イレギュラーネットワークにおける
適応型ルーティングに関する研究

平成18年度

上 樂　明 也

論文要旨

パーソナルコンピュータ (PC) を，Myrinet などの高速なシステムエリアネットワーク (SAN) で相互結合することにより構築される PC クラスタは，近年の半導体製造技術の進歩に伴なう PC の飛躍的な性能向上と低価格化により，コストパフォーマンスに優れた高性能並列分散コンピューティング環境として急速に普及が進んでいる．

SAN は，point-to-point リンクにより相互結合された高速スイッチ群から構成され，従来の大規模並列計算機と同様に，バーチャルカットスルー方式 (VCT 方式) またはワームホール方式 (WH 方式) によるパケット転送とデッドロックフリールーティングを用いることにより，低レイテンシかつ高バンド幅の通信を実現している．一方で，SAN は，拡張性および耐故障性の要求を満たすために，大規模並列計算機とは異なり，トポロジとして結合方式に制限がないイレギュラーネットワークをサポートすることが多い．しかし，イレギュラーネットワークでは，経路保証とデッドロックフリーの実現のための制約が厳しいため，大規模並列計算機で用いられるメッシュやトーラスなどのレギュラーネットワークに比べて，ルーティングアルゴリズムの設計が難しい．このため，既存のルーティングアルゴリズムの多くは，トポロジ上にスパニングツリーのマッピングを行ない，ツリー構造の接続性と非循環の性質を利用して経路保証とデッドロックフリーを実現する方式を採用している．

Up*/Down* ルーティングは，スパニングツリーベースの代表的な適応型ルーティングアルゴリズムであり，仮想チャネルやバッファ等のハードウェアを付加することなしに適用可能であることから，Autonet および Myrinet などの SAN において利用されている．Up*/Down* ルーティングは，マッピングしたスパニングツリーの階層構造に基づいて，ネットワークを 1 次元の方向 (up/down) を持つ有向グラフに見立て，最も単純な 1 次元の Turn モデルを適用することにより，デッドロックフリールーティングを実現している．しかし，1 次元の Turn モデルでは，トポロジの不規則性に基づく禁止ターン分布の偏りが大きくなるため，トラフィックの分散が困難となる．このため，Up*/Down* ルーティングは，効率的にネットワークのバンド幅を利用することが難しい．

本論文では，Up*/Down* ルーティングにおける上記の問題を解決するため，Up*/Down* ルーティングにおける 1 次元 Turn モデルを拡張した，2 次元 Turn モデルに基づく適応型ルーティングアルゴリズムである Left-up first turn (L-turn) ルーティングおよび Right-down last turn (R-turn) ルーティングを提案し，フリットレベルの相互結合網シミュレータを C++ 言語で実装して，確率モデルシミュレーションにより，評価を行なった．L-turn ルーティングおよび R-turn ルーティングは，1 次元有向グラフを拡張した H/V グラフと呼ばれる 2 次元有向グラフを導入し，ターンのパターン数および禁止ターン選択における自由度の増加を実現する．そして，H/V グラフに対して，2 次元 Turn モデルの適用によるシステマティックな手法を用いて，より均等な禁止ターンの分散を実現し，トラフィック分散とスループット向上の実現を図る．

L-turn ルーティングおよび R-turn ルーティングは，Up*/Down* ルーティングと同様に，付加的なハードウェアに依存しないため，任意のトポロジの SAN に適用可能な汎用性の高い手法である．

確率モデルシミュレーションの結果，L-turn ルーティングは，最も高いスループットを達成し，Up*/Down* ルーティングに対し最大で約 80% のスループット向上を実現することが確認された．一方，R-turn ルーティングは，全体的に，最も低いスループットを示すことが確認された．L-turn ルーティングと R-turn ルーティングは，ほぼ同等の禁止ターン分散を実現するが，選択された禁止ターンのパターンの違いにより，L-turn ルーティングでは，ツリーの葉方向にトラフィックが分散されやすくなるのに対し，R-turn ルーティングではホットスポットが発生しやすいルート方向にトラフィックが集中してしまうことがわかった．これより，スループット向上のためには，より均等な禁止ターンの分散と葉方向へのトラフィック分散の両立が重要であり，この条件を満たす L-turn ルーティングが，優れたルーティングアルゴリズムであることがわかった．

Abstract

Network-based parallel processing using clusters of personal computers (PCs), which are interconnected by system area networks (SANs), has been researched as potential cost-effective parallel-computing environments.

SANs consist of switches connected with point-to-point links, uses wormhole routing (WH) or virtual cut-through switching (VCT) as its switching technique to provide low-latency and high-bandwidth communications like those of interconnection networks in massively parallel computers. Unlike the interconnection networks used in massively parallel computers, SANs usually accept arbitrary topologies so as to provide extensibility and dependability to cope with low-reliability commodity PCs. The interconnection adaptivity, however, makes it difficult to establish paths that are free of deadlocks. A deadlock-free routing algorithm is thus crucial for making efficient use of network resources, yet the current deadlock-free routing algorithms in massively parallel computers with regular topologies cannot be directly employed in most cases. Thus, traditional routing algorithms usually employ connectivity and acyclicity of mapped spanning-tree to ensure deadlock-freedom and connectivity in their target topologies.

Up*/Down* routing is the most popular spanning-tree based deadlock-free adaptive routing algorithm. Up*/Down* routing guarantees deadlock-freedom by using one-dimensional (up/down) turn model approach, and does not require additional hardware such as virtual channels or buffers. However, Up*/Down* routing tends to generates unbalanced traffic because the one-dimensional turn model always generate unbalanced prohibited turns, and thus it leads to poor throughput.

This thesis introduces the systematic approach for designing deadlock-free adaptive routing algorithms called "left-up first turn (L-turn) routings" and "right-down last turn (R-turn) routings". The L-turn routings and the R-turn routings are based on a two-dimensional turn model to guarantee deadlock-freedom and make the paths as uniformly distributed as possible by selecting well-distributed prohibited turns. The extended two-dimensional directed graph, called H/V graph, provides the extra degree of freedom for the selection of prohibited turns. The L-turn routings and the R-turn routings can be applied to any networks in which Up*/Down* routing is used because they do not require additional hardware.

A flit-level interconnection network simulator written in C++ was used for evaluating Up*/Down* routing and the proposed routings by probabilistic simulation model. Results of simulations show that the L-turn routings achieved the highest and up to 80% improvement on throughput. On the other hand, the throughputs of R-turn routings were the lowest. Although the L-turn routings and the R-turn routings achieve almost the same degree

of uniformity about the distribution of prohibited turns, there is difference in an approximate tendency in which the traffic is more likely to be distributed. The traffic would be distributed toward the leaf nodes when the L-turn routings are employed, and thus it leads to better traffic balance and throughput. However, the traffic would be distributed toward the root node when the R-turn routings are employed, and thus it leads to unbalanced traffic and poor throughput. Therefore, better throughput results from uniformly distributing the prohibited turns by which the traffic would be more distributed toward the leaf nodes, and the L-turn routings meet this condition.

目 次

図 目 次

表 目 次

第1章　緒 論

近年の半導体製造技術の進歩に伴ない，パーソナルコンピュータ (PC) およびワークステーション (WS) の飛躍的な性能向上と低価格化が続いている．これに伴ない，安価な PC を 高速なシステムエリアネットワーク (SAN) [Mae91, RS91, N.J95, Myra, PFH01, I.T04, TSJ+99, 西 宏 00, STH+00, NKN+01] で相互結合することにより構築される PC クラスタは，大規模並列計算機に代わるコストパフォーマンスに優れた高性能並列分散コンピューティング環境として急速に普及が進んでいる．現在，世界の上位 500 のスーパーコンピュータ [TOP] の中でも，PC クラスタが占める割合は年々増加を続けており，当面この傾向が続くものと考えられる．

PC クラスタの性能に影響を与える構成要素の一つである SAN は，高速な point-to-point リンクにより相互結合された高速スイッチ群から構成される高性能ネットワークであり，バーチャルカットスルー方式 (VCT 方式) [KK79] またはワームホール方式 (WH 方式)[DS87] などの，従来の大規模並列計算機向けのアーキテクチャに基づくパケット転送とデッドロックフリールーティングを行なうことにより，低レイテンシ，高バンド幅，高信頼性の通信を実現している．このため，大規模並列計算機のネットワークと同様に，ルーティングアルゴリズムが SAN の構成および性能に影響を与える重要な要素となる．

SAN におけるルーティングアルゴリズムは，大規模並列計算機のネットワークと同様に，固定型ルーティングと適応型ルーティングに分類される．固定型ルーティングは，出発地スイッチから目的地スイッチまで，常に同じ経路を用いてパケット転送を行なう手法である．固定型ルーティングは，既存の高速スイッチの多くで採用されており，次の利点を持つ．

- シンプルかつ実装が容易
- パケットが送信順に必ず到達する性質 (FIFO 性) を持つ

一方で，パケット転送中に代替経路を利用することができないため，次の欠点を持つ．

- ネットワークの資源を効率的に利用できない場合がある
- 故障箇所をその場で迂回することができない (耐故障性を持たない)

適応型ルーティングは，出発地スイッチから目的地スイッチまでに複数の経路を用意し，状況に応じて，動的に経路を選択してパケット転送を行なうことができる手法である．このため，適応型ルーティングは次の利点を持つ．

- 混雑箇所を回避し，ネットワーク資源を効率よく利用することが可能
- 故障箇所の動的な迂回による耐故障性の実現が可能

一方で，ルーティングにおける経路選択などの処理が複雑になるため，次の欠点を持つ．

- 実装が複雑
- FIFO 性の保証が困難

SAN では，各スイッチが point-to-point リンクにより相互接続されているため，パケットが出発地スイッチから目的地スイッチに到達するまでの過程において，途中経路にある複数のスイッチを経由する必要がある．SAN において，パケットがルーティングアルゴリズムによって決定される移動先隣接スイッチについての情報を取得するための手法は，ソースルーティング方式と分散ルーティング方式に分類される．ソースルーティング方式では，出発地スイッチにおいて，ルーティングアルゴリズムに基づいて目的地スイッチに到達するまでの全経路情報が計算され，パケットのヘッダに格納される．パケットのヘッダがスイッチに到達するたびに，ヘッダに格納された経路情報に従って，次の移動先スイッチの決定と転送を行ない，これを目的地スイッチに到達するまで繰り返す．ソースルーティング方式では，各経由スイッチにおける経路決定処理が簡素化されるため，その分スイッチの実装コストを抑えられるという利点を持つ [Myra]．しかし，全経路情報を格納する分，ヘッダ長のオーバヘッドが大きくなる．また，目的地スイッチまでの経路が出発地スイッチで決定されるため，適応型ルーティングを用いる場合，経路選択は，出発地スイッチにおいてだけ可能となり，途中経路において動的に経路を選択することができない (この場合，固定型ルーティングとなる) という制限を持つ．分散ルーティング方式では，パケットが各スイッチに到達するたびに，スイッチに実装された経路計算用ハードウェアにより，ルーティングアルゴリズムに基づいた移動先隣接スイッチが決定される．一般的には，スイッチに用意されたルーティングテーブルを参照する table-lookup 方式が用いられる [Mae91, I.T04]．分散ルーティング方式では，中間スイッチにおいて，動的に経路を選択することが可能であるため，適応型ルーティングの長所をフルに活用することができる．また，ヘッダに格納される経路情報が小さくて済むという利点も持つ．しかし，スイッチの実装がより複雑となるため，ソースルーティング方式に比べて，実装コストの面では劣る．

PC クラスタでは，拡張性，耐故障性および可用性が重視される．このため，SAN は，トポロジとして結合方式に制限がないイレギュラーネットワークをサポートすることが多い．近年，大規模並列計算機に匹敵する高性能を重視する SAN のトポロジとしては，大量のポートを持つスイッチと大量の中間リンクにより構成される Fat ツリーなどのレギュラーネットワークを用いる場合が多い [Myra, PFH01, FFA+02]．しかし，少量のポートを持つスイッチから構成される低コストの構成を取る場合や，クラスタ間同士を相互接続する場合などのケースにおいては，依然としてイレギュラーネットワークが必要となると考えられる．また，専用のクラスタではなく，高速なネットワークを用いて机上に配置された PC や WS を接続し，専用のクラスタシステムと同様の性能を実現することを目的としたシステムも提案されており [TSJ+99, 西 宏 00, STH+00, NKN+01]，このような場合は，物理的な配置の制約からトポロジとしてイレギュラーネットワークが必須となる．

イレギュラーネットワークでは，経路保証とデッドロックフリーの実現のための制約が厳しいため，大規模並列計算機で用いられるメッシュやトーラスなどのレギュラーネットワークに比べて，ルーティングアルゴリズムの設計が難しい．このため，既存のルーティン

グアルゴリズムの多くは，トポロジ上にスパニングツリーのマッピングを行ない，ツリー構造の接続性と非循環の性質を利用して経路保証とデッドロックフリーを実現する方式を採用している．

Up*/Down* ルーティング [Mae91][1]は，スパニングツリーベースの代表的な適応型ルーティングアルゴリズムであり，

- スイッチまたは PC に，仮想チャネルや専用バッファ等のハードウェアを付加することなしに適用可能
- 固定型ルーティングとしての利用も可能

などの理由から高い汎用性を持ち，Autonet および Myrinet などのネットワークにおいて利用されている．Up*/Down* ルーティングは，マッピングした breadth first search (BFS) スパニングツリーの階層構造に基づいて，ネットワークを 1 次元の方向 (up/down) を持つ有向グラフに見立て，最も単純な 1 次元の Turn モデルを適用することによりデッドロックフリールーティングを実現している．しかし，1 次元の Turn モデルでは，トポロジの不規則性に基づく禁止ターン分布の偏りが大きくなるため，非最短経路の割合が増加し，トラフィックの分散が困難となる．このため，Up*/Down* ルーティングは，効率的にネットワークのバンド幅を利用することが難しい．Up*/Down* ルーティングの改良案として，ヒューリスティックルールに基づいた depth first search (DFS) スパニングツリーを利用して，禁止ターンの削減を図る手法 [JAJ00, JA00] も提案されているが，Up*/Down* ルーティングが本質的に抱える禁止ターン分布の偏りに関する問題が残ったままとなっている．

Up*/Down* ルーティングにおける問題を改善し，トラフィックの分散を実現するために提案された既存のルーティングアルゴリズムは，大きく次の 2 つに分類される．

(a) 仮想チャネルおよびスパニングツリーに依存しない手法 [LVT96, QNR99]

(b) 仮想チャネルまたは専用バッファを利用する手法 [SD00, FJ00, JPMJ02, JMPJ02, SLT02, JPJ$^+$02, MAH03]

しかし，これらの手法は，汎用性に欠けるという問題を持つ．前者の手法は，経路計算の計算量が現実的でない [LVT96]，または，適用可能なトポロジが限定される [QNR99]，という問題を持つ．また，後者の手法は，スイッチの付加的なハードウェアの利用を前提としているため，適用可能なネットワークが限定されるという問題を持つ．このため，これらの手法は，高い汎用性を持つ Up*/Down* ルーティングの代替手法として常に選択できるわけではない．

そこで，本研究では，Up*/Down* ルーティングと同等の汎用性と Up*/Down* ルーティングが抱える問題の改善を実現するために， Up*/Down*ルーティングにおける 1 次元 Turn モデルを拡張した 2 次元 Turn モデルに基づく適応型ルーティングアルゴリズムである left-up first turn (L-turn) ルーティング および right-down last turn (R-turn) ルーティングを提案する [MAAH01, AMAH02, 上樂 03]．L-turn ルーティングおよび R-turn

[1]パケットが up 方向に必要ホップ数移動した後，down 方向に移動するため，名称に*という表現を用いている．

ルーティングは，既存の 1 次元 (垂直方向だけ) の有向グラフを拡張した 2 次元 (垂直方向と水平方向) の有向グラフである H/V グラフを用いる．H/V グラフでは，各チャネルに割当てる論理方向が 2 つから 4 つに増加するため，スイッチを通過するパケットの方向転換 (ターン) のパターンが従来の 2 個から 6 倍となる 12 個に増加する．これにより，禁止ターン選択の自由度が増加し，より均等な禁止ターン分散の実現が可能となる．H/V グラフに対して 2 次元 Turn モデルをシステマティックに適用することにより，分散を考慮した禁止ターンの集合を持つ L-turn ルーティングおよび R-turn ルーティングの定義を行ない，トラフィック分散およびバンド幅の向上を図る．L-turn ルーティングおよび R-turn ルーティングは，仮想チャネルや専用バッファ等の付加的なハードウェアを必要としないため，Up*/Down* ルーティングと同等の高い汎用性を持つ．これにより，容易に Up*/Down* ルーティングの代替手法として適用可能な現実性の高い手法となっている．

本論文の構成は次の通りである．

第 2 章では，本研究の基礎となった SAN におけるトポロジとルーティングアルゴリズムに関する基本要素について説明し，第 3 章でイレギュラーネットワーク向けの既存のルーティングアルゴリズムにおける問題点と既存の SAN の実現例について述べる．第 4 章では，2 次元 Turn モデルの適用に基づく適応型ルーティングアルゴリズムである L-turn ルーティングと R-turn ルーティングを提案し，第 5 章にてフリットレベル相互結合網シミュレータを用いた確率モデルシミュレーションにより評価を行った結果を示す．最後に第 6 章にて結論を述べる．

本研究で取り組んだ研究の要約を表 1.1 に示す．

表 1.1: 本研究の要約

従来技術の問題点	Up*/Down* ルーティングが用いられるが，トラフィックが偏るため，バンド幅の有効利用ができない．
目的	Up*/Down*ルーティングと同等の高い汎用性とトラフィック分散によるスループット向上を実現する適応型ルーティングの提案
提案技術	2 次元 Turn モデルに基づいてパケットの禁止ターン分布を分散させる L-turn ルーティングと R-turn ルーティング
効果	スループットの向上

第2章　SAN のトポロジと ルーティングアルゴリズム

SAN は，性能面および機能面に関する次の特徴を備えた PC クラスタ向けの高性能ネットワークとして定義される．

- 高バンド幅，低レイテンシ
  - 高速なリンクと高速なスイッチを用い，PC とスイッチおよびスイッチ間を point-to-point に接続
  - WH 方式または VCT 方式による低レイテンシのパケット転送
  - インテリジェントなネットワークインタフェースによるプロトコル処理オーバヘッドの低減
- 高信頼性通信
  - デッドロックフリールーティングによりパケット廃棄を回避
  - 極めて低いエラーレート
- 高拡張性
  - 数十から数千の PC を接続可能
  - トポロジとして，レギュラーネットワークまたはイレギュラーネットワークをサポート

本章では，本研究の基礎となる SAN におけるトポロジとルーティングアルゴリズムについて述べる．まず，第 2.1 節で，SAN で用いられるトポロジであるイレギュラーネットワークとレギュラーネットワークについて述べる．そして，第 2.2 節で，SAN におけるルーティングアルゴリズムについて，パケット転送に関連する基本的な要素と共に述べる．

2.1　トポロジ

2.1.1　ネットワークモデル

SAN は，図 2.1 のような point-to-point リンク (各リンクは互いに反対の方向に向かう 2 つの単方向物理チャネル (双方向チャネル) から成る) により相互結合されたスイッチ群により構成される．各スイッチは複数のポートを持ち，各ポートは他のスイッチまたは PC との接続に用いられる．

SAN におけるルーティングアルゴリズムの設計は，各スイッチ間のパケット転送経路を対象とするため，通常，SAN はグラフ $G(N, C)$ で表される [JSL02]．ここで，N はグラフのノード (スイッチ) の集合，C はグラフのエッジ (各スイッチを相互結合するリンク) の集合をそれぞれ表す．例えば，図 2.1 の SAN は，図 2.2 に示すグラフで表すことができる．

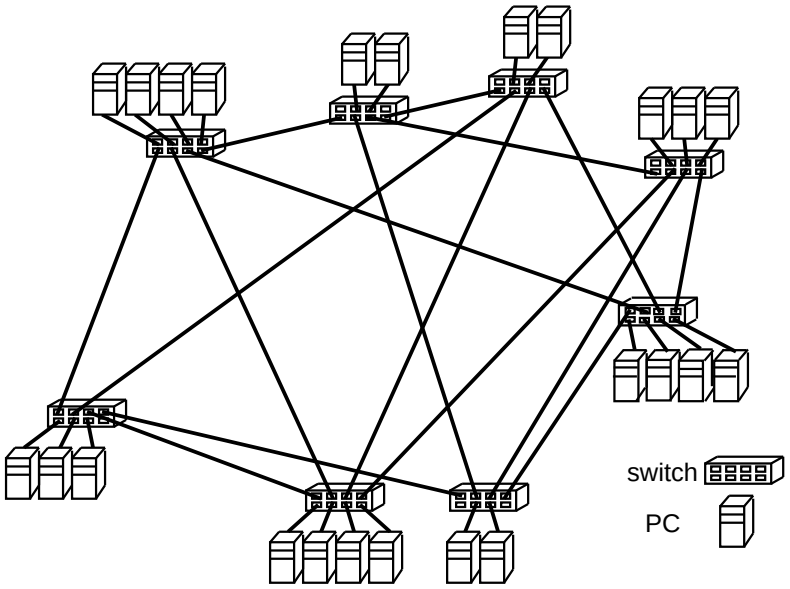


図 2.1: SAN の構成例

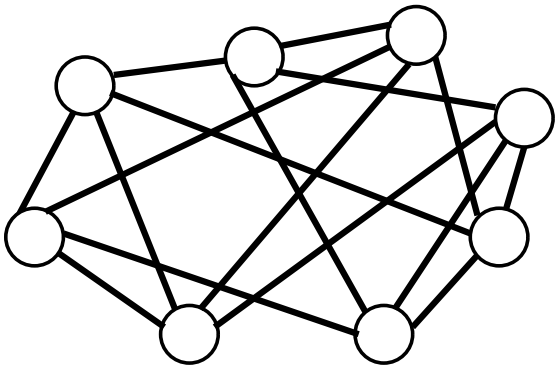


図 2.2: 図 2.1 に対応するグラフ G

各スイッチ間の結合パターンは，トポロジによって定まる．以下，SAN における代表的なトポロジについて説明する．

2.1.2　イレギュラーネットワーク

イレギュラーネットワークは，図 2.2 および図 2.3 のように，各スイッチ間の結合パターンに制限が無いトポロジである．

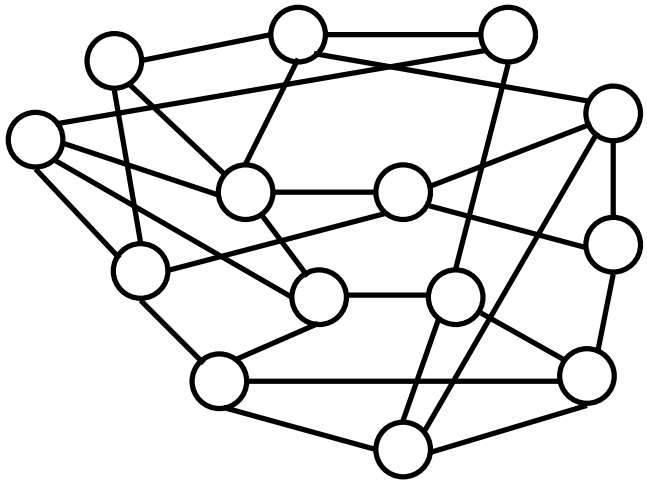


図 2.3: イレギュラーネットワーク

イレギュラーネットワークは，結合パターンに制限が無いため，次のような長所を持つ．

- リンクおよびスイッチの追加を柔軟かつ容易に行なえるため，拡張性および可用性が高い．
- 冗長なリンクおよびスイッチを追加することなどにより，耐故障性を持たせることができる．

しかし，結合パターンに制限が無いことにより，次のような問題も抱えている．

- レギュラーネットワーク向けに開発された既存の効率的なルーティングアルゴリズムが適用できない．
- 効率的なパケット転送を行なうためのルーティングアルゴリズムの設計が難しい．

このため，イレギュラーネットワークにおいては，ルーティングアルゴリズムが，特に性能に大きな影響を与える重要な要素となる．

イレギュラーネットワークをサポートする SAN としては，Autonet [1][Mae91], Myrinet [N.J95, Myra], InfiniBand[I.T04] および RHiNET[TSJ+99, 西 宏 00, STH+00, NKN+01] などが挙げられる．これらの SAN では，次に説明するレギュラーネットワークも選択可能である．

[1]Autonet は，Local Area Network (LAN) としての利用を目的として開発されたネットワークであるが，SAN と共通する特徴を多数持つため，本論文では SAN の一例として扱っている．

2.1.3　レギュラーネットワーク

レギュラーネットワークは，各スイッチ間が，特定の規則に基づいて結合されるトポロジである．レギュラーネットワークは，通常，イレギュラーネットワークよりも高いバイセクションバンド幅[2]を持ち，ネットワークの性能が特に重視される大規模並列計算機では一般的に用いられている．SAN においても，大規模並列計算機に匹敵する高性能を目的とする場合には，レギュラーネットワークが用いられることが多い．

以下，代表的なレギュラーネットワークについて説明する．

2.1.3.1　n 次元メッシュ

スイッチ間を格子状に接続したトポロジをメッシュと呼ぶ．図 2.4 は，4×4 スイッチ構成の 2 次元メッシュを示している．図 2.4 の例では各次元のスイッチ数は等しいが，次元毎にスイッチ数が異なる構成も可能である．メッシュは，ネットワークの端に位置するスイッチとそれ以外のスイッチでは隣接するスイッチ数が異なるという特徴を持つ．これまで多くの並列計算機において，2 次元メッシュもしくは 3 次元メッシュが用いられており，Intel Paragon[Int91], Stanford DASH[Dea92], MIT J-Machine[MDW93], MIT Reliable Router[Wea94] などで採用されている．

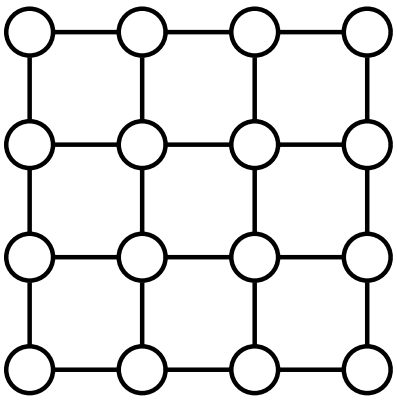


図 2.4: 2 次元メッシュ(4×4 スイッチ)

2.1.3.2　n 次元 トーラス

メッシュにおいて，各次元の端のスイッチ同士を接続し，すべてのスイッチにおける隣接スイッチの数を等しくしたトポロジをトーラスと呼ぶ．図 2.5 は，4×4 スイッチ構成の 2 次元トーラスを示している．トーラスを一般化したものは，k-ary n-cube と呼ばれ，k は各次元のリングのスイッチ数を表し，n は次元数を表している．図 2.5 のトーラスは，4-ary

[2]ネットワークを 2 つのサブネットワーク (それぞれノード数が等しい) に分割する際に切断される最小数のリンクのバンド幅の合計値を指す．バイセクションバンド幅が高いほど，トポロジの転送性能が高くなるが，実装コストも高くなる．

2-cube となる．近年の大規模並列計算機においては，3 次元トーラスが用いられる場合が多く，Cray T3D[Oed93], Cray T3E[ST96], Cray XT3[D.H] および BlueGene/L[ea02] などで採用されている．

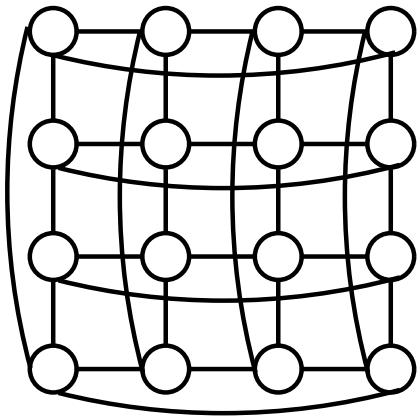


図 2.5: 2 次元トーラス (4×4 スイッチ)

2.1.3.3　Fat ツリー

Fat ツリー [C.E85] は，図 2.6 に示すように，複数のルートを持つ多重化されたツリーであり，ツリーのルート方向へのリンク数 p，ツリーの葉方向へのリンク数 q および階層数 r の組合せ (p, q, r) で定義される．Fat ツリーでは，PC は葉スイッチにおいてだけ接続され，一般的に q^r 個の PC を接続することができる．図 2.6 は，$(2, 4, 2)$ の Fat ツリーであり，16 台の PC が接続可能となっている．

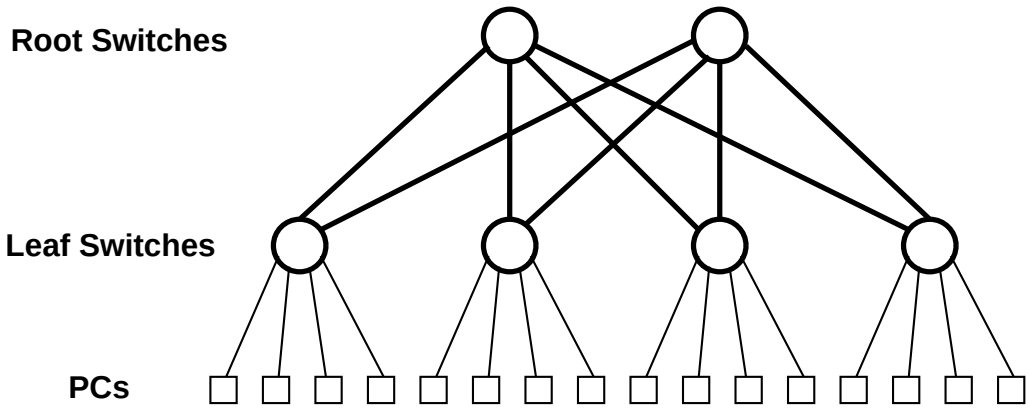


図 2.6: Fat ツリー $(2, 4, 2)$

近年，SAN におけるレギュラーネットワークとして Fat ツリーが用いられる場合が多い．代表例としては，QsNET[PFH01, FFA+02]，Myrinet, InfiniBand, などの SAN が挙げられる．

2.2　ルーティングアルゴリズム

パケットが，出発地スイッチから目的地スイッチに到達するまでに経由するスイッチおよびチャネル (物理チャネルまたは仮想チャネル) は，ルーティングアルゴリズムによって決定される．ここでは，SAN におけるルーティングアルゴリズムについて，パケット転送に関連する基本的な要素と共に説明する．

2.2.1　パケット転送方式

SAN において，ネットワークを介して PC 間でやりとりされるメッセージは，パケットの形で転送される．パケットは，物理チャネルもしくは仮想チャネルを通じて，隣接スイッチ間またはスイッチと PC 間で転送され，出発地から 1 つ以上のスイッチを経由して目的地まで到達する．パケットの形式は, システムによって異なるが, 一般的には, 図 2.7 のように，目的地 PC の番号 (ルーティング情報)，パケット長などの情報を含むヘッダ部分とデータ本体から成る．多くのマシンでは, 物理チャネルは，8 bit から大きいもので 64bit 程度のデータ幅を持つが, 1 回の転送では，パケット全体を送り切ることができない. 物理チャネルに 1 クロックで挿入することができる単位をフリットと呼び, 1 つのパケットは, フリットを単位として転送される.

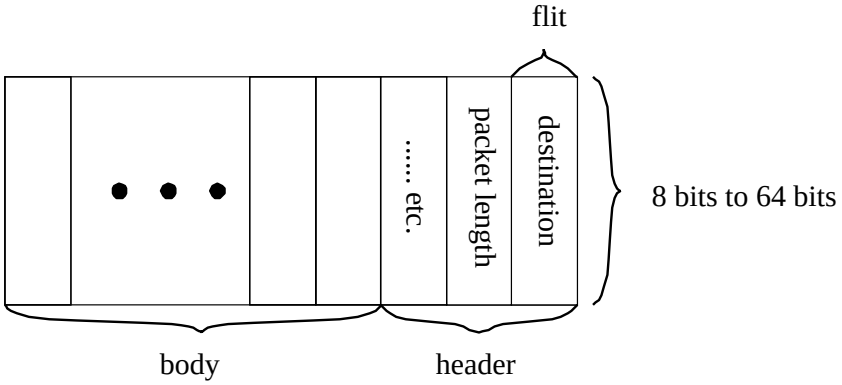


図 2.7: パケットの構成

パケット長は，固定の場合と可変を許す場合があるが, 可変長の場合でも最大長は決まっている. 可変長パケットは, ヘッダ内にパケット長を入れておき, 各スイッチ，PC でパケットの終わりを検出できるようにしておくのが普通である.

パケット転送方式は, 次の 3 方式に大別される．

(a) Store-and-Forward 方式 (SF 方式)
各スイッチは，パケット全体を格納することができるチャネルバッファを持つ. 図 2.8(a) に示すように, 各スイッチはパケット全体をチャネルバッファに受けとってから, 順に次のスイッチに渡していく.

(b) **ワームホール方式** (WH **方式**)[DS87]
各スイッチは，基本的には, **数フリット分を格納することのできるチャネルバッファを持つ**. **図** 2.8(b) **に示すように**, **パケットの先頭は**, **送り先のチャネルバッファが空いている限り**, **次々と先のスイッチに進んでいく．パケットは複数のスイッチのチャネルバッファの列にまたがって格納され**, **全体がいも虫のように前進する．先頭が進もうとするバッファが**, **他のパケットによって使われていた場合**, **パケットの進行はそこでストップし**, **チャネルバッファが空くのを待って前進を再開する**.

(c) **バーチャルカットスルー方式** (VCT **方式**)[KK79]
SF **方式同様**, **各スイッチはパケット全体を格納することのできるチャネルバッファを持つ**. **しかし**, **ワームホール方式同様，パケットの先頭は**, **本体の到着を待つことなしに次々と先のスイッチに進んでいく**. **パケットの先頭が**, **他のパケットによってブロックされた場合**, **パケット本体の転送は停止することなしに**, **先頭フリットのあるスイッチのチャネルバッファに格納される**.

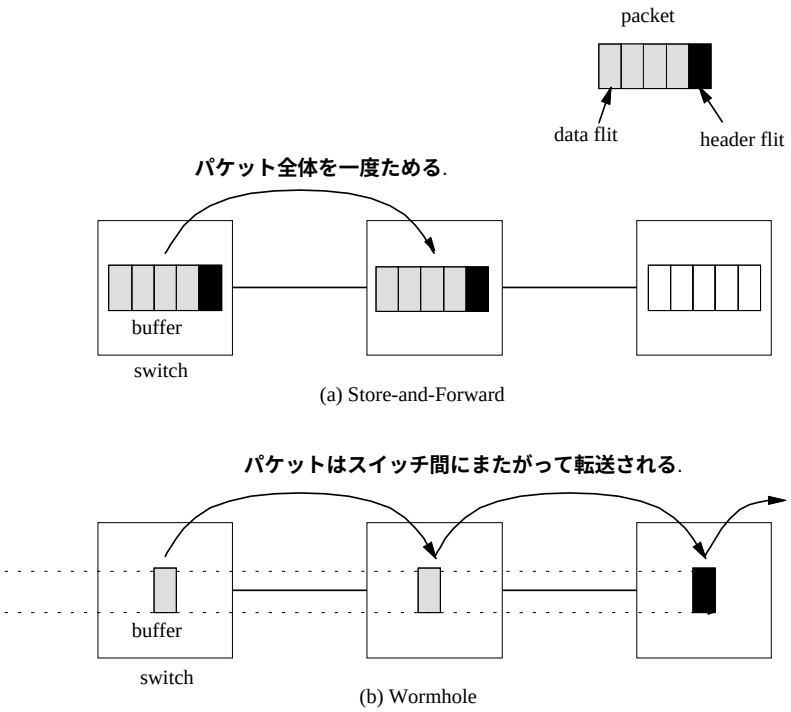


図 2.8: **パケット転送方式**

SAN **では，パケットの転送を開始する場合**, **専用のハンドシェーク線を使ってハンドシェークを取るが**, **転送を開始した後**, **クロックに同期して** 1 **フリットごとに転送を行なっていく**.

