3H−5

分散制御型全対全通信結合網

野口泰生　武理一郎　横田治夫
(株) 富士通研究所

1はじめに

マルチプロセッサ上での並列処理において通信のオーバヘッドは大きな問題である。この分野での研究として局所性の高いアルゴリズムの開発や高性能なネットワークの開発が盛んに行われいる。しかしデータベースのジョインやソートなどの演算をマルチプロセッサ上で並列処理で行う場合には全ノードが全ノードと一斉に通信する事態は避けらない。またこのような場合には、いかに高性能なネットワークがあったとしても各ノードがランダムに通信を行う限りパケットの衝突などが起こりネットワークのバンド幅を使いきることができない。

われわれのアプローチは衝突の原因となるランダムな通信を避けて、ネットワークの幾何学的な構造にうまく写像できる通信パタンを用いて通信の効率を上げることである。以前報告したドラゴンルーチングは上述のジョインやソートの全対全通信をグループに分解し超立方体構造に最適に写像するアルゴリズムである。このアルゴリズムを実行するネットワーク、ドラゴンネットにより全対全通信を効率よく処理することが可能になった［1、2、3］。

しかしドラゴンネットでは集中制御を前提としていたのでスイッチを切り替える度に通信準備時間が必要であった。またこのような制御機構は規模が大きくなるにつれ実現が難しくなる。

今回この欠点を補うためインテリジェントなスイッチを用いて分散制御でドラゴンルーチングを実現する新ドラゴンネットを考案しトランスピュータでシュミレーションを行った。

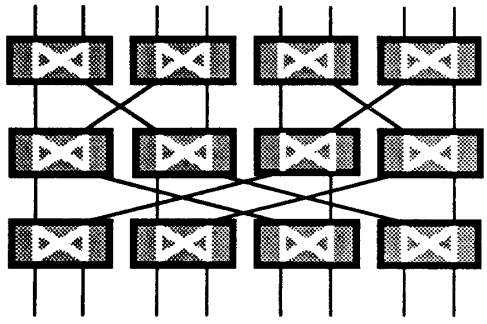

図1　分散型ドラゴンネットの構成

All-to-all Communication Network
Controled By Local Inteligence
Yasuo Noguchi,Riichirou Take,Haruo Yokota
Fujitsu Laboratories Ltd.

2分散制御型ドラゴンネット

分散制御型ドラゴンネットは図1のようなbinary n-cubeの構成をとっている。このトポロジはOmegaに代表されるlogn段の多段結合網と同じものである。

binary n-cubeネットワークでのドラゴンルーチングは、図2のようにフェーズsの時、点vは点v⊕sにパケットを送信する。フェーズが0からn−1までを遷移すると各点は自分を含む全ての点に1回づつ送信し全対全通信が処理される。ドラゴンルーチングによれば衝突を起こさずかつ全てのリンクを使いきるため、最適時間で全対全通信を処理することができる。但し、各点からのメッセージの送信順は予めスケジュールされてしまい各点の自由にはならない。

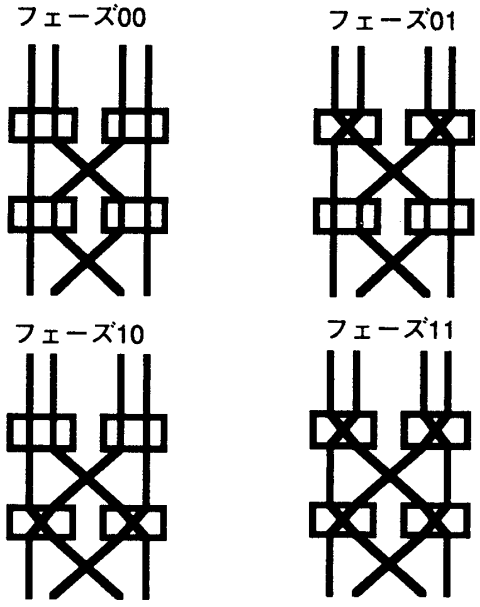

図2　ドラゴンルーチング

スイッチエレメントはドラゴンルーチングを実現するために図3のような自律制御機構を持っている。この制御機構は以下のような特徴を持っている。

1分配器と集約器はそれぞれ局所フェーズを持っている。

2分配器はリンクからパケットを受信し、局所フェーズからドラゴンルーチングにより定まる集約器に送信する。
3集約器も局所フェーズから定まる分配器からパケットを受信し、リンクに送信する。
4分配器と集約器はハンドシェークでパケットを送受信する。
5パケットにフェーズ 切り替えタグがあると分配器及び集約器はそれぞれの局所フェーズを進める。切り替えタグ自身は前のフェーズのパスを進む。

