

超並列計算機における耐故障性を持つデッドロックフリールーティング

鯉 淵 道 紘[†]

数千ノード規模の超並列計算機を安定して運用するためには、結合網に対して耐故障性を持たせることが必要となってくる。本稿では、耐故障性を提供するために、スパニングツリーを基にした L-Turn ルーティングをこれらの結合網で用いられているメッシュおよびトーラストポロジに適用する手法を提案する。提案手法は、メッシュおよびトーラスにおいて West-First Turn モデルと同じ分散したデッドロックフリーな最短経路を取ることができる。一方で、提案手法は、仮想チャネルを追加することなく、West-First Turn モデルでは対処できないあらゆる故障箇所を迂回する経路を取ることができる。さらに、その経路群は West-First Turn モデルに近いので、トラフィックを分散させることができる。

A Fault-Tolerant Deadlock-Free Routing in Massively Parallel Computers

MICHIMIRO KOIBUCHI[†]

Interconnection networks are required to be equipped with fault-tolerance so that massively parallel computers with thousands of nodes work stably and safely. In order to provide dependability, this paper presents a methodology to establish efficient balanced paths of the L-Turn routing, which is based on a spanning-tree, on mesh and torus topologies used in these interconnection networks. The proposed method can make a same path set as that of the west-first turn model, which would be minimal and distributed, on mesh and torus. On the other hand, the proposed method can provide paths avoiding any fault, which the west-first turn model cannot avoid, without additional virtual channels. Then, since the path set is similar to that of the west-first turn model, the traffic will be well-distributed.

1. はじめに

数千ノード規模の超並列計算機を安定して運用するためには、結合網に対して耐故障性を持たせることが必要となってくる。結合網ではリンク故障とルータ故障への対処が必要になる。最も簡単な対処法は、冗長なハードウェアを持たせ、故障箇所を物理的に置きかえることである。しかし、冗長なハードウェアが必要となるため、そのコストが高くなる問題がある。

そこで、故障箇所を検出した後、そこを迂回する経路を設定することで既存のプロセス、新たに投入される計算プログラムを健全なノードに配送する方法が検討されてきた。これらの方法は、静的な回復モデルと動的な回復モデルのどちらかを基にしている。静的な回復モデルでは、故障が発生した場合、すべてのプロセスを一時停止させ、故障箇所を迂回する経路を設定した後、システムを再起動させる。この静的な回復モデルは、チェックポイントを用いて故障による影響を抑えることでより効果を発揮する。一方、動的な回復モデルでは、故障が発生してもシステムを停止することなく、故障箇所を迂回する経路を設定し直し、故障箇所を以後使用しないようにする。動的な回復法は、ネットワーク中の古い経路を使用しているパケットと新たな経路を使用しているパケット

が混在することになるため、パケット間のデッドロックフリーを保証することが難しい。また、トポロジを物理的に分割できる領域に区切り、故障箇所のある領域内のみにおいて静的な対処法を採用する中間的な方法も考えられる。

いずれのモデルにおいても、故障箇所を迂回するルーティングアルゴリズムが必要となる。既存の耐故障ルーティングの研究は、並列計算機で採用されている規則性を持つトーラスなどのトポロジの特徴を基に考えられた耐故障デッドロックフリールーティングと、任意のトポロジに適用することができるシステムエリアネットワーク (SAN) 向けのデッドロックフリールーティング¹⁾に分けられる。前者は、仮想チャネルやチャネルバッファを増やすことでデッドロックフリーを保証する方法が多い。一方で、後者は新たな仮想チャネルは必要ないが、メッシュやトーラスにおいてスループットが著しく落ちる問題点がある²⁾。その他に両者を組み合わせる方法も提案されている³⁾。一般的に結合網のコストパフォーマンスと耐故障性はトレードオフの関係にあるため、結合網の性能を落とさずに耐故障性を提供するデッドロックフリールーティングの開発は難しい。

超並列計算機の結合網はスイッチング技術としてワームホール方式もしくは非同期ワームホール方式を採用している。そのため、これらの結合網では、パケット間にデッドロックが発生しないことを保証するルーティングアルゴリズムが通常用いられる。