この際，WH **方式 の場合，受信バッファのオーバフローを抑えるために，先頭フリットがブロックされていないか，専用のハードウェアで監視する必要がある．一方** SF **方式の場合，パケットを受けとりつつ，次のスイッチに送ることはせず，受信バッファのオーバ**

フローも起きないためソフトウェア処理が可能である．

SAN では，PC 間の低レイテンシ通信を実現するため，通常，WH 方式もしくは VCT 方式が用いられる．これは，WH 方式および VCT 方式のパケット転送時間が，SF 方式に比べてずっと短いためである．これは次の式より明らかである．

ここで，ネットワークにおいて，スイッチ間の距離の最大値を示す直径を D，パケットヘッダのフリット数を F_h，本体 (データ部) のフリット数を F_b とし，1 フリットを 1 クロックで転送可能であるとする．SF 方式では，全スイッチで一通り，パケットを格納する必要があるため，パケット転送に要する時間は

$$(F_h + F_b) * D$$

となる．これに対して，WH 方式および VCT 方式では，各スイッチは，ヘッダの格納だけが必要なので

$$F_h * D + F_b$$

となる．通常，ヘッダは数フリットですむため，上記 2 式において，F_h の値は，F_b よりもずっと小さくなる．つまり，SF 方式では，転送遅延が直径 D の影響を大きく受けるのに対し，WH 方式および VCT 方式では，転送遅延が直径 D の影響を受けないことを示している．

VCT 方式では，先頭フリットが他のパケットにブロックされた場合も，後続のフリットは停滞せずに進む．このため，VCT 方式はブロックされたパケットが他のパケットの進行を妨げることが少ない点で，WH 方式より優れている．一方，WH 方式では，スイッチの内部に，パケットヘッダ分のバッファを用意すればよいため，バッファサイズを最小に抑えることができる．このため，WH 方式は，スイッチのハードウェア量の点で，VCH 方式よりも優れている．

2.2.2　仮想チャネル

WH 方式の問題点は，パケットの先頭がブロックされると，そのパケットは複数のスイッチのバッファを占有しながら停止してしまう点にある．この場合問題となるのは，図 2.9 のように，ブロックされたパケット A によりバッファが占有されるため，進行方向のバッファが空いているパケット B もブロックされてしまう点である．

そこで，図 2.10 に示すように，スイッチ内に別のバッファを設け，そのバッファが空いているかどうかを判断するハンドシェーク線を独立に設ける．このようにすると，パケット B は，空いている方のバッファを利用して，ブロックされることなしに先に進むことができるようになる．この方法は，ちょうど一車線しかない道路では，右折する車によって後続車がすべてブロックされてしまうのが，二車線にして右折レーンを設けることにより，ブロックがなくなるのに似ている．

新たに設けられたバッファは，バッファの量を増やすだけでなく，独立のハンドシェーク線を用いて，独立にパケット転送を行なうことが必要である．この手法では，それぞれのバッファにより，スイッチ間に仮想チャネルと呼ぶ仮想的な転送経路を作ることができる．仮想チャネルを利用したスイッチ間の転送制御を，仮想チャネルフロー制御と呼ぶ [Dal92]．仮想チャネルフロー制御を行なうことにより，複数の仮想チャネルで共有される

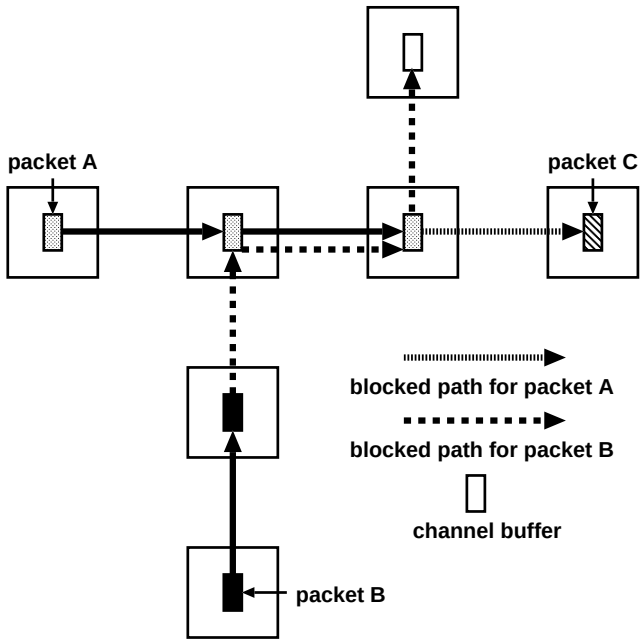


図 2.9: **パケットのブロック**

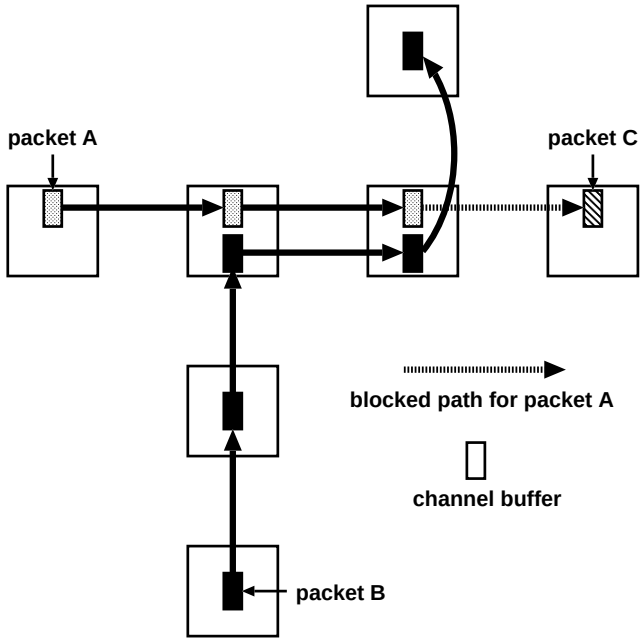


図 2.10: **仮想チャネルの利用によるブロックの回避**

物理チャネルの利用率が向上し，物理チャネル数を増やすことなしに, 結合網の転送容量を飛躍的に上げることができる．図 2.10 では 2 本の仮想チャネルを使っているが，必要に応じて何本も設けることが可能である．

仮想チャネルは，物理チャネル利用率の向上という長所を持つが，仮想チャネルバッファとハンドシェーク線の追加よるハードウェア量の増加という短所も抱えている．このため，仮想チャネルは，必ずしもすべての SAN でサポートされているわけではない．

2.2.3　デッドロック

SAN におけるルーティングアルゴリズムでは，高性能かつ高信頼性の通信を実現するために，デッドロックに対する処理が特に重要となる [JSL02, 天野 96]．

デッドロックとは，ネットワークを通過中のパケットが，起こる可能性がない事象を待ち続けることにより，転送することが不可能となる状態のことをいう．

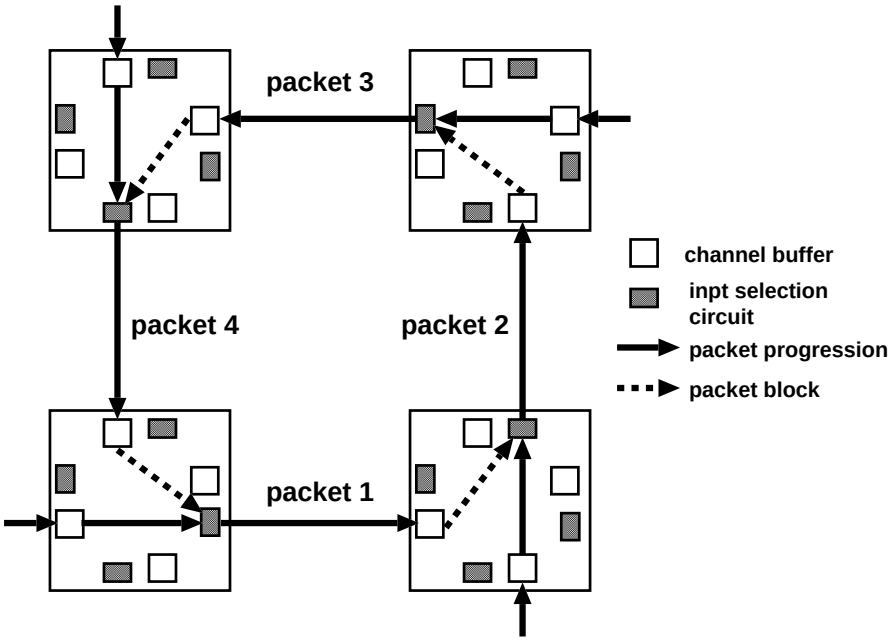


図 2.11: デッドロックの例

デッドロックが発生するのは, パケットが利用するスイッチのチャネルバッファ間に，論理的な循環構造が形成されるためである．図 2.11 にデッドロックの例を示す．図 2.11 では，4 つのパケットが，それぞれ進行方向のチャネルバッファが空くのを待っているが，互いに他のパケットが必要とするチャネルバッファを循環的に占有しあっているため，先に進むことができなくなっている．デッドロックは，チャネルバッファが有限かつ循環構造を形成する可能性がある限り，WH 方式に限らず，VCT 方式，SF 方式でも生じるが，複数のチャネルバッファを占有した状態でパケットがブロックされる WH 方式では特に発生しやすい．

2.2.4　デッドロックリカバリー方式とデッドロックフリー方式

デッドロックの対応策として次の 2 つの方式がある．

(a) デッドロックリカバリー方式

(b) デッドロックフリー方式

デッドロックリカバリー方式は，デッドロックが発生した場合，パケットの廃棄，再送処理等を行なうことにより，パケット転送を保証する方法である．しかし，一般的にデッドロックリカバリー方式は，

- デッドロックの検出およびデッドロックからの回復機構などが複雑になるためソフトウェアのオーバヘッドが大きくなる
- デッドロックが頻繁に発生すると性能が極端に低下する，

などの問題があるため，様々な研究 [ST97, ST99, KT95b, KT95a, KTJ96] が行われているものの，SAN において実装された例は，ほとんどない．

一方，デッドロックフリー方式は，ルーティングアルゴリズムによって定められるチャネル利用方法に制限を課し，パケット転送時に，図 2.11 のような循環構造の形成を防ぐことにより，デッドロックの発生を無くす方式である．このようなルーティングアルゴリズムは，デッドロックフリールーティングと呼ばれ，SAN では，高性能，高信頼性の通信を実現するために必要である．このため，本論文においては，デッドロックフリールーティングを対象としている．

2.2.5　固定型ルーティングと適応型ルーティング

SAN におけるデッドロックフリールーティングアルゴリズムは，大規模並列計算機と同様に，固定型ルーティングと適応型ルーティングの 2 つに分類される．

固定型ルーティングは，出発地スイッチと目的地スイッチが決まると，常に同じ経路を用いてパケット転送を行なう方法である．固定型ルーティングは次の長所を持つ．

- 経路選択のための機能が単純であるため，スイッチの実装コストが抑えられ，高速化がしやすい．
- パケットが送信順に必ず到達する性質 (FIFO 性) を持つ．
- 経路が固定されているため，パケット配送エラーの検出が容易である．

このため，固定型ルーティングは，既存の高速スイッチ [Myra, PFH01, FFA+02] の多くで採用されている．しかし，パケット転送中に動的に経路を選択することができないため，次の短所を持つ．

- 混雑箇所の回避ができないため，ネットワークの資源を効率良く利用できない場合がある，

- 故障箇所をその場で迂回することができない (耐故障性を持たない)．

適応型ルーティングは，出発地スイッチから目的地スイッチまでに複数の経路を用意し，途中経路で動的な経路選択を行なってパケットを転送する方法である．適応型ルーティングの機能は，適応型アルゴリズムと出力選択機構 (output selection function: OSF) の 2 段階に分けられる．前者は，対象ネットワーク上のすべてのスイッチ間におけるデッドロックフリーな経路の集合 (各スイッチ間の経路は 1 つ以上存在する) を提供する機能である．一方，後者は，適応型アルゴリズムによって提供される隣接スイッチへの経路候補から一つの経路を提供する機能であり，転送中のパケットが隣接スイッチに移動する際に実行される [BP89, Wu96, Wu99]．これらは互いに独立した機能であり，また，一般的に適応型ルーティングの性能は適応型アルゴリズムによって決定される場合が多いことから，前者を指して適応型ルーティングという場合も多い．本論文においても，適応型ルーティングにおける適応型アルゴリズムを対象としている．

適応型ルーティングは，固定型ルーティングと対照的に次の点が長所となる．

- 混雑箇所を迂回することにより，空きチャネルを効率よく利用できる．
- 故障箇所をその場で迂回することが可能である (耐故障性を持つ)．

このため，適応型ルーティングは，チャネル利用率の向上や動的な耐故障性の実現を目的とする場合に用いられる [Mae91]．一方，適応型ルーティングでは，固定型ルーティングと対照的な次の点が短所とされる．

- 出力選択機能のため，実装が複雑となり，高速化がしにくい．
- FIFO 性の保証が難しい．

しかし，前者の問題については，実用化された Autonet で用いられているような単純な選択機構 [Mae91, AMAH01] を用いて，並列に出力チャネルの状態をチェックするなどにより，実用可能なレベルに抑えることが可能であると考えられている．また，後者の問題についても，近年，複雑なハードウェアを用いずに，適応型ルーティングの機能面および性能面におけるアドバンテージを維持しつつ FIFO 性を保証するために，FIFO 転送法 [MJJ$^+$05]，一対制限法 [MJJ$^+$05] が提案されている．両手法とも，各出発地 PC 毎に，1 つの目的地 PC に一度に送信可能なパケット数を制限する．FIFO 転送法では，同一目的地 PC への (前の) パケットの ACK (acknowledgment) を受信した後に，次のパケットをネットワークに注入することにより，目的地 PC におけるパケットのソートなしに FIFO 性を実現する．一方，一対制限法では，同一目的地 PC へ一度に送信可能なパケット数を最大で 2 つとすることにより注入制限を緩和して性能向上を図る．2 パケット間で out-of-order 転送が生じるため，FIFO 性の実現のために目的地 PC におけるパケットのソートを必要とするが，ACK と NACK (negative acknowledgment) を併用することにより，ソートのために必要となるバッファサイズを数パケット分に抑えている．

以上より，将来的には，チャネル利用率の向上と動的な耐故障性が重視される場面などにおいて，適応型ルーティングの需要が高まると考えられる．

2.2.6　ソースルーティング方式と分散ルーティング方式

パケットが，ルーティングアルゴリズムによって決定される経路情報を取得する方式は，ソースルーティング方式と分散ルーティング方式に分類される．

ソースルーティング方式では，各出発地スイッチにおいて，ルーティングアルゴリズムに基づいて，目的地スイッチに到達するまでの全経路情報が計算される．通常，全経路情報は，起動時またはトポロジ構成が変化した際などに計算され，各出発地スイッチまたは PC 上のルーティングテーブルに格納される．目的地までの経路情報は，パケット生成時に，ルーティングテーブルから取得され，パケットヘッダに格納される．ネットワークに注入されたパケットは，ヘッダが途中スイッチに到達するたびに，ヘッダに格納された経路情報に従って，次の移動先スイッチの決定と転送を行ない，これを目的地スイッチに到達するまで繰り返す．ソースルーティング方式では，各経由スイッチにおける経路決定処理が簡素化されるため，スイッチの実装コストを抑えられるという利点を持つ．しかし，全経路情報を格納するため，その分ヘッダ長のオーバヘッドが大きくなる．また，目的地スイッチまでの経路が出発地スイッチで決定されるため，途中経路において動的に経路を選択することができないという制限を持つ．このため，ソースルーティング方式におけるパケット転送は，固定型ルーティングとなる．ソースルーティング方式において，適応型ルーティングにより提供される複数経路を利用することは可能であるが，この場合，経路選択は出発地スイッチにおいてだけ可能となり，パケット転送は固定型ルーティングにより行なわれる．

分散ルーティング方式では，パケットが経路途中のスイッチに到達するたびに，スイッチに実装された経路決定用ハードウェアにより，ルーティングアルゴリズムに基づいて移動先隣接スイッチが決定される．一般的には，スイッチのルーティングテーブルに格納された計算済の経路情報を参照する table-lookup 方式が用いられることが多い．分散ルーティング方式では，中間スイッチにおいて，動的に経路を選択することが可能であるため[3]，適応型ルーティングの長所を十分に活用することができる．また，ヘッダに格納される経路情報が小さくて済むという利点も持つ．一方，スイッチの実装がより複雑となるため，ソースルーティング方式に比べて，実装コストの面で劣る．

[3]ただし，動的な経路選択が可能かどうかはスイッチの実装に依存する．

第3章　関連研究

本章では，イレギュラーネットワーク向けに開発された既存のルーティングアルゴリズムとその問題点について述べる．また，第 4 章で提案する適応型ルーティングアルゴリズムである L-turn および R-turn ルーティングの適用対象となる代表的な SAN の実現例についても述べる．

3.1　イレギュラーネットワーク向けの既存のルーティングアルゴリズム

従来より，第 2.1 節で述べたメッシュやトーラスなどのレギュラーネットワークを対象とする多くの固定型および適応型ルーティングアルゴリズムが，大規模並列計算機における利用を目的として提案されてきた [DS87, LH91, GN92, DA93, Dua93, Dua94] [1]．しかし，これらのルーティングアルゴリズムは，対象とするトポロジの規則性に依存しているため，結合パターンに規則性を持たないイレギュラーネットワークには適用することができない．ここでは，SAN のイレギュラーネットワークに適用可能な既存のルーティングアルゴリズムに焦点を当てて説明する．

イレギュラーネットワーク向けのルーティングアルゴリズムは，大きく分けて，仮想チャネルや専用バッファなどの，スイッチまたは PC 上の付加的なハードウェアに依存しない手法と，依存する手法に分類することができる．前者の例としては，Up*/Down* ルーティング [Mae91]，DFS スパニングツリーベースの Up*/Down* ルーティング [JAJ00], Smart ルーティング [LVT96], Adaptive-Trail ルーティング [QNR99], などがあり，後者の例としては，構造化チャネル法 (Structured Buffer Pool)[MJ80], Minimal ルーティング [SD00], LASH ルーティング [SLT02], In-Transit バッファを利用する手法 [JPMJ02], 複数スパニングツリーを利用する手法 [JPMJ02], DL ルーティング [MAH03] などが挙げられる．これらのうち，後者の各ルーティングアルゴリズムは，次の長所を持つ．

- ほとんど，もしくはすべての経路において，トポロジ上の最短経路を取ることができる．
- 仮想チャネルの利用により，物理チャネルをより効率的に利用することができる．

このため，前者のルーティングアルゴリズムに比べて，優れたトラフィック分散と高スループットを実現している．しかし，これらのルーティングアルゴリズムは，付加的なハードウェアに依存するため，汎用性に欠け，適用可能なネットワークが限定されるという問題を持つ．そこで，本論文では，付加的なハードウェアに依存しないイレギュラーネットワー

[1] これらのルーティングアルゴリズムの詳細については，[舟橋 99] でまとめられている．

ク向けのルーティングアルゴリズムを対象とする．以下，そのような既存の 4 つのルーティングアルゴリズムについて説明し，その後，それらの問題点について述べる．

3.1.1　Up*/Down* ルーティング

Up*/Down* ルーティングは，イレギュラーネットワーク向けの代表的な適応型ルーティングであり，Autonet[Mae91] や Myrinet[N.J95] などの SAN において利用されている．

Up*/Down* ルーティングは，任意のトポロジに対して，デッドロックフリーと任意のスイッチ間の経路を保証するために，対象ネットワークに対してスパニングツリーのマッピングを行なう．スパニングツリーとは，ネットワーク内のすべてのスイッチを含むツリーのことである．スパニングツリーは，(1) 循環が存在しない，(2) 任意のスイッチ間の経路が常に存在する，という特性を持ち，Up*/Down* ルーティングでは，この特性を利用することにより，デッドロックフリーと経路保証を実現する．スパニングツリーの構築方法は，breadth first search (BFS) に基づく．例として，Autonet では，minimum depth スパニングツリー (MDST)[Mae91] および propagation order スパニングツリー (POST)[RS91] などの BFS スパニングツリー構築アルゴリズムがある．これらは共に，スパニングツリーの高さが最小となることを念頭に置いている (POST では必ず最小となることが保証されないが，ほとんどの場合に最小となることが Autonet では，確認されている [RS91])．ここでは，POST によるスパニングツリー構築アルゴリズムを簡単に示す．

(1) 全スイッチの中からスパニングツリーのルートスイッチを選択する．

(2) ルートスイッチは，すべての隣接スイッチに join 要求メッセージを送信し，要求を受諾したスイッチを子としてスパニングツリーに追加する．

(3) あるスイッチの子となったスイッチは，同様にすべての隣接スイッチに join 要求メッセージを送信し，要求を受諾したスイッチ (既にスパニングツリーに含まれているスイッチは要求を拒否する) を自身の子としてスパニングツリーに追加する．

(4) 全スイッチがスパニングツリーに含まれるまで (3) の作業を繰り返す．

スパニングツリー構築の完了後，ネットワーク上のすべてのチャネル (ツリーを構成するチャネル (tree channel) とそれ以外のチャネル (outer channel)) に対して，次の規則に基づいて up または down の方向を割当て，1 次元の方向を持つ有向グラフを構築する．

(1) up 方向を，次の 2 つの条件のいずれかを満たすチャネルに対して割当てる．なお，ここではスイッチ数を n とし，各スイッチには 0 から $n-1$ までの，一意の整数の ID が割当てられているとする．

  (a) 接続先のスイッチが接続元のスイッチよりもルートスイッチに近い．

  (b) 接続先のスイッチと接続元のスイッチのルートスイッチからの深さが同一であり，かつ，接続先のスイッチの ID が接続元のスイッチの ID よりも小さい．

(2) down 方向を残りのすべてのチャネルに対して割当てる．

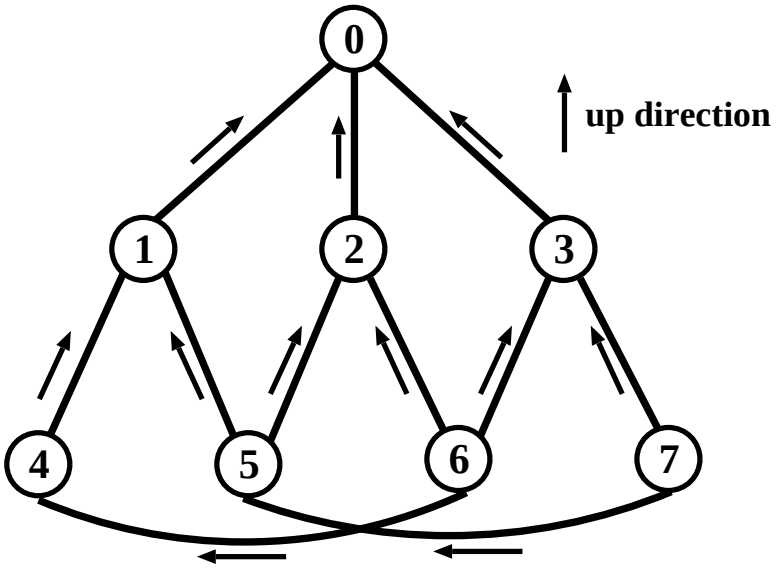


図 3.1: BFS スパニングツリーに基づいた有向グラフ

以上の手順により，図 3.1 のような有向グラフが構築される．

上記の方向割当てにより，有向グラフにおけるパケットの方向転換 (ターン) は，up 方向から down 方向 のターンと down 方向から up 方向 のターンの 2 つとなる．したがって，有向グラフ内で形成されるすべての循環構造は，これら 2 つのターンの連鎖により形成される．Up*/Down* ルーティングは，デッドロックフリーと経路保証の実現のために，次の規則に従って，ルーティングを行う．

> すべてのパケットは，0 回以上 up 方向に移動した後に， 0 回以上 down 方向に移動して目的地スイッチまで到達する．

この条件により，パケットは down 方向から up 方向へのターンができなくなるため，これによりすべての循環構造が除去され，デッドロックフリーが保証される．また，up 方向から down 方向へのターンを許可することにより，パケットは常にスパニングツリーのチャネルを経由した移動が可能となるため，これにより経路保証が実現される．

Up*/Down* ルーティングは，上記のように，パケットのターンにより形成される循環構造に着目して，デッドロックフリーを実現しているため，対象ネットワークのチャネルの実装 (物理チャネルまたは仮想チャネル) に依存することがない．

Up*/Down* ルーティングは，上記の条件を守る限り，自由に経路を選択できるが，通常は，任意の非最短経路を許した場合に発生しやすいホットスポット形成を抑えて効率的なルーティングを行なうために，選択可能な経路のうち最短となる経路だけを選択する．ただし，常にトポロジ上の最短経路を取ることができるわけではないので，Up*/Down* ルーティングは，非最短型の適応型ルーティングとなる．例えば，図 3.1 において，スイッチ 5 からスイッチ 0 へパケットを転送する場合には，スイッチ 1 またはスイッチ 2 を経由してスイッチ 0 まで到達することができるので，すべての最短経路を選択することができる．これに対して，スイッチ 3 からスイッチ 5 へパケットを転送する場合には，down 方

向から up 方向への移動が必要となるため，スイッチ 7 を経由するトポロジ上の最短経路は選択することができず，スイッチ 0 → 1 またはスイッチ 0 → 2 を経由する非最短経路しか選択することができない．

Up*/Down* ルーティングの経路計算のための計算量は，スイッチ数を n とすると $O(n^2)$ となる．なお，Up*/Down* ルーティングでは，1 つ以上の経路が選択可能であるため，適応型ルーティングが可能となっているが，各出発地スイッチ，目的地スイッチ間の経路を 1 つに固定することにより，固定型ルーティングとして利用することも可能となっている．このため，Up*/Down* ルーティングは，固定型ルーティングだけをサポートしているネットワークにおいても利用可能であるという長所も持つ．

3.1.2　DFS スパニングツリーベースの Up*/Down* ルーティング

Up*/Down* ルーティングの性能は, 各チャネルに対する方向の割当て方に大きく影響されるため，スパニングツリーおよび有向グラフの構築アルゴリズムが重要となる．そこで，前節で述べた BFS スパニングツリーを利用する手法の改良案として，Sancho らによる，ヒューリスティックルールに基づいた depth first search (DFS) のスパニングツリーを利用する手法が提案された [JAJ00, JA00]．

この手法は，BFS スパニングツリーベースの手法に比べて，禁止ターン数を減少することにより性能向上を図っており，禁止ターン数の削減は，次の 2 つの特徴により実現される．

- 各スイッチに対して一意のラベルを割当てることにより，BFS スパニングツリーにおける同階層スイッチ間の冗長な禁止ターンを削除する．
- DFS スパニングツリー構築時に，禁止ターン数がより少なくなるようなヒューリスティックルールを利用する．

前節で述べた通り，BFS スパニングツリーベースの手法では，同階層スイッチ間のチャネルに対する方向の割当てが，ランダムに配置されるスイッチの ID に依存しているため，同階層チャネル間に冗長な禁止ターンが発生する可能性があるという問題が存在する．この問題の具体例として， BFS スパニングツリーベースの手法を 9 スイッチのイレギュラーネットワークに適用した図 3.2 を示す．

図 3.2 では，同階層スイッチ間を結ぶ太線で示された一連のリンク $(5 \rightarrow 7 \rightarrow 6 \rightarrow 8 \rightarrow 5)$ がリング状に接続されている．これらのリンク間では，スイッチ 7 と 8 の 2 箇所において down から up のターンが発生するため，これら 2 つのスイッチにおいて禁止ターンを課すことにより循環構造が除去されている．しかし，1 つのリング内の循環構造を除去するためには，1 箇所でだけ禁止ターンを課せばよいことは明らかであり，例えば，スイッチ 7 とスイッチ 6 の位置を入れ替えることにより，スイッチ 7 における禁止ターンを無くすことができる．しかし，BFS スパニングツリーベースの手法では，同階層スイッチの各 ID はランダムに決定されるため，この問題を解決することは難しい．

このような問題の解決と，更なる禁止ターンの削減のために，次の手順に基づいて DFS スパニングツリーを用いた Up*/Down* ルーティングを構築する．

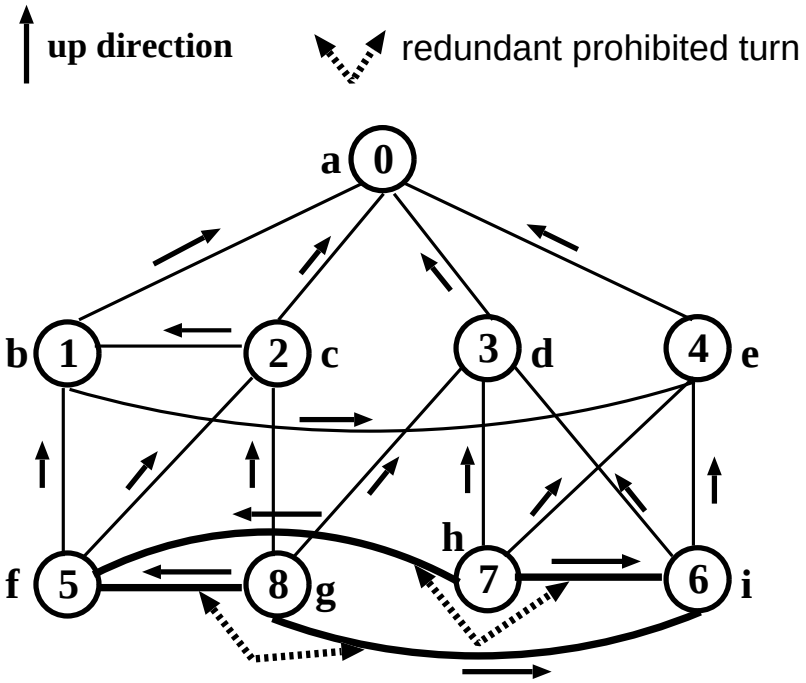


図 3.2: BFS **スパニングツリーを用いた** Up*/Down* **ルーティングにおける冗長禁止ターン**

(1) DFS **スパニングツリーの構築**

(2) **各スイッチへのラベルの割当て**

(3) **各チャネルへの方向の割当て**

3.1.2.1　DFS スパニングツリーの構築

DFS **スパニングツリーの構築は，ルートスイッチを選択した後，ルートスイッチを起点とした深さ優先ベースの探索を行ない，訪問順に残りのスイッチをスパニングツリーに組み込むことにより行なわれる．この手続きにより，まず，ルートスイッチを始点とし，すべての隣接スイッチが訪問済となったスイッチを終点とする** main branch **と呼ばれるパスが構築される．**main branch **にすべてのスイッチが組込まれなかった場合には，未訪問の隣接スイッチを持つスイッチを起点とした** secondary branch **と呼ばれるパスを同様にして構築し，これをすべてのスイッチがスパニングツリーに組込まれるまで再帰的に行なう．探索時に，次に訪問可能な隣接スイッチが複数存在する場合があるが，隣接スイッチの選択は禁止ターンの削減を目的としたヒューリスティックルールに基づいて行なわれる．これについては，第** 3.1.2.4 **節で述べる．**

例として，図 3.2 **のネットワークに対しては，図** 3.3 **のような** DFS **スパニングツリーが構築される．なお，図** 3.3 **では，後述のラベル (整数値) との混同を避けるため，図** 3.2 **の各スイッチとの対応をアルファベットの** ID **で表している．**

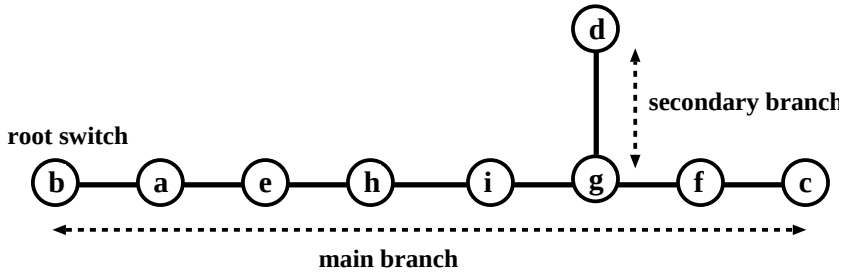


図 3.3: DFS スパニングツリー

3.1.2.2　各スイッチへのラベルの割当て

DFS スパニングツリーの各スイッチに対して，一意となる整数 (ラベル) を割当てる．ラベルの割当ては，main branch に含まれるスイッチに対しては，スパニングツリーを構築する際の訪問順に従い，0 から始まる昇順の整数を割当てる．一方，secondary branch に含まれるスイッチでは，訪問順に従い，分岐スイッチのラベルから始まる降順の整数を割当てる．これにより，secondary branch の葉スイッチに対しては，その branch の中で最も小さい整数が割当てられる．

図 3.4 に，図 3.3 の DFS スパニングツリーにおけるラベル割当ての例を示す．

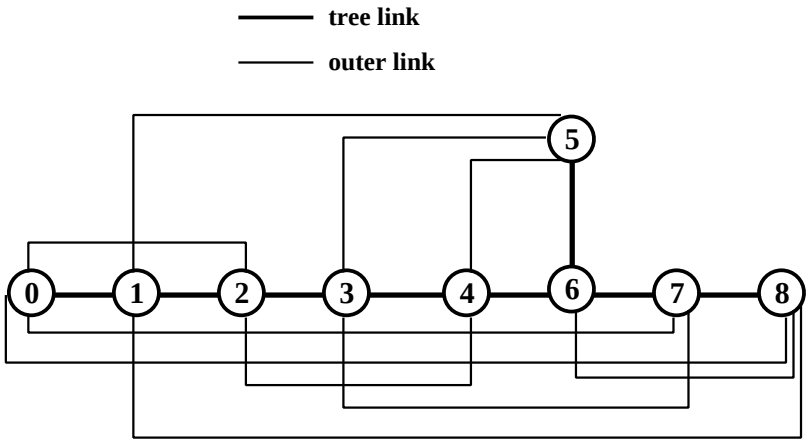


図 3.4: 各スイッチへのラベルの割当て

なお，図 3.4 において，太い線は，スパニングツリーを構成するリンク (tree link) を示し，細い線は，スパニングツリー構成外のリンク (outer link) を示している．

3.1.2.3　各チャネルへの方向の割当て

各スイッチに割当てられたラベルに基づいて，すべてのチャネルに対して，方向 (up または down) の割当てが行なわれる．接続先のスイッチのラベルが，接続元のスイッチのラ

ベルより大きいチャネルに対して up 方向を割当て，その逆となるチャネルに対して down 方向を割当てる．

$L(x)$ をスイッチ x に割当てられたラベルを返す関数とすると，上記の割当てにより，すべての循環構造において，$L(y) > L(x) < L(z)$ が成立する (down 方向 から up 方向への移動が発生する) 3 つの連続したスイッチ x, y, z が存在する．Up*/Down* ルーティングでは，down 方向から up 方向への移動を禁止するので，$y \rightarrow x \rightarrow z$ および $z \rightarrow x \rightarrow y$ の移動が禁止される．このため，すべての循環構造が除去され，デッドロックフリーが保証される．

DFS スパニングツリーを用いる場合も，スパニングツリーを経由することにより常に任意のスイッチ間でパケット転送が可能であるため，BFS スパニングツリーベースの手法と同様に，経路保証が実現されている．また，各スイッチに対して，昇順または降順に基づく一意のラベルが割当てられたことにより，BFS スパニングツリーベースの手法において問題となった同階層チャネル間の冗長な禁止ターンが発生しなくなる．

例として，図 3.4 のグラフに対して方向を割当てた有向グラフを図 3.5 に示す．

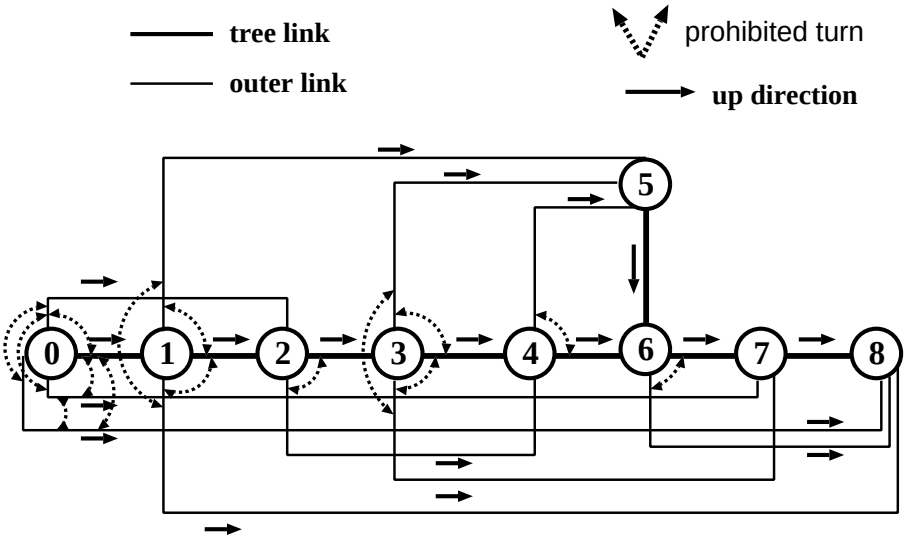


図 3.5: DFS スパニングツリーを用いた Up*/Down* ルーティングにおける各チャネルへの方向の割当てと禁止ターン

図 3.5 では，図 3.2 の BFS スパニングツリーベースの手法で発生した同階層チャネル間の冗長な禁止ターンが存在しないことがわかる．実際に，図 3.2 では，34 個 (17 ペア) の禁止ターンが存在するのに対し，図 3.5 では 30 個 (15 ペア) となっており，DFS スパニングツリーベースの手法により禁止ターン数の削減が実現されている．

