博 士 論 文

階層型相互結合網のワームホール・ルーティングに関する研究

指導教官　堀口 進 教授

北陸先端科学技術大学院大学
情報科学研究科情報システム学専攻

三浦 康之

平成 14 年 1 月 22 日

要 旨

近年の 3 次元 IC の発展により，3 次元コンピュータが研究されている．階層型相互結合網 TESH(Tori connected mESHes) および H3D トーラス (Hierarchical 3-D torus) は，数千〜数万プロセッサ規模の 3 次元超並列コンピュータ用階層型相互結合網として提案された．

これまで様々な研究により，階層型相互結合網である TESH や H3D トーラスはネットワーク距離に優れ，VLSI 上実装に適した結合網であることが示されているが，実際のパケット通信を行う際の動的ネットワーク通信特性に関する詳細な検討は行われていない．TESH 上でパケット通信を行うとデッドロックが発生するため，適切なリンク配置を行った上で仮想チャネルを付加することにより論理的にデッドロックを回避する．したがって，そのために必要なリンク配置，仮想チャネル数の導出およびルーティング法の検討が必要となる．また，TESH 上でスループットを向上させ，アプリケーションの実質的な実行時間を向上させるための方法として，ネットワーク上の数経路を選択可能とする適応ルーティングや，データ配置法の工夫がある．そこで本論文では，TESH および H3D トーラス上における適切なメッセージ通信方式を提案し，それによる動的通信性能を評価する．

まず TESH ネットワークの適切なリンク配置法として一列配置を提案し，デッドロックを回避するために必要な仮想チャネル数を導出した．その結果，TESH の必要仮想チャネル数が 4 本または 2 本であることを証明した．シミュレーションにより動的通信性能についての評価を行った結果，3 階層の TESH は他のネットワークに比べて高い通信性能を示した．また，提案された固定ルーティングの手法をもとに階層型相互結合網 TESH におけるルーティング性能の向上のための適応ルーティングのいくつかの手法を提案し，固定ルーティングに比べて高いスループットが実現できることをシミュレーションにより明らかにした．さらに，TESH 上における新たなマッピング法を提案した．シミュレーションによって実験を行った結果，3 階層の TESH で，同サイズのメッシュに比べて実行時間を短縮できることが明らかになった．最後に，階層型相互結合網 H3D トーラスのデッドロック回避法を提案し，必要な仮想チャネル数を導出した．その結果として，必要仮想チャネル数が 2 本であることを証明した．また，シミュレーションによって 4096PE を持つ H3D トーラスの動的性能評価を行った結果，ほぼ同数のリンク数と PE 数を持つ TESH に比べて，若干勝る平均通信時間およびスループットを実現していることを示した．

目 次

図 目 次

表 目 次

第 1 章

緒言

1.1 研究の背景と目的

近年，気象予測，物理シミュレーション，人工知能，画像処理など，多くの分野において大規模な並列処理に対する要求が高まっており，それに伴って数多くの並列計算機が開発されている [1][2][3][4].

また，工業用 3 次元メモリシステムの 3 次元 IC 技術 [5][6] が発展しており，3 次元コンピュータが研究されている．Little らは，5 個のウェーハスタックとして組織された 32×32 のセルラアレイを開発した [7]．このスタックは，アキュムレータとシフタの二種類のウェーハにより構成されている．ダイの大きさはおよそ 1 立方インチで，10 メガヘルツでのスループットは約 600MOPS である．縦方向結線については，Campbell[8] らや，Carson[5] による研究が報告されている．近年では，栗野ら [9] がさらに高度な 3 次元構築技術について議論している．3 次元コンピュータの構築にあたって重大な障害となるのは，縦方向の結線のためのエリアコストである．各結線は，$300\mu m \times 300\mu m$ の面積を必要とする．そのため，これらの縦方向結線を最小化することが，3 次元実装において重要である．階層型相互結合網 TESH(Tori connected mESHes)[10] および H3D トーラス (Hierarchical 3-D torus)[11] は，数千〜数万プロセッサ規模の超並列計算機向けに提案された階層型相互結合網で，下位階層にメッシュ，上位階層にトーラスを用いた階層型構造となっている．

TESH に関しては，これまで様々な研究により，TESH や H3D トーラスがネットワーク距離に優れ，VLSI 上実装に適した結合網であることが示されている [10][12][13]．また，これらの階層型相互結合網を用いたいくつかのアプリケーションについて，TESH に適したマッピングの方法が提案されている．Jain ら [10] は，TESH 上におけるデジタルフィルタの効率的なマッピング法を提案している．さらに，Peng ら [14] は，TESH 上でのウェーブ

レット変換の並列化に関する提案を行っている．しかしながら，メッセージ通信による動的特性に関する検証は行われていない．

相互結合網のメッセージ通信方式として，ハードウェア量や通信スループット等の点で優位なワームホールルーティング [15] がある ．ワームホールルーティングは，パケットをより小さな単位のフリットに分割し，フリットをパイプライン状に転送する手法である．ワームホールルーティングは，ネットワーク中のレイテンシを抑えることができ，ハードウェアコストがそれほど大きくならないことから，並列計算機上でメッセージを転送する際の手法として広く用いられている．

ワームホールルーティングではデッドロックが頻繁に発生するため，仮想チャネルを付加することによりデッドロックを回避する必要がある．この場合，論理的にデッドロックの起こらないネットワークの構築が可能となるが，そのために必要な仮想チャネル数の導出およびルーティングアルゴリズムの検討が必要となる．また，ワームホールルーティングを用いたネットワーク上で，アプリケーションの実質的な実行時間を向上させるための方法として，ネットワーク上の複数経路を選択可能とする適応ルーティングや，データ配置法の工夫等が考えられる．これらの方法を適用することにより，並列計算機システムのさらなる性能向上が可能であると考えられる．

本論文の目的は，階層型ネットワーク上における適切なルーティング方式を提案することである．第一に，TESH および H3D トーラス上でデッドロックの起こらないルーティングを行うための固定ルーティングアルゴリズムの提案および必要な仮想チャネル数の導出を行う．次に，性能向上の方法として適応ルーティングアルゴリズムを提案する．また，信号処理などの分野で広く用いられている高速フーリエ変換について，結合網中におけるメッセージの衝突を低く抑えることのできるデータ配置法を提案する．

本研究では，階層型相互結合網 TESH および H3D トーラスが，VLSI 上への実装性に優れているだけでなく動的通信特性においても優れた結合網であることを明らかにする．

1.2 本論文の構成

本論文の構成は次の通りである．第 2章では，これまでに提案されている超並列計算機の相互結合網を紹介し，その評価方法および従来の相互結合網の問題点について述べる．第 3章では，相互結合網の代表的なルーティング方式であるワームホールルーティングについて紹介し，ワームホールルーティングでデッドロックを回避するために必要な技術としての仮想チャネルおよび，仮想チャネルを用いたデッドロック回避法を紹介する．第 4章では，階層型相互結合網 TESH のルーティング法を提案する．まず，効率の良い上位レベルリン

クの配置法を提案し，固定型ルーティングのための必要仮想チャネル数を導出する．また，TESH の適応ルーティング法を提案する．最後に，シミュレーションによる動的通信性能の評価を行う．第 5章では，階層型相互結合網 H3D トーラスの固定ルーティングのための必要仮想チャネル数を導出する．また，シミュレーションによる動的通信性能の評価を行う．第 6章は結論である．

第 2 章

超並列計算機の相互結合網

2.1 はじめに

並列計算機の相互結合網は，直線，リング，メッシュ，トーラス，ツリー網など様々なものが存在する [16][17]．これらの相互結合網には，それぞれ利点と欠点が存在するため，各結合網の欠点を補完するための手段として，再帰型相互結合網や階層型相互結合網などの，階層構造を持った相互結合網が提案されている．

本章では，これまでに提案されている超並列計算機の相互結合網を紹介し，その評価方法および従来の相互結合網の問題点について述べる．また，大規模 VLSI 実装に適した結合網として，TESH(Tori connected mESHes) および H3D トーラス (Hierarchical 3-Dimensional torus) について説明する．

2.2 相互結合網

基本的な相互結合網を図 2.1に示す．結合網の最も基本的なものとして，直線，リング，メッシュ，トーラス，ツリー網などが存在する．これらの結合網のうち，メッシュやトーラスは格子網と呼ばれる．格子網はさまざまな利点を持つことから広く利用されている結合網であり，k-ary n-cube として総称されている．

図 2.2に k-ary n-cube の構成例を示す．各ノードの番号を n 桁の k進数で表現して，それぞれの桁方向をリングで結んだ結合網である．このようにして構成される結合網は，n 次元のトーラス網となるため，ハイパートーラスとも呼ばれる．

n や kを，それぞれある特定の値に決めることにより，特殊な結合網を構成することも可能である．例えば，$n = 1$ とした k-ary 1-cube は k個のノードを持つリング網となる．ま

た，$k = 2$ とした 2-ary n-cube は，図 2.2(c) に示すような n 次元ハイパーキューブ網と呼ばれる結合網となる．

ハイパーキューブ網は，n 次元の多次元立方体状の結合網である．ノードを二進数で表現し，それぞれの桁が 1bit だけ異なるノード同士を結合することにより構成される．ハイパーキューブは，後述するようにネットワーク距離等の特性が優れているため，かつては最もよく用いられた結合網である．

2.2.1 階層構造を持つ相互結合網

相互結合網には，後に述べるようにそれぞれ利点と欠点が存在する．そこで，各結合網の欠点を補完するための手段として，再帰型相互結合網や階層型相互結合網 [18] といった，階層構造を持った相互結合網が提案されている．

再帰型相互結合網は，ある特定の結合網の一部のノードを利用して同型の結合網を構成した結合網の総称である．RDT[19] や SRT[20] は，再帰型相互結合網と呼ばれる．後述するように，再帰型相互結合網は結合網の直径を短くすることができる．

階層型相互結合網とは，ある特定の結合網のノードに相当する部分に他の結合網を埋め込むことにより階層構造を持たせた結合網の総称である．CCC[21] 等は階層型相互結合網に分類される．後に述べるが，階層型相互結合網は，階層構造を持たない結合網に比べて，直径を短くし，次数や二分割幅を少なめに保つことができるという特徴がある．

代表的な再帰型相互結合網や階層型相互結合網の例として，以下のようなものが挙げられる．

CCC

CCC の構成例を図 2.3に示す．CCC はハイパーキューブ網のノードをリング網で置き換えた構造を持つ階層型相互結合網である [21]．このような結合網では，各ノードがリング網のための二本のリンクの他に，ハイパーキューブ網のための一本のリンクを持つ．

RDT

RDT の構成例を図 2.4に示す．RDT は，2 次元トーラスに付加リンクを付け加えて直径を低く抑えた結合網である [19]．まず，2 次元トーラスのあるノード (x, y) に対して，± 2 離れたノードに対して付加リンクを加え，45 度傾いた目の粗いトーラスを構成する．この上位トーラス上に，同様の方法で次々に上位トーラスを構成してゆくのが完全 RDT である．このような形の RDT は，ノード数が増えるにしたがって次数が大きくなるので，実

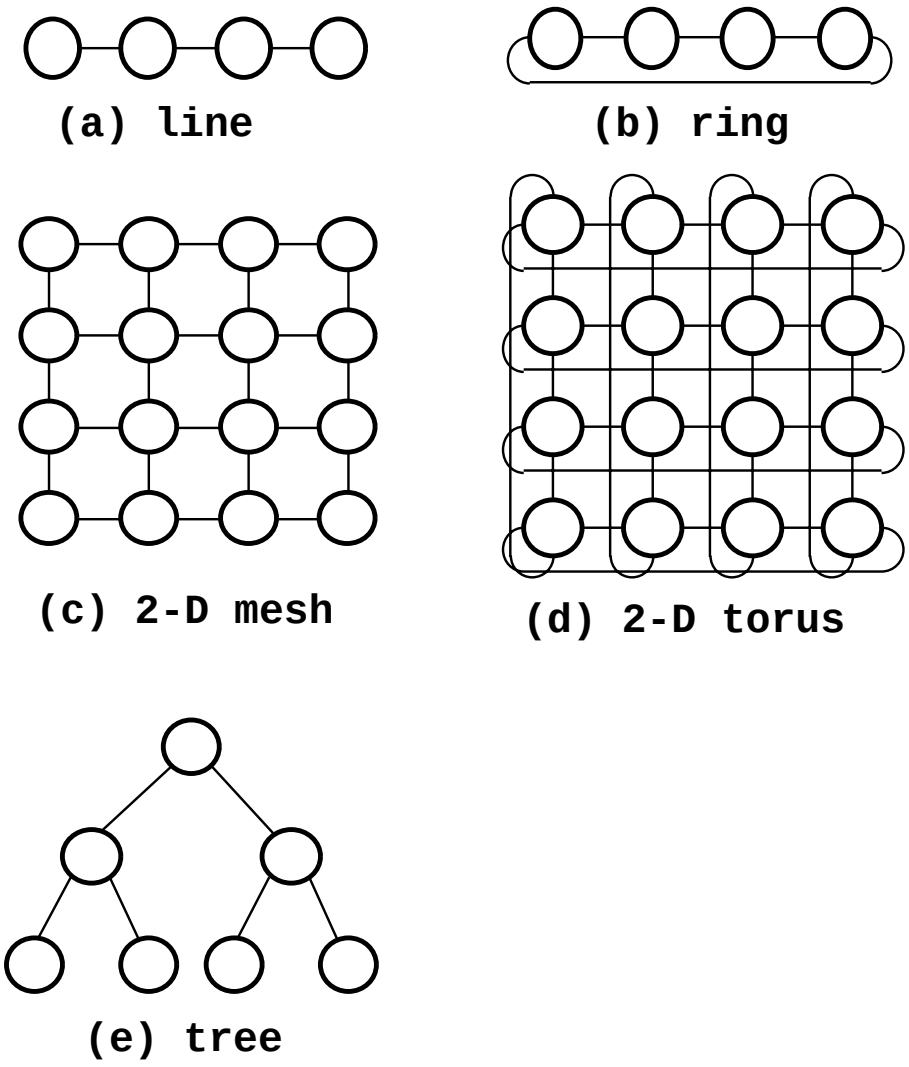


図 2.1: 基本的な相互結合網

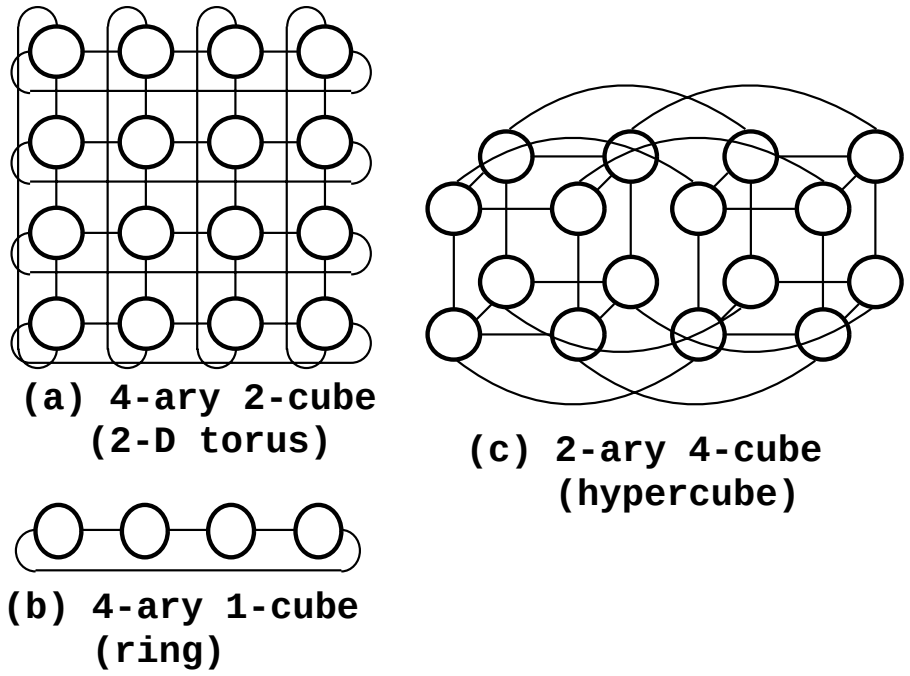


図 2.2: k-ary n-cube

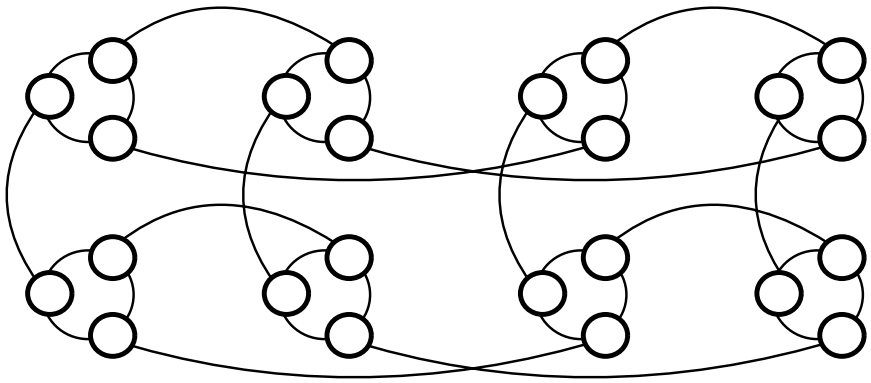


図 2.3: CCC 網の構成

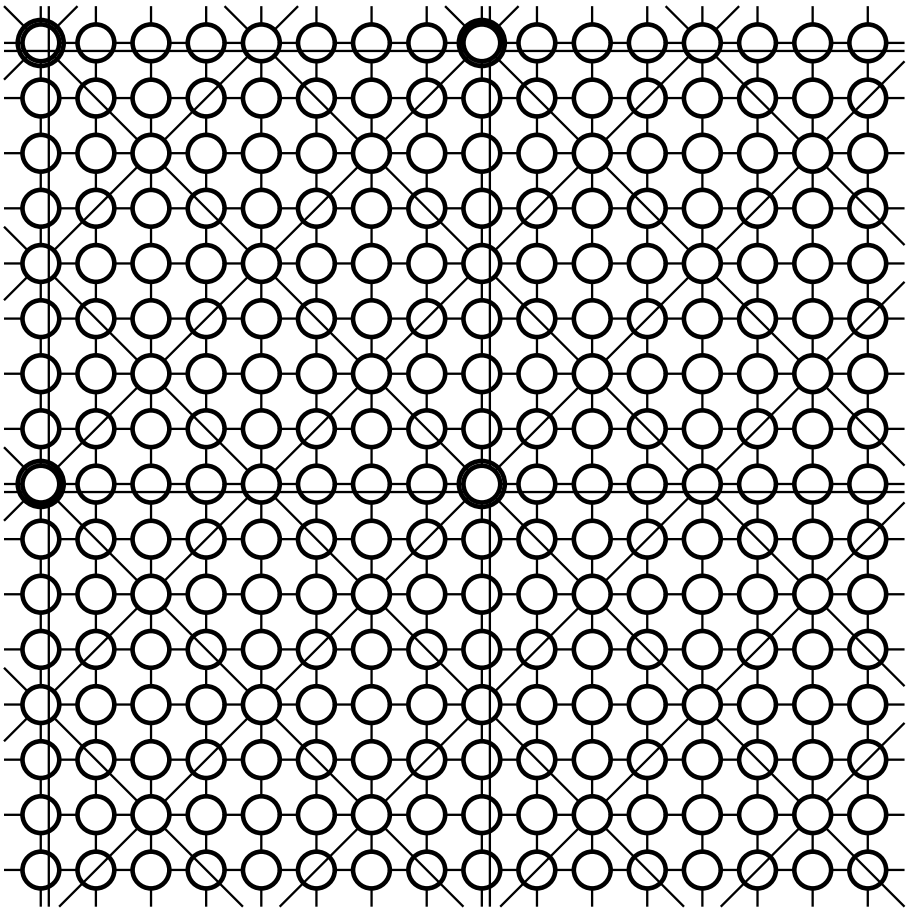


図 2.4: 完全 RDT 網の構成

際には上位トーラスをそれぞれ別のノードが受け持つ.

SRT(Shift Recusive Torus)

SRT は，トーラス結合網に対してバイパスリンクを再帰的に付加することによって構成された結合網である.

一次元 SRT は，図 2.5で示すように，一次元トーラス網（リング網）の，2^n離れた n レベルノード同士を結合した n レベルのリンクにより，再帰的なネットワークを構成している．また，このような一次元 SRT を 2 次元方向に拡張させることにより，2 次元 SRT を構成することができる.

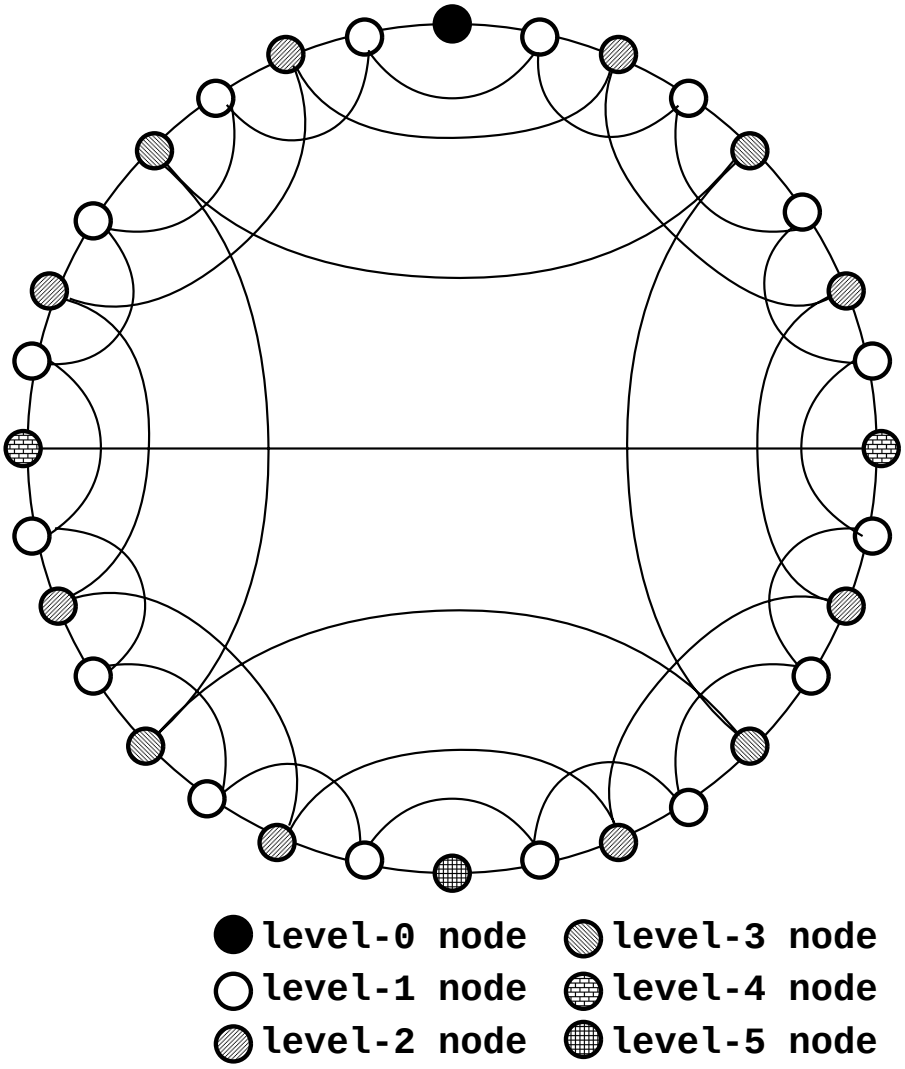


図 2.5: 32 ノードからなる基本 1D-SRT の構成

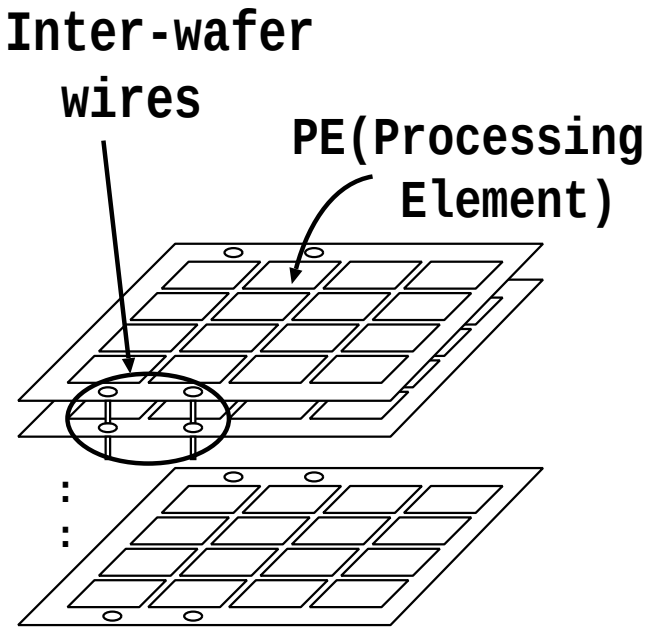


図 2.6: ウェーハスタック構造

2.3 階層型相互結合網 TESH

2.3.1 ウェーハスタック構造

VLSI の大規模化や，大規模 VLSI の再構成技術の発達などにより，ウェーハスタック構造などのように大規模 VLSI を複数用いた超並列計算機システムが実現可能となりつつある．Little ら [7] は，5 個のウェーハスタックとして組織された 32×32 のセルラアレイを開発した．このスタックは，アキュムレータとシフタの二種類のウェーハにより構成されている．ダイの大きさはおよそ 1 立方インチで，10 メガヘルツでのスループットは約 600MOPS である．

ウェーハスタック構造は，図 2.6のように一つのウェーハに複数の PE(Processing Element) を搭載し，複数のウェーハを重ね合わせる実装方法である．ウェーハ間の配線は，$300\mu m \times 300\mu m$ と，ウェーハ内の配線に比べてレイアウト面積が大きくなるため，いかに配線数を少なくするかという点が問題となる．ウェーハ間の配線コストを抑えるには，配線数の少ない階層型相互結合網が有効である．

2.3.2 TESH

階層型相互結合網 TESH(Tori connected mESHes)[10][12][13] は，ウェーハスタック構造に適した結合網として提案された．TESH は，下位階層にメッシュ，上位階層にトーラスを用いる階層型の相互結合網にすることによりウェーハ間の配線数を抑えている．ウェーハ間の配線数を抑え，かつ通信の局所性を利用することで良好なネットワーク性能を持つ．

2.3.3 H3D トーラス

H3D トーラスは，TESH よりもさらに大規模な並列計算機システムのための相互結合網として提案された．H3D トーラスは，下位階層に 3 次元メッシュ，上位階層に 3 次元トーラスを使用することにより，TESH に比べて多数のノードを持つことが可能となり，大規模な並列計算システムの実装が容易に可能となる．

2.4 相互結合網の指標

相互結合網には，さまざまな評価基準があり，それらの評価基準に基づいて優劣が評価される．以下に，いくつかの評価基準を述べる．

直径

あるノードから別のあるノードまでに通過する最短リンク数をノード間の距離と呼ぶ．直径とは，ノード間の距離の最大値である．直径が少ない結合網は，少ない転送で目的のノードへ到達するため，優れた結合網とされる．

次数

各ノードに接続されているリンク数の最大値を次数と呼ぶ．次数が多い結合網は実現のためのコストが大きくなるため，次数が大きくなることは望ましくない．

二分割幅

結合網全体をノードが等しい二つの結合網に分割した場合の，分割線にまたがるリンクの数．結合網によっては分割方法により分割線にまたがるリンク数が異なるので，その場合またがるリンク数が最小となるような分割方法によって評価する．一般に二分割幅が大きいと結合網の転送容量が大きくなるが，他方でリンク数が増えるため，リンク数の節約のために二分割幅が小さいことが望ましいとされることがある．

ルーティングの容易性

送信元 PE と受信先 PE が与えられた時，スイッチの制御が容易である必要がある．そのため，単純なアルゴリズムで経路を選択できることが望まれる．

拡張性

結合網のサイズが大きくなっても容易にスケールアップできる結合網であることが望まれる．完全結合のような，サイズが大きくなるにつれてリンク数や次数が大きくなるような結合網は拡張性に乏しいとされる．

他の結合網の埋め込み可能性

ある結合網の中に，他の結合網の埋め込みが可能なら，その結合網に特化したアルゴリズムを効率的に利用できる．そのため，他の結合網の埋め込みが可能な結合網であることが望ましいとされている．

耐故障性

特定の PE や配線が故障した場合，他の PE や配線でバイパスできるなどということが要求される．

表 2.1に，各結合網のパラメータを示す．3 次元以上のメッシュやトーラス，ハイパーキューブは次数や外部リンク数が大きくなるため，大規模 VLSI 実装に適さない．また，2 次元のメッシュやトーラスでは，直径が大きくなる．一方で，再帰型相互結合網の RDT や階層型相互結合網の TESH，H3D トーラスなどは，次数を小さく保ちつつ，直径が短くなっている．

2.5 まとめ

本章では，これまでに提案されている超並列計算機の相互結合網を紹介し，その評価方法および従来の相互結合網の問題点を述べた．

並列計算機の相互結合網は，大きく分けて直接網と間接網に分けられており，さまざまな間接網について詳細に説明した．基本的な間接網の多くは，ネットワークのノード数が大きくなるにつれて，直径が増えることや，次数が大きくなるといった問題がある．そのため，これらの結合網の欠点を補完するための結合網として，再帰型相互結合網や階層型相互結合網などの相互結合網が提案されている．本章では，これら再帰型相互結合網や階層型相互結合網の構成，各パラメータ等について述べた．最後に，大規模 VLSI 実装に適した結合網として，TESH および H3D トーラスについて説明した．

表 2.1: 各結合網の N ノードにおけるパラメータ

結合網	直径	次数	
リング	$O(N)$	2	
2D メッシュ，トーラス	$O(\sqrt{N})$	4	
3D メッシュ，トーラス	$O(\sqrt[3]{N})$	6	2D メッシュ，トーラスを内包
ハイパーキューブ	$O(\log N)$	$O(\log N)$	メッシュ，トーラス，リングを内包
ツリー	$O(\log N)$	3	中心付近に負荷が集中
RDT	$O(\log N)$	6	メッシュ，トーラスを内包
CCC	$O(\log N)$ 以上	3	
HypernNet	$O(\log N)$	6	直径でハイパーキューブに劣る
TESH	$O(\log N)$	4	ウェーハスタック構造に適する
H3D トーラス	$O(\log N)$	6	
一次元 SRT	$O(\log N)$	4	

これらの結合網を用いて実際に並列計算機システムを構築する場合，適切な方式によってメッセージをやり取りすることになる．第 3 章ではこれらの結合網上でメッセージをやり取りするための方式について説明し，実際にメッセージをやり取りするにあたっての問題点を述べる．

第 3 章

相互結合網のワームホール・ルーティング

3.1 はじめに

パケット転送方式の一つであるワームホールルーティングは，ストアアンドフォワードやバーチャルカットスルーのような他の転送方式に対していくつかの利点を持った方式である．反面，パケット同士の衝突によるブロッキングが多くなるため，結合網のデッドロックが起こりやすくなる．デッドロックに対処するための手法にはさまざまなものが提案されているが，仮想チャネルを付加して論理的にデッドロックを防ぐ方法がワームホールルーティングに対して一般的に用いられている．

本章では，まず 3.3節でパケット転送方式の一つであるワームホールルーティングの手法と，他の転送方式と比較したワームホールルーティングの利点について説明する．次に，3.4節で，デッドロック回避のための有効な方法として，チャネル依存グラフを使う方法を説明し，3.4 節でメッシュ網やリング網におけるチャネル依存グラフを用いたデッドロックフリーの証明を述べる．3.5 節では，デッドロックフリーなルーティングの中でも単一経路のみを選択する固定ルーティングと，複数経路を選択可能な適応ルーティングを説明し，適応ルーティングの利点を述べる．

3.2 第一世代並列計算機と第二世代並列計算機

第一世代の並列計算機では，１ホップルーティングするたびにメインメモリにメッセージを格納する．その際の通信制御はプロセッサが担当する．第二世代以降の並列計算機ではルータという，ルーティング専用のハードウェアが登場し，より複雑な通信制御を行うことが可能となった．物理的な構成の違いにともない，第一世代と第二世代以降の並列計

算機では，パケット通信方式が大きく異なっている．第二世代以降の並列計算機は，ルータの登場により，後述のワームホールルーティングが可能となっている．

3.3 パケット通信方式

多くの並列計算機でメッセージ通信をする時には，通信するメッセージを送信先等のタグを付加したパケットにまとめ，パケットをさらにフリットと呼ばれるものに分割して，クロックに同期して１フリットごとに転送を行なってゆく．フリットとは１クロックの転送で送ることのできるメッセージの最小単位で，8〜64bits 程度の大きさである．ノード間のパケット転送方式は，ストアアンドフォワード，ワームホールルーティング [15] ，バーチャルカットスルー [22] の 3 つに大別される．

3.3.1 ストアアンドフォワード

ストアアンドフォワードとは，図 3.1で示されるように，各ノードにパケット全体を保持できるバッファを持ち， 各ノードがパケット全体を受け取ってから次のノードへ送信する方式である．

3.3.2 ワームホールルーティング

ワームホールルーティングは，図 3.2で示されるように，パケットをより小さな単位のフリットに分割し，フリットをパイプライン状に転送する手法で，並列計算機上でメッセージを転送する際の手法として広く用いられている．ワームホールルーティングには，以下のような特長がある．

- パケット全体を保存するためのバッファが必要ないため，ストアアンドフォワードやバーチャルカットスルーに比べて少ないバッファサイズで実現できる．実際，ワームホールルーティングでパケット全体を保存するためのバッファを持つことは少ない．

