直接結合型クラスタ並列計算機のネットワークインターフェイスにおけるメモリモデルの検討.

加藤 渉†, 三浦 康之†, 渡辺 重佳†,
湘南工科大学†

1. はじめに

近年、複数のコンピュータを結合し高い処理速度・信頼性を得ることを目的とした様々なクラスタ並列計算機が実用化されている。

クラスタ並列計算機において、クラスタ間との通信性能を向上させることは演算性能を向上させる方法の一つとされている。ハードウェアによる専用の高速ネットワークや、ソフトウェアによる通信制御の改善など様々な方法で通信性能を向上させる研究が行われている[1]-[4]。

本研究では、リング網を用いた小規模向けのクラスタ並列計算機を想定し、通信性能の向上を計る。本発表は、ネットワークインターフェイスに搭載されたメモリを効率良く利用するために、1系統のメモリを、チャネルを用いずに使用する新たなメモリ管理手法を提案する。クラスタ並列計算機の通信で多用されるショートパケットに対し、シミュレーションを行うことで比較・検討を行う。

2. クラスタ型並列計算機

本発表ではクラスタ並列計算機のノード数を4～32台と小規模のものを想定している。構築する際には、比較的安価に構築できるようハードウェア量を少なくする。そのため、トポロジはバス型かリング型が望ましい。バストポロジは、複数のノードが1本を共有するため、トラフィック量の増加によりパケットの衝突が高くなり通信性能の低下が起こりうることや、ネットワークの局所的な通信が不可能である。リングトポロジは、ノードの障害がネットワークに影響する。しかし、小規模を想定しているためノードによる障害が起こる影響が少ないことや、双方向の通信の確保ができるためリングトポロジを採用した。

A Memory model for Network Interface of direct-connected cluster computer
†Wataru Kato, †Yasuyuki Miura, †Shigeyoshi Watanabe
†Shonan Institute of Technology

3. 通信方式

3-1 従来手法による通信

双方向通信をリングトポロジでデットロックを起こさず通信するためには、2個のチャネルが必要になる。そのため、メモリを2つに分割するのが一般的である。2本のチャネルを使用した通信手法を用いた構成を図1に示す。

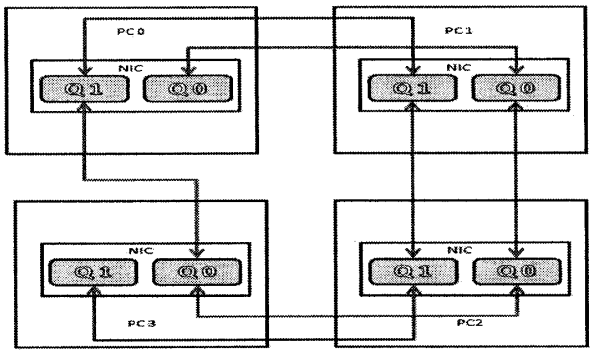

図1 2本のチャネルによる構成図

図1はノード数4の構成図になる。NICに搭載されているメモリを2つのチャネルに分割している。また、矢印は通信経路であり、双方向通信を行うためシミュレーションでは逆回りも同様の操作を行っている。

パケットの送信時の処理の流れを下記に示す。

1) 現在いるノードのPCIバスが空いている場合、キューにパケットを送信する。PCIバスが混雑している場合、パケットは待機する。
2) パケットの送信先ノードの情報を元に、送信先の経路を決定する。送信先のキューの情報を所得し、空いている場合は送信する。キューが満杯の場合は現在いるキューで待機する。
3) 目的のノードにパケットがたどり着いた場合、PCIバスが空いている場合、パケットに到着情報を付与する。PCIバスが混雑している場合、キューで待機する。
4) 1～3までに必要と時間を測定する。

3-2 提案手法による通信

従来手法では、デットロックを防ぐためにチ

ャネルを 2 本使用ためメモリを分割する必要がある。しかし、限られたメモリを有効に活用するためには、メモリを 2 つに分けずに可能な限りメモリ領域を共有することが望ましい。チャネル用のキューの容量を減らし、残りの容量を共有することで効率化を図る。図 2 に提案手法を用いた構成図を示す。

