4B－3

超並列 *Teraflops* マシン *TS/1*

～フォールトトレラントルーティング～[1)]

菅野 伸一　田邊 昇　鈴木 真樹　小柳 滋[2)]
RWCP[3)]超並列東芝研究室[4)]

1　はじめに

超並列計算機TS/1では、最大65536台のプロセッシングエレメント(PE)を変形3次元トーラスネットワークで接続する。

トーラスネットワークを採用した超並列計算機では、通信におけるレイテンシーが大きくなるので、小レイテンシーで高スループットなルーティング方式が求められている。また、PE数の大きな超並列計算機では、計算機全体を構成するPEのうちどれか一つでも故障している確率は一般の計算機よりも高く、かつ故障部分が大きくなければ補修をせずに動作をさせたいという要求がある。

本発表では、多次元トーラスネットワークにおける高スループットかつ故障にある程度耐え得るルーティングについて検討し、併せてTS/1におけるインプリメントについても説明する。

2　従来のルーティング方式

2.1　ワームホールルーティング

ワームホールルーティングでは、先頭の宛先情報にしたがって中継ルートを決定して伝送する。メッセージ伝送中はその中継ルートは占有され他のメッセージの伝送を行うことはできない。

ワームホールルーティングは、レイテンシーが小さいという特徴があるが、デッドロックが起こる可能性があるという問題がある。

Dally らは、仮想チャネルを用いてメッセージ伝送路においてチャネル依存関係グラフ上でのループを排除することによってデッドロックを回避する方式(文献[1])、ならびに仮想チャネルの多重化を行うことによってスループットを向上する方法(文献[2])の提案を行っている。仮想チャネルは物理的にはFIFOとして実装されるので、仮想チャネルを増やすことは、FIFOすなわちハードウェア量の増加となる。

2.2　故障回避ルーティング

Linder らは、仮想チャネルを用いて多次元トーラスネットワークにおいてワームホールルーティングを行う際に故障部分を回避してルーティングを行う方法を提案している(文献[3])。この方法では、故障部分の回避を行うために仮想チャネルを必要数だけ用意しなければならず、故障部分が無いときには全く利用されない仮想チャネルが存在する。このため、仮想チャネルの利用効率が悪くなり、結果として仮想チャネルの実装数が多くてもスループットの向上には寄与しないという問題がある。

3　提案する方法

ここでは、故障回避ルーティングを実現しつつスループットの向上を図るため、仮想ネットワークによるルーティング方式を提案する。

3.1　仮想ネットワーク

仮想ネットワークとは、図1のように複数の仮想チャネルをグルーピングしたものであり、それぞれの仮想ネットワークでは、文献[1]と同様な手法でデッドロックが起きないことを保証する。また、物理ネットワークの上に複数の仮想ネットワークを配置することが可能である。

前述した *Dally* や *Linder* による方式では仮想チャネルが、それぞれスループット向上あるいは故障回避のためという目的にのみ利用されていた。本方式では、静的あるいは動的に需要に応じて有限な物理資源である仮想チャネルを仮想ネットワークに配分できるようにしているため、仮想チャネルの利用効率が向上する。

[1)] Massively Parallel Teraflops Machine "TS/1", – Fault Tolerant Routing –

[2)] Shin-ichi KANNO, Noboru TANABE, Masaki SUZUKI, Shigeru OYANAGI

[3)] Real World Computing Partnership(新情報処理開発機構)

[4)] (株) 東芝 研究開発センター 内

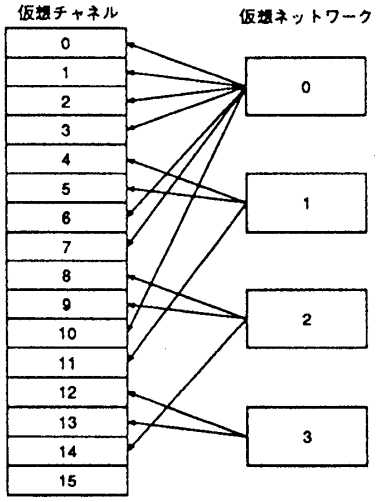

図 1: 仮想チャネルのマッピング

3.2 故障回避ルーティング

2次元ネットワークでの仮想ネットワークを用いた故障回避の方法を図2に示す。図では PE-A から PE-G にメッセージを伝送することを想定しているが、転送経路中に故障 PE や故障リンクがあるために、故障部分を回避しなければならない。

メッセージの中継処理において、故障回避のためにルート変更を必要があれば仮想ネットワーク番号が現在使用している仮想ネットワークより 1 だけ大きい仮想ネットワークを使用してメッセージを転送する。このようにすることにより、仮想ネットワーク間の依存関係ではループが構成されておらず、前述したように仮想ネットワーク内ではデッドロックが起きないことが保証されているので、本方式は故障回避を行ってもやはりデッドロックフリーである。

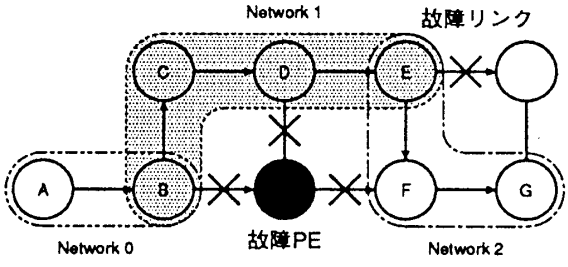

図 2: 仮想ネットワークによる故障回避

4 TS/1 のルーターの構成

TS/1 のルーターは、以下に述べるような機能を持つことを念頭に設計を行う。

1. 故障回避ルーティング

TS/1 では、これまで説明したような仮想ネットワークを使用したルーティング機構を組み込む。仮想チャネルの仮想ネットワークへの割り当ては動的に行う予定である。

仮想ネットワークを増加すれば複雑な故障パターンにも対応できるようになるが、仮想ネットッワーク情報はメッセージヘッダーで伝達する必要があるので、仮想ネットワークを増やすと他のメッセージヘッダー情報を圧迫することになり好ましくない。そこで *TS/1* では、クロスバースイッチを用いた故障回避機能を備えている (文献[4]) ことを考慮して、仮想ネットワークの数はそれ程大きくとらない方針である。

2. Wavefront Array への対応

Wavefront Array の高スループットを生かすため、文献 [4] で紹介されているヘッダー省略機構や短縮ヘッダーの取扱を行い、また、その高い性能を活かすために通信路 1 本あたり、500MByte/Sec の通信性能を持たせる。

3. クロスバーへのルーティング

TS/1 は、3 次元トーラスネットワークの他に 1 次元分にクロスバースイッチを装備している。このような構造に対応し、故障回避や非隣接 PE への通信行うときにクロスバースイッチを利用したルート設定を行うようなルーティング機構を組み込む。

5 まとめ

ワームホールルーティングにおいて、デッドロックを回避しつつスループットの向上と故障回避ルーティングを行うルーティング方式の提案を行い、*TS/1* のルーターチップのインプリメントについて説明を行った。

今後は、本ルーティング方式の故障回避ルート選択アルゴリズムの検討、故障回避性能、スループットの定量的検討を行う予定である。

参考文献

[1] W.J.Dally et al. : IEEE trans. COM-36, No. 5, pp. 547-553, MAY 1987.

[2] W.J.Dally : "Virtual-Channel Flow Control", Int'l. Symp. Computer Artitecture(ISCA'90)

[3] D.H.Linder et al. : IEEE trans. COM-40, No. 1, JAN 1991

[4] 田邊, 菅野, 鈴木, 小柳: 第 48 回情報処理学会全国大会,4B-02, MAR 1994