

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

三浦 康之, 阿部 亨, 堀口 進

(miu, beto, hori@jaist.ac.jp)

北陸先端科学技術大学院大学

階層型相互結合網の一種である TESH(Tori connected mESHes) 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、プロセッサのリンク数と直径を小さくし、通信の局所性を利用したネットワークである。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、適切な方法によるリンクの配置が必要となる。本稿では、ネットワーク距離および仮想チャネル数を最小にするために基本モジュール間リンクを一列に配置する方法を提案する。また、シミュレーションにより TESH における動的通信性能の評価を行う。その結果、TESH のネットワーク性能が、同サイズのメッシュに比べて優れていることを明らかにする。

Virtual-Channel Flow Control of Hierarchical Interconnection Network: TESH

Yasuyuki Miura, Tohru Abe, Susumu Horiguchi

(miu, beto, hori@jaist.ac.jp)

Japan Advanced Institute of Science and Technology

A hierarchical interconnection network: TESH(Tori connected mESHes) consists of torus interconnection between meshes as a basic module(BM), and allows exploitation of computational locality. To implement a wormhole routing on TESH, an appropriate method is required to allocate virtual links on a basic module, since the number of virtual channels depend on allocation policy. This paper addresses a link allocation policy that minimizes the network diameter and the number of virtual channels. Dynamic communication performances are simulated for TESH and mesh networks. It is seen that the dynamic communication performance of TESH is better than mesh interconnection.

1 はじめに

VLSI 技術の向上にともない、数十から数千のプロセッサを持つマルチプロセッサシステムが実現可能となってきた。近年様々な並列計算機が構築され、アプリケーションの並列化がなされてきている。多くのアプリケーションは通信の局所性を有しているため、その特性を生かした階層型のネットワークを構築することにより、低コストで比較的高性能な通信が可能となる。

階層型相互結合網の一種である TESH(Tori connected mESHes)[1] 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。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。

本稿では、3階層からなる TESH 相互結合網上で、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を行う手法を提案する。また、TESH の直径を算出し、デッドロックフリーを保証するために必要な仮想チャネルの数を導出する。さらに、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。

2 階層型相互結合網 TESH

図1に2階層の TESH の結合図を示す。階層型相互結合網 TESH[1] は三次元 VLSI/ULSI への実装を考慮した結合網 [2][3] である。TESH は、最下位レベルネットワークである基本モジュールをメッシュ、上位レベルネットワークをトーラスにより結合したネットワークである。

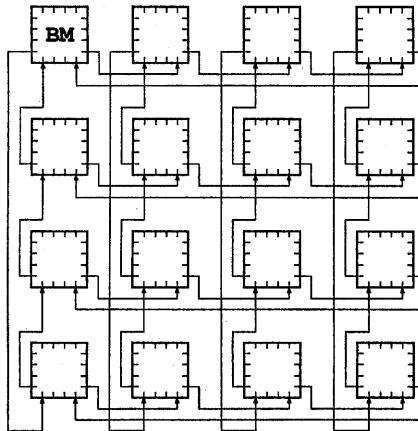


図 1: TESH の構成例

三次元実装では、一つのウェーハに複数の PE を搭載し、複数のウェーハを重ね合わせて実装する。その際、ウェーハ間の配線のために必要なレイアウト面積が大きいので、ウェーハ間の配線はなるべく少ないことが望ましい。TESH は、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ネットワーク性能を保っている。

TESH(m, L, q) は、次のように定義される。

- 基本モジュールは $2^m \times 2^m$ メッシュにより構成。

- L レベルの階層構造により構成される。

- 各レベルにつき 2^q 本の BM 間リンクを持つ。

TESH のルーティングは、上位レベルの転送が先に行われ、次に下位レベル転送が順次行われる。また、同レベル間では垂直方向の転送が先に実行される。

3 デッドロックフリー・ルーティング

3.1 基本モジュール間のリンク配置

基本モジュール間を結ぶ上位レベルリンクは、各基本モジュールの周囲に存在する。そこで、上位レベルリンクをフルに使用して固定ルーティングを行う場合、図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。このようにすると、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必要最小限に抑えることが可能となる。

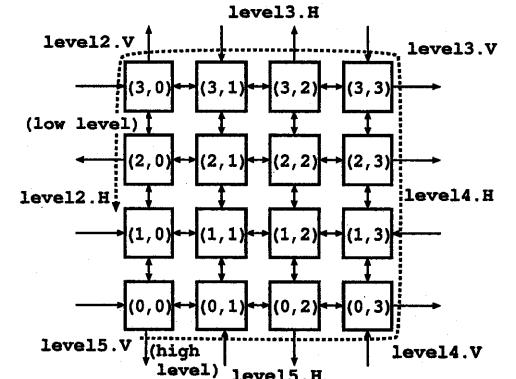


図 2: 上位レベルリンクの配置

TESH のネットワーク直径 $D_{\text{TESH}(m, L, q)}$ は、以下のように 4 ステップから導出できる。

1. レベル L の基本モジュール間リンクに到達するまでの転送回数 D_1 は、目的とするリンクが 4 端 $((0,0)(0,3)(3,0)(3,3)$ のいずれか) にある場合 $q = 0$ で $D_1 = 2 \times (2^m - 1)$
 $q = 1$ で $D_1 = (2^m - 1) + (2^m / 2 - 1)$
 $q = 2$ で $D_1 = 2 \times (2^m / 2 - 1) = 2^m - 2$ となる。

