ハイパクロスバ・ネットワークにおける適応ルーティングの実現方法とその評価*

1B－3

曽根 猛、板倉 憲一、朴 泰祐、中澤 喜三郎、中村 宏†
筑波大学 電子・情報工学系‡
{sone,itakura,taisuke,nakazawa,nakamura}@arch.is.tsukuba.ac.jp

1. はじめに

超並列計算機において、ネットワークはシステムの性能を左右する大きな要因の一つである。各種ネットワーク・トポロジの中で、パイパクロスバ・ネットワーク（HXB）は高速かつ柔軟な転送性能を持つ[1]。同ネットワークにおいて、従来はデッドロック回避のために固定ルーティングによるメッセージ転送が行なわれてきたが、適応ルーティングを行なうことでより高い性能の向上が期待される。

本研究では、Virtual Channel を用いて HXB におけるデッドロック・フリーな適応ルーティングの実現方法を提案し、その性能評価を行なう。

2. ハイパクロスバ・ネットワーク

n 次元 HXB は、1 次元方向の各 PU をクロスバ・スイッチ（XB）で完全結合しそれを n 次元空間に配置したものであり、各次元毎に任意のサイズをとることができる。各 PU は EX と呼ばれる小規模なクロスバ・スイッチで各次元方向の XB と接続されているので、n 次元 HXB においては最大 n 個の XB を通過すること（n ステップ）でメッセージ転送ができる。図 1 に 3 次元 HXB の構成を示す。

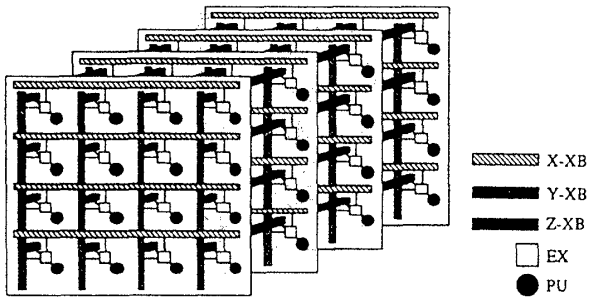

図 1: 3 次元 HXB（4 × 4 × 4）

3.Virtual Channel による適応ルーティング

HXB において Wormhole ルーティングを行なう際、従来はデッドロック回避のために固定ルーティングが用いられてきたが、固定ルーティングではメッセージが衝突した場合に他の空いているチャネルを利用せず、その場でブロックしてしまう。そこで、動的に空きチャネルを見つけ、そちらに転送をするような適応ルーティングを用いることでより性能を向上できる。ただし、そのまま適応ルーティングを行なうとデッドロックを起こしてしまうので、Virtual Channel[2] を導入することによりデッドロック・フリーを保証する適応ルーティングを提案する。

n 次元の HXB において、各 EX および各 XB の出力部に n 枚のバッファを用意し、一つの物理チャネルを n 本の Virtual Channel として利用する。Virtual Channel の割り当ては、そのメッセージがこれまでに経由した XB の数によって決定する。n 次元の HXB では、メッセージは最大 n 個の XB を経由して転送されるため、n 本の Virtual Channel を用いれば複数のメッセージ間で互いにブロックし合うことがなくなる。例えば、3 次元 HXB において X→Y→Z の経路で転送されるメッセージ A と Y→Z→X の経路で転送されるメッセージ B は中継点の EX が一致した場合は互いにブロックし合う可能性がある。しかし、Virtual Channel を用いるとメッセージ A は $X_1 \rightarrow Y_2 \rightarrow Z_3$ の経路で、メッセージ B は $Y_1 \rightarrow Z_2 \rightarrow X_3$ の経路で転送されるため、互いにブロックし合うことはなくなり、デッドロック・フリーな適応ルーティングが可能となる。これは、以下のようにして証明できる。

PU を P_{xyz}、Virtual Channel を $Cvdis$ で表す。ここで、v は Virtual Channel 番号、d は次元番号、i は次元 d における destination EX の番号、s はチャネルの source EX の番号である。メッセージは各 XB を通過するとき、過去に通過していない任意の次元を選択できる。その際、v が現在のステップ数と等しい Virtual Channel を使用するものとする。これにより、v は単調増加するので Virtual Channel 番号 $Cvdis$ も単調増加となる。よって、channel dependency graph においてサイクルがなくなるので、このルーティングはデッドロック・フリーである[2]。