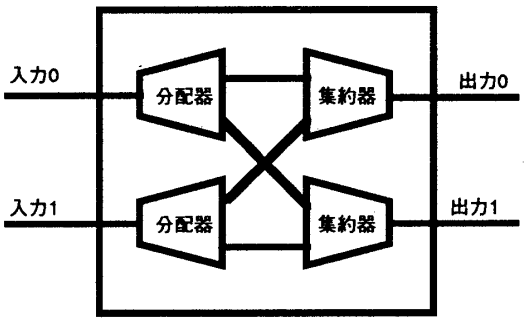

図3　スイッチエレメント

このネットワークでは分配器から集約器へ、集約器から分配器へとパケットが媒介して同期が全体に連鎖する。このため自然にフェーズの逐次性が保持される。

リンクとスイッチエレメントはダブルバッファを用いて並行動作する。このため分配器と集約器の間でパケットをコピーする時間やフェーズ切り替えの時間はリンクによるパケットの転送時間に隠すことができる。

3 実験

ドラゴンネットをトランスピュータネットワークでシュミレートした。ネットワークは16ノード構成で32個のスイッチエレメントを持つ。

スイッチエレメント1個にT800を1個用いた。分散制御機構はすべてOCCAMで記述した。分配器および集約器はOCCAMのプロセスで実現した。両プロセスのハンドシェークはOCCAMのチャネルで実現した。

図4の縦軸は全対全通信時におけるリンク1本あたりのパケット転送速度である。横軸はスイッチエレメントのバッファ長である。Vmaxは本実験系におけるT800の転送速度の上限である。T800の転送速度の上限は1.7MB/s程度であるがマザーボードからT800のリンクを引き出すために市販のスイッチをつかっているので1.3MB/sとなっている。

バッファ長が500Byte以上ではドラゴンネットの通信効率V/Vmaxは98%以上あり、全対全通信時でもネットワークのバンド幅を使いきっていることがわかる。バッファ長が小さいとき通信効率が下がるのはダブルバッファの効果が少ないのでスイッチエレメント自体がCPUネックになっているせいである。

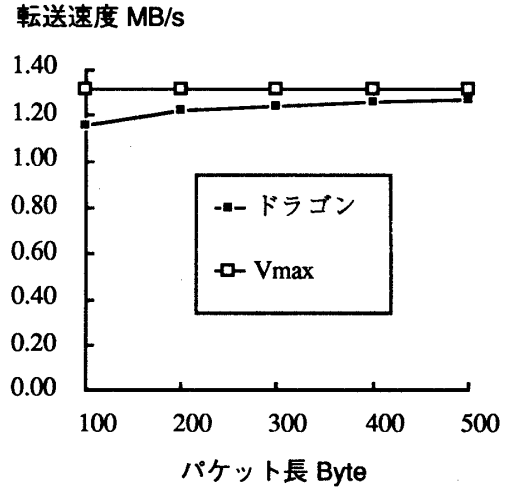

図4　ドラゴンネットの通信効率

4 まとめ

全対全通信を最適に実行するネットワークについて述べた。

このネットワークは
1 binary n-cubeのトポロジを持つ。
2 全対全通信を衝突を起こさずかつ全てのリンクを使いきることによって最大効率で実行する。
3 スイッチエレメンは自律的に同期ししフェーズの逐次性を保証する。

さらにトランスピュータを用いたシミュレーションによりこのネットワークが有効であることを検証できた。

参考文献

［1］武、超立方体形ネットワークに於ける全点対全点通信の最適ルーティング法、情報処理学会第35回全国大会、1987

［2］山根、武、Parallel Partition Sort for Database Machine、Proc. IWDM、1987

［3］野口、武、ソート及びハッシュジョインの並列処理、情報処理学会第38回全国大会、1989