Vol.2010-HPC-126 No.11
2010/8/3

通信量を考慮したデッドロック回避ルーティング方式

中島 耕太[†1]　成瀬 彰[†1]
住元 真司[†1]　久門 耕一[†1]

本稿では，各ターンの通信量を考慮したデッドロック回避ルーティング方式であるターン追加法を提案する．本手法は，スイッチにおける入力ポートと出力ポートの組であるターンを一部禁止することでターンのループを取り除きデッドロックを回避する手法の一種である．この際，通信量が大きいターンから順に許可していき，ターンによるループが生じる場合に当該ターンを禁止することで，できるだけ通信量が小さいターンを禁止する．これにより負荷分散の良いルーティングを得る．

本手法をランダムネットワークに適用し，評価した．その結果，代表的なデッドロック回避ルーティング手法である Up*/Down*法と比較して，負荷分散の良い経路が得られることを確認した．また，スイッチ数 100 のランダムネットワークにおいて，スループットを平均 2.08 倍改善できることを確認した．

A Deadlock Avoidance Routing Method Based on Network Traffic

Kohta NAKASHIMA,[†1] Akira NARUSE,[†1]
Shinji SUMIMOTO[†1] and Kouichi KUMON [†1]

This paper describes a proposal of turn addition method that is a routing method to avoid deadlock using network traffic information. A turn is defined as a pair of input-output ports in a switch. The turn addition method avoids deadlock by prohibited turns which break turn loops. In order to select prohibited turns from lighter traffic turn, it selects allowed turn from heaviest traffic order.

We apply the turn addition method to random network routing. In the evaluation result, the turn addition method generate better load balance routing than Up*/Down* method, and it can achieve 2.08 times higher throughput in 100 switches random networks than Up*/Down* method.

1. はじめに

近年，多数の計算サーバを高速ネットワークで接続したクラスタシステムが広く用いられている．クラスタシステムでは，並列計算性能を高めるために，並列計算を実行する計算サーバ間を高速に接続することが求められる．このため，大規模でも高い帯域が確保できる InfiniBand[1] による Fat Tree トポロジーが広く採用されている．Fat Tree における標準的なルーティングでは，トポロジーとルーティングの性質により，デッドロックの回避が保証される．

一方，最近では，システムの多様化により，MPI 通信のような並列計算用の通信だけでなく，ファイルシステム接続用通信や管理用通信をクラスタネットワーク経由で行う事例[2] が増加している．また，多様なアプリケーションに対応するため，複数の種類の計算サーバ群を相互に接続する事例[3),4)] も増加している．このような様々な構成に対応するためには，Fat Tree における標準的なルーティングのようにトポロジーに依存するデッドロック回避手法は採用できない．したがって，トポロジーに依存しないデッドロック回避ルーティングが必要である．

さらに，このようなクラスタネットワークにおいては，トポロジーとサーバ間の通信量の関係からネットワーク内の通信量には偏りが生じる場合がある．デッドロック回避のために，ルーティングには一部制限が与えられるため，通信量が大きい箇所にはできるだけ制限を与えないことが好ましい．

このように多様なネットワークに対して柔軟に対応するためには，トポロジーに依存せずにデッドロックを回避しつつ通信量を考慮したルーティングを実現する手法が必要である．そこで，本稿では，ターン通信量を考慮したデッドロック回避ルーティング方式であるターン追加法を提案する．本手法は，スイッチにおける入力ポートと出力ポートの組であるターンを一部禁止することによりデッドロック回避を実現する手法の一種である．まず，事前にサーバ間通信量に基づきターン通信量を算出する．そして，通信量の大きいターンから順に許可していき，ターンによるループが生じる場合に当該ターンを禁止する．これにより，できるだけ通信量が小さいターンを禁止することで負荷分散の良いルーティングを得る．

本手法をランダムネットワークに適用し，評価した．この結果，従来の代表的なトポロ

†1 (株) 富士通研究所
Fujitsu Laboratories Ltd.

Vol.2010-HPC-126 No.11
2010/8/3

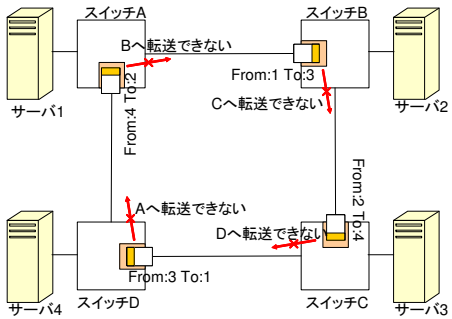

図 1　デッドロックの発生

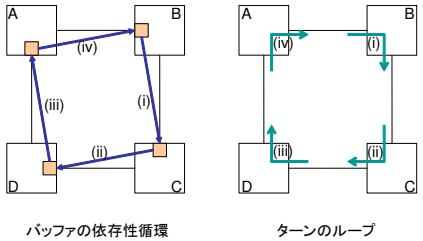

図 2　バッファの依存性循環とターンのループ

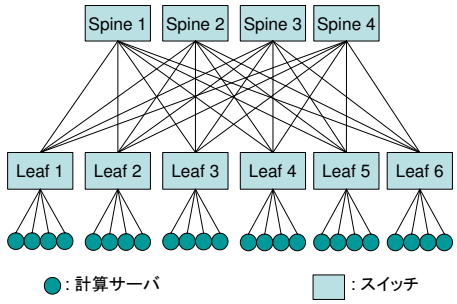

図 3　Fat Tree 構成

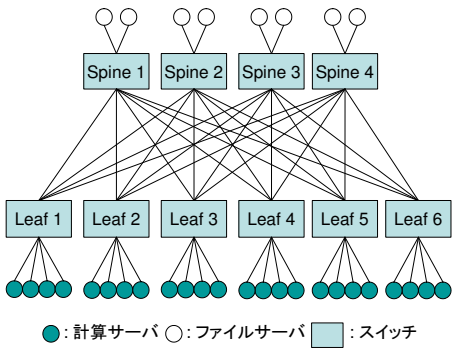

図 4　Spine スイッチにファイルサーバを接続

ジー非依存デッドロック回避ルーティング手法である Up*/Down*法と比較して，負荷分散が良い経路が得られることを確認した．また，スイッチ数 100 のネットワークにおいて，スループットを平均 2.08 倍改善できることを確認した．

2. 課　　題

本章では，クラスタネットワークにおけるルーティングの課題について述べる．

2.1 デッドロック

ループを含むトポロジーを持つネットワークでは，デッドロックが発生する可能性があることが広く知られている．例えば，**図 1** のようなリング上のネットワークにおいて，各サーバがそれぞれ対角線上のスイッチに接続されるサーバへ同時に送信する場合を考える．すなわちサーバ 1 → 3，サーバ 2 → 4，サーバ 3 → 1，サーバ 4 → 2 の送信を同時に実行する．このとき，全てのスイッチが時計回り方向の経路でパケットを転送すると，それぞれ時計回り方向に隣接する次のスイッチの入力バッファにパケットが到着する．仮に，各スイッチの入力バッファの容量が 1 パケット分だとすると，スイッチはそれぞれさらに次のスイッチへ転送しようとするものの，時計回り方向に隣接するスイッチの入力バッファには空きがなく，転送することができない．図 1 の状況では，パケットが破棄されない限り，全てのパケットの転送ができなくなる．これがデッドロックである．

スイッチが次のスイッチへパケットを転送するためには，次のスイッチの入力バッファに空きがある必要がある．すなわち，ある入力バッファから次のスイッチの入力バッファには依存性があり，この依存性が循環するとデッドロックが発生する可能性がある．

入力バッファの依存性とターンの関係について説明する．ターンとは，スイッチにおける入力ポートから出力ポートへの転送である．出力ポートは対向スイッチの入力ポートと 1 対 1 に対応するため，次の入力バッファへの依存性とターンは 1 対 1 に対応する．また，バッファの依存性循環は，ターンがループを形成することに 1 対 1 に対応する．したがって，**図 2** に示すように，バッファの依存性循環とターンのループは同値であり，ターンのループが存在するとき，デッドロックが発生する可能性がある．

デッドロックが発生すると入力バッファのパケットが破棄されない限り，パケットの転送ができなくなる．仮にタイムアウトによりパケットを破棄するネットワークであっても大幅な性能低下を引き起こす可能性が高い．したがって，デッドロックが発生する可能性があるルーティングは回避しなければならない．

2.2 クラスタネットワーク

クラスタシステムにおけるネットワークはまず第一に計算サーバ間を高速に接続することが求められる．このため，**図 3** のような Fat Tree 構成が広く採用されている．

一方で，最近では，クラスタネットワーク上にファイルサーバや管理サーバといった計算サーバ以外のノードを接続する事例が増加している．このようなノードは，計算サーバ間のように高い通信性能が求められるわけではないため，コストに見合う構成を考慮する必要がある．例えば，図 3 のような構成において，各スイッチが全て 8 ポートだとすると，Spine 側のスイッチには空きポートがある．そこで**図 4** のように，この空きポートにファイ

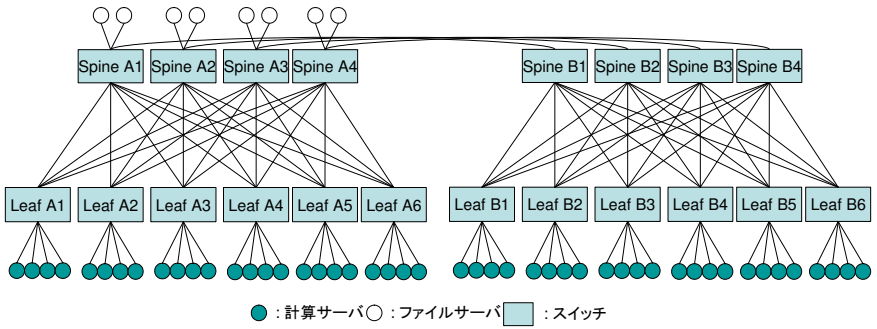

図 5　2 つのクラスタを接続した構成

ルサーバを接続する構成も考えられる．さらに 2 つのクラスタがファイルサーバを共有するために，**図 5** のように 2 つのクラスタネットワークを接続した構成も求められる場合がある．このように，最近では，クラスタネットワークのトポロジーは複雑化している．

また通信量について考慮すると，図 4 のような構成でも，各リンクの通信量は概ね均等である．しかし，ターンを通過する通信量には偏りがある．Leaf → Spine → Leaf と転送される Spine 側スイッチのターンは計算サーバ間通信が経由するため通信量は大きく，Spine → Leaf → Spine と転送される Leaf 側スイッチのターンはファイルサーバ間通信が経由するため通信量は小さい．デッドロック回避のためには一部のターンに制限が与える必要があるため，ターン通信量が高い Spine 側スイッチには制限をなるべく与えないことが好ましい．

2.3　ルーティングの課題

2.1 節と 2.2 節の議論より，クラスタネットワークにおけるルーティングの課題は以下の 3 つである．

- デッドロックの回避
- トポロジー非依存
- ターン通信量の考慮

まず，安定した通信性能を実現するために，デッドロックを回避するルーティングでなければならない．次に多様なネットワークトポロジーに柔軟に対応するためにはトポロジーに依存しないルーティング方式が好ましい．そして，ネットワーク内のターン通信量を考慮したルーティングが好ましい．そこで，本稿ではこの 3 つの課題を解決するルーティング方式について考察する．

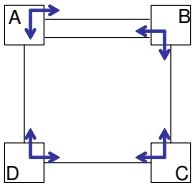

図 6　チャネル追加法

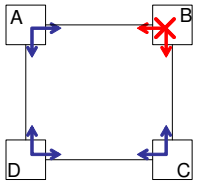

図 7　ターン禁止法

3. デッドロック回避方式

デッドロックを回避するルーティング方式はこれまで広く研究されている．基本的には以下の 2 つの方式によりデッドロックを回避している．

- チャネル追加法
- ターン禁止法

チャネル追加法[5)–8)] では，**図 6** のようにチャネルを追加することにより，ターンのループを防ぐ．図 6 では，物理的にリンクを 1 つ加えているが，実際には，1 つのリンクに入力バッファを複数持たせる仮想チャネルを用いることで仮想的に図 6 の構造を実現する．ルーティングの自由度は高いが，入力バッファを複数持たせる必要があり，多くのハードウェア資源を必要とする．また InfiniBand でチャネル追加法を使用するためには，InfiniBand Verbs(API) を使用するプログラムが直接使用する仮想チャネルを指定する必要がある．このため，MPI 実装のような通信ミドルウェアが仮想チャネル指定に対応する必要があり，実際の適用は難しい．

ターン禁止法[9)–11)] では，**図 7** のようにルーティングに使用するターンの一部を禁止することでターンのループを防ぐ．ルーティングの自由度は低下するが，追加のハードウェア資源は不要である．そこで本稿では，チャネル追加法は使用せず，ターン禁止法によってのみデッドロック回避を行う．

Up*/Down*法[9)] はターン禁止法の一種であり，最も広く使用されている手法の 1 つである．Up*/Down*法では，まず，ネットワーク上の 1 つのスイッチを頂点とし，各スイッチと頂点との距離を算出する．そして各リンクにおいて頂点へ近づく方向を Up 方向と定める．頂点との距離が同一である場合はノード番号が若い方向を Up 方向とする．そして，Down 方向の入力から Up 方向への出力となるターンを禁止することによりターンのループを回避している．この手法は，どのようなトポロジーにも適用でき，比較的単純な操作で禁止

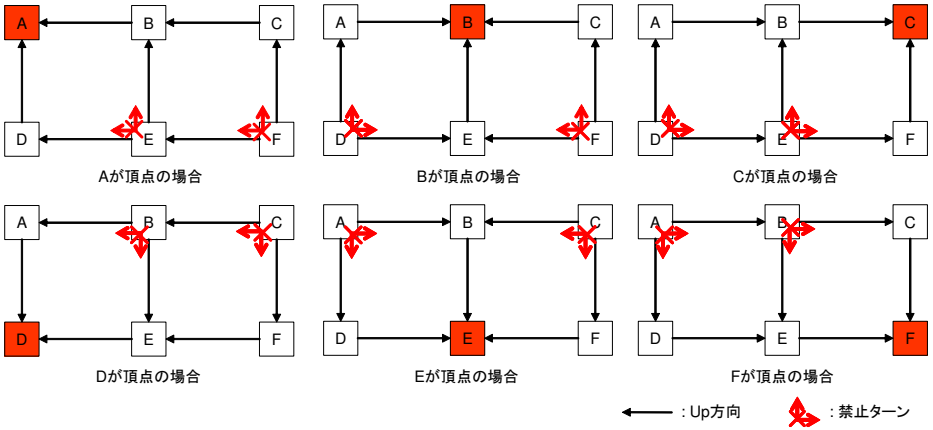

図 8　Up*/Down*法による禁止ターンの算出例

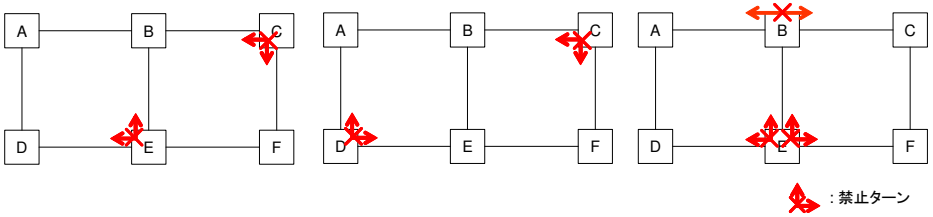

図 9　Up*/Down*法では算出できない禁止ターンの例

ターンを算出できるため幅広く用いられている．しかし，頂点を定めると禁止ターン箇所は一意に決定され，自由度が低い．例えば**図 8** の事例では，スイッチ A〜F のいずれかを頂点とした場合の禁止ターン設置パターンしかなく，**図 9** のようなパターンは算出できない．したがって，通信量の高い箇所に禁止ターンを設置せざるを得ない場合が生じやすい．このような問題が生じにくいような自由度の高い禁止ターンの算出方式が必要である．

4. ターン追加法

本章では，提案するデッドロック回避ルーティング手法であるターン追加法について説明する．

4.1 禁止ターン決定方法

ターン追加法では，ターンの一部を禁止することでデッドロックを回避する．この際，できるだけルーティングに与える影響の少ないターンを禁止したい．そこで，通信量の大きいターンから順番に許可/禁止の判別を行う．既に許可ターンされてるターン群に対して，当該ターンを追加し，ターンのループが生じるかどうかを判別する．ループする場合には当該ターンを禁止ターンとし，ループしない場合は許可ターンとする．このようにして禁止ターンを決定すれば，通信量の大きいターンが許可されやすく，通信量の小さいターンが禁止されやすくなるため，ルーティングに与える影響の少ないターンを禁止しやすいと考えられる．

具体的には以下の手順により，禁止ターンを決定する．

(1)　ターン通信量の算出
(2)　ターンの追加

以降，各手順について説明する．

4.1.1 ターン通信量の算出

ターン通信量は次のようにして算出する．

(i)　サーバ間通信量の定義
(ii)　禁止ターンなし時のルーティング算出
(iii)　ターン通信量の算出

まず，サーバ間の通信量をあらかじめ決定する．例えば，計算サーバ間の通信量は大きい値を設定し，ファイルサーバとの通信量は小さい値を設定する．この通信量とは単位時間あたりの転送データ量であり，スループットと同等の概念である．

次に，ルーティング対象のネットワークにおいて，禁止ターンが存在しないと仮定した場合のルーティングを算出する．そして，そのルーティングにおいて全ての通信ペアについて通過するターンを特定し，通過するターンに対し，設定した通信量を加算する．

ターン通信量の算出例を**図 10** を用いて説明する．図 10 では，禁止ターンなし時のルーティングが A から F への経路は A → C → E → F，B から F への経路は B → C → E → F と算出された場合を示している．したがって，A から F への経路ではターン ACE と CEF を経由し，B から F への経路ではターン BCE と CEF を経由する．A から F への通信量が 1.00，B から F への通信量が 0.50 であるとすると，ターン ACE には 1.00 を，ターン BCE には 0.50 を，ターン CEF には 1.50 を加算する．このようにして，各ターンの通信量を算出する．

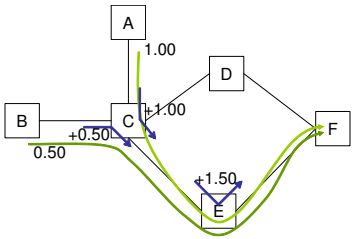

図 10　ターン通信量の算出

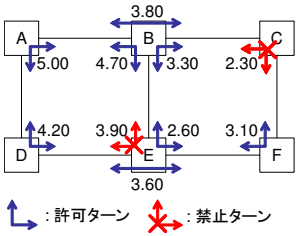

図 11　ターン追加法

4.1.2　ターンの追加

通信量が大きいターンから順に許可/禁止の判別を行う．これまでの許可ターン群に対し，当該ターンの追加によりターンのループが形成される場合は当該ターンを禁止ターンとし，ループが形成されない場合は許可ターンと判別する．この操作は，双方向のターンをペアとして取り扱う．すなわち，あるスイッチにおけるポート 1 からポート 2 へのターンを許可する際には，同時にポート 2 からポート 1 へのターンも許可する．同様に，片方向が禁止となる場合は，逆方向も禁止にする．双方向のターンをペアとして取り扱うため，ターン通信量は，双方向の合計値を用いる．

図 11 のようなネットワークの事例について具体的に説明する．図 11 における双方向の矢印はそれぞれターンを示している．また，ターンに記されている数字は双方向合計のターン通信量である．

最も通信量が大きいターンはターン DAB（通信量: 5.00）である．そこでまずターン DAB を許可ターン群に追加する．ループは生じないので，ターン DAB を許可ターンと判別する．同じ要領で通信量が大きいターンから順にターン ABE（4.70）と EDA（4.20）を追加する．いずれのターンを追加しても，その時点ではループは生じないのでこれらを許可ターンと判別する．

次に通信量が高いターン BED（3.90）を追加すると，既に許可ターンとなっているターン EDA，DAB，ABE とループを形成する．したがって，ターン BED を禁止ターンと判別する．

さらに，ターン ABC（3.80），DEF（3.60），EBC（3.30），CFE（3.10），FEB（2.60）の順にターンを追加する．いずれのターンを追加しても，その時点ではループは生じないのでこれらを許可ターンと判別する．

そして，ターン BCF（2.30）を追加すると，既に許可ターンとなっているターン FEB，EBC，BCF とループを形成する．したがって，ターン BCF を禁止ターンと判別する．

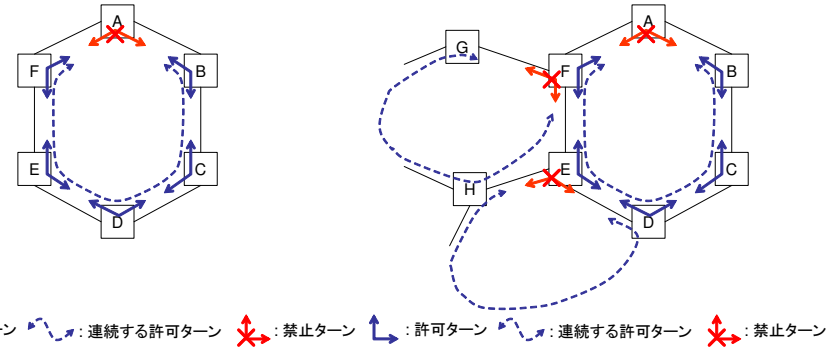

図 12　同一ループ上に存在するスイッチ間

図 13　同一ループ上に存在しないスイッチ間

このよう，通信量が大きい順にターンを追加し，ループが生じるターンを禁止ターンに決定する．このように禁止ターンを決定することで，できるだけ通信量が少ないターンを禁止ターンに指定する．

4.2　到達性保証

ターン追加法では，ターンが禁止ターンに分類される場合には，必ずそのターンを追加することによってループを形成する連続する許可ターンが存在する．双方向のターンをペアで取り扱っているため，禁止ターンによって到達できなくなった経路は必ず連続する許可ターンを利用して逆向きに迂回できる．以降，詳細について説明する．

4.2.1　同一ループ上に存在するスイッチ間

図 12 のような対象ループ状に存在するスイッチ間の到達保証性を考える．ターン ABC，BCD，CDE，DEF，EFA が許可ターンである場合，ターン FAB は禁止ターンとなる．この場合，反時計回りの連続する許可ターンが存在するため，いずれのスイッチ間も到達可能である．

4.2.2　同一ループ上に存在しないスイッチ間

次に**図 13** のような場合について考える．具体的にはループ外のスイッチ G からスイッチ A〜F への到達保証性を考える．ターン GFE が許可ターンであれば，明らかに G から A〜F へは到達可能である．

次にターン GFE が禁止ターンである場合を考える．ターン GFE が禁止ターンである場合は，ターン GFE を追加することでループを形成する連続する許可ターンが存在する必要

表 1　ランダムネットワーク諸元

スイッチ数	10	20	30	40	50	60	70	80	90	100
スイッチのポート数	20									
スイッチ間接続ポート数	10									
スイッチあたりの接続サーバ数	10									
総サーバ数	100	200	300	400	500	600	700	800	900	1,000
総スイッチ間接続リンク数	50	100	150	200	250	300	350	400	450	500

がある．したがって，その経路を経由して G から E へは到達可能である．

仮に，ターン HED が許可ターンであれば，G から A〜F へは到達可能となる．ターン HED が禁止ターンである場合は，先ほどと同様の理由により，ターン HED を追加することでループを形成する連続する許可ターンが存在し，これにより，D へ到達可能である．

この議論から，途中に禁止ターンが存在する経路であっても，その禁止ターンに対応する連続する許可ターンを用いて G から A〜F へは到達可能であることがわかる．このように，ターン追加法では，到達性が保証される．

5. 評　　価

5.1 評 価 項 目

ターン追加法の性能を評価するため，ランダムネットワークにおけるルーティングの性能評価を行った．性能評価では，以下の 2 つを評価指標とした．

(1)　リンク通信量のばらつき
(2)　ネットワークスループット

まず，各リンクの通信量のばらつきを評価する．一部のリンクに通信負荷が集中するようなネットワークは，集中箇所がボトルネックとなるため高い性能が得られない．したがって，各リンクの通信量が均等であり，ばらつきが小さいほどよいルーティングと言える．

次に，ネットワークスループットを評価する．ネットワークに接続する全サーバから同じ通信量を送出する場合，ネットワークに対して送出できる通信量をネットワークスループットとする．このスループットが高いほどよいルーティングと言える．

ターン追加法の有効性を確認するため，ターン追加法により禁止ターンを決定した場合（add）の他に Up*/Down*法により禁止ターンを決定した場合（updown）と，ネットワークに禁止ターンが存在しないと仮定してルーティングを行った場合（none）と比較を行う．

5.2 評価に用いるネットワーク

評価に用いるランダムネットワークの諸元を**表 1** に示す．各スイッチにおけるスイッチ-スイッチ間接続用ポート数を 10 とし，スイッチ数を 10 から 100 まで 10 個刻みで変化させた各場合においてそれぞれ 10 種類のネットワークを生成する．したがって合計 100 種類のネットワークを作成する．スイッチ-スイッチ間はランダムに 2 つのスイッチのポートを選択し接続することで，ランダムネットワークを生成する．

各スイッチにはそれぞれ 10 台サーバを接続する．したがって，スイッチ数 10 の場合は 100 台，スイッチ数 100 の場合は 1,000 台のサーバを接続する．

5.3 評 価 方 法

100 種類のランダムネットワークにおいて，以下の手順にて評価指標を算出する．

(1)　ルーティングの算出
(2)　各リンクの通信量の算出
(3)　評価指標の算出

以降，詳細を説明する．

5.3.1 ルーティングの算出

ネットワークに接続される各サーバが自分以外のすべてのサーバと均等に通信する場合において，最短経路で通信し，ネットワークの各リンクの通信負荷ができるだけ分散するようにルーティングを行う．

「none」のルーテイングは，以下の手順にて決定する．

(i)　最短経路の列挙
(ii)　経路の選択
(iii)　選択した経路への重み付け

まず，対象となるサーバ間の最短経路を列挙する．この最短経路が経路の候補となる．

次に，候補となる各経路について，リンクの通信量を比較し，最も通信量が少ない経路を選択する．通信量の比較は，経路を構成する全リンクのうち最も通信量が大きいリンク同士を比較し，最も低い経路を選択する．もし，最も通信量が大きいリンクの通信量が同一である場合は，2 番目のリンク同士，3 番目のリンク同士と順に比較する．このようにして，できるだけボトルネック箇所を回避するような経路を選択する．

そして，選択した経路を経由するリンクに対してサーバ間の通信量を加算する．本評価では，各サーバが自分以外のすべてのサーバと均等に通信する場合を仮定しているので，加算する通信量は，どのサーバ間も等しい値を用いる．

例えば，**図 14** に示すような通信量である場合，s から d1 の経路を選択する場合には，経路 A → B → D と経路 A → C → D が最短経路であり，経路の候補となる．最も通信量が

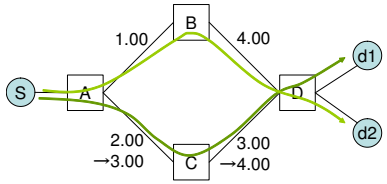

図 14 ボトルネック箇所を回避するルーティング

大きいリンクの通信量は，経路 A → B → D では 4.00，経路 A → C → D では 3.00 であるので，経路 A → C → D を選択する．そして，選択した経路を経由するリンクに通信量を加える．この事例では，A → C と C → D のリンクに 1.00 を加えている．次に，s から d2 への経路を選択する場合は，最も通信量が大きいリンクの通信量は，経路 A → B → D では 4.00，経路 A → C → D では 4.00 である．そこで 2 番目に通信量が大きいリンクの通信量を比較すると，経路 A → B → D では 1.00，経路 A → C → D では 3.00 であるので，経路 A → B → D を選択する．

「add」と「updown」の場合のルーティングの算出は以下の手順で行う．

(1) ターン通信量の算出
(2) 禁止ターンの特定
(3) できるだけ均等なルーティングの算出
　(i) 最短経路の列挙
　(ii) 経路の選択
　(iii) 選択した経路への重み付け

まず,「none」のルーティングに基づいて各ターンの通信量を算出し，これに基づき禁止ターン箇所を算出する．

次に，禁止ターンを特定する.「add」の場合は，4 章の手順により禁止ターンを特定する.「updown」の場合は，まず，ネットワーク上のスイッチを 1 つ選択し，3 章で説明した Up*/Down*法の手順により禁止ターンを特定する．そして，禁止ターンの通信量の合計値を算出する．各スイッチを頂点とした場合について，それぞれこの操作を行い，禁止ターンの通信量の合計値が最も少ないスイッチを頂点とした場合の禁止ターンを採用する．

そして,「none」の場合と同様に (i), (ii), (iii) の手順でネットワークの各リンクの通信負荷ができるだけ分散するようにルーティングを行う．但し，最短経路を列挙する際に，禁止ターンを経由する経路は選択肢から除外する．このようにして，各場合におけるルーティングを算出する．

5.3.2 各リンクの通信量の算出

各場合におけるルーティングに基づき，各リンクの通信量を算出する．通信量の算出は以下のように行う．各サーバは自分以外のすべてのサーバに対し，合計で 1.00 の通信量を送信すると仮定する．サーバ数を N とすると，サーバ間の通信量は $1.00/(N-1)$ である．自分以外のすべてのサーバから受信するので，各サーバは合計で 1.00 を受信する．

すべてのサーバ間において，5.3.1 項で得られたルーティングに基づき経由するリンクを特定し，経由するリンクにサーバ間の通信量を加算する．この操作により，各リンクの通信量を算出する．

5.3.3 評価指標の算出

リンクの通信量のばらつきは，各ネットワークにおけるリンク通信量の標準偏差を用いて算出する．ネットワーク毎にネットワークに送出する通信量が異なるため，リンク通信量の平均値により，標準偏差を正規化した値（変動係数）を算出し，これをリンクの通信量のばらつきの指標とする．

ネットワークスループットは，リンク通信量の最大値を用いて算出する．リンク通信量が最大となるリンクはネットワークのボトルネック箇所である．リンク通信量の最大値が M であるとし，各リンクは最大 1.00 の通信量を流せると仮定すると，ボトルネック箇所の通信量を 1.00 にするためには各サーバが $1.00/M$ を送出すればよい．したがって、ネットワークスループットは，サーバ数を N とすると $N * 1.00/M$ と算出できる．但し，ネットワークに送出する通信量はサーバの台数に比例するため，ネットワークスループットをサーバ数で正規化した値（1 サーバあたりのネットワークスループット）である $1.00/M$ を評価指標とする．

5.4 評 価 結 果

5.4.1 リンク通信量のばらつき

リンク通信量のばらつきを示す評価指標である変動係数を算出した．各スイッチ数における 10 種類のネットワークの平均を**図 15** に示す．スイッチ数が 20 以上の場合において,「updown」よりも「add」の方が変動係数が小さく，リンク通信量のばらつきが小さいことがわかる．また，ネットワーク規模が大きいほどその差は大きくなる．このことから，ランダムネットワークにおいては，Up*/Down*法と比較してターン加算法の方が負荷分散の良いルーティングが得られやすいことがわかる．

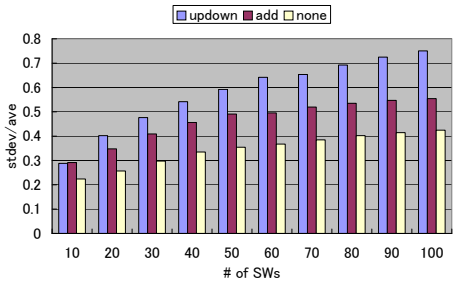

図 15 リンク通信量のばらつき（変動係数）

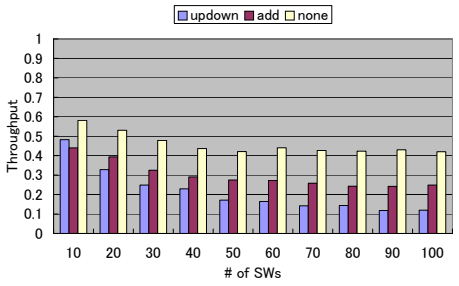

図 16 1 サーバあたりのスループット

5.4.2 ネットワークスループット

スループットを示す評価指標である 1 サーバあたりのネットワークスループットを算出した．各スイッチ数における 10 種類のネットワークの平均を図 15 に示す．

スイッチ数が 20 以上の場合において，「updown」よりも「add」の方が 1 サーバあたりのネットワークスループットが高いことがわかる．また，ネットワーク規模が大きいほどその差は大きくなり，スイッチ数 100 の場合において，スループットは平均 2.08 倍に改善される．このことから，ランダムネットワークにおいては，Up*/Down*法と比較してターン追加法の方がスループットの高いルーティングが得られやすいことがわかる．

6. おわりに

本稿では，ターン通信量を考慮したデッドロック回避ルーティング方式であるターン追加法を提案した．本手法では，通信量の大きいターンから順に許可していき，ターンによるループが生じる場合に当該ターンを禁止することで，できるだけ通信量が小さいターンを禁止する．これにより負荷分散の良いルーティングを得る．

本手法をランダムネットワークに適用し，評価した．その結果，従来の代表的なデッドロック回避ルーティング手法である Up*/Down*法と比較して，負荷分散が良い経路が得られることを確認した．負荷分散の改善によりネットワークのスループットを改善し，スイッチ数 100 のネットワークにおいて，スループットを平均 2.08 倍改善できることを確認した．

ターン追加法は，トポロジーに依存しない禁止ターン算出手法であるため，幅広いネットワークに適用できる．これにより，より柔軟なクラスタネットワーク設計が可能になる．

今後の課題として，実用的なネットワークに対するターン追加法の適用と評価がある．

参 考 文 献

1) InfiniBand Architecture Specification Release 1.2, InfiniBand Trade Association, http://www.infinibandta.org.
2) RIKEN Integrated Cluster of Clusters, http://accc.riken.jp/ricc.html
3) 日本原子力研究開発機構，http://www.jaea.go.jp/
4) 「日本原子力研究開発機構様の新スーパーコンピュータシステムが稼動」，富士通株式会社，プレスリリース，http://pr.fujitsu.com/jp/news/2010/03/1.html (2010).
5) J. Duato : “A New Theory of Deadlock-Free Adaptive Routing in Wormhole Networks,” IEEE Transaction on Parallel and Distributed Systems, Vol.4, No.12. (1993).
6) T. Skeie, et al.: “Layered Shortest Path (LASH) Routing in Irregular System Area Networks,” Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS’02).
7) T. Skeie, et al.: “LASH-TOR: A Generic Transition-Oriented Routing Algorithm,” Proceedings of the Tenth International Parallel and Distributed Processing Symposium (IPDPS’04).
8) O. Lysne, et al.: “Layered Routing in Irregular Networks,” IEEE Transactions on Parallel and Distributed Systems, Vol.17, No.1 (2006).
9) Schroeder, M.D., et al.: “Autonet: A High-Speed, Self-Configuring Local Area Network Using Point-to-Point Links,” IEEE Journal on selected areas in communications. Vol.9, No.8 (1991).
10) D. Starobinski, et al.: “Application of Network Calculs to General Topologies Using Turn-Prohibition,” IEE/ACM Transactions on Networking, Vol.11, No.3 (2003).
11) C. J. Glass and L. L. NI: “The Turn Model for Adaptive Routing,” Journal of the Assocation for Computing Machinery, Vol.41, No.5, pp 874–902 (1994).
12) OpenSM, OpenFabrics Alliance, http://www.openfabrics.org.