[†] 国立情報学研究所
National Institute of Informatics

本稿では、任意のトポロジに適用することができるデッドロックフリールーティングである L-Turn ルーティングをメッシュおよびトラスに適用する手法を提案する¹⁾。提案手法は、メッシュおよびトラスにおいて West-First Turn モデルと同じ分散した最短経路を取ることができる。一方で、提案手法は、仮想チャネルやチャネルバッファを追加することなく、West-First Turn モデルでは対処できないいかなる故障箇所をも迂回する経路を取ることができる。さらに、その経路群は West-First Turn モデルとほぼ同じであるため、トラフィックを分散させることが期待できる。

一般的に、メッシュなどの規則性のあるトポロジにおいて、次元順ルーティング⁴⁾などの単純なルーティングアルゴリズムは、ルーティングテーブルを用いずにハードウェアレベルで実装することが多い。このような場合にも本提案手法は West-First Turn モデルと同じ、もしくは近い経路群を提供することができるため、トンネリングなどの若干のフォーマットの改良⁵⁾などで対応できると考えられる。また、耐故障性を提供することを前提に結合網を実装した場合、柔軟に経路を設定するため、ルーティングテーブルを持たせたルータの設計も考えられる。L-Turn ルーティングは元々 $C \times N \mapsto C$ ルーティング機能により表現されるが、提案手法を用いた場合、West-First Turn モデルと同様、もしくは近い経路群を取る。よって、 $N \times N \mapsto C$ ルーティング機能のテーブル、もしくはそれに若干のリネーミング技術を組み合わせることで簡単に実装することができる。

以降、2章ではメッシュおよびトラスにおける L-Turn ルーティングの適用方法を示す。3章では、その実装方法を示し、4章では、結論と今後の課題を述べる。

2. L-Turn ルーティングを用いた経路設定

本章では、結合網において故障箇所を迂回する L-Turn ルーティングの適用方法を提案する。ルータ故障は接続しているすべてのリンクが故障している場合とみなすことができるため、以後リンク故障に対する経路設定に焦点をあてる。

2.1 West-First Turn モデル

メッシュやトラスにおいて、仮想チャネルの追加なしに耐故障性を提供する高性能デッドロックフリールーティングとして Turn モデル⁶⁾が提案されている。Turn モデルはチャネルバッファが論理的に循環構造を作ること避けることでデッドロックフリーを保証する。例えば、2次元メッシュにおいて Turn モデルを適用した例を図1に示す。2次元メッシュにおける循環構造は、図1.(a)に示す二種類になる。各循環の進路変更を1つずつ禁止し、デッドロックを防いだ場合を図1.(b)に示す。図中の点線の進路変更が禁止されている。このルーティング方法は、

先に西(-x)方向にパケットを送ることから west-first と呼ばれる。デッドロックを防ぐ切り方は1つではなく、たとえば図1.(c)に示す切り方(north-last)でも、デッドロックを防ぐことができる。

West-First Turn モデルは、図2のノード S_1 から D_1 への経路のように縦方向(y方向)のリンク故障に対して迂回する経路を設定することができる。しかし、ノード S_2 から D_2 への経路のように横方向(x方向)のリンク故障に対して迂回する経路を設定することができない。同様に North-Last Turn モデルでは縦方向(y方向)のリンク故障に対して迂回する経路を設定することができない。そのため、West-First Turn モデルでは高い耐故障性を提供することが難しい。