- パケット長が大きい場合，転送時間がネットワーク距離に依存しにくくなる．

上記の特徴から，並列計算機のメッセージ送信には従来のストアアンドフォワード方式に代わって頻繁に使用されている．

3.3.3 バーチャルカットスルー

ストアアンドフォワード方式と同様，各ノードがパケット全体を保持できるバッファを持つが，ワームホールルーティングと同様，パケット本体の到着を待たずに先頭が先へ進むことができる．ワームホールルーティングと違い，パケット全体を保持できるバッファを持つため，パケットの先頭がブロックされた場合でもパケット後続が先頭に進路を妨害されずに先に進むことができるという利点がある一方，膨大なバッファを必要とするという欠点がある．この方式は，ワームホールルーティングと同等の流れでパケットの送受信が行われるため，ワームホールルーティングの一種として扱うことがある．

3.3.4 各方式の比較

ストアアンドフォワードおよびワームホールルーティングの転送時間に関して議論する．T_cをクロックのサイクルタイム，Dをネットワーク距離，L をパケット長，Wをバンド幅とすると，ストアアンドフォワードの転送時間 T_{SF}は次式で示される．

$$T_{SF} = T_c \left(D \times \frac{L}{W} \right) \tag{3.1}$$

これに対し，ワームホールルーティングの転送時間 T_{WH}は次式で示される．

$$T_{WH} = T_c \left(D + \frac{L}{W} \right) \tag{3.2}$$

このように，ワームホールルーティングではパケット長が長くなるに従って転送時間がネットワーク上の距離に依存しにくくなる．また，ストアアンドフォワードと比べて短い転送時間でルーティングを行えることが明らかである．したがって，可能であればワームホールルーティングを用いる方がネットワーク中のレイテンシ軽減の上では効果的である．

iPSC,NCUBE 等の第一世代の並列計算機では，ほとんどがストアアンドフォワード方式を用いていた．これは，これらのマシンでは専用のルータを持たず，パケットの転送を CPU が制御していたためである．その後，第二世代以降の並列計算機では専用のルータを持つようになり，パケット後続部の到着を待たずにパケット先頭部を送るという複雑な動作が可能になり，ワームホールルーティングやバーチャルカットスルーが用いられるようになった．

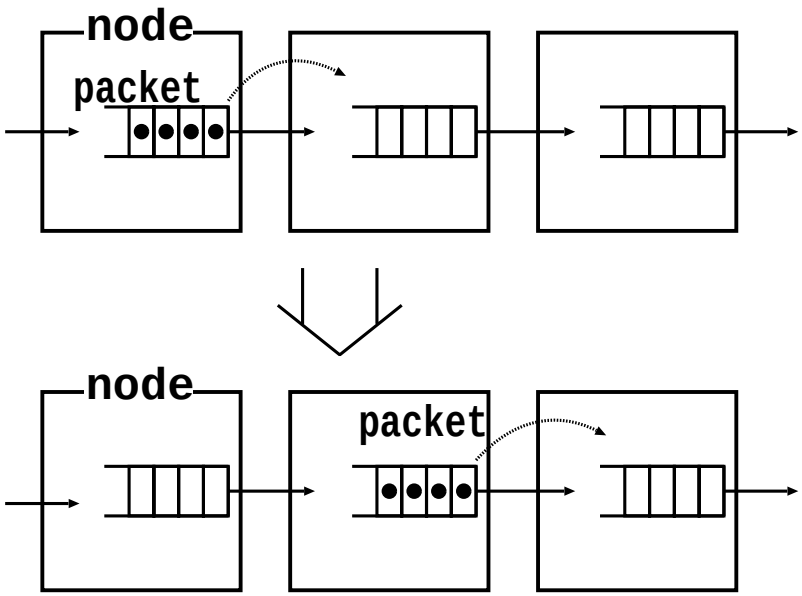


図 3.1: ストアアンドフォワード

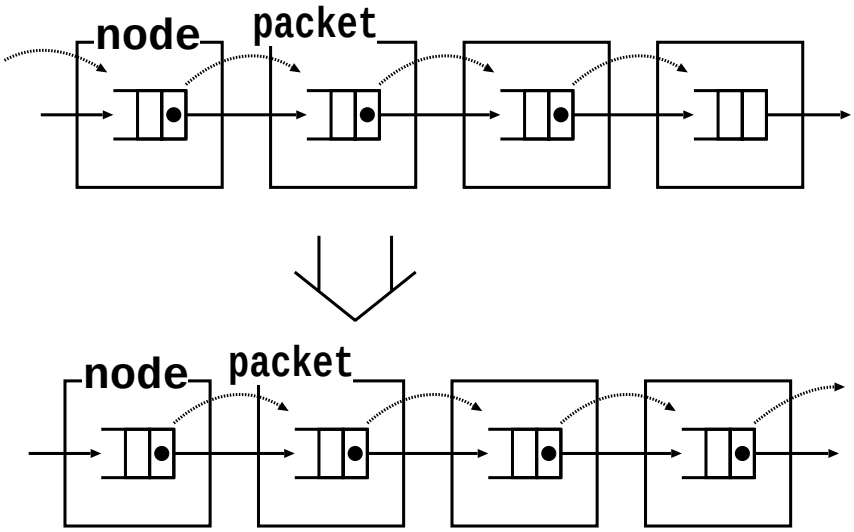
3.4 デッドロック

ワームホールルーティングは複数のノードにわたってパケットがまたがるため，ストアアンドフォワードに比べてパケット同士の衝突回数が多くなる．そのため，デッドロックの発生頻度が高くなるという問題点がある．したがって，何らかの方法でデッドロックに対処することが必要となる．

図 3.3にデッドロックの概念図を示す．図 3.3に示すように，パケット 1 がノード 0 からノード 1 を通ってノード 2 へ向かい，パケット 2 がノード 1 からノード 2 を通ってノード 0 へ向かい，パケット 3 がノード 2 からノード 0 を通ってノード 1 へ向かう場合，三つのパケットが相互に進行を妨害し合うため，全体としてまったく動きが取れなくなる．このような循環依存の状態をデッドロックと呼ぶ．デッドロックは，ワームホールルーティングに限らず，ストアアンドフォワードやバーチャル・カット・スルー等，どのようなルーティング法でも起こりうるが，ワームホールルーティングではパケット同士で進路を妨害し合うことが多いため，特にデッドロックの発生頻度が多くなる．

デッドロック回避の方法として，おおまかに分けて以下の二つの方法がある．

1. 結合網の内部でデッドロックを検出して，デッドロックの起こったパケットを除去し再送する方法



ノード間にまたがってパケットが転送される

図 3.2: ワームホールルーティング

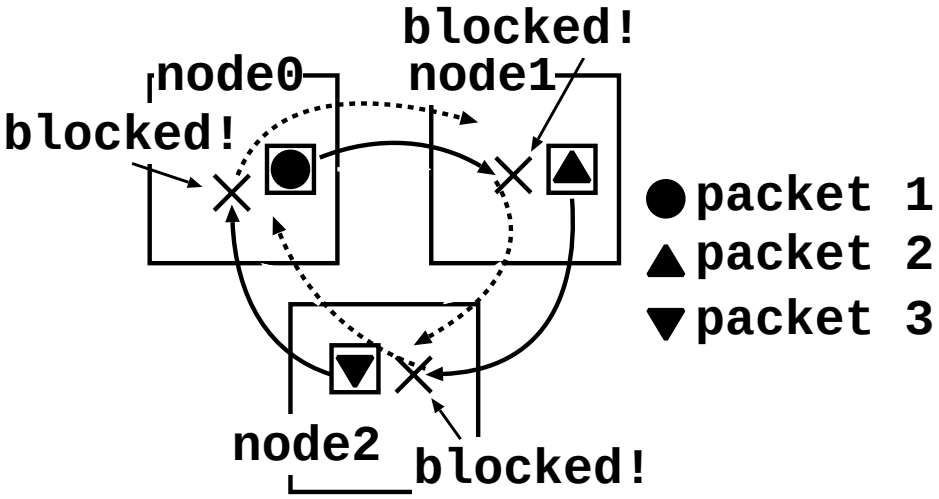


図 3.3: デッドロック

2. ルーティングの方向を制限したり，経路を増やすことにより論理的にデッドロックを防ぐ方法

これらのうち，前者の方法は，デッドロックの発生頻度が高いと性能の低下が大きくなるため，ワームホールルーティングを行うシステムでは主に後者の方法が用いられている．

後者の方法としては，以下のようなアプローチが考えられる．

- ルーティングの経路・方向を制限して，パケットの循環依存を断ち切る．e-cube ルーティング [23] や Turn モデルによるルーティングの制限方法 [24] 等が挙げられる．
- リンクに複数の仮想チャネル [25] を設け，循環依存が起こらないように利用する．構造化バッファ／チャネル法 [26]，double Y-Channel ルーティング [27]，Duato の必要十分条件 [28] 等が挙げられる．

仮想チャネルは，原理的にはワームホールルーティング，ストアアンドフォワード，バーナルカットスルーのいずれにも用いることができるが，ワームホールルーティングに用いるのが最も効果的である．それは次のような理由による．第一に，ワームホールルーティングは，必要となるバッファサイズが小さく，仮想チャネルを付加するために必要なコストが少なくて済むことである．ストアアンドフォワードやバーチャルカットスルーで仮想チャネルを付加するためには，最大パケット長と同じ数のバッファを用意しなければならない．一方，ワームホールルーティングでは，パケット全体を格納するバッファを必要としない．したがって，他の 2 つの方法よりも少ないコスト消費で仮想チャネルを付加することができる．

第二に，ワームホールルーティングは，仮想チャネルを付加することによる性能の向上が大きいことが挙げられる．3.3節で述べたように，ワームホールルーティングでは一つのノードでパケット全体を保持することができないためパケットの衝突が起こりやすい．そのため，仮想チャネルを利用したブロック回避の機会も多くなり，性能向上が期待できるからである．

3.5 チャネル依存グラフによるデッドロックの回避

デッドロック回避のための一般的な方法として，チャネル依存グラフを使う方法がある [28][29]．一般にネットワーク G は，PE N とチャネル C によるグラフ $G(N, C)$ で記述される．チャネル依存グラフは，ネットワーク中の各チャネル間の依存関係をグラフにより表現したものである．チャネル間の依存関係は，ルーティング関数およびルーティング副関数を用いて定義される．

3.5.1 ルーティング関数

プロセッサ間でやりとりされるパケットの転送経路は，ネットワーク中で自由に選ばれるわけではなく，通常は経路選択に制限が加えられる．そこで，各 PE からの直接の送信先 PE となる可能性のある PE を出力する関数としてルーティング関数を定義する．ルーティング関数 $R'(x,y) = \{c_1, c_2, ..., c_m\}$ とは，ノード x にあり，ノード yを目的地とするパケットの次の行き先チャネル $c_1, c_2, ..., c_m$を出力とする関数である．

ルーティング副関数 $R_1(x,y)$ は，ルーティング関数 R'から特定のチャネルを除いた関数で，以下の式により定義される．

$$R_1(x,y) = R'(x,y) \bigcap C_1 \tag{3.3}$$

ここで $C_1 \in C$は，デッドロックフリーを保証するために最低限必要なチャネルの集合である．以降本論文では C_1を，固定チャネルと呼ぶ．つまり，$C - C_1$は，適応ルーティング用の予備のチャネルとなる．本論文では，このようなチャネルを以降余剰チャネルと呼ぶことにする．

3.5.2 チャネル依存グラフ

ルーティング副関数が前節のように決定される時，以下の二つの条件のいずれかを満たすチャネル c_iと c_jは，互いに依存関係があると呼ぶ．ここで，s_iおよび d_iは，チャネル c_iの入力側ノードおよび出力側ノードである．

1. 直接依存関係

$$\begin{cases} c_i \in R_1(s_i, y) \\ c_j \in R_1(d_i, y) \end{cases} forsomey \in N \tag{3.4}$$

2. 間接依存関係

$$\begin{cases} c_i \in R_1(s_i, y) \\ c_1 \in R'(d_i, y) \\ c_{t+1} \in R'(d_t, y) \quad t = 1, \cdots, k-1 \\ c_j \in R_1(d_k, y) \end{cases} forsomey \in N \tag{3.5}$$

直接依存関係は，ルーティングを通して，チャネル c_iにあるパケットが，次のホップで直接 c_jを選択する可能性があることを意味している．間接依存関係は，チャネル c_iにある

パケットが，余剰チャネルを通した数ホップのルーティングの後に c_j へ到達する可能性があることを意味している．

以上を用いて，以下の手順によりチャネル依存グラフが作られる．

1. 固定チャネル C_1 をノードとする．

2. 依存関係がある二つのノード c_i，c_j について，c_i から c_j への有向枝を引く．

以上のようにして作成したチャネル依存グラフに巡回がなければ，ネットワークにデッドロックが起こらない．チャネル依存グラフに巡回がないことを示すためには，それぞれのチャネルに番号を振り，依存関係のあるチャネル同士で必ず昇順または降順になることが示されれば良い．そこで本論文では，ルーティング関数 R_1 で使用されるチャネル（C_1 に含まれるチャネル）に番号を振る方法でデッドロック・フリーを証明する．チャネル依存グラフに巡回がなければ以下のルールに従って，図 3.4のように各チャネルに順序をつけることができる．なお，図 3.4薄く示された部分は結合網そのものを表わし，濃く示された部分がチャネル依存グラフを表わす．

1. チャネル依存グラフのノード（チャネル）のうち，受信先ノードを持たないノード（グラフの終点となっているノード）をシンクチャネルと呼び，最低位の順序とする．

2. 各々のノードについて，そのすべての送信元先ノード（有向枝の元のノード）よりも高い順序をつける．

3. 上記の手順を，すべてのノードについて行う．

そのようにして順序をつけられたチャネルのうち，フリットを含むチャネルの中で最高位の順序のチャネル c に着目する．そのチャネルの有向枝の先にある各チャネルはフリットを含まないため，c に含まれるフリットは，進路を妨害されることはない．したがってネットワークにデッドロックは発生しないことが証明される．

上記のようなチャネル依存グラフを描けばネットワークのデッドロック・フリーを証明することができるが，チャネル依存グラフは非常に枝の多いグラフになることが多いため，固定ルーティングのような比較的単純なルーティングの証明をする際には，有向枝を省略してチャネルに番号を割り振るだけで証明を済ませることが多い．また，固定ルーティングでは，すべてのチャネルを固定チャネルとするため，間接依存関係を省略して考えることが多い．実際，すべてのチャネルが固定チャネルの時には，直接依存関係のみでデッドロック・フリーの十分条件の証明は十分である [28].

3.5.3　メッシュ網におけるデッドロック回避

e-cube ルーティングを用いたメッシュ網のチャネル依存グラフを図 3.5に示す．図 3.5(a) が 2×2 メッシュ，図 3.5(b) が 3×3 メッシュの例である．2×2 メッシュでは明らかにチャネル依存グラフに巡回がないことが分かる．3×3 メッシュは，図の凡例のように，色の薄いチャネルから昇順になるように各チャネルに順番をつけると，有向グラフの矢印に沿って昇順になるような順番となるため，やはりチャネル依存グラフに巡回はないことが分かる．したがって，双方ともデッドロックは起こらない．

3.5.4　仮想チャネルの付加

ルーティングの経路を変えずにデッドロックを防ぎたい場合や，経路を変えただけではデッドロックを防げない場合には，各配線ごとに複数の転送経路を設け，おのおのを別個のチャネルとして扱う方法が有効である．この際，物理的な配線を複数設けるのはハードウェア上の制約上問題があるので，実際には一組の配線を複数の仮想チャネルで共有して使用することになる．

図 3.6に仮想チャネルの概念図を示す．図 3.6は，入力側と出力側のノードと，両ノード間のリンクを示しており，クロスバスイッチやバッファ等を囲む太い線が，それぞれ入力側と出力側のノードを示している．図 3.6に示すように，一本のリンクを入出力側にある複数のバッファが共有して使用している．これらのうち，ハンドシェーク線を共有して入力側と出力側のノードにまたがる二つのバッファの組が，一本の仮想チャネルとして使用される．ハンドシェーク線は，入力側と出力側のバッファ間の調停に使用される．図 3.6では仮想チャネルが複数本設けられており，これらが互いに時分割で一本のリンクを共有して使用する．複数の仮想チャネル間の調停は，マルチプレクサにより行われる．

3.5.5　リング網におけるデッドロック回避

仮想チャネルの付加によりデッドロックを回避するための方法の一例として，リング網におけるデッドロック回避法を紹介する．リング網の構成を図 3.7に示す．図 3.7のように，リング網は循環構造をもっているため，通常は各配線に二本の仮想チャネルを付加してデッドロック回避する．図 3.8に，二本の仮想チャネルを付加したリング網を示す．図 3.8中の薄く示された矢印がリンク，濃い点線で示された二本の矢印がそれぞれ仮想チャネルのチャネル L とチャネル H を示す．二本の仮想チャネルを図 3.8のように配置し，PE3 → PE0 のチャネルを通るときにチャネル L からチャネル H に移行というルールを設けることによってデッドロックを回避する．ここで，PE3 → PE0 の転送は，一般にラウンドトリップと呼

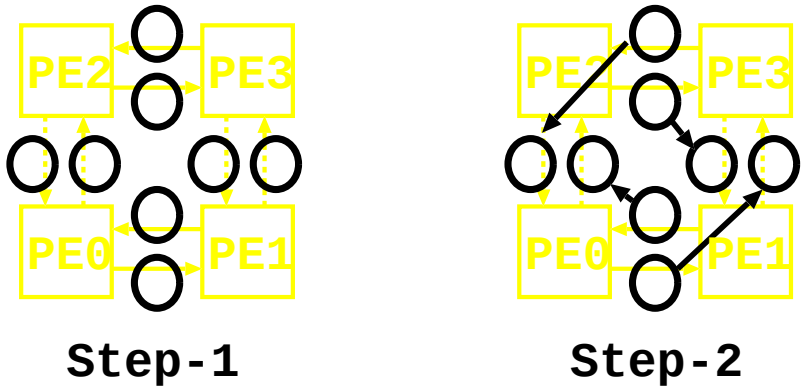


図 3.4: チャネル依存グラフの作成

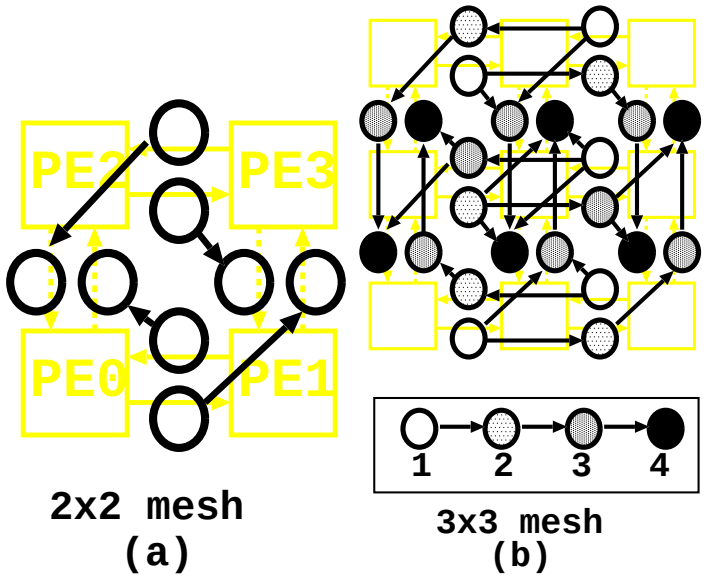


図 3.5: e-cube ルーティングのチャネル依存グラフ

ばれ，ラウンドトリップを行うチャネルをラップアラウンドチャネルと呼ぶ．リング網のチャネル依存グラフを図 3.9に示す．図で薄く示された部分が二本の仮想チャネルを付加したリング網で，濃い線で示されたグラフがチャネル依存グラフを示す．図 3.9のように，各チャネルが循環しないため，デッドロックは起こらない．

3.5.6 階層型相互結合網のデッドロック回避

一般の相互結合網については，チャネル依存グラフが循環を持たないことを証明することによってデッドロックが起こらないことを証明することができる [28]．そのためには，それぞれのチャネルに番号を振り，依存関係のあるチャネル同士で必ず昇順または降順になることが示されれば良い．このことは階層型相互結合網についても同様であるが，階層構造を持たない結合網に比べて構造が複雑なため，通常の方法ではデッドロック・フリーの証明が困難である．多くの階層型相互結合網では，ルーティングも複数の階層に分けて行われる．多くの場合個々の階層は既存のネットワークの形をしているため，個々の階層を分割した上でデッドロックフリーについて考察する方法が効果的である．

一般に，相互結合網 Iは，全ノードの集合 Nと，全チャネルの集合 Cによるグラフ $I = \mathrm{G}(N, C)$ と定義することができる．また，これらの中の複数のチャネル $c_i, c_j \in C$について，c_iの受信元ノードと c_jの受信元ノード，c_iの送信先ノードと c_jの送信先ノードがそれぞれ同一となるチャネルが存在する可能性がある．これらは一つのリンク $l \in \Lambda$中の仮想チャネルとして存在することになる．ただし，Λをリンク全体の集合とする．

以下，必要な定理について述べる．

定理 3.1　*ルーティング全体の流れを，ルーティングの順序に沿った任意の n 個のフェーズに分割できるものと仮定する．フェーズ 1 からフェーズ n の各フェーズでデッドロックがおこらなければ，ネットワーク全体でもデッドロックがおこらない．*

証明　フェーズ i における各チャネルのチャネルにつけられる k桁の番号を

$$ch_i = (ch_{i1}, ch_{i2}, \cdots, ch_{ik})$$

と置く．フェーズ i がデッドロックフリーならば ch_iはルーティングに従って昇順になるように番号をつけることができる．ここで，ネットワーク全体のチャネル番号を $k+1$ 桁の数である $(i, ch_{i1}, ch_{i2}, \cdots, ch_{ik})$ とすると，ルーティングがフェーズ番号の順序通りに進むことから，ネットワーク全体でルーティングに従って昇順に番号が割り振られる．したがって，ネットワーク全体でもデッドロックがおこらない．　□

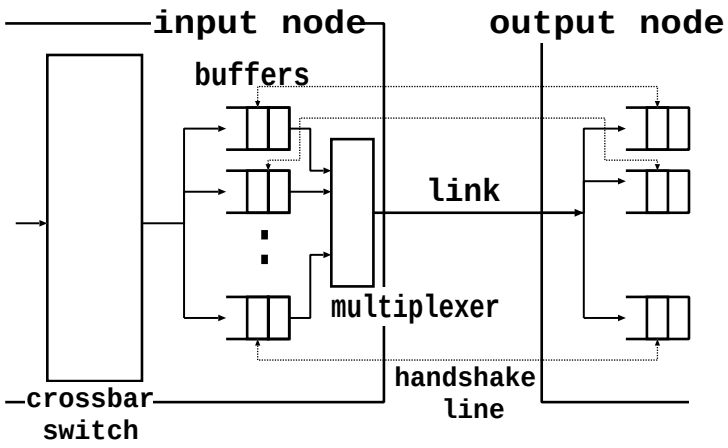


図 3.6: 仮想チャネルの概念図

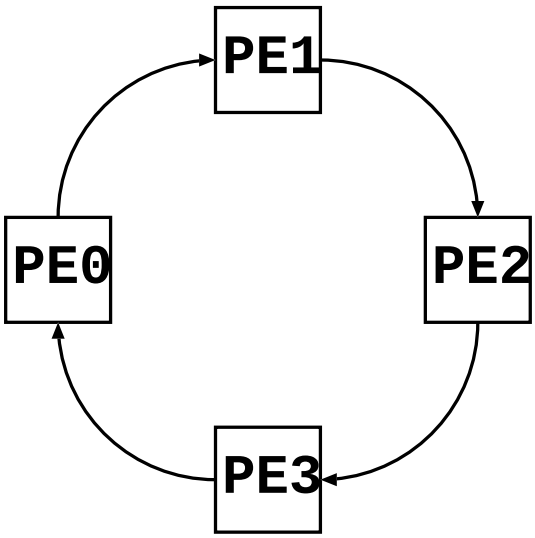


図 3.7: リング網

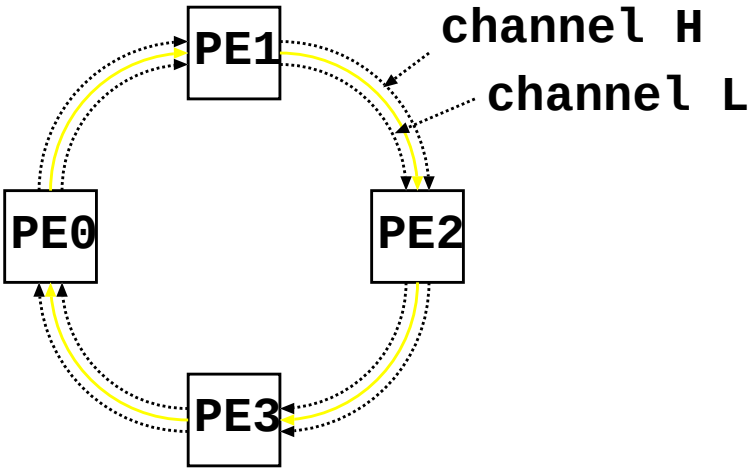


図 3.8: 二本の仮想チャネルを付加したリング網

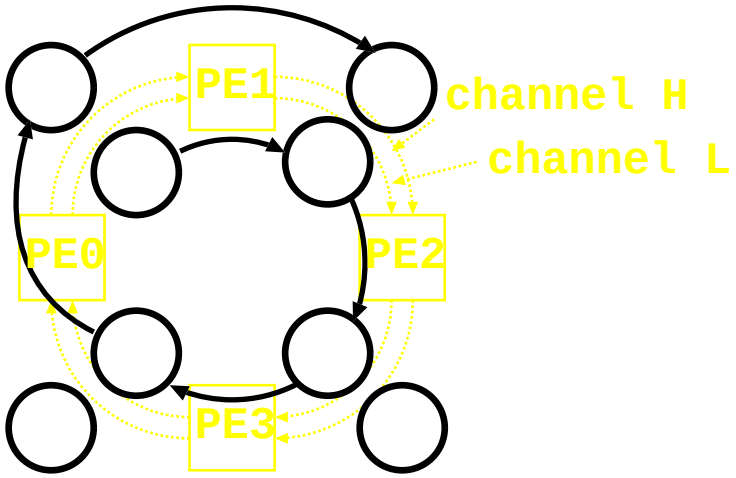


図 3.9: リング網のチャネル依存グラフ

リンク lにおける必要仮想チャネル数 $\mathrm{channels}(l)$ は，以下の定理 2 によって導かれる．

定理 3.2　各リンクにおける必要仮想チャネル数は，各フェーズにおいて，該当するチャネルに必要な仮想チャネル数の和で示される．すなわち，リンク lにおいてフェーズ kがデッドロック・フリーであるために必要な仮想チャネル数を $\mathrm{C}_p(k,l)$ とすると，

$$\mathrm{channels}(l) \;=\; \sum_{k=1}^{n} \mathrm{C}_p(k,l) \tag{3.6}$$

となる．

証明

あるネットワークが n 個のフェーズに分かれていると仮定する．各フェーズは，それぞれが独立したネットワークであると考えられるので，おのおのについて必要仮想チャネル数を各リンクごとに定義することができる．ここで，リンク lにおいてフェーズ kで必要な仮想チャネル数を $\mathrm{C}_p(k,l)$ とする．同様に，リンク lにおいてフェーズ $j(j \neq k)$ で必要な仮想チャネル数は $\mathrm{C}_p(j,l)$ とあらわすことができる．この場合，$\mathrm{C}_p(k,l)$ と $\mathrm{C}_p(j,l)$ で数えられているチャネルは，いずれも同一のリンクに属するチャネルなので，一つのリンクに $\mathrm{C}_p(k,l)$+$\mathrm{C}_p(j,l)$ 個のチャネルが配置されることになる．同様に，すべてのフェーズについて考えると，一つのリンクに$\sum_{k=1}^{n} \mathrm{C}_p(k,l)$ 個のチャネルが配置されていることになる．□

ネットワークに必要な仮想チャネル数は，すべての物理リンクに必要な仮想チャネル数の最大値で示される．すなわち，ネットワークがデッドロック・フリーであるためにリンク lにおいて必要な仮想チャネル数を $\mathrm{channels}(l)$ とした場合，ネットワーク自体の必要チャネル数 CHは

$$CH \;=\; \max_{\forall l \in \Lambda}\{\mathrm{channels}(l)\} \tag{3.7}$$

で示される．

式 (3.7) は，大規模 VLSI へシステムを搭載する際に非常に大きな意味を持つ．結合網が均一ではない階層型相互結合網の場合，結合網の各所によってリンク一つあたりに必要な仮想チャネル数が異なる．しかし，大規模 LSI 上に実装する場合，故障したノードを回避しつつ自律再構成を行うことになるため，すべてのノードが均一な構造である必要がある．そのため，すべてのノードに対して 1 ノードごとの最大の仮想チャネルを搭載する必要がある．したがって，CHをいかに小さくすることができるかが問題となる．

3.5.7　仮想チャネルの効果

ワームホールルーティングの欠点として、パケット同士の衝突によるブロッキングが多いという点があげられる。ワームホールルーティングでは、パケットの先頭がブロックされると、そのパケットが複数のバッファを占有したまま停止してしまうため、図 3.10のように、停止したパケットAが他のパケットBをブロックしてしまうということが起こりやすい。これに対し、図 3.11のように仮想チャネルを設けると、パケットBは仮想チャネルによりブロックされることなく先に進むことができる。

一般に、仮想チャネルの数が多いほどパケットブロックの頻度が少なくなり、リンクの利用効率が向上するが、リンクのキャパシティ以上にパケットを処理することはできないので、ある程度以上仮想チャネルを増やすと、利用効率が上がりにくくなる [25]。

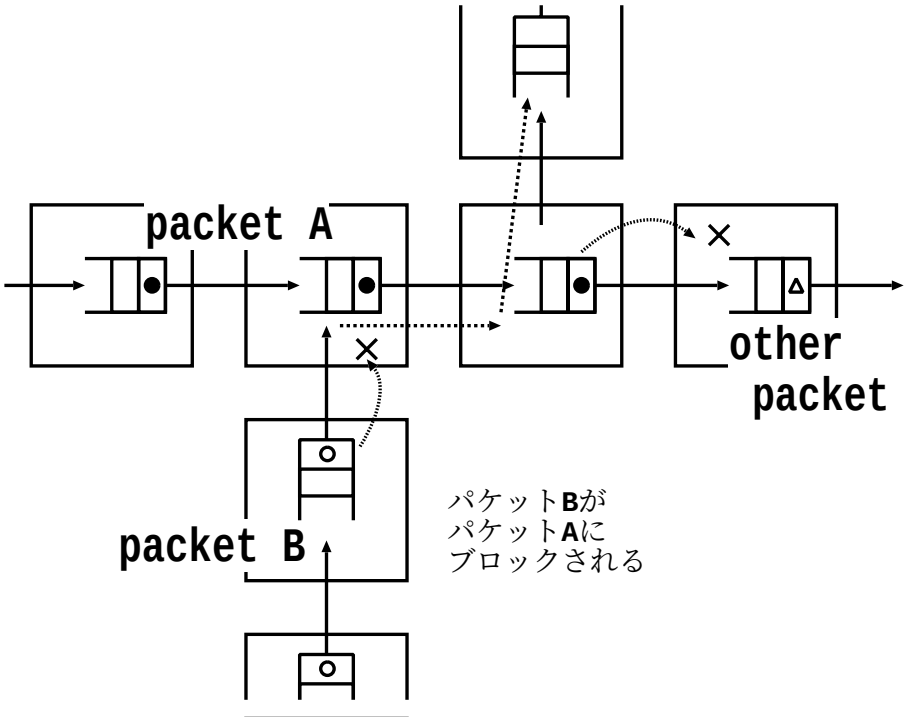


図 3.10: パケットのブロック

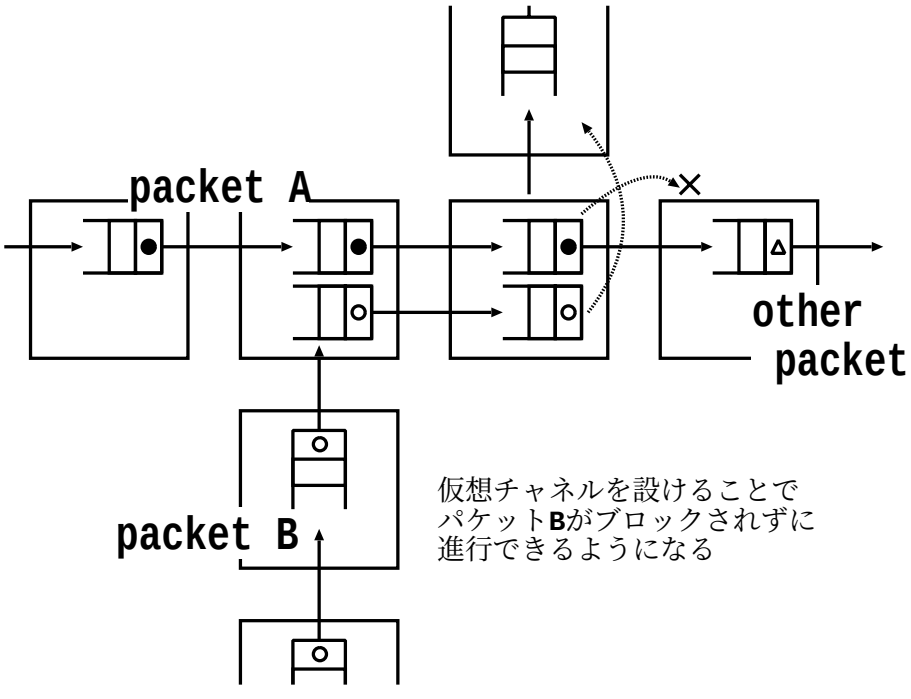


図 3.11: 仮想チャネルの利用によるブロックの回避

3.6　固定ルーティングと適応ルーティング

相互結合網上のルーティングには，出発地から目的地まで必ず同じ経路を通る固定ルーティングと，出発地から目的地までの経路を複数通り選ぶことのできる適応ルーティングがある．固定ルーティングは実装が単純なため，実際の並列計算機でしばしば用いられるが，その一方で，途中経路が混雑したり故障で使えなくなった場合，その経路を迂回することができない．その点適応ルーティングでは複数経路を選択できるため，混雑や故障を回避して経路を選ぶことができる．そのため，適応ルーティングはネットワークのトラフィックが混雑した際にスループットが向上し，また途中経路が故障した際にシステム全体が停止することなく稼働するという利点がある．