図 3.5 の有向グラフにおいて，自身以外のすべてのスイッチから up 方向だけの移動により到達可能なスイッチは，スイッチ 8 (図 3.3 のスイッチ c) である．つまり，図 3.5 の有向グラフ上のルーティングにおけるルートスイッチは，スイッチ 8 となる．これは，スパニングツリー構築時の深さ優先探索におけるルートスイッチであるスイッチ 0(図 3.3 のスイッチ b) とは異なる．このように，DFS スパニングツリーベースの Up*/Down* ルーティングでは，スパニングツリー構築時とルーティング時で，ルートとなるスイッチがそ

れぞれ異なる．

DFS スパニングツリーベースの Up*/Down*ルーティングは，BFS スパニングツリーベースの手法と同様に，選択可能な経路のうち最短となる経路だけを選択する．このため，経路計算のための計算量は，スイッチ数を n とすると同様に $O(n^2)$ となる．

3.1.2.4　DFS スパニングツリー構築時のヒューリスティックルール

第 3.1.2.1 節で述べたように，DFS スパニングツリー構築における探索時に，訪問可能な隣接スイッチが複数存在する場合，禁止ターンの減少を目的とした次のヒューリスティックルールに基づいて訪問スイッチの選択が行なわれる．

訪問スイッチ選択ヒューリスティックルール

> 既にスパニングツリーに組み込まれているスイッチとの接続数の最も多いスイッチを選択する．該当スイッチが複数存在する場合は，該当スイッチの中で，残りのスイッチとの average topological distance (スイッチ間のトポロジ上の最短距離) が最も大きいスイッチを選択する．

このヒューリスティックルールの狙いは，各スイッチにおいて形成される禁止ターンのパターンとして，図 3.6 の (c) のような，禁止ターンが集中 (図では 6 つ存在) するパターンを避け，可能な限り，(a) または (b) のような禁止ターンが少ない (図では 0 または 2 つ) パターンを選択することにある．これにより，冗長な禁止ターンの削除に加えて，更なる禁止ターンの削減を図っている．

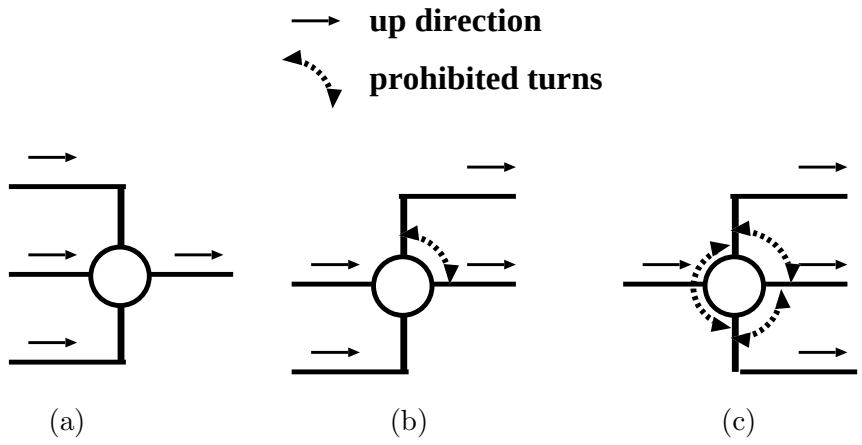


図 3.6: スイッチに接続されるリンクの方向と禁止ターン

なお，Sancho らは，ルートスイッチの選択が Up*/Down* ルーティングの性能に影響を受けることに着目し，ルートスイッチの選択における次のようなヒューリスティックルールについての提案 [JA00] も行なっている．

ルートスイッチ選択ヒューリスティックルール

対象ネットワークのすべてのスイッチに対して，各々をルートとした場合の

(a) **ルーティングアルゴリズムによって決定される各スイッチ対の最短経路，**

(b) crossing path(**各チャネルを通過する経路数の中の最大値を示す**)，

(c) average distance(**各スイッチ対の最短経路の平均距離を示す**)

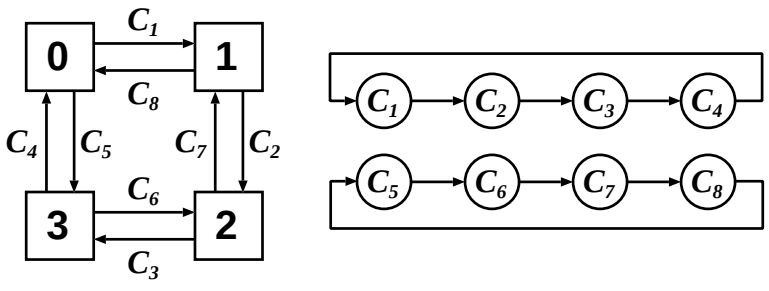
をそれぞれ計算する．これらのうち，crossing path **が最小値となるルートスイッチを選択する．もし，**crossing path **の最小値が同数の場合，**average distance **が最小値となるルートスイッチを選択する．**

このヒューリスティックルールの狙いは，crossing path **を最小とすることによってトラフィックの分散を図ることであり，また，**average distance **についても考慮することによって，**crossing path **が同数の場合には，より多くの最短経路が選択可能となることを図っている．このヒューリスティックルールが必要とする計算量は，スイッチ数を** n **とすると，** $O(n^3)$ **である．**

ルートスイッチ選択のヒューリスティックルールは，DFS **スパニングツリーベースの** Up*/Down* **ルーティングだけでなく，**BFS Up*/Down* **ルーティングを始めとするスパニングツリーベースの任意のルーティングアルゴリズムに対して適用可能であり，一般性の高い手法となっている．**

3.1.3　Smart ルーティング

Smart **ルーティング** [LVT96] **は，スパニングツリーでなく，**channel dependency graph (CDG) **の構築をベースとしてデッドロックフリーを実現する適応型ルーティングアルゴリズムである．**CDG **とは，対象とするトポロジにおける** channel dependency **を表した有向グラフである．例として，図** 3.7(a) **の** 2×2 **スイッチの** 2 **次元メッシュにおける** CDG **は，図** 3.7(b) **のように表される．図** 3.7(b) **において，各ノードが図** 3.7(a) **の各チャネルを表し，各チャネルが，図** 3.7(a) **の各チャネル間の** channel dependency **を表す．**



(a)　　(b)

図 3.7: 2 **次元メッシュ**(2×2 **スイッチ**) **の** CDG

Smart ルーティングは，対象トポロジにおける CDG を構築した上で，CDG 内で形成される循環構造を一つずつ識別し，循環構造を形成するいずれか 1 つの channel dependency を切断することにより循環構造の除去を行なう．トラフィックの分散を実現するために，循環構造除去のために切断される channel dependency は，average path lengths (すべてのスイッチ間の最短経路長の平均値) が最小となるものが選択される．この選択の際に，循環構造を形成するすべての channel dependency について，channel dependency を切断した場合の average path lengths の計算が breadth first search (切断された channel dependency は移動不可とする) によりそれぞれ行なわれる．上記の手続きは，CDG において循環構造が形成されなくなるまで行なわれる．なお，適切でない channel dependency の切断を行なった結果，手続きの途中でどの channel dependency を切断しても，あるスイッチ間における経路が切断されてしまう循環構造が識別される場合がある．このような場合においてだけ，循環構造除去の手続きを tree mode によって最初からやり直す．tree mode では，まず，トポロジ上で breadth first search によるツリー構造 (backbone) の構築を行なう．そして，経路保証の実現のために，ツリーを構成するリンク間の channel dependency は切断しないものとして，同様に循環構造の識別と除去の手続きが行なわれる．

1 つの循環構造除去のために 1 つの channel dependency を切断する際，多くの場合，それ以外の複数の循環構造も同時に除去される．このため，上記の手続きの結果切断された channel dependency の幾つかは，切断しなくても循環構造が形成されることがない．そのような channel dependency は，経路選択の自由度を高めるために，切断を解除される．

上記の手続きのための計算量は，スイッチ数を n とすると $O(n^9)$ となる．ただし，平均的には $O(n^4)$ で収まるとされている．

Smart ルーティングでは，切断されていない channel dependency を利用することにより各スイッチ間で 1 つ以上の経路を選択することができるため，適応型ルーティングが可能となる．また，CDG の構築をベースとしてデッドロックフリーを実現するため，Up*/Down* ルーティングと同様に，仮想チャネルなどの付加的なハードウェアに依存することがない．

3.1.4　Adaptive-Trail ルーティング

Adaptive-Trail ルーティング [QNR99] は，Eulerian trail をベースとした adaptive trail と呼ばれるパスを用いてデッドロックフリーを実現する適応型ルーティングアルゴリズムである．

Eulerian trail は，対象トポロジ上のすべてのリンクをそれぞれ一度だけ辿ることにより構築されるパスであり，各リンクが双方向チャネルから成るネットワークでは，互いに反対の方向に向かう 2 つの Eulerian trail が形成される．各 Eulerian trail では，すべてのスイッチが循環の無い単方向の CDG に沿って接続されているため，Eulerian trail 上でルーティングを行なうことにより，任意のスイッチ間の経路とデッドロックフリーが保証される．しかし，Eulerian trail 上だけのルーティングでは，各スイッチ間においてトポロジ上の最短経路を選択できない場合が多く，また，選択可能な経路数も限られる．

Adaptive-Trail ルーティングは，各 Eulerian trail に shortcut を追加した adaptive trail と呼ばれる パスを構築し，2 つの adaptive trail のいずれかの上でルーティングを行なう

ことにより，より多く，かつ，分散された最短経路とそれによるトラフィック分散の実現を図る．adaptive trail に追加される shortcut は，free-style shortcut, destination shortcut, source shortcut の 3 つに分類される．これらのうち，free-style shortcut は任意のパケットが利用可能であるが，destination shortcut と source shortcut は，利用可能となるパケットが制限される．前者は，接続先のスイッチを目的地とするパケットだけを対象とし，後者は，接続元のスイッチを出発地とするパケットだけを対象としている．追加された各 shortcut について，デッドロック発生の可能性がチェックされ，発生の可能性があると判定された shortcut は除去される．これにより，adaptive trail に沿ったルーティングは，デッドロックフリーが保証される．また，adaptive trail 上では，各スイッチ間で 1 つ以上の経路を選択することができるため，同様に適応型ルーティングが可能となる．

Adaptive-Trail ルーティングの経路計算のための計算量は，チャネル数を m とすると $O(m^2)$ となる．Adaptive-Trail ルーティングは，Eulerian trail をベースとしているため，Up*/Down* ルーティングと同様に，仮想チャネルなどの付加的なハードウェアに依存することがない．しかし，対象トポロジ上で Eulerian trail が構築可能であることの必要十分条件は，すべてのスイッチの degree が 偶数である，または，2 つのスイッチだけが奇数の degree を持つ，となる．このため，この条件を満たさないトポロジに対しては適用することができない．また、各トポロジにおける対処可能な故障パターンが限定されるため、耐故障性が若干落ちる。

3.1.5　既存のルーティングアルゴリズムにおける問題点

前述の付加的なハードウェアに依存しない 4 つのルーティングアルゴリズムは，スパニングツリーをベースとする手法 (BFS および DFS Up*/Down* ルーティング) と，それ以外の手法 (Smart ルーティングおよび Adaptive-Trail ルーティング) の 2 つに分類される．

Up*/Down* ルーティングでは，up と down の 2 つの方向だけの 1 次元有向グラフをベースとしているため，図 3.8 のように，禁止ターンが形成される 1 つのスイッチ上の 2 リンク間において，互いに反対の方向に向かう禁止ターンのペアが必ず形成される．

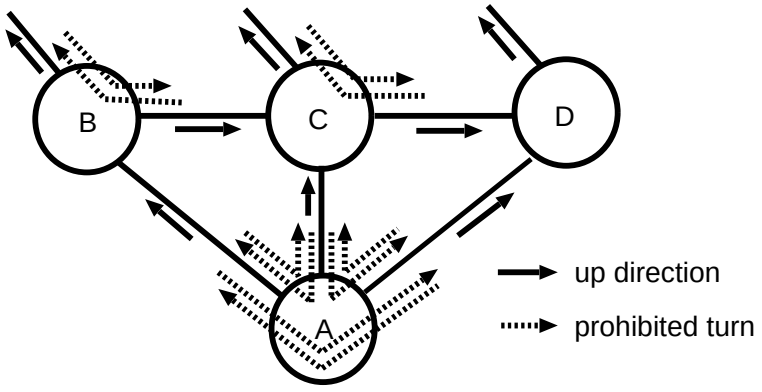


図 3.8: Up*/Down* ルーティングにおける禁止ターンペアの形成

図 3.8 において，スイッチ B と C に 1 つ，スイッチ A には 3 つの禁止ターンペアがそ

れぞれ形成されている．このような禁止ターンの偏りにより，ネットワーク内のトラフィックに偏りが生じやすくなり，ネットワークバンド幅の有効利用が困難になるという問題が発生する．

この問題を改善するためには，禁止ターンの集中を緩和する必要があるが，第 3.1.1 節で述べたように，1 次元有向グラフでは，up → down および down → up の 2 つのターンしか存在せず，また，経路保証のために up → down のターンを禁止することができないため，本質的に，禁止ターンを分散させるための選択の余地がない．

DFS ベースの手法では，ヒューリスティックルールに基づいてスパニングツリーを構築することにより，図 3.8 のような同一スイッチ上における禁止ターンペアの集中を可能な限り避け，禁止ターンを削減することにより性能向上を図っている．しかし，図 3.5 のように，常に回避することができるわけではなく，また，禁止ターンペアの形成は 1 次元有向グラフの利用を起因とするため，DFS ベースの手法は上記の問題の根本的な解決となっていない．

一方，Smart ルーティングと Adaptive-Trail ルーティングは，CDG または Eurelian trail の構築をベースとすることにより，Up*/Down* ルーティングにおけるトラフィック集中の改善を図っている．しかし，これらの手法は，汎用性の面でそれぞれ大きな問題を抱えている．まず，Smart ルーティングは，計算量が最悪の場合 $O(n^9)$ となり，これは現実的な許容範囲を超えている．また，Adaptive-Trail ルーティングは，トポロジによっては，Eulerian trail が構築できないため，任意のトポロジに適用可能ではない．

各ルーティングアルゴリズムについては，表 3.1 の通りにまとめられる．表 3.1 の計算量において，n はスイッチ数，m はチャネル数を指す．

表 3.1: 付加的なハードウェアに依存しないルーティングアルゴリズムの比較

	BFS Up*/Down*	DFS Up*/Down*	Smart	Adaptive Trail
スパニングツリー利用	yes	yes	no	no
禁止ターン集中	high	medium	-	-
トポロジフリー	yes	yes	yes	no
計算量	$O(n^2)$	$O(n^2)$	$O(n^9)$	$O(m^2)$

以上より，既存の各ルーティングアルゴリズムは，それぞれ性能面または汎用性における問題を抱えている．しかし，現実には，高い汎用性を持つ Up*/Down* ルーティングが，性能面における問題点を抱えつつも，イレギュラーネットワーク向けのルーティングアルゴリズムとして一般的に利用されている．そこで，本研究では，Up*/Down* ルーティングと同等の高い汎用性を実現し，Up*/Down* ルーティングにおけるトラフィック集中の問題を解決する適応型ルーティングアルゴリズムである L-turn および R-turn ルーティングを提案する．

L-turn および R-turn ルーティングは，高い汎用性を実現するために，Up*/Down* ルーティングと同様に，スパニングツリーと有向グラフの構築をベースとしている．しかし，2 次元に拡張された有向グラフを導入し，2 次元 Turn モデルを適用することにより，Up*/Down* ルーティングにおけるトラフィック集中の問題改善を図っている．

3.2　SAN の実現例

本節では，Up*/Down* ルーティングおよび第 4 章で提案する L-turn および R-turn ルーティングの適用対象となる代表的な SAN の実現例として，Autonet, Myrinet, QsNET, InfiniBand および RHiNET について説明する．

3.2.1　Autonet

Autonet [Mae91, RS91] は，10 Mbps のイーサネットに代わる，より高速かつ実用的な LAN の実現を目的として開発されたネットワークである．Autonet では，高性能，高可用性および耐故障性を実現するために，SAN の基本となる様々な技術が用いられている．

各スイッチ間は，バス接続ではなく，より高速な 100Mbps の 全二重 point-to-point リンクで接続される．各スイッチは，12 ポートのクロスバを持ち，低レイテンシ転送の実現のため，cut-through 方式によるパケット転送が行なわれる．リンク長は，同軸ケーブルで 100 m, 光ファイバで 2 km までサポートしており，バッファオーバフローの発生を防ぐために，受信 FIFO バッファが半分以上埋まった時に，送信側に対してパケット転送停止の信号を送る start-stop フロー制御を利用している．

Autonet は，任意のトポロジをサポートしており，トポロジの状態を定期的に監視することにより，スイッチやリンクの状態が変化 (追加，故障など) した際に，自動的に再構成 (トポロジ情報の取得およびルーティングテーブルの更新など) を行なう．これにより，高い可用性と耐故障性が実現されている．トポロジ情報の取得やルーティングテーブルの更新は，各スイッチの制御用プロセッサ上で実行される Autopilot と呼ばれるソフトウェアにより行なわれる．任意のトポロジをサポートするために，ルーティングアルゴリズムは，分散ルーティング方式による 適応型の Up*/Down* ルーティングが用いられている．これにより，複数経路を利用した効率的なパケット転送が可能となっている．

3.2.2　Myrinet

Myrinet [N.J95, Myra] は，Myricom 社により開発された現在の主要な SAN の 1 つであり，高性能，高可用性を要求する PC クラスタを中心に広く用いられている．

現在の Myrinet の主要バージョンである Myrinet-2000 は，2Gbps の高速な point-to-point リンクにより相互接続された 16 ポートまたは 32 ポートのクロスバを持つスイッチから構成される．パケット転送方式として高速な WH 方式を用い，また，任意のトポロジおよび自動的な再構成をサポートしている．ルーティングアルゴリズムは，ソースルーティング方式による Up*/Down* ルーティングが用いられる．ソースルーティング方式であるため，パケットは途中スイッチで動的な経路選択を行なうことはできないが，出発地，目的地間に複数の経路が存在する場合は，出発地スイッチにおいて経路選択を行なうことができる．

Myrinet では Glenn's Messages (GM) および Myrinet Express (MX) と呼ばれるローレベルメッセージパッシングシステムが用意されており，これにより，トポロジ情報の取

得およびルーティングテーブルの計算，セキュアなユーザレベルゼロコピー通信[2]および高信頼性のメッセージパッシング，などが実現されている．

Myrinet-2000 のネットワークインタフェースは，LANai X [Myrb] と呼ばれる制御用チップを備えている．LANai X には，LANai コア (プロセッサとパケットインタフェースを持つ)，X-port と呼ばれるマルチプロトコルポート，ローカルバス，PCI-X インタフェースなどが実装されており，プロセッサ上で動作する Myrinet Control Program (MCP) により GM API の処理などが行なわれる，

3.2.3　QsNET

QsNET[PFH01, FFA+02] は，Quadrics 社により開発された PC クラスタ向けの主要な SAN の 1 つである．

現在の主要バージョンは，QsNET II であり，8.5Gbps の高速な point-to-point リンクにより相互接続された 8 ポート (それぞれ 2 つの仮想チャネルを持つ) の Elite4 スイッチから構成される．パケット転送方式として，WH 方式を用い，トポロジは Fat ツリーだけをサポートしている．このため，Fat ツリー上を階層構造に沿って，単純に，出発地の葉スイッチから up し，目的地の葉スイッチまで down するルーティングアルゴリズム[3]が用いられ，Myrinet と同様に，ソースルーティング方式により実装されている．

QsNET II のネットワークインタフェースは，Elan4 と呼ばれる通信制御用プロセッサを持つ．Elan4 は，64bit の RISC プロセッサ，DMA エンジン，MMU (メモリマネージメントユニット)，32kbyte キャッシュメモリ，PCI-X および SDRAM インタフェース，などにより構成される．Elan4 は，高位の通信ライブラリをホストプロセッサの介在無しに高速に処理するなどの，低レイテンシ，高バンド幅通信のための様々な機能を備えている．

3.2.4　InfiniBand

InfiniBand[I.T04] は，PC クラスタにおける PC 間通信，およびサーバクラスタにおけるサーバ I/O 間通信などにおける利用を目的として標準化された高性能 I/O ネットワークである．InfiniBand の標準化は，IBM, Intel, Hewlett-Packard, Microsoft などの多数の企業の参加により設立された InfiniBand Trade Association (IBTA) により進められており，プロプライエタリな Myrinet や QsNET などと異なり，オープンな規格となっているのが大きな特徴である．

現在の InfiniBand の規格は，2004 年 10 月に発表された InfiniBand Architecture (IBA) Specification 1.2 であり，Voltaire 社[4], SilverStorm Technologies 社[5]などにより製品化が行なわれている．

[2] PC 間の通信の際，通常，ホスト内ではシステムコールを介してカーネルが主記憶からネットワークインタフェースへ複数回のコピーを通してデータを転送する．これに対し，システムコールなどのオーバヘッドを避けるために，ユーザプロセスが直接ネットワークインタフェースにアクセスして通信を行う方法をユーザレベル通信と呼ぶ．また，ホスト内のデータコピーの回数を減らすためにネットワークインタフェースが主記憶のデータを直接読み書きする方法をゼロコピー通信と呼ぶ．ユーザレベルゼロコピー通信とはユーザレベルで実現するゼロコピー通信のことである．

[3] Up*/Down* ルーティング，L-turn および R-turn ルーティングによりエミュレートすることが可能

[4] http://www.voltaire.com/

[5] http://www.silverstorm.com/

InfiniBand は，他の SAN と同様に，point-to-point リンクで相互接続されたスイッチベースのネットワークである．リンクあたりのデータ転送レートは，2Gbps であり，4 本または 12 本のリンクを並列に利用することにより，8Gbps または 24Gbps に拡張することが可能である．また，IBA 1.2 では，DDR(Double Data Rate) および QDR(Quad Data Rate) により更に 2 倍，4 倍となるデータ転送レートが実現されている．

InfiniBand ネットワークの構成単位はサブネットと呼ばれ，サブネットは，エンドノード (PC または I/O デバイス)，スイッチ，スイッチ間リンク，サブネットマネージャにより構成される．また，エンドノードとリンク間のインタフェースは Channel Adapter(CA)[6] と呼ばれ，各 CA の各ポートと各スイッチに対して，サブネット内のルーティングに用いられる Local Identifiers (LID) がサブネットマネージャにより割当てられる．

InfiniBand では，任意のトポロジが選択可能であり，ルーティング (実装はベンダ依存) は，各スイッチが持つルーティングテーブルを参照する分散ルーティング方式となる．また，パケット転送方式としては cut-through 方式が用いられる．ただし，利用経路は，目的地 CA ポートの LID により一意に定まるため，固定型ルーティングとなり，Myrinet, QsNET と同様に，出発地 CA においてだけ複数経路が選択可能となる．

InfiniBand においても Up*/Down* ルーティングは適用可能であるが，途中スイッチにおける出力チャネル決定が，目的地の LID だけをインデックスとして行なわれるため，Autonet や Myrinet と異なり，そのままでは適用することはできない．このため，(1) 最短経路の割合をある程度犠牲にする方式 [JAJ01]，もしくは，(2) destination renaming[7] を実装する方式 [PJJ01] などを用いて適用が可能となる．

InfiniBand では，仮想チャネルに相当する最大 15 本の仮想レーンをデータトラフィックに使用することができる．仮想レーンは，Quality of Service (QoS), トラフィッククラスの分離などの他に，デッドロック回避のためにも利用可能となっている．このため，仮想チャネルを必要とするルーティングアルゴリズムを実装することも可能となっている．

3.2.5　RHiNET

RWCP High Performance Network (RHiNET)[TSJ+99, 西 宏 00, STH+00, NKN+01] は，RWCP，日立製作所および慶應義塾大学天野研究室により開発されたネットワーク[8] であり，高速な光インターコネクトと高速なスイッチにより，商用の SAN に匹敵する高性能，高信頼性通信を実現している．RHiNET は，マシンルーム内の PC 間接続だけでなく，オフィス，もしくはビルのフロア内の PC 間接続に焦点を当てている．RHiNET の実装としては，これまでに，RHiNET-1, RHiNET-2 および RHiNET-3 が開発されている．ここでは，RHiNET-3 について述べる．

RHiNET-3 は，ネットワークインタフェースのコントローラである Martini と高速な光リンクで相互接続された RHiNET-3/SW スイッチにより構成される．RHiNET-3 では，任意のトポロジが利用可能であり，ルーティングは，固定型の構造化チャネル法が用いら

[6] PC の NIC に相当する Host CA (HCA) と I/O デバイスの NI に相当する Target CA (TCA) に分類される

[7] 経路選択を柔軟に行うために，同一の目的地に対して複数の識別子を与え，経路制御をおこなう方法

[8] 提案者は SAN をマシンルームなどで，トポロジに制限を与え，短いリンク長で集中配線したネットワークであると狭義し，RHiNET がこの SAN と LAN の特徴を持つという点で Local Area System Network (LASN) と呼んでいる．

れる．このために，リンクあたり 32 本の仮想チャネルが用意されている．ルーティングとしては，Up*/Down* ルーティングを適用することも可能である．また，ルーティング方式は，分散ルーティング方式とソースルーティング方式の両方をサポートしている．

RHiNET-3/SW は，0.14 μm CMOS エンベッデッドアレイで構成される 1 チップスイッチであり，高速な 10 Gbps のリンクバンド幅を持つ．また，リンクレベルのエラー検出と修正，再送機構を搭載し，エラーレートの高い安価な媒体を用いた場合にもハードウェアのレベルで信頼性を確保し，通信のソフトウェアオーバヘッドの削減を実現する．1km のリンク長をサポートするため，フロー制御として credit based 方式を採用している [NKN+01]．

Martini は，ユーザレベルゼロコピー通信をサポートするためにユーザメモリ領域のプロテクション，アドレス変換機構などの機能をすべてハードウェアで高速に処理する．また，ハードウェアで実装されていない通信処理をコアプロセッサのソフトウェアで実現するといった高い柔軟性も併せ持つ．

3.2.6　SAN の実現例のまとめ

最後に，上記 5 つの SAN の実現例についてまとめたものを，表 3.2 に示す．

表 3.2: 既存の SAN の比較

	Autonet	Myrinet	InfiniBand	QsNET	RHiNET
トポロジフリー	yes	yes	yes	no	yes
仮想チャネル利用可	no	no	yes	yes	yes
ルーティングアルゴリズム	適応型	固定型	固定型	固定型	固定型
ルーティング方式	分散	ソース	分散	ソース	分散/ソース
Up*/Down*適用可	yes	yes	yes	-	yes
L-turn/R-turn 適用可	yes	yes	yes	-	yes

L-turn および R-turn ルーティングは，表 3.2 に示すように，理論上，イレギュラーネットワークをサポートする既存の SAN において適用可能である．これは，前述の通り，L-turn および R-turn ルーティングは，Up*/Down* ルーティングと同等の高い汎用性を持つため，Up*/Down* ルーティングが適用可能なネットワークであれば，同様に適用可能となるためである．なお，QsNET で用いられる Fat ツリー上のルーティングについても，Up*/Down* ルーティング，L-turn および R-turn ルーティングによりエミュレートが可能であるが，これらを適用する意義に欠けるため，対象外としている．

第4章　L-turn/R-turn ルーティング

第 3.1.5 節で述べたように，Up*/Down* ルーティングにおける主な問題点は，次の 2 点である．

- 禁止ターンが形成されるスイッチ上では，常に禁止ターンのペアが発生するためトラフィックの偏りが発生しやすい
- 禁止ターン選択に自由度がないため禁止ターン分散の余地がない

これらの問題点は，1 次元有向グラフにおける次の 2 つの特性に起因している．

- up と down の 2 つの方向だけが存在
- パケットの移動に伴ない発生するターンが 2 つだけに限定される

そこで，本章では上記の問題点を改善するために，Up*/Down* ルーティングで利用されている 1 次元有向グラフを拡張して H/V (Horizontal/Vertical) グラフと呼ばれる 2 次元有向グラフを導入する [AMAH02, 上樂 03]．H/V グラフの導入により，形成可能なターンの数は従来の 6 倍である 12 パターンに増加するため，トラフィックの分散を考慮した，より柔軟な禁止ターンの選択を行なうことが可能となる．

H/V グラフに基づいてトラフィック分散を考慮したルーティングアルゴリズムを設計する際には，次の 2 つの条件を満たすことが重要となる．

- 必要最低限の禁止ターンだけを選択して，デッドロックフリーを実現する
- 可能な限り，分散された禁止ターン集合を選択する

H/V グラフの導入により，トラフィック分散実現のためのより柔軟な禁止ターンの選択を行なう余地が生まれるが，一方で，1 次元有向グラフに比べてより多くの複雑な循環構造が形成されるため，その識別が難しくなるという問題も生じる．このため，H/V グラフに対して上記 2 つの条件を考慮した禁止ターン選択を行なう方法は，極めて直観的な選択による Up*/Down* ルーティングに比べて非常に複雑なものとなる．そこで，本章では，2 次元 Turn モデルの適用によるシステマティックな手法を用いてそのような禁止ターン選択を行ない，トラフィックの分散を図る L-turn ルーティングおよび R-turn ルーティングと呼ぶ適応型デッドロックフリールーティングアルゴリズムを提案する．L-turn および R-turn ルーティングは，仮想チャネルやバッファなどの付加的なハードウェアを必要としないため，Up*/Down* ルーティングが適用可能なすべての SAN において適用可能となっている．

以降，第 4.1 節 で H/V グラフの構築手順を示し，第 4.2 節で，2 次元 Turn モデルの適用により L-turn および R-turn ルーティングを導出するシステマティックな手法を示す．

そして，第 4.3 節で，L-turn および R-turn ルーティングの開発における研究過程と筆者が担当した作業内容を示し，本論文の位置付けについて述べる．

4.1　H/V グラフの構築

horizontal direction と vertical direction の 2 つの方向を持つ 2 次元有向グラフである H/V グラフは，次の 3 つのステップにより構成される．

(1) BFS スパニングツリーを構築する．

(2) 各スイッチに 2 次元座標を割当てる．

(3) 各チャネルに 2 次元方向を割当てる．

4.1.1　BFS スパニングツリーの構築

まず，Up*/Down* ルーティングと同様にして，BFS スパニングツリーの構築を行なう．そして，構築した BFS スパニングツリーの階層構造を反映した座標である depth を各スイッチに対して割り当てる．depth は，各スイッチのルートスイッチからの最短距離を示し，各チャネルの垂直軸における方向 vertical direction (up および down) の決定に用いられる．ルートスイッチ自身の depth は 0 であり，同階層の各スイッチに対しては，それぞれ同一の depth が割当てられる．

定義 1 (depth) 各スイッチのルートスイッチからの最短距離を *depth* とする．

例として，9 スイッチ構成のイレギュラーネットワークに対する depth の割当てを図 4.1 に示す．図の実線と破線はそれぞれスパニングツリーを構成するリンク (tree link) とそれ以外のリンク (outer link) を示している．

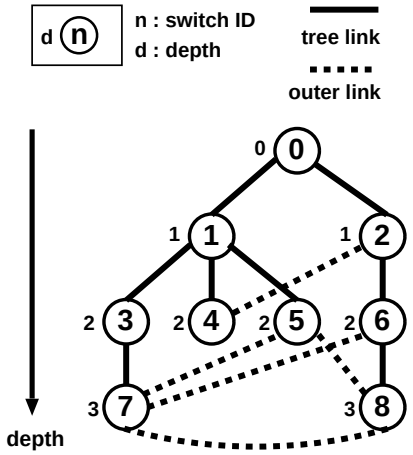


図 4.1: depth の割当て

4.1.2　各スイッチへの 2 次元座標の割当て

次に，2 次元有向グラフを構築するための拡張として，depth に加えて各スイッチに対し horizontal spread を割当て，水平軸における方向 horizontal direction (left および right) の概念を導入する．

horizontal spread は，構築したスパニングツリー上で，前順走査[1]を行なったときの訪問順序であり，走査における訪問順にしたがって，0 から始まる昇順の値が各スイッチに割当てられる．

定義 2 (horizontal spread) スパニングツリー上で，ルートスイッチを起点とした前順走査を行ない，訪問順にしたがって各スイッチに割当てる *0* から始まる昇順の値を *horizontal spread* とする．

horizontal spread を前順走査により割当てている理由は，スパニングツリーにおける経路保証と各スイッチに対する一意の depth および horizontal spread の組合せ (2 次元座標) の割当てを実現するために，次の 2 つの条件を満たす必要があるためである．

(a) 子スイッチの horizontal spread が常に親スイッチよりも大きい．

(b) 各スイッチの horizontal spread は互いに異なる．

前者の条件により，depth と同様にして，horizontal spread が小さくなる方向に進むことにより任意のスイッチからルートスイッチに到達可能となり，また，大きくなる方向に進むことによりルートスイッチから任意のスイッチへ到達可能であることが保証される．一方，後者の条件により，depth と horizontal spread の組合せが同一となるスイッチが存在しないことが保証される (同じ depth を持つスイッチは存在するが，同じ horizontal spread を持つスイッチは存在しない)．

horizontal spread は，図 4.2 に示すように，直観的には，スパニングツリー上の水平方向における座標を表すものであり，各チャネルの horizontal direction および，同じ depth を持つスイッチ間の vertical direction の決定に用いられる．

以上より，各スイッチに対して horizontal spread (h) と depth (d) の組合せから成る一意の 2 次元座標 (h, d) を割当てることが可能となる．

例として，図 4.1 のネットワークに対する horizontal spread および 2 次元座標の割当ては，図 4.2 の通りとなる．

前順走査においては，次に訪問するスイッチとして 2 つ以上の子スイッチが選択可能となる場合があるため，複数の選択ポリシーが適用可能である．このため，同一ネットワークに対する horizontal spread の割当て方は複数通り存在し，これにより異なる有向グラフが構築されうる．訪問スイッチの選択ポリシーについては，第 4.2.5 節で説明する．

[1] ルートスイッチを起点として，スパニングツリー上の各スイッチを一つずつ訪問することを走査と呼ぶ．親スイッチを訪問してから，子スイッチを順番に訪問する走査を前順走査 [石畑 89] と呼ぶ．

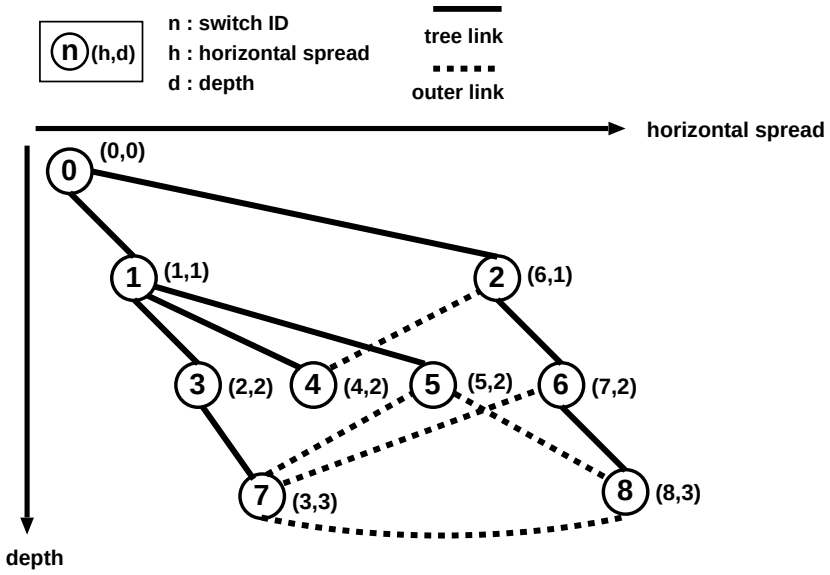


図 4.2: depth と horizontal spread の割当て

4.1.3　各チャネルへの 2 次元方向の割当て

各スイッチに割当てられた 2 次元座標を基に，各チャネルに対する horizontal direction と vertical direction の割当てを行ない，H/V グラフの構築に必要となる H/V direction の割当てを行なう．