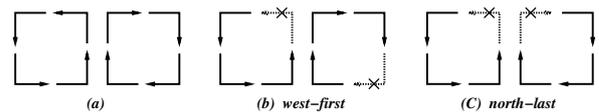


図1 Turn モデル

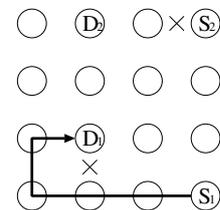


図2 West-first における故障箇所の迂回

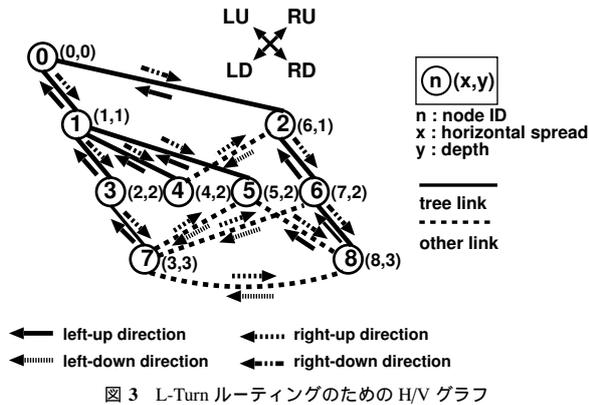
2.2 Left-Up First Turn ルーティング

われわれは、West-First Turn モデルを任意のトポロジにおいて適用することができるように拡張した Left-up First Turn (L-Turn) ルーティングを提案した¹⁾。L-Turn ルーティングは、図3のように任意のトポロジに適用するためにスパニングツリーを用いた2次元グラフ(H/Vグラフ)を基に禁止ターンを設定する。H/Vグラフは、結合網のリンクを辺、ルータを頂点としている。また、リンクは異なる向きの2つの単方向チャネルから構成されるものとする。

L-Turn ルーティングはメッシュやトラスのような規則網にも適用できるため、H/Vグラフを再構築することであらゆるリンク故障に対処することができる。よって、L-Turn ルーティングはメッシュやトラスにおいて高い耐故障性を提供することができる。

2.3 ツリー構成アルゴリズム

L-Turn ルーティングは Up*/Down* ルーティングと同様にスパニングツリーの構成方法により経路長、その分散具合が大きく影響される。しかし、メッシュやトラスにおいて任意のトポロジに適用することができる Breadth-First



Search スパニングツリーを用いた場合、トポロジに特化したルーティング⁴⁾並みの高い性能を達成することが難しい²⁾。

そこで、本稿では L-Turn ルーティングのために、メッシュやトラスに適した 2 次元グラフである H/V グラフの構築法を次のように提案する。

- (1) スパニングツリーを次のようにして構成する。
 - (a) ルートをノード 0 とする。
 - (b) ノード 0 から、+y 方向のリンクのみを辿り、ノードをツリーに追加する。
 - (c) ツリーに含まれる各ノードから +x 方向のリンクのみを探索し、ノードを追加する。
- (2) 2 次元座標を割当てるためツリーにおいて深さ優先探索を行う (Horizontal Spread の割当)。この探索において、複数のノードが選択可能であった場合、孫ノードの数が多き子ノードを優先する。
- (3) ルートから各ノードまでの最短距離である *depth* と、Horizontal Spread の値からノードの座標を求め、この座標から各チャンネルに方向を割当てる。

そして、この H/V グラフにおいて L-Turn ルーティングの禁止ターンを課す。4×4、および 6×6 メッシュにおいてこの手順で構築したチャンネルの向きを持つ H/V グラフとその上での L-Turn ルーティングの禁止ターンの例を図 4 に示す。

定理 1 2 次元メッシュにおいて提案手法により構築した L-Turn ルーティングは West-First Turn モデルと同じとなる。□

証明 $n \times n$ 2 次元メッシュにおいて +x, -x 方向のチャンネルは、提案手法により構築した H/V グラフでは、right-down(RD), left-up(LU) チャンネルとなる。また、図 4.(a) のノード 0, 4, 8, 12 のように -x 方向のチャンネルを持たないノードの +y, -y 方向のチャンネルは right-down(RD), left-up(LU) チャンネルとなる。その他の +y, -y 方向のチャンネルは left-down(LD), right-up(RU) チャンネルである。よって、RU チャンネルから LU チャンネルへの禁止ターンと LD チャンネルから LU チャンネルへの禁止ターンは West-First Turn

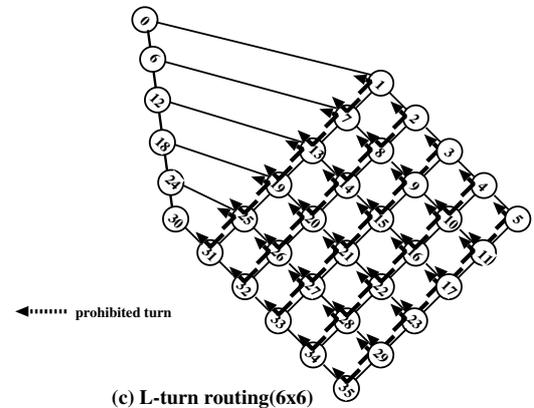
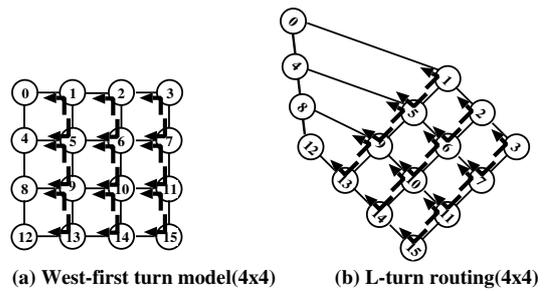