3.4節で述べたルーティングのうち，e-cube ルーティングは固定ルーティングに相当する．それに対し，Turn モデル [24] によるルーティングや，構造化バッファ／チャネル法 [26]，Duato の必要十分条件 [28] 等は，適応ルーティングのために開発されたアルゴリズムである．

3.7　まとめ

本章では，ワームホールルーティングおよび，仮想チャネルを用いたデッドロック回避の方法について述べた．

パケット転送方式の一つであるワームホールルーティングは，他の転送方式に対して多くの利点を持った方式である．反面，パケット同士の衝突が多くなるため，結合網のデッドロックが起こりやすくなる．デッドロックに対処するための手法にはさまざまなものが提案されているが，仮想チャネルを付加して論理的にデッドロックを防ぐ方法がワームホールルーティングに対して一般的に用いられている．その際，チャネル依存グラフを作り，チャネル依存グラフが順序関係を持つことを証明できれば，その結合網がデッドロックフリーであることを証明することができる．

第 4 章では，階層型相互結合網の一種である TESH について，効率の良いルーティング方式を提案し，チャネル依存グラフを用いたデッドロックフリーの証明を行う．

第 4 章

階層型相互結合網 TESH

4.1 はじめに

階層型相互結合網 TESH は，下位階層にメッシュ，上位階層にトーラスを用いることにより，双方の結合網の特長を有しつつ通信の局所性を利用したネットワークである．ウェーハ間の配線数を抑え，かつ通信の局所性を利用することで良好なネットワーク性能を持つ．

TESH を用いてマルチプロセッサシステムを実装するためには，デッドロックを回避するために仮想チャネル [25] を複数付加する必要がある．この時に必要な仮想チャネルの数は基本モジュール間リンクの配置の仕方により異なるため，適切な方法によってリンクを配置する必要がある．また，各々のリンク配置法ごとに異なる方法で仮想チャネルの選択を行わなければならない．

デッドロック回避に最低限必要な固定ルーティングに加え，複数経路を選択可能な適応ルーティングを行うことで，ネットワークの性能向上をはかることも可能である．固定ルーティングで必要な仮想チャネルに必要とされるハードウェアの量がさほど大きなものにならないことから，さらに仮想チャネルの数を増やしてネットワークのスループットや遅延時間の軽減を図る余地がある．メッシュやトーラスを含む，k-ary n-cube の適応ルーティング法は，これまで数多く提案されてきた [24][28][30][31]．これらの手法を TESH に適用することができれば，TESH のさらなる性能向上が可能となることが期待される．これらの手法の一部は，TESH の各階層ごとに適用されるローカルな適応ルーティングとしてそのまま適用することが可能である．一方，TESH のルーティングでは上位レベル転送の各次元ごとに基本モジュール（BM）内の過渡的なルーティングを含むため，上記の手法を上位レベルネットワーク全体を利用したグローバルな適応ルーティング法としてそのまま適用することは不可能である．そのため，グローバルな適応ルーティング法か適用できる条

件について細かく検証し，最も効率的な条件のもとで可能な適応ルーティング法を検討する必要がある．

本章ではまず，TESH について，少ないホップ数で通信が可能となるような基本モジュール間リンクの配置法を提案する．また，異なる二種類の配置法について，デッドロックフリーを保証するために必要なルーティング法を提案し，必要な仮想チャネルの数を導出する．提案した配置法により，ランダム通信および FFT の通信パターンによるシミュレーションを行い，動的通信性能について検討する．さらにこれらの固定ルーティング法を基にした TESH のための適応ルーティング法として CS 法，LS 法，DDR 法，DCS 法，BM-adapt の五つの方法を提案する．これらの適応ルーティング法について，いくつかの通信パターンによるシミュレーションを行い，動的通信性能について検討する．最後に，TESH 中のトーラスネットワーク上において，ラウンドトリップを利用して，FFT においてよりメッセージ間の衝突の少ないデータのマッピング法を提案する．トーラス結合を利用した新しいマッピング法により，メッセージ通信によるオーバーヘッドを削減して FFT を高速に実行できることを示す．

4.2 階層型相互結合網 TESH

4.2.1 ネットワーク構成

階層型相互結合網 TESH は三次元 VLSI/ULSI への実装を考慮した結合網である．TESH は，レベル 1 ネットワークをメッシュ状に構成している．これを基本モジュール（BM）とよび，BM 内の各 PE を結合しているリンクを BM 内リンクとよぶ．各 BM は $2^m \times 2^m$ のサイズで構成される [10]．本論文では主に 4×4 のサイズの BM（$m = 2$）について議論する．$2^m \times 2^m$ 個の下位レベルネットワークをトーラス状に接続して上位レベルネットワークを構成する．上位レベルネットワークを構成するためのリンクをここでは BM 間リンクとよぶ．図 4.1 に，二階層の TESH(2,2,0)（三つのパラメータについては後述）を構成した例を示す．図 4.1 では，BM 一つあたり最大 16 本使用できる BM 間リンクのうち 4 本を使用し，全部で 256 個の PE を結合している．三階層の TESH では，さらに 4 本の BM 間リンクを使用して二階層の TESH 同士を結合する．この場合，全部で 4096 個の PE を接続することができる．このようにして L 階層の TESH を構成すると，上位階層ネットワークは $k = 2^m$，$n = 2(L-1)$ の k-ary n-cube となる [10]．

各レベルにつき複数組の BM 間リンクを設けることも可能である．各レベルにつき 2^q 組（つまり 4×2^q 本）の BM 間リンクを設けた場合，最大で $L_{max} = 2^{m-q} + 1$ 階層まで設けることができる．パラメータ m, L, q を用いると，さまざまな種類の TESH を定義することがで

きる．そこで，TESH の種類を表わすために TESH(m, L, q) と表す．なお，TESH(m, L, q) の PE 数 Nは $N = 2^{2mL}$となる [10].

TESH(m, L, q) の PE は，(4.1) 式に示す 2^m進数でアドレス付けされる．

$$\begin{aligned} n &= n_{2L-1}n_{2L-2}...n_1n_0 \\ &= (n_{2L-1}n_{2L-2})...(n_1n_0) \end{aligned} \tag{4.1}$$

(4.1) 式から，i 番目の組である $(n_{2i-1}n_{2i-2})$ は，レベル $i-1$ のサブネットワーク位置となることが分かる．

例えば，$m = 2$ で 3 階層 TESH （4096PE）の場合，四進数で $n = n_5n_4n_3n_2n_1n_0$ のように表現され，n_5n_4は 3 レベルネットワーク，n_3n_2は 2 レベルネットワーク，n_1n_0は BM 内の PE の位置を各々示す．図 4.1 中の番号は，このようにしてアドレス付けされた 2 レベルネットワーク中の BM のアドレス (n_3, n_2) を示している．

$n^1 = (n^1_{2L-1}n^1_{2L-2}...n^1_1n^1_0)$ を含む BM と $n^2 = (n^2_{2L-1}n^2_{2L-2}...n^2_1n^2_0)$ を含む BM が

$$\exists i\{n^1_i = (n^2_i \pm 1) \bmod 4 \wedge \forall j(j \neq i \ \rightarrow \ n^1_j = n^2_j)\} \tag{4.2}$$

を満たした時，リンクで結合する．ただし，$i, j \geq 2$ とする．

(4.2) 式で，n^1と n^2を含む BM アドレスの各桁を比較した時，差が±1 または±3 となる桁が一つ存在して残りの桁は同一の値を持つ時，双方の BM は結合リンクを持つ．例えば，図 4.1 中の BM$(0, 0)$ は，$n_3 = 0$ で，かつ $n_2 = 1$ または $n_2 = 3$ となる BM$(0, 1)$ および BM$(0, 3)$ と，$n_2 = 0$ で，かつ $n_3 = 1$ または $n_3 = 3$ となる BM$(1, 0)$ および BM$(3, 0)$ の，あわせて 4 個の BM と接続する．

4.2.2 BM 間のリンク配置

BM 間リンクは，各 BM の外周部の PE が持つ．どのレベルの BM 間リンクをどの PE に配置するかは自由に決められるが，直径を低く保つことやルーティングを単純化するという観点から，BM の四隅の PE（$n_1 = \{0, 3\}, n_0 = \{0, 3\}$）から出ている二本のリンクは同一レベルの一組のリンクとして使用することが望ましい．また, BM の側面の PE を使用する時は, ホップ数を少なく保つため隣り同士の PE から出ているリンク同士を一組とする．したがって， BM の側面の PE を使用する時（BM 一個あたりの BM 間リンク数が 8 本を越える時）と，BM の側面の PE を使用しない時（BM 一個あたりの BM 間リンク数が 8 本以下の時）では，リンク配置の仕方が異なる．

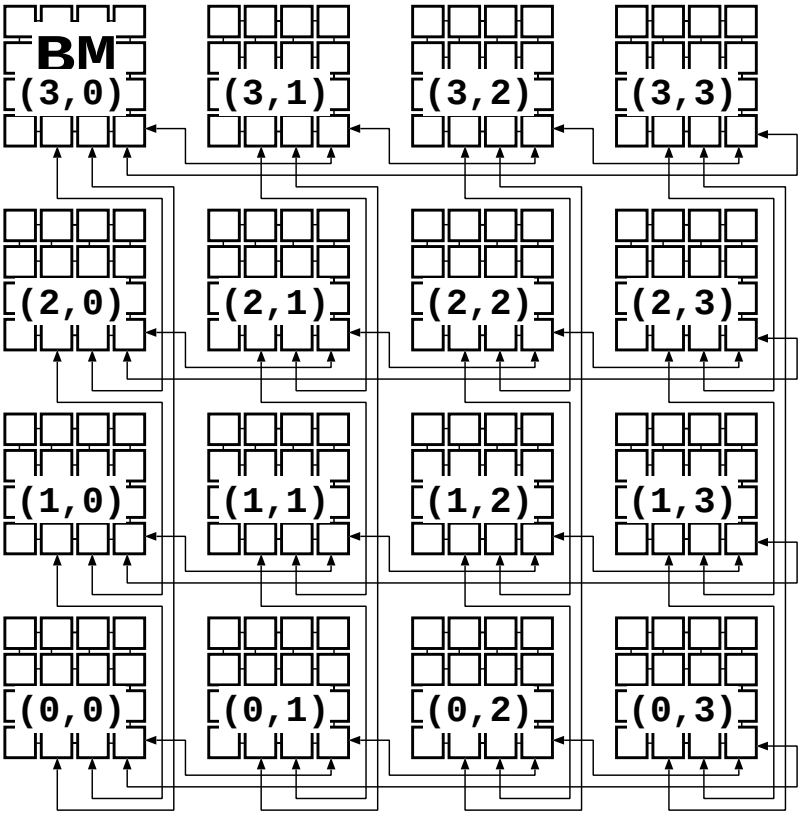


図 4.1: TESH(2,2,0) の構成例

BM 間リンクが 8 本を越える場合

BM の角と側面の PE を使用して固定ルーティングを行う場合，直径を短かくするために，BM 内で隣り合う二つの PE から出ているリンクを同一レベルの一組のリンクとして使用する．さらに，図 4.2のようにリンクを上位レベルから下位レベルまで一列に配置する．

以下に一列配置の定義について述べる．

1. BM 間リンクは 2^q組のグループからなり，各グループには $4 \times (L-1)$ 本のリンクがある．

2. 各リンクは，グループ番号 $g(1 \leq g \leq 2^q)$，レベル番号 $l(2 \leq l \leq L)$，次元 $d(d \in \{\mathrm{V,H}\})$ および向き$\delta(\delta \in \{+,-\})$ によって $(g,l,d\delta)$ とラベル付けされる．

3. グループ g のリンク $(g,2,\mathrm{H}+)$ およびリンク $(g,2,\mathrm{H}-)$ は，四隅のいずれかに配置される（以後，これらのリンクが同じ PE に配置されることを『リンク $(g,2,\mathrm{H}+/-)$ が配置される』と表現する）．

4. 各リンクは，リンク $(g,2,\mathrm{H}+/-)$ を起点として lの小さい順に BM のまわりを時計回りに $(g,l,\mathrm{H}+/-),(g,l,\mathrm{V}+),(g,l,\mathrm{V}-)$,
$(g,l+1,\mathrm{H}+/-)$ の順に配置される．

5. 隣接する BM 間は $(g,l,d+)$ と $(g,l,d-)$ により結合される．

6. $q \geq 1$ の場合，異なるグループの BM 間リンクは BM の中心点を挟んで点対象に配置される．

なお，文中の向き+はアドレスが昇順になる向き，向き−はアドレスが降順になる向きを表しており，次元 V,Hは，それぞれ縦方向リンクまたは横方向リンクであることを示している．

図 4.2に，レベル 3 の TESH における BM 間リンクとそのラベルを示す．BM 間リンクの一列配置により，上位レベルネットワークから下位レベルネットワークへの移動に要するホップ数を低く抑えることが可能となる．さらに，$q \geq 1$ では，上位レベル間の転送に BM の中心部の PE（(1,1)，(1,2)，(2,1)，(2,2)）を使用することがなくなる上に，ルーティングの方向が限定されることから，仮想チャネルの数を低く抑えることが可能となる．

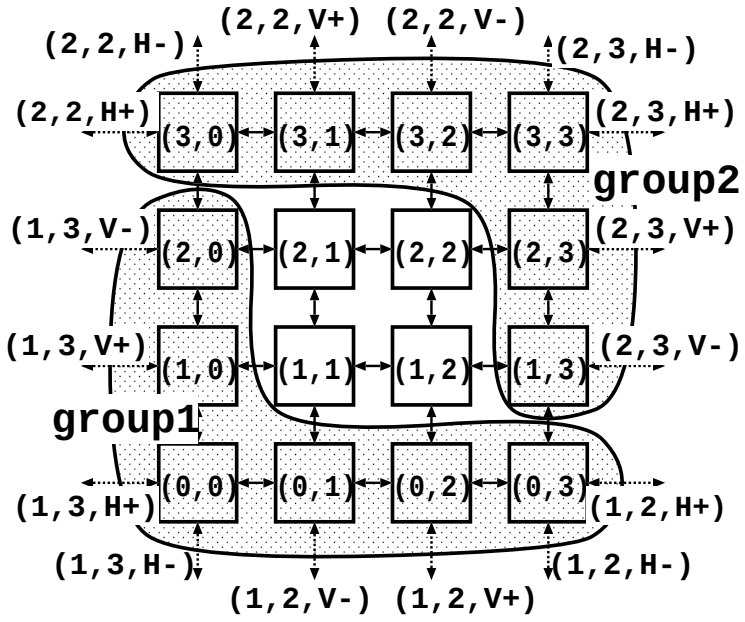


図 4.2: BM 間リンクの配置

BM 間リンクが 8 本以下の場合

実際に TESH を何らかのシステムで使用する際，BM 間リンクの数に制限があることやルーティングの煩雑さを避けることを考えて，BM 一個あたりの BM 間リンクを 8 本以下に抑えて使用することが多いと考えられる．その場合，直径を低く保ち，かつルーティングを単純化するために BM の四隅の PE（$n_1 = \{0,3\}, n_0 = \{0,3\}$）から出ているリンクのみを使用することになる．その場合，考えられる TESH ネットワークは，$TESH(m,2,0)$, $TESH(m,2,1)$ および $TESH(m,3,0)$ のいずれかである．以下に，それらの TESH の上位レベルリンクの配置法を述べる．

1. 各リンクを，グループ番号 $g(1 \leq g \leq 2^q)$，レベル番号 $l(2 \leq l \leq L)$，次元 $d(d \in \{\mathrm{V,H}\})$ および向き$\delta(\delta \in \{+,-\})$ によって $(g,l,d\delta)$ とラベル付けする．

2. リンク $(1,2,\mathrm{H}+/-)$ を PE(0,3) へ，リンク $(1,2,\mathrm{V}+/-)$ を PE(3,3) へ，それぞれ配置する．

3. $TESH(m,3,0)$ では，リンク $(1,3,\mathrm{H}+/-)$ を PE(3,0) へ，リンク $(1,3,\mathrm{V}+/-)$ を PE(0,0) へ，それぞれ配置する．

4.2.3 ルーティングアルゴリズム

TESH の固定ルーティングのアルゴリズムについて述べる．

固定ルーティングは，最上位レベル転送から最下位レベル転送まで順に行われる．すなわち，目標となる BM 間リンクまでの転送を行い，BM 間リンクを用いた転送を行うという手順を最上位レベルから順に繰り返す．BM 間リンクを用いた転送は，各レベルで縦方向転送→横方向転送という順に行う．なお，縦・横方向ともにアドレスが昇順になる+方向とアドレスが降順になる−方向があり，受信先 PE までの距離のより近い方向を選択することになる．ここで，縦の+・−方向をそれぞれ V+方向・V−方向，横方向のそれをそれぞれ H+方向・H−方向と表現する．最後に，目的の BM に到着した時，BM 内の目的の PE への転送を行う．BM 内の転送は，x 方向と y 方向それぞれについて+，−の向きがあり，それぞれを x+，x−，y+，y−と表現する．

$q \geq 1$ の場合，同じレベルの BM 間リンクが複数存在する．その場合は最も近いリンクを選択する．たとえば，図 4.2で BM 内のアドレス $(2,1)$ にあるパケットを 3 レベル縦方向リンクへ転送する場合，アドレス $(2,0)$ から出ているリンクが 3 レベル縦方向リンクとして最も近いため，まず $(2,0)$ へ転送する．

TESH における固定ルーティングアルゴリズムを図 4.3に示す．ここで，送信元 PE のアドレス s を $s_{2L-1}s_{2L-2}...s_1s_0$，受信先 PE のアドレス d を $d_{2L-1}d_{2L-2}...d_1d_0$ とする．関数 get_group_number は，グループ番号を取得する関数である．関数 get_group_number の引数は，送信元 PE のアドレス s，受信先 PE のアドレス d および，向き+/−を区別する変数 routedir となる．関数 outlet_x および outlet_y は，それぞれリンク $(g,l,d\delta)$ が存在する PE の x 座標および y 座標を取得する関数である．引数は，第一引数から順に g,l,d,δを使用する．

ルーティングアルゴリズム中の (a) の部分では，ルーティングタグの値をもとに，次に選択する上位レベルリンクを決定し，上位レベルリンクのある PE への BM 内部の転送を行う．ルーティングアルゴリズム中の (b) の部分では，上位レベルリンクのある PE にパケットが到着した後，必要なホップ数だけ BM 間転送を行っている．最後に，目的の BM に到着した時にルーティングアルゴリズム中の (c) の部分によって目的の PE までの BM 内転送が行われる．

図 4.3のアルゴリズムによる転送例を図 4.4に示す．図 4.4中の (a)，(b)，(c) の部分は，それぞれ図 4.3のルーティングアルゴリズム中の (a)，(b)，(c) に該当する．図 4.4の例では，送信元 PE のアドレスが (10)(12)，受信先 PE のアドレスが (31)(12) である．PE(10)(12) を出発したパケットは，最初に (a) の BM 内転送によって，リンク $(0,2,\mathrm{V}+)$ の入口にあたる PE(10)(02) へ向かう．BM 間リンクの入口に到達したパケットは，(b) の BM 間転送に

よって PE(20)(01) に行き (a) の BM 内転送によって PE(20)(02) へ向かう．同様の手順によって (b)，(a) の転送を繰り返して，リンク $(0, 2, \mathrm{H}+/-)$ の入口にあたる PE(30)(03) へ向かい，さらに (b) の BM 間転送によって PE(31)(03) へ行く．PE(31)(03) は受信先 PE と同じ BM の中にあるので，最後に (c) の BM 内転送によって受信先 PE である PE(31)(12) へ到達する．

固定ルーティングにより転送を行った場合，送信元 PE からの最初の BM 内転送（図 4.3 の (a) の最初のループにおける BM 内転送，または図 4.4のソース PE から BM 間リンクに至る (a) の矢印）と受信先 PE までの最後の BM 内転送（図 4.3, 図 4.4の (c)）は，BM 内の中心部のリンクを通過する可能性があるが，それ以外の BM 内転送は BM 周囲のリンクのみを通過する．そこで，仮想チャネル割り当ての都合上，これ以降はこれらを分けて考える．

4.3 TESH の固定ルーティング

4.3.1 デッドロックフリー

本節では，TESH の固定ルーティングにおいてデッドロックを回避する方法を説明する．1節で述べたように，BM の側面の PE を使用する時（BM 一個あたりの BM 間リンク数が 8 本を越える時）と，BM の側面の PE を使用しない時（BM 一個あたりの BM 間リンク数が 8 本以下の時）では，リンク配置法が異なる．そのため，各々のリンク配置ごとに異なるデッドロック回避法を考える必要がある．

BM 間リンクが 8 本を越える場合

TESH では，チャネルの循環によるデッドロックが発生する可能性がある．デッドロックを回避するために，これまでさまざまな方法が提案されている [23] [24][26][27][28]．3 章で述べたように，デッドロック回避のためにはルーティングに制約を与える方法と複数の仮想チャネルを付加する方法がある．TESH のような複雑な構造を持つ結合網でルーティングに制約を与える方法を用いた場合，制約が厳しくなることによって著しい性能低下を招くおそれがあるため，本稿では，物理リンクに複数の仮想チャネルを付加する方法 [23][26][27] を適用する．

以下，4.2.3節で示したルーティングアルゴリズムがデッドロックフリーであることを保証するために必要な仮想チャネル数について考察する．ここでは仮想チャネル割り当ての都合上，4.2.3節で示したルーティングアルゴリズムを場合分けする．目的の BM 間リンク

```
Routing Algorithm for a Level-L TESH:

Routing(s,d);
source;s={s_{2L-1},s_{2L-2},...,s_0};  destination;d={d_{2L-1},d_{2L-2},...,d_0};
tag;t_{2L-1},t_{2L-2},...,t_0;     group;g ;

  for i = 2L-1:2;
    if (d_i-s_i+2^m) mod 2^m <= 2^m/2 then
            routedir = plus; t_i= (d_i-s_i+2^m) mod 2^m;
    else  routedir = minus; t_i= 2^m- (d_i-s_i+2^m) mod 2^m; endif;

    g = get_group_number(s,d,routedir);

    while(t_i != 0) do
      if i is even number then
              outlet_node_x= outlet_x(g,i/2+1,H,routedir);
              outlet_node_y= outlet_y(g,i/2+1,H,routedir);endif;
      if i is odd number then
              outlet_node_x= outlet_x(g,i/2+1,V,routedir);
              outlet_node_y= outlet_y(g,i/2+1,V,routedir);endif;
      BM_routing(outlet_node_x, outlet_node_y);

      if routedir = plus then send packet to next BM;
      else             send packet to previous BM; endif;

      t_i= t_i- 1;
    endwhile;
  endfor;

    BM_routing(d_1,d_0);
end.


BM_routing(d_x, d_y);
source;s_x,s_y; destination;d_x,d_y;
tag;t_x,t_y;

  t_x= d_x- s_x;
  t_y= d_y- s_y;
  while(t_y != 0) do
    if t_y> 0 then move packet to upper node; t_y= t_y- 1; endif;
    if t_y< 0 then move packet to lower node; t_y= t_y+ 1; endif;
  endwhile;
  while(t_x != 0) do
    if t_x> 0 move packet to right node; t_x= t_x- 1; endif;
    if t_x< 0 move packet to left node; t_x= t_x+ 1; endif;
  endwhile;
end.
```


図 4.3: TESH のルーティングアルゴリズム

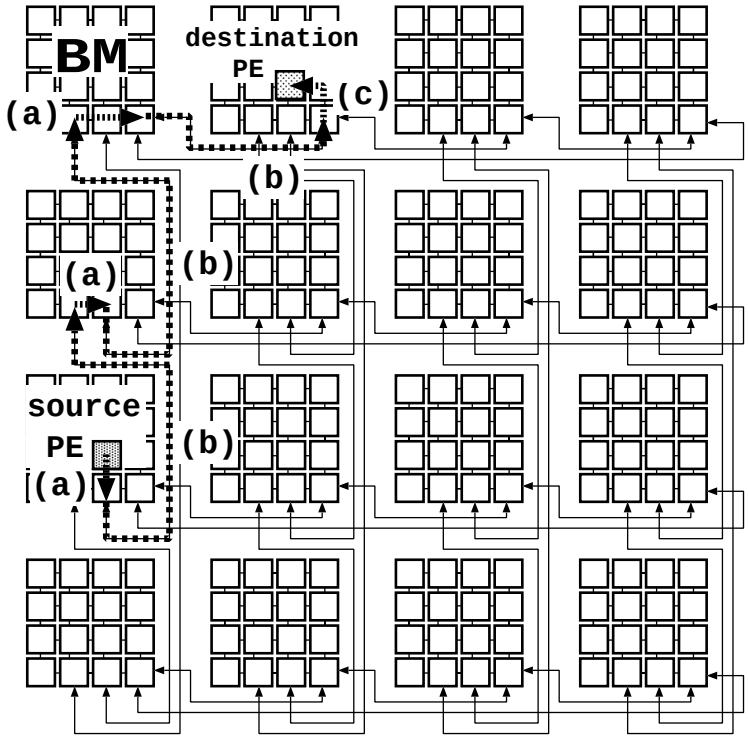


図 4.4: TESH の転送例

へ向かうまでのBM内転送（図4.3の(a)）を，ループの最初のイタレーションとそれ以外に分け，さらに目的のBMに到着後の受信先PEまでの最後のBM内転送（図4.3の(c)）を分けて考える．すると，以下の3つのランクに分けることができる．

ランク1　送信元PEから，最初のBM間リンクに到達するまでのBM内転送（図4.3の(a)の最初のイタレーション）

ランク2　受信先PEの存在するBMに到達するまでのBM間転送（図4.3の(a)の残りおよび(b)）

ランク3　受信先PEの存在するBMに到達してから，受信先PEまでのBM内転送（図4.3の(c)）

すると，パケットの転送は（ランク1）→（ランク2）→（ランク3）
という順序で行われることになる．ランク2についてはトーラスの形状をしているので最低2つのチャネルを必要とする．ランク1とランク3はメッシュ状をしているので1つのチャネルで良い．そこで，ランク1の転送用チャネルとしてチャネル0，ランク2の転送用チャネルとしてチャネル1とチャネル2，ランク3用としてチャネル3を割り当てる．
ここで定理4.1が成り立つ．

定理4.1　ランク1を，送信元PEから最初のBM間リンクに到達するまでのBM内転送，ランク2を，受信先PEの存在するBMに到達するまでのBM間転送，ランク3を，受信先PEの存在するBMに到達してから受信先PEまでのBM内転送とする．ランク1にチャネル0，ランク2にチャネル1とチャネル2，ランク3にチャネル3を割り当て，以下の条件でランク2のチャネルを使い分けるとき，TESHの固定ルーティングはデッドロックフリーとなる．

（条件1）ランク1からランク2への移行時にはチャネル1を使用
（条件2）トーラスのラウンドトリップ時にチャネル2に移動
（条件3）各レベルについて，縦方向または横方向転送が終了した時点でチャネル1に移動

証明
図4.3で示したルーティングアルゴリズムによりチャネルに循環が生じないことを示すため，各チャネルにチャネル番号を割り当てる．
パケットの転送は（ランク1）→（ランク2）→（ランク3）という順序で行われることになるため，各ランクごとにチャネル番号を割り当てればデッドロックフリーが証明されることになる．ランク1とランク3については，以下のようにチャネル番号を定める．

$$\begin{cases} (0, n_1), & \text{y+方向のチャネル,} \\ (1, 4-n_1), & \text{y−方向のチャネル,} \\ (2, n_0), & \text{x+方向のチャネル,} \\ (3, 4-n_0), & \text{x−方向のチャネル,} \end{cases}$$

なお，n_1, n_0は，各チャネルの送信元側 PE アドレスの下位 2 桁を表わす．つまり，あるチャネルが PEn^sから PEn^dの間を結合するチャネルなら，n_1, n_0はそれぞれ n_1^s, n_0^sとなる．また，チャネル番号は左側が上位の桁になっており，チャネル番号の大小関係の比較は左側の数字から順に行う．

ランク 2 については，上位レベルチャネル（ここでは BM 間リンクのチャネルの他に同レベル同次元の BM 間リンクの間を結ぶ BM 内チャネルも含む）の他に異なるレベル・次元のチャネル間を結ぶ BM 内チャネルを持つ．そこで，以下のようにチャネル番号 (l', c_h, n') を割り当てる．

ここで

$$l' = \begin{cases} (L-l)\times 4, & \text{レベル } l \text{縦方向の} \\ & \text{上位レベルチャネル,} \\ (L-l)\times 4+1, & \text{レベル } l \text{縦方向のリンクと} \\ & \text{レベル } l \text{横方向リンクを結ぶ} \\ & \text{BM 内のチャネル,} \\ (L-l)\times 4+2, & \text{レベル } l \text{横方向の} \\ & \text{上位レベルチャネル,} \\ (L-l)\times 4+3, & \text{レベル } l \text{横方向リンクと} \\ & \text{レベル } l-1 \text{ 縦方向リンクを} \\ & \text{結ぶ BM 内のチャネル,} \end{cases}$$

$c_h =$（使用した仮想チャネル），
（1:チャネル 1，2:チャネル 2）

l'は，レベルや次元が変わるごとに番号が昇順に変わる．また c_hは，各次元においてトーラスのラウンドトリップ時に番号が上昇する．

次に n'は，番号が昇順になるように各 PE のアドレス番号または全 PE 数 Nに対するアドレス番号の補数を割り振る．アドレス番号として，2 章で定義したアドレス n を使用した場合，ルーティングにしたがってアドレスが単調増加するためには V+が V−よりも右または上にある必要があるが，実際には BM の左側面と上側面では，前者が後者の左または下にあるため，アドレスの一部を付け変えた新アドレスνを導入する．ここで用いられるアドレスνは，2 章で定義されたアドレス n について，$n_1 = 3 \wedge n_0 = 1, 2$ となる PE（BM の上側面の PE）の n_0を $4 - n_0$に置き換え，$n_0 = 0 \wedge n_1 = 1, 2$ となる PE（BM の左側面の PE）の n_1を $4 - n_1$ に置き換えてつけられるアドレスである．このようにしてνを定めると，BM 中の PE$(1, 0)$ と PE$(2, 0)$，PE$(0, 1)$ と PE$(0, 2)$ のアドレスがそれぞれ置き変わる．

以上により，n'は以下のように定められる．

$$n' = \begin{cases} \nu, & \text{V+方向，または} \\ & \text{H+方向のチャネル,} \\ N - \nu, & \text{V−方向または} \\ & \text{H−方向のチャネル,} \end{cases}$$

以上のようにチャネルの番号を定めると，ルーティングに従って単調にチャネル番号が増加するため，デッドロックフリーが証明される． □

TESH(2,3,1) の場合，必要なチャネル数は図 4.5のようになる．図 4.5に示される数字が，必要な仮想チャネルの数である．図 4.5において，(A) のリンクではランク 1・ランク 3 で 1 つずつチャネルを使用し，ランク 2 で 2 つのチャネルを使用することになる．また，(B) の部分でランク 1，ランク 3 およびランク 2 の 1 つのチャネルを使用することになるため，合わせて 3 つのチャネルが一つの物理リンクを共有することになる．

BM 間リンクが 8 本以下の場合

一般的な TESH のルーティングの手順を述べる．

固定ルーティングは，最上位レベル転送から最下位レベル転送まで順に行われる．すなわち，目標となる BM 間リンクまでの転送を行い，BM 間リンクを用いた転送を行うという手順を最上位レベルから順に繰り返す．BM 間リンクを用いた転送は，各レベルで縦方向転送→横方向転送という順に行う．最後に，目的の BM に到着した時，BM 内の目的の PE への転送を行う．BM 内の転送は，x 方向と y 方向それぞれについて+，−の向きがあ

