

並列計算機用ネットワークのルーティング方式

2X-5

堀江 健志 池坂 守夫 石畑 宏明

(株)富士通研究所

1. はじめに

我々は、数値計算、映像生成の高速実行を目的とした高並列計算機を開発している。既に、256台の要素プロセッサ(セル)から構成される並列計算機CAP-256¹を開発し、その有用性を確認した。この経験に基づき、現在、さらに大規模で高性能な並列計算機を開発中である。

本稿では、相互結合網(ネットワーク)のルーティング方式を提案し、その方式を用いて、アレイ型ネットワークを評価する。

2. ネットワークの設計方針

ネットワークの設計にあたり、数値演算と映像生成への応用を考え、以下の点に留意した。

- (1)任意セル間通信における中継処理を各セルのCPUの介在なく、ハードウェアで高速に行う(自動ルーティング)。
- (2)メッセージの送信から受信までの遅延時間を小さくする(小レイテンシ)。
- (3)ネットワーク全体で通信が行われているとき、ネットワークの一部分の輻輳によりネットワーク全体の性能が低下することがない(高スループット)。
- (4)1セルに物理的に接続されるワイヤの数、1ボードに複数のセルをのせたときのボードに接続されるワイヤ数等のハードウェア規模を小さくし、1000台規模のセル構成がとれる(拡張性)。

3. ルーティング方式の原理

2.の設計方針から、レイテンシが小さいことを特徴とするワームホールルーティングに構造化バッファプールのアルゴリズムを取り入れることにより、小レイテンシかつ高スループットを実現するルーティング方式を考案した。ここでは、本方式について述べる。

3.1 ワームホールルーティング²

ワームホールルーティングでは、メッセージのヘッダが入力チャンネルから出力チャンネルへ中継ルートをつくりながら、メッセージが送り出されていく。ストア・アンド・フォワード・ルーティングでは、中継ノードがメッセージ全体をストアするのに対し、ワームホールルーティングでは

、フリットと呼ぶ数バイト(ビット)のデータのみが中継ノードにストアされる。あるセルがメッセージのヘッダを受信すると、中継ルートのチャンネルを選択し、フリットをそのチャンネルへ転送する。後続のフリットはヘッダのフリットが選択したルートと同じルートに転送されていく。

ワームホールルーティングの特徴は、レイテンシが小さい点にある。一方、1つのメッセージが転送されている間、そのメッセージが使用しているチャンネルをブロックするので、①デッドロックの発生、②スループットの低下、を起こす可能性がある。

3.2 提案するルーティング方式

本論文で提案するルーティング方式は、ワームホールルーティングに構造化バッファプール³のアルゴリズムを取り入れることにより、デッドロックの発生とスループットの低下を回避している。これは、1つのメッセージが転送されている間、チャンネルをブロックすることがないからである。以下、図1の例を用いて、この方式について述べる。

各セルは、フリットをストアするためのバッファを用意する。このバッファは、「セル間の最大距離+1」個のフリットをストアできる大きさである。

4つのセルが単方向のチャンネルで接続されているとすると、各セルは、4つのフリットをストアできる大きさのバッファを持つ。

バッファの使用方法は、セル①からセル②へメッセージを送信するときは、「セル①のクラス1⇒セル②のクラス2」、セル①からセル③へメッセージを送信するときは、「セル①のクラス1⇒セル②のクラス2⇒セル③のクラス3」、セル①からセル④へメッセージを送信するときは、「セル①のクラス1⇒セル②のクラス2⇒セル③のクラス3⇒セル④のクラス4」である。セル②、セル③、セル④のときも同じように、「クラス1⇒クラス2⇒クラス3⇒クラス4」、を使用する。

各セルのクラスにフリットがストアされると、次のセルにそのフリットを転送する。転送するとき、データのクラスとデータそのものを転送する。どのクラスのフリットを転送してもよく、フリット転送ごとに転送するクラスをかわることができる。

本アルゴリズムは、まず、ワームホールルーティングにおけるデッドロックを回避している。これは、どのセルからもクラス1⇒クラス2⇒クラス3⇒クラス4という経路が存在し、ループを形成しないからである。

次に、本アルゴリズムにおいて、全てのセルが、同時に、右回りに、①⇒③、②⇒④、③⇒①、④⇒②、へ転送する
ときを考えると、すべてのチャネルを用いて転送が行われ、
スループットの低下を引き起こすことはないことがわかる。
本方式は、チャネルが両方向の場合、あるいは、セルの接
続チャネル数が多い場合にも適用することができる。

4. 評価

本ルーティング方式は、アレイ型ネットワークに適用
することができる。ネットワークのアーキテクチャを検討す
るため、次元の異なるアレイ型ネットワークのスループッ
ト性能をシミュレーションにより求める。なお、アレイの
端どうしはトラス状に接続されている。

