並列計算機 KUMP/D の相互結合網

5L－5

馬場 朗　津留 健治　川野 哲生
冨安 洋史　谷口 倫一郎　雨宮 真人

九州大学総合理工学研究科

1 はじめに

我々は並列計算機 KUMP/D[1] を製作中であるが、ネットワークは並列計算機の性能を左右する重要な構成要素の一つであり、対象とする計算機の実行モデルを十分に考慮して設計されなければならない。本稿では KUMP/D のネットワークの実装方式について報告する。

2 並列計算機 KUMP/D

KUMP/D は MIMD 型の細粒度並列計算機であり、プロセッサエレメント (PE) は 2D トーラス網で結合されている。各 PE は Datarol-II アーキテクチャ[1] に基づいて設計されており、KUMP/D では市販 CPU と通信・同期処理を行う付加ハードウェアで実現される。

KUMP/D は細粒度実行を行い、パケットを非常に短い固定長パケットとして送出するので、ネットワークに大きな負荷を与える。MUMP/D で用いられる Datarol-II アーキテクチャは耐レイテンシ性能において優れた性能を有するので、ネットワークはレイテンシの短さよりもスループットを重視した設計をすることができる。

3 ネットワーク

3.1 設計方針

製作中の KUMP/D は PE の実現方式を検証するための試作版であり、ルータチップはいくつかのアルゴリズムを実験的に実装できるように FPGA を使用している。FPGA はゲートアレイ等に較べ、利用可能なハードウェア量が少ない。結果的にこのハードウェア量の不足はスループットの低下を引き起こすと考えられる。設計の際には FPGA のハードウェアを有効活用することにより、この性能低下を可能な限り抑え、高スループットを実現することを目標とする。

最適なルーティングアルゴリズムは、ネットワークの大きさに依存する。従って、ネットワークの大きさが変化した場合には、ルーティングアルゴリズムは最適ではなくなり、ネットワーク性能は低下する。そこで、KUMP/D はパワーオン時に FPGA が ROM から読み込む内部配線情報を変更することによって、ネットワークの大きさに適したアルゴリズムを使い分ける。その結果、ネットワークサイズが変化した場合にもハードウェア資源を有効活用し、優れたネットワーク性能を得ることができる。

3.2 転送アルゴリズム

デッドロック回避とスループット獲得のためにバッファを多重化する手法として、バッファ予約方式 [2] や、バッファ予約方式のバッファ数を削減した negative hop 方式 [2] がある。バッファ予約方式では、パケットが最初にネットワークに投入されたときにはクラス 0 のバッファを使用し、その後はホップする度にバッファのクラスを 1 ずつ増加させていく。その結果、チャネル依存グラフにサイクルが含まれることは無くなりデッドロックを回避できる。このときに必要とされるクラス数はネットワークの各次元あたりの直径+1 である。negative hop 方式ではクラスの増加は 2 ホップに一回であり、クラス数を約半分に削減することができる。

KUMP/D のルータチップが構成可能なバッファクラス数は、パケットサイズと FPGA の容量から計算すると、ポートあたり最大 5 なので、最大構成時の 16×16 ネットワークでは negative hop 方式を用いる。このときのクラス数はポートあたり平均 4.5 クラスである。

同アルゴリズムを 8×8 ネットワークに適用すると、使用するクラス数は 2.5 となり、このクラス数の減少はチャネルブロック増加によるスループット低下を招く。この問題を解決するために、8×8 ネットワークではバッファ予約方式を用いることにより、使用していないバッファを利用する。その結果、使用クラス数は 3.5 になる。

4×4 ネットワークでは利用可能なバッファ数は更に減少する。バッファ予約方式を用いてもクラス数は 1.5 である。いま、各次元方向で独立して使われているバーチャルチャネルを連結する。クラスの使用法は、(a) パケットがネットワークに投入された時使用するクラスは目的地までの距離に等しい、(b) パケットはホップする度にクラスを 1 ずつ減少する、である。こ

Interconnection Network for the Parallel Computer KUMP/D
Akira Baba ,Kenji Tsuru, Tetsuo Kawano, Hiroshi Tomiyasu, Rin-ichiro Taniguchi, Makoto Amamiya
Department of Information Systems, Graduate School of Engineering Sciences, Kyushu University
6-1, Kasuga-koen, Kasuga, Fukuoka 816 JAPAN

[1] Kyushu University Multi-media Processor on Datarol-II

のときのクラス数は、x-y ルーティングを行う場合には、x 方向のポートでは 3.5、y 方向では 1.5 であり、平均すると 2.5 になる。以上のようにクラス数を増やすことによってチャネルブロックを減らすことが可能となる。