2. 各レベルにおける基本モジュール間転送回数 D_2 は、縦方向と横方向の基本モジュール間リンクのうちいずれか一方が 4 端にある場合

$$D_2 = 3 \times (2^m - 1)$$

となる。

3. 目標の基本モジュールに到達した後の、目的ノードまでの転送回数 D_3 は、レベル 2 の基本モジュール間リンクが 4 端にある場合

$$D_3 = 2 \times (2^m - 1)$$

4 端から n 離れた所にある場合は

$$D_3 = 2 \times (2^m - 1) - n$$

となる。

4. 各レベルの転送の間に行われる基本モジュール内転送の転送回数 D_4 は、基本モジュールの 4 端のリンクから向かう場合は 1 となる。それ以外は 2 となる。 $m = 2$ の場合、各レベルについて 4 端からと 4 端以外からの転送が各一回ずつあるので、 $D_4 = 3$ となる。ただし、レベル 2 の転送は、最後の転送を D_3 とみなすので $D_{4,2LV} = 1$ となる。

以上を用いると TESH のネットワーク直径は

$$D_{\text{TESH}(m,L,q)} = D_1 + D_2 \times (L - 1) + D_4 \times (L - 2) + D_{4,2LV} + D_3$$

となる。

表 1 に、TESH の各パラメータと直径の関係およびメッシュの直径を示す。メッシュに比べて直径が小さくなっていることが分かる。

本手法によりリンクの割り当てを行うと、上記のうち 4 による転送回数を最小限に抑えることができる。

3.2 ルーティングアルゴリズム

図 2 に示すリンク配置における、TESH の固定ルーティング手法について述べる。固定ルーティングを行う場合、PE 間の転送は以下の順序で行われる。

- フェイズ 0
基本モジュール間リンクまでの最初の基本モジュール内転送

表 1: TESH の直径

ノード数	TESH			メッシュ	
	m	L	q	直径	直径
256	2	2	2	17	30
4096	2	3	1	31	126
65536	2	4	0	51	510

- フェイズ 1
レベル L の縦方向上位レベル転送
- フェイズ 2
レベル L の横方向リンクに向かうまでの基本モジュール内転送
- フェイズ 3
レベル L の横方向上位レベル転送
- フェイズ 4L-5
レベル 2 の横方向上位レベル転送
- フェイズ 4L-4
目的地 PE までの基本モジュール内転送

上記のように、フェイズ番号が奇数の時は上位レベル転送、偶数の時は基本モジュール内転送となる。上記の各フェイズを図示すると図 3 になる。図中丸で囲んだ数字が、フェイズ番号である。

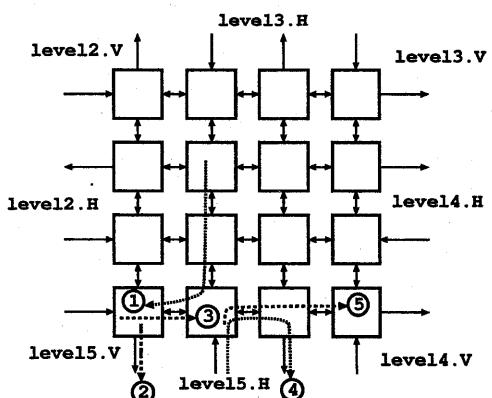


図 3: 各フェイズの転送例

3.3 デッドロックフリー

ここでは、3.2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。

定理1 TESH の固定ルーティングはデッドロックフリーである。

証明) 各フェイズにおいてデッドロックフリーが保証されれば、TESH の固定ルーティング手法がデッドロックフリーであることを証明できる。そこで、各フェイズにそって、次のようなチャネル番号を割り当てる。

((フェイズ番号), (フェイズ内のチャネル番号))

チャネル番号は各フェイズで昇順になるように割り当てることが可能なため、全体のデッドロックフリーが保証される。□

次に、必要なチャネル数について考察すると以下のようない定理2が得られる。

定理2 基本モジュール間のリンクを一列に配置した場合、デッドロックフリーを保証するために必要な仮想チャネル数の最大値は5である。

証明)

フェイズ0およびフェイズ $4L-4$ の転送に使用するチャネルの番号を、メッシュの場合と同様にして転送経路に対して昇順になるように割り当てる。フェイズ $2N+1$ ($N=0, 1, 2, \dots, 2L-3$)の転送に使用するチャネル番号についても、トーラスのチャネル番号割り当てと同様のルールに従って割り当てる。

最後に、フェイズ $2N$ ($N=1, 2, \dots, 2L-3$)の転送に使用するチャネル番号を割り当てる。この場合、転送経路は一つに決まるため、チャネル番号はそのフェイズの転送の最初のPEからの距離で決定できる。以上のようにチャネル番号を割り当ると、チャネル番号が転送経路に対して必ず昇順になるためデッドロックフリーが保証される。

固定ルーティングのフェイズ0、フェイズ $4L-4$ は、メッシュのルーティングと同じになり、フェイズ $2N$ は転送経路が一方向に決まっているので必要な仮想チャネルの数は1である。また、フェイズ $2N+1$ は、トーラスの形状をしているので必要な仮想チャネルの数は2となる。基本モジュール間リンクを一

列に配置した場合、これらのうちフェイズのうち $2N$ に