2G－1

ハイパクロスバ網における迂回ルーティング・アルゴリズムに関する研究*

田辺 雅明　原田 智紀　朴 泰祐†
筑波大学 電子・情報工学系‡

1. はじめに

これまで提案されてきた数多くの超並列計算機用ネットワーク・トポロジーの中で，ハイパクロスバ網は特に優れた転送性能を持っていることが確認されている．ハイパクロスバ網ではデッドロック・フリーを保証するために，従来固定ルーティングが用いられてきた．しかし，固定ルーティングでは転送経路上に１箇所でも故障チャネルが存在すると，そこから先へメッセージを転送することはできず，システム全体が稼働しなくなるという問題点がある．

また，高い計算性能の要求から近年数千台規模の超並列計算機が登場してきたが，そのチャネル総数は数万本にも達し，チャネルに故障が発生する可能性は決して無視できないものとなっている．

本研究では，文献 [1] により提案された Static Algorithm をハイパクロスバ網に適用し，ハイパクロスバ網には従来無かった耐故障性を持たせることで，経路上に発生した故障に対処するルーティング手法を提案する．また，転送性能の評価によりその有効性を示す．

2. ハイパクロスバ網

n 次元のハイパクロスバ網（以下 HXB と省略）は，n 次元の直交座標の各格子点上に PU を並べて，直交座標と平行に配置したクロスバ（以下 XB と省略）と，PU と XB を繋ぐエクスチェンジャ（以下 EX と省略）により PU 間を間接的に結合したネットワークである．図１に３次元 HXB (4×4×4) を示す．XB と EX はクロスバスイッチで構成されており，これにより XB は各次元方向の EX 同士を完全結合させることができ，EX は各次元の XB と PU を完全結合させることができる．

3. HXB における迂回ルーティング

HXB における迂回ルーティングとは，１度転送した次元方向にメッセージを再び転送することであり，次元オーダーを破って高次元から低次元への移動が起こる．従って，迂回ルーティングを行なうとデッドロックの危険が生じるため，デッドロック・フリーを保証するための何らかの手法が必要となる．

3.1 デッドロック・フリーを実現する手法

本研究では，文献 [1] において提案された Static Algorithm を HXB に適用して，デッドロック・フリーを実現する．この手法では，DR (Dimension Reversal) 番号を各バーチャル・チャネルに対応させることでネットワークを複数のクラスに分ける．DR 番号とはパケットが高次元から低次元へ移った回数を表すものである．

HXB における次元オーダー・ルーティングはデッドロック・フリーであることが既に分かっている [2]．従って，あるクラス上で次元オーダー・ルーティングを行なっている間はデッドロック・フリーが保証されるが，これを破るとデッドロックの危険が生じる．そこで，次元オーダーを守らないルーティングを行なう時には，DR 番号を１つ増やし，転送を行なうクラスを１つ上げれば，デッドロックは解消できる．

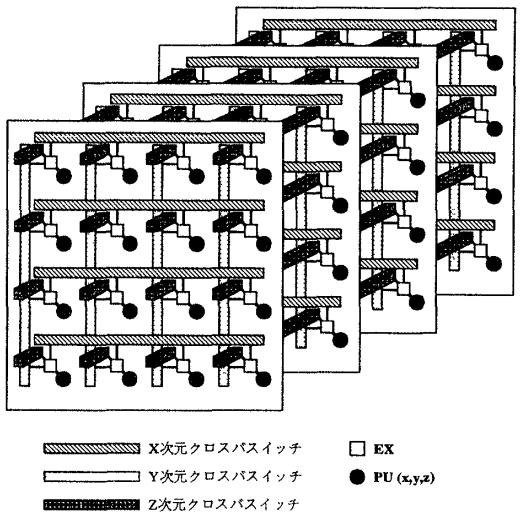

図 1: ３次元 HXB (4×4×4)

3.2 迂回ルーティングのアルゴリズム

DR 番号を用いてデッドロックを解消する手法は，適応ルーティングにも適用できる．そこで，余ったバーチャル・チャネルの有効利用も考慮に入れ，本研究では，２種類の迂回ルーティング・モデルを提案する．

<Fix>　次元オーダーの固定ルーティングに迂回ルーティングを組み合わせたモデル

<Adaptive>　適応ルーティングに迂回ルーティングを組み合わせたモデル

経路上に故障も混雑もない場合では次元オーダーの固定ルーティングを行なう．

混雑にぶつかると，<Fix>ではチャネルが空くのを待ち，<Adaptive> では適応ルーティングを行なう．

EX において故障にぶつかると，<Fix><Adaptive> 共に，まず，転送可能な他の最短経路を探し，無ければ迂回ルーティングを行なう．XB においては直ちに迂回ルーティングを行なう．XB での迂回先はランダムに決定する．ただし，迂回の際に，１ステップ前に通過した EX，XB には出力しないことにする．これは，EX については効率良く迂回を行なうためであり，XB については閉ループを作るのを避けるためである．

3.3 対処できない故障

本研究において提案する迂回ルーティング手法はどのような故障にも対処できるわけではない．本手法では対処できない故障として以下のものがある．

1. EX において１ステップ前に通過した XB 以外の

*Detour Routing Algorithm on Hyper-Crossbar Network
†Masaaki Tanabe, Tomoki Harada and Taisuke Boku
‡Institute of Information Sciences and Electronics, University of Tsukuba

次元方向への出力チャネルがすべて故障している場合

2. DR 番号が最大値に達した後, 経路上に故障がある場合

1 の条件は EX において集中して故障が発生したという状況であり, この状況をルーティング手法のみで解決するのは難しいと思われる. 従って, この条件により耐故障性が損なわれるとは考えにくい. また, 2 の条件は十分なバーチャル・チャネルを用意すれば対処できるが, 多くの故障を抱えたままでシステムの運用を続けるのは難しいと思われる. よって本手法はある程度妥当なものであると考えられる.