3.3 仕様

ルータは 4 方向に隣接するルータとの間の 4 本のリンクと自 PE 用のリンク、計 5 本のリンクをもつ。各リンクは 24bit 幅であり、これを 12bit 単方向のリンク 2 本として用いる。その他の仕様を表 1に示す。

表 1: ルータ仕様

最大構成台数	256 台
パケット形式	72bit 固定長 (ヘッダ部 8bit)
バッファクラス数	最大 5
ルーティング方式	x-y ルーティング
フロー制御方式	Store and Forward

4 評価

ネットワークのスループット性能を検証するため、論理レベルのシミュレーションを行った。

4.1 ランダム通信

ランダム通信時の平均スループットを図 1に示す。4×4,8×8,15×15 [2] ネットワークについてアルゴリズムを変化させてシミュレーションを行った。図中の"_neg" は negative hop アルゴリズムを表し、"_ext"は 8×8 ではバッファ予約方式を、4×4 ではバッファ予約方式を拡張したアルゴリズムを表す。図 1の縦軸はリンク (12bit 幅) が 1clock に転送したパケットの bit 数であり、スループットを表す。横軸はパケットのネットワークへの投入間隔 (clock) である。図 1より、ランダム通信時の最大スループットは、図中の最大値 3.4bit/(link · clock) からパケットヘッダを除いた 3.0bit/(link · clock) である。

4.2 1 対 1 通信

1 対 1 通信時のスループットをランダム通信時と同様に計測したところ、スループットは 6.0bit/(link · clock であった。ヘッダ部を除いたスループットは 5.3bit/(link · clock) である。

1 対 1 通信時のスループットはローカルポートのスループットの最大値に等しく、この最大値は PE のパケット処理機構の送受信能力によって決定されている。

5 考察

ランダム通信時のスループットの飽和は、8×8_neg、8×8_ext、15×15_neg 方式では、リンクのスループット飽和によるものである。また、8×8 の場合の 2 つのグラフからクラス数を増やしたことによって、最大スループットが向上したことが分かる。4×4 での最大スループットは、パケットの平均転送距離が短いため、自 PE とルータ間のスループットに支配される。この結果、リンクに要求されるスループットは 3.0bit/(link · clock) であるので、4×4_ext 方式は理想値の約 90%を満たしているといえる。4×4_neg 方式では、ネットワークに投入されたパケットの使用するクラス数が固定されていることによって、チャネルブロックが頻発し、スループットが低下していると思われる。

[2] 16×16 はシミュレータの問題により計測不能であった

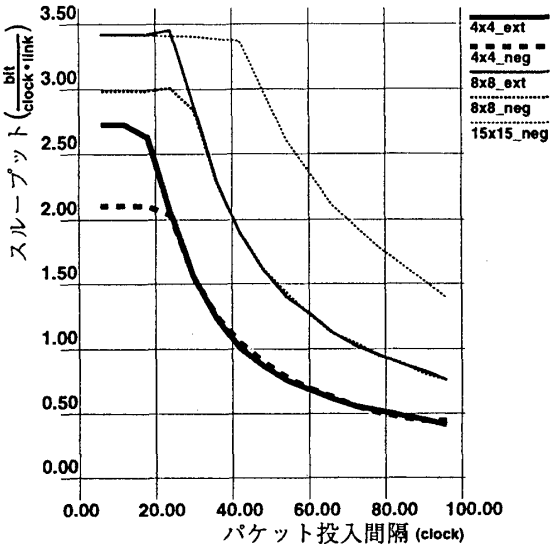

図 1: ランダム通信時の平均スループット

6 おわりに

ネットワークの大きさの変化に対して、バッファの使用アルゴリズムを変化させることにより、スループットの低下を防ぐことが可能となる。その結果、ランダム通信を行った場合でも、1 対 1 通信時の約 56%のスループットが得られた。

また、現時点でのルータチップの動作周波数は 4.0MHz であるが、まだ改善余地は多く残されており、速度向上が今後の課題となる。

参考文献

[1] 川野哲生, 日下部茂, 谷口倫一郎, 雨宮真人. "細粒度マルチスレッド処理向けプロセッサ Datarol-II の構成とその評価". 情報処理学会論文誌, Vol. 36, No. 7, pp. 1700–1708, 1995.

[2] Anujan Varma and C.S.Raghavendra. *"Interconnection Networks for Multiprocessors and Multicomputers Theory and Practice"*, pp. 242–257. IEEE Computer Society Press, 1994.