また、HXB における経路選択は EX 上で行なわれるが、EX における出力が決まると XB における出力も一意に決まる。そこで、EX は接続している XB の全出力バッファの状態を監視し、EX の出力バッファと XB の出力バッファがともにフリーのときのみメッセージを転送するように制御する必要がある。さらに、実際の Wormhole ルーティングではメッセージの先頭が XB の出力に到達するまでに遅延が生じるので、それまでの間はバッファを予約するような制御も必要となる。

4. シミュレーションによる性能評価

評価はランダム転送を対象とし、以下の仮定のもとで計算機シミュレーションを用いて行なう。

システム中の全 PU は完全に非同期に動作し、PU はメッセージ転送の命令を発行したらすぐ次の動作に移ることができる（send&forget）。各 PU は以下の動作をシミュレーション時間内、繰り返すものとする。各 PU はランダムに相手 PU を選んで一定長のメッセージを転送する。1 つのメッセージを転送する毎に一定時間の内部処理を行なう（今回は 50clock）。もし、メッセージを転送する際に PU からネットワークへの出力がビジーのときはその間、内部処理を行ない、出力がフリーになったらメッ

*An Adaptive Routing Algorithm on Hyper-Crossbar Network
†Takeshi SONE, Ken'ichi ITAKURA, Taisuke BOKU, Kisaburo NAKAZAWA, Hiroshi NAKAMURA
‡Institute of Information Sciences and Electronics, University of Tsukuba

セージを転送する。また、PU に到着したメッセージは割り込み処理により適宜受信されるものとする。通常、メッセージの送受信は DMA コントローラによって行なわれるので、送信処理と受信処理を同時にできるものとした。

ルーティングは、転送方式を Wormhole ルーティング、転送経路の決定を適応ルーティングとした。また、メッセージ間のチャネル競合の調停はランダムに処理することが望ましいが、今回は実装の簡単さからメッセージが PU から送出された時刻により優先度がつけられるとした。

システムの規模は 1024PU の 3 次元 HXB ($8 \times 8 \times 16$) を対象とした。XB および EX の各チャネルのバンド幅を 1 に正規化し、1flit/clock でメッセージが転送されるものとする（メッセージ長の単位は flit)。3 次元 HXB では、最大 3 つの XB とその途中の 4 つの EX を経由するので PU 間の最大距離は 8 となる。

評価は 1PU 当たりの理想的なメッセージ転送の場合をスループット 1 とし、正規化して行なった。理想的な場合とは、メッセージがオーバーヘッド 0 で送出され無衝突転送が行なわれる場合で、T clock のシミュレーションでは 1PU から転送されるメッセージの総量は T flit となる。ここでは、10000clock のシミュレーションを行なったので、各 PU が転送した平均総メッセージ量を 10000 で割ったものをスループットとした。

適応ルーティングにおいて、x, y, z の各方向からメッセージが到着することを想定して、PU↔EX 間の物理チャネルの本数を 1 本、2 本、3 本と変化させた。固定ルーティング（FIX）と適応ルーティング（ADP）におけるメッセージ長とスループットの関係を図 2 に示す。凡例中の数値 (1,2,3) は PU↔EX 間の物理チャネルの本数を表す。