図 4 2 次元メッシュにおける West-First Turn モデルと L-Turn ルーティング

モデルの禁止ターンとなる。さらに、RU チャンネルから LD チャンネルへのターン、LD から RU へのターンが存在しないため、LU チャンネルを含まない循環は存在しない。よって、論文¹⁾ で定めた循環検出アルゴリズムによって禁止される可能性のある LD チャンネルから RD チャンネルへのターンはすべて許可される。従って、提案手法により構築した H/V グラフにおける L-Turn ルーティングと West-First Turn モデルは同じ禁止ターンとなる。□

2.4 リンク故障による 2 次元グラフの部分再構成

提案した H/V グラフは、リンクが故障した場合、全ノード対の経路を保証するために再構成を必要とすることがある。

H/V グラフはスパニングツリーを基に構築されているため、スパニングツリーを構成するリンクが故障しない限り L-Turn ルーティングの各ノード間の経路は保証される。そのため、スパニングツリーに属さないリンクが故障した場合、H/V グラフを再構成する必要はない。また、この場合、禁止ターン群は West-First Turn モデルと同じであるため、故障リンクを避けた経路群も同じとなる。

一方で、スパニングツリーに属しているリンクが故障した場合、経路を保証するためには H/V グラフを再構成する必要が生じる。例えば、図 4.(b) において、ノード 4, 5 の間のリンクが故障した場合、H/V グラフにおいてノード 5 からノード 0 への経路が消滅してしまう。しかし、BFS ツリーにより H/V グラフを完全に再構成した場合、経路が偏り性能が落ちてしまう²⁾。そこで、故障したリンクに

接続している2つのノードのうち、*depth* の値が大きい(ルートから遠い)ノードのスパニングツリーに属していないチャンネルの向きを変更する。この部分的にツリーを再構成する手法により、変更となる経路群を小さく抑える。

すべてのノード対の経路を保証するためには、スパニングツリーに属した上位ノード(親ノード)へのリンクが必要となるため、対象となるノードから少なくとも1つのLUチャンネルが必要となる。提案したH/Vグラフにおいて各ノードは、スパニングツリーに属していない高々1本のLDチャンネルおよび高々1本のRUチャンネル、親ノードへの1本のLUチャンネル、および子ノードへの高々1本のRDチャンネルにより構成される。親ノードへの1本のLUチャンネルが故障した場合、1)スパニングツリーに属していないチャンネルの向きを変更し、2)そのうちの1つがLU方向、かつ3)他のチャンネルの向きを変更させないという条件下では、再構成法は次の4通りとなる。

LUチャンネルとRUチャンネル

LU方向のチャンネルが故障し、LDチャンネルが存在する場合、LDチャンネルをLUチャンネルに変更する。例えば、図4.(b)において、ノード4,5の間のリンクが故障した場合、図5の部分再構成となる。なお、論文¹⁾に従い、対象とするノードのH/V座標を変更する方法を定めることで、結果的にチャンネルの向きを変更させることが正式な手順である。しかし、ここでは、簡略化のためチャンネルの向きを直接変更する方法を示す。

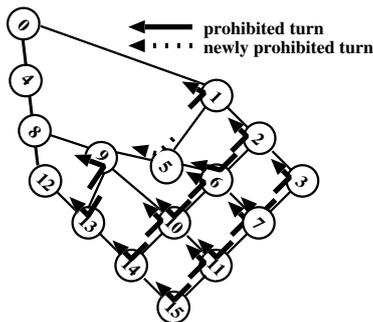


図5 LUチャンネルとRUチャンネルを用いた再構成の例

この場合、ノード9においてノード5からノード8への禁止ターンがなくなり、一方でノード5において、ノード1からノード9へのターンが新たに禁止される。

LUチャンネルとRDチャンネル

この再構成法ではスパニングツリーに属さないLDチャンネルをRUチャンネルに、RUチャンネルをRDチャンネルに変更する。例えば、図4.(b)において、ノード4,5の間のリンクが故障した場合、図6の部分再構成となる。