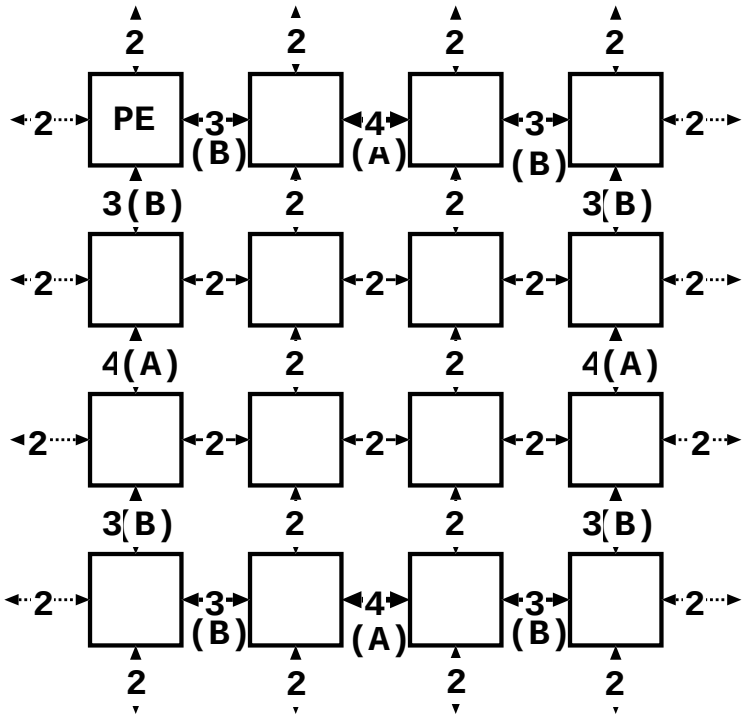


図 4.5: TESH の最大仮想チャネル数

り，それぞれを x+，x−，y+，y−と表現する．なお，BM 内リンクを用いた転送は，必ず y+もしくは y−方向のルーティングを行い，次に x+もしくは x−方向の転送を行う．

なお，$q \geq 1$ の場合同じレベルの BM 間リンクが複数存在する．その場合は最も近いリンクを選択する．

一般に，L レベルの TESH のルーティングは以下の 3 フェーズに分けることができる．

フェーズ 1 送信元 PE から BM の四隅 ($n_1 = 0$ or 3 または $n_0 = 0$ or 3 を満たす PE) に到達するまでの BM 内転送

フェーズ 2 レベル $j(2 \leq j \leq L)$ の転送

フェーズ 3 BM 間リンクの出口から受信先 PE までの BM 内転送

ただし，転送の最初のホップで BM の四隅に到達する時は，フェーズ 1 は無視する．

また，フェーズ 2 は，以下のサブフェーズに分けられる．

サブフェーズ 2.i.1 $L-i$ レベル縦方向 BM 間リンクの入口の PE に到達するまでの BM 内転送

サブフェーズ 2.i.2 $L-i$ レベル縦方向 BM 間リンクを使用した BM 間転送

サブフェーズ 2.i.3 $L-i$ レベル縦方向転送を終え，i レベル横方向 BM 間リンクの入口の PE に到達するまでの BM 内転送

サブフェーズ 2.i.4 $L-i$ レベル横方向 AM 間リンクを使用した BM 間転送

ここで，$0 \leq i \leq L-2$ である．BM 内転送を行うサブフェーズでは，通過しない BM 間リンクの，入口 PE を途中で通過するケースがある．このような場合，その PE までの転送は，通過しない BM 間リンクまでの BM 内転送として扱う．たとえば，3 レベルの TESH における BM 内転送で，2 レベル縦方向リンクの入口の PE を通過して 2 レベル横方向リンクへ向かうような場合，2 レベル縦方向リンクの入口の PE まではサブフェーズ 2.1.1，それ以降はサブフェーズ 2.1.3 となる．

すべての送信元 PE と受信先 PE の組み合わせにおいて必ずしもこれらのフェーズすべてが含まれるとは限らないが，これらのうちいくつかが含まれる場合必ず上記の順序を守ることになる．たとえば，レベル 2 の転送（フェーズ 2.1）を終えた後にレベル 3（フェーズ 2.0）の転送が始まるというようなことは起こらない．

TESH ネットワークのデッドロック・フリーを証明する．TESH のルーティングはフェーズ 1〜フェーズ $L+1$ の $L+1$ 個のフェーズに分けることができる．これら各々について仮

想チャネルを用意してデッドロックを防止する．TESH の各フェーズは，メッシュまたはリング網の形状をしているので，まずはじめにメッシュやリングのデッドロック・フリーの証明を行う．

補題 4.1　二次元メッシュで y方向$\rightarrow x$ 方向の順にルーティングを行うものとする．この二次元メッシュはデッドロック・フリーである．

証明　以下のようにチャネル番号を割り振ればチャネル依存グラフにそってチャネル番号が昇順に割り振られるので，デッドロック・フリーが証明される．

$$\begin{cases} (0, n_1), & \text{y+方向のチャネル,} \\ (1, 4-n_1), & \text{y−方向のチャネル,} \\ (2, n_0), & \text{x+方向のチャネル,} \\ (3, 4-n_0), & \text{x−方向のチャネル,} \end{cases}$$

□

補題 4.2　以下の条件でランク 2 のチャネルを使い分け，2 個のチャネルを使用したリング網はデッドロックフリーである．
（条件 1）ルーティング開始時はチャネル 0
を使用
（条件 2）ラウンドトリップ時にチャネル 1 に移動
証明　以下のようにチャネル番号を割り振ればチャネル依存グラフにそってチャネル番号が昇順に割り振られるので，デッドロックフリーが証明される．

$$\begin{cases} (ch, n_i), & +\text{方向のチャネル,} \\ (ch, 4-n_i), & -\text{方向のチャネル,} \end{cases}$$

$c_h =$（使用した仮想チャネル），
（0:チャネル 0，1:チャネル 1）
□

定理 3.2 で示した理論を用いた TESH のデッドロックフリー・ルーティングを考える．

定理 4.2　BM 一個あたりの BM 間リンクが 8 本以下であるとする．このようなチャネ

ル数 2 の TESH はデッドロック・フリーである．

証明　各フェーズで必要となる仮想チャネルの数を考える．定理 3.1 により，各フェーズでデッドロックフリーが保証されれば TESH ネットワーク全体でデッドロックフリーが保証される．フェーズ 1 およびフェーズ 3 は，メッシュの形（あるいはその一部）をしているため，補題 4.1 により必要な仮想チャネル数は 1 となる．

フェーズ 2 のサブフェーズ 2.i.1 および 2.i.3 では BM の周囲のチャネルを使用する．これらのチャネルは経路が一通りなので，サブフェーズ 2.i.2 および 2.i.4 の入口までの距離 d を用いてチャネル番号を $2^m - d$ と定めればデッドロックフリーが証明される．したがって必要な仮想チャネル数は 1 である．フェーズ 2 のサブフェーズ 2.i.2 および 2.i.4 はリング状をしているため，補題 4.2 により必要な仮想チャネル数は 2 である．

うち，フェーズ 2 のサブフェーズ 2.i.2 および 2.i.4 は BM 間リンクを使用したチャネルで，二つ以上のフェーズで同一のリンクを共有しないので，すべての BM 間リンクの必要仮想チャネル数は 2 となる．また，フェーズ 2 のサブフェーズ 2.i.1，2.i.3 およびフェーズ 1 は BM 内リンクを使用したチャネルであるうち，フェーズ 1 は BM の側面を除いた部分のチャネルのみであり，サブフェーズ 2.i.1，2.i.3 は BM の側面のチャネルのみである．さらに，サブフェーズ 2.i.1，2.i.3 に含まれる各サブフェースは，それぞれ BM 内の別々の場所に位置するため，二つ以上のフェーズで同一のリンクを共有しない．これらのサブフェーズ用チャネルは，それぞれがフェーズ 3 と同じリンクを共有するので，BM 内リンクの必要仮想チャネル数の最大値は 2 となる．

CHの定義により，TESH ネットワークに必要な仮想チャネル数は，これらの数字の最大値となるので，ネットワーク全体で必要な仮想チャネル数は 2 となる．　□

4.3.2　チャネルバッファの増加による影響

仮想チャネルの構造を図 4.6に示す．図中，buffers と記述されている部分がチャネルバッファに相当し，この部分に一度に格納できるフリットの数がチャネルバッファのサイズである．ワームホールルーティングでは，図 4.6中のチャネルバッファ中に，分割されたパケットの一部であるフリットが格納され，互いに送信と受信を繰り返しつつパイプライン状にパケットが送られる．この時，パケットサイズに比べてチャネルバッファのサイズが小さいと，ブロッキングなどによってパケットが止まった時パケットの後続の流れも止まってしまい，複数の PE にパケットがとどまることになる．すると，後続フリットがさらに他のパケットの進行を妨害する結果になり，ネットワークのスループットが低下するというホットスポット現象が起こる．特に TESH のような階層型相互結合網の場合，BM 間リン

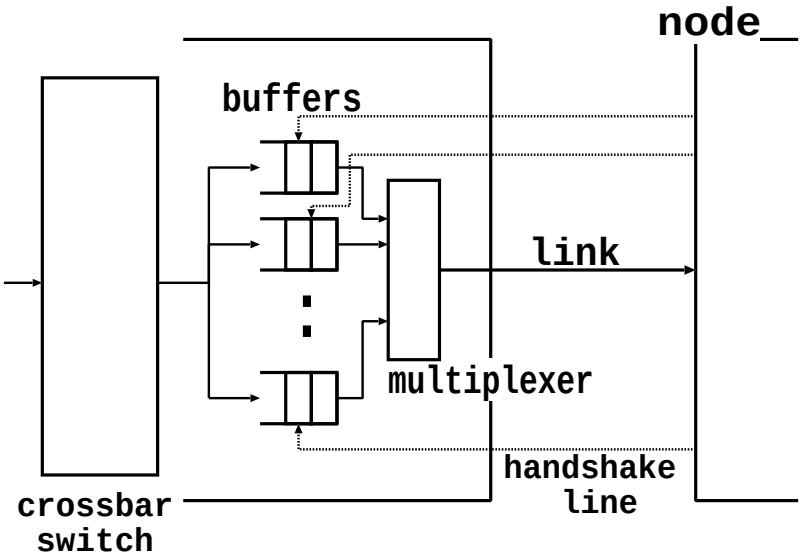


図 4.6: 仮想チャネル

クが混雑して多くのパケットが BM 手前でブロックされるため，後続フリットが BM 内で他のパケットの進行を妨害し，結果として BM 内のパケットの進行まで妨害され，大きな性能低下を招くことが考えられる．

そこでチャネルバッファのサイズを増やせば，後続フリットが後方の PE で止められることが減り他のパケットの進行を妨害することが少なくなるため，性能の向上に寄与することが期待される．

4.3.3 最大ホップ数

4.2.3節で示したルーティングを行った時の TESH の通信性能を評価するため，$m = 2$ の場合について，BM 間リンクを一列配置した時の TESH の最大ホップ数を導出する．TESH の最大ホップ数 $D_{\mathrm{TESH}(m,L,q)}$は，以下のように 4 ステップから導出できる．

1. 送信元 PE を出発したパケットは，最初リンク $(g, L, \mathrm{V}+/-)$ を通過する．これらは，で述べた一列配置の定義 3. および定義 4. により，必ず BM の辺にある（四隅にはない）．特に $q = 2$ では，どの PE にも隣接して $(g, L, \mathrm{V}+/-)$ が存在する．したがって，送信元 PE からレベル L の基本モジュール間リンクに到達するまで（図 4.3(a) の最初のループに相当）の転送回数 D_1は

$$D_1 = \begin{cases} 5, & \text{for } q = 0, \\ 3, & \text{for } q = 1, \\ 1, & \text{for } q = 2, \end{cases} \tag{4.3}$$

となる．

2. 上位階層は 4×4 のトーラスを構成しているので，各レベルの BM 間リンクにおける最大転送回数は縦横両方合わせて $2 + 2 = 4$ となる．ただし，縦方向では中継 BM で一回だけ BM 内転送が含まれるので，各レベルにおいて BM 間を移動するために必要な転送回数 D_2は，

$$D_2 = 2 + 2 + 1 = 5, \tag{4.4}$$

となる．

3. 各レベルの転送の間に行われる BM 内転送の回数 D_3 は，（縦方向）→（横方向）と（横方向）→（次のレベルの縦方向）でそれぞれ二回づつとなるので

$$D_3 = 2, \tag{4.5}$$

となる．

4. 目的の基本モジュールに到達した後の，目的 PE までの転送回数 D_4は，レベル 2 の横方向基本モジュール間リンクが四隅にあるので

$$D_4 = 6, \tag{4.6}$$

となる．

L レベルの TESH では，D_2の転送が $L-1$ 回，D_3の転送がレベル 2 で 1 回，レベル 3 以上で 2 回の合わせて $2(L-2)+1 = 2L-3$ 回起こるので，TESH の最大ホップ数は

$$\begin{aligned} D_{\mathrm{TESH}(m,L,q)} &= D_1 + D_2 \times (L-1) \\ &\quad + D_3 \times (2L-3) + D_4 \end{aligned} \tag{4.7}$$

となる．

表 4.1: TESH の最大ホップ数

PE 数	結合網	格子サイズ	ホップ数	次数
256	2D メッシュ	16×16	30	4
	2D トーラス	16×16	16	4
	3D メッシュ	$8 \times 8 \times 4$	17	6
	ハイパーキューブ	2^8	8	8
	TESH(2,2,2)		14	4
4096	2D メッシュ	64×64	126	4
	2D トーラス	64×64	64	4
	3D メッシュ	$16 \times 16 \times 16$	45	6
	ハイパーキューブ	2^{12}	12	12
	TESH(2,3,1)		25	4

表 4.1に，TESH の各パラメータと最大ホップ数の関係およびメッシュの最大ホップ数を示す[1]．表 4.1より，ハイパーキューブを除く他の結合網に比べて TESH のホップ数やリンク次数が小さくなっていることが分かる．ハイパーキューブはホップ数で TESH に勝るが，リンク次数が大きくなる．

4.4 固定ルーティングによる動的通信性能評価

4.4.1 シミュレーション条件

4096PE からなる TESH ネットワーク上でシミュレーションによる動的通信性能の評価を行う．TESH(2,3,1) の基本モジュール間リンクは，図 4.2に示すように，グループ 1 とグループ 2 の 2 組の BM 間リンクを持つ．

シミュレーションは，BM 間リンクを $2^1 = 2$ 組持つ TESH(2,3,1) と，BM 間リンクを $2^0 = 1$ 組持つ TESH(2,3,0)，およびメッシュ結合について行う．メッシュは，Y 方向→X 方向の順にアドレスを合わせる e-cube ルーティング [23] によりデッドロックを回避す

[1]メッシュ，トーラスおよびハイパーキューブは，最もよく用いられる次元順ルーティング [15] による最大ホップ数を想定して評価しているが，それ以外の方法でも最短距離を通るルーティング法ならば最大ホップ数は同じになるため，同様に評価できる．

る．なおシミュレーションは，ランダム通信と特定の通信パターンが必要な最大値問題，Non-Uniform Transfer の 3 種類で行う．

本実験では，TESH(2,3,1) の仮想チャネル数は 4，TESH(2,3,0) の仮想チャネル数は 2 または 4 である．メッシュの仮想チャネル数は 1，2 または 4 としている．パケットの転送方式はワームホールルーティング [15] とし，サイズの大きなメッセージなども一つのパケットで転送出来るものとしている．なお，仮想チャネルのアービトレーション法はラウンドロビンとした．

評価基準は，平均転送時間およびスループットである．まず，Tクロックサイクルの間に特定確率でパケットを送信し，その間の全パケットの転送時間と PE が受け取ったフリットの数を記録する．次に，それらの値をもとにパケットの平均転送時間およびスループットを算出し，グラフにプロットする．以上の処理を，パケット送信確率を変えて複数回実行する．パケットの転送時間は，パケットの先頭がソース PE を出発してから，パケットの最後尾がディスティネーション PE に到着するまでの時間であらわされ，平均転送時間は転送時間の平均値で示される．スループットは，シミュレーション中に全 PE が受け取ったフリットの数を，PE 数および Tで割ったもので，1 クロックサイクルに 1PE が受け取るフリット数の平均値である．なお，パケット長は 16 フリット（2 フリットのヘッダフリットを除く）とし，$T = 20000$ としている．

4.4.2 ランダム通信

各 PE で，パケットの発生確率を変えながら，受信先 PE がランダムなパケットを送信したときの，スループットに対する平均転送時間を図 4.7，図 4.8，図 4.9および図 4.10に示す．図 4.7は，TESH(2,3,1) とメッシュの比較，他は TESH(2,3,0) とメッシュの比較である．また，図 4.7，図 4.8，図 4.10はチャネルバッファのサイズを 2 フリットにした場合，図 4.9は 20 フリットにした場合の実験結果である．図 4.10は仮想チャネルを 4 個に増やした場合の結果である．

各図で，横軸はスループット，縦軸は転送時間である．1 チャネルのメッシュと 2 チャネルのメッシュを比較した場合，後者の方が横軸の伸びが大きい．これは，チャネル数が多いネットワークの方が高いスループットが得られることを示している．メッシュと TESH を比較した場合，メッセージ生成率の低い部分では TESH の平均転送時間はメッシュの半分以下となる．これは，TESH のネットワーク距離がメッシュに比べて短いためである．また，図 4.7より，TESH(2,3,1) はチャネル数 1 のメッシュと比較すると負荷が飽和する点でのスループットが高くなる．チャネル数 4 のメッシュに対してはスループットは低いが，その差は小さい．

また，図 4.8と図 4.9を比較すると，チャネルバッファのサイズが 2 フリットの場合については，負荷が飽和する点で，TESH(2,3,0) は 1 チャネルのメッシュに比べてやや高いスループットが得られるが，2 チャネルのメッシュに比べてかなり低いスループットにとどまる結果となっている．一方，チャネルバッファのサイズを 20 フリットにした場合においては，TESH(2,3,0) と 2 チャネルのメッシュの比較でも，得られるスループットの差ほとんどないことが分かる．このことから，TESH(2,3,0) においてチャネルバッファを増やす方法は有効であることが分かる．

図 4.8と図 4.10を比較すると，仮想チャネルが 4 個に増加したことに対するメッシュの最大スループットの増加がさほどではないのに対して TESH の最大スループットが大きく増加していることが分かる．同じ仮想チャネル数でも，TESH はメッシュに比べて使用できる仮想チャネルの制限が強いため，仮想チャネルの増加によるリンクの利用効率増加の効果を受けにくい．そのため，メッシュに比べて仮想チャネルの増加によるスループット向上の余地が大きい．

パケット長を変化させた時の TESH(2,3,0) および 64×64 メッシュのスループットを図 4.11に示す．なお，本実験では，各 PE がパケットを送信後，ただちに次のパケットを送信するものとして実験を行っている．実験結果から，パケット長が大きくなるに従って，チャネルバッファのサイズを 20 とした TESH のスループットが低下していることが分かる．これは，パケットが長くなると，チャネルバッファがパケット全体を収容しきれなくなるために，ホットスポット効果を十分に緩和できないことが原因であると思われる．ただし 100 フリット以上の大きなパケットについても，チャネルバッファを増やした TESH は増やさない TESH に比べて 1.33 倍程度スループットが増加している．

ランダム通信における BM 間リンクの利用効率を測定した．ランダム通信では，全パケットの 255/256 は BM 間リンクのいずれかを通過するため，BM 間リンクの利用効率を測定することにより，TESH の理論的な最大性能を知ることができる．まず，TESH ネットワーク中の全 BM 間リンクについて，Tクロックサイクルの間にリンクが使用された回数を記録する．次に，記録された値をもとに利用効率を求める．以上の動作を，パケットの送信確率を変えながら複数回実行する．利用効率は，全 BM 間リンクが使用された回数の合計を，全 BM 間リンク数および Tで割った値として定義されるもので，1 クロックサイクル中に各 BM 間リンクが使用される確率を示している．

スループットに対する BM 間リンクの利用効率を図 4.12に示す．図 4.12の，横軸はスループット，縦軸は利用効率である．図に示すように，利用効率とスループットはチャネル数に関係なく比例関係を示す．図 4.12から，BM 間リンクが最大効率で使用された場合，スループットはほぼ 5.5×10^{-2}程度となることが分かる．

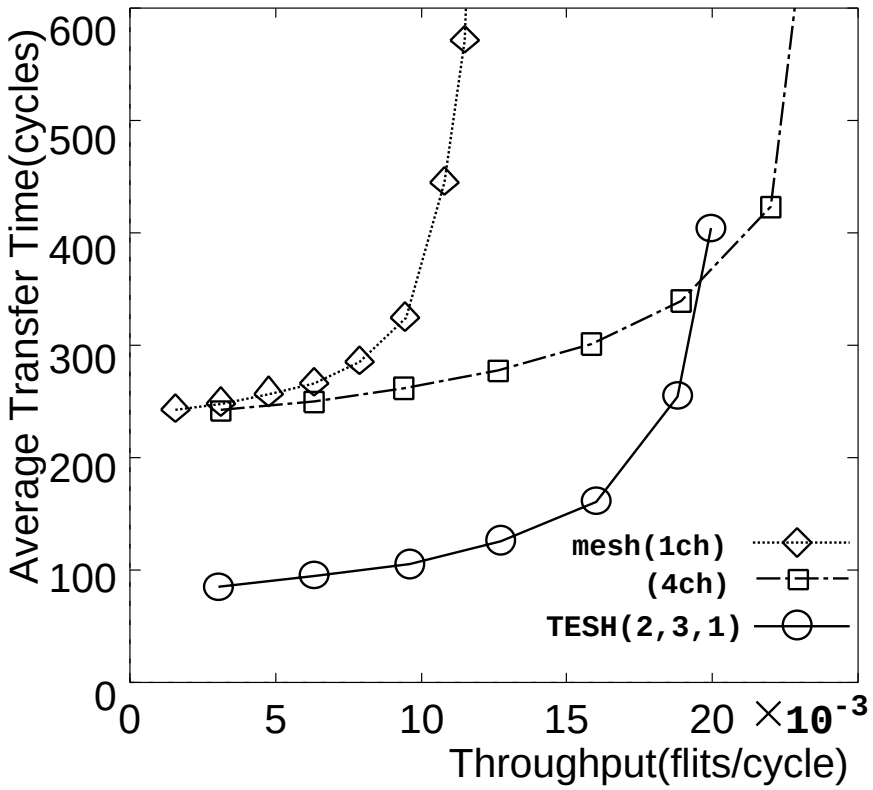


図 4.7: TESH(2,3,1) におけるランダム通信の平均転送時間

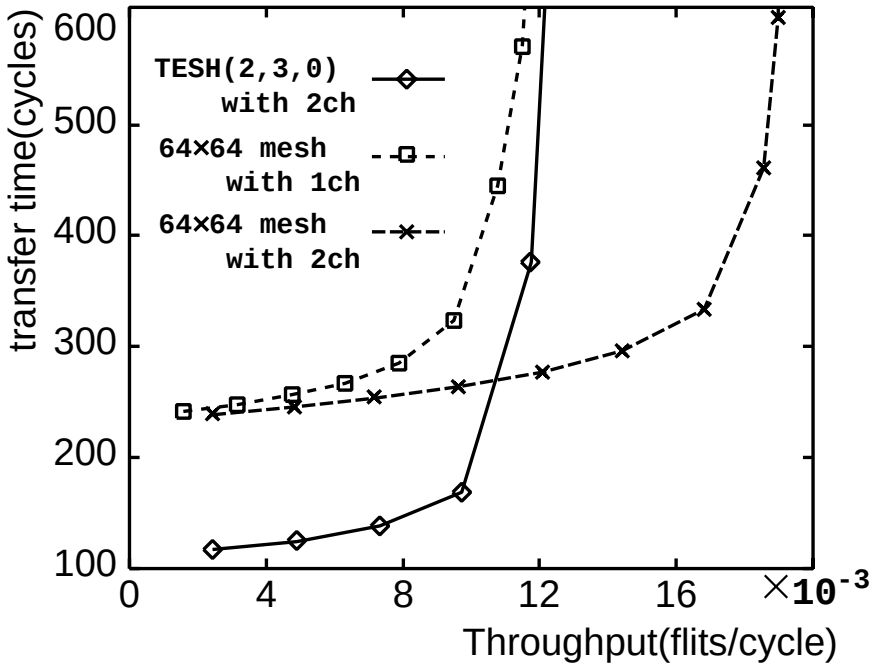


図 4.8: TESH(2,3,0) におけるランダム通信の平均転送時間（バッファサイズ 2）

4.4.3 最大値問題

最大値問題の通信パターンをシミュレーションし，実行時間を測定する．最大値問題は，各 PE に置かれた d 個のデータの値を比較して，大きい方の値を他の PE に送るという処理を繰り返す．本実験では，c 回目の比較を 2^{c-1}離れた PE 同士で行うと仮定し，4096PE の場合 12 回，256PE の場合 8 回の転送を行っている．k回目の通信は，2^{k-1}離れたデータとの間で行う．そこで，N個 $(N < d)$ の PE に，2^{p-1} $(p > \log N)$ 個離れたデータ同士が同じ PE に位置するようにデータを配置すれば，通信の回数は $\log N$回となる．この場合，通信パターンが局所性を持つので，大きい方の値のみを転送すれば充分である．なお，本実験では，値の比較に 40 サイクルを要すると仮定して実験を行っている．

実験結果を図 4.13に示す．図 4.13より，TESH とメッシュの実行時間に差が見られる．最大値問題では，通信距離の長くなるメッシュより TESH の方が実行時間が短い．また，256PE を持つ 16×16 メッシュと TESH(2,2,2) の実行時間の差と，4096PE を持つ 64×64 メッシュと TESH(2,3,1) の実行時間の差を比較した場合，後者の差の方が大きくなる．このように，より多くの PE を多階層で構成した TESH の方がメッシュと比較した性能差は大きなものとなる．これは，PE 数の多い高階層の TESH ほど，メッシュに比べた TESH のネットワーク距離が小さくなるという理由によるものである．

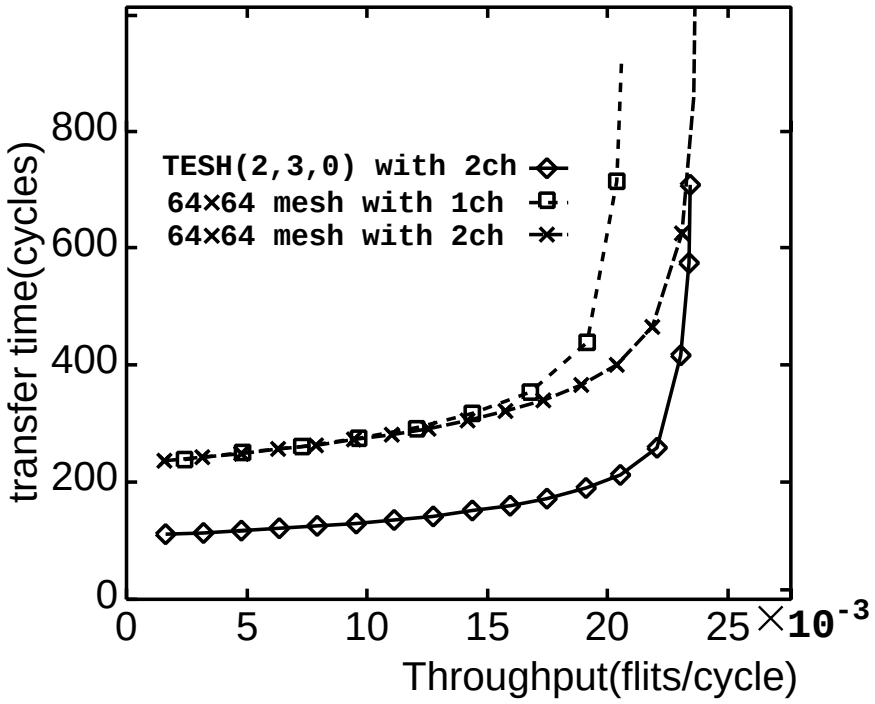


図 4.9: TESH(2,3,0) におけるランダム通信の平均転送時間（バッファサイズ 20）

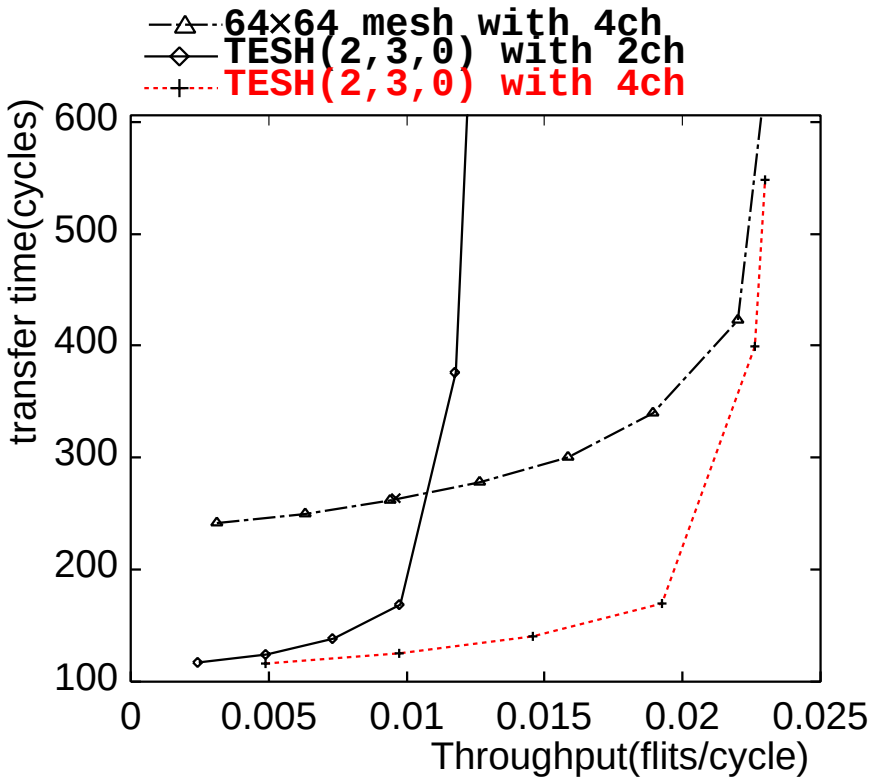


図 4.10: TESH(2,3,0) におけるランダム通信の平均転送時間（仮想チャネルを 4 個に増やした場合

4.4.4 Non-Uniform Transfer

実際に TESH 上で並列処理を行う場合，通信の局所性を利用した近隣の PE との通信の割合が多くなるため，不均一な通信パターンについて検証する必要がある．Non-uniform transfer の評価では，ランダム通信と同様にパケットをランダムに送信し，転送時間およびスループットを測定してグラフにプロットする．Non-uniform transfer はランダム通信と異なり，すべてのパケットを以下の二種類に分け，両者の比率を変えて実験を行っている．

BM 内通信パケット 受信先 PE と送信元 PE を同一の BM 内からランダムに選んだパケット

BM 間通信パケット 受信先 PE と送信元 PE を異なる BM からランダムに選んだパケット

なお実験は，BM 内通信パケットと BM 間通信パケットの比を 1 : 1 にした場合について行っている．このような通信パターンでは，ランダム通信に比べて BM 内通信パケットの