4. 計算機シミュレーション

4.1 ネットワーク・シミュレータ生成システム INSPIRE

本研究では, 計算機シミュレーションを用いて, ルーティング手法の転送性能評価を行なう. シミュレーションには本学で開発されたネットワーク・シミュレータ生成システム INSPIRE[3] を用いる.

4.2 評価モデル

<Fix> と <Adaptive> の 2 つのモデルに対して転送性能の評価を行なう.

ネットワークの規模は 4×4×4(64PU) の 3 次元 HXB とし, バーチャル・チャネルは 8 本とする. ネットワーク中の全チャネルのバンド幅は 1[flit/clock] とする. メッセージの転送方式には wormhole 方式を用いる. 故障は EX と XB 間の各出力チャネルにのみ存在するものとする. これらはネットワーク中に一様に分布しており, 固定的である. メッセージ長は 5[flit] と 20[flit] の 2 種類とし, ネットワークに最高に負荷をかけた状態で一様ランダム転送を行なう.

4.3 性能評価指標

本研究では, 性能評価指標としてスループットを用いる. スループットは PU がネットワークを通じて送受信できた総メッセージ転送量である. システム中の各 PU が毎クロック, 1[flit] 分のメッセージを受信できた場合を 1 として, 1PU あたりの平均メッセージ受信量をそれに対する比率で表す.

4.4 評価結果と考察

図 2 に故障率の上昇に対する各モデルのスループットの変化を示す.

故障率が上昇すると総じてスループットは低下している. しかしながら, <Adaptive> は <Fix> と比較して, メッセージ長が 5[flit] の場合も 20[flit] の場合も共に高い転送性能を維持していることが分かる. これは, 迂回するメッセージによって故障箇所近隣に混雑が引き起こされるが, その箇所を適応ルーティングによって回避し, メッセージの衝突を緩和しているためであると思われる.

また, メッセージ長が長くなるとメッセージの衝突によるネットワーク中での待ち時間が長くなるため, 適応ルーティングの効果が大きくなり, 5[flit] の場合より 20[flit] の場合の方が <Adaptive> と <Fix> の転送性能の差が広がっていると思われる.

しかし, 故障が存在しない条件の下では <Adaptive> が <Fix> とほぼ同程度となっている. これは, <Adaptive> で EX において先読みと予約 [2] を行なうことで XB で

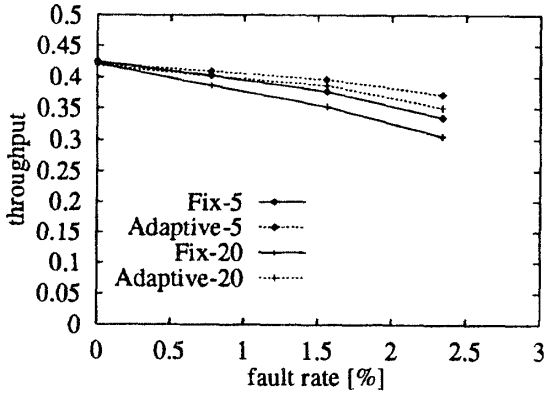

図 2: 故障発生時の性能

の衝突を無くしたにも関わらず, EX での衝突率が高くなっているためであるためと考えている.

一方, 耐故障性の限界は図中の端点 (2.34 %) が示している. しかし, 故障率がこの値以下であれば, どのような故障にも対処できることを指しているのではない. この値は故障の分布状況や転送パターンによって変化すると思われ, あくまで, 確率統計上の値である. しかし, 十分長いシミュレーションにより得られた結果であるから, 両モデルの耐故障性を示すには有効な指標である.

結果によると, 両モデル共, 耐故障性の限界は故障率約 2.3 % となっており, モデル間の差は現れていない. しかし, 両モデルにおける限界点での各メッセージの DR 番号 (0~7) の平均は, <Fix> では 0.24, <Adaptive> では 0.67 となっており, <Fix> の方が耐故障性の消耗が少ない. 従って, 同じ故障率で限界に達したとはいえ, <Fix> は <Adaptive> よりも耐故障性の余力を残していることが分かる.

以上の結果から, <Fix> ではより高い耐故障性を実現し, <Adaptive> では耐故障性を持たせるとともに転送性能の急激な低下を抑えることができている.

5. おわりに

本稿では HXB における迂回ルーティング手法を 2 種類提案し, 転送性能の評価を行なった. HXB において従来用いられてきた固定ルーティングでは, 全く耐故障性が無かったことを考えると, 高い耐故障性を持っていることが確認できた. また, 適応ルーティングと合わせて用いることで, 故障発生後のスループットの急激な低下を抑えることができることも示せた.

謝辞

本研究に関し, 貴重な御意見をいただいた筑波大学西川博昭助教授ならびに坂井修一助教授, アーキテクチャ研究室諸氏に深く感謝します.

参考文献

[1] William J. Dally and Hiromichi Aoki, "Deadlock-Free Adaptive Routing in Multicomputer Networks Using Virtual Channels", IEEE Trans. Parallel Distrib. Syst., vol. 4, No. 4, pp. 466-475, April 1993

[2] 曽根 猛 他, "ハイパクロスバ・ネットワークにおける virtual channel の動的選択による適応ルーティング", 並列処理シンポジウム JSPP'95 論文集, pp. 249-256, 1995

[3] 原田 智紀 他, "並列処理ネットワークのための性能評価用シミュレータ生成系 INSPIRE", 情報研報 ARC-113-9, pp. 65-72, 1995