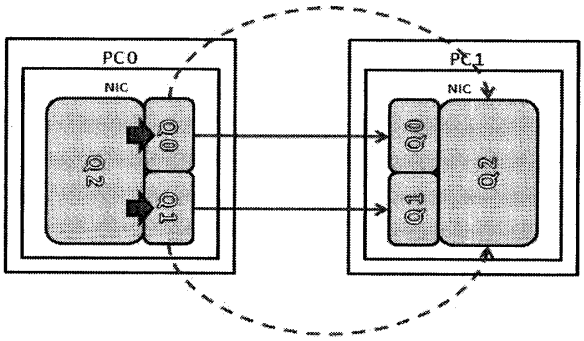

図2 提案手法による構成図

図 2 では、Q0・Q1 を非共有領域、Q2 を非共有領域としている。送信は非共有領域で行う。受信は非共有領域で行われるが、非共有領域があふれた場合に共有領域に情報が送られる。

パケットの送信時の処理の流れを下記に示す。

1) 現在いるノードの PCI バスが空いている場合、共有領域または非共有領域にパケットを送信する。PCI バスが混雑している場合、パケットは待機する。
2) パケットの送信先ノードの情報を元に、送信先の経路を決定する。送信先の非共有領域の情報を所得し、空いている場合は送信する。空いていない場合は、送信先の共有領域へ送信する。どちらも満杯の場合は現在いるキューで待機する。
3) パケットが目的のノードにたどり着いた時、PCI バスが空いているならばパケットに到着情報を付与する。PCI バスが混雑している場合、共有領域または非共有領域で待機する。

4. 実験

4.1 実験方法

提案手法の性能評価のため、シミューレーションプログラムによる実験を行った。シミュレーションプログラムは、パケットの生成確率を指定することで指定した個数のパッケットがランダムに生成される。生成されたパケットは、現在いるノードの位置と送信先ノードの位置がランダムで決定される。転送方式はストア＆フォワード方式で行われる。

実験はノード数 16 のリング網、メモリ容量 256word で行う。したがって、従来手法は 1 キューにつき 128word となる。提案手法では、共有領域と非共有領域の合計が 256word となるよう各領域を割り振る。

表1 2つの手法のメモリ割り当て

	従来手法	提案手法	
QUEUE 0	128 word	64 word	16 word
QUEUE 1	128 word	64 word	16 word
QUEUE 2	0 word	128 word	224 word

4.2 実験結果

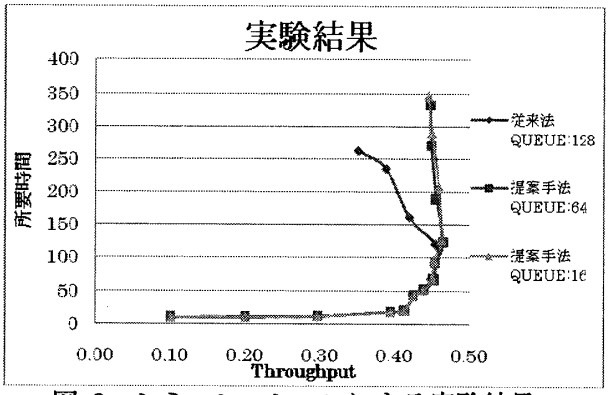

図 3 シミュレーションによる実験結果

従来手法はネットワーク中のパケットが増大すると大幅にスループットが低下することがある。これは、キューにパケットが満たされてキューが満杯状態になるとパケットの移動が阻害されることが原因だと考えられる。提案手法では、パケットが増大しても共有領域へ退避することが可能となるため大スループットの低下が抑えられる。

5. まとめ

本稿では、ネットワークインターフェイスに搭載されているメモリを効率的に利用するための提案を行った。その結果として、従来手法に比べネットワークが混雑したさい、提案法ではスループットの低下が抑えられることが明らかになった。

6. 参考文献

[1] 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

[2] 青木圭一，山際伸一，和田耕一，小野雅晃，Maestro2 クラスタネットワーク向けメッセージパッシングライブラリの開発と評価

[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] 小畑 正貴公子,FPGA における差動信号入出力を用いた PC クラスタ用ネットワークインターフェイス