この場合、ノード9においてノード5からノード8への禁止ターンがなくなり、一方でノード1において、ノ

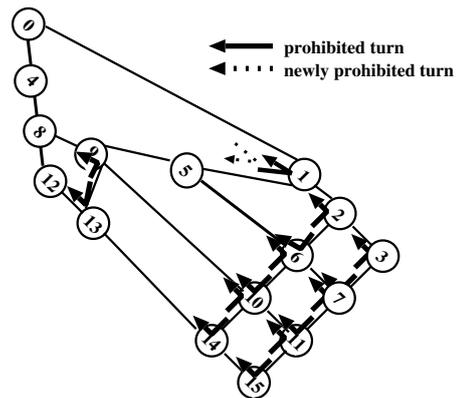


図6 LUチャンネルとRDチャンネルを用いた再構成の例

ド0からノード5へのターンが新たに禁止される。

2つのLUチャンネル

スパニングツリーの属さないLDチャンネルをLUチャンネルに、RUチャンネルをLUチャンネルに変更する。例えば、図4.(b)において、ノード4,5の間のリンクが故障した場合、図7の部分再構成となる。

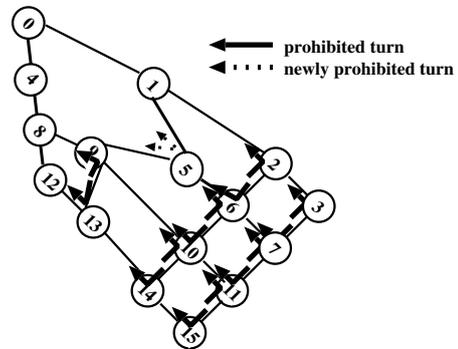


図7 2つのLUチャンネルを用いた再構成の例

この場合、ノード9においてノード5からノード8への禁止ターンとノード1におけるノード5からノード0への禁止ターンがなくなる。一方でノード5において、ノード1とノード9の間の2つのターンが新たに禁止される。

LDチャンネルとLUチャンネル

RUチャンネルをLUチャンネルに変更する。例えば、図4.(b)において、ノード4,5の間のリンクが故障した場合、図8の部分再構成となる。

この場合、ノード1においてノード5からノード0への禁止ターンがなくなり、一方でノード5において、ノード9からノード1へのターンが禁止される。

なお、図4.(b)のノード1,2,3,13,14,15のようにスパニングツリーに属していないチャンネルが1本の場合、それをLUチャンネルとした再構成を取る必要があるため、選択肢は1つとなる。また、スパニングツリーに属してい

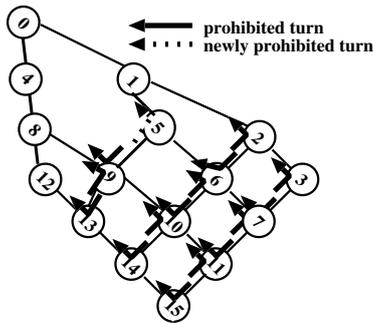


図 8 LD チャネルと LU チャネルを用いた再構成の例

ないリンクが 0 本の場合、子ノードへのチャンネルを LU チャネルとする再構成のみが可能となる。

定理 2 4 つのいずれの部分再構成においても、L-Turn ルーティングはすべてのノード対の経路を保证する。□
 証明 この部分再構成により、H/V ツリーが構成される。よって、論文¹⁾の定理 2 により、任意のノード間の経路が保証される。□

すべての部分再構成において、West-First Turn モデルで禁止されている 1 つもしくは 2 つのターンが許可される一方、1 つもしくは 2 つのターンが新たに禁止される。これは、West-First Turn モデルでは迂回できない故障リンクを L-Turn ルーティングにおいて迂回する経路を設定するためである。しかし、そのため若干 West-First Turn モデルと経路が異なる場合が生じる。

2.5 2次元トラスへの拡張

これまでの議論は、2次元メッシュを基に行ってきた。しかし、2次元トラスに拡張することも可能である。2次元トラスは、wraparound チャンネルが存在するため、Turn モデルを用いる場合、2本の仮想チャンネルが必要になる。