まず，次のようにして horizontal direction (left/right) を各チャネルに割当てる．

定義 3 (horizontal direction) 座標 (x_s, y_s) から座標 (x_d, y_d) に向かうチャネルにおいて，次のように *horizontal direction* を定める．

(a) $x_s > x_d$, ならば *left* 方向を割当て，

(b) $x_s < x_d$, ならば *right* 方向 を割当てる．

次に，同様にして vertical direction (up/down) を各チャネルに割当てる．

定義 4 (vertical direction) 座標 (x_s, y_s) から座標 (x_d, y_d) に向かうチャネルにおいて，次のように *vertical direction* を定める．

(a) $(y_s > y_d) \vee ((y_s = y_d) \wedge (x_s < x_d))$, ならば *up* 方向を割当て，

(b) $(y_s < y_d) \vee ((y_s = y_d) \wedge (x_s > x_d))$, ならば *down* 方向を割当てる．

そして，各チャネルに対して，次のように 4 つの方向から成る H/V direction を割当てる．

定義 5 (H/V direction) *H/V direction* **は，***horizontal direction (h)* **と** *vertical direction (v)* **の組み合わせ** $HV(h,v)$ **により次のように定める．**

(a) $HV(left, up)$ **に対し** *left-up(LU)* **方向 を割当て，**

(b) $HV(left, down)$ **に対し** *left-down(LD)* **方向 を割当て，**

(c) $HV(right, up)$ **に対し** *right-up(RU)* **方向 を割当て，**

(d) $HV(right, down)$ **に対し** *right-down(RD)* **方向 を割当てる．**

以上をまとめたものを表 4.1 **に示す．**

表 4.1: H/V direction **の定義**

	$x_s > x_d$	$x_s < x_d$
$y_s > y_d$	LU	RU
$y_s = y_d$	LD	RU
$y_s < y_d$	LD	RD

本論文では，以降，H/V direction *dir* **を持つチャネルを** *dir* **チャネルと呼ぶ．**
各チャネルに対して H/V direction **を割当てることにより，**2 **次元有向グラフである** H/V **グラフが構築される．例として，図** 4.1 **におけるネットワークの** H/V **グラフは，図** 4.3 **の通りとなる．**

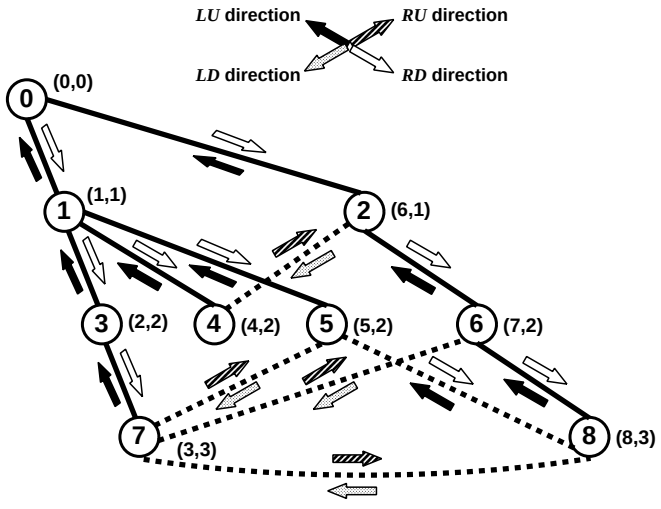


図 4.3: H/V **グラフ**

H/V **グラフにおいて，スパニングツリーを構成するチャネルだけで構成される部分グラフを** H/V **ツリーと呼ぶ．**

なお，同 depth スイッチ間チャネルの割当てを表 4.1 のように定めている理由は，一意の方向割当てを行なうことにより，第 3.1.2 節で述べた BFS Up*/Down* ルーティングにおいて問題となる同階層スイッチ間の冗長な禁止ターンの発生を抑制することと，禁止ターンの分散を実現しやすくするためである．後者については，第 4.2.4 節で説明をする．

4.2　2 次元 Turn モデルによるルーティングアルゴリズムの設計

以下，H/V グラフに対する 2 次元 Turn モデルの適用により，L-turn および R-turn ルーティングを導出するシステマティックな設計手順を説明する．

4.2.1　Turn モデル

デッドロックが生じるのは， 結合網内のチャネルバッファが論理的な循環構造を作ってしまうためということを第 2.2.3 節で述べた．Glass らによる Turn モデル [GN92] は，パケットがルーティング中に行なう方向転換 (ターン) のパターンを制限して，循環構造の形成を防ぐことにより，デッドロックフリールーティングアルゴリズムを設計する方法である．このモデルは，メッシュやトーラスなどのレギュラーネットワークへの適用を念頭に置いて提案されているが，パケットのターンによって形成される論理的な循環構造の除去に着目しているため，結合網のトポロジに依存することがない．このため，イレギュラーネットワークを含む任意のトポロジに対して適用可能となっている[2]．

Turn モデルによるデッドロックフリールーティングアルゴリズムの設計は，次に示す手順に従って行なわれる．ここで，チャネルは物理チャネル，仮想チャネルのいずれの場合においても適用可能である．なお，ここで述べる手順の一部 (ラップアラウンドチャネルや 180 度のターンの考慮) は，メッシュおよびトーラスへの適用を念頭に置いたものとなっているが，他の任意のトポロジを対象とする場合でも，適宜調整をした上で，ほぼ同様の手順が適用可能である．

(a) パケットが転送される方向に基づいてチャネルを分類する．各スイッチが 1 つの物理的な方向に対し v 個のチャネルを持つ場合，これらは，v 個の論理的な方向として区別する．

(b) ある方向から別の方向へのターンのすべてのパターンを識別する．0 度と 180 度のターン[3]は無視する (0 度のターンは，複数の仮想チャネルを利用する場合だけ考慮する)．

(c) ターンの連鎖によって構成されるすべての循環構造を識別する．一般的には，トポロジ上の各平面における最も単純な循環構造をそれぞれ識別すればよい．

[2]イレギュラーネットワークへの適用においては，基本的に有向グラフを構築する必要がある

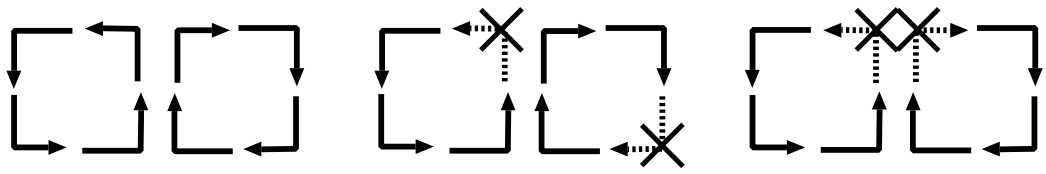
[3]仮想チャネルを切り替えて同一方向に移動する際に発生するターンを 0 度のターンと呼ぶ．一方，ある方向から正反対となる方向に移動する際に発生するターンを 180 度のターンと呼ぶ．ここで，方向とは論理的な方向を指す．

(d) 識別されたすべての循環構造を防ぐために，各循環構造について 1 つのターンを禁止し，最低限必要なだけのターンを禁止する．循環構造はいくつかの循環の複合で生じる場合があるので，禁止するターンは慎重にチェックして決めなければならない．

(e) トーラスの場合は，ネットワークの端で折り返しを行うラップアラウンドチャネルが存在するが，ラップアラウンドチャネルを使ったターンも，循環構造を形成しないようにした上で可能な限り許可する．

(f) 0 度あるいは 180 度のターンを，循環構造を形成しないようにした上で，可能な限り許可する．

例として，2 次元メッシュにおける Turn モデルの適用を考える．2 次元メッシュにおいて可能な循環構造は，図 4.4(a) に示す 2 種類になる．ここで，Turn モデルに従い，各循環構造に含まれるターンを 1 つずつ禁止し，デッドロックを防いだ場合を図 4.4(b) に示す．図において，点線のターンが禁止されている．図の禁止ターンに基づくルーティング方法は，先に西方向にパケットを送ることから West-first ルーティングと呼ばれる．デッドロックを防ぐための禁止ターンの選択肢は 1 つではなく， たとえば図 4.4(c) に示す切り方 (North-last ルーティング) も選択可能である．



(a)　(b) West-first　(c) North-last

図 4.4: Turn モデル (2 次元メッシュ)

しかし，どのような選択をしてもよいというわけではない．図 4.5 に示すように禁止ターンを定めると，防いだはずの循環構造が，8 の字型に複合して新たな循環構造が生じてしまう．従って，このような状況を配慮しつつ禁止ターンを選択する必要がある．

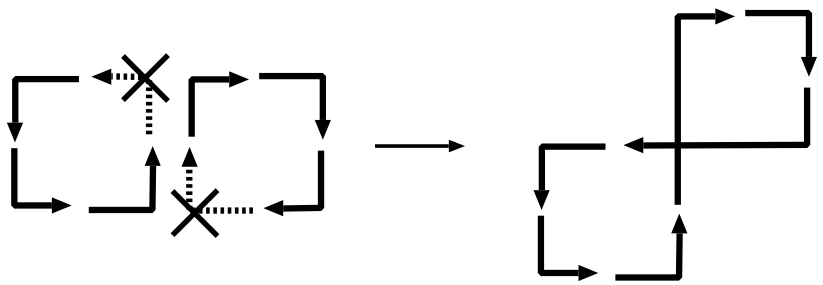


図 4.5: Turn モデル (2 次元メッシュ) における失敗した切り方

2 次元メッシュにおける一般的な固定ルーティングである e-cube ルーティング [DS87] を Turn モデルで考えると，図 4.6 に示すように，この循環構造のうち 4 つのターンしか許していないことになる．この場合，確かに循環構造を防ぐことができるのでデッドロックを生じないが，制限のし過ぎで代替経路を失ってしまうことがわかる．

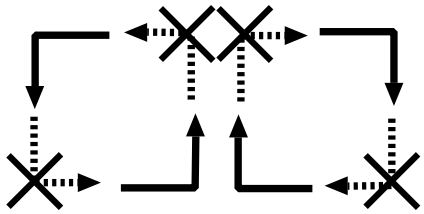


図 4.6: e-cube ルーティングの Turn モデル (2 次元メッシュ)

Turn モデルにより生成された West-first または North-last ルーティングを使えば，パケットは 6 通りのターンを行なうことが許されるため，図 4.7 に示すような故障箇所や混雑箇所を迂回する適応型ルーティングが可能となる．

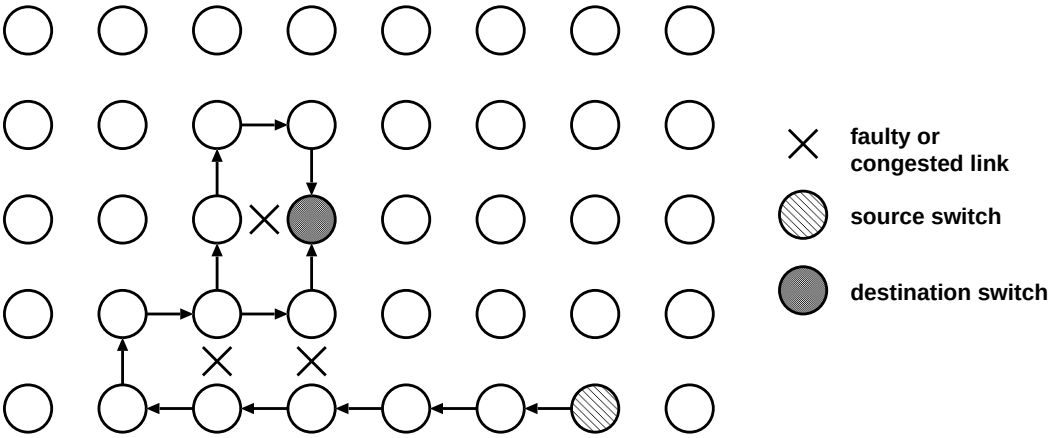


図 4.7: West-first ルーティングによる故障 (混雑) 箇所の回避

先に述べたように，Turn モデルは，有向グラフの構築をベースとすることにより，イレギュラーネットワークに対しても適用可能であり，Up*/Down* ルーティングは，up および down の 2 つの方向を持つ最も単純な 1 次元の Turn モデルの適用例として考えることができる (この場合，180 度のターンが識別対象となる)．しかし，Turn モデルの適用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方法に関する研究は，本研究がなされるまでほとんど行なわれておらず，2 次元の Turn モデルの適用例は本研究が初めてである．

4.2.2　H/V グラフにおける Turn モデルの適用手順

前節で述べたように，各チャネルに対して方向を割当てた後の Turn モデルによるデッドロックフリールーティング設計手順は，大きく分けて，次の 4 つのステップから成る

STEP1 パケット転送時に形成可能なすべてのターンを識別する．

STEP2 識別されたターンの連鎖により形成される循環構造をすべて識別する．

STEP3 識別されたすべての循環構造除去のために，最低限必要となる禁止ターン集合を選択する．

STEP4 循環構造の形成に関与しないターンを可能な限り許可する．

H/V グラフでは，パケット転送のために，180 度のターンが必要となるため，STEP1 において，180 度のターンについても識別するものとする．仮想チャネルを利用しないため，0 度のターンについては考慮する必要がない．

H/V グラフでは，180 度のターンを含むことにより，複雑なターンの連鎖による循環構造が多数形成されるため，直観的な循環構造の識別が可能である 2 次元メッシュおよびトーラスなどのレギュラーネットワークに比べて，すべての循環構造の識別と適切な禁止ターンの選択による循環の除去 (上記の STEP2 と STEP3 に相当) が難しい．そこで，厳密かつ効率的にすべての循環構造の検出と除去を行なうために，上記の STEP2 と STEP3 の手続きを次のように融合して適用する．

(a) H/V グラフ上で最も単純な循環構造を識別する

(b) (a) で検出された各循環構造を除去するために，最低限必要となる禁止ターン集合を選択する．

(c) (b) で選択された禁止ターン集合を除く残りのターンにより形成可能なすべての循環構造を識別する．

(d) (c) で検出された各循環構造を除去するために，最低限必要となる禁止ターン集合を選択し，残りすべての循環構造を除去する．

なお，トラフィックの均等な分散を実現するために，上記手順において，可能な限り，禁止ターンの分布が分散されるように，禁止ターンの選択を行なう．

更に，STEP4 において，特定の循環構造を探索により検出するアルゴリズムを適用することにより，冗長な禁止ターンを除去してルーティングの自由度の低下を抑え，より均等なトラフィック分散の実現を図る．

H/V グラフにおける Turn モデル適用手順を表 4.2 にまとめる．

以下，上記の手順に従って, L-turn および R-turn ルーティングを設計するングを設計する手順を具体的に説明する．

表 4.2: H/V グラフにおける Turn モデル適用手順

STEP1	パケット転送時に形成可能なすべてのターンを識別する．	
STEP2	形成可能なすべての循環構造の識別とそれらを除去するために最低限必要な禁止ターンを選択する．	
	(a)	H/V グラフ上で最も単純な循環構造を識別する．
	(b)	(a) で検出された各循環構造において，最低限必要な禁止ターンを選択し循環構造の除去を行なう．
	(c)	(b) で選択された禁止ターン集合を除いた残りのターンにより形成可能なすべての循環構造を識別する．
	(d)	(c) で検出された各循環構造において，最低限必要な禁止ターンを選択し残りすべての循環構造の除去を行なう．
STEP3	特定の循環構造を探索により検出するアルゴリズムを適用して冗長な禁止ターンを除去する．	

4.2.2.1　準備

最初に，以降で用いられる表記を次に示す．

定義 6 (ターン) スイッチ到着時のパケット転送方向 p_dir とスイッチ通過後のパケット転送方向 n_dir により形成されるターンを T_{p_dir,n_dir} と表す．

定義 7 (ターン連鎖) ターン T_i を伴なうパケット転送直後に，ターン T_j を伴なう転送が可能である場合のターンの連鎖を $TD(T_i, T_j)$ と表す．

定義 8 (循環構造) H/V グラフにおいて，$\{TD(T_i, T_j) \mid j = (i+1) \bmod n, i = 0, 1, ..., n-1\}$ のターン連鎖を形成する n 個のターンの集合 $\{T_0, T_1, ..., T_{n-1}\}$, により循環構造が形成される場合，その循環構造を $C(T_0, T_1, ..., T_{n-1})$ と表す．

4.2.2.2　ターンの識別 (STEP1)

H/V グラフにおいて，ある H/V direction へ移動した後，その他の H/V direction へ移動した際に形成可能なすべてのターンを図 4.8 に示す．先に述べたように，0 度のターンは無視している．図 4.8 より，H/V グラフでは 4 つの H/V direction が存在するため，形成可能なターンは全部で 12 パターンとなる．

4.2.2.3　循環構造の識別と禁止ターンの選択 (STEP2)

図 4.8 に示したターンの連鎖により形成されるすべての循環構造の識別とその除去に必要な禁止ターンの選択を，第 4.2.2 節で示した手順に基づいて行なう．

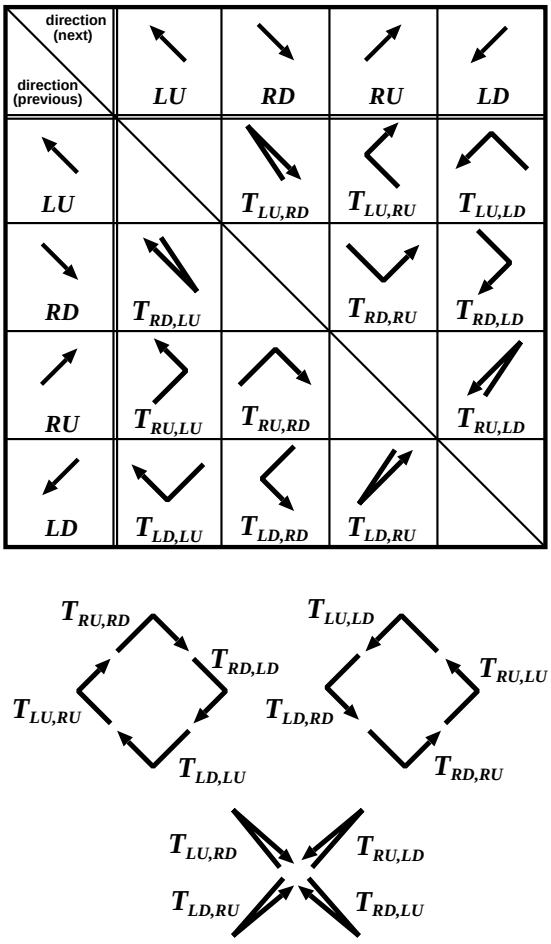


図 4.8: H/V **グラフにおける形成可能ターン**

STEP2-(a) 循環構造の検出 (1 回目)

スパニングツリーベースの有向グラフでは，ツリー構造の部分グラフ (2 **つ以上の** tree link **で接続された** 3 **つ以上のスイッチから成る**) **に** 1 **つの** outer link **を追加することにより常に循環構造が形成される．例として，図** 4.9 **は，**2 **つの** tree link **と** 1 **つの** outer link **から構成される最も単純な循環構造を表している．**

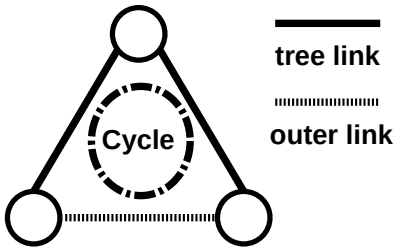
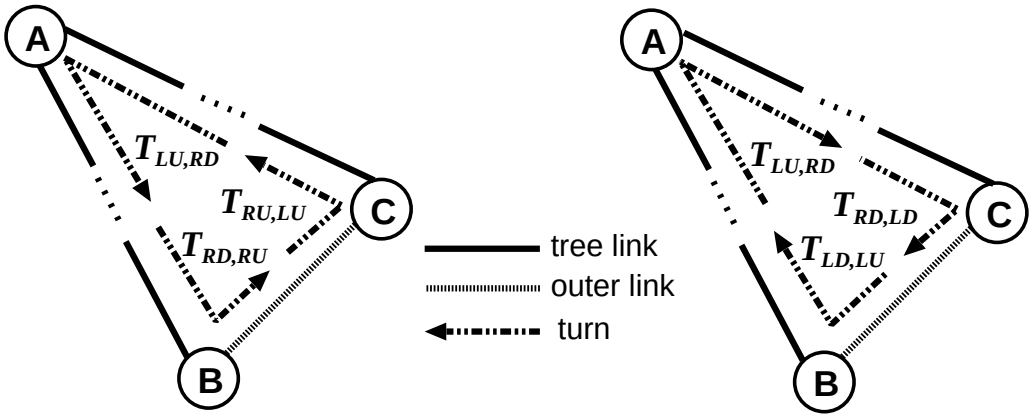


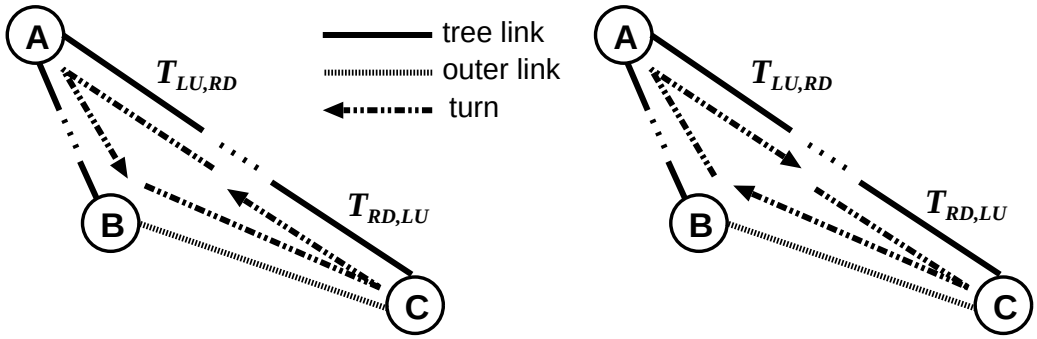
図 4.9: **スパニングツリーベースの有向グラフにおける最も単純な循環構造**

H/V グラフにおいて，図 4.9 に相当する部分グラフは，図 4.10 と図 4.11 に示す 2 つの部分グラフに分類される．2 つの部分グラフの違いは，子スイッチであるスイッチ B とスイッチ C の垂直方向における相対位置の違いである．



(a) $C_1(T_{LU,RD}, T_{RD,RU}, T_{RU,LU})$　　(b) $C_2(T_{LU,RD}, T_{RD,LD}, T_{RU,LU})$

図 4.10: H/V グラフにおける循環構造 (C_1, C_2)



(a) $C_3(T_{LU,RD}, T_{RD,LU})$　　(b) $C_3'(T_{LU,RD}, T_{RD,LU})$

図 4.11: H/V グラフにおける循環構造 (C_3, C_3')

図 4.10 と図 4.11 において，各部分グラフには，互いに反対方向となる 2 つの循環構造がそれぞれ形成される．図 4.10(a) における循環構造は，$C_1(T_{LU,RD}, T_{RD,RU}, T_{RU,LU})$ であり，図 4.10(b) における循環構造は，$C_2(T_{LU,RD}, T_{RD,LD}, T_{LD,LU})$ である．同様に，図 4.11(a) における循環構造は，$C_3(T_{LU,RD}, T_{RD,LU})$ であり，図 4.11(b) における循環構造は，$C_3'(T_{LU,RD}, T_{RD,LU})$ となる．ただし，循環構造 C_3 と C_3' は論理的に同一であるため，C_3 だけを考慮すればよい．

STEP2-(b) 禁止ターンの選択 (1 回目)

STEP2-(a) で識別した 3 つの循環構造の形成を防ぐために，各循環構造内の 1 つのターンを次のポリシーに基づいて禁止する．

(a) ターン $T_{LU,RD}$ を禁止しない．

(b) 可能な限り，選択した禁止ターンの組合せにより，図 4.12 のような同一スイッチ上の禁止ターンのペアが発生しないようにする．

上記の (a) において，ターン $T_{LU,RD}$ を禁止しない理由は，H/V グラフ上の任意のスイッチ間の経路を保証するために，任意のスイッチから LU 方向の tree channel を 0 回以上辿って任意の目的地スイッチの祖先[4]となるスイッチに到達可能であり，かつ，祖先スイッチに到達後に，RD 方向の tree channel を 0 回以上用いて任意の目的地スイッチに到達可能である，という条件を満たす必要があるためである．

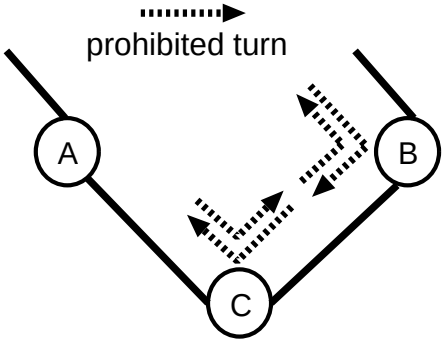


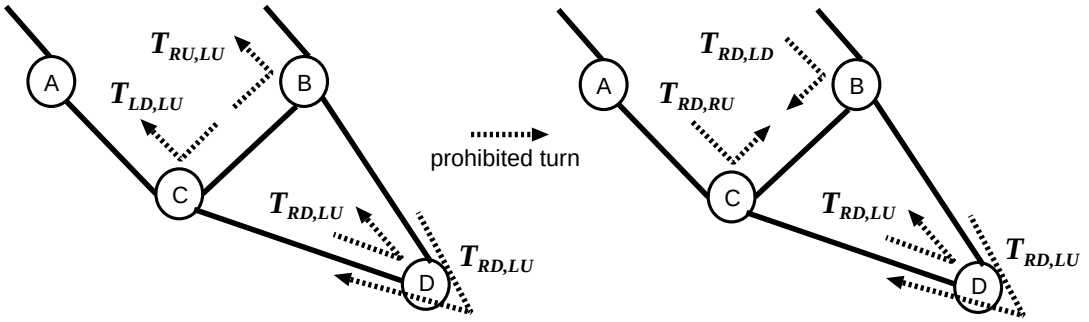
図 4.12: H/V グラフにおける禁止ターンの偏り

上記の (a),(b) のポリシーを考慮すると，循環 C_1 および C_2 を破るために禁止するターンの集合は，$\{T_{RU,LU}, T_{LD,LU}\}$ または $\{T_{RD,RU}, T_{RD,LD}\}$ となり，循環 C_3 を破るための禁止ターンは $T_{RD,LU}$ となる．したがって，図 4.10 および図 4.11 に示したすべての循環構造を破るために禁止するターン集合は，2 通り選択可能となる．これらを次のように定める．

$$\begin{aligned} P_1 &= \{T_{LD,LU}, T_{RU,LU}, T_{RD,LU}\}, \\ P_2 &= \{T_{RD,RU}, T_{RD,LD}, T_{RD,LU}\} \end{aligned}$$

図 4.13(a),(b) に，禁止ターン集合 P_1 および P_2 による禁止ターン分布の例を示す．図より，循環構造 C_1 および C_2 を破るために選択した禁止ターンについて，分散が実現されていることがわかる．一方，循環構造 C_3 を破るための禁止ターン $T_{RD,LU}$ については，偏りが発生してしまうが，C_3 を破るためには，これ以外に選択肢が無いので，この場合はやむをえないものとする．

[4]ここでは，スイッチ間の関係を親族関係の用語を用いて説明している．あるスイッチから見て，親，親の親，... となるスイッチをまとめて祖先と呼ぶ．



(a) $T_{LD,LU}, T_{RU,LU}, T_{RD,LU}$ (P_1)　　(b) $T_{RD,RU}, T_{RD,LD}, T_{RD,LU}$ (P_2)

図 4.13: H/V **グラフにおける禁止ターン集合** (P_1, P_2)

STEP2-(c) 循環構造の検出 (2 回目)

続いて，STEP2-(b) **で選択した禁止ターン集合以外のターンの連鎖により形成される循環構造の識別を行なう．以下，禁止ターン集合として** P_1 **または** P_2 **を選んだ場合の手順をそれぞれ並行して示す．**

禁止ターン集合 P_1 **を除く残りの** 9 **パターンのターン集合は**

$$\begin{aligned} Q_{1a} &= \{T_{LU,n\_dir} \mid n\_dir \in \{LD, RU, RD\}\}, \\ Q_{1b} &= \{T_{p\_dir,n\_dir} \mid p\_dir, n\_dir \in \{LD, RU, RD\}, p\_dir \neq n\_dir\} \end{aligned}$$

の 2 **種類のターン集合に分類することができる．同様に，禁止ターン集合** P_2 **を除く残りの** 9 **パターンのターン集合は**

$$\begin{aligned} Q_{2a} &= \{T_{p\_dir,RD} \mid p\_dir \in \{LU, LD, RU\}\}, \\ Q_{2b} &= \{T_{p\_dir,n\_dir} \mid p\_dir, n\_dir \in \{LU, LD, RU\}, p\_dir \neq n\_dir\} \end{aligned}$$

の 2 **種類のターン集合に分類することができる．**

ここで次の定理が成り立つ．

定理 1 ターン集合 Q_{1a} **に属するターンを含む循環構造には，禁止ターン集合** P_1 **に属するターンが必ず含まれる．**□

証明　ターン集合 Q_{1a} **に属するターン** T_{1a} **を含み，かつ禁止ターン集合** P_1 **に属するターンを含まない循環構造が形成可能であると仮定する．**T_{1a} **は** LU **方向 からその他の** H/V direction **へのターンであるので，このとき，ターン** T_{1a} **の直前に連鎖して循環を形成するターンは，ターン集合** $\{T_{p\_dir,LU} \mid p\_dir \in \{LD, RU, RD\}\}$ **に属するものでなければならない．しかし，このターン集合は，禁止ターン集合** P_1 **と同一であるため先の仮定に矛盾する．ゆえに，ターン集合** Q_{1a} **に属するターンを含む循環構造には，禁止ターン集合** P_1 **に属するターンが必ず含まれる．**□

定理 2 ターン集合 Q_{2a} に属するターンを含む循環構造には，禁止ターン集合 P_2 に属するターンが必ず含まれる．□

証明　ターン集合 Q_{2a} に属するターン T_{2a} を含み，かつ禁止ターン集合 P_2 に属するターンを含まない循環構造が形成可能であると仮定する．T_{2a} は RD 方向以外の H/V direction **から RD 方向へのターンであるので，このとき，ターン T_{2a} の直後に連鎖して循環を形成するターンは，ターン集合 $\{T_{RD,n\_dir} \mid n\_dir \in \{LU, LD, RU\}\}$ に属するものでなければならない．しかし，このターン集合は，禁止ターン集合 P_2 と同一であるため先の仮定に矛盾する．ゆえに，ターン集合 Q_{2a} に属するターンを含む循環構造には，禁止ターン集合 P_2 に属するターンが必ず含まれる．□**

定理 1 **より，禁止ターン集合 P_1 を選択することにより，ターン集合 Q_{1a} に属するターンを含むすべての循環構造についても除去される．このため，禁止ターン集合 P_1 を選択した場合に形成可能な循環構造は，ターン集合 Q_{1b} に属する LU 方向を伴なわないターンだけで構成されるものに絞られる．定理** 2 **から，同様の議論により，禁止ターン集合 P_2 を選択した場合に形成可能な循環構造は，ターン集合 Q_{2b} に属する RD 方向を伴なわないターンだけで構成されるものに絞られる．**

残りすべての循環構造を識別するために，ターン集合におけるターン間の依存関係を示す turn dependency graph (TDG) **を導入する．**TDG D **は，$D = G(V, E)$ で表わされ，V は形成可能なターン集合 $V = \{T_1, T_2, ..., T_n\}$ を表し，E は V に属する** 2 **つのターン間で形成可能なターン連鎖の集合 $E = \{TD_1, TD_2, ..., TD_m\}$ を表す．**

図 4.14 **および図** 4.15 **に，ターン集合 Q_{1b} および Q_{2b} における** TDG D_1**，D_2 をそれぞれ示す．図** 4.14 **および図** 4.15 **において，各ノードは，各々のターン集合に属するターンの** 1 **つを表し，各ノード間を結ぶチャネルは** 2 **つのターン間の連鎖を表している．**

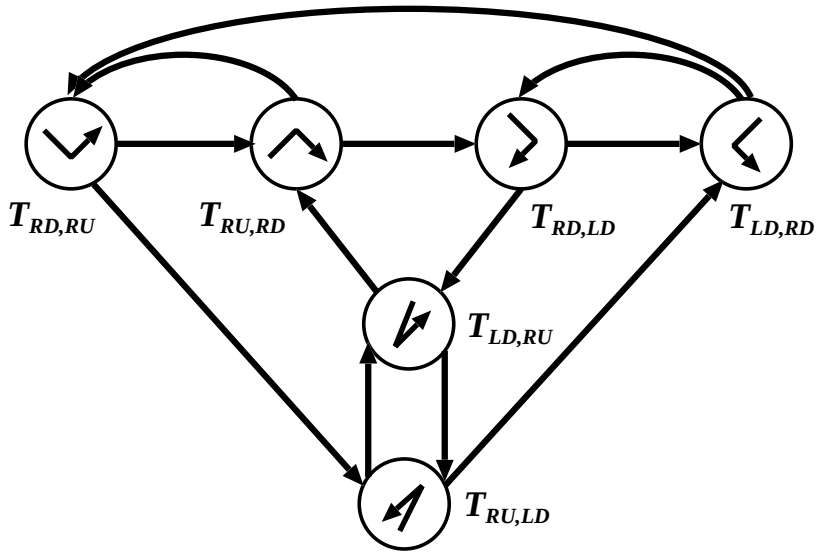


図 4.14: **ターン集合 Q_{1b} における** TDG D_1

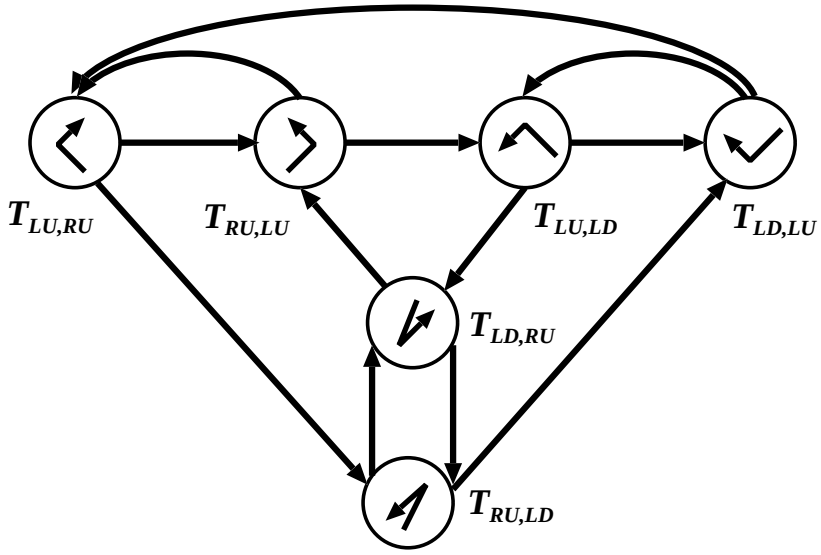


図 4.15: ターン集合 Q_{2b} における TDG D_2

TDG D_1 におけるターン間の依存関係により形成される循環構造には，例として，図 4.16 に示す $C_r(T_{RD,RU}, T_{RU,RD}, T_{RD,LD}, T_{LD,RU}, T_{RU,LD}, T_{LD,RD})$ のような循環構造が含まれている．図 4.16 の循環構造 C_r では，例えば，ターン集合 $\{T_{RD,RU}, T_{RU,LD}, T_{LD,RD}\}$ を取り除いても，残りのターンにより，図 4.17(a) に示す循環構造 $C_{m1}(T_{RU,RD}, T_{RD,LD}, T_{LD,RU})$ が維持される．同様に，C_r からターン集合 $\{T_{LD,RU}, T_{RU,LD}\}$ を除いても，4.17(b) に示す循環構造 $C_{m2}(T_{RD,RU}, T_{RU,RD}, T_{RD,LD}, T_{LD,RD})$ が維持される．本論文では，このような循環構造を冗長循環構造と呼ぶ．