4.1 ネットワーク・シミュレータ

性能評価を目的としたソフトウェアによる「ネットワ
ーク・シミュレータ」を開発した。

ネットワーク・シミュレータは、 n 次元のネットワ
ークのとき、まず一次元内のセルでルーティングを行い、次に、
二次元内のセルでルーティング、最後に n 次元内のセルで
ルーティングというようにルーティングしていく。次元内
のルーティングは、3.で述べた方式により、静的な最短ル
ーティングを行う。ただし、1チャネルを双方向に使用す
る。

4.2 スループット性能

スループット性能として、ネットワーク使用率を求める。
ネットワーク使用率は、CPUがネットワークヘデータを
送信しようとしたとき、データを送信することができる確
率である。ネットワークが輻輳状態になるとネットワーク
がデータを送信できなくなり、使用率は低下する。ネット
ワーク使用率は、全セルの平均の値とする。いつも送信で
きたとすると、ネットワーク使用率は100%となる。

送信は、可能な限り送信する。受信は、メッセージが到
着するとただちにデータを受信するものとする。

図2に、台数一定のもとで、次元を変えたときのネット
ワーク使用率を示す。メッセージの長さは8ワードとする。
図から例えば1024台構成のときの二次元アレイ構成と
ハイパーキューブとの性能比は、1対2.5であり、これは、
平均セル間距離の比5対16よりも小さい値になっているの
がわかる。

1チャネルのビット幅が同じであれば、ハイパーキュー
ブが二次元あるいは三次元アレイよりも優れているのは明
らかである。しかし、ハイパーキューブはハードウェア量
が多く、結果として1チャネルのビット幅を小さくする必
要がある。例えば、1台のセルに接続されるワイヤ数を等
しいとすると、1024台構成のとき二次元アレイとハイ
パーキューブの1チャネルのビット幅の比は、10対4にな

り、提案したルーティング方式を用いれば、スループット
においてほぼ同程度の性能が得られることになる。なお、
レイテンシと拡張性においては、二次元アレイの方が優れ
ているのは明らかである。

5. おわりに

本稿で述べたルーティング方式を用いて、小レイテンシ
かつ高スループットを満たすネットワークを構築できる見
通しを得た。現在、本ルーティングを実現するルーティン
グチップを開発中である。

参考文献

- 1) 石井他：高並列計算機CAP, 信学論D, J71-D, 8
- 2) Dally, W.J.: A VLSI Architecture for Concurrent
Data Structures, Kluwer, Hingham, MA, 1987.
- 3) Merlin, P.M.: Deadlock avoidance in store-and-
forward networks-I, IEEE Trans. commun., COM-28

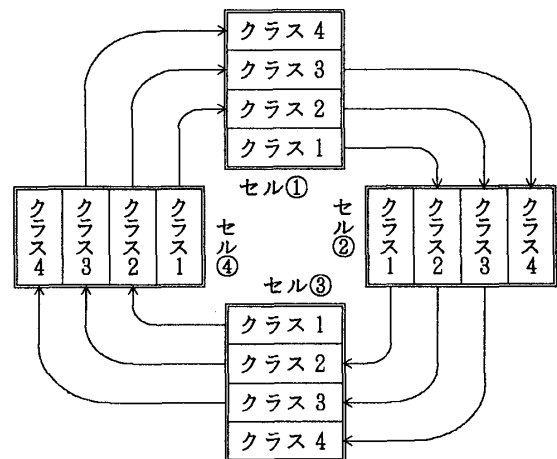


図1 ルーティング方式のバッファ使用方法

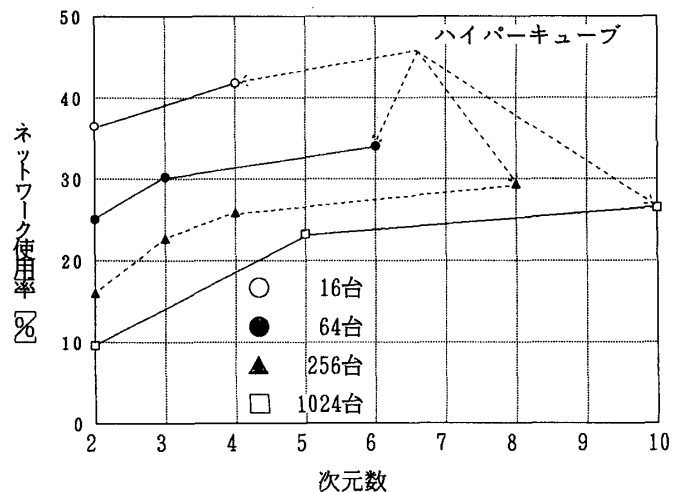


図2 次元数とネットワーク使用率