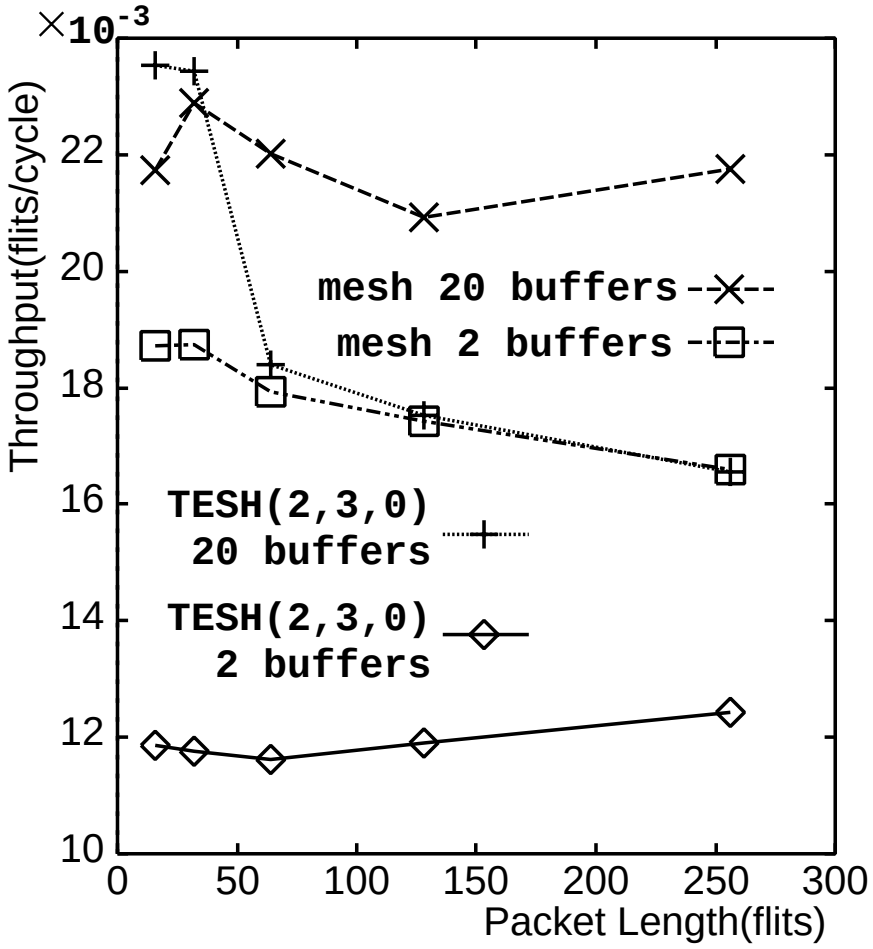


図 4.11: パケット長に対するスループットの変化

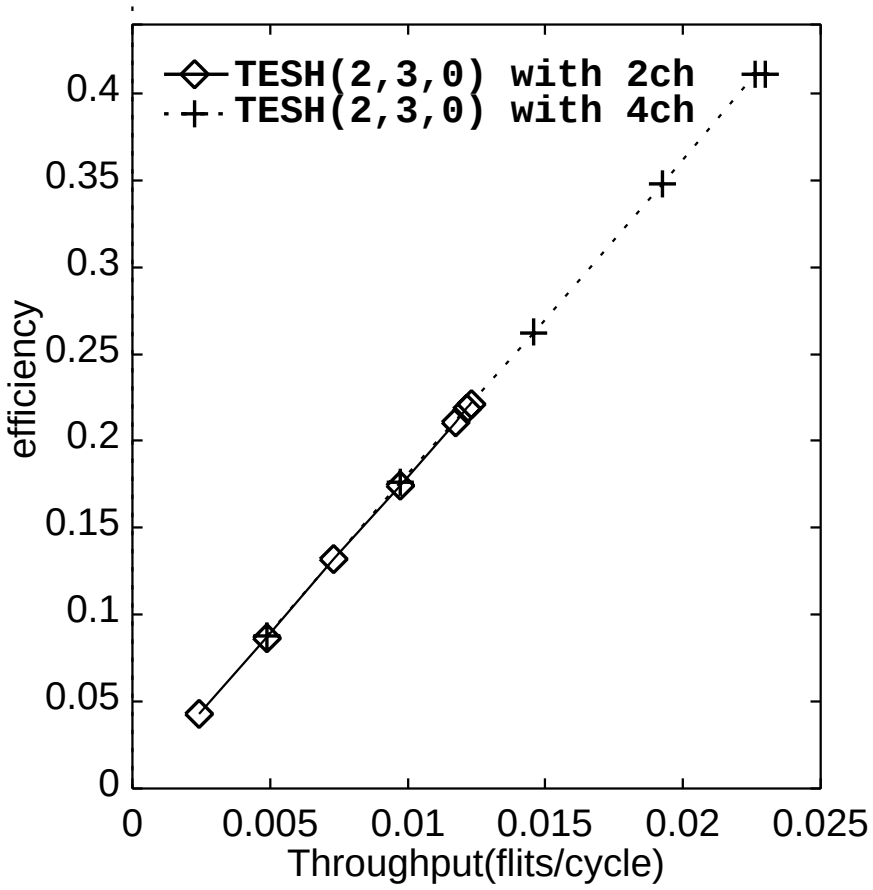


図 4.12: リンクの利用効率

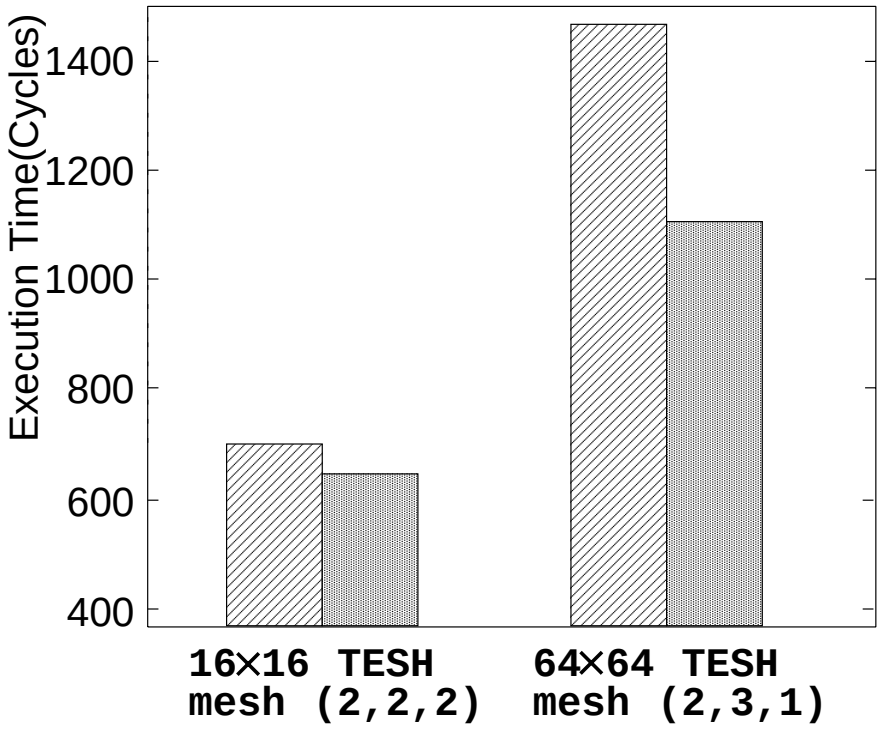


図 4.13: 最大値問題の実行時間

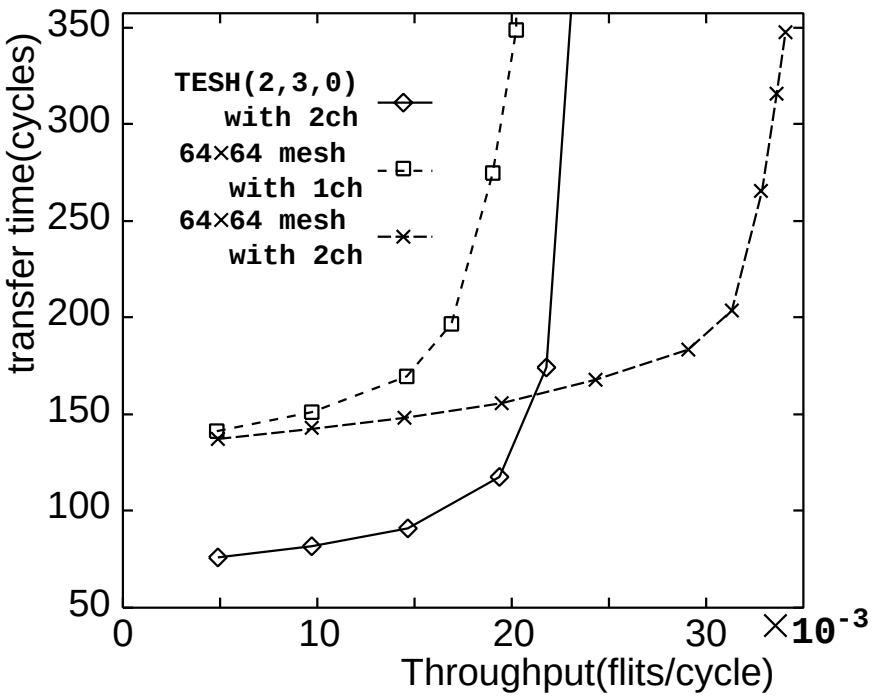


図 4.14: 1:1 の Non-Uniform Transfer の平均転送時間（バッファ数 2）

割合が高くなる．

チャネルバッファのサイズを 2 フリットにした時の平均転送時間を図 4.14，20 フリットにした時の平均転送時間を図 4.15に示す．全体的に，ランダム通信の場合と比較して，メッシュと比較した TESH の最大スループットが若干高めになっている．これは，TESH においてトラフィックの混雑する BM 間リンクを使用しない通信がランダム通信に比べて多いため，ランダム通信で問題となるような，BM 間リンクの混雑によるスループットの低下が若干抑えられているためと思われる．ただし，チャネルバッファ数 2 の実験ではやはり Non-uniform transfer のケースでもチャネル数 2 のメッシュに比べると最大スループットが低くなる．チャネルバッファ数を 20 にした場合，TESH とチャネル数 2 のメッシュのスループットがほぼ同じとなっている．このことは，通信パターンが均一でない場合においても，同様にチャネルバッファ数を増やす方法が有効であることを示している．

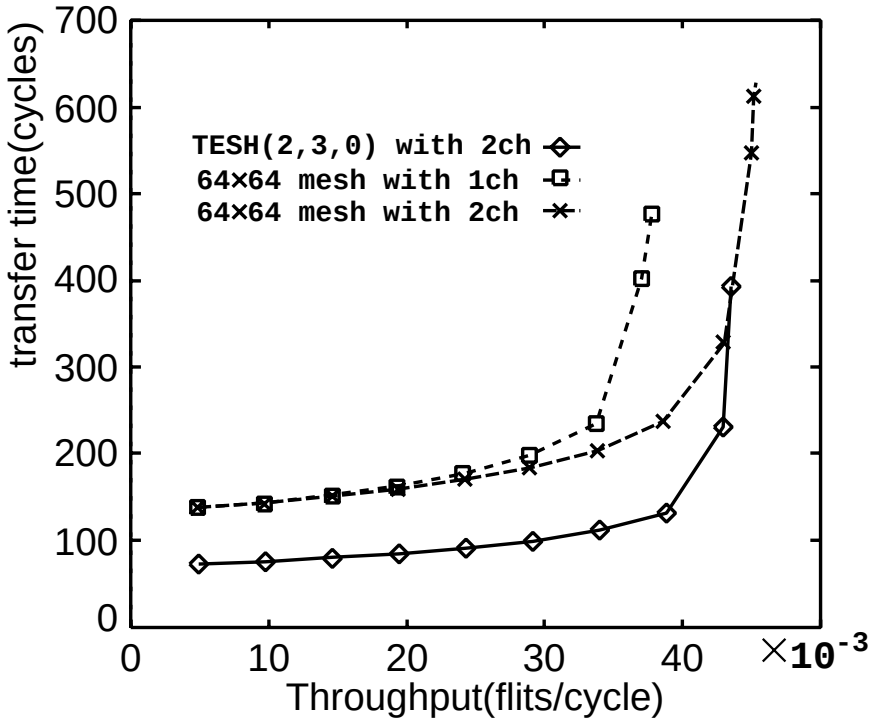


図 4.15: 1:1 の Non-Uniform Transfer の平均転送時間（バッファ数 20）

4.5 TESH の適応ルーティング

多くの結合網では，経路を複数選択する適応ルーティングにより，ネットワークの性能向上，耐故障性の向上などが期待される．TESH は，メッシュとトーラスを組み合わせた相互結合網である．そのため，メッシュやトーラス向けに提案された従来の適応ルーティングをいくつか組み合わせて適用することが可能である．本節では，メッシュとトーラス向けの適応ルーティングを応用したいくつかの TESH の適応ルーティングの手法と，適応ルーティングにおいてデッドロックを回避する方法を説明する．

BM 間リンクにおけるチャネルの動的選択 (CS 法)

リング網のルーティングには，一方向につき通常 2 本の仮想チャネルを必要とする．CS(Channel Select) 法は，これら 2 本の仮想チャネルを特定条件下で自由に選択する手法である．

双方向リング網の固定ルーティングがデッドロック・フリーであることは，補題 4.2 によって証明される．

PE(0) から PE(2) へのルーティング等のようにルーティングの途中でラップアラウンドチャネルを使用しない場合，チャネル L のみが使用される．そのため，本来は使用しないチャネル H に途中で移動した場合や最初からチャネル H を使用することも可能である．ラップアラウンドチャネルを使用する場合も，PE(2) から PE(0) へのルーティング等のように，ラップアラウンドチャネルを通過した時点でルーティングが終了する場合にはやはりチャネル L のみが使用される．そのため，やはり途中でチャネル H に移動した場合や最初からチャネル H を使用した場合も仮想チャネル番号が昇順になる．

4PE リング網の固定ルーティングにおいて，チャネル L のみが使用されるのは以下の二つの条件のいずれかを満たす時である．

- ルーティングの途中でラップアラウンドチャネルを使用しない場合
- ラップアラウンドチャネルを通過した時点でルーティングが終了する場合

また，上記のような条件の時，ルーティング経路に沿って仮想チャネル番号が昇順になるのは，以下の条件のいずれかを満たす時である．

- チャネル L のみを使用した場合
- チャネル H のみを使用した場合
- ルーティングの途中でチャネル L からチャネル H に移動した場合

CS 法は，このような条件下でチャネルを有効に使用するためのアルゴリズムである．CS 法のアルゴリズムでは定理 1 で示した二つの条件に加えて，以下の条件 3 を加える．

(条件 3) チャネル L にいるパケットが以下の条件を満たすとき，パケットはチャネル H を選択することが可能である

- ルーティングの途中でラップアラウンドチャネルを使用する予定がない
- ラップアラウンドチャネルを通過した時点でルーティングが終了する予定である

以下に，CS 法がデッドロック・フリーである証明を述べる．

補題 4.3　以下の条件 1 および条件 2 により 2 つのチャネルを使い分けた双方向リング網はデッドロックフリーである．

(条件 1) ルーティング開始時はチャネル L を使用

(条件 2) ラップアラウンドチャネルの通過直後にチャネル H に移動

(条件 3) チャネル L にいるパケットが以下の条件を満たすとき，パケットはチャネル H を選択することが可能である

- ルーティングの途中でラップアラウンドチャネルを使用しない
- ラップアラウンドチャネルを通過した時点でルーティングが終了する

証明　補題 4.2 と同じ方法でチャネル番号を割り振ればチャネル依存グラフにそってチャネル番号が昇順に割り振られるので，デッドロックフリーが証明される．

□

定理 4.3　BM 一個あたりの BM 間リンクが 8 本以下であるとする．CS 法を用いたこのようなチャネル数 2 の TESH はデッドロックフリーである．

証明

定理 4.2 における TESH のデッドロックフリーの証明中の BM 間リンクのリング網のデッドロックフリーの証明において，補題 4.2 の代わりに補題 4.3 を用いることにより証明される．□

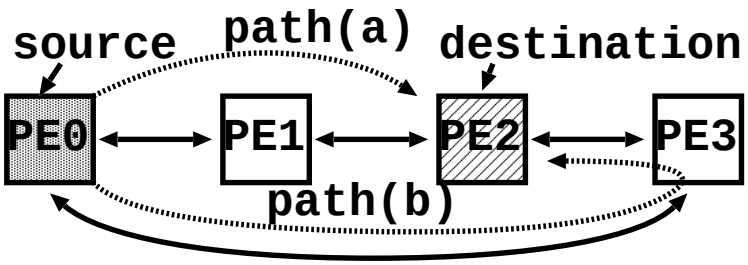


図 4.16: LS 法によるリンクの選択

4.5.1 同一レベルリンクにおける複数方向の選択 (LS 法)

各 BM 間リンクは，2^m個の PE を持つリング状をしている．そのため，$2^m/2$ 個離れた PE 同士の通信では+方向と–方向のいずれのリンクを通っても等距離となる．そこで，このような条件下では両方のリンクを使用できるものとしたのが LS(Link Select) 法である．たとえば，4PE を持つリング網で PE0 から PE2 へルーティングを行う場合，図 4.16のように経路 (a) と経路 (b) のいずれを選択しても 2 ホップで PE2 へ到着する．そのため，経路 (a) と経路 (b) の両方を選択可能とするアルゴリズムである．

+方向と–方向の双方のリンクのうち空いているものを選択する条件は，以下の通りである．

$$|s - d| = 2 \tag{4.8}$$

ここで，s および d はそれぞれソース PE とディスティネーション PE のアドレスを示している．

このようなルーティングアルゴリズムがデッドロックフリーであることは，4.3.1節の補題 4.2 と同様の方法で証明される．

補題 4.4　以下の条件でランク 2 のチャネルを使い分け，2 個のチャネルを使用したリング網はデッドロックフリーである．

（条件 1） ルーティング開始時はチャネル 0 を使用

（条件 2） ラウンドトリップ時にチャネル 1 に移動

（条件 3） 以下の条件を満たすとき，+方向と–方向の双方のチャネルから空いているものを選択し，それ以外の場合はディスティネーション PE が近い方向のチャネルを選

択する．

$$|s - d| = 2$$

ここで，s および d はそれぞれソース PE とディスティネーション PE のアドレスを示すものとする．

証明　以下のようにチャネル番号を割り振ればチャネル依存グラフにそってチャネル番号が昇順に割り振られるので，デッドロックフリーが証明される．

$$\begin{cases} (ch, n_i), & +\text{方向のチャネル}, \\ (ch, 4 - n_i), & -\text{方向のチャネル}, \end{cases}$$

$c_h =$（使用した仮想チャネル），
（0:チャネル 0，1:チャネル 1）
□

定理 4.4　BM 一個あたりの BM 間リンクが 8 本以下であるとする．LS 法を用いたこのようなチャネル数 2 の TESH はデッドロックフリーである．

証明
定理 4.2 における TESH のデッドロックフリーの証明中の BM 間リンクのリング網のデッドロックフリーの証明において，補題 4.2 の代わりに補題 4.2' を用いることにより証明される．□

4.5.2　次元逆転ルーティングによる BM 間リンクの動的選択 (DDR 法)

TESH の固定ルーティングでは，BM 間リンクを通る順序が，上位レベル→下位レベル，縦方向→横方向と決まっている．

一般的な k-ary n-cube でも固定ルーティングでは使用されるリンクの次元順は決まっているが，予め決められた次元順と逆順にルーティングを行う次元逆転ルーティング [31] を用いることにより，リンクが使用される順序を逆転する方法が提案されている．本稿では，

うち動的次元逆転ルーティングを応用して，TESH の適応ルーティングアルゴリズムを考える．

k-ary n-cube の次元逆転ルーティングとしては，静的逆転ルーティング (Static Dimension Reversal) と動的逆転ルーティング (Dynamic Dimension Reversal) の二種類が提案されている．本論文では，うち仮想チャネルを効率良く使用できる動的逆転ルーティング（以下 DDR）を用いる．

DDR 法において，各パケットは DR(Dimension Reversal，次元逆転) 数と呼ばれる値を持つ．パケットの DR 数は，パケットがサブフェーズ $2.p$ から，より順序の低いサブフェーズ $2.q$,（$q < p$）へ移動した回数として定義される．ただし，p および qは $p_1.p_0$，$q_1.q_0$の形式となっているため，p_1，q_1を上位桁，p_0，q_0を下位桁とみなして順序を比較する．DR 数は，以下のように割り当てられる．

1. すべてのパケットの DR 数の初期状態は 0 である．

2. パケットがサブフェーズ $2.p$ に属するチャネル C_iから，より順序の低いサブフェーズ $2.q$,（$q < p$）に属するチャネル C_jへ移動した時，パケットの DR 数をインクリメントする．

DDR 法では，全チャネルが適応ルーティングチャネルと，固定ルーティングチャネルに分けられる．各パケットは最初，適応ルーティングチャネルを使用し，適応ルーティングを行う．パケットがチャネルを獲得した時には，チャネルに対して自身の DR 数を記録する．デッドロックを回避するため，$p \geq q$の時，DR 数が p であるパケットは DR 数が qであるチャネルが空くまで待つことができないという制限を加える．パケットの全出力チャネルが，自分と同じかより少ない値を持つパケットに占領されている場合，パケットは固定チャネルに移動する．すべての進路を，自パケット以下の DR を持つパケットにブロックされた時，パケットは固定ルーティングチャネルへと移動する．固定ルーティングチャネルでは，固定ルーティングを行う．固定ルーティングチャネルでは固定ルーティングを行い，以降適応ルーティングへは戻らない．

適応チャネルにおける適応ルーティングの流れは以下のようになる．k-ary n-cube における DDR ルーティングでは，パケットが適応ルーティングチャネル内において全ての次元を選択することができる．TESH は階層型相互結合網のため，上位レベルの k-ary n-cube を構成する各リンクが，同一 BM 内の異なる箇所に散在する．そのため，k-ary n-cube と異なり，パケットの進路をいくつもの BM 間リンクから自由に選択することはできない．しかし，固定ルーティングによる BM 内転送の途中で，何度か BM 間リンクが存在する PE を通過するため，BM 内転送を中断して BM 間リンクによる BM 間転送を行うことができる．BM 内における BM 間リンクの 4 つの入力側 PE は，4.1のように位置している．BM

内ルーティングにおいてパケットがこれらの PE を通過する時，以下に示す 2 つの経路を選ぶことができる．

経路 1 BM 内転送を中断して BM 間リンクを選択する．

経路 2 BM 内転送を続ける．

以上の条件を満たす時，経路 1 を優先的に選択する．なお，経路 1 を選択した場合本来の次元順を破る次元逆転が起こる．

図 4.17に，DDR を用いた TESH の適応ルーティングの一例を述べる．図 4.17の，網掛けの PE が送信元 PE，BM 中の実線太字の矢印が，固定ルーティングを行う場合のパケットの進路である．この例では，転送途中においてパケットがリンク $(1,3,V+/-)$ 及びリンク $(1,3,H+/-)$ を通過するものと仮定する．固定ルーティングでは，フェイズ 1 において，パケットがリンク $(1,3,V+/-)$ の入口に送られる．しかし図 4.17の例では，その途中でリンク $(1,3,H+/-)$ を持つ PE を通過するため，そこでまずリンク $(1,3,H+/-)$ が他のパケットに占有されずに使用可能な状態にあるか否かの判定を行う．仮に使用可能なら，リンク $(1,3,V+/-)$ へ行く前にリンク $(1,3,H+/-)$ の転送を行う．使用可能でないなら，経路 2 を選択して BM 内ルーティングを継続する．従って，フェイズ 1 の転送では図 4.17の太字実線の経路の他に，太字点線の経路に沿ったルーティングが可能となる．

次に，DDR 法を用いた TESH のデッドロックフリー・ルーティングを考える．

定理 4.5　DDR 法により適応ルーティングを行う TESH について，固定チャネルで用いられるルーティングアルゴリズムがデッドロック・フリーであるとする．このような TESH のルーティングアルゴリズムはデッドロックフリーである．

証明

ネットワークにデッドロックが発生していると仮定すると，同じ集合 P内の他のパケットに使用されているチャネルの解放を待つパケットの集合 Pが存在する．また，P中には $\mathrm{DR}(p_{\max}) \geq \mathrm{DR}(q) \forall \in P$を満たすパケット $p_{\max}$が存在する．ここで，$next(p_i)$ を，p_iが次に選択するチャネルとし，rを，チャネル $c_n = next(p_{\max})$ の DR 数とすると，$r \leq \mathrm{DR}(p_j)$ となる．なお，チャネル c_nはパケット p_jが使用しているものとする．しかしながら，$\mathrm{DR}(p_{\max}) \geq \mathrm{DR}(p_j) \geq (r)$ なので，$p_{\max}$は $next(p_{\max})$ が解放されるまで待つことはできない．パケット $p_{\max}$は固定チャネルに向かい，固定チャネルではデッドロックが起こらないので，ネットワーク中にデッドロックは起こらない．

□

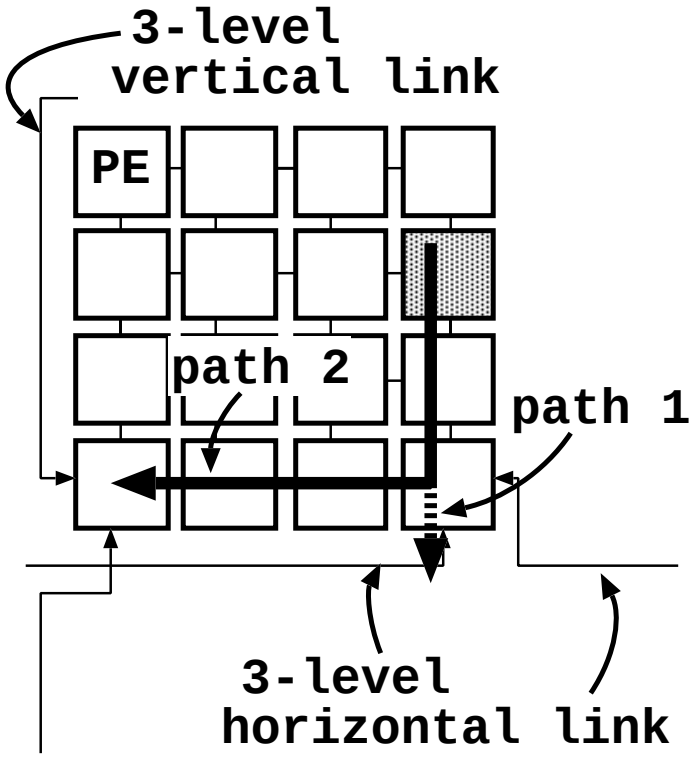


図 4.17: DDR を用いた TESH の適応ルーティング

4.5.3 TESH(2,3,1) におけるチャネルの動的選択 (DCS 法)

図 4.18に，TESH$(2,3,1)$ における BM 内の各リンクの必要チャネル数を示す．それぞれのリンクの必要チャネル数の最大値は 4 だが，実際にはリンクのネットワーク上での位置によって必要チャネル数が異なる．このような結合網を用いたシステムを，故障回避を行う大規模 VLSI 上へ搭載すると，おのおののリンクのネットワーク上での位置が各 PE の故障状態によって変化するため，あらかじめ最大数のチャネルを用意する必要がある．本論文では，それらの余剰のチャネルを，同一リンクの他のチャネルの代わりとして使用することにより，ネットワーク中で使用可能な仮想チャネルを増やす DCS(Dynamic Channel Select) 法によってネットワーク性能の向上をはかる．

一般にネットワーク G は，PE Nとチャネル Cによるグラフ $G(N,C)$ で記述される．多くの場合，PE 間に一本ないしは二本のリンクを配置し，一本のリンクに複数の仮想チャネルが含まれるという形をとる．PE u と PE vの間に配置されるリンク l_{uv}に含まれるチャネルの集合を $C_{uv} \subseteq C$とあらわす．TESH$(2,3,1)$ においては $|C_{uv}| = 4$ である．また，ネットワーク G のルーティング関数 Rは，PE x に存在するメッセージが，ディスティネーション PE yに到達するために次に選択し得るチャネルの集合として，下式のように定義される．

$$R(x,y) \in out(x) \tag{4.9}$$

なお，式中の $out(x)$ は，PE x を入力側 PE とするチャネルの集合である．

C_{uv}に含まれるチャネルのうち，実際に固定ルーティングに使われる固定チャネルの集合を，$C'_{uv} \subseteq C_{uv}$とする．すなわち

$$C' = \bigcup_{\forall p} \bigcup_{x,y \in N} R(x,y) \tag{4.10}$$

$$C'_{uv} = C' \cap C_{uv} \tag{4.11}$$

とする．当然，TESH$(2,3,1)$ では $|C'_{uv}| \leq 4$ となる．また，固定ルーティングに使われない余剰チャネルは $C_{uv} - C'_{uv}$で示される．

以下に，本論文で提案するルーティング法について述べる．

本手法は，余剰チャネルを自由に選択する手法である．本論文で新たに提案される適応ルーティングのルーティング関数 R'は，基本となるルーティング関数 Rで選ばれるチャネルに，余剰チャネルを加えたものなので，

$$R'(x,y) = R(x,y)$$

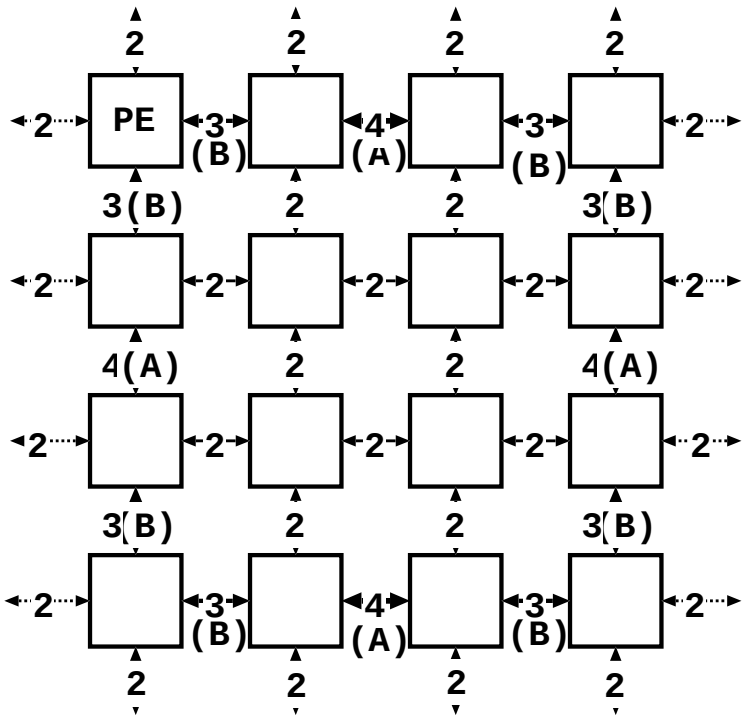


図 4.18: TESH の最大仮想チャネル数

$$+ \bigcup_{v \in R(x,y)} (C_{xv} - C'_{xv}) \tag{4.12}$$

で示される．

チャネルの選択は，固定チャネル C' に含まれるチャネルを優先し，固定チャネルが塞がっている場合のみ余剰チャネル $C - C'$ のチャネルを使用する．

定理 4.6　TESH ネットワークの固定ルーティングのルーティングがデッドロックフリーであるとする．この時，DCS 法によるルーティングは，デッドロックフリーである．

証明

(4.11) 式で示されるルーティング関数は，必ずリンク l_{xv} を通過し，リンク l_{xv} の出力側 PE x' に到達する．この時使用されるチャネル c は，以下の二種類が考えられる．

(1) C'_{xv} に含まれる固定チャネル

(2) $C_{xv} - C'_{xv}$ に含まれる余剰チャネル

また，PE x に到達するまでに

(i) チャネル $c_0 \in C'$を通過したことがある

(ii) チャネル $c_0 \in C'$を通過したことがない

場合の 2 通りが考えられる．この組み合わせのうち，チャネル間の順序関係が問題になるのは (1) かつ (i) の場合のみである．

ルーティング関数 Rがデッドロックフリーであると仮定すると，R'と Rは同じ経路の異なるチャネルを使用することになる．そのため，R'で使用されたチャネル $c_0, c \in C'$は，Rでも使用される．したがって $c_0 > c$ である．よって，R'に関するチャネル依存グラフは循環を持たない． □

以上から，チャネルの動的選択の手法はデッドロックフリーであることが分かる．

上記の適応ルーティングは，各リンクごとのローカルなルーティングアルゴリズムである．TESH の DCS 法では，上記のような同一リンクからの複数チャネル選択を，すべてのリンクについて行う．BM 間リンクの場合，固定ルーティングに必要な仮想チャネル数は 2 個であり，ルーティングアルゴリズムに従って 2 個の固定チャネルのうち片方を選択する．DCS 法では，固定ルーティングで選ばれる固定チャネルの他に，2 個の余剰チャネルを選択できる．BM 内リンクでは，場所によって必要仮想チャネル数が異なる．必要仮想チャネル数が 3 個のリンクでは固定ルーティングで選ばれる固定チャネルの他に，1 個の余剰チャネルを選択することが可能である．

4.5.4 BM に対する適応ルーティング (BM-adaptive)

メッシュ網では，バイパス用のチャネルを用いて，最短距離を保証する範囲で転送経路を自由に選ぶ適応ルーティングが可能であることが知られている．本手法では，余剰のチャネルをバイパス用のチャネルとして，基本モジュール（Basic Module,BM）に対して Duato の fully adaptiverouting による適応ルーティングを行う．

本手法のチャネル選択は以下のようになる．まず，次元順ルーティングのためのチャネルを固定チャネルとし，これらのチャネルが空いている場合，e-cube ルーティングに従って（上，下方向）→（右，左方向）の優先順位で適切なチャネルが選択される．固定チャネルが塞がっている場合，固定ルーティングでは用いられない余剰チャネルの中から適切なチャネルを選択する．その際の選択基準は以下の通りである．なお，ここで x_1, x_0は，それぞれメッセージの現在値 PE x の BM 内の縦方向アドレスと横方向アドレス，y_1, y_0は，ディスティネーション PE yの BM 内アドレスの縦方向アドレスと横方向アドレスである．

1. $x_1 < y_1$のとき，上方向リンク中のチャネルを選択

2. $x_1 > y_1$のとき，下方向リンク中のチャネルを選択

3. $x_0 < y_0$のとき，右方向リンク中のチャネルを選択

4. $x_0 > y_0$のとき，左方向リンク中のチャネルを選択

なお，上記のうち複数の条件が重複した時は，番号の小さい方の条件を優先する．

以上のことを，TESH(2,3,1) を例にして述べる．TESH(2,3,1) でリンク一つあたりに必要なチャネル数は，4.18のように，ネットワーク上におけるリンクの位置により異なる．ここで，BM の中心部付近のリンクを例にとると，フェーズ 1，フェーズ 3 の二つの目的のためにそれぞれ 1 個ずつ仮想チャネルを必要とする．そこで，4 個の仮想チャネルを以下のように使い分ける．

チャネル 1 送信元 PE から，最初の BM 間リンクに到達するまでの BM 内転送（フェーズ 1）

チャネル 2 受信先 PE の存在する BM に到達してから，受信先 PE までの BM 内転送（フェーズ 3）

チャネル 3 余剰チャネル

チャネル 4 余剰チャネル

適応ルーティングでは，フェーズ 1 ではチャネル 1 が選ばれる．チャネル 1 が塞がっている場合，チャネル 3 かチャネル 4 がランダムに選ばれる．また，フェーズ 3 の転送ではまずチャネル 2 が選ばれ，チャネル 2 が塞がっている場合，チャネル 3 かチャネル 4 が選ばれる．

次に，BM に対する適応ルーティングの手法がデッドロックフリーであることを証明する．

定理 4.7　TESH ネットワークの固定ルーティングがデッドロックフリーであるとする．この時，フェーズ 1 およびフェーズ 3 のルーティングに Duato の fully adaptiverouting を適用した適応ルーティングはデッドロックフリーである．

証明フェーズ 2 は固定ルーティングとまったく同じ方法でルーティングを行うため，定理 4.6 によりフェーズ 2 のデッドロックフリーが保証される．

フェーズ 1 およびフェーズ 3 では，固定チャネルと余剰チャネルを用いて 3.5節で述べたようにチャネル依存グラフを作る．

まず，フェーズ 1 で用いられるチャネル全体の集合を C とし，固定チャネルを，C_1 とする．次に，3.5節で述べた直接依存関係および間接依存関係に基づき，チャネル依存グラフを作り，以下のように各固定チャネルに番号を割り振る．

$$\begin{cases} (0, n_1), & \text{y+方向のチャネル}, \\ (1, 4 - n_1), & \text{y−方向のチャネル}, \\ (2, n_0), & \text{x+方向のチャネル}, \\ (3, 4 - n_0), & \text{x−方向のチャネル}, \end{cases}$$