定義 9 (冗長循環構造) 循環を形成するターンの集合から，いずれかのターン集合を除去した場合に，残りのターンの組み合わせにより循環が維持されるような循環構造を冗長循環構造と定める．

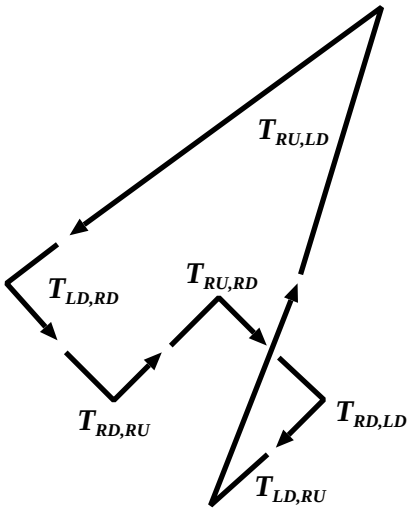
循環構造 C_{m1} および C_{m2} では，循環を形成するターン集合のどのターンを除いても，残りのターンの組み合わせにより循環構造が形成されることがない．このような循環構造を，最小循環構造と定める．

定義 10 (最小循環構造) 循環を形成するターンの集合から，任意のターン集合を除去しても，残りのターンの組み合わせにより循環構造が形成されることがない循環構造を最小循環構造と定める．

TDG において形成可能なすべての循環構造を破るためには，TDG におけるすべての最小循環構造を破ればよい．

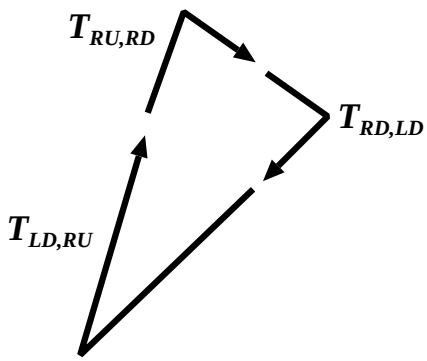
そこで，TDG におけるターン集合について，次のアルゴリズムに基づく TDG 上の探索を行なうことにより，形成可能なすべての最小循環構造の識別を行なう．

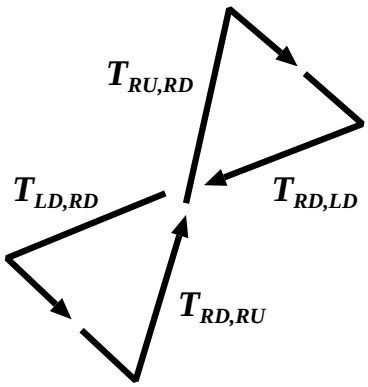
まず，準備として，探索アルゴリズムにおいて用いられる用語の説明を行なう．



$C_r(T_{LD,RD}, T_{RD,RU}, T_{RU,RD}, T_{RD,LD}, T_{LD,RU}, T_{RU,LD})$

図 4.16: TDG D_1 **における冗長循環構造の例**





(a) $C_{m1}(T_{RU,RD}, T_{RD,LD}, T_{LD,RU})$　(b) $C_{m2}(T_{RD,RU}, T_{RU,RD}, T_{RD,LD}, T_{LD,RD})$

図 4.17: TDG D_1 **における最小循環構造の例**

- **探索ステート**
  探索プロセスが起点ノード (ターン) から現在地ノードに至るまでに経由したノードリスト (ターンリスト) を表す．探索プロセスは，ノードを訪問する度に，訪問したノード (ターン) を探索ステートの末尾に追加する．探索プロセスが 1 つ前のノードに戻る場合には，探索ステート末尾のノードを除去する．

- **通過済探索ステートリスト**
  TDG **上の各チャネル (ターン連鎖) が保持するリストを表す．探索プロセスがチャネルを通過した際に，探索プロセスの探索ステートが通過済探索ステートリストに追加される．探索プロセスは，次のいずれかの条件を満たすチャネルを通過することができない．**

  (a) **通過済探索ステートリストに探索プロセスの探索ステートが登録されている，**

  (b) **移動先のノードが探索ステートに含まれている (訪問済である)．ただし、移動先が起点ノードの場合は該当しない．**

TDG $D = G(V, E)$ **に対する探索アルゴリズムは次の通りとなる．**

TDG における探索アルゴリズム

(1) **ノード集合** V **のうち，探索が完了していないノードを選択して起点ノードとし，**(2) **に進む．すべてのノードについて探索が完了した場合，探索を終了する．**

(2) **起点ノードから隣接ノードに向かうチャネルのうち，選択可能なチャネルがあれば，そのチャネルを通過して隣接ノードを訪問し，**(3) **に進む．起点ノードから隣接ノードに向かうすべてのチャネルを起点とする探索が完了した場合，現起点ノードにおける探索を完了し** (1) **に戻る．**

(3) **現在地ノードが起点ノードでない場合，現在地ノードから隣接ノードに向かう選択可能なチャネルがあれば，そのチャネルをたどって隣接ノードを訪問し，**(3) **を繰り返す．**

すべてのチャネルが選択不可である場合，一つ前のノードに戻る．1 **つ前のノードが起点ノードである場合は** (2) **に戻り，それ以外の場合は，**(3) **を繰り返す．**

現在地ノードが起点ノードである場合，探索ステートにおいて，次の 4 **通りのターンが行なわれた回数をそれぞれ数える．**

(t1) up **方向から** down **方向 へのターン**
($T_{LU,LD}, T_{LU,RD}, T_{RU,LD}, T_{RU,RD}$ **のいずれか**)

(t2) down **方向から** up **方向**
($T_{LD,LU}, T_{LD,RU}, T_{RD,LU}, T_{RD,RU}$ **のいずれか**)

(t3) left **方向から** right **方向**
($T_{LU,RU}, T_{LU,RD}, T_{LD,RU}, T_{LD,RD}$ **のいずれか**)

(t4) right 方向から left 方向
($T_{LU,RU}, T_{LU,RD}, T_{LD,RU}, T_{LD,RD}$ のいずれか)

そして，上記の 4 通りのターンが行なわれた回数が次のいずれに該当するかを判定する．

(n1) 4 通りのターンのいずれかが行なわれていない
→ 探索ステートは循環構造を形成しない

(n2) 4 通りのターンがそれぞれ 1 回以上行なわれており，かつ，いずれかのターンが 2 回以上行なわれている
→ 探索ステートは循環構造を形成するが，冗長なターンを含むため最小循環構造ではない．

(n3) 4 通りのターンがそれぞれ 1 回ずつ行なわれている
→ 探索ステートは最小循環構造を形成する．

判定後，(n3) に該当する場合だけ，最小循環構造リストに探索ステートの循環構造を追加する (既に含まれている場合を除く)．その後，一つ前のノードに戻って (3) を繰り返す．

上記の探索アルゴリズムを TDG D_1 に適用することにより，ターン集合 Q_{1b} に属するターンによって形成されるすべての最小循環構造は，図 4.18 における次の 4 つの循環構造となる．

- $C_4(T_{RU,RD}, T_{RD,LD}, T_{LD,RU})$,
- $C_5(T_{RD,RU}, T_{RU,LD}, T_{LD,RD})$,
- $C_6(T_{LD,RU}, T_{RU,LD})$,
- $C_7(T_{RD,RU}, T_{RU,RD}, T_{RD,LD}, T_{LD,RD})$

同様に，TDG D_2 に対して探索アルゴリズムを適用することにより，ターン集合 Q_{2b} に属するターンによって形成されるすべての最小循環構造は，図 4.19 における次の 4 つの循環構造となる．

- $C_8(T_{LD,RU}, T_{RU,LU}, T_{LU,LD})$,
- $C_9(T_{LD,LU}, T_{LU,RU}, T_{RU,LD})$,
- $C_{10}(T_{LD,RU}, T_{RU,LD})$,
- $C_{11}(T_{LD,LU}, T_{LU,RU}, T_{RU,LU}, T_{LU,LD})$

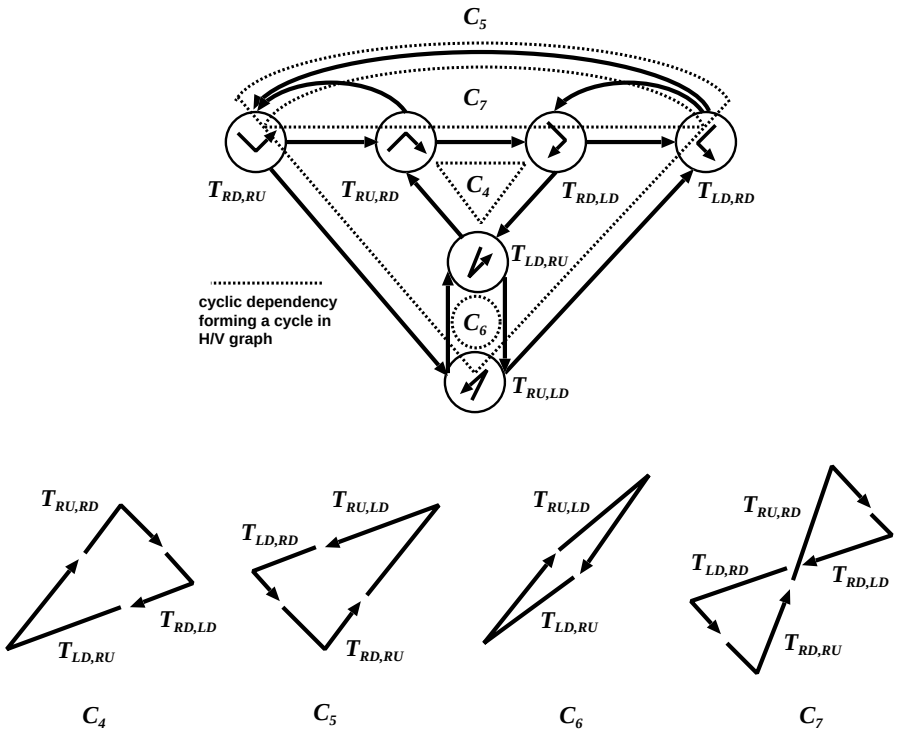


図 4.18: TDG D_1 **における最小循環構造** (C_4, C_5, C_6, C_7)

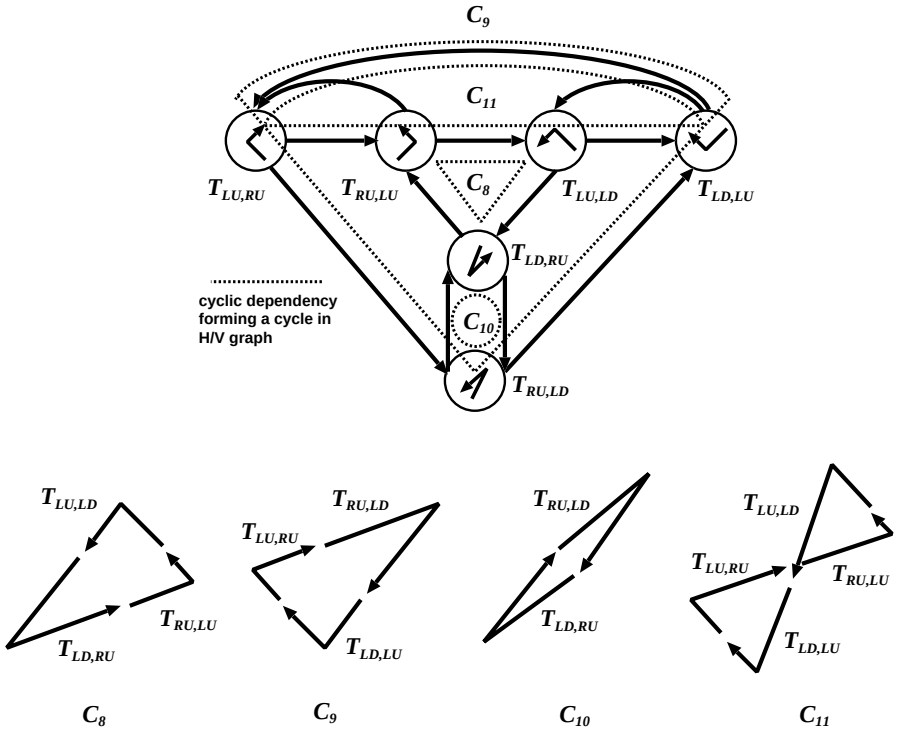


図 4.19: TDG D_2 **における最小循環構造** $(C_8, C_9, C_{10}, C_{11})$

STEP2-(d) 禁止ターンの選択 (2 回目)

STEP2-(c) で識別されたそれぞれ 4 つの循環構造を破るために，前述の選択ポリシーに基づいて禁止ターンの選択を行なう．

まず，禁止ターン集合 P_1 を選択した場合，ターン集合 Q_{1b} によって形成される 4 つの循環構造 $\{C_1, C_2, C_3, C_4\}$ を破るための禁止ターン集合としては，次の 2 通りの禁止ターン集合が選択される．

$$\begin{aligned} P_{1a} &= \{T_{LD,RU}, T_{LD,RD}\}, \\ P_{1b} &= \{T_{RU,LD}, T_{RU,RD}\} \end{aligned}$$

図 4.20 に上記の 4 つの循環構造とそれらを破るための禁止ターン集合 P_{1a} および P_{1b} に属するターンをそれぞれを示す．

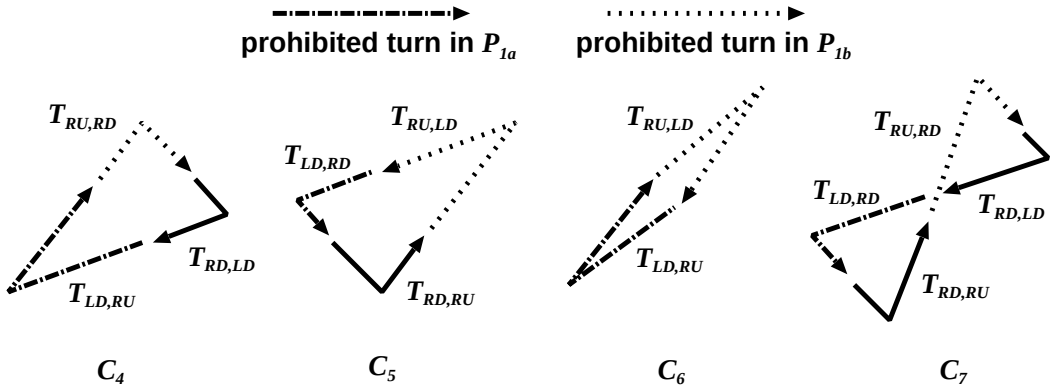


図 4.20: H/V グラフにおける禁止ターン集合 (P_{1a}, P_{1b})

次に，禁止ターン集合 P_2 を選択した場合，ターン集合 Q_{2b} によって形成される 4 つの循環構造 $\{C_5, C_6, C_7, C_8\}$ を破るための禁止ターン集合としては，次の 2 通りの禁止ターン集合が選択される．

$$\begin{aligned} P_{2a} &= \{T_{LD,RU}, T_{LU,RU}\}, \\ P_{2b} &= \{T_{RU,LD}, T_{LU,LD}\} \end{aligned}$$

図 4.21 に上記の 4 つの循環構造とそれらを破るための禁止ターン集合 P_{2a} および P_{2b} に属するターンをそれぞれを示す．

最終的に，禁止ターン集合 P_1 を選択した場合，次の 2 通りの禁止ターン集合が選択される．

$$\begin{aligned} P_1 + P_{1a} &= \{T_{LD,LU}, T_{RU,LU}, T_{RD,LU}, T_{LD,RU}, T_{LD,RD}\}, \\ P_1 + P_{1b} &= \{T_{LD,LU}, T_{RU,LU}, T_{RD,LU}, T_{RU,LD}, T_{RU,RD}\} \end{aligned}$$

禁止ターン集合 P_1 により，LU 方向を伴なうターンを含む循環構造が破れ，禁止ターン集合 P_{1a} または P_{1b} によりその他のターンを含む循環構造が破れる．これにより，H/V

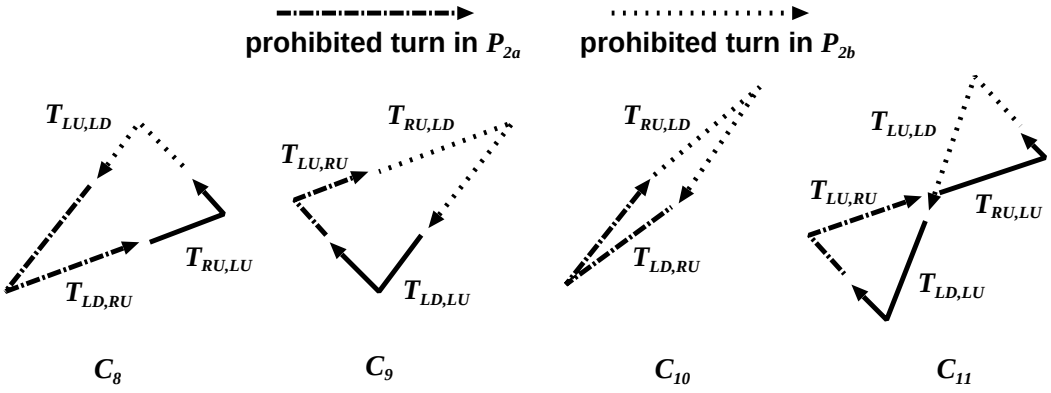


図 4.21: H/V **グラフにおける禁止ターン集合** (P_{2a}, P_{2b})

グラフにおいて形成可能なすべての循環構造が破れ，デッドロックフリーであることが保証される．

STEP2-(b) **で禁止ターン集合** P_2 **を選択した場合についても，同様に，次の** 2 **通りの禁止ターン集合が定められる．**

$$\begin{aligned} P_2 + P_{2a} &= \{T_{RD,RU}, T_{RD,LD}, T_{RD,LU}, T_{LD,RU}, T_{LU,RU}\}, \\ P_2 + P_{2b} &= \{T_{RD,RU}, T_{RD,LD}, T_{RD,LU}, T_{RU,LD}, T_{LU,LD}\} \end{aligned}$$

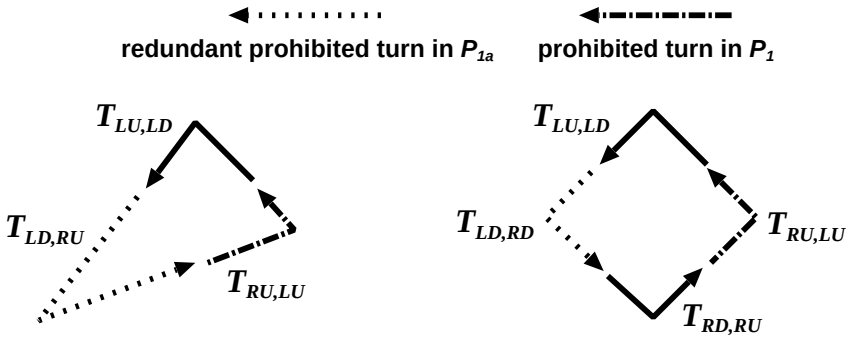
禁止ターン集合 P_2 **により** RD **方向を伴なうターンを含む循環構造が破れ，禁止ターン集合** P_{2a} **または** P_{2b} **によりその他のターンを含む循環構造が破れる．これにより，**H/V **グラフにおいて形成可能なすべての循環構造が破れ，同様にデッドロックフリーであることが保証される．**

4.2.2.4　循環構造検出アルゴリズムによる冗長禁止ターンの削除 (STEP3)

前節 (STEP2) **において，**H/V **グラフにおいて形成可能なすべての循環構造を破るために** 4 **つの禁止ターン集合を導出した．ここでは，各禁止ターン集合における一部の禁止ターンは，循環構造の形成に関与しない場合があることを示し，トポロジ毎にそのような冗長な禁止ターンを削除するための手順を示す．なお，ここでは，禁止ターン集合として** $P_1 = \{T_{LD,LU}, T_{RU,LU}, T_{RD,LU}\}$ **と** $P_{1a} = \{T_{LD,RU}, T_{LD,RD}\}$ **を選択した場合についての手順を述べるが，その他の禁止ターン集合を選択した場合についても同様にして考えることができる．**

禁止ターン集合 P_{1a} **に属する** 2 **つの禁止ターン** $\{T_{LD,RU}, T_{LD,RD}\}$ **は 図** 4.18 **の** 4 **つの循環構造** C_4, C_5, C_6, C_7 **を破るために必要とされる．しかし，これら** 2 **つのターンを含む循環構造は，図** 4.22 **のように，禁止ターン集合** P_1 **に属するターンを含んでいる場合がある．**

図 4.22 **における** 2 **つの循環構造には，禁止ターン集合** P_1 **および禁止ターン集合** P_{1a} **に属するターンがそれぞれ** 1 **つずつ含まれている．このような場合，**P_{1a} **に属するターンを禁止せずとも，**P_1 **に属するターンを禁止することにより循環構造は除去される．この**



(a) $C(T_{LD,RU}, T_{RU,LU}, T_{LU,LD})$　　(b) $C(T_{LD,RD}, T_{RD,RU}, T_{RU,LU}, T_{LU,LD})$

図 4.22: 冗長な禁止ターンを含む循環構造

ような循環構造が形成されるトポロジにおいて，禁止ターン集合 P_{1a} に属するターンをすべて禁止してしまうと，ルーティングの自由度が低下し，不要な非最短経路の増加とトラフィックの偏りの原因となりうる．

このような冗長な禁止ターンを削除するために，トポロジ毎に，H/V グラフにおける図 4.18 の 4 つの循環構造 (除去のために，禁止ターン集合 P_{1a} を必要とする) を検出し，検出された循環構造に含まれる場合にだけ，禁止ターン集合 P_{1a} に属するターンを個別に禁止する．循環構造の検出は，H/V グラフ上で探索ベースのアルゴリズムを適用することにより行なう．

以下，循環構造検出アルゴリズムについて述べる．

循環構造の検出は，H/V グラフにおいて，次の 2 つの条件のいずれか，または両方を満たす各スイッチをそれぞれ起点として深さ優先探索を行なうことにより行なわれる．

(a) **禁止ターン集合 P_{1a} に属するターン $T_{LD,RD}$ が形成可能である**
(1 つ以上の RU チャネルおよび RD チャネルが接続されている)．

(b) **禁止ターン集合 P_{1a} に属するターン $T_{LD,RU}$ が形成可能である**
(2 つ以上の RU チャネルが接続されている)．

探索において，隣接スイッチの訪問に利用される出力チャネルは，次の条件を満たす場合に選択可能であるとし，利用後に通過済マークをつける．

(1) **LU チャネルではない (P_1 に含まれる禁止ターンを形成しない)，**

(2) **通過済マークがついてない，**

(3) **過去の探索において禁止された，ターン集合 P_{1a} に属するいずれかのターンを形成しない**

探索の手順を次に示す．探索は 2 つの手順から成り，起点となるスイッチが条件 (a) を満たす場合に手順 1 を，条件 (b) を満たす場合に手順 2 をそれぞれ実行する．

手順 1: 条件 (a) に該当するスイッチを起点とした探索

(1) **起点スイッチから隣接スイッチに向かう RD チャネルのうち，通過済マークがついていないものを選び，到達可能な隣接スイッチを訪問する．その後，(2) に進む．選択可能な RD チャネルがなければ，すべてのチャネルの通過済マークを消去し，探索を完了する．**

(2) **現在地スイッチが起点スイッチであり，かつ，最後に通過したチャネルが LD チャネルであるならば，循環構造が検出されたことになる．この場合，到着時に通過した LD チャネルと出発時に通過した RD チャネルの間に形成されるターン $T_{LD,RD}$ を禁止する．その後，直前のスイッチに戻り，(2) を繰り返す．**

それ以外の場合には，選択可能な出力チャネルがあれば，深さ優先探索に基づいて，隣接スイッチを訪問し，(2) を繰り返す．選択可能な出力チャネルがない場合には，直前のスイッチに戻る．直前のスイッチが起点スイッチであれば (1) に戻り，そうでなければ (2) を繰り返す．

手順 2: 条件 (b) に該当するスイッチを起点とした探索

手順 1 を次の条件で置き換えて実行する．

(1) **起点スイッチからの最初の訪問には，ターン $T_{LD,RU}$ を形成するチャネルのうち，起点スイッチから出る方向となる RU チャネルを用いる．**

(2) **循環構造検出時には，ターン $T_{LD,RU}$ が禁止される．**

手順 1 により検出される循環構造は，ターン $T_{LD,RD}$ を含み，かつ，禁止ターン集合 P_1 に属するターンを含まないので，循環構造 C_5 または C_7 のいずれかとなる．同様に，手順 2 により検出される循環構造は，ターン $T_{LD,RU}$ を含み，かつ，禁止ターン集合 P_1 に属するターンを含まないので，循環構造 C_4 または C_6 のいずれかとなる．

上記のアルゴリズムは，禁止ターン集合として $P_1 + P_{1b}$, $P_2 + P_{2a}$ および $P_2 + P_{2b}$ のいずれかを選択した場合についても，検出の対象となる禁止ターン集合 P_{1a} を，$P_{1b} = \{T_{RU,LD}, T_{RU,RD}\}$, $P_{2a} = \{T_{LD,RU}, T_{LU,RU}\}$ および $P_{2b} = \{T_{RU,LD}, T_{LU,LD}\}$ にそれぞれ置き換えることにより同様にして実行可能である．ただし，禁止ターン集合として P_2 を選択した場合には，探索における禁止ターン集合 P_1 を P_2 に置き換える．

図 4.23 に，禁止ターン集合として $P_1 + P_{1a}$ を選択した場合に，循環構造検出アルゴリズムにより検出される循環構造の例を示す．

図 4.23 において，5 つのターン $T_1, ..., T_5$ は，ターン集合 P_{1a} に属するターンであり，これらのターンが形成されているスイッチ 5,6,7,9 が前述の探索の起点となる．図 4.23 より，循環構造検出アルゴリズムにより検出される循環構造は，循環 C_1 だけであるため，5 つのターンのうち，循環 C_1 に含まれるターン T_5 だけを禁止すればよいことがわかる．

スイッチ数を n，スイッチあたりのリンク数を l とすると，探索アルゴリズムの計算量は $O(n^2 * l)$ となる．

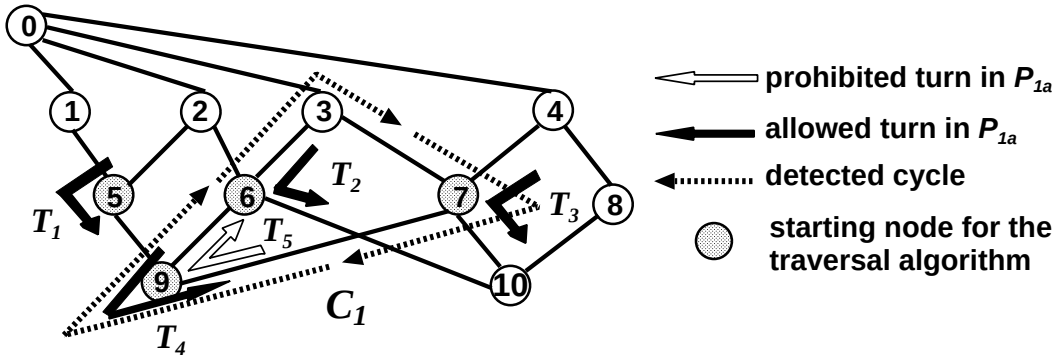


図 4.23: 循環構造検出アルゴリズムにより検出される循環構造

4.2.3　L-turn/R-turn ルーティングの定義

第 4.2.2.3 節で定めた 4 通りの禁止ターン集合と第 4.2.2.4 節で定めた循環構造検出アルゴリズムに基づいて，4 つの適応型デッドロックフリールーティングアルゴリズムを次のように定義する．

まず，循環構造検出アルゴリズムを対象トポロジに適用することにより禁止される全ターン集合 DP を次のように表す．

$$DP = DA(H, P_{stat}, P_{cond})$$

DA は，対象とする H/V グラフ H において，常に禁止されるターン集合を P_{stat}，検出対象とする禁止ターン集合を P_{cond} として循環構造検出アルゴリズムを適用した結果禁止されるすべてのターン集合 DP を返す関数とする．

次に，ターン集合 $P_1 = \{T_{LD,LU}, T_{RU,LU}, T_{RD,LU}\}$ を禁止ターンとする 2 つのルーティングアルゴリズムを定義する．

ターン集合 P_1 を禁止することにより，LU 方向に向かうすべてのターンが禁止されるため，この場合，LU 方向への移動は最初に行なう必要がある．そこで，このようなルーティングアルゴリズムをまとめて **L-turn (Left-up first turn)** ルーティングと呼ぶ．

L-turn/α ルーティング　H/V グラフ上で，ターン集合 $P_1 = \{T_{LD,LU}, T_{RU,LU}, T_{RD,LU}\}$ に属するターンを禁止し，ターン集合 $P_{1a} = \{T_{LD,RU}, T_{LD,RD}\}$ を検出対象とする循環構造検出アルゴリズムの適用により，ターン集合 $DP_{1a} = DA(H, P_1, P_{1a})$ を禁止するルーティングアルゴリズムを **L-turn/α** ルーティングと呼ぶ．

L-turn/β ルーティング　H/V グラフ上で，ターン集合 $P_1 = \{T_{LD,LU}, T_{RU,LU}, T_{RD,LU}\}$ に属するターンを禁止し，ターン集合 $P_{1b} = \{T_{RU,LD}, T_{RU,RD}\}$ を検出対象とする循環構造検出アルゴリズムの適用により，ターン集合 $DP_{1b} = DA(H, P_1, P_{1b})$ を禁止するルーティングアルゴリズムを **L-turn/β** ルーティングと呼ぶ．

次に，ターン集合 $P_2 = \{T_{RD,RU}, T_{RD,LD}, T_{RD,LU}\}$ を禁止ターンに含む 2 つのルーティングアルゴリズムを定義する．

ターン集合 P_2 を禁止することにより，RD 方向からその他の方向に向かうすべてのターンが禁止されるため，RD 方向への移動は最後に行なう必要がある．そこで，このようなルーティングアルゴリズムをまとめて **R-turn (Right-down last turn)** ルーティングと呼ぶ．

R-turn/α ルーティング　H/V グラフ上で，ターン集合 $P_2 = \{T_{RD,RU}, T_{RD,LD}, T_{RD,LU}\}$ に属するターンを禁止し，ターン集合 $P_{2a} = \{T_{LD,RU}, T_{LU,RU}\}$ を検出対象とする循環構造検出アルゴリズムの適用により，ターン集合 $DP_{2a} = DA(H, P_2, P_{2a})$ を禁止するルーティングアルゴリズムを **R-turn/α ルーティング**と呼ぶ．

R-turn/β ルーティング　H/V グラフ上で，ターン集合 $P_2 = \{T_{RD,RU}, T_{RD,LD}, T_{RD,LU}\}$ に属するターンを禁止し，ターン集合 $P_{2b} = \{T_{RU,LD}, T_{LU,LD}\}$ を検出対象とする循環構造検出アルゴリズムの適用により，ターン集合 $DP_{2b} = DA(H, P_2, P_{2b})$ を禁止するルーティングアルゴリズムを **R-turn/β ルーティング**と呼ぶ．

L-turn ルーティングおよび R-turn ルーティングにおける許可ターンと禁止ターン集合を図 4.24 にまとめて示す．図 4.24 において，実線は許可ターン，破線は禁止ターン，点線は，循環構造検出アルゴリズムにより禁止となりうるターンを示す．

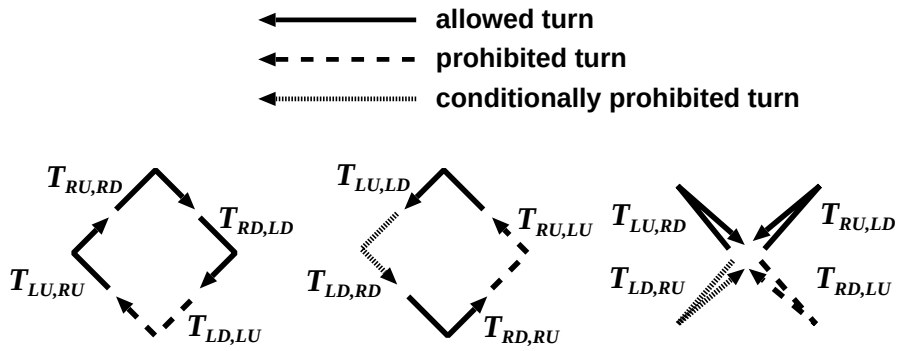
定理 3 *L-turn* ルーティングおよび *R-turn* ルーティングはデッドロックフリーである □

証明　H/V グラフにおいて形成可能なすべての循環構造は，各ルーティングアルゴリズムにおいて選択された禁止ターン集合により除去される．ゆえに，L-turn ルーティングおよび R-turn ルーティングは，デッドロックフリーである □

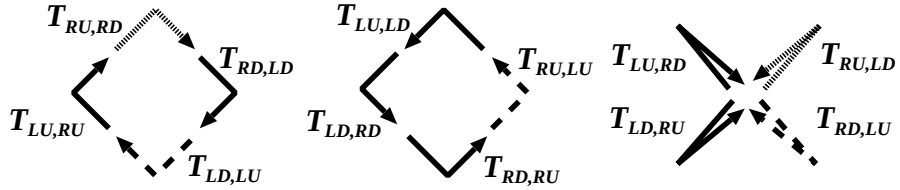
定理 4 *L-turn* ルーティング および *R-turn* ルーティングでは任意のスイッチ間の経路が保証される．□

証明　H/V ツリーに属するチャネルの方向は，LU 方向または RD 方向だけである．RD 方向から LU 方向へのターンは禁止されているので，H/V ツリー内で形成可能なターンは $T_{LU,RD}$ だけとなる．ターン $T_{LU,RD}$ は禁止されていないので H/V ツリーにおいては任意のノード間でのパケット転送が保証される．ゆえに，L-turn ルーティング および R-turn ルーティングでは，任意のスイッチ間の経路が保証される．□

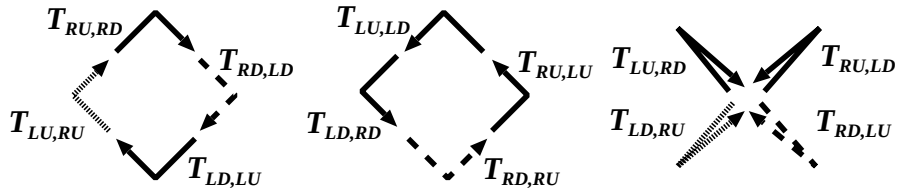
図 4.25 は，4×4 スイッチの 2 次元メッシュにおいて，BFS Up*/Down* ルーティング，L-turn ルーティング (α および β で同一)，2 次元メッシュ向けの West-first ルーティングを適用した場合の禁止ターンの分布をそれぞれ示している．図 4.25(a) と図 4.25(b) を比較すると，いずれも禁止ターンの数は同じであるものの，Up*/Down* ルーティングでは禁止ターンの偏りが発生しているのに対し，L-turn ルーティングでは禁止ターンの分散が実現されていることがわかる．また，図 4.25(b) と図 4.25(c) を比較すると，両者の禁止ターンの分布は同一となっていることが分かる．これは，トポロジによっては，L-turn ルーティングが，一般的に高い性能を示す 2 次元メッシュに特化したルーティングアルゴリズムと同等の禁止ターンの分散を実現しうることを示す．