L-Turn ルーティングも 2本の仮想チャンネルを用いることで West-First Turn モデルとほぼ同様の経路をとることができる。

そのための拡張を次に示す。

- (1) wraparound チャンネルをスパニングツリーの構成外リンクとして、前節において提案した手順で H/V グラフを構築する。例えば、6×6 2次元トラスに適用した例を図 9 に示す。
- (2) 図 9 に示したように L-Turn ルーティングをこの H/V グラフに適用する。この場合、メッシュの場合と違い、論文¹⁾の循環除去アルゴリズムにより、LD チャネルと RU チャネル間のターンが新たに禁止される。
- (3) 禁止ターンを高々 1 個含んだ各ノード間の最短経路を探索する。ただし、禁止ターンを含む場合、そのノードにおいて使用する仮想チャンネル番号を +1 増加させる。また、複数の最短経路が存在する場

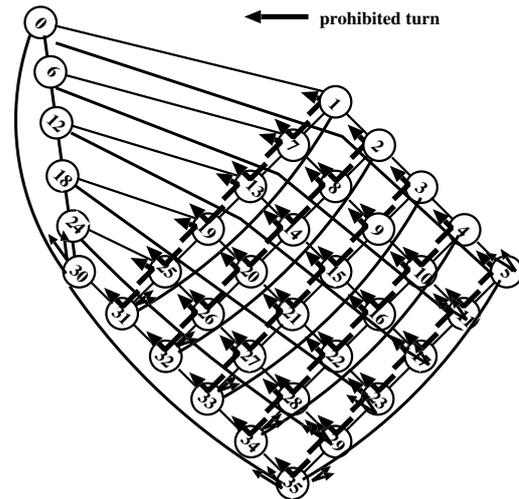


図 9 36 ノードのトラスにおける L-Turn ルーティング

合、禁止ターンを含まない経路のみを選択する。

仮想チャンネル番号を増加させることで禁止ターン上のチャンネルを利用する方法は、次元逆順ルーティングにおいて採用されており、それをホップ数の削減に利用するアイデアは DL ルーティングで採用されている。よって、仮想チャンネル番号の切り換えによりデッドロックフリーが保証できることを示す証明はここでは略す。

トラスにおいて L-Turn ルーティングでは、wraparound チャンネルにおいて禁止ターンが集中する。よって、L-Turn ルーティングは、wraparound チャンネルを 2 回使用する経路の選択肢が、West-First Turn モデルに比べて若干減る。例えば、図 9 においてノード 28 からノード 7 へは West-First Turn モデルのみノード 28, 34, 4, 10, 10, 14, 6, 7 を経由する経路が選択可能である。しかし、この点以外は提案した H/V グラフの構築法を用いることにより West-First Turn モデルと同じ物理経路を取ることができる。また、リンク故障に対する H/V グラフの再構成については、メッシュの場合と同様に行うことができる。

3. 実装方法

3.1 固定型ルーティング

超並列計算機の結合網におけるルーティングアルゴリズムは、各ノード対にただ 1 つの経路を用いる固定ルーティングと動的に経路を選択することができる適応型ルーティングの 2 つに分けられる。前者に比べ、後者はリンクバンド幅を活用する技術として幅広く研究が行われている。しかし、適応型ルーティングを導入することにより、次の新たな 2 つの問題がネットワークに発生するためすべての結合網が適応型ルーティングを採用しているわけではない; 1) In-Order パケット配送を保証していない; 2)

複数経路の中から1つの経路を選択する機能を付加することにより、スイッチの複雑さが増す。

そこで、本節では、提案した耐故障ルーティングを固定ルーティングとして実装する方法を示す。これまで、メッシュやトラスにおいて最短経路を取る高性能ルーティングとして次元順ルーティング⁴⁾が提案されている。West-First Turn モデルおよび提案手法にしたがって経路を設定したL-Turn ルーティングは次元順ルーティングの経路を含む。そのため、基本的に次元順ルーティングと同じ経路を選択するものとする。ただし、リンク故障による部分再構成により、その経路が取れない場合、次の方針で経路を設定する；中間ルータにおいて次元順ルーティングの経路の出力方向が選択できない場合、選択可能な最短経路上の出力方向の中から、+x,-x,+y,-y 方向の順で選択する。これにより、故障箇所を迂回する経路以外、ほとんどすべての経路は次元順ルーティングと同じ経路を取ることができる。

3.2 適応型ルーティング

Turn モデルのみを適応型ルーティングとして採用している結合網の場合、2章で提案した手法によりL-Turn ルーティングをそのまま使用すれば良い。

しかし、Duato のプロトコルなどのチャネル依存グラフ (CDG) においてチャネル間の循環を許す自由度の高い適応型ルーティングを採用している結合網も実装されている。この場合、最短経路に限定した West-First Turn モデルや次元順ルーティングを escape path として用い、fully adaptive path と併用して用いられる。これらは1つの物理チャネルを複数の仮想チャネルに分割し、各仮想チャネルを escape path と fully adaptive path に割当ててことで実装する。そして、2次元メッシュやトラスにおいて、最短経路を用いることにより、escape path と fully adaptive path を動的に切り換えてもデッドロックフリーを保証することができる。

しかし、リンク故障が発生した場合、メッシュやトラスにおいて故障前と同じホップ長の経路がとれなくなる。そのため、この場合、fully adaptive path から escape path への変更は許されるが、逆は禁止しなければデッドロックフリーは保証できない。よって、提案手法によるL-Turn ルーティングを用いた場合、仮想チャネル間の切り替えはこれに従う必要がある。これは、不規則なトポロジにおける escape path を用いた Silla と Duato の適応型ルーティングの考え方と同じである。

3.3 ルーティングテーブルとパケットフォーマット

超並列計算機の相互結合網において、柔軟な経路設定を可能にするためにルーティングテーブルを持たせてある場合、提案手法の実装は簡単である。この場合、L-Turn ルーティングは元々 $C \times N \mapsto C$ ルーティング機能を必要とするが、提案手法を用いた経路群は、West-First Turn モ

デルと同様に $N \times N \mapsto C$ ルーティング機能のテーブル、もしくは若干のレネーミング技術を組み合わせることで実装することができる。一方で、メッシュやトラスなどの規則的なトポロジにおいて、次元順ルーティングなどの単純な手法に限定した場合、ルーティングテーブルを用いずにハードウェアで実装する方法も考えられる。この場合、迂回経路など一部の経路は複数の中継ノードを目的地としてパケットに埋め込み、各中継ノードに到着した時点でその情報を削除するトンネリング⁵⁾を用いる等で実装することができる。

4. む す び

数千ノード規模の超並列計算機を安定して運用するためには、結合網に対して耐故障性を持たせることが必要となってくる。本稿では、耐故障性を提供するために、スパニングツリーを基にしたL-Turn ルーティングをメッシュおよびトラスに適用する手法を提案した。提案手法は次の4つの特徴を持つ。

- 故障がない場合、既存のデッドロックフリールーティングアルゴリズム (次元順および West-First Turn モデル) と同じ物理経路を取ることができる。
- 仮想チャネルおよびチャネルバッファの追加なしにあらゆる故障箇所を迂回することができる。
- リンクの故障箇所によらず、分散した経路群を採用することができるため、高スループットが期待できる。
- 実装が容易である。

今後は、適応型ルーティングもしくは固定ルーティングとしてシミュレーション、実機の評価を通して提案手法の有効性を定量的に示す予定である。

参 考 文 献

- 1) 上樂 明也, 鯉淵 道紘, 天野 英晴: 2次元 Turn モデルに基づくイレギュラーネットワーク向けルーティングアルゴリズムの設計と評価, 情報処理学会論文誌コンピューティングシステム, Vol. 44, No. SIG11(ACS 3), pp. 157-168 (2003).
- 2) Silla, F. and Duato, J.: Is it Worth the Flexibility Provided by Irregular Topologies in Networks of Workstations?, *Proceedings of Workshop Comm. and Architectural Support for Network-Based Parallel Computing*, pp. 47-61 (1999).
- 3) 吉永 努, 細越 洋行, 曾和 将容: 耐故障性を考慮した k-ary n-cube 用適応デッドロック 回復ルーティング, 情報処理学会論文誌コンピューティングシステム, Vol. 45, No. SIG11 (2004).
- 4) Dally, W. and Seitz, C.: Deadlock-Free Message Routing in Multiprocessor Interconnection Networks, *IEEE Transaction on Computers*, Vol. 36, No. 5, pp. 547-553 (1987).
- 5) 松谷 宏紀, 鯉淵 道紘, 山田 裕, 天野 英晴: チップ内ネットワークにおけるリンク故障時の経路設定方法, 電子情報通信学会技術研究報告 FIIS-05-156 (2005).
- 6) Glass, C.J. and Ni, L.M.: The Turn Model for Adaptive Routing, *Proceedings of International Symposium on Computer Architecture*, pp. 278-287 (1992).