すると，余剰チャネルを使わない固定ルーティングについてはデッドロックフリーであることが分かる．

適応ルーティングでは，余剰チャネルが使用されるため，間接依存関係を用いて証明する．つまり，間接依存関係を含んだチャネル依存グラフ中に，上式で定めたチャネル番号が有向グラフに従って降順になる部分がないことが証明できれば良い．

チャネル番号が降順になるのは，チャネル依存グラフ中に，以下のような間接依存関係が含まれている場合である．

1. （y−方向チャネル）→（y+方向チャネル）
2. （x−方向チャネル）→（x+方向チャネル）
3. 同一方向のチャネルで，より番号の小さいチャネルへの間接依存関係
4. （x 方向チャネル）→（y 方向チャネル）

ソース PE とディスティネーション PE の位置関係によってルーティングの方向が定まるので，1. や 2. のケースのように，y−方向チャネルと y+方向チャネル，および x−方向チャネルと x+方向チャネルを同じパケットが選択することはあり得ない．また，パケットが，より番号の小さい同一方向チャネルを選択するためには，y−方向チャネルと y+方向チャネルや，x−方向チャネルと x+方向チャネルを同じパケットが選択しなければならないのでので，やはりあり得ない．

x 方向チャネルであるチャネル c_i と y 方向チャネルであるチャネル c_j の間に間接依存関係か存在すると仮定すると，間接依存関係の定義により，c_i，c_j のいずれも固定チャネルとなる．仮定より，y 方向の固定チャネルが選択できるのは y 方向のルーティングが終了して以降なので，x 方向の固定チャネル c_i と y 方向の固定チャネル c_j の間の間接依存関係はあり得ない．

したがって，フェーズ 1 およびフェーズ 3 のデッドロックフリーが保証される．

□

4.5.5　各適応ルーティングの比較検討

TESH ネットワークは，TESH(2, 3, 0) のように BM 一個あたりの BM 間リンク数が 8 本以下の場合と，TESH(2, 3, 1) のように BM 一個あたり 8 本を越える BM 間リンクを持つ場合でネットワークの構成が大きく異なる．そこで，上記二つの場合それぞれについて，前節で述べた 5 種類の適応ルーティングの適用可能性を議論する．

BM 間リンクが 8 本を越える場合

BM 間リンクが 8 本を越えるケースでは必要最大仮想チャネル数が 4 個となるため，DCS 法が適用できる．また，BM-adapt も問題なく適用可能である．CS 法も適用可能である．ただし，前述の DCS 法を適用した場合，DCS 法によりリンクが十分に有効利用されるので，その効果は薄いと思われる．LS 法については，BM 側面の BM 間リンクのルーティング経路がやや複雑なため，単純には適用できない．DDR 法の適用も可能である．

BM 間リンクが 8 本以下の場合

CS 法，LS 方および DDR 法は，問題なく適用可能である．ただし，BM 間リンクが 8 本以下のケースでは，必要仮想チャネル数が少ないことから固定ルーティングで使用されない余剰チャネルが少ないため，使用する仮想チャネルを増やさない限り BM-adapt および DCS 法による効果は期待できない．使用する仮想チャネルを増やした場合，よりリンク選択の自由度の高い DDR 法の方が有効である．このことから，BM 間リンクが 8 本以下のケースでは CS 法，LS 方および DDR 法の組み合わせが有効であると考えられる．

以上に述べた適用可能性をまとめたものを表 4.2に示す．表中の記号は，○が可能，×が不可能，△は，可能だが効果的ではないことを示している．

表 4.2: 適応ルーティングの適用可能性

BM ごとのリンク数	CS 法	LS 法	DDR 法	DCS 法	BM-adapt
8 本より多い	○	×	○	○	○
8 本以下	○	○	○	△	△

4.6 適応ルーティングによる動的通信性能の評価

4.6.1 シミュレーション条件

4.6.2 シミュレーション条件

4096PE からなる TESH ネットワーク上でシミュレーションによる動的通信性能の評価を行う．シミュレーションに用いる TESH 網は，BM 間リンクを $2^1 = 2$ 組持つ TESH(2,3,1) と，BM 間リンクを $2^0 = 1$ 組持つ TESH(2,3,0) である．シミュレーションは，TESH(2,3,0) の評価では，TESH(2,3,0) の固定ルーティング，CS 法，LS 法，および DDR 法の三つのルーティング法について行っている．また，TESH(2,3,1) の評価では，TESH(2,3,1) の固定ルーティング，DCS 法 (以降 adaptive1) および DCS 法+BM-adapt(以降 adaptive2) の三つのルーティング法について行っている．シミュレーションは，全 PE で均一に通信を行うランダム通信と，一定の確率で特定の PE（ここでは PE 0）と通信を行う hotspottransfer，近隣の PE との通信の比率が高い local transfer の 3 種類で行う．

評価基準は，平均転送時間およびスループットである．まず，Tクロックサイクルの間に特定のパケット送信要求確率でパケットを送信し，その間の全パケットの転送時間と PE が受け取ったフリットの数を記録する．次に，それらの値をもとにパケットの平均転送時間およびスループットを算出し，グラフにプロットする．以上の処理を，パケット送信要求確率を変えて複数回実行する．パケットの転送時間は，パケットの先頭がソース PE を出発してから，パケットの最後尾がディスティネーション PE に到着するまでの時間であらわされ，平均転送時間は転送時間の平均値で示される．スループットは，シミュレーション中に全 PE が受け取ったフリットの数を，PE 数および Tで割ったもので，1 クロックサイクルに 1PE が受け取るフリット数の平均値である．なお，パケット長は 16 フリットとし，$T = 20000$ としている．本実験では各チャネルが 2 フリットのバッファを保持できるものとしている．

4.6.3 uniform transfer

各 PE で均一に通信を行ったときの平均転送時間を図 4.19〜図 4.25に示す．図 4.19〜図 4.23は TESH(2,3,0) と他の結合網との比較，図 4.25は TESH(2,3,1) と他の結合網との比較である．また，横図の軸はスループット，縦軸は転送時間である．

図 4.19，図 4.20は，CS 法，LS 法，DDR 法の各手法を用いた TESH と，固定ルーティングの TESH との比較である．図の折れ線は，それぞれ TESH(2,3,0) の固定ルーティング，CS 法，LS 法，CS 法と LS 法を併用した適応ルーティングを示している．図 4.19に示すように，CS 法と LS 法は，いずれも固定ルーティングに比べて若干高いスループットとなっている．また，CS 法と LS 法を併用した適応ルーティングは，単独で用いた場合に比べて高いスループットを実現している．

図 4.20は，それぞれ TESH(2,3,0) の固定ルーティング，DDR 法のみを用いた場合，3 種類の適応ルーティング法すべてを用いた場合を示している．

図 4.20に示すように，DDR 法で最大スループットを示す部分では，DDR 法を用いない手法に比較して短い転送時間で大きなスループットを得ている．また，DDR 法とローカルな適応ルーティングの二種類の手法を組み合わせた適応ルーティングでは，他の組み合わせと比較して最大のスループットを実現している．DDR 法を用いた手法では，フリットの送信要求頻度がある程度以上大きくなると，それ以降スループットが急激に減少する．これは，ネットワーク中に送り込まれるパケットが増えるに従って固定ルーティングチャネルに殺到するパケットが増加するため，固定ルーティングチャネルが混雑することが原因となる．このような現象を防ぐためには，Throttling を導入する必要がある [31].

フリットの送信要求確率に対する実際のスループットを図 4.21に示す．図 4.21の横軸はフリット送信要求率，縦軸はスループットである．図のように，DRR 法を用いないルーティングアルゴリズムに比べて DRR 法を用いた適応ルーティングでは最大スループットが高くなっている．また，送信要求確率が 0.03〜0.04 付近で，DDR 法を用いた適応ルーティングではスループットが低下している．

図 4.22，4.23は，TESH の固定ルーティングと DDR 法のみを用いた TESH，CS 法，LS 法，DDR 法をすべて用いた TESH との比較である．うち，図 4.22は仮想チャネル数を 6 個（固定チャネル 2 個+適応チャネル 4 個）にした場合の比較，図 4.23は仮想チャネル数を 8 個（固定チャネル 2 個+適応チャネル 6 個）にした場合の比較である．図??，4.22，4.23中の各々をグラフの値を比較すると，仮想チャネル数に関わらず最大スループットでは，すべて用いた TESH，DDR 法のみを用いた TESH，TESH の固定ルーティング，メッシュの順になっている．このことから，DDR 法がネットワークの最大スループット向上に寄与していると言える．

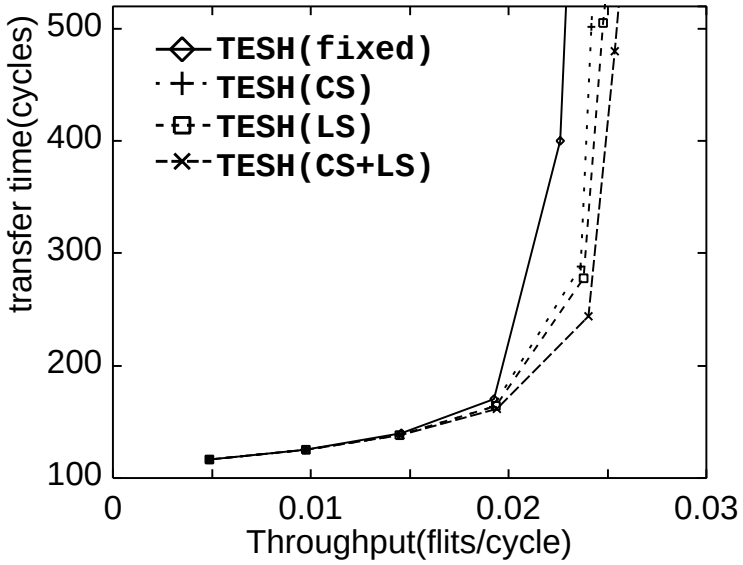


図 4.19: CS，LS 法における uniform transfer の平均転送時間

仮想チャネル数を増やすに従って TESH の各アルゴリズムとメッシュとの差が大きくなる．これは，TESH の固定ルーティングで最低限必要な 2 個のチャネルではリンクを十分効率的に利用できていないことを示している．通常，仮想チャネルを増やすに従ってリンクの利用効率が上昇するが，ある程度以上の仮想チャネル数になると頭打ちになり，それ以上仮想チャネルを増やしてもスループットは上がらなくなる．メッシュの場合 4 個のチャネルでほぼ頭打ちになるが，TESH ではさらに仮想チャネルを増やすことで，リンクの利用効率を向上させることが可能である．また，3 種類の適応ルーティング法を用いた TESH と DDR 法のみを用いた TESH を比較すると，仮想チャネルの数を増やすに従って両者の差は小さなものとなる．これは，CS 法が固定チャネルのみを使用した適応ルーティング法であることによる．仮想チャネルが増えるに従って，固定チャネルが使用される割合が少なくなるため，固定チャネルを利用した適応ルーティングの効果が少なくなる．

同じ仮想チャネル 8 個の条件で，固定ルーティングチャネルを変えた DDR 法のスループットを比較したグラフを図 4.24に示す．グラフは，それぞれ固定ルーティングチャネルを 2 個，4 個，6 個，8 個にしたものを示す．固定ルーティングチャネルを 8 個にした DDR 法では適応ルーティングチャネルを持たないため，固定ルーティングのみを行う．図 4.24 より，スループットが最大となる部分において，固定ルーティングチャネル数が 2 個，4 個，6 個，8 個の DDR 法の順に最大スループットが高くなっていることが分かる．このことから，適応ルーティングチャネルの比率を増やすことが最大スループットの向上に役立つこ

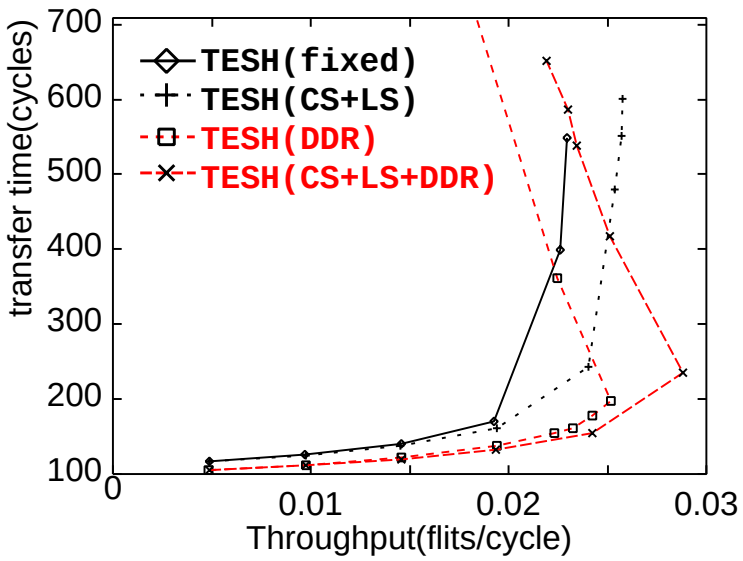


図 4.20: DDR 法における uniform transfer の平均転送時間

とが分かる．また，固定ルーティングチャネルが 8 個のグラフを除いたいずれのグラフも，最大スループットの部分からスループットが低下しているが，固定ルーティングチャネル数が 2 個の DDR 法に比べて，4 個や 6 個の DDR 法のスループットの低下はゆるやかであることが分かる．トラフィックが混雑するに従って適応ルーティングチャネルが混雑して固定ルーティングチャネルに移動するパケットの数が多くなるため，あらかじめ固定ルーティングチャネルを多めに用意することにより，ある程度混雑を解消することができるのが理由であると考えられる．

図 4.25は，TESH(2,3,1) における固定ルーティングと適応ルーティングの比較である．図 4.25中の，fixed は TESH の固定ルーティング，adaptive1 は DCS 法のみを用いた適応ルーティング，adaptive2 は DCS 法と BM-adaptive を用いた適応ルーティングである．図 4.25 で分かるように，DCS 法を用いた適応ルーティングは，固定ルーティングに比べて最大スループットが高くなっている．このことから，DCS 法は BM 間リンクを効率的に使用していると考えることができる．一方，BM-adaptive を用いた適応ルーティングは，用いないルーティング法に比べて若干のスループット向上を実現できる．ただし，その差は小さなものであることから，BM 内の適応ルーティングはネットワーク性能の向上に，それほど多く寄与するものではないことが分かる．

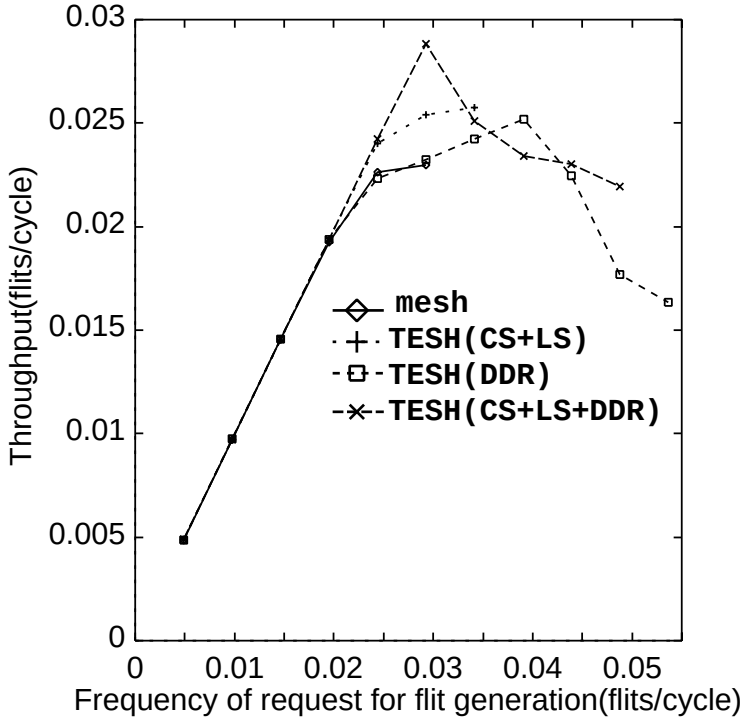


図 4.21: uniform transfer のスループット

4.6.4 hotspot transfer

hotspot transfer は，特定の PE と PE_h通信を行うホットスポットパケットを P_hの確率で含む通信パターンである．本実験では，PE_hを，ソース PE アドレス $\mathrm{PE}(n_5, n_4)(n_3, n_2)(n_1, n_0)$ に対して $\mathrm{PE}(n_5, n_4)(0, 0)(0, 0)$ とし，$P_h = 0.1$ としている．

hotspot transfer のシミュレーション結果を図 4.26，図 4.27に示す．図の横軸がスループット，縦軸が平均転送時間である．uniform transfer と違い，これらの図ではすべての手法について，ある程度のパケット送信要求確率以上で若干スループットが低下している．これは，パケット送信要求確率が上昇するに従ってホットスポットパケットが増えるためにネットワークが混雑することが原因である．特定の PE に集中するホットスポットパケットパケットは，その他のパケットに比べてネットワーク中の滞在時間が長くなるため，パケット送信要求確率が上がるにしたがってホットスポットパケットがネットワーク中の多くのバッファを占有することになる．

図 4.26の折れ線は，それぞれ TESH(2,3,0) の固定ルーティング，CS 法，LS 法，CS 法と LS 法を併用した適応ルーティングを示している．固定ルーティングと CS 法ではスループットに大きな差が見られなかったが，LS 法では固定ルーティングと比較して若干のス

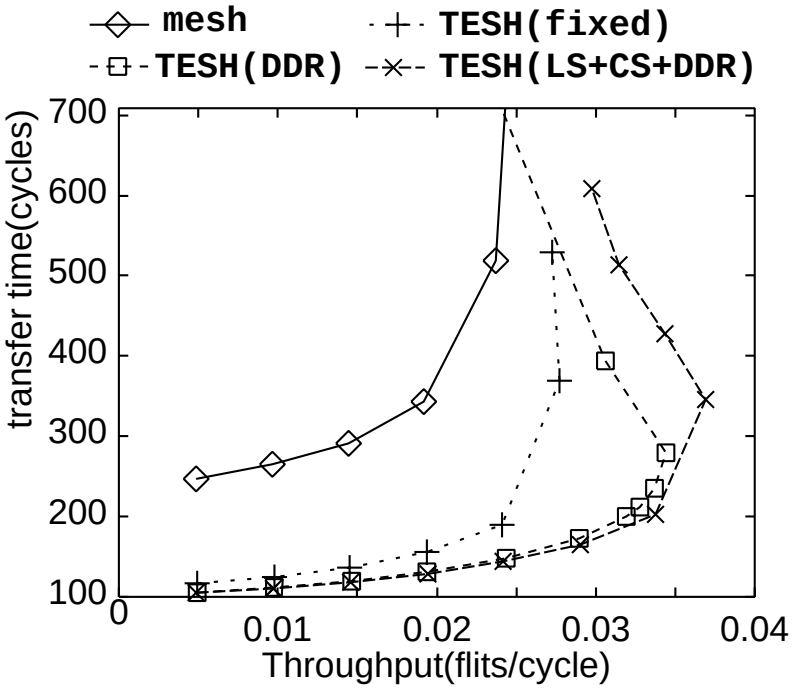


図 4.22: TESH(2,3,0) における uniform transfer の平均転送時間（チャネル数 6 の DDR 法による比較）

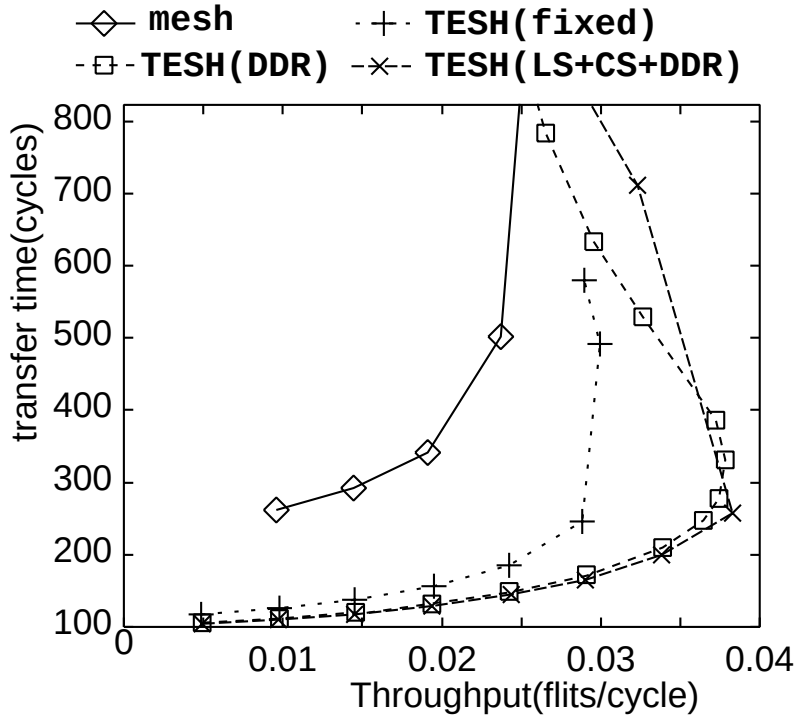


図 4.23: TESH(2,3,0) における uniform transfer の平均転送時間（チャネル数 8 の DDR 法による比較）

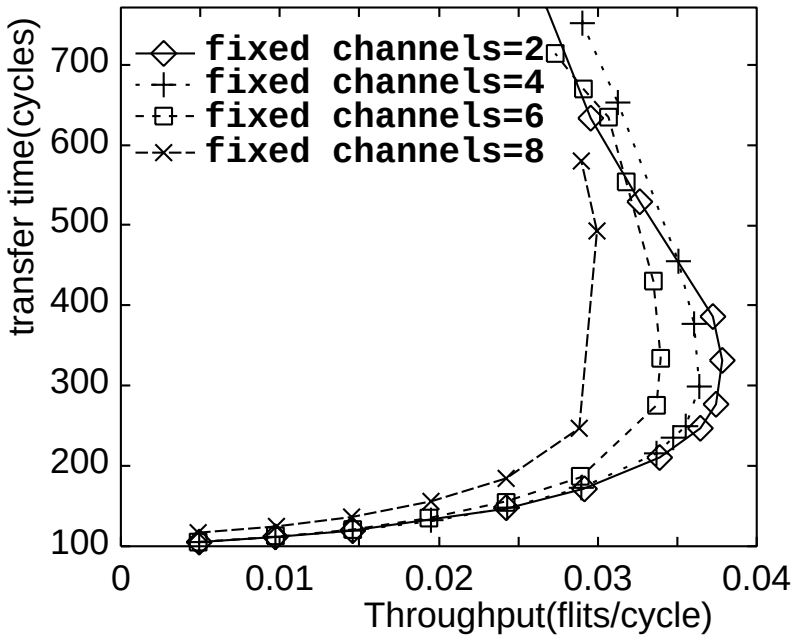


図 4.24: TESH(2,3,0) における uniform transfer の平均転送時間（DDR 法の固定チャネル数による比較）

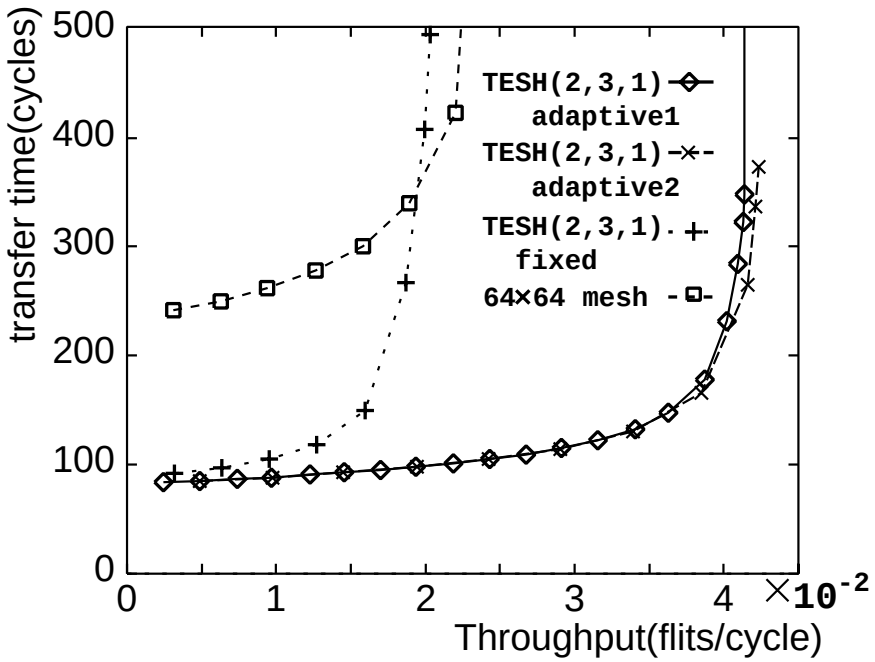


図 4.25: TESH(2,3,1) におけるランダム通信の平均転送時間

ループット向上が見られた．LS 法では，CS 法や固定ルーティングと異なり混雑している BM 間リンクを避けて通ることができるため，hotspot transfer のようにトラフィックが混雑するパターンでは効果を発揮する．

図 4.27の折れ線は，それぞれ TESH(2,3,0) の固定ルーティング，DDR 法，CS 法+LS 法に加えて，DDR 法を併用した適応ルーティングを示している．

DDR 法のようなグローバルな適応ルーティングでは，ローカルな適応ルーティングに比べて BM 間リンクの選択の幅が増えるため，hotspot transfer の最大スループットがさらに上昇する．

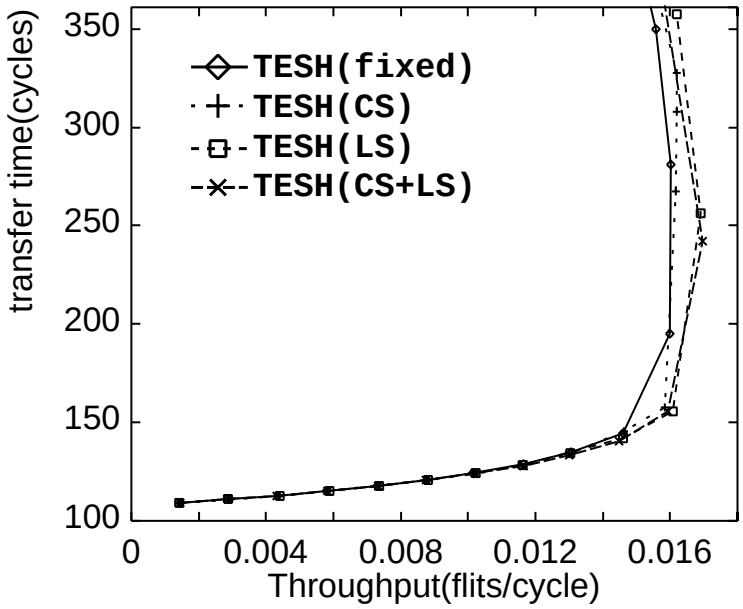


図 4.26: CS，LS 法における hotspot transfer の平均転送時間

4.6.5 Non-Uniform transfer

Non-Uniform transfer は，送信先 PE を送信元 PE と同一の BM から選ぶ BM 内通信をある一定の確率で実行するランダム通信のパターンである．このような通信パターンは，通信に局所性を持つアプリケーションをマッピングした場合などに起りうる [14].

メッシュの Non-Uniform transfer では，64×64 メッシュを 4×4 の細かいメッシュに分割して BM と見なし，TESH の Non-Uniform transfer と同様に一定の確率で BM 内通信を行うものとした．本実験での BM 内通信の比率は 80 %としている．

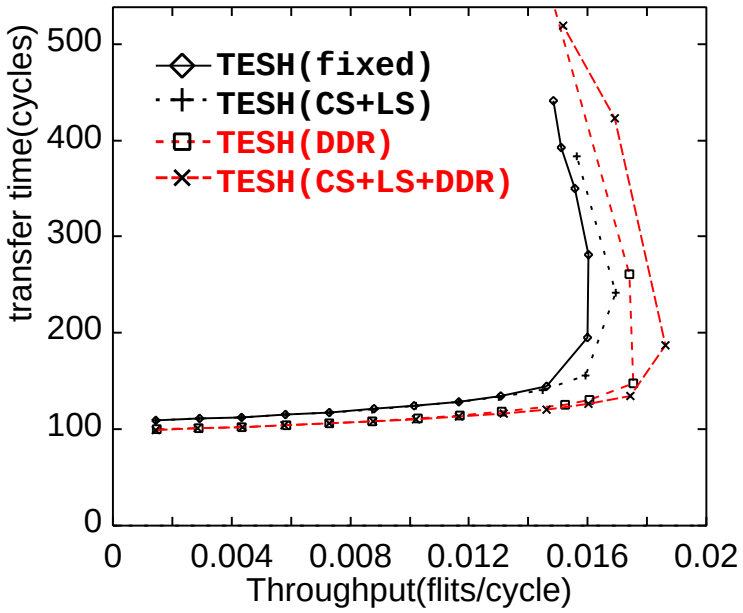


図 4.27: DDR 法における hotspot transfer の平均転送時間

平均転送時間を図 4.28に示す．図 4.28の横軸はスループット，縦軸は転送時間である．uniform transfer の実験と同様，低スループットの部分で TESH(2,3,1) の平均転送時間はメッシュよりも短くなる．ただし，両者の差は uniform transfer ほど大きくならない．これは，BM 内通信で両ネットワークの転送時間に差が生じないためである．

固定ルーティングの TESH は他に比べて低スループットの部分でグラフが飽和している．これは，固定ルーティングでは複数のチャネルを選択できないという理由による．一方のメッシュでは 4 個のチャネルを使用してパケットのブロッキングを抑えているため，より高いスループットが得られている．TESH の他の手法でも複数のチャネルを使用できるため，やはり固定ルーティングに比べて高いスループットとなる．

この結果から，固定ルーティングの TESH ネットワークは，同サイズのメッシュに比べて局所通信が苦手である一方，余剰チャネルを有効利用することによりこの弱点を緩和できることが分かる．

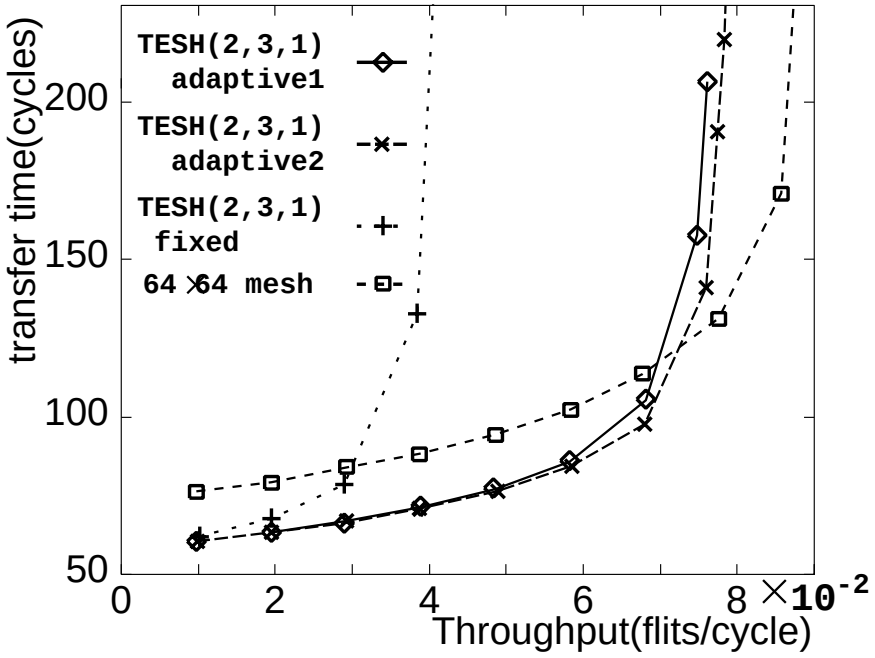


図 4.28: Non-Uniform transfer の平均転送時間

4.7 アルゴリズムのマッピング

4.7.1 FFT アルゴリズム

高速フーリエ変換 (FFT)[32] は，画像処理や信号処理等において幅広い応用がなされているアルゴリズムである．そのため，並列計算機上における FFT の高速化に関する研究が数多くなされている [33][34][35][36]．これらはいずれも，一般的な格子網で用いることのできる並列化手法である．

FFT アルゴリズムには，データを時間領域から周波数領域に変換する 1 次元 FFT の他に，画像データのような 2 次元のデータを，行方向と列方向それぞれについて周波数領域に変換する 2 次元 FFT がある．FFT は，離散フーリエ変換 (DFT) の計算量を削減するために考えられたアルゴリズムである．$N = 2^m$点の入力データ列を，$x(n)(0 \leq n \leq N-1)$ とする．1 次元 DFT は，以下のような式で示される．

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn} \tag{4.13}$$

ここで，$W_N = \exp(-2\pi j/N)$ である．FFT アルゴリズムでは，この計算と同等の処理を

m 段階のバタフライ演算の繰り返しによって行う．各段階の計算は以下の通りである．

$$\begin{aligned} X_i(k) &= X_i(u_0,\cdots,u_{i-2},u_{i-1},k_{m-i-1},\cdots,k_0) \\ &= X_{i-1}(u_0,\cdots,u_{i-2},0,k_{m-i-1},\cdots,k_0) \\ &+ X_{i-1}(u_0,\cdots,u_{i-2},1,k_{m-i-1},\cdots,k_0)W \end{aligned} \tag{4.14}$$