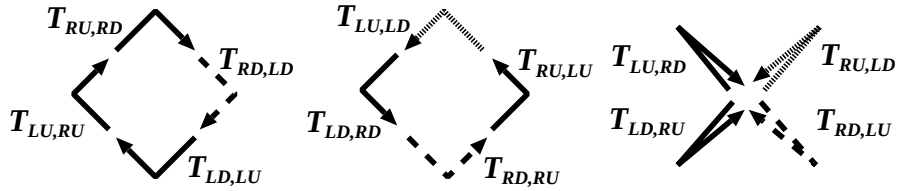
(a) L-turn/α ルーティング



(b) L-turn/β ルーティング

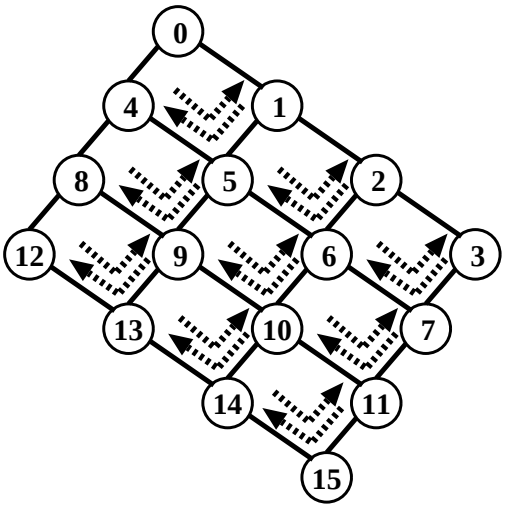


(c) R-turn/α ルーティング

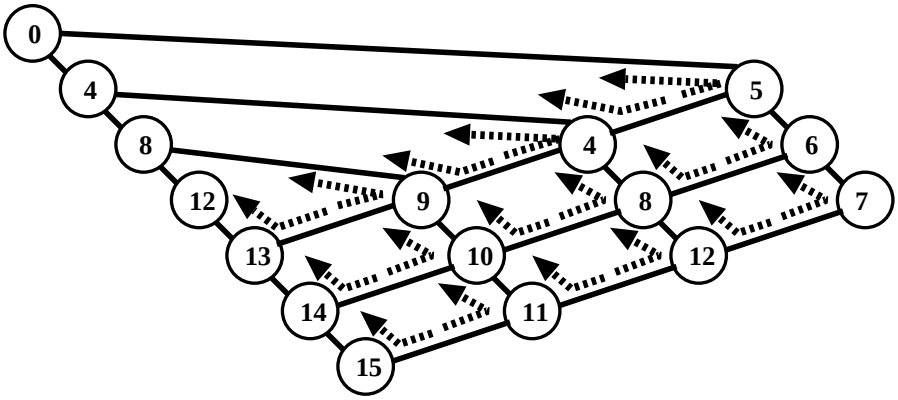


(d) R-turn/β ルーティング

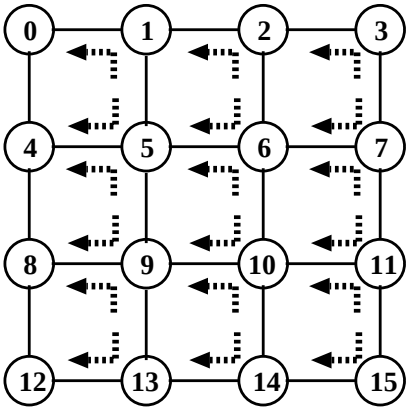
図 4.24: L-turn/R-turn **ルーティングにおける許可ターンと禁止ターン集合**



(a) Up*/Down* ルーティング



(b) L-turn ルーティング



(c) West-first ルーティング

図 4.25: 2 **次元メッシュ**(4×4 **スイッチ**) **における禁止ターン**

L-turn および R-turn ルーティングでは，Up*/Down* ルーティングと同様に，禁止ターンを行なわない限り，任意の経路を選択して適応型のルーティングを行なうことが可能である．しかし，効率的なルーティングの実現のため，各ルーティングアルゴリズムでは，同様に，任意のスイッチ間における選択可能な経路のうち最短となるものだけを基本的に選択するものとする．

各ルーティングアルゴリズムにおける各スイッチ間の経路計算は，次のように行なわれる．ここでは，L-turn/α の場合について述べるが，提案した他のルーティングアルゴリズムについても同様にして行なわれる．

(a) 対象トポロジの H/V グラフ上の各スイッチにおいて，ターン集合 P_1 に属するターンを形成するチャネル間の移動を禁止する．

(b) H/V グラフにおいて，第 4.2.2.4 節で示した条件 (a),(b) のいずれか，または両方を満たす全スイッチにおいて循環構造検出アルゴリズムを順に適用し，循環構造の検出を行なう．検出された循環構造において，ターン集合 P_{1a} に属するターンを形成するチャネル間の移動を禁止する．

(c) H/V グラフの全スイッチにおいて，ダイクストラのアルゴリズム [E.W59] を適用し，全スイッチ間の最短経路を算出する．探索においては，禁止されたチャネル間の移動はできないものとする．

上記の経路計算のための計算量は，スイッチ数を n とすると $O(n^2)$ となる．最短経路が複数存在する場合の経路選択方法については，Up*/Down* ルーティングと同様に，対象とするネットワークで用いられるスイッチの実装に依存する．

図 4.26 に，L-turn/α および BFS Up*/Down* ルーティングの経路例を示す．図 4.26 は，16 スイッチ構成の H/V グラフにおけるスイッチ番号 11 からスイッチ番号 10 へのパケット転送において，各ルーティングアルゴリズムにより選択可能なすべての経路を示している．図 4.26 において，各ルーティングアルゴリズムは，共に 4 通りの経路を持っている．しかし，Up*/Down* ルーティングの経路はすべて 5 ホップを要し，かつ，ルートスイッチを必ず通らなくてはならないのに対し，L-turn/α ルーティングの経路はすべて，3 ホップを要するだけであり，また，ルートスイッチを通る必要が無く，経路も分散されている．

4.2.4　同 depth スイッチ間チャネルの方向割当ての効果

第 4.1.3 節で述べたように，H/V グラフにおける図 4.27(a) のような同 depth スイッチ間チャネルの方向割当ては，図 4.27(b) に示すように，right 方向に向かうチャネルに対して right-up，left 方向に向かうチャネルに対して left-down としている．方向割当てとしては，図 4.27(c) に示すような，逆のケースも考えられる．

前者の方向割当てを行なっている理由は，この場合，図 4.27(b) のような禁止ターンの分散が常に実現されるためである．これに対し，後者の方向割当てでは，図 4.27(c) のように，$T_{RD,LU}$ による禁止ターンの集中が常に発生してしまう．

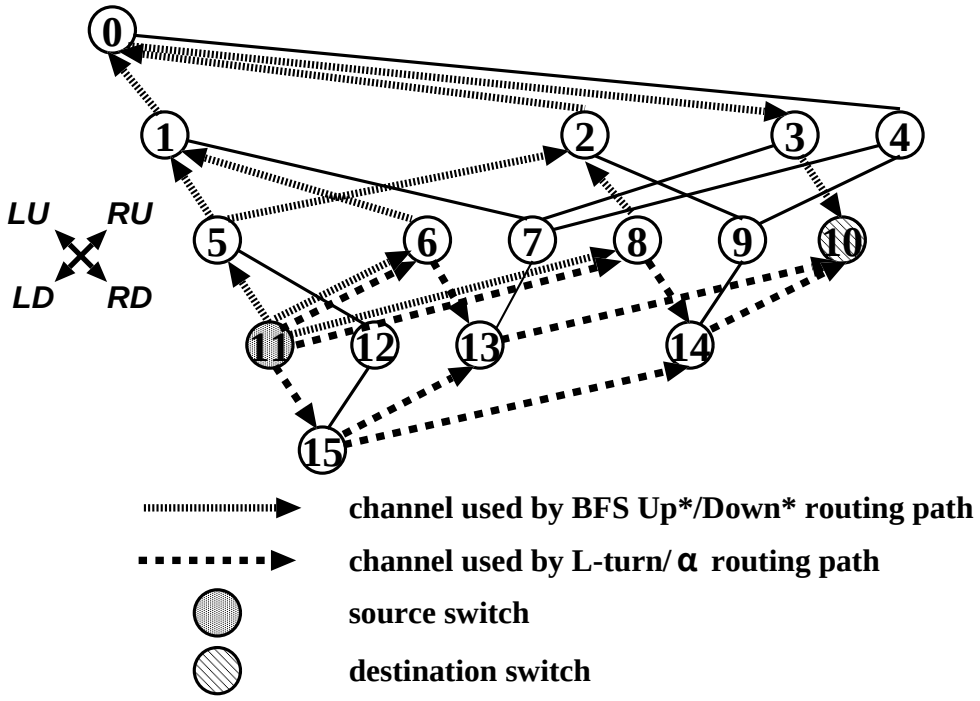


図 4.26: BFS Up*/Down* および L-turn/α ルーティングの経路例

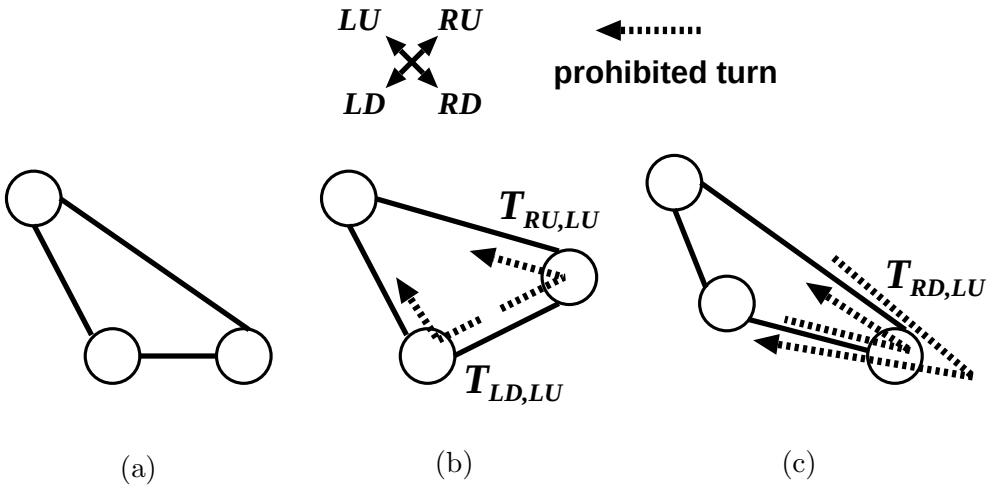


図 4.27: 同 depth スイッチ間チャネルの方向割当ての違いによる
L-turn ルーティングの禁止ターン分布の違い

なお，図 4.27 では L-turn ルーティングの場合の禁止ターン例を示しているが，R-turn ルーティングの場合も同様にして，禁止ターンが分散されている．

4.2.5　H/V グラフ構築時の前順走査における訪問スイッチ選択ポリシー

L-turn および R-turn ルーティングでは，第 4.1.2 節で述べたように，H/V グラフ構築時に行なう前順走査において，次に訪問する子スイッチとして 2 つ以上の子スイッチが選択可能となる場合がある．この選択は，訪問スイッチ選択ポリシーにより行なわれる．選択ポリシーの違いにより，同一ネットワークに対して異なる horizontal spread の割当てがなされるため, これにより異なる H/V グラフが構築されうる．

訪問スイッチ選択ポリシーの例として，次のポリシーが挙げられる．

(a) random
訪問スイッチをランダムに選択する．

(b) less child-node first
サブツリー以下の子スイッチの数が最小となるスイッチを選択する．

(c) more child-node first
サブツリー以下の子スイッチの数が最大となるスイッチを選択する．

(d) more upper-channel first
サブツリー以下の子スイッチが持つ up 方向の スパニングツリー構成外チャネル数の合計が最大となるスイッチを選択する．

random 以外の選択ポリシーでは，同じ条件を満たすスイッチが複数存在する場合には，それらの中からランダムに選択をするものとしている．これらのうち more upper-channel first が，禁止ターンのより均等な分散の実現に適しており，他のポリシーに比べて，平均的にスループットが向上することが確認されている．これは，more upper-channel first では，他のポリシーに比べて，図 4.27(b) のような禁止ターンの分散を実現しやすいためである．

これらの選択ポリシーの違いによる性能への影響については，第 5 章で述べる．

4.2.6　既存のルーティングアルゴリズムとの比較

ここでは，L-turn および R-turn ルーティングと，第 3.1 節で述べた付加的なハードウェアに依存しない既存のルーティングアルゴリズムについてまとめたものを，表 4.3 に示す．表 4.3 の計算量において，n はスイッチ数，m はチャネル数を指す．

表 4.3 に示すように，付加的なハードウェアに依存しないルーティングアルゴリズムは，Turn モデル (スパニングツリー) をベースとするか否かで，大きく 2 つに分類することができる．

表 4.3: 付加的なハードウェアに依存しないルーティングアルゴリズムの比較

	L-turn R-turn	BFS Up*/Down*	DFS Up*/Down*	Smart	Adaptive Trail
スパニングツリー利用	yes (BFS)	yes (BFS)	yes (DFS)	no	no
Turn モデルベース	yes (2D)	yes (1D)	yes (1D)	no	no
禁止ターン集中	low	high	medium	-	-
トポロジフリー	yes	yes	yes	yes	no
計算量	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(n^9)$	$O(m^2)$

4.2.7　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分類

イレギュラーネットワーク向けのルーティングアルゴリズムは，一般的に，仮想チャネルの利用およびスパニングツリーの利用の有無で大きく分類される．スパニングツリーベースのルーティングアルゴリズムである Up*/Down* ルーティングおよび本章で提案した L-turn および R-turn ルーティングは，先に述べたように，Turn モデルをベースとするルーティングアルゴリズムであるため，有向グラフにおける方向の次元数に基づいて更に分類することができる．そこで，第 3.1 節で述べたイレギュラーネットワーク向けのルーティングアルゴリズムおよび L-turn および R-turn ルーティングの分類を，Turn モデルの視点による分類を追加した上で行なうと，図 4.28 のように分類することができる．

理論上，有向グラフの次元数は，n 次元 ($n > 2$) に拡張することが可能であると考えられる．そこで，図 4.28 では，n 次元有向グラフ (n 次元 Turn モデル) を追加している．図 4.28 のように，仮想チャネルを利用するルーティングアルゴリズム [MJ80, SD00, SLT02, JPMJ02, MAH03] は，パケット転送中の仮想チャネルの切り替え (virtual channel transitions) の有無で更に分類される．理論上は，スパニングツリーおよび有向グラフをベースとすることにより，仮想チャネルを用いた Turn モデルベースのルーティングアルゴリズムを実装することが可能である．しかし，仮想チャネルを用いたルーティングアルゴリズムにおいては，図 4.28 の DL ルーティングおよび Minimal ルーティングなどのように，Turn モデルベースの手法を独自の手法と組み合わせて併用しているケースがあり，明確に Turn モデルとしての分類を行なうことが難しい．そのため，図 4.28 では 仮想チャネル利用時の Turn モデルとしての分類については省略をしている．

4.3　研究の過程と本論文の位置付けについて

L-turn および R-turn ルーティングは，筆者と鯉渕の共同研究の成果であり，鯉渕の博士論文 [鯉渕 02] の一部においても述べられている．以下，L-turn および R-turn ルーティングに関する研究の過程と筆者が担当した作業内容について述べ，本論文の位置付けについてまとめる．

まず，L-turn および R-turn ルーティングの発端は，Up*/Down* ルーティングにおける 2 つの方向を 4 つに増加させた上で禁止ターン選択を行なうことにより，イレギュラーネットワークにおいて，より効率的なルーティングを実現できるのではないか，という鯉

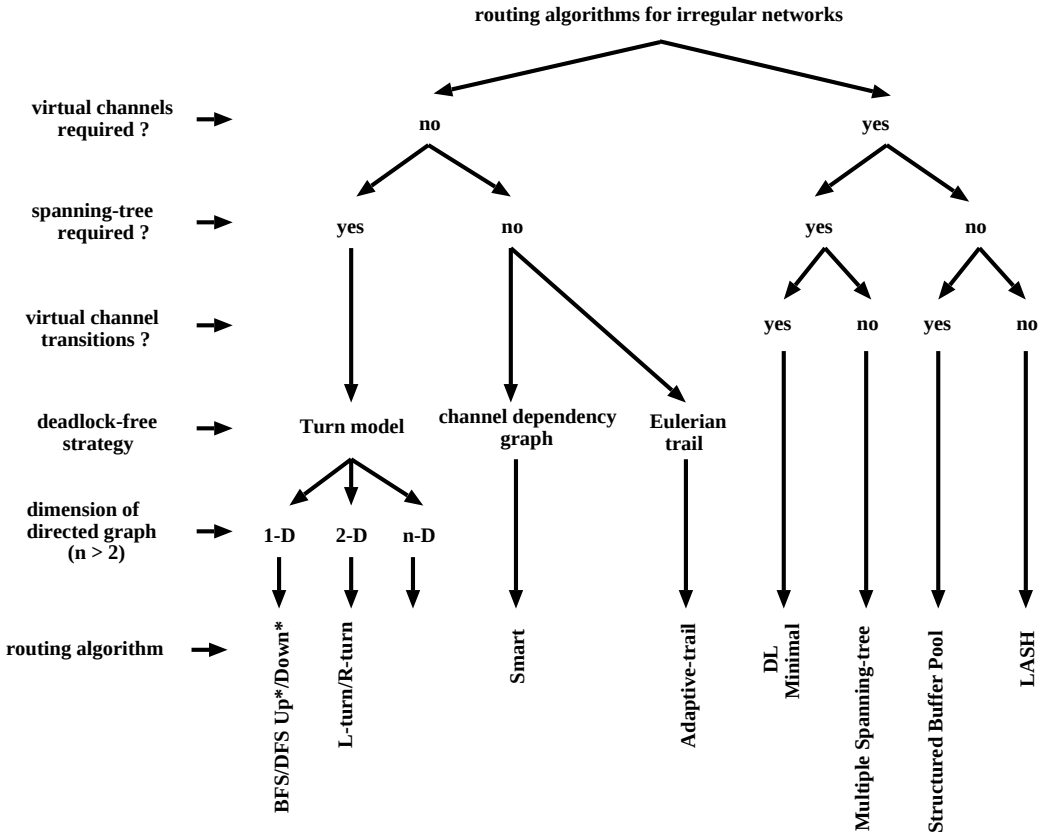


図 4.28: **イレギュラーネットワーク向けルーティングアルゴリズムの分類**

渕の提案によるものである．そこでまず，筆者と鯉渕により，前順走査を利用した 2 **次元有向グラフである** H/V **グラフの構築手順が確立された．当初は，**H/V **グラフにおける** Turn **モデルの適用手順が確立されておらず，まず，第** 4.2.1 **節で述べた** 2 **次元メッシュにおける** Turn **モデルと同様の直観的な選択により，** LU **方向へ向かうターン集合だけを禁止ターンとして選択したルーティングアルゴリズム**[5]**が鯉渕により提案され，確率モデルシミュレーションによる予備評価が行なわれた．予備評価およびこれ以降の評価では，筆者が実装したフリットレベルの相互結合網シミュレータが用いられた．**

予備評価の結果，選択した禁止ターン集合だけではデッドロックが発生することが確認されたため[6]**，その後，筆者と鯉渕により，シミュレーションの結果を基にして，形成可能な循環構造の識別とデッドロックフリー実現のために必要な** 2 **つの禁止ターンの追加が行なわれた．これにより，**L-turn/α **とほぼ同等の禁止ターンを課す (冗長禁止ターンの除去は行なわれていない)**1 **つのデッドロックフリールーティングアルゴリズムが導出された．しかし，改めて予備評価を行なった結果，イレギュラーネットワークにおけるスループッ**

[5] **これが** L-turn **ルーティングの原形となった**

[6] **図** 4.24 **で示されている通り，**L-turn **ルーティングでは，**LU **方向へ向かうターン以外に更に** 2 **つのターンを禁止する必要がある**

ト向上は予想よりも低い値となった．分析の結果，この原因は，冗長禁止ターンが存在するためであることが確認された．そこで，更なる性能向上実現のために，筆者により冗長禁止ターン除去のための循環構造検出アルゴリズムが提案された．これにより，L-turn/α に相当する 1 つのデッドロックフリールーティングアルゴリズムが確立し，評価の結果，高い性能向上が確認された [MAAH01]．

しかし，この時点ではまだ本論文で述べた H/V グラフにおける Turn モデルの適用手順が確立されておらず，L-turn/β, R-turn/α および R-turn/β ルーティングの定義がなされていなかった．そこで，その後筆者により，H/V グラフにおけるシステマティックな 2 次元 Turn モデルの適用手順が提案され，これにより本論文で提案した 4 つのデッドロックフリールーティングアルゴリズムの定義がなされた [AMAH02, 上樂 03]．

以上より，主に筆者が担当した作業内容をまとめると次の通りとなる．

- H/V グラフ構築手順の確立
- 循環構造検出アルゴリズムの確立
- H/V グラフにおける 2 次元 Turn モデル適用手順の確立
  (4 つのルーティングアルゴリズムの導出)
- 評価用フリットレベル相互結合網シミュレータの実装

鯉渕の博士論文 [鯉渕 02] では，文献 [AMAH02] 時点の研究成果に関する内容をベースとして L-turn および R-turn ルーティングについての記述がなされているが，この中で筆者が担当した作業内容は上記の通りである．本論文は，文献 [AMAH02, 上樂 03] の内容をベースとして，主に次の点を改善および補足した内容となっている．

- L-turn および R-turn ルーティングの構築手順における 2 次元 Turn モデル適用手順の具体化 (第 4.2 節)
- より多くの評価指標を用いた性能評価 (第 5 章)
  - 禁止ターン分散の度合いを示す静的な評価指標の分析
  - 経路分散の度合いを示す静的な評価指標の分析
- L-turn および R-turn ルーティングの性能に影響する要素の評価
  - ルートスイッチ選択ポリシー (第 5.2.4 節)
  - H/V グラフ構築時の前順走査における訪問スイッチ選択ポリシー (第 4.2.5 節，第 5.2.5 節)
- L-turn および R-turn ルーティング をソースルーティング方式 (固定型ルーティング) として実装した場合の評価 (第 5.3 節)

4.4　まとめ

本章では，Up*/Down* ルーティングにおいて問題となるトラフィックの偏りを改善するために，より均等なトラフィックの分散の実現を目的とする適応型ルーティングアルゴリズムである L-turn ルーティングおよび R-turn ルーティングの提案を行なった．L-turn および R-turn ルーティングは，Up*/Down* ルーティングで利用されているスパニングツリーベースの 1 次元有向グラフを拡張した H/V グラフと呼ばれる 2 次元有向グラフを利用する．H/V グラフの導入により，形成可能なターンの数は従来の 6 倍である 12 パターンに増加し，これによりトラフィックの分散を考慮したより柔軟な禁止ターンの選択を行なってデッドロックフリーを実現することが可能となる．そして，H/V グラフに対して，禁止ターンの分散を考慮した 2 次元 Turn モデルをシステマティックな手法で適用することにより，L-turn および R-turn ルーティングが導出される．この際，循環構造検出アルゴリズムを導入することにより，冗長な禁止ターンを削除し，更なる性能向上の実現を図っている．L-turn および R-turn ルーティングは，スパニングツリーの構築と 2 次元 Turn モデルの適用をベースとすることにより，Up*/Down* ルーティングと同等の高い汎用性を実現しており，任意の SAN およびトポロジに適用可能となっている．

第5章　評価

本章では，確率モデルシミュレーションにより L-turn および R-turn ルーティングと BFS および DFS Up*/Down* ルーティングの性能評価を行なった結果を示す．

以降，第 5.1 節で評価環境を示し，第 5.2 節と第 5.3 節で，各ルーティングアルゴリズムを分散ルーティング方式 (適応型ルーティング) およびソースルーティング方式 (固定型ルーティング) により実装した場合の評価結果をそれぞれ順に示す．

5.1　評価環境

5.1.1　相互結合網シミュレータ

性能評価のために，フリットレベルでのパケット転送をシミュレートする相互結合網シミュレータを C++ により実装した．このシミュレータは，並列計算機，SAN などにおける様々な相互結合網の性能評価を目的としており，トポロジ，ルーティングアルゴリズム，パケット転送方式，パケット長，リンク間レイテンシなどの相互結合網に関する各パラメータを選択可能となっている．汎用性を高めるため，相互結合網の一般的な構成要素であるルータ，リンク，パケット，ネットワークインタフェースなどが，これらの動作をシンプルにシミュレートした基本クラスとして実装されており，対象とするネットワークに応じて，適宜，スイッチ間結合パターン (トポロジ) の記述，各クラスの機能拡張 (対象とするルーティングアルゴリズムの実装など) を行なうことにより所望するシミュレータが実装される．本論文の評価で用いたシミュレータは，単純なスイッチ間のパケット転送動作をシミュレートしており，任意のトポロジ，スパニングツリーベースの各ルーティングアルゴリズムを選択可能となっている．

このシミュレータは，シミュレーション方式として確率モデルシミュレーションを用いる．確率モデルシミュレーションは, メモリアクセスのパターンなどを乱数モデルに基づいて発行して評価を行う方法であり，相互結合網やメモリシステムなどの評価を行なう際によく用いられる．シミュレーション方式としては，確率モデルシミュレーションの他に，トレースドリブンシミュレーションおよび命令レベルシミュレーションなどが挙げられる．トレースドリブンシミュレーションは, 実機上でアプリケーションプログラムを実行してメモリ参照のアドレスのトレースデータをとり, それをシミュレータに入力して評価を行なう方法である. また，命令レベルシミュレーションは, CPU のインストラクションのレベルまでソフトウェアでシミュレートして, 実機と同様の環境を構築して評価する方法である．これらの方式は，確率モデルシミュレーションに比べて，より現実的な評価を行うことができるが，シミュレータの開発が困難であり，また，シミュレーションの実行時間が膨大になるため，SAN のルーティングアルゴリズムに関する性能評価では確率モデル

シミュレーションが用いられることが多い．このため，本研究においても，性能評価に確率モデルシミュレーションを用いている．

5.1.2　シミュレーション条件

L-turn および R-turn ルーティングと BFS および DFS Up*/Down* ルーティングについて，(1) ルーティング方式，(2) トポロジ，(3) トポロジサイズ，(4) トラフィックパターン，の組み合わせを変えてそれぞれ評価を行なった．

これらは次の通りに指定した．

(1) ルーティング方式

  (a) 分散ルーティング方式 (適応型ルーティング)

  (b) ソースルーティング方式 (固定型ルーティング)

(2) トポロジ

  (a) イレギュラーネットワーク

  (b) 2 次元メッシュ

  (c) 2 次元トーラス

(3) トポロジサイズ

  (a) 16 スイッチ (メッシュとトーラスでは 4×4 スイッチ構成)

  (b) 64 スイッチ (メッシュとトーラスでは 8×8 スイッチ構成)

(4) トラフィックパターン

  (a) uniform

  (b) bit-reversal

ソースルーティング方式の場合，各スイッチ間の経路は Sancho の経路選択アルゴリズム [JA00] を用いて決定された 1 つの経路だけを常に用いるようにした．このアルゴリズムは，crossing path が可能な限り小さくなるように経路選択を行なうという点が特徴であり，これにより，固定型ルーティングにおける効率的なトラフィック分散の実現を図っている．ソースルーティング方式における評価は，適応型ルーティングである各ルーティングアルゴリズムを固定型ルーティングとして利用した場合の性能比較を目的としている．

イレギュラーネットワークについては，それぞれ 20 パターンの異なるトポロジを生成してシミュレーションを行ない，それらの結果の平均値について評価を行なった．スパニングツリーは，BFS スパニングツリーとして，第 3.1.1 節で述べた，minimum depth スパニングツリー (MDST)，DFS スパニングツリーとして，第 3.1.2 節で述べた，ヒューリスティックルールに基づいた DFS スパニングツリーを用いた．前者は，BFS Up*/Down* ルーティングと L-turn および R-turn ルーティングの構築に用い，後者は，DFS Up*/Down* ルーティングの構築に用いた．各ルーティングアルゴリズムにおいて，スパニングツリー

のルートスイッチは，第 3.1.2.4 節で述べた，crossing path と average distance の値により決定するヒューリスティックルールにより選択した．また，L-turn および R-turn ルーティングの H/V グラフ構築時の前順走査における訪問スイッチ選択は，第 4.2.5 節で述べた more upper-channel first を用いて行なった．

ルートスイッチの選択と訪問スイッチの選択が L-turn および R-turn ルーティングの性能に影響を与えることを，評価結果の一例を挙げてそれぞれ 第 5.2.4 節 および第 5.2.5 節 でそれぞれ示す．

シミュレーションにおいて，各パケットの目的地は，次の 2 つのトラフィックパターンにより決定した．

- uniform
  すべての目的地はランダムに決定され，均一に分散される．

- bit reversal
  まず，各 PC に 0 から $n-1$ (n は PC 数) までの昇順の 2 進数の番号を割当てる．2 進数の番号（$a_0, a_1, \cdots, a_{n-2}, a_{n-1}$）を持つ PC は，自身の番号のビット列を逆順に並べた番号（$a_{n-1}, a_{n-2}, \cdots, a_1, a_0$）の PC を目的地とする．なお，ビット列を逆順に並べた番号が，自身と同じである場合には，全ビットを反転した番号の PC にパケットを送るものとした．

次に，シミュレーションにおけるその他の共通パラメータを述べる．

表 5.1: 共通シミュレーションパラメータ

実行時間	500,000 クロック
スイッチのポート数	8
スイッチあたりの接続 PC 数	4
チャネル数	物理チャネル 1 本
パケット長	128 フリット
パケット転送方式	VCT 方式
1hop に要するフリット転送時間	最低 23 クロック
OSF(分散ルーティング方式時)	ランダム

シミュレータのスイッチのポート数は，既存の Myrinet スイッチ M3F-SW8 [1] および RHiNET-2/SW[STH+00] を想定して，8 ポートとした．8 ポートのうち，4 ポートは各々異なる PC に直結し，残りの 4 ポートは他のスイッチとの接続に利用した．パケットの先頭フリットがスイッチに到着してから隣接スイッチに転送されるまでのフリット転送時間は，ルーティングを行ないクロスバを通過可能となるまでに最短で 20 クロック，スイッチ内のクロスバの移動に 1 クロック，スイッチ間の移動に 2 クロック要するものとした．L-turn ルーティングと R-turn ルーティングは，仮想チャネルを持たない SAN を主な対象としているため，シミュレーションにおいて，各スイッチは，1 本の物理チャネルだけを

[1] http://www.myrinet.com/myrinet/product_list.html

利用するものとした．分散ルーティング方式の場合，選択可能な物理チャネルの中からランダムに出力物理チャネルを選択するランダム選択機構 [SB97] を OSF として使用した．

5.1.3　評価指標

次の指標について評価を行った．

スループット

全 PC がクロックあたりに受信するフリット数の平均値を受信トラフィックとし，飽和時点の受信トラフィック値をスループットとした．受信トラフィックは，全 PC が毎クロックに 1 フリット受信する場合を 1 とした．

レイテンシ

出発地の PC が，パケットの先頭フリットを NIC の入力バッファに挿入した時刻を t_0，目的地の PC の NIC がパケットの末尾のフリットを受け取った時刻を t_1 とする．ここで，$t_1 - t_0$ をレイテンシとした．

経路制限の度合いを示す静的な評価指標

- MPR(Minimal Path Rate)
  全経路のうち，トポロジ的な最短経路と等しい経路の割合 (%) を MPR とした．
- PT(Prohibited Turns)
  各スイッチにおける禁止ターン数の平均値を PT とした．

上記の 2 つの評価指標について，MPR はより大きいほど，PT はより小さいほど，ルーティングアルゴリズムによる経路制限がより小さいことを示す目安となる．

禁止ターン分散の度合いを示す静的な評価指標

- $SDPT$(Standard Deviation of Prohibited Turns)
  PT の標準偏差を $SDPT$ とした．
- PPT(Pairs of Prohibited Turns)
  各スイッチにおける禁止ターンペア数の平均値を PPT とした．

上記の 2 つの指標は，より小さいほど，禁止ターンがより均等に分散されていることを示す目安となる．

経路分散の度合いを示す静的な評価指標

- $CPUP$(Crossing routing Paths on UP channel)
  up **方向に向かう各チャネルを通過する経路数** (crossing routing path) **の平均値を** $CPUP$ **とした．**$CPUP$ **は，ルートスイッチ方向へのトラフィック量の目安となる．**

- $CPDW$(Crossing routing Paths on DoWn channel)
  down **方向に向かう各チャネル上の** crossing routing path **の平均値を** $CPDW$ **とした．**$CPDW$ **は，葉スイッチ方向へのトラフィック量の目安となる．**

5.2　分散ルーティング方式における評価結果

まず，分散ルーティング方式 (適応型ルーティング) における評価結果を各トポロジについて，順に示す．

5.2.1　イレギュラーネットワークにおける評価

(1) スループットの評価

16 **および** 64 **スイッチのイレギュラーネットワークにおける各ルーティングアルゴリズムの** uniform **および** bit-reversal **トラフィックにおけるスループットの平均値を，表** 5.2 **に示す．**

表 5.2 **より，すべての条件において，**2 **つの** L-turn **ルーティングのいずれかがもっとも高いスループットを実現していることがわかる．**BFS Up*/Down* **ルーティングに対するスループット向上は，**16 **スイッチの場合に約** 7**~** 9%**，**64 **スイッチの場合に約** 22**~** 29%**となっており，**DFS Up*/Down* **ルーティングに対する性能向上は，**16 **スイッチの場合に約** 3**~** 9%**，**64 **スイッチの場合に約** 8**~** 14% **となっている．これより，スイッチ数が大きくなるほど** L-turn **ルーティングの効果が大きくなっていることがわかる．**L-turn **ルーティングのスループットは，**uniform **トラフィックの場合よりも** bit-reversal **トラフィックの場合の方がより高くなっているが，この傾向は，他のルーティングアルゴリズムについても同様となっている．**L-turn/α **と** L-turn/β **のスループット差は，最大でも** 64 **スイッチの場合の約** 3% **であり，これらは，ほぼ同等のスループットを達成しているといえる．**

一方，2 **つの** R-turn **ルーティングは，**16 **スイッチの場合に** BFS Up*/Down* **ルーティングとほぼ同等のスループットを示すにとどまり，**64 **スイッチの場合には，もっとも低いスループットにまで落ちこんでいることがわかる．**L-turn **ルーティングとは対照的に，スイッチ数が大きくなるほどスループット低下が大きくなっており，特に，**64 **スイッチの** bit-reversal **トラフィックでは，**BFS Up*/Down* **ルーティングに対して，最大となる約** 16%**のスループット低下となっている．**R-turn/α **と** R-turn/β **のスループット差は，最大でも** 64 **スイッチの場合の約** 3% **であり，**L-turn **ルーティングと同様に，ほぼ同等のスループットを実現しているといえる．**

DFS Up*/Down* **ルーティングは，**16 **スイッチの** bit-reversal **トラフィックを除いて，**2 **つの** L-turn **ルーティングに次ぐスループットを実現しており，**L-turn **と同様に，スイッチ数が大きくなるほどスループット向上の度合いが大きくなっていることがわかる．**

表 5.2: イレギュラーネットワークにおける平均スループット

	16 スイッチ		64 スイッチ	
	Uniform	Bit-reversal	Uniform	Bit-reversal
BFS Up*/Down*	0.1050	0.1332	0.0357	0.0389
DFS Up*/Down*	0.1090	0.1334	0.0383	0.0451
L-turn/α	0.1124	0.1435	0.0434	0.0486
L-turn/β	0.1122	0.1450	0.0438	0.0500
R-turn/α	0.1053	0.1343	0.0331	0.0336
R-turn/β	0.1032	0.1347	0.0340	0.0338

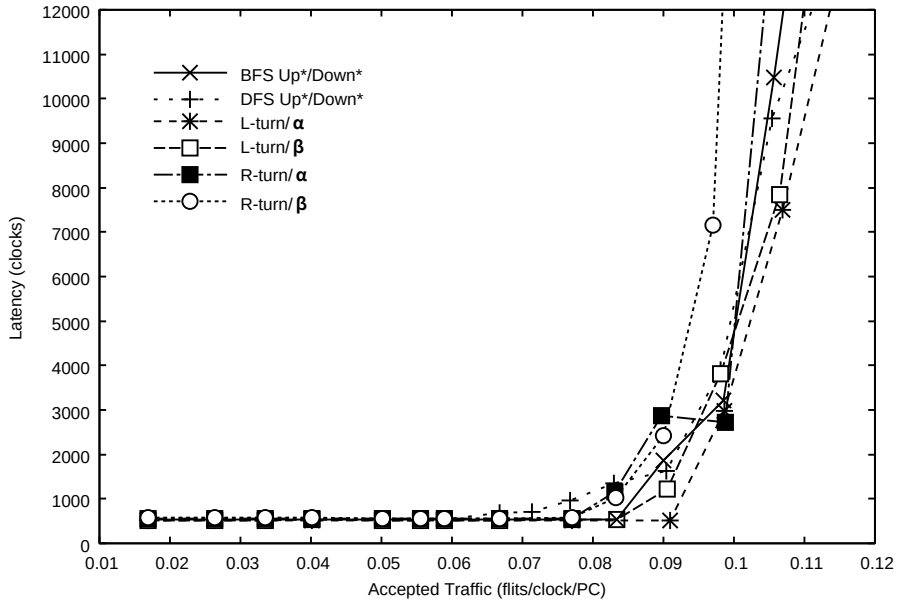
(2) レイテンシの評価

次に，16 および 64 スイッチのイレギュラーネットワークにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおける受信トラフィックとレイテンシの関係を示したグラフを，図 5.1 および図 5.2 にそれぞれ示す．図 5.1 および図 5.2 は，ランダムに生成した 20 のトポロジのうち，各 L-turn ルーティングと R-turn ルーティングのレイテンシが平均に近いトポロジにおける結果を示している．図において，各ルーティングアルゴリズムのレイテンシの優劣の傾向はスループットとほぼ同様であり，L-turn ルーティングがすべての条件でもっとも低いレイテンシを実現する一方で，R-turn ルーティングがもっとも高いレイテンシとなっている．

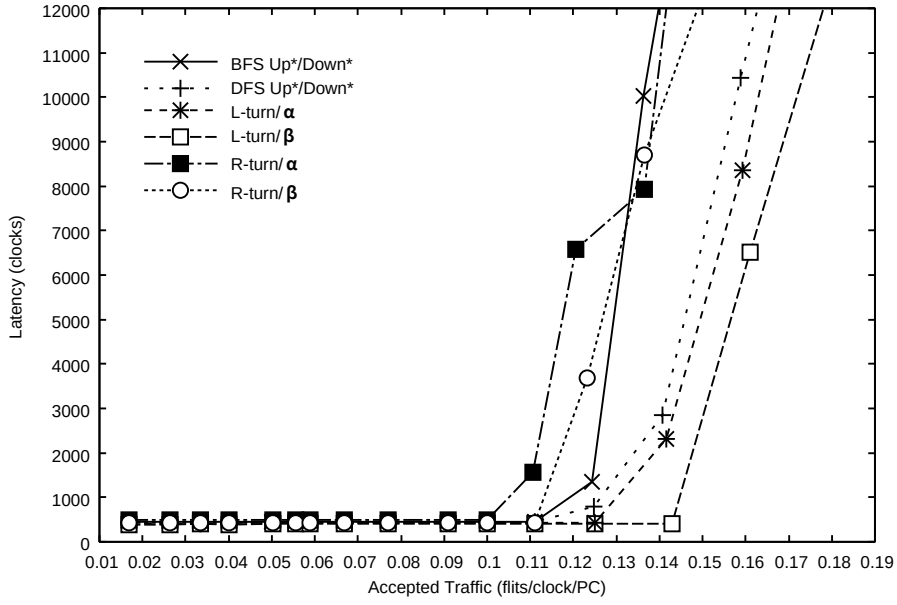
(3) 静的な評価指標の評価

次に，16 および 64 スイッチのイレギュラーネットワークにおける，各ルーティングアルゴリズムの静的な評価指標について，表 5.3 および表 5.4 にそれぞれ示す．まず，表 5.3 において，経路制限の度合いを示す MPR と PT については，DFS Up*/Down* ルーティングが若干優れた値を示し，その他のルーティングについては，ほぼ同等であることがわかる．これより，DFS Up*/Down* ルーティングは，禁止ターン数を減らすことにより，他のルーティングよりも多くの最短経路を確保していると考えられる．これに対し，禁止ターン分散の度合いを示す $SDPT$ と PPT については，各 L-turn および R-turn ルーティングが，各 Up*/Down* ルーティングに比べて大幅に小さな値を実現していることがわかる．各 L-turn および R-turn ルーティングの $SDPT$ と PPT は，ほぼ同等であり，BFS Up*/Down* ルーティングに対して，$SDPT$ については約 40%，PPT については約 80% の減少を実現し，同様に，DFS Up*/Down* ルーティングに対しては，それぞれ約 25%，約 75%の減少を実現している．これより，L-turn および R-turn ルーティングは，Up*/Down* ルーティングに比べて，より均等な禁止ターン分散を実現しているといえる．

しかし，ほぼ同等の $SDPT$ と PPT を実現しているにもかかわらず，表 5.2 において，L-turn ルーティングがスループット向上を実現する一方で，R-turn ルーティングは，Up*/Down* ルーティングと同程度のスループットの実現にとどまっている．この原因は，経路分散の度合いを示す $CPUP$ と $CPDW$ の違いによるものと考えられる．表 5.3 に

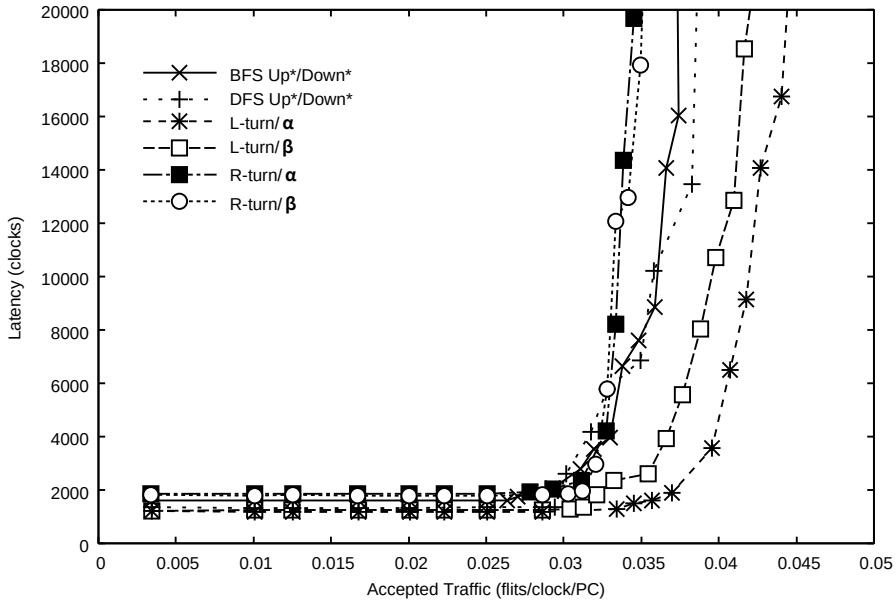


(a) Uniform

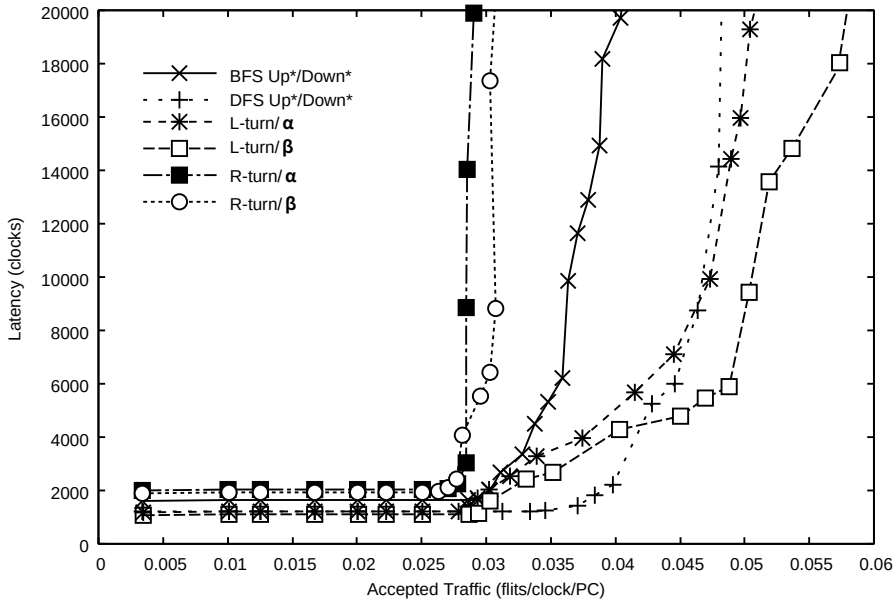


(b) Bit-reversal

図 5.1: イレギュラーネットワーク (16 スイッチ) における受信トラフィックと平均レイテンシ



(a) Uniform



(b) Bit-reversal

図 5.2: イレギュラーネットワーク (64 スイッチ) における受信トラフィックと平均レイテンシ

おいて，L-turn ルーティングでは，$CPUP$ より $CPDW$ の値が大きくなっていることから，ルートスイッチ方向よりも葉スイッチ方向により多くのトラフィックが流れやすくなっていると考えられる．これにより，L-turn ルーティングでは，ルートスイッチ周辺のホットスポット発生が緩和され，葉スイッチ周辺により多くのトラフィックが分散されたことにより，スループットの向上を実現したものと考えられる．一方，R-turn ルーティングでは，$CPDW$ より $CPUP$ の値が大きくなっていることから，L-turn ルーティングと対照的に，ルートスイッチ方向にトラフィックが流れやすくなり，ルートスイッチ周辺のホットスポットがより発生しやすくなっているものと考えられる．この原因として，次のような R-turn ルーティングの禁止ターンの特性が考えられる．R-turn ルーティングでは，ターン $T_{RD,LU}$ を除く LU 方向へのターンをすべて許可しているため，これによりパケットがルートスイッチ方向に集中しやすくなるものと考えられる．一方，RD 方向からその他の方向へのターンは禁止されているため，対照的に葉方向にはパケットが転送されにくくなっているものと考えられる．これらの特性により，R-turn ルーティングでは，L-turn ルーティングと異なり，スループット向上が実現できなかったものと考えられる．

なお，各 Up*/Down* ルーティングにおける $CPUP$ と $CPDW$ は等しくなっているが，これは， Up*/Down* ルーティングでは，(1) up と down の 2 つの方向しか存在しない，(2) 任意のスイッチ間には，互いに反対方向に向かう対照的な 2 つの経路が存在する，という特性のためである．スパニングツリーベースの有向グラフでは，葉スイッチ周辺よりもルートスイッチ周辺に近づくにつれ利用可能な経路が少なくなるため，$CPUP$ と $CPDW$ が等しい場合，ルートスイッチ方向にトラフィックが偏りやすくなると考えられる．

表 5.3: イレギュラーネットワーク (16 スイッチ) における静的な評価指標

	MPR	PT	$SDPT$	PPT	$CPUP$	$CPDW$
BFS Up*/Down*	89.6	3.181	3.723	1.591	11.40	11.40
DFS Up*/Down*	92.9	2.863	3.061	1.431	11.73	11.73
L-turn/α	88.9	3.175	2.264	0.366	10.76	12.54
L-turn/β	88.8	3.172	2.379	0.388	10.66	12.78
R-turn/α	88.8	3.184	2.300	0.375	12.54	10.78
R-turn/β	88.6	3.163	2.328	0.369	12.64	10.74

次に，表 5.4 についてみると，64 スイッチの場合の各ルーティングアルゴリズムの優劣の傾向は，16 スイッチの場合の表 5.3 と同様であり，各 L-turn および R-turn ルーティングがもっとも均等な禁止ターンの分散を実現していることがわかる．ただし，64 スイッチの場合には，各ルーティングアルゴリズムの MPR が 16 スイッチの場合に比べて約 20~25% 減少していることから，トポロジ的な最短経路が確保しにくくなっているといえる．これにより，スイッチ数が増加した場合には，禁止ターンのより均等な分散と葉スイッチ方向へのトラフィックの分散が，スループット向上のためにより重要となるものと考えられる．表 5.2 において，64 スイッチの場合に，L-turn ルーティングのスループット向上の割合が増加する一方で，R-turn ルーティングのスループット低下の割合が増加した原因は，このためであると考えられる．

表 5.4: イレギュラーネットワーク (64 スイッチ) における静的な評価指標

	MPR	PT	$SDPT$	PPT	$CPUP$	$CPDW$
BFS Up*/Down*	64.2	2.994	3.626	1.497	86.16	86.16
DFS Up*/Down*	72.9	2.602	2.454	1.301	85.54	85.54
L-turn/α	66.9	2.890	2.288	0.316	82.94	91.63
L-turn/β	67.1	2.866	2.271	0.306	82.47	91.69
R-turn/α	67.0	2.886	2.281	0.316	91.14	82.38
R-turn/β	67.0	2.863	2.269	0.302	91.96	82.74

以上より，表 5.2 において，L-turn ルーティングが DFS Up*/Down* ルーティングよりも高いスループットを実現していることから，スループットの向上には，禁止ターン数の削減よりも禁止ターンのより均等な分散による効果の方が大きいと考えられる．ただし，先に述べた L-turn ルーティングと R-turn ルーティングの性能差の原因から，スループット向上のためには，より均等な禁止ターンの分散と葉スイッチ方向へのトラフィック分散の両立が重要であると考えられる．

5.2.2　2 次元メッシュにおける評価

(1) スループットの評価

4×4 および 8×8 スイッチの 2 次元メッシュにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおけるスループットを，表 5.5 に示す．イレギュラーネットワークにおける結果と同様に，すべての条件において，L-turn ルーティングがもっとも高いスループットを実現しており，スループット向上はスイッチ数が大きいほどより大きくなっている．BFS Up*/Down* ルーティングに対するスループット向上は，イレギュラーネットワークの場合よりも大きく，約 10~ 50%の向上となっている．同様に，R-turn ルーティングは，すべての条件において，BFS Up*/Down* ルーティングに対してスループットが低下しており，スループット低下はスイッチ数が大きいほどより大きくなっている．BFS Up*/Down* ルーティングに対するスループット低下は，同様に，イレギュラーネットワークの場合よりも大きく，約 10~ 30%の低下となっている．L-turn/α と L-turn/β および R-turn/α と R-turn/β のスループットがそれぞれ同じ値となっているが，これは，4×4 および 8×8 スイッチ構成の 2 次元メッシュにおいて，これらによる全スイッチ間の経路 (禁止ターン分布) がそれぞれまったく同じものとなっているためである．

イレギュラーネットワークにおける結果と異なり，2 次元メッシュでは，DFS Up*/Down* ルーティングのスループット低下が顕著となっている．16 スイッチの uniform トラフィックの場合を除いたすべての条件で，もっとも低いスループットを示し，特に，BFS Up*/Down* ルーティングに対しては，最大で約 40%，L-turn ルーティングに対しては，最大で約 60%の低下となっている．

表 5.5: 2 次元メッシュにおけるスループット

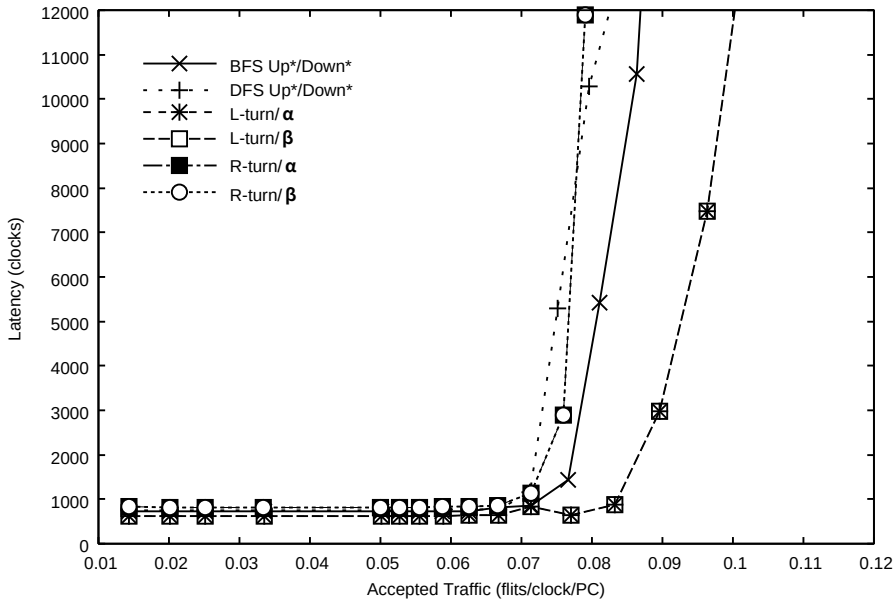
	4×4 スイッチ		8×8 スイッチ	
	Uniform	Bit-reversal	Uniform	Bit-reversal
BFS Up*/Down*	0.0863	0.0877	0.0357	0.0380
DFS Up*/Down*	0.0796	0.0601	0.0215	0.0226
L-turn/α	0.0963	0.1069	0.0510	0.0575
L-turn/β	0.0963	0.1069	0.0510	0.0575
R-turn/α	0.0791	0.0769	0.0257	0.0300
R-turn/β	0.0791	0.0769	0.0257	0.0300

(2) レイテンシの評価

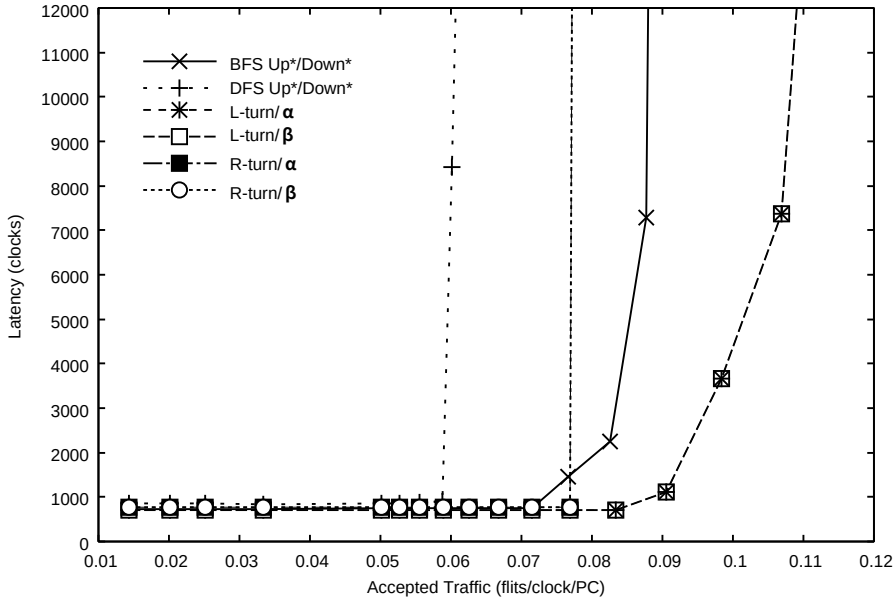
次に，4×4 および 8×8 スイッチの 2 次元メッシュにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおける受信トラフィックとレイテンシの関係を示したグラフを，図 5.3 および図 5.4 にそれぞれ示す．図より，各ルーティングアルゴリズムのレイテンシの優劣の傾向は，イレギュラーネットワークと同様に，スループットとほぼ同様であり，L-turn ルーティングが，すべての条件でもっとも低いレイテンシを実現している．R-turn ルーティングは，DFS Up*/Down* ルーティングに比べると，ほぼ同等またはより低いレイテンシを実現しているが，BFS Up*/Down* ルーティングに対しては，より高いレイテンシとなっている．

(3) 静的な評価指標の評価

次に，4×4 および 8×8 スイッチの 2 次元メッシュにおける，各ルーティングアルゴリズムの静的な評価指標について，表 5.6 および表 5.7 にそれぞれ示す．表 5.6 および表 5.7 より，経路制限に関する MPR と PT については，8×8 スイッチにおける DFS Up*/Down* ルーティングをのぞいて，各ルーティングアルゴリズムは，ほぼ同等の値を示すことがわかる．特に，MPR の値から，2 次元メッシュにおける経路のほとんどは，トポロジ的な最短経路となることがわかる．これに対し，禁止ターン分散の度合いを示す $SDPT$ と PPT については，8×8 スイッチにおける BFS Up*/Down* ルーティングの $SDPT$ をのぞいて，イレギュラーネットワークの場合と同様に，各 L-turn および R-turn ルーティングが，もっとも小さい値を実現しており，特に PPT については，4×4 スイッチで 0，8×8 スイッチで Up*/Down* ルーティングに対して 約 92% の減少となり，大幅に禁止ターンのペアを削減していることがわかる．これより，イレギュラーネットワークの場合と同様に，L-turn ルーティングと R-turn ルーティングが，より均等な禁止ターンの分散を実現しているといえる．また，経路分散に関する各ルーティングアルゴリズムの $CPUP$ と $CPDW$ の差についても，イレギュラーネットワークと同様の傾向となっていることがわかる．このため，同様に，L-turn ルーティングにおいてはスループットが向上し，R-turn ルーティングにおいてはスループットが低下したものと考えられる．

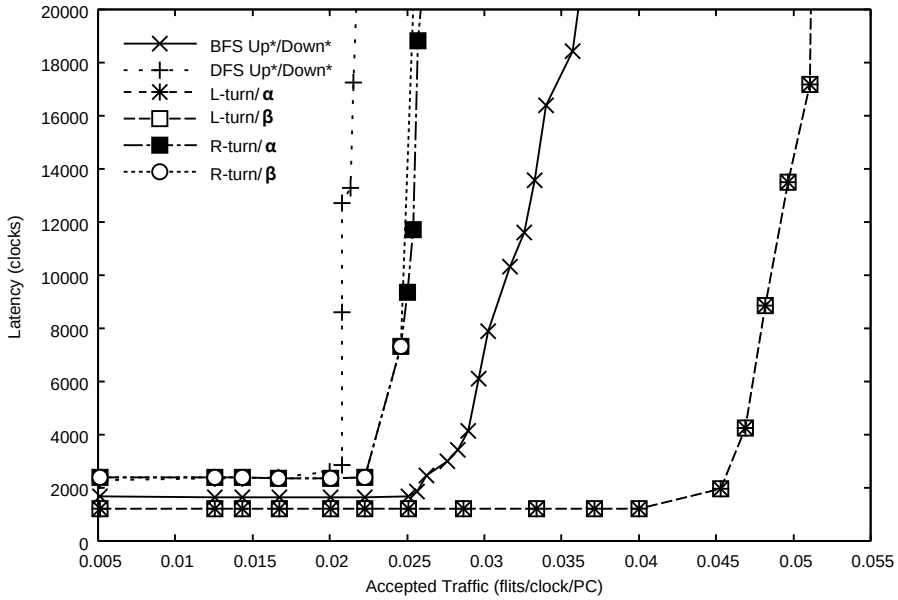


(a) Uniform traffic

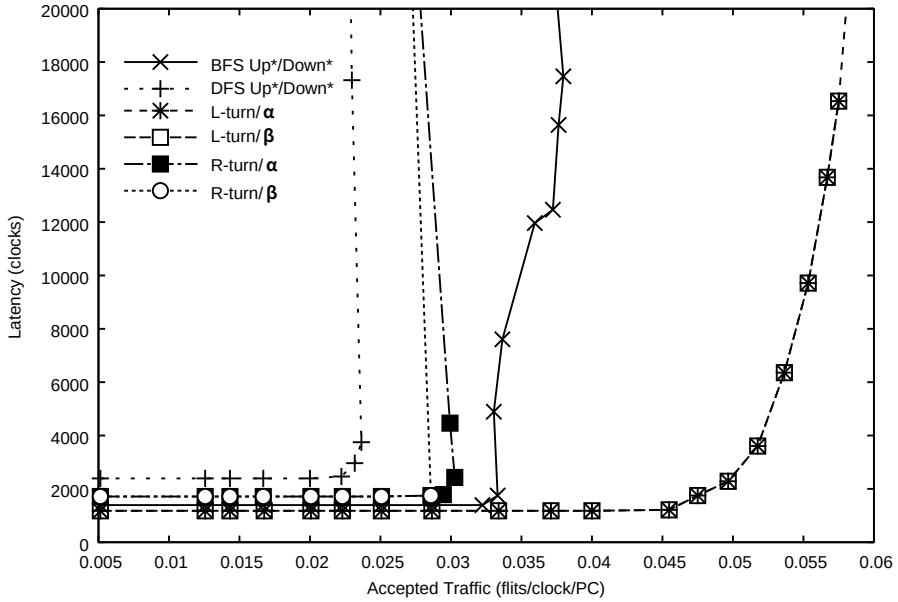


(b) Bit-reversal traffic

図 5.3: 2 **次元メッシュ**(4×4 **スイッチ**) **における受信トラフィックと平均レイテンシ**



(a) Uniform traffic



(b) Bit-reversal traffic

図 5.4: 2 **次元メッシュ**(8×8 **スイッチ**) **における受信トラフィックと平均レイテンシ**

表 5.6: 2 次元メッシュ(4×4 スイッチ) における静的な性能指標

	MPR	PT	$SDPT$	PPT	$CPUP$	$CPDW$
BFS Up*/Down*	100	1.125	0.992	0.563	35.67	35.67
DFS Up*/Down*	100	1.125	0.992	0.563	31.17	31.17
L-turn/α	100	1.125	0.781	0	28.17	36.08
L-turn/β	100	1.125	0.781	0	28.17	36.08
R-turn/α	100	1.125	0.781	0	36.08	28.17
R-turn/β	100	1.125	0.781	0	36.08	28.17

表 5.7: 2 次元メッシュ(8×8 スイッチ) における静的な性能指標

	MPR	PT	$SDPT$	PPT	$CPUP$	$CPDW$
BFS Up*/Down*	100	1.531	0.847	0.766	1671.43	1671.43
DFS Up*/Down*	83.5	1.750	1.953	0.875	622.96	622.96
L-turn/α	99.0	1.578	0.932	0.063	1075.21	1774.31
L-turn/β	99.0	1.578	0.932	0.063	1075.21	1774.31
R-turn/α	99.0	1.578	0.932	0.063	1774.31	1075.21
R-turn/β	99.0	1.578	0.932	0.063	1774.31	1075.21

5.2.3　2 次元トーラスにおける評価

(1) スループットの評価

4×4 および 8×8 スイッチの 2 次元トーラスにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおけるスループットを，表 5.8 に示す．表より，その他のトポロジにおける結果と同様に，すべての条件において，L-turn ルーティングがもっとも高いスループットを実現しており，また，スループット向上はスイッチ数が大きいほどより大きくなっていることがわかる．BFS Up*/Down* ルーティングに対するスループット向上は，もっとも大きく，約 16~ 83%の向上となっている．一方，R-turn ルーティングは，同様に，一部の条件をのぞいて，BFS Up*/Down* ルーティングに対してスループットが低下していることがわかる．しかし，BFS Up*/Down* ルーティングに対するスループット低下は，2 次元メッシュよりは小さく，最大で 約 20%の低下となっている．2 次元メッシュほど顕著ではないが，2 次元トーラスにおいても DFS Up*/Down* ルーティングのスループットは，一部をのぞいて，BFS Up*/Down* ルーティングに比べて低下しており，BFS Up*/Down* ルーティングに対しては，最大で約 20%，L-turn ルーティングに対しては，最大で約 55% の低下となっている．

(2) レイテンシの評価

次に，4×4 および 8×8 スイッチの 2 次元トーラスにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおける受信トラフィックとレイテンシの関係を示したグラフを，図 5.5 および図 5.6 にそれぞれ示す．

表 5.8: 2 次元トーラスにおけるスループット

	4×4 スイッチ		8×8 スイッチ	
	Uniform	Bit-reversal	Uniform	Bit-reversal
BFS Up*/Down*	0.1195	0.1356	0.0386	0.0383
DFS Up*/Down*	0.1201	0.1080	0.0362	0.0320
L-turn/α	0.1385	0.1590	0.0623	0.0655
L-turn/β	0.1392	0.1574	0.0583	0.0700
R-turn/α	0.1200	0.1088	0.0316	0.0331
R-turn/β	0.1211	0.1104	0.0330	0.0393

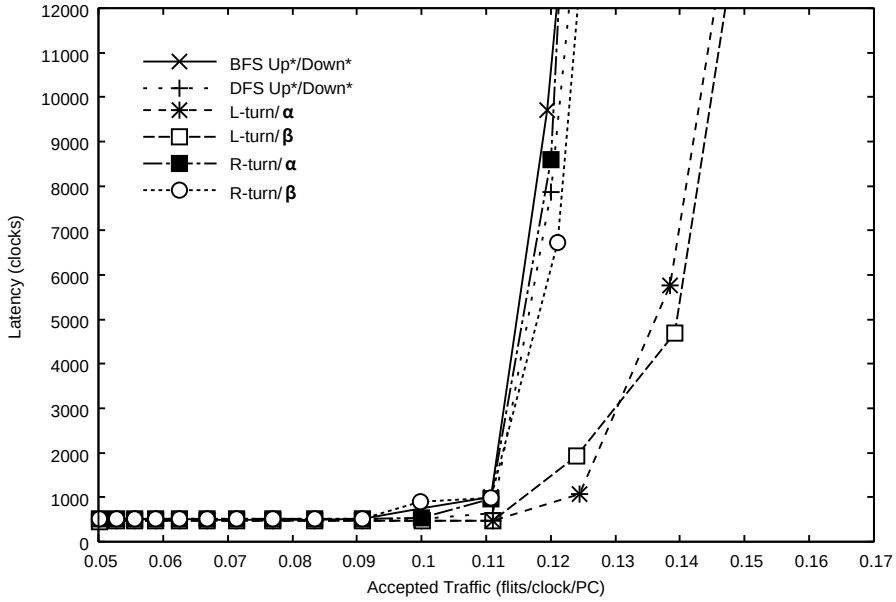
図より，各ルーティングアルゴリズムのレイテンシの優劣の傾向は，その他のトポロジと同様に，スループットの場合とほぼ同様であり，L-turn ルーティングがすべての条件でもっとも低いレイテンシを実現している．

(3) 静的な評価指標の評価

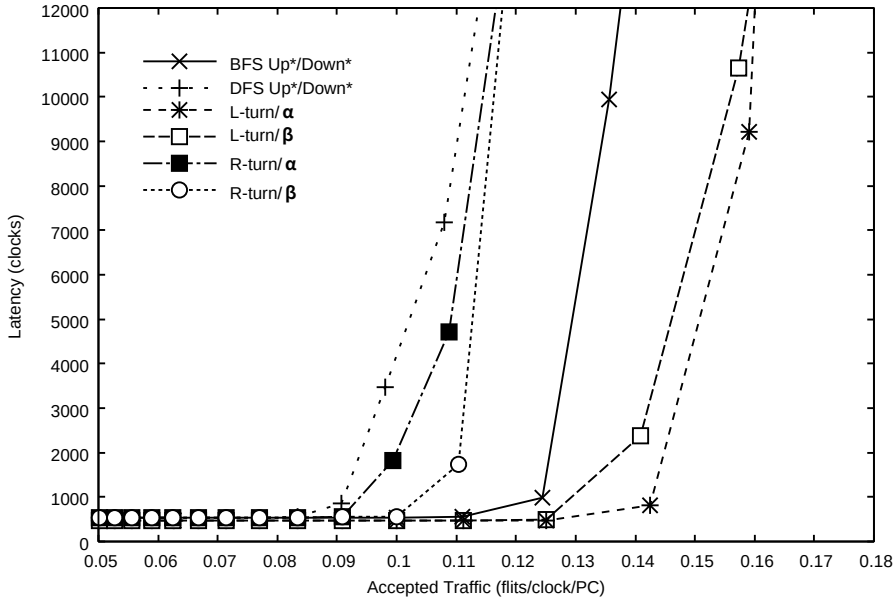
次に，4×4 および 8×8 スイッチの 2 次元トーラスにおける，各ルーティングアルゴリズムの静的な評価指標について，表 5.9 および表 5.10 にそれぞれ示す．表 5.9 および表 5.10 より，経路制限に関する MPR と PT については，各ルーティングアルゴリズムは，ほぼ同等の値を示すことがわかる．これに対し，禁止ターン分散の度合いを示す $SDPT$ と PPT については，その他のトポロジと同様に，各 L-turn および R-turn ルーティングが，もっとも小さい値を示しており，同様に，より均等な禁止ターンの分散を実現していることがわかる．また，経路分散に関する各ルーティングアルゴリズムの $CPUP$ と $CPDW$ の差についても，その他のトポロジと同様の傾向であり，同様に，L-turn ルーティングにおけるスループット向上と R-turn ルーティングにおけるスループット低下につながっているものと考えられる．

表 5.9: 2 次元トーラス (4×4 スイッチ) における静的な性能指標

	MPR	PT	$SDPT$	PPT	$CPUP$	$CPDW$
BFS Up*/Down*	100	3	3.240	1.5	22.00	22.00
DFS Up*/Down*	100	3	3.240	1.5	22.25	22.25
L-turn/α	100	3	2.208	0.438	19.75	29.06
L-turn/β	100	3	2.208	0.438	19.75	29.06
R-turn/α	100	3	2.208	0.438	29.06	19.75
R-turn/β	100	3	2.208	0.438	29.06	19.75

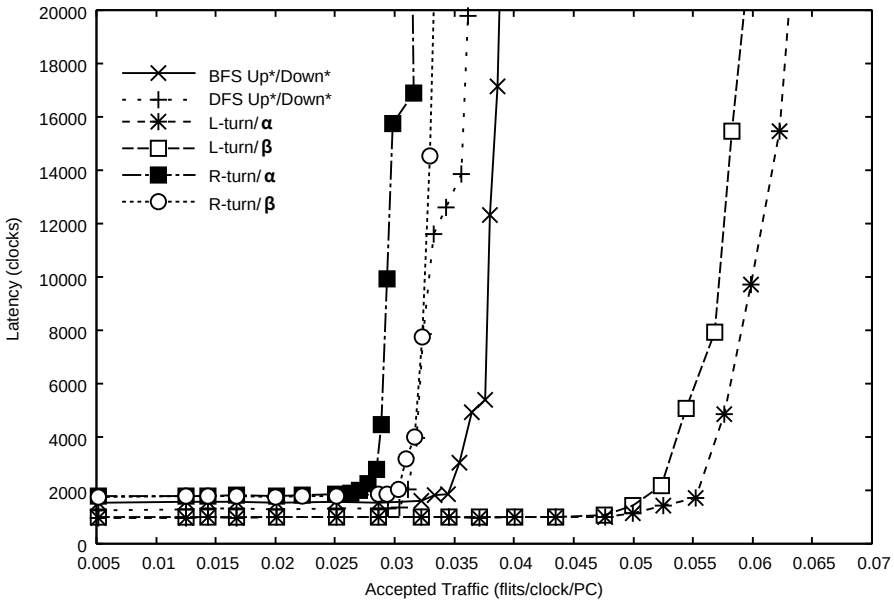


(a) Uniform traffic

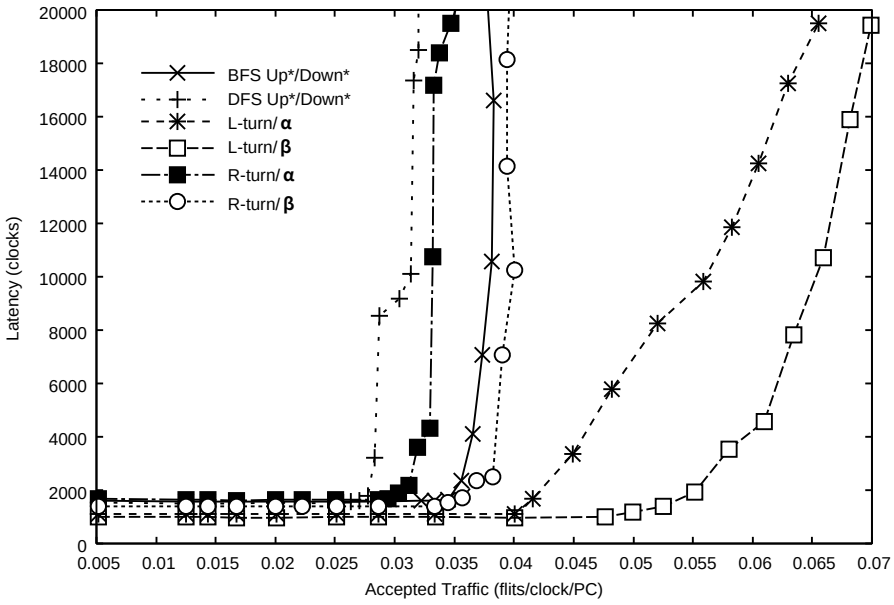


(b) Bit-reversal traffic

図 5.5: 2 次元トーラス (4×4 スイッチ) における受信トラフィックと平均レイテンシ



(a) Uniform traffic



(b) Bit-reversal traffic

図 5.6: 2 **次元トーラス** (8×8 **スイッチ**) **における受信トラフィックと平均レイテンシ**

表 5.10: 2 次元トーラス (8×8 スイッチ) における静的な性能指標

	MPR	PT	$SDPT$	PPT	$CPUP$	$CPDW$
BFS Up*/Down*	81.9	2.5	2.264	1.25	1014.56	1014.56
DFS Up*/Down*	88.7	2.5	2.264	1.25	654.25	654.25
L-turn/α	86.5	2.516	1.601	0.234	411.07	689.68
L-turn/β	86.3	2.516	1.649	0.234	405.73	691.76
R-turn/α	86.5	2.516	1.601	0.234	689.68	411.07
R-turn/β	86.3	2.516	1.649	0.234	691.76	405.73

5.2.4　ルートスイッチ選択の影響

スパニングツリーの構造は，ルートスイッチの選択により変化するため，Up*/Down* ルーティングの性能は，ルートスイッチ選択の影響を受ける [JA00]．ここでは，スパニングツリーベースである L-turn および R-turn ルーティングの性能も，同様にルートスイッチ選択により影響を受けることを示す．

表 5.11 および表 5.12 は，16 スイッチおよび 64 スイッチのイレギュラーネットワークにおける L-turn/α および R-turn/α ルーティングについて，ルートスイッチの選択を，(R1) 先の評価で用いた Sancho のヒューリスティックルールに従ってルートスイッチを決定，(R2) 同ヒューリスティックルールにおいて，最悪の場合の選択を行なうように変更してルートスイッチを決定，という正反対の 2 つのポリシーに基づいて行なった場合の，スループットと静的な評価指標を示している (L-turn/β および R-turn/β においても同等の結果となるため，ここでは省略する)．各トポロジサイズにおいて，先の評価と同じ 20 の異なるトポロジが用いられており，トラフィックパターンとして，uniform トラフィックを用いている．

表 5.11: イレギュラーネットワーク (16 スイッチ，Uniform) における平均スループットと静的な評価指標

	Throughput	MPR	PT	$SDPT$	PPT
L-turn/α(R1)	0.1124	88.9	3.175	2.264	0.366
L-turn/α(R2)	0.1022	86.8	3.256	2.553	0.453
R-turn/α(R1)	0.1053	88.8	3.184	2.300	0.375
R-turn/α(R2)	0.0919	87.1	3.234	2.515	0.431

表 5.11 および表 5.12 より，L-turn/α および R-turn/α ルーティングにおいて，ルートスイッチ選択を R2 により行なった場合は，いずれも R1 に対して，スループットとすべての静的な評価指標が劣ることがわかる．特に，R2 を用いた場合の L-turn/α ルーティングのスループットは，R1 を用いた場合の R-turn/α ルーティングのスループットに対して，16 スイッチの場合で劣っており，64 スイッチの場合についてもほぼ同じ程度までに減少することがわかる．

以上より，L-turn および R-turn ルーティングの性能は，ルートスイッチ選択により影響を受けることが確認された．

表 5.12: イレギュラーネットワーク (64 スイッチ，Uniform) における
平均スループットと静的な評価指標

	Throughput	MPR	PT	$SDPT$	PPT
L-turn/α(R1)	0.0434	66.9	2.890	2.288	0.316
L-turn/α(R2)	0.0346	64.7	2.950	2.381	0.357
R-turn/α(R1)	0.0331	67.0	2.886	2.281	0.316
R-turn/α(R2)	0.0276	64.7	2.948	2.384	0.355

5.2.5　訪問スイッチ選択ポリシーの影響

ここでは，第 4.2.5 節で述べた，H/V グラフ構築時の前順走査における訪問スイッチ選択ポリシーが性能に与える影響を示し，4 つの選択ポリシーのうち，more upper-channel first がもっとも高い性能を実現することを示す．

表 5.13 および表 5.14 は，16 スイッチおよび 64 スイッチのイレギュラーネットワークにおける L-turn/α および R-turn/α ルーティングについて (L-turn/β および R-turn/β においても同等の結果となるため，同様に省略した)，異なる 4 つの訪問スイッチ選択ポリシー，(V1) random, (V2) less child-node first, (V3) more child-node first (V4) more upper-channel first, を適用した場合のスループットと静的な評価指標をそれぞれ示している．各トポロジサイズにおいて，先の評価と同じ 20 の異なるトポロジが用いられており，トラフィックパターンとして，uniform を用いている．

表 5.13: イレギュラーネットワーク (16 スイッチ，Uniform) における
平均スループットと静的な評価指標

	Throughput	MPR	PT	$SDPT$	PPT
L-turn/α(V1)	0.1022	86.5	3.306	2.533	0.447
L-turn/α(V2)	0.0946	86.7	3.216	3.206	0.738
L-turn/α(V3)	0.1022	86.8	3.281	2.551	0.459
L-turn/α(V4)	0.1124	88.9	3.175	2.264	0.366
R-turn/α(V1)	0.0912	86.7	3.300	2.526	0.441
R-turn/α(V2)	0.0924	86.7	3.216	3.204	0.738
R-turn/α(V3)	0.0927	87.0	3.247	2.488	0.425
R-turn/α(V4)	0.1053	88.8	3.184	2.300	0.375

表 5.13 および表 5.14 より，L-turn/α および R-turn/α ルーティングにおいて，訪問スイッチ選択ポリシーとして more upper-channel first (V4) を用いた場合のスループットとすべての静的な評価指標は，その他 3 つの選択ポリシーを用いた場合よりも優れていることがわかる．特に，less child-node first (V2) を用いた場合の L-turn/α のスループットは，V4 を用いた場合の R-turn/α のスループットに対して，16 スイッチおよび 64 スイッチの場合でともに劣っており，訪問スイッチ選択ポリシーの違いによる影響が少なくないことがわかる．

以上より，H/V グラフ構築時の前順走査における訪問スイッチ選択ポリシーが，L-turn

および R-turn ルーティングの性能に影響を与えており，4 つの選択ポリシーのうち，more upper-channel first がもっとも高い性能を実現するということが確認された．

表 5.14: イレギュラーネットワーク (64 スイッチ，Uniform) における平均スループットと静的な評価指標

	Throughput	MPR	PT	$SDPT$	PPT
L-turn/α(V1)	0.0353	65.0	2.960	2.417	0.352
L-turn/α(V2)	0.0307	64.1	2.878	2.917	0.603
L-turn/α(V3)	0.0340	64.4	2.959	2.443	0.373
L-turn/α(V4)	0.0434	66.9	2.890	2.288	0.316
R-turn/α(V1)	0.0284	64.9	2.962	2.418	0.355
R-turn/α(V2)	0.0275	64.1	2.882	2.931	0.608
R-turn/α(V3)	0.0276	64.4	2.944	2.423	0.370
R-turn/α(V4)	0.0331	67.0	2.886	2.281	0.316

5.3　ソースルーティング方式における評価結果

次に，ソースルーティング方式 (固定型ルーティング) における評価結果を各トポロジについて，順に示す．

5.3.1　イレギュラーネットワークにおける評価

16 および 64 スイッチのイレギュラーネットワークにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおけるスループットの平均値を，表 5.15 に示す．表 5.15 より，分散ルーティング方式 (適応型ルーティング) の場合と同様に，L-turn ルーティングが，すべての条件で最も高いスループットを実現していることがわかる．BFS Up*/Down* ルーティングに対するスループット向上は，最大で約 25% であり，分散ルーティング方式の場合に比べて若干小さくなっているが，これは，選択経路が 1 つに固定されたことがトラフィック分散能力に影響を与えているためと考えられる．また，R-turn ルーティングについてもこれまでの傾向と同様に，一部を除いてスループットが低下していることがわかる．以上より，複数経路の選択ができない固定型ルーティングにおいても，L-turn および R-turn ルーティングのトラフィック分散能力がスループットに反映されているものと考えられる．

なお，ここでは省略するが，各ルーティングアルゴリズムのレイテンシについても，分散ルーティング方式の場合と同様に，スループットにおける優劣の傾向がほぼ同様に反映されることが確認されている．また，各ルーティングアルゴリズムの静的な評価指標については，分散ルーティング方式の場合と同等となるため，同様に省略する (2 次元メッシュおよび 2 次元トーラスについても同様)．

表 5.15: イレギュラーネットワークにおける平均スループット

	16 スイッチ		64 スイッチ	
	Uniform	Bit-reversal	Uniform	Bit-reversal
BFS Up*/Down*	0.1036	0.1442	0.0359	0.0394
DFS Up*/Down*	0.1047	0.1441	0.0376	0.0482
L-turn/α	0.1107	0.1476	0.0430	0.0486
L-turn/β	0.1102	0.1465	0.0436	0.0492
R-turn/α	0.1043	0.1384	0.0331	0.0324
R-turn/β	0.1047	0.1407	0.0329	0.0318

5.3.2　2 次元メッシュにおける評価

4×4 および 8×8 スイッチの 2 次元メッシュにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおけるスループットを，表 5.16 に示す．表 5.16 より，2 次元メッシュにおいても同様にして，L-turn ルーティングが，すべての条件で最も高いスループットを実現していることがわかる．R-turn ルーティングについては，8×8 スイッチの場合に，これまでと同様にスループットが低下しているが，4×4 スイッチの場合には，各 Up*/Down* ルーティングに対して高いスループットを実現していることがわかる．これは，この条件においては，Sancho の経路選択アルゴリズムの適用による経路分散の効果が R-turn ルーティングにおけるトラフィックの集中を改善する方向にうまく働いたためと考えられる．

表 5.16: 2 次元メッシュにおけるスループット

	4×4 スイッチ		8×8 スイッチ	
	Uniform	Bit-reversal	Uniform	Bit-reversal
BFS Up*/Down*	0.0681	0.0834	0.0295	0.0362
DFS Up*/Down*	0.0681	0.0728	0.0286	0.0298
L-turn/α	0.1045	0.1172	0.0398	0.0474
L-turn/β	0.1045	0.1172	0.0398	0.0474
R-turn/α	0.0873	0.0961	0.0274	0.0310
R-turn/β	0.0873	0.0961	0.0274	0.0310

5.3.3　2 次元トーラスにおける評価

4×4 および 8×8 スイッチの 2 次元トーラスにおける各ルーティングアルゴリズムの uniform および bit-reversal トラフィックにおけるスループットを，表 5.17 に示す．表 5.17 より，2 次元トーラスにおいても同様にして，L-turn ルーティングが，すべての条件で最も高いスループットを実現していることがわかる．一方，これまでと異なり，R-turn ルーティングについても，すべての条件で，BFS Up*/Down* ルーティングに対して高いス

ループットを実現していることがわかる．これは，2 次元メッシュの場合と同様に，Sancho の経路選択アルゴリズムの影響によるものと考えられる．

表 5.17: 2 次元トーラスにおけるスループット

	4×4 スイッチ		8×8 スイッチ	
	Uniform	Bit-reversal	Uniform	Bit-reversal
BFS Up*/Down*	0.0945	0.1067	0.0292	0.0325
DFS Up*/Down*	0.1047	0.1352	0.0316	0.0406
L-turn/α	0.1201	0.1547	0.0519	0.0552
L-turn/β	0.1198	0.1826	0.0514	0.0560
R-turn/α	0.1052	0.1351	0.0309	0.0407
R-turn/β	0.1058	0.1151	0.0314	0.0422

以上より，L-turn および R-turn ルーティングをソースルーティング方式 (固定型ルーティング) として適用した場合も，分散ルーティング方式 (適応型ルーティング) の場合と同等の効果が得られることがわかった．これにより，L-turn ルーティングは，固定型ルーティングとして適用しても性能向上が実現可能であるといえる．

5.4　まとめ

本章では，L-turn および R-turn ルーティングと BFS および DFS Up*/Down* ルーティングの性能評価を確率モデルシミュレーションにより行なった．シミュレーションの結果，L-turn ルーティングは，すべての条件で最も高いスループットを実現し，BFS Up*/Down* ルーティングに対して，イレギュラーネットワークにおいては最大で約 30%，レギュラーネットワークにおいては最大で約 80%のスループット向上を実現することが確認された．一方，R-turn ルーティングは，対照的にほとんどの条件で，最も低いスループットを示す結果となった．禁止ターンの分散に関する静的な評価指標から，L-turn ルーティングと R-turn ルーティングは，ほぼ同等の禁止ターン分散を実現することが確認されたが，選択された禁止ターン集合のパターンの違いにより，L-turn ルーティングでは，葉スイッチ方向にトラフィックが分散されやすくなるのに対し，R-turn ルーティングでは，ホットスポットが発生しやすいルートスイッチ方向にトラフィックが集中してしまうことがわかった．これより，スループット向上のためには，より均等な禁止ターンの分散と葉スイッチ方向へのトラフィック分散の両立が重要であることがわかった．また，L-turn および R-turn ルーティングは，ソースルーティング方式 (固定型ルーティング) として適用した場合も，分散ルーティング方式 (適応型ルーティング) の場合と同等の効果が得られることがわかった．これにより，L-turn ルーティングは，適応型ルーティングとしてだけでなく，固定型ルーティングとして適用した場合も性能向上が可能である有効なルーティングアルゴリズムであるといえる．

第6章　結 論

近年，従来の大規模並列計算機に代わる高性能並列分散コンピューティング環境として PC クラスタの普及が急速に進んでいる．PC クラスタにおける PC 間の接続に用いられる SAN は，拡張性および耐故障性などの要求からトポロジとして，イレギュラーネットワークをサポートすることが多い．イレギュラーネットワークにおけるルーティングアルゴリズムは，SAN の性能に影響を与える重要な要素であり，これまで多くの提案がなされている．その中でも，Up*/Down* ルーティングは，任意の SAN およびトポロジに適用可能であり，汎用性の高い適応型ルーティングアルゴリズムとして，一般的に利用されている．しかし，Up*/Down* ルーティングは，1 次元の有向グラフをベースとしているために禁止ターン分布の偏りが大きくなり，高スループット実現のためのトラフィックの分散が困難となるという問題を持つ．

本研究では，Up*/Down* ルーティングにおける上記の問題の解決と，Up*/Down* ルーティングと同等の高い汎用性の実現の両立を目的として，適応型ルーティングアルゴリズムである L-turn および R-turn ルーティングを提案し，確率モデルシミュレーションにより評価を行なった．L-turn および R-turn ルーティングは，Up*/Down* ルーティングで利用されているスパニングツリーベースの 1 次元有向グラフを拡張した H/V グラフと呼ばれる新たな 2 次元有向グラフを利用する．H/V グラフの導入により，形成可能なターンの数は従来の 6 倍である 12 パターンに増加し，これにより，トラフィックの分散を考慮したより柔軟な禁止ターンの選択を行なってデッドロックフリーを実現することが可能となる．そして，H/V グラフに対して，禁止ターンの分散を考慮した 2 次元 Turn モデルをシステマティックな手法で適用することにより，L-turn および R-turn ルーティングが定義される．この際，循環構造検出アルゴリズムを導入することにより，冗長な禁止ターンを削減し，更なる性能向上の実現を図っている．L-turn および R-turn ルーティングは，スパニングツリーの構築と 2 次元 Turn モデルの適用をベースとすることにより，Up*/Down* ルーティングと同等の高い汎用性を実現している．

確率モデルシミュレーションの結果，L-turn ルーティングは，全体的に最も高いスループットを実現し，BFS Up*/Down* ルーティングに対して，最大で約 80%のスループット向上を実現することが確認された．一方，R-turn ルーティングは，全体的に，最も低いスループットを示す結果に終わった．禁止ターンの分散に関する静的指標の評価から，L-turn ルーティングと R-turn ルーティングは，ほぼ同等の禁止ターン分散を実現することが確認されたが，選択された禁止ターン分布の違いにより，L-turn ルーティングでは，ツリーの葉スイッチ方向にトラフィックが分散されやすくなるのに対し，R-turn ルーティングではホットスポットが発生しやすいルートスイッチ方向にトラフィックが集中してしまうことがわかった．これより，スループット向上のためには，より均等な禁止ターンの分散と葉スイッチ方向へのトラフィック分散の両立が重要であることがわかった．この条

件を満たす L-turn **ルーティングは，適応型ルーティングとして適用しただけでなく，固定型ルーティングとして適用した場合も最も高い性能を示し，**Up*/Down* **ルーティングに代わる優れたルーティングアルゴリズムであることが確認された．**

謝辞

本研究の機会を与えてくださり，絶えず御指導頂いた慶應義塾大学理工学部 天野 英晴教授に深く感謝致します．

また，本研究をまとめるにあたり，本論文の草稿を丁寧に査読していただき，貴重な御助言を頂いた慶應義塾大学理工学部 寺岡 文男教授，山本 喜一助教授，西 宏章専任講師に深く感謝致します．

本研究を共に行った国立情報学研究所 鯉渕 道紘助手には，数々のご助言をいただき大変お世話になりました．深く感謝致します．

在学中絶えず御指導いただき，精神的な面においても大きく支えていただいた北野共生システムプロジェクト (ERATO-SORST) 舟橋 啓博士には大変お世話になりました．深く感謝致します．

また，普段より御助言，御協力頂いた慶應義塾大学理工学部情報工学科天野研究室の皆様，北野共生システムプロジェクト (ERATO-SORST) の皆様に心より感謝致します．

最後に，私の長い研究生活を支えてくれた両親，家族に深く感謝致します．

2007 年 3 月

論文目録

【本研究に関する論文】

1. 公刊論文

1. (**掲載決定済**:)Akiya Jouraku and Michihiro Koibuchi and Hideharu Amano: "An Effective Design of Deadlock-Free Routing Algorithms Based on 2-D Turn Model for Irregular Networks", IEEE Transaction on Parallel and Distributed Systems, Mar. 2007

2. **上樂 明也**, **鯉渕 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ハイパフォーマンスコンピューティングシステム** Vol.44 No.SIG11 (ACS 3), pp.157-168, Aug. 2003

3. **鯉渕 道紘**, **舟橋 啓**,**上樂 明也**, **天野 英晴**: "L-turn routing: Irregular Network **における** Adaptive Routing", **情報処理学会論文誌ハイパフォーマンスコンピューティングシステム** Vol.42 No.SIG9 (HPS 3), pp.119-134, 2001

2. 国際会議，査読付きシンポジウム

4. **上樂 明也**, **鯉渕 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**", **先進的計算基盤システムシンポジウム**, SACSIS 2003 **論文集**, pp.37-44, May. 2003

5. Akiya Jouraku and Michihiro Koibuchi and Akira Funahashi and Hideharu Amano: "Routing Algorithms based on 2D Turn Model for Irregular Networks", the Sixth International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN'02), pp.289-294, May. 2002

6. **舟橋 啓**, **鯉渕 道紘**,**上樂 明也**: "Irregular Network **における** Adaptive Routing **の提案**", **並列処理シンポジウム** JSPP'2001 **論文集**, pp.247-254, Jun. 2001

7. Michihiro Koibuchi and Akira Funahashi and Akiya Jouraku and Hideharu Amano: "L-turn routing: An Adaptive Routing in Irregular Networks", the 2001 International Conference on Parallel Processing (ICPP'01), pp.384-393, Sep. 2001

3. 研究会

8. **上樂 明也**, **鯉渕 道紘**, **舟橋 啓**, **天野 英晴**: “L-turn routing：Irregular Network **における** Adaptive Routing”, **電子情報通信学会技術研究報告** CPSY2001-12, pp.89-96, Apr. 2001

【その他の論文】

1. 公刊論文

9. Michihiro Koibuchi and Kenichiro Anjo and Yutaka Yamada and Akiya Jouraku and Hideharu Amano: “A Simple Data Transfer Technique Using Local Address for Networks-on-Chips”, IEEE Transaction on Parallel and Distributed Systems, Volume 17, Number 12, pp.1425-1437, Dec. 2006

10. **山田 裕**, **天野 英晴**, **鯉渕 道紘**, **上樂 明也**, **安生 健一郎**: “**リコンフィギャラブルプロセッサアレー向けチップ内接続網**: Fat H-Tree”, **電子情報通信学会論文誌** D1, VOL.J89-D, No.9, pp.1923-1934, Sep. 2006

11. Michihiro Koibuchi and Akiya Jouraku and Hideharu Amano: “Path selection algorithm: the strategy for designing deterministic routing from alternative paths”, PARALLEL COMPUTING, Volume 31, Issue 1, pp.117-130, Jan. 2005

12. Hideharu Amano and Akiya Jouraku and Kenichiro Anjo: “A Dynamically Adaptive Hardware on Dynamically Reconfigurable Processor”, IEICE TRANSACTIONS on Communications, Vol.E86-B, No.12, pp.3385-3391. Dec. 2003

13. **鯉渕 道紘**,**上樂 明也**, **天野 英晴**: “**イレギュラーネットワークにおける仮想チャネルを用いた固定ルーティング**”, **情報処理学会論文誌ハイパフォーマンスコンピューティングシステム** Vol.43, No.SIG 6 (HPS 5), pp.112-121, 2002.

14. Yulu Yang and Akira Funahashi and Akiya Jouraku and Hiroaki Nishi and Hideharu Amano and Toshinori Sueyoshi: “Recursive Diagonal Torus: An Interconnection Network for Massively Parallel Computers”, IEEE Transaction on Parallel and Distributed Systems, Volume 12, Number 7, pp.701-715, Jul. 2001

15. Akira Funahashi and Michihiro Koibuchi and Akiya Jouraku and Hideharu Amano: “The Impact of Output Selection Function on Adaptive Routing”, ISCA Information: An International Journal, Vol 4, No.4, pp.541-550, 2001

2. 国際会議，査読付きシンポジウム

16. Tomohiro Otsuka and Michihiro Koibuchi and Akiya Jouraku and Hideharu Amano: “VLAN-based Minimal Paths in PC Cluster with Ethernet on Mesh and Torus”, the International Conference on Parallel Processing (ICPP’05), pp.567-576, Jun. 2005

17. Kenichiro Anjo and Yutaka Yamada and Michihiro Koibuchi and Akiya Jouraku and Hideharu Amano: “BLACK-BUS: A New Data-Transfer Technique using Local Address on Networks-on-Chips”, 18th International Parallel and Distributed Processing Symposium (IPDPS’04), pp.10-17, Apr. 2004

18. Michihiro Koibuchi and Akiya Jouraku and Konosuke Watanabe and Hideharu Amano: “Descending Layers Routing: A Deadlock-Free Deterministic Routing using Virtual Channels in System Area Networks with Irregular Topologies”,
Proceedings of the International Conference on Parallel Processing (ICPP’03), pp.527-536, Oct. 2003

19. Akira Funahashi and Akiya Jouraku and Hideharu Amano: “Adaptive routing on the Recursive Diagonal Torus.”, ISCA 12th International Conference on Parallel and Distributed Computing Systems (PDCS’99), pp.171-177, Aug. 1999

3. 研究会

20. **上樂 明也**, **舟橋 啓**, **西村 克信**,, **天野 英晴**: “**相互結合網** RDT **における** adaptive routing”,
電子情報通信学会技術研究報告 CPSY97-110, pp.66-74, Jan. 1998

21. **上樂 明也**, **舟橋 啓**, **鯉渕 道紘**, **若林 正樹**, **天野 英晴**: “**命令レベルシュミレーションによる** adaptive routing **の評価**”, **情報処理学会技術研究報告** 2000-ARC-137, 2000-HPC-80, pp.47-52, Mar. 2000

参考文献

[AMAH01] A.Funahashi, M.Koibuchi, A.Jouraku, and H.Amano. The Impact of Output Selection Function on Adaptive Routing. In *Proceedings of International Conference on Computers And Their Applications*, pp. 241–246, March 2001.

[AMAH02] A.Jouraku, M.Koibuchi, A.Funahashi, and H.Amano. Routing Algorithms Based on 2D Turn Model for Irregular Networks. In *Proceedings of the International Symposium on Parallel Architectures, Algorithms, and Networks*, pp. 289–294, June 2002.

[BP89] S. Badr and P. Podar. An Optimal Shortest-Path Routing Policy for Network Computers with Regular Mesh-Connected Topologies. *IEEE Transactions on Computers*, Vol. 38, No. 10, pp. 1362–1371, October 1989.

[C.E85] C.E.Leiserson. "Fat-trees: Universal networks for hardware-efficient supercomputing". *IEEE Transactions on Computers*, Vol. C-34, No. 10, pp. 892–901, October 1985.

[DA93] W. J. Dally and H. Aoki. Deadlock-Free Adaptive Routing in Multicomputer Networks Using Virtual Channels. *IEEE Transaction on Parallel and Distributed Systems*, Vol. 4, No. 4, pp. 466–475, 1993.

[Dal92] W. J. Dally. Virtual-channel flow control. *IEEE Transaction on Parallel and Distributed Systems*, Vol. 3, No. 2, pp. 194–205, 1992.

[Dea92] D.Lenoski and et al. The Stanford DASH multiprocessor. *IEEE Transactions on Computers*, Vol. 25, No. 3, pp. 63–79, 1992.

[D.H] D.H. Brown Associates, Inc. Cray XT3 MPP Delivers Scalable Performance. *available from the Cray Inc., http://www.cray.com/producsts/xt3/.*

[DS87] W. J. Dally and C. L. Seitz. Deadlock-Free Message Routing in Multiprocessor Interconnection Networks. *IEEE Transactions on Computers*, Vol. 36, No. 5, pp. 547–553, May 1987.

[Dua93] J. Duato. A New Theory of Deadlock-Free Adaptive Routing in Wormhole Networks. *IEEE Transaction on Parallel and Distributed Systems*, Vol. 4, No. 12, pp. 1320–1331, 1993.

[Dua94] J. Duato. A Necessary And Sufficient Condition For Deadlock-Free Adaptive Routing In Wormhole Networks. *Proceedings of the International Conference on Parallel Processing*, Vol. 1, pp. 142–149, 1994.

[ea02] NR Adiga et al. An Overview of the Blue Gene/L Supercomputer, NR. In *Proceedings of IEEE/ACM Conference on Supercomputing*, pp. 1–22, November 2002.

[E.W59] E.W.Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, Vol. 1, pp. 269–271, October 1959.

[FFA+02] F.Petrini, W.C Feng, A.Hoisie, S.Coll, and E.Frachtenberg. The Quadrics network: high-performance clustering technology. *IEEE Micro*, Vol. 22, No. 1, pp. 46–57, 2002.

[FJ00] F.Silla and J.Duato. On the Use of Virtual Channels in Networks of Workstations with Irregular Topology. *IEEE Transactions on parallel and distributed systems*, Vol. 11, No. 8, pp. 813–828, 2000.

[GN92] C. J. Glass and L. M. Ni. The Turn Model for Adaptive Routing. *Proceedings of International Symposium on Computer Architecture*, pp. 278–287, 1992.

[Int91] Intel. *Paragon XP/S Product Overview*. Beaverton, OR, Supercomputer Systems Division, 1991.

[I.T04] I.T.Association. Infiniband architecture. specification volume 2 release 1.2. *available from the InfiniBand Trade Association, http://www.infinibandta.org/*, October 2004.

[JA00] J.C.Sancho and A.Robles. Improving the Up*/Down* Routing Scheme for Networks of Workstations. In *Proceedings of the European Conference on Parallel Computing*, pp. 882–889, August 2000.

[JAJ00] J.C.Sancho, A.Robles, and J.Duato. A New Methodology to Compute Deadlock-Free Routing Tables for Irregular Networks. In *Proceedings of Communication and Architectural Support for Network-Based Parallel Computing*, pp. 45–60, January 2000.

[JAJ01] J.C.Sancho, A.Robles, and J.Duato. Effective Strategy to Compute Forwarding Tables for InfiniBand Networks. In *Proceedings of the International Conference on Paralel Processing*, pp. 48–57, January 2001.

[JMPJ02] J.Flich, M.P.Malumbres, P.Lopez, and J.Duato. Removing the latency overhead of the ITB mechanism in COWs with source routing. In *Proceedings of Euromicro Workshop on Parallel, Distributed and Network-based Processing*, pp. 463–470, 2002.

[JPJ+02] J.Flich, P.Lopez, J.C.Sancho, A.Robles, and J.Duato. Improving InfiniBand Routing through Multiple Virtual Networks. In *Proceedings of International Symposium on High Performance Computing*, pp. 49–63, May 2002.

[JPMJ02] J.Flich, P.Lopez, M.P.Malumbres, and J.Duato. Boosting the Performance of Myrinet Networks. *IEEE Transactions on Parallel and Distributed Systems*, Vol. 13, No. 7, pp. 693–709, July 2002.

[JSL02] J.Duato, S.Yalamanchili, and L.Ni. *Interconnection Networks: an engineering approach*. Morgan Kaufmann, 2002.

[KK79] P. Kermani and L. Kleinrock. Virtual cut-through: A new computer communication switching techniques. *Computer Networks*, Vol. 3, No. 4, pp. 267–286, 1979.

[KT95a] K.V.Anjan and T.M.Pinkston. An efficient fully adaptive deadlock recovery scheme: DISHA. In *Proceedings of International Symposium on Computer Architecture*, pp. 201–210, June 1995.

[KT95b] K.V.Anjan and T.M.Pinkston. DISHA: A deadlock recovery scheme for fully adaptive routing. In *Proceedings of International Parallel Processing Symposium*, pp. 537–543, April 1995.

[KTJ96] K.V.Anjan, T.M.Pinkston, and J.Duato. Generalized theory for deadlock-free adaptive routing and its application to Disha Concurrent. In *Proceedings of International Parallel Processing Symposium*, pp. 815–821, April 1996.

[LH91] D. H. Linder and J. C. Harden. An adaptive and fault tolerant wormhole routing strategy for k-ary n-cubes. *IEEE Transaction on Computer*, Vol. 40, No. 1, pp. 2–12, 1991.

[LVT96] L.Cherkasova, V.Kotov, and T.Rokicki. Fibre channel fabrics: evaluation and design. In *Proceedings of the 29th Hawaii International Conference on System Science*, January 1996.

[MAAH01] M.Koibuchi, A.Funahashi, A.Jouraku, and H.Amano. L-turn routing: An adaptive routing in irregular networks. In *Proceedings of the International Conference on Parallel Processing*, pp. 374–383, September 2001.

[Mae91] M.D.Schroeder and al et. Autonet: a high-speed, self-configuring local area network using point-to-point links. *IEEE Journal on Selected Areas in Communications*, Vol. 9, pp. 1318–1335, 1991.

[MAH03] M.Koibuchi, A.Jouraku, and H.Amano. Descending Layers Routing: A Deadlock-Free Deterministic Routing using Virtual Channels in System Area

Networks with Irregular Topologies. In *Proceedings of the International Conference on Parallel Processing*, pp. 527–536, October 2003.

[MDW93] M.Noakes, D.A.Wallach, and W.J.Dally. The j-machine multicomputer: An architectural evaluation. In *Proceedings of International Symposium on Computer Architecture*, pp. 224–235, May 1993.

[MJ80] M.P.Merlin and J.P.Schweitzer. Deadlock Avoidance in Store-and-Forward Networks. *IEEE Transactions on Computers*, Vol. COM-28, No. 3, pp. 345–354, 1980.

[MJJ+05] M.Koibuchi, J.C.Martinez, J.Flich, A.Robles, P.Lopez, and J.Duato. Enforcing In-Order Packet Delivery in System Area Networks with Adaptive Routing. *Journal of Parallel and Distributed Computing (JPDC)*, Vol. 65, pp. 1223–1236, October 2005.

[Myra] Myricom, Inc. http://www.myri.com/.

[Myrb] Myricom, Inc. http://www.myri.com/vlsi/.

[N.J95] N.J.Boden and et al. Myrinet: A Gigabit-per-Second Local Area Network. *IEEE Micro*, Vol. 15, No. 1, pp. 29–35, 1995.

[NKN+01] S. Nishimura, T. Kudoh, H. Nishi, J. Yamamoto, K. Harasawa, N. Matsudaira, S. Akutsu, K. Tasho, and H. Amano. RHiNET-3/SW: an 80-Gbit/s high-speed network switch for distributed parallel computing. In *Hot Interconnect*, pp. 119–123, 2001.

[Oed93] W. Oed. The Cray Research Massively Parallel Processing System: Cray T3D. *Cray Research*, 1993.

[PFH01] F. Petrini, W.C. Feng, and A. Hoisie. The Quadrics network (QsNet): high-performance clustering technology. In *Proceedings of Hot Interconnects*, pp. 125–130, August 2001.

[PJJ01] P.Lopez, J.Flich, and J.Duato. Deadlock-free Routing in $InfiniBand^{TM}$ through Destination Renaming. In *Proceedings of the International Conference on Parallel Processing*, pp. 427–434, September 2001.

[QNR99] W. Qiao, L. M. Ni, and T. Rokicki. Adaptive-Trail Routing and Performance Evaluation in Irregular Networks Using Cut-Through Switches. *IEEE Trans. on Parallel and Distributed Systems*, Vol. 10, No. 11, pp. 1138–1158, November 1999.

[RS91] T.L. Rodeheffer and M.D. Schroeder. Automatic reconfiguration in Autonet. *Technical Report SRC research report 77,DEC*, September 1991.

[SB97] L. Schwiebert and R. Bell. The Impact of Output Selection Function Choice on the Performance of Adaptive Wormhole Routing. In *Proceedings of International Conference on Parallel and Distributed Computing Systems*, pp. 539–544, October 1997.

[SD00] F. Silla and J. Duato. High-Performance Routing in Networks of Workstations with Irregular Topology. *IEEE Transactions on parallel and distributed systems*, Vol. 11, No. 7, pp. 699–719, 2000.

[SLT02] T. Skeie, O. Lysne, and I. Theiss. Layered Shortest Path (LASH) Routing in Irregular System Area Networks. In *Proceedings of International Parallel and Distributed Processing Symposium*, pp. 162–169, April 2002.

[ST96] S. L. Scott and G. T.Horson. The Cray T3E network: adaptive routing in a high performance 3D torus. In *Proceedings of Hot Interconnects IV*, pp. 147–156, August 1996.

[ST97] S.Warnakulasuriya and T.M.Pinkston. Characterization of deadlocks in interconnection networks. In *Proceedings of IEEE Symposium on Parallel and Distributed Processing*, pp. 80–86, April 1997.

[ST99] S.Warnakulasuriya and T.M.Pinkston. characterization of deadlocks in irregular networks. In *Proceedings of the International Conference on Parallel Processing*, pp. 75–84, October 1999.

[STH+00] S.Nishimura, T.Kudoh, H.Nishi, J.Yamamoto, K.Harasawa, N.Matsudaira, S.Akutsu, K.Tasho, and H.Amano. High-speed network switch RHiNET-2/SW and its implementation with optical interconnections. In *Hot Intercoonect*, pp. 31–38, August 2000.

[TOP] TOP500 Supercomputing Sites. http://www.top500.org/.

[TSJ+99] T.Kudoh, S.Nishimura, J.Yamamoto, H.Nishi, O.Tatebe, and H.Amano. RHiNET: A network for high performance parallel computing using locally distributed computing. In *Proceedings of IWIA*, pp. 69–73, November 1999.

[Wea94] W.J.Dally and et al. The reliable router: A reliable and high-performance communication substrate for parallel computers. In *Proceedings of the Workshop on Parallel Computer Routing and Communications*, pp. 241–255, May 1994.

[Wu96] J. Wu. An Optimal Routing Policy for Mesh-Connected Topologies. *Proceedings of International Conference on Parallel Processing*, Vol. 1, pp. 267–270, 1996.

[Wu99] J. Wu. Maximum-shortest-path (MSP): an optimal routing policy for mesh-connected multicomputers. *IEEE Transaction on Reliability*, Vol. 48, No. 3, pp. 247–255, 1999.

[**鯉渕** 02] **鯉渕 道紘**. **システムエリアネットワークにおけるルーティングに関する研究**. 2002 **年度慶應義塾大学大学院博士論文**, 2002.

[**舟橋** 99] **舟橋 啓**. **相互結合網における適応型ルーティングアルゴリズムに関する研究**. 1999 **年度慶應義塾大学大学院博士論文**, 1999.

[**上樂** 03] **上樂 明也**, **鯉渕 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. SIG 11 (ACS 3), pp. 157–168, August 2003.

[**西 宏** 00] **西 宏章**, **西村 信治**, **多昌 廣治**, **工藤 知宏**, **天野 英晴**. **効率良い並列処理をサポートするローカルエリア向けネットワークスイッチ**. **電子情報通信学会論文誌**, Vol. J83D-I, No. 7, 2000.

[**石畑** 89] **石畑清**. **アルゴリズムとデータ構造**. **岩波書店**, 1989.

[**天野** 96] **天野英晴**. **並列コンピュータ**. **昭晃堂**, 1996.