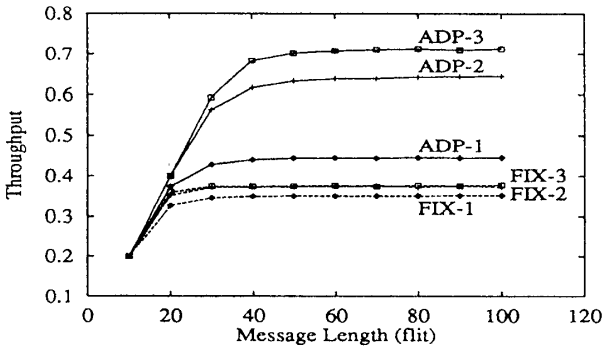

図 2: 適応ルーティングによる性能向上

図 2 において、固定ルーティングと適応ルーティングのどちらの場合もメッセージ長が長くなるにつれてスループットが増加している。これは以下の理由による。メッセージ長が短いときは、PU の内部処理が終らないうちにメッセージが相手 PU に到着するためネットワーク中の総転送量が少ない。それに対し、メッセージ長が長くなるとメッセージが相手 PU に到着しないうちに次のメッセージを転送しようとするためネットワーク中にメッセージが満たされ、総転送量が増加するからである。

固定ルーティングと適応ルーティングを比較すると、PU↔EX 間の物理チャネルを変化させた場合のいずれにおいても、適応ルーティングの方が良い性能を示している。メッセージ長が短いときはネットワークが空いているので固定ルーティングと適応ルーティングで性能差は見られないが、メッセージ長が長くなりネットワークが混雑してくるにつれて適応ルーティングは固定ルーティングに比べて、大幅に性能が向上することがわかる。

PU↔EX 間の物理チャネルの本数を変化させた場合について、適応ルーティングでは、かなりの性能向上が見られるが、固定ルーティングでは、ほとんど性能は変わらない。固定ルーティングにおいては $x \rightarrow y \rightarrow z$ の順でルーティングを行なうとしたため、PU から送出されたメッセージはいつでも最初に x 方向に移動し、EX の x 方向の出力においてほとんどのメッセージが衝突する。また、ランダム転送であるため PU に到着するメッセージのほとんどが EX の z 方向から移動してくる。以上の理由により固定ルーティングでは、PU↔EX 間の物理チャネルが複数本あったとしても有効に使われないため、性能が向上しない。次に、適応ルーティングにおいて、PU↔EX 間の物理チャネルを 1 本から 2 本にしたとき大幅に性能が向上している。これは、1PU からネットワークに送出された 2 つのメッセージは適応ルーティングにより異なる次元方向に移動することができるため衝突する可能性がほとんどなくなる。さらに、メッセージは x, y, z の各次元方向から PU に到着するため、PU↔EX 間の 2 本の物理チャネルを有効に利用しているからである。それに対し、PU↔EX 間の物理チャネルを 2 本から 3 本にしたときは先の場合に比べると性能向上は鈍っている。これはランダム転送で 1PU に同時に 3 方向からメッセージが到着する可能性が低いことと、ネットワークが非常に混雑しネットワーク中に溜っているメッセージが多くなっていることの 2 つによるものと考えられる。

5. まとめと今後の課題

本研究では、従来、固定ルーティングのみを対象として評価されてきた HXB において、Virtual Channel を利用した適応ルーティングの手法を提案し、計算機シミュレーションを用いてその有効性を評価した。結果として、適応ルーティングを行なうことにより従来の固定ルーティングに比べ大幅な性能向上が達成できることが分かった。

HXB における他の性能向上の手法として、従来の Wormhole ルーティングの代わりに Virtual Cut-Through ルーティングを適用する方法がある [3]。これは、本研究における適応ルーティングと相反するものではないので、今後は、今回提案した適応ルーティングの手法を Virtual Cut-Through ルーティングにおいて応用した場合の性能を評価していく予定である。

謝辞

本研究に関し貴重な御意見を頂いた筑波大学西川博昭助教授ならびに中澤研究室諸氏に深く感謝します。なお、本研究の一部は文部省科学研究費（奨励 (A)05780225）および創成的基礎研究（05NP0601）の補助による。

参考文献

[1] 朴 泰祐 他, “ハイパクロスバ・ネットワークの性能評価”, 信学技報 CPSY93-40, pp.41-48, 1993 年

[2] W.J.Dally,et.al., “Deadlock-Free Message Routing in Multiprocessor Interconnection Networks”, IEEE Trans. Computer, Vol.C-36, No.5, pp.547-553, May 1987

[3] 三島 健 他, “ハイパクロスバ・ネットワークにおけるバーチャル・カット・スルーの性能評価”, 第 48 回情報処理全国大会, 1994 年 3 月