ここで，$W=(-1)^{u_{i-1}}W_N^{u_{i-2}2^{m-2}+\cdots+u_0 2^{m-k}}$である．

(4.15) 式のようにすると，i 回目の FFT データのやり取りは，2^{m-i}離れたデータと行うことになる．

2 次元 FFT のアルゴリズムには，行方向と列方向それぞれについて 1 次元 FFT を行う方法と，4 入力のバタフライ演算を行い，行方向と列方向を一度に計算する方法の 2 通りの方法がある．このうち，後者の方法は前者に比べて計算量を削減することができるので，後者の方法を採用する．$N=N'\times N'(N'=2^{m'})$ 点の入力のデータ列を $x(r;s)(0\le r\le N'-1,\ 0\le s\le N'-1)$ とおくと，2 次元 DFT は，以下のような式で示される．

$$X(k;l)=\sum_{r=0}^{N'-1}\sum_{s=0}^{N'-1}x(r;s)W_{N'}^{kr}W_{N'}^{ls} \tag{4.15}$$

2 次元 FFT アルゴリズムでは，この計算と同等の処理を m'段階のバタフライ演算の繰り返しによって行う．各段階の計算は以下の通りである．

$$\begin{aligned} X_i(k;l) &= X_i(u_0,\cdots,u_{i-2},u_{i-1},k_{m'-i-1},\cdots,k_0) \\ &= X_{i-1}(u_0,\cdots,u_{i-2},0,k_{m'-i-1},\cdots,k_0; \\ &\quad v_0,\cdots,v_{i-2},0,l_{m'-i-1},\cdots,l_0) \\ &+ X_{i-1}(u_0,\cdots,u_{i-2},1,k_{m'-i-1},\cdots,k_0; \\ &\quad v_0,\cdots,v_{i-2},0,l_{m'-i-1},\cdots,l_0)W_a \\ &+ X_{i-1}(u_0,\cdots,u_{i-2},0,k_{m'-i-1},\cdots,k_0; \\ &\quad v_0,\cdots,v_{i-2},1,l_{m'-i-1},\cdots,l_0)W_b \\ &+ X_{i-1}(u_0,\cdots,u_{i-2},1,k_{m'-i-1},\cdots,k_0; \\ &\quad v_0,\cdots,v_{i-2},1,l_{m'-i-1}, \\ &\quad \cdots,l_0)W_aW_b \end{aligned} \tag{4.16}$$

ここで，$W_a=(-1)^{u_{i-1}}W_{N'}^{u_{i-2}2^{m'-2}+\cdots+u_0 2^{m'-i}}$，$W_b=(-1)^{v_{i-1}}W_{N'}^{v_{i-2}2^{m'-2}+\cdots+v_0 2^{m'-i}}$である．

4.7.2 FFT データのマッピング

メッシュ上へのマッピング

2 次元メッシュの PE 数を $P = P_x \times P_y$とすると，PE 番号は x 方向，y 方向のアドレスで表記して PE$(y;x)$ で示すことができる．$P_x = P_y$としてすると，この時の PE 数は $P = 2^r$である．これを二進数で表現すると，$y = b_{r-1} \cdots b_{r/2}$，$x = b_{r/2-1} \cdots b_0$として PE$(b_{r-1}, \cdots, b_{r/2}; b_{r/2-1}, \cdots, b_0)$ となる．

また，対象データの数を $N = 2^p$とすると，FFT の対象となるデータは，2 進数表記で $X_0(n_{p-1}, n_{p-2}, \cdots, n_1, n_0)$ となる．

通常，メッシュやトーラスなどの格子網上に FFT データをマッピングする場合，Stage-by-stage 法または Multi-stage 法を用いる．

Stage-by-stage 法は，データ $X_i(n_{p-1}, n_{p-2}, \cdots, n_1, n_0)$ を PE$(n_{p-1}, n_{p-2}, \cdots, n_{p-r})$ または PE$(n_{r-1}, n_{r-2}, \cdots, n_1, n_0)$ に割り付ける方法である．Multi-stage 法は，1PE に割付けられた 2^{p-r}個のデータを用いて，$p-r$段のバタフライ演算を PE 間通信なしで行い，バタフライ演算終了後にデータを再配置するという処理を $p/p-r$回繰り返す方法である．この場合，$2(p-r)j$段のバタフライ演算終了後の j回目のデータ再配置で，データ

$$X_{(p-r)j}(u_0, \cdots, u_{p-(p-r)j-1}, k_{(p-r)j-1}, \cdots, k_0)$$

を，データ

$$X_{(p-r)(j-1)}(u_0, \cdots, u_{p-(p-r)(j-1)-1}, k_{(p-r)(j-1)-1}, \cdots, k_0)$$

が配置されている PE に割り当てる．その結果，上記のデータは
PE$(u_0, \cdots, u_{p-(p-r)j-1}, k_{(p-r)(j-1)-1}, \cdots, k_0)$
に割り当てられることになる．

表 4.3に，各手法の通信回数および通信を行う FFT データの総数を示す．表 4.3に示すように，Multi-stage 法は，Stage-by-stage 法に比べて通信の回数そのものは多くなるが，一回の通信で転送するデータ数が少なくなる．したがって，両者のうちどちらが効率的かは，使用する計算機環境や 1PE ごとのデータ数によって異なる．

TESH 上へのデータ割付け

TESH(m,L,q) 上での入力データの数を $N = 4^p$，PE 数を $P = 2^{2mL}$とすると，TESH の PE 番号は 2 進数表記で PE$(n_{4L-1}, n_{4L-2}, \cdots, n_1, n_0)$ と書ける．Stage-by-stage 法ではデータ X$_0(n_{p-1}, n_{p-2}, \cdots, n_1, n_0)$ を PE$(n_{4L-1}, n_{4L-2}, \cdots, n_1, n_0)$ に割り付ける．すると，データ

表 4.3: 各手法の通信回数およびデータ総数

	N/P	2	4	8	16
stage-by-stage	通信回数	12	12	12	12
	データ数	24	48	96	192
Multi-stage	通信回数	12	18	32	48
	データ数	12	18	32	48

$\mathrm{X}_0(n_{p-1}, n_{p-2}, \cdots, n_1, n_0)$ はレベル $i-1$ サブネットワーク $(n_{4L-1}, n_{4L-2}, \cdots, n_{4i-3}, n_{4i-4})$ に割り付けられるため，サブネットワーク間の通信パターンは従来の二次元格子網に対する FFT の割り当て法による通信パターン [33] と同じものになる [37]．この時，$4(L-i)+1$ 回目から $4(L-i)+4$ 回目までのサブネットワーク間の通信パターンは図 4.29のようになる．なお，図中の□はレベル $i-1$ のサブネットワークを示し，図全体はレベル i のサブネットワークを示す.

また Multi-stage 法では，j回目のデータ再配置で，データ

$$X_{(p-r)j}(u_0, \cdots, u_{p-(p-r)j-1}, k_{(p-r)j-1}, \cdots, k_0)$$

を，データ

$$X_{(p-r)(j-1)}(u_0, \cdots, u_{p-(p-r)(j-1)-1}, k_{(p-r)(j-1)-1}, \cdots, k_0) \tag{4.17}$$

が配置されている PE に割り当てる．その結果，上記のデータは
$\mathrm{PE}(u_0, \cdots, u_{p-(p-r)j-1}, k_{(p-r)(j-1)-1}, \cdots, k_0)$
に割り当てられることになる．このようにすると，初期のステージでは近接 PE との通信が多くなるため，TESH 上で効率的にメッセージ通信を行うことができる．

割付け順序の入れ替え

本論文では，$m=2$ の TESH に関して，上記の Stage-by-stage 法のデータの割付け順序を入れ替え，メッセージ間の衝突を抑える方法を提案する．ここで，データ X_0，PE の番号をそれぞれ 4 進数で表記して，
$\mathrm{X}_0(n_{p/2-1}, n_{p/2-2}, \cdots, n_3, n_2, n_1, n_0)$，$\mathrm{PE}(n_{L-1}, n_{L-2}, \cdots, n_1, n_0)$
と表記することにする．この時，
$\mathrm{X}_0(n'_{p/2-1}, n'_{p/2-2}, \cdots, n'_3, n'_2, n_1, n_0)$ を $\mathrm{PE}(n_{L-1}, n_{L-2}, \cdots, n_1, n_0)$ に割り付ける．ここで，$n'_i (p/2-1 \geq i \geq 2)$ は，以下の式に従う．

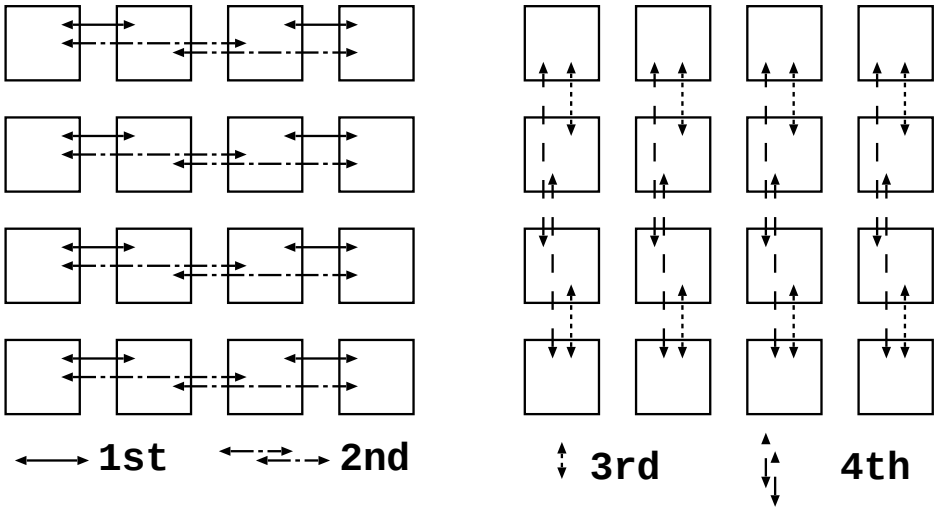


図 4.29: FFT の通信パターン（Stage-by-stage 法）

$$n_i' = \begin{cases} n_i, & (n_i = 0 \ \text{ or } \ n_i = 1), \\ 3, & (n_i = 2), \\ 2, & (n_i = 3), \end{cases} \tag{4.18}$$

以上のようにデータを割り付けた時の，TESH の BM 間通信の例を図 4.30 に示す．なお，図中の□は BM であり，図中の通信パターンは，BM 間通信に限ったものである．従来は図 4.29のように通信を行っていたものを，図 4.30のような通信パターンになるようにデータを入れ替えている．

4.7.3 シミュレーションによる通信性能評価

シミュレーション条件

4096PE からなる TESH ネットワーク上でシミュレーションによる評価を行う．評価方法は，シミュレータ上で FFT を実行した時の実行時間を比較する．本実験では，バタフライ演算に要する時間を 600 サイクル，通信の前後処理に要する時間を 300 サイクルとした．また，一つの FFT データにつき 32 フリットの長さを持つものとして実験を行った．ルーティング法は，4.3節で述べた固定ルーティングを用いる．そのため，TESH(2,3,1) の仮想チャネル数は 4，TESH(2,3,0) の仮想チャネル数は 2 である．また，メッシュの仮想チャネル数は，TESH(2,3,1) との比較では 4，TESH(2,3,0) との比較では 2 としている．シミュ

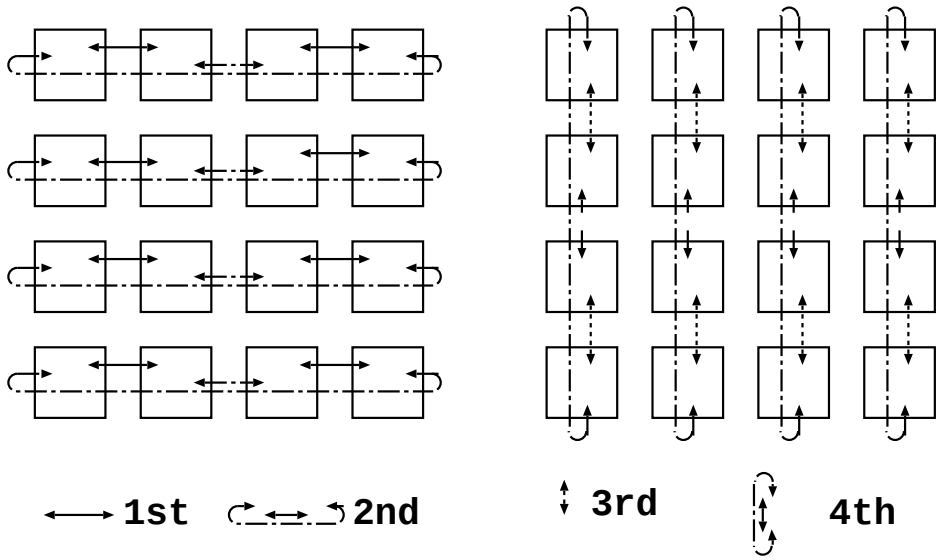


図 4.30: FFT の通信パターン（割付け順序を入れ替えた場合）

レーションに用いたアルゴリズムは，メッシュでは Multi-stage 法と Stage-by-stage 法の二種類，TESH(2,3,1) および TESH(2,3,0) では Multi-stage 法と Stage-by-stage 法，データの入替えを行った Stage-by-stage 法の 3 種類である．

実行時間

TESH およびメッシュ上で FFT をシミュレートしたときの，データ数に対する実行時間を図 4.31，図 4.32に示す．図 4.31は TESH(2,3,1) と 4 チャネルを有するメッシュとの比較，図 4.32は TESH(2,3,0) と 2 チャネルを有するメッシュとの比較である．

図 4.31に示すように，メッシュ上に比べて TESH(2,3,1) 上における実行時間は改善されている．これは，TESH がメッシュに比べて，遠方の PE への通信ホップ数が短く，メッセージの衝突が少なくなるためである [37]．メッシュ上で二つの手法を比較すると，Muiti-stage 法の方が短い時間で実行を終了している．一方，TESH(2,3,1) 上で二つの手法を比較すると，1PE あたりのデータ数が少ない場合は Muiti-stage 法の方が若干短い時間で実行を終了しているが，1PE あたりのデータ数が増えるとこの関係は逆転する．メッシュ上で Muiti-stage 法の方が実行時間が短くなるのは，Muiti-stage 法では通信するデータサイズが小さいため通信そのものに要する時間が短いためである．一方，TESH(2,3,1) 上で 1PE あたりのデータ数が多い場合に逆の結果となるのは，Muiti-stage 法ではデータ数が増えるにしたがって通信回数が増えるため，通信のためのセットアップ時間が実行時間に影響を及ぼすためで

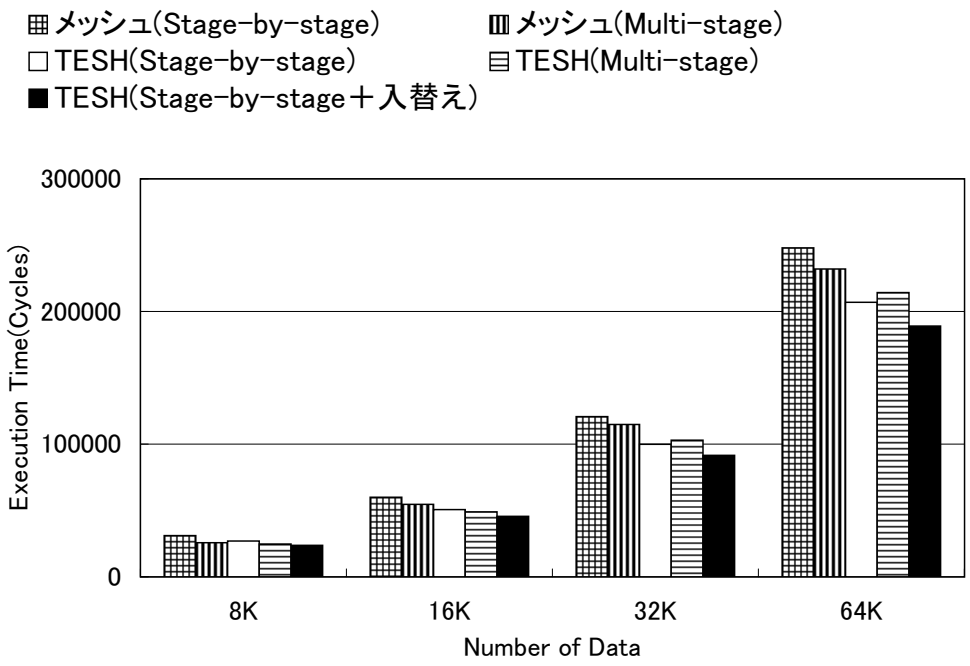


図 4.31: TESH(2,3,1) 上における FFT の実行時間

ある．また，割付け順序の入替えを行った Stage-by-stage 法では，単純な Stage-by-stage 法に比べて実行時間が改善されている．これは，割付法の改良により，メッセージ間のブロッキングによる遅延が緩和されたことによるものである．

また，図 4.32に示すように，TESH(2,3,0) とメッシュを比較すると，同じ手法同士では実行時間にほとんど差が出ない．このようにメッシュに対する TESH の優位性が得られないのは，TESH(2,3,1) に比べて TESH(2,3,0) は BM 間リンクが少なく，メッセージの衝突が多くなるため，トータルでより多くの実行時間を要するためである．このような場合でも，Stage-by-stage 法で割付け順序を入れ替えを行うとメッシュや他の手法に比べて実行時間が短くなっている．

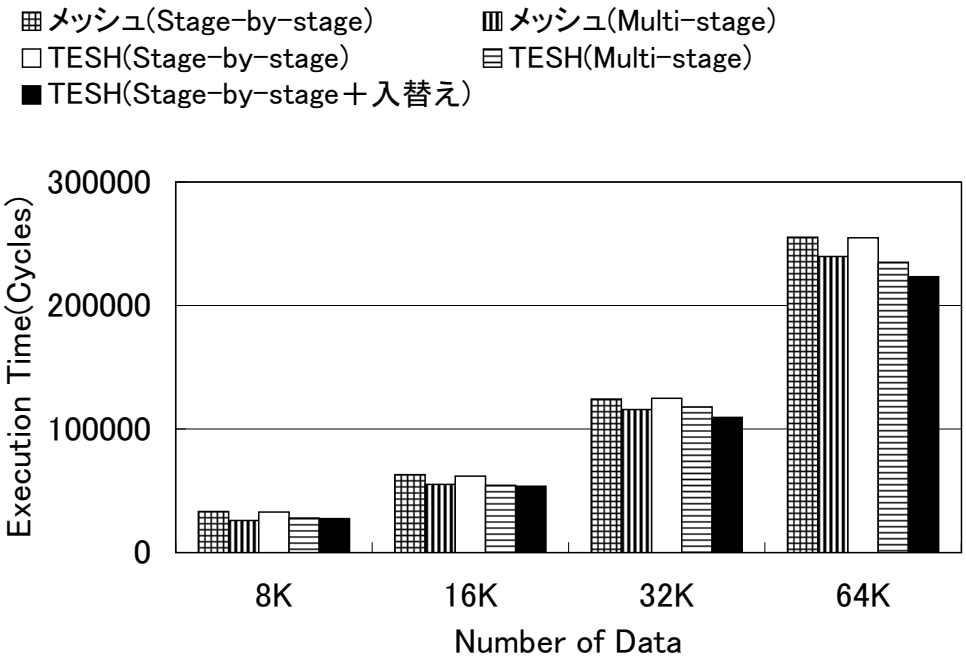


図 4.32: TESH(2,3,0) 上における FFT の実行時間

4.7.4 速度向上率

同一アルゴリズムについての、TESH 上におけるメッシュ上での実行に対する速度向上率を図 4.33，図 4.34に示す．Multi-stage アルゴリズム、Staga-by-stage アルゴリズム、Staga-by-stage+入替のアルゴリズムの速度向上率 S_{multi}、S_{stage}、S_{re-map}、は、それぞれ下式により示される。

$$S_{multi} = T_{mesh,multi}/T_{TESH,multi} \tag{4.19}$$

$$S_{stage} = T_{mesh,stage}/T_{TESH,stage} \tag{4.20}$$

$$S_{re-map} = T_{mesh,multi}/T_{TESH,re-map} \tag{4.21}$$

ここで、$T_{x,multi}$、$T_{x,stage}$、$T_{x,re-map}$は、それぞれ Multi-stage アルゴリズム、Staga-by-stage アルゴリズム、Staga-by-stage+入替のアルゴリズムの、ネットワーク x 上における実行時間を示している。なお、x は mesh か TESH のいずれかである。図 4.33は TESH(2,3,1) 上における速度向上率，図 4.34は TESH(2,3,0) 上における速度向上率であり，TESH 上の Multi-stage 法と，割付け順序の入れ替えを行わなかった Stage-by-stage 法，行った Stage-by-stage 法の，それぞれメッシュ上での実行時間に対する速度向上比である．

TESH(2,3,0) では，割付け順序の入れ替えを行わなかった時には，TESH はメッシュに対してほとんど速度向上が見られない．それに対して割付け順序の入れ替えを行った時の速度向上比はほぼ 1.2 倍程度に伸びている．TESH(2,3,1) では，割付け順序の入れ替えを行わなかった場合，Multi-stage 法では 1.1 倍，Stage-by-stage 法では 1.2 倍程度の速度向上が見られた．対して，割付け順序の入れ替えを行った Stage-by-stage 法では約 1.3 倍程度の速度向上が見られた．このことから，割付け順序の入れ替えは TESH 上で FFT を実行するにあたって有効な手法であることが分かる．

4.8 まとめ

本章では，階層型相互結合網 TESH におけるデッドロックフリーなルーティング法を提案した．

第一に，BM 一個あたりの BM 間リンク数が 8 本を越える場合と越えない場合の 2 つの場合について，適切なリンク配置の方法を提案し，それぞれについてデッドロックを回避するために必要な仮想チャネル数を導出した．その結果，BM 間リンク数が 8 本を越える TESH に適したリンクの配置法として一列配置を提案した．また，BM 間リンク数が 8 本を越え

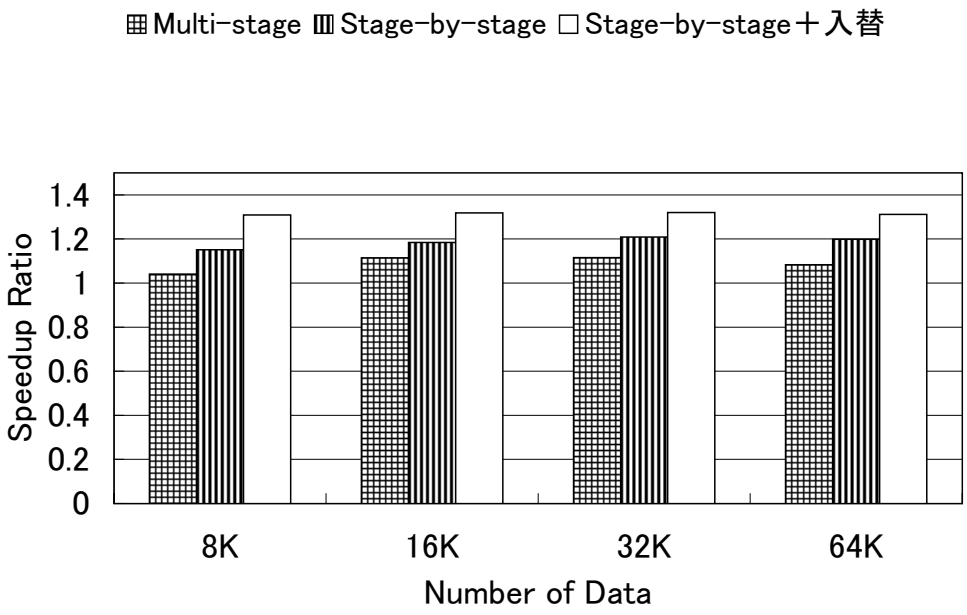


図 4.33: TESH(2,3,1) 上における速度向上率

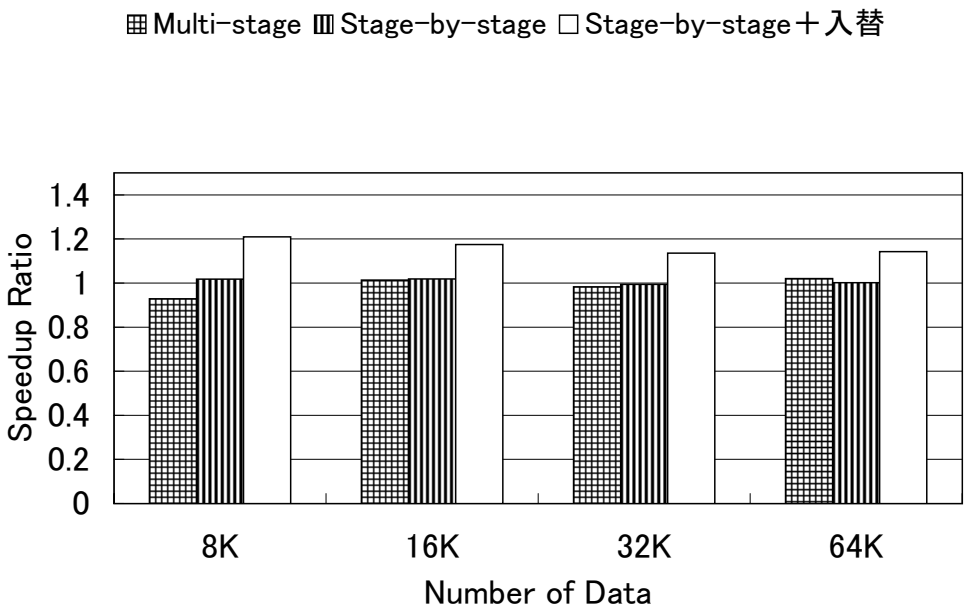


図 4.34: TESH(2,3,0) 上における速度向上率

る TESH の必要仮想チャネル数が 4 本，越えない TESH の必要仮想チャネル数が 2 本であることを証明した．これらは，TESH そのもののサイズには依存しない値である．さらに，シミュレーションにより動的通信性能についての評価を行った．その結果，TESH(2,3,1) はメッシュに比べて良好な通信性能を有することを示した．また，TESH(2,3,0) はチャネルバッファ数が多い場合，チャネルバッファ数が少ないものに比べて大きな性能向上が得られることを示したまた，チャネル数を増やすことによって TESH のスループットが大きく向上することを示し，同サイズのメッシュに比べて良好な通信性能を有することを示した．

第二に，TESH におけるルーティング性能の向上のための適応ルーティングの手法として，cs 法，ls 法，DDR 法を提案し，提案手法がデッドロック・フリーを保証することを証明した．さらに，シミュレーションにより固定ルーティングと提案手法の比較検討を行った．その結果，いくつかの通信パターンについて固定ルーティングに比べて低い平均通信時間および高いスループットが実現できることが明らかになった．

第三に，TESH 上における 1 次元 FFT のトーラス結合を利用した新たなマッピング法を提案した．本マッピング法では，単純なマッピング法に比べて通信距離が短くなり，かつメッセージ間の衝突が減るため，実行時間を短縮することができた．シミュレーションによって実験を行った結果，4096 個の PE を持つ 3 階層の TESH で，同サイズのメッシュに比べて約 1.3 倍程度実行時間を短縮することができた．

第 5 章

階層型相互結合網 H3D トーラス

5.1 はじめに

TESH は，2 次元メッシュと 2 次元トーラスの階層構造により，直径やウェハー間結線数を小さくすることができる結合網である．しかしながら，TESH の直径が大きくなり PE 数が増えるに従って，縦方向配線の制限が厳しくなる．そこで，提案されたのが，階層型 3 次元トーラス（Hierarchical 3-Dimensional Torus, H3D トーラス）である．H3D トーラスは，TESH よりもさらに大規模な並列計算機システムのための相互結合網として提案された．H3D トーラスは，下位階層に 3 次元メッシュ，上位階層に 3 次元トーラスを使用することにより，TESH に比べて多数のノードを持つことが可能となり，大規模な並列計算システムの実装が容易に可能となる．H3D トーラスは，BM や上位レベルネットワークに 3 次元メッシュ／トーラスを使用しているので，TESH に対して提案したルーティング法でデッドロックを回避することが可能である．

本章では，H3D トーラス上でデッドロックを回避するために必要な仮想チャネル数を導出する．さらに，固定ルーティングを用いた H3D トーラスネットワークの動的通信性能を評価する．

5.2 階層型相互結合網 H3D トーラス

5.2.1 ネットワーク構成

H3D トーラスは，レベル 1 ネットワークである基本モジュール (BM) として 3 次元メッシュ$(m \times m \times m)$ を用いた階層型相互結合網である．これらの BM は，レベル 2 ネット

ワークである 3 次元トーラス $(n \times n \times n)$ によって結合される．

$m = 4$ の場合について考える．この場合，BM の PE 番号は，4 進数によって以下のようにアドレス付けされる．数字は，最下位桁からそれぞれ，x 方向，y方向，z方向のアドレスをあらわす．

$$A = (a_z)(a_y)(a_x), \quad (a_z, a_y, a_x = 0, 1, 2, 3) \tag{5.1}$$

BM 内においてこのようにしてアドレス付けされた各 PE を，以後 PE(a_z, a_y, a_x) と表記する．

上位レベルネットワークを構成するためのリンクをここでは BM 間リンクとよぶ．3 階層の H3D トーラスは，$n \times n \times n$ 個の 2 階層 H3D トーラスをさらに 3 次元トーラスで結合して作られる．

図 5.1に，$4 \times 4 \times 4$ メッシュにより構成された BM の構成例を示す．BM 中の，$a_y = 0, 3$，$a_x = 0, 3$ である各PE は，BM 間リンクとの結合に用いられる．この例では，PE$(a_z, 0, 0)$ はレベル 2 ネットワーク，PE$(a_z, 0, 3)$ はレベル 3 ネットワーク，PE$(a_z, 3, 3)$ はレベル 4 ネットワーク，PE$(a_z, 3, 0)$ はレベル 5 ネットワークと，それぞれ結合する．また，それら各々の，$a_z = 0, 1, 2$ である 3 つの PE は，それぞれ z 方向，y 方向，x 方向の 2 本のリンクと結合されている．

$n = 4$ とした場合，レベル L^iネットワーク中の各 PE は，BM の場合と同様，4 進数によって以下のようにアドレス付けされる．数字は，最下位桁からそれぞれ，x 方向，y方向，z方向のアドレスをあらわす．

$$A^{L_i} = (a_z^{L_i})(a_y^{L_i})(a_x^{L_i}), \quad (a_z^{L_i}, a_y^{L_i}, a_x^{L_i} = 0, 1, 2, 3) \tag{5.2}$$

以上のような場合のレベル 2H3D トーラスの構成は図 5.2のようになる．

なお，各レベルにつき複数組のリンクを設けることも可能である．その場合，各レベルにつき 2^q組（つまり 4×2^q本）のリンクを設けることとなる．

5.2.2 ルーティングアルゴリズム

BM 内の各 PE や，上位レベルの各サブネットは，4 進数によってアドレス付けされている．レベル L H3D トーラスの PE アドレスは，以下のように与えられる．

$$A = \prod_{L_i=1}^{L} (a_z^{L_i}\ a_y^{L_i}\ a_x^{L_i}) \tag{5.3}$$

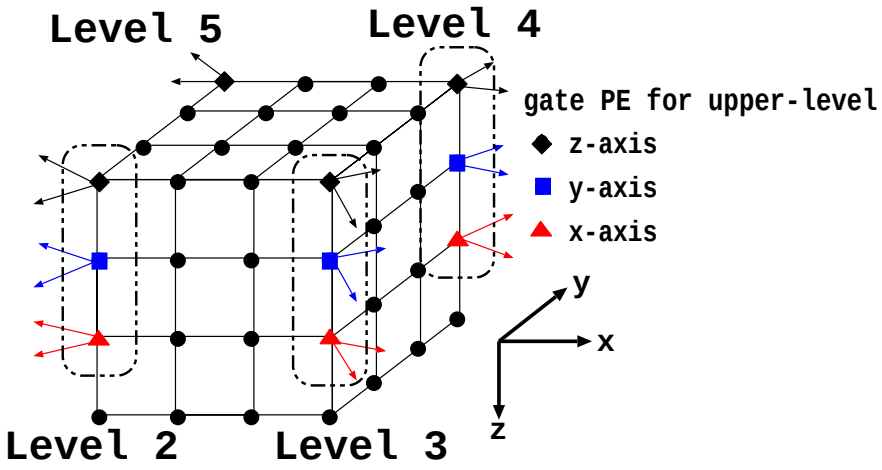


図 5.1: H3D トーラスの BM 構成

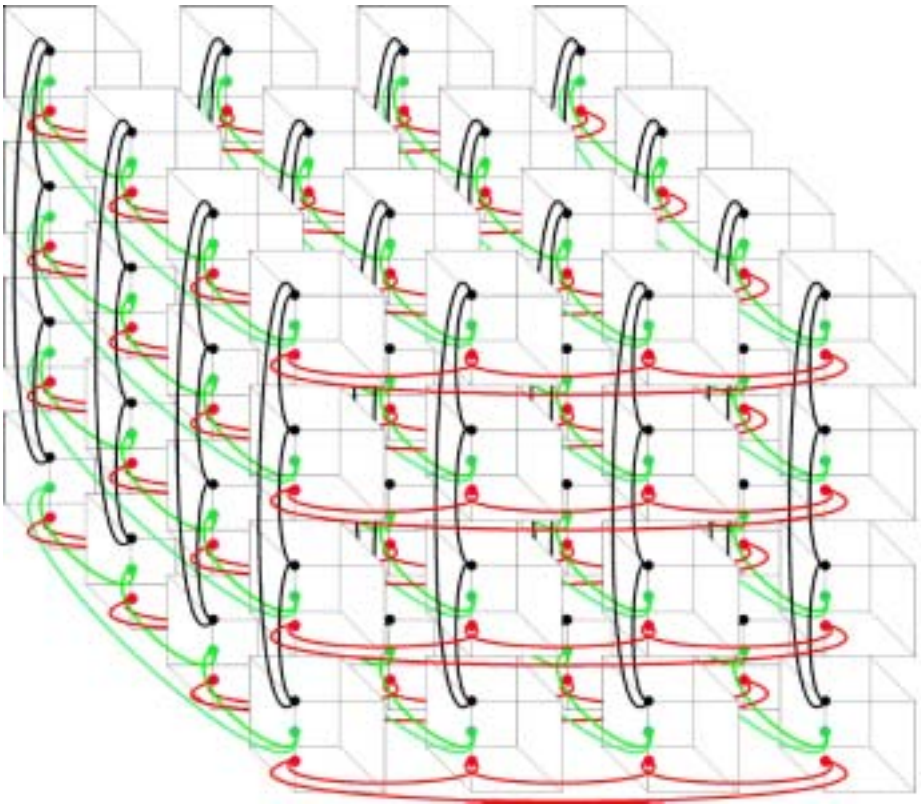


図 5.2: 2 レベル H3D トーラスの構成

H3D トーラスのルーティングは，TESH と同様，上位レベルから下位レベルの順に行われる．最初に，最上位レベルの転送により，目的のサブネットワークへのルーティングを行い，次に，サブネットワーク中のルーティングを行う．このような手順を，最終的な目的地に到達するまで繰り返す．各レベルのルーティングの順序は，z 方向，y 方向，x 方向の順序である．

