワークとクラスタの接続方法

8PEからなるクラスタに対して2重に通信路を接続するために、4PEに1つのネットワークノードを割り当て、それぞれのPEにはネットワークインタフェース(NIU)と呼ぶコプロセッサ構成のインタフェースを付ける。

KL1の処理において、通信はメッセージ転送の形態を取るが、メッセージの大きさは、通信内容により大きく異なり、数10バイト程度の大きさのパケットが多いと考えられる。

したがって、PIM/pのネットワークでは、可変長のパケットを扱えるようにし、最大パケット長を256バイトに設定した。

NIUとRTRとの間には、クラスタ内のバックパネルを用いて専用接続することで、共有バスに対する負荷を減らす。(図1) 1つのRTRを4PEとのみ接続することにより、ネットワークノードを構成するネットワークルータ(RTR)のピンネックも緩和される。2つのRTRは、独立にハイパーキューブを構成する。従って、システム全体は、2重化されたハイパーキューブ構造となる。NIUとRTRとの間の通信速度は

20MB/sである。この様に、NIとRTRとの間のデータスループットを最大80MB/sとすることで、ハイパーキューブの持つネットワーク性能を無駄無く使うことが出来る。

3. ネットワークインタフェースユニット(NIU)

NIUは、CPUのコプロセッサとしてCPUボード上にある。CPUとNIUとの間は、コプロセッサ命令によりデータ転送を行う。NIU内部には、256バイトの packets バッファが送受それぞれ2つ有る。

送信時にNIUはCPUからデータ転送の要求を受けると packets ヘッダを付けてCPUから送られてきた転送データと共にRTRに送出する。NI内部のバッファに空きが無い時には、CPUからNIへのデータ転送を要求を受けた時点でCPUにバッファフルを通知し、バッファが空いた時点でCPUへ割込により空きを通知する。

受信時には、RTRから packets を受けると内部のバッファに格納した後CPUに対して通知を行う。KL1でのデータ転送の単位は、タグ付きデータ、命令コード、制御情報など大きさが一定していない。このようなデータをCPUに負担を与えずに効率良く転送するために、CPUとNIUとのデータ転送巾は、1/2/4バイト巾をコプロセッサ命令により選択できるようにした。

4. ネットワークルータ(RTR)

RTRは、ルータボード上に実装され、NIU向きに4ポート、他のクラスタのRTRに対して5ポートの出力を持つ。PIM/pでは、1きょう体に8クラスタが実装されるが、きょう体内部は20MB/sの電氣的結合、きょう体間は200Mbpsの光リンクによる結合を考えている。

ネットワーク上で、デッドロックを起こさないためには、

- (1)ルーティングを制限する。
- (2)構造化バッファプール法を採用する。

の2点が考えられるが、本ネットワークでは、ネットワークスループットを低下させるホットスポットを回避するために動的ルーティングを採用し、構造化バッファプール法を用いる予定である。シミュレーションによれば、これにより50%程度のスループットの向上が得られる。しかし、パケットの順序性が保証できないため、ソフトウェアのオーバーヘッドが増加するため、現在ネットワークの性能向上を

計るシミュレーションを行っている。ネットワークを用いた負荷分散をサポートするために負荷分散に関するソフトウェアの負担を軽くするために、ネットワークによる動的負荷分散機構を検討中である。

謝辞

日頃、ご指導御助言を頂く ICOT 内田第4研究室長、富士通研究所林人工知能研究部長ならびにICOT、富士通のPIM研究開発メンバに感謝いたします。

参考文献

- [1] 後藤他、並列推論マシンPIM/pの概要、本大会予稿集
- [2] 篠木他、並列推論マシンPIM/pの要素プロセッサ、本大会予稿集
- [3] Gelernter, D. A DAG-Based Algorithm for Prevention of Store-and-Forward Deadlock in Packet Networks, IEEE Trans. Comput., Vol. C-30, No. 10, pp.709-715

図1. クラスタとネットワークの接続

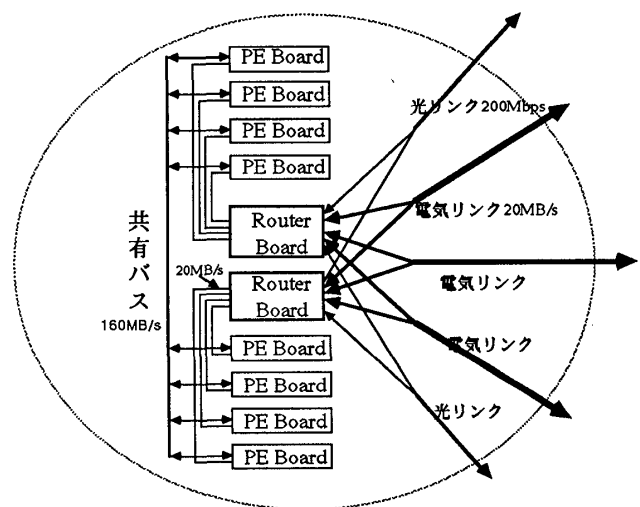


図2. ネットワークによるクラスタ間接続