H3D トーラスのルーティングは，送信元 PE アドレスと受信先 PE アドレスによって決められる．ここで，送信元 PE のアドレス s を $s_{n-1}s_{n-2}...s_1s_0$，受信先 PE のアドレス d を $d_{n-1}d_{n-2}...d_1d_0$ とする．PE($s_{n-1}s_{n-2}...s_1s_0$) から PE($d_{n-1}d_{n-2}...d_1d_0$) へのルーティングのため，タグ t_iを，以下のように決定する．

$$t_i = d_i - s_i,\ (i = 0, \cdots, n-1) \tag{5.4}$$

これらを用いて，H3D トーラスのルーティングアルゴリズムは図 5.3のように決められる．H3D トーラスにおける，$PE_{(123)(211)}$と $PE_{(333)(111)}$の間のルーティングの例を示すと，以下のようになる．まず，z 方向のレベル 2 リンクのゲート PE にあたる $PE_{(123)(000)}$へ移動する．ついで，$PE_{(123)(000)}$のパケットを z 方向のレベル 2 リンクに沿って $PE_{(323)(000)}$まで移動させる．同様の方法で y 方向のレベル 2 リンクのゲート PE にあたる $PE_{(323)(100)}$まで移動させ，y 方向のレベル 2 リンクに沿って $PE_{(333)(100)}$まで移動する．最後に，BM 内転送を行い，転送を終える．

5.3 デッドロック・フリー

5.3.1 H3D トーラスへの適用

H3D トーラスのデッドロック回避に必要な仮想チャネルの数を導出するために，転送の流れをいくつかのフェーズに分割する．

一般に，L レベルの TESH のルーティングは以下の 3 フェーズに分けることができる．

フェーズ 1 送信元PE から，xy 平面と平行な側面を除く BM の側面 ($a_y^1 = 0$ or $3, a_x^1 = 0$ or 3 のいずれかを満たす PE) に到達するまでの BM 内転送

フェーズ 2 レベル $j(2 \leq j \leq L)$ の転送

フェーズ 3 BM 間リンクの出口からディスティネーション PE までの BM 内転送

Routing-H3D-torus();
source node adress:$s_{n-1}, s_{n-2}, \cdots, s_0$
destination node adress:$d_{n-1}, d_{n-2}, \cdots, d_0$
tag:$t_{n-1}, t_{n-2}, \cdots, t_0$,
movedir:moving direction
for i=n-1:3;
if(($t_i > 0$ and $2 \geq t_i$) or ($t_i < 0$ and $t_i = -3$)), movedir = +; endif;
if(($t_i > 0$ and $t_i = 3$) or ($t_i < 0$ and $t_i = -1$)), movedir = -; endif;
if(movedir= + and $t_i > 0$), distance=t_i; endif;
if(movedir= + and $t_i < 0$), distance=$t4+_i$; endif;
if movedir= -, distance=1; endif;
j=1 mod 3;
while($t_i \neq 0$ or distance $\neq 0$) do
if j=2, gate-node=z-axis gate-node of Llevel-⌈i/3⌉; endif
if j=1, gate-node=y-axis gate-node of Llevel-⌈i/3⌉; endif
if j=0, gate-node=x-axis gate-node of Llevel-i/3+1; endif
send packet to next BM; distance=distance-1;
endwhile;
endfor;
BM-tagt_2, t_1, t_0=receiveing node address(r_2, r_1, r_0)-destination(d_2, d_1, d_0)
while($t_2 \neq 0$)do
if $t_2 > 0$, move packet to +z node;$t_2 = t_2 - 1$; endif;
if $t_2 < 0$, move packet to -z node;$t_2 = t_2 + 1$; endif;
endwhile;
while($t_1 \neq 0$)do
if $t_1 > 0$, move packet to +y node;$t_1 = t_1 - 1$; endif;
if $t_1 < 0$, move packet to -y node;$t_1 = t_1 + 1$; endif;
endwhile;
while($t_0 \neq 0$)do
if $t_0 > 0$, move packet to +x node;$t_0 = t_0 - 1$; endif;
if $t_0 < 0$, move packet to -x node;$t_0 = t_0 + 1$; endif;
endwhile;
end;

図 5.3: H3D トーラスのルーティングアルゴリズム

ただし，転送の最初のホップで BM の側面に到達する時は，フェーズ 1 は無視する．
また，フェーズ 2 は，以下のサブフェーズに分けられる．

サブフェーズ 2.i.1 $L-i$ レベル z方向 BM 間リンクの入口の PE に到達するまでの BM 内転送

サブフェーズ 2.i.2 $L-i$ レベル z方向 BM 間リンクを使用した BM 間転送

サブフェーズ 2.i.3 前フェーズの転送を終え，i レベル y方向 BM 間リンクの入口の PE に到達するまでの BM 内転送

サブフェーズ 2.i.4 $L-i$ レベル y方向 BM 間リンクを使用した BM 間転送

サブフェーズ 2.i.5 前フェーズの転送を終え，i レベル x 方向 BM 間リンクの入口の PE に到達するまでの BM 内転送

サブフェーズ 2.i.6 $L-i$ レベル x 方向 BM 間リンクを使用した BM 間転送

ここで，$0 \leq i \leq L-2$ である．

BM 内転送を行うサブフェーズでは，通過しない BM 間リンクの，入口にある PE を途中で通過するケースがある．このような場合，その PE までの転送は，通過しない BM 間リンクまでの BM 内転送として扱う．たとえば，3 レベルの TESH における BM 内転送で，2 レベル縦方向リンクの入口の PE を通過して 2 レベル横方向リンクへ向かうような場合，2 レベル縦方向リンクの入口の PE まではサブフェーズ 2.1.1，それ以降はサブフェーズ 2.1.3 となる．

次に，これら各々について仮想チャネルを用意してデッドロックを防止することを考える．

補題 5.1　以下の条件でランク 2 のチャネルを使い分け，2 個のチャネルを使用したリング網はデッドロック・フリーである．
（条件 1）ルーティング開始時はチャネル 0
を使用
（条件 2）ラウンドトリップ時にチャネル 1 に移動

証明　以下のようにチャネル番号を割り振ればチャネル依存グラフにそってチャネル番号が昇順に割り振られるので，デッドロック・フリーが証明される．

$$\begin{cases} (ch, n_i), & +\text{方向のチャネル}, \\ (ch, 4 - n_i), & -\text{方向のチャネル}, \end{cases}$$

$c_h =$（使用した仮想チャネル），
（0:チャネル 0，1:チャネル 1）　□

補題 5.2　二次元メッシュで e-cube ルーティングを行うものとする．この二次元メッシュはデッドロック・フリーである．

証明　以下のように チャネル番号を割り振ればチャネル依存グラフにそってチャネル番号が昇順に割り振られるので，デッドロック・フリーが証明される．

$$\begin{cases} (0, n_1), & \text{y+方向のチャネル}, \\ (1, 4 - n_1), & \text{y−方向のチャネル}, \\ (2, n_0), & \text{x+方向のチャネル}, \\ (3, 4 - n_0), & \text{x−方向のチャネル}, \end{cases}$$

□

補題 5.3　三次元メッシュで e-cube ルーティングを行うものとする．この三次元メッシュはデッドロック・フリーである．

証明　以下のようにチャネル番号を割り振ればチャネル依存グラフにそってチャネル番号が昇順に割り振られるので，デッドロック・フリーが証明される．

$$\begin{cases} (0, n_1), & \text{z+方向のチャネル,} \\ (1, 4 - n_1), & \text{z−方向のチャネル,} \\ (2, n_1), & \text{y+方向のチャネル,} \\ (3, 4 - n_1), & \text{y−方向のチャネル,} \\ (4, n_0), & \text{x+方向のチャネル,} \\ (5, 4 - n_0), & \text{x−方向のチャネル,} \end{cases}$$

□

以上の理論を用いた H3D トーラスのデッドロックフリー・ルーティングを考える．

定理 5.1　チャネル数 2 の H3D トーラスはデッドロック・フリーである．

証明　各フェーズで必要となる仮想チャネルの数を考える．各フェーズでデッドロック・フリーが保証されれば H3D トーラスネットワーク全体でデッドロック・フリーが保証される．

フェーズ 1 は 3 次元メッシュの形状をしている．したがって，補題 5.3 より，必要な仮想チャネル数は 1 となる．

フェーズ 2 のサブフェーズ 2.i.1，2.i.3 および 2.i.5 では BM の周囲のチャネルを使用する．これらのチャネルは 2 次元メッシュの形状をしているので，補題 5.2 より，必要な仮想チャネル数は 1 となる．フェーズ 2 のサブフェーズ 2.i.2，2.i.4 および 2.i.6 はリング状をしているため，補題 5.1 により必要な仮想チャネル数は 2 である．

フェーズ 3 は 3 次元メッシュの形状をしている．したがって，補題 5.3 より，必要な仮想チャネル数は 1 となる．

以上のうち，フェーズ 2 のサブフェーズ 2.i.2，2.i.4 および 2.i.6 は BM 間リンクを使用したチャネルで，二つ以上のフェーズで同一のリンクを共有しないので，すべての BM 間リンクの必要仮想チャネル数は 2 となる．また，フェーズ 2 のサブフェーズ 2.i.1，2.i.3 および 2.i.5 は BM 内リンクを使用したチャネルで，二つ以上のフェーズで同一のリンクを共有しない．これらのサブフェーズ用チャネルは，それぞれがサブフェーズ 1.1 およびサブ

フェーズ 3.1 と同じリンクを共有する．フェーズ 1 とフェーズ 2 の 2 つのフェーズはお互いに重なり合わないので，BM 内リンクの必要仮想チャネル数の最大値は，フェーズ 3 と合わせて 2 となる．

H3D トーラスネットワークに必要な仮想チャネル数は，これらの数字の最大値となるので，ネットワーク全体で必要な仮想チャネル数は 2 となる． □

5.4 動的通信性能の評価

5.4.1 シミュレーション条件

2 階層の H3D トーラスについて，シミュレーションによる動的通信性能の評価を行う．

シミュレーションに使用した結合網は，2 レベルリンクを一組持つ H3D トーラスと，四組持つ H3D トーラス，TESH(2,3,0) チャネル数 1 および 2 の 64×64 メッシュ，チャネル数 2 の 64×64 トーラスである．これらの結合網の PE 数は，すべて 4096PE で同数である．

TESH(2,3,0) のルーティングは，4.3節で説明した固定ルーティング，64×64 メッシュ及び 64×64 トーラスのルーティングは，Y 方向→X 方向の次元順ルーティング [15] としている．なおシミュレーションは，全 PE で均一に通信を行うランダム通信について行った．

評価は，パケットの平均転送時間およびスループットを用いる．パケットの転送時間は，パケットの先頭フリットがソース PE を出発してから，パケットの最後尾がディスティネーション PE に到着するまでの時間であらわされ，平均転送時間は転送時間の平均値で示される．なお，時間の単位はクロックサイクルである．スループットは，1 クロックサイクルに 1PE が受け取るフリット数の平均値である．本実験では，20000 サイクルの間に送信したフリット数を比較している．なお，パケット長は 18 フリット（うち，2 フリットはヘッダ）としている．本実験では各チャネルが持つチャネルバッファのサイズを 2 フリットとしている．

5.4.2 ランダム通信

平均転送時間を図 5.4に示す．図 5.4で，横軸はスループット，縦軸は転送時間である．結果が示すように，BM 間リンクを一組しか持たない H3D mesh は他の二つのネットワークに通信性能で大きく差をつけられている．これは，2 レベルリンク付近でトラフィックが混雑するためである．

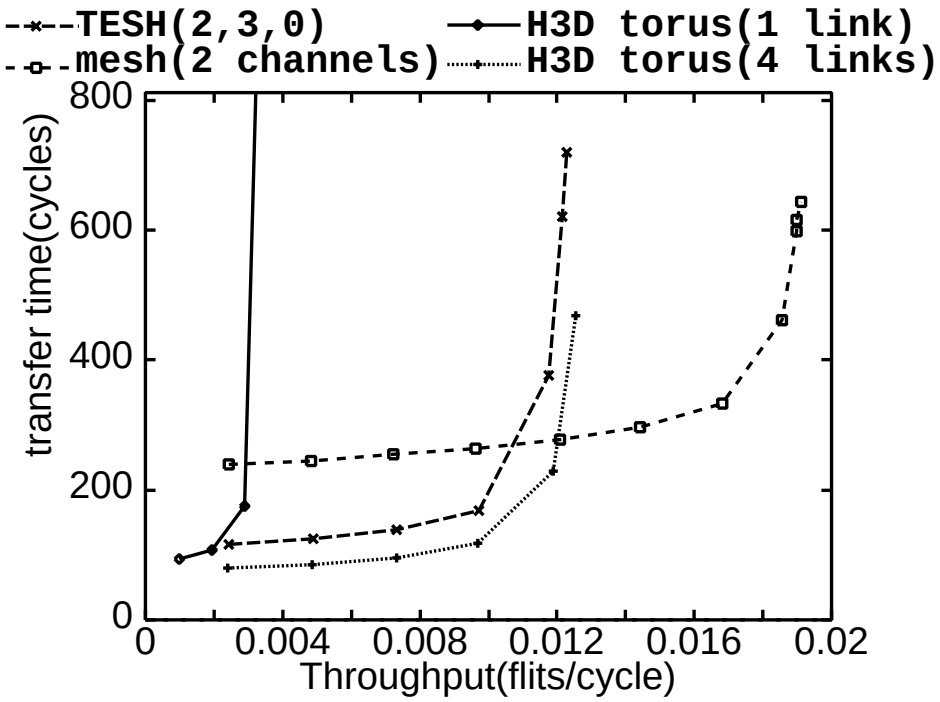


図 5.4: H3D トーラスにおけるランダム通信の平均転送時間

BM 間リンクを四組持った H3D トーラスは，TESH(2,3,0) に若干勝る平均通信時間およびスループットを実現している．この両者のウェハー間配線の数や直径を比較して，H3D トーラスの方がやや少なくなることを考え合わせると，4096PE 規模では，H3D トーラスの方が TESH に比べて若干優れているということが言える．

5.5 適応ルーティングアルゴリズムの適用

H3D トーラスは TESH と同様，階層型相互結合網であるため，TESH の適応ルーティング法のいくつかを H3D トーラスに適用することが可能である．

CS 法や LS 法は，TESH の BM 間リンクであるリング網に対して適用した適応ルーティングアルゴリズムである．TESH と同様，H3D トーラスの BM 間リンクもリング網を持つため，CS 法や LS 法はそのままの形で適用することができる．

DDR 法は，TESH に対するグローバルなルーティングアルゴリズムとして提案された適

応ルーティングアルゴリズムである．TESH と H3D トーラスでは BM 内における BM 間リンクの配置が異なるが，TESH と同様に BM の一部に BM 間リンクが配置される構造をしており，BM 内転送の途中で，BM 間リンクの入口にあたる PE を複数箇所にわたって通過するため，TESH に対する DDR 法と同じアルゴリズムを使用できる．

5.6 まとめ

本章では，階層型相互結合網 H3D トーラスのデッドロック回避法を提案し，必要な仮想チャネル数を導出した．その結果として，必要仮想チャネル数が 2 本であることを証明した．

また，シミュレーションによって 4096PE を持つ H3D トーラスの動的性能評価を行った．その結果，ほぼ同数のリンク数と PE 数を持つ TESH に比べて，若干勝る平均通信時間およびスループットを実現していることを示した．

今後の課題として，さらに PE 数を増やした H3D トーラスによる通信性能の評価，および H3D トーラスのための適応ルーティングが挙げられる．

第 6 章

結言

6.1 はじめに

本論文では，3 次元大規模 VLSI 向け階層型相互結合網である TESH およぼ H3D トーラスにおける適切なリンク配置を提案し，ルーティング法を提案し，それによる動的通信性能を評価した．本章では，本論文の結論を各章ごとにまとめ，さらに今後の課題について述べる．

6.2 本論文の結論

はじめに，従来提案されている超並列計算機の相互結合網を紹介し，その評価方法および従来の相互結合網の問題点を述べた．

並列計算機の基本的な相互結合網には，直線，リング，メッシュ，トーラス，ツリー網などなどのさまざまなものが存在する．そのため，これらについて詳細に説明した．また，これらの結合網の欠点を補完するための結合網として，再帰型相互結合網や階層型相互結合網などの相互結合網の構成，各パラメータ等について述べた．最後に，大規模 VLSI 実装に適した結合網として，TESH(Tori connected mESHes) および H3D トーラス (Hierarchical 3-Dimensional torus) について説明した．

次に，ワームホールルーティングおよび，仮想チャネルを用いたデッドロック回避の方法について述べた．

パケット転送方式には，ストアアンドフォワード，ワームホールルーティング，バーチャルカットスルーと，大きく分けて三種類があり，うちワームホールルーティングは，他の転送方式に対して多くの利点を持った方式であることを述べた．その反面，パケット同士

の衝突によるブロッキングが多くなるため，結合網のデッドロックが起こりやすくなることを述べた．デッドロックに対処するための手法として，論理的にデッドロックを防ぐ方法がワームホールルーティングに対して一般的に用いられていることを述べ，そのための方法として，チャネル依存グラフを作り，チャネル依存グラフが順序関係を持つことを証明できれば，その結合網がデッドロック・フリーであることを証明することができることを述べた．

さらに，階層型相互結合網 TESH におけるデッドロックフリーなルーティング法を提案した．TESH のデッドロックフリーの証明には，全体のルーティングを，順序関係を持ちそれ自体が基本的な相互結合網の形状をしている複数のフェーズに分割し，各フェーズでデッドロックが起らないことを証明するという方法を用いた．

第一に，BM 一個あたりの BM 間リンク数が 8 本を越える場合と越えない場合の 2 つの場合について，適切なリンク配置の方法を提案した．BM 間リンク数が 8 本を越える場合，各レベルのリンクを一列に配置する一列配置の方法により，ホップ数および必要仮想チャネル数を少なく保つ方法を提案した．また，それぞれについてデッドロックを回避するために必要な仮想チャネル数を導出した．その結果，BM 間リンク数が 8 本を越える TESH の必要仮想チャネル数が 4 本，越えない TESH の必要仮想チャネル数が 2 本であることを証明した．これらは，TESH そのもののサイズには依存しない値である．さらに，シミュレーションにより動的通信性能についての評価を行った．ランダム通信，最大値問題，Non-Uniform Transfer について動的通信性能評価を行った結果，TESH(2,3,1) はメッシュに比べて良好な通信性能を示すことを示した．また，TESH(2,3,0) はチャネルバッファ数が多い場合や仮想チャネル数が多い場合，チャネルバッファ数が少ないものやチャネル数が少ないものに比べて大きな性能向上が得られ，同サイズのメッシュに比べて良好な通信性能を示すことを示した．

第二に，階層型相互結合網 TESH におけるルーティング性能の向上のための適応ルーティングの手法として，上位レベルリンクにおいて複数チャネルを動的に選択する CS 法，同一レベルリンクにおいて複数方向を選択する LS 法，上位レベル転送に動的次元逆転ルーティングを用いた DDR 法，余剰の上位レベルチャネルの動的選択法，BM に対して Duato の適応ルーティングを用いる方法を提案し，提案手法がデッドロック・フリーを保証することを証明した．さらに，uniform transfer，hotspot transfer，Matrix Transpose，local transfer についてシミュレーションにより固定ルーティングと提案手法の比較検討を行った．その結果，いくつかの通信パターンについて固定ルーティングに比べて低い平均通信時間および高いスループットが実現できることが明らかになった．

第三に，大規模相互結合網 TESH 上における 1 次元 FFT のトーラス結合を利用して，一部データを置き換えるマッピング法を提案した．本マッピング法では，単純なマッピング法

に比べて通信距離が短くなり，かつメッセージ間の衝突が減るため，実行時間を短縮することができた．シミュレーションによって実験を行った結果，4096 個の PE を持つ 3 階層の TESH で，同サイズのメッシュに比べて約 1.3 倍程度実行時間を短縮することができた．

また，階層型相互結合網 H3D トーラスのデッドロック回避法を提案し，必要な仮想チャネル数を導出した．H3D トーラスのデッドロックフリーの証明には，TESH と同様に全体のルーティングを複数のフェーズに分割し，各フェーズでデッドロックが起らないことを証明するという方法を用いた．その結果として，必要仮想チャネル数が 2 本であることを証明した．また，ランダム通信のシミュレーションによって 4096PE を持つ H3D トーラスの動的性能評価を行った．その結果，ほぼ同数のリンク数と PE 数を持つ TESH に比べて，若干勝る平均通信時間およびスループットを実現していることを示した．

6.3 今後の課題

今後の課題として，第一に H3D トーラスの性能評価が挙げられる．H3D トーラスは，より多数の PE のためのシステムとして提案された結合網である．これまでの研究で，4096PE を超えた数での並列処理に適していることが指摘されている．そこで，PE 数を増やした H3D トーラスによる性能評価を行う．具体的には，$m = 1$ とした場合の評価を行う．提案された固定ルーティングは，どのようなサイズの H3D トーラスにも適用可能であるが，$m = 1$ とした場合，上位レベルがハイパーキューブと同じになるので，デッドロック防止のための方策が若干変わってくる．

第二に，H3D トーラスの適応ルーティングが挙げられる．H3D トーラスは TESH と似たような構造をしているため，本論文で提案した TESH の適応ルーティング法を応用することができる．これにより，ネットワークのスループットおよび耐故障性の向上を実現することができる．

6.4 おわりに

本論文では，3 次元大規模 VLSI 向けの階層型相互結合網として TESH およ H3D トーラスに着目し，それらの結合網を用いて並列計算機を実現するにあたって，メッセージのルーティングに伴い問題となるデッドロックを解決し，より効率的なメッセージルーティングのための様々な手法を提案した．

本論文で提案されたルーティング法により，ネットワークのサイズによらず，TESH ネットワークを用いた効率的なメッセージのルーティングが可能となる．また，TESH と同様に

H3D トーラスの固定ルーティング法を提案したことにより，より規模の大きな H3D トーラスの動的通信性能評価や適応ルーティング法の開発が可能になる．

今後に残された課題としては，さらに PE 数をふやした H3D トーラスの性能評価と適応ルーティング法の開発が挙げられる．

謝辞

本研究は北陸先端科学技術大学院大学 情報科学研究科の堀口進教授のもとで行われました．本研究を進めるにあたり，御指導御鞭撻を賜った堀口進教授に心から感謝すると共に御礼申し上げます．

副テーマの課題について御指導を賜りました松澤照男教授に，深く御礼申し上げます．本研究全般に渡り，貴重な御助言を賜りました阿部亨助教授に，心から感謝致します．

日比野靖教授ならびに吉岡良雄教授には，御教示御検討を頂きましたことを深く感謝致します．

林亮子助手，井口寧助手ならびに山森一人助手（現宮崎大学助教授）には，日頃から有意義な御指導御助言を賜りましたことを深く感謝致します．

最後に，堀口研究室の皆様には大変お世話になりました．心より御礼申し上げます．

参考文献

[1] J.Kuskin et.al. The Stanford FLASH Multiprocessor. In *ISCA94*, pp. 302–313, 1994.

[2] Anant Agarwal et.al. The MIT Alewife Machine: Architecture and Performance. In *Proc. of ISCA'95*, pp. 2–13, 1995.

[3] H.Tanaka. *The Massively Parallel Processing System JUMP-1.* Ohmsha IOS Press, 1986.

[4] 石畑, 稲野, 堀江, 清水, 池坂. 高並列計算機 AP1000 のアーキテクチャ. 電子情報通信学会論文誌, Vol. J75-D-I, No. 8, pp. 637–645, 1992.

[5] J.Carson. The Emergence of Stacked 3D Silicon and Impacts on Microelectronics Systems Integration. In *IEEE Int'l Conf. on Innovative Systems in Silicon*, pp. 1–8, 1996.

[6] S.Horiguchi. Wafer Scale Integration. In *Proc. 6th International Microelectronics Conference*, pp. 51–58, 1990.

[7] M.J Little, J.Grinberg, S.P.Laub, J.G.Nash, and M.W.Yung. The 3-D Computer. In *IEEE Int'l Conf. on Wafer Scale Integration*, pp. 55–64, 1989.

[8] Michael L.Campbell, Scott T.Toborg, and Scott L.Taylor. 3-D Wafer Stack Neurocomputing. In *IEEE Int'l Conf. on Wafer Scale Integration*, pp. 67–74, 1993.

[9] H.Kurino, T.Matsumoto, K.H.Yu, N.Miyakawa, H.Tsukamoto, and M.Koyanagi. Three-dimensional Integration Technology for Real Time Micro-vision Systems. In *IEEE Int'l Conf. on Innovative Systems in Silicon*, pp. 203–212, 1997.

[10] V.K.Jain, T.Ghirmai, and S.Horiguchi. TESH:A New Hierarchical Interconnection Network for Massively Parallel Computing. *IEICE Transactions*, Vol. E80-D, No. 9, pp. 837–846, 1997.

[11] S.Horiguchi and T.ooki. Hierarchical 3D-Torus Interconnection Network for Massivery Parallel Computers. Jaist research report(is-rr-2000-022), JAIST, 2000.

[12] V.K.Jain, T.Ghirmai, and S.Horiguchi. Reconfiguration and Yield for TESH: A New Hierarchical Interconnection Network for 3-D Integration. In *IEEE Proceeding of International Conference Wafer Scale Integration*, pp. 288–297, 1996.

[13] V.K.Jain and S.Horiguchi. VLSI Considerations for TESH: A New Hierarchical Interconnection Network for 3-D Integration. *IEEE Trans on Very Large Scale Integration(VLSI) Syatems*, Vol. 6, No. 3, pp. 346–353, 1998.

[14] Dongming Peng and Mi Lu. Optiamally Embedding Discerete Wavelet Transform into TESH Connected Parallel Processors via I/O Index Space Data Dependence Analysis. In *Proceedings of PDCAT2000*, pp. 125–132, 2000.

[15] L.M.Ni and P.K.McKinley. A Survey of Wormhole Routing Techniques in Direct Networks. *Proc of the IEEE*, Vol. 81, No. 2, pp. 62–76, 1993.

[16] 富田眞治. 並列コンピュータ工学. 昭晃堂, 1996.

[17] 天野英晴. 並列コンピュータ. 昭晃堂, 1995.

[18] Y.R.Potlapalli. Trend in InterconnectionNetworks Topologies: Hierarchical Networks. In *Proc. of ICPP'95*, pp. 24–29, 1995.

[19] 楊, 天野, 柴村, 末吉. 超並列計算機向き結合網:RDT. 電子情報通信学会論文誌, Vol. J78-D-I, No. 2, pp. 118–128, 1995.

[20] 井口寧, 堀口進. 超並列計算機用プロセッサ結合網 SRT. 電子情報通信学会技術研究報告 (cpsy96-45), 電子情報通信学会, 1996.

[21] F.P.Preparata and J.Vuillemin. The Cube-Connected Cycles: A Versatile Network for Parallel Computation. *Comm. ACM*, Vol. 24, No. 5, pp. 300–309, 1981.

[22] P.Kermani and L.Kleinrock. Virtual Cut-Through: A New Computer Communication Switching Techniques. *Computer Networks*, Vol. 3, No. 4, pp. 267–286, 1979.

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

[24] C.J.Glass and L.M.Ni. Maximally Fully Adaptive Routing in 2D Meshes. In *ISCA92*, pp. 278–287, 1992.

[25] W.J.Dally. Virtual-Channel Flow Control. *IEEE Trans on Parallel and Destributed Systems*, Vol. 3, No. 2, 1992.

[26] M.P.Merlin and J.P.Schweitzer. Deadlock Avoidance in Store-and-Forward Networks - 1: Store and Forward Deadlock. *IEEE Trans. on Comm.*, Vol. COM-28, No. 3, pp. 345–354, 1980.

[27] D.H.Linder and J.C.Harden. An Adaptice and Fault Tolerant Wormhole Routing Strategy for k-ary n-cubes. *IEEE Trans. on Computers*, Vol. C-40, No. 1, pp. 2–12, 1991.

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

[29] E. Fleury and P.Fraigniaud. A General Theory for Deadlock Avoidance in Wormhole-Routing Networks. *IEEE Trans. Parallel and Distributed Systems*, Vol. 9, No. 7, pp. 626–638, 1998.

[30] C.S.Yang and Y.M.Tsai. Adaptive Routing in k-ary n-cube Multicomputers. In *Proc. of ICPADS '96*, pp. 404–411, 1996.

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

[32] J.W.Cooley and J.W.Tukey. An Algorithm for the Machine Calculation of Complex Fourier Series. *Math. Compt.*, Vol. 19, pp. 297–301, 1965.

[33] 武田利浩, 丹野州宣, 堀口進. 分散共有メモリ型並列計算機上での並列 FFT アルゴリズムの実装評価. 情報処理学会研究報告 (97-hpc-68-7), 情報処理学会, 1997.

[34] 澤邊知子, 藤井哲郎, 小野定康. 格子型結合並列処理システムにおける 2 次元 FFT. 電子情報通信学会論文誌, Vol. J73-A, No. 7, pp. 1290–1293, 1990.

[35] K.Tanno and T.Takeda. Parallel 2-D FFT Algorithms on an Eight-Neighbor Processor Array. *T.IEE Japan*, Vol. 114-C, No. 5, 1994.

[36] 国枝博昭, 伊藤和人. メッシュ結合マルチプロセッサシステムにおける 2 次元 FFT アルゴリズム. 電子情報通信学会論文誌, Vol. J71-A, No. 7, pp. 1424–1431, 1988.

[37] 三浦康之, V.K.Jain, 堀口進. 階層型ネットワーク TESH におけるデッドロックフリー・ルーティング. 情報処理学会論文誌, Vol. 41, No. 5, pp. 1370–1378, 2000.

研究業績

査読付き論文

1. 三浦康之, 堀口進, Vijay K, jain,
   「階層型ネットワーク TESH におけるデッドロックフリー・ルーティング」,
   情報処理学会論文誌, Vol.41, No.5, pp.1370-1378, May 2000

2. 三浦康之, 堀口進, 「階層型相互結合網 TESH の適応型ルーティング」,
   情報処理学会論文誌（準備中）

査読付き国際会議発表論文

1. Yasuyuki Miura, Susumu Horiguchi,
   "A Deadlock-Free Routing for Hierarchical Interconnection Network:TESH",
   Proceedings of the Fourth International Conference on High Performance Computing in Asia-Pacific Region, Vol.1, pp.128-133, 2000

査読付きシンポジウム

1. 三浦康之, 堀口進
   「階層型相互結合網 TESH の動的通信性能」,
   Joint Symposium on Parallel Processing 2001 論文集, pp.239-246, 2001

学会発表

1. 三浦康之, 阿部亨, 堀口進,
   「ワームホールルーティングにおける仮想チャネルフロー制御」,
   情報処理学会研究報告 (98-HPC-74-11), pp.59-64, 1998

2. 三浦康之, 阿部亨, 堀口進,
   「階層型ネットワーク TESH における仮想チャネルフロー制御法」,
   情報処理学会研究報告 (99-ARC-133-8), pp.43-48, 1999

3. 三浦康之, 阿部亨, 堀口進,
   「ワームホールルーティングにおける仮想チャネルのフロー制御」,
   情報処理学会第 57 回全国大会論文集, Vol.1, pp.48-49, 1998.

4. 三浦康之, 阿部亨, 堀口進,
   「階層型相互結合網 TESH の動的通信性能」,
   1999 年電子情報通信学会ソサイエティ大会講演論文集, pp.36, 1999.9

紀要等

1. Yasuyuki Miura, Susumu Horiguchi,
   "A Deadlock-Free Routing for Hierarchical Interconnection Network:TESH",
   JAIST Research Report, IS-RR-2000-004A, pp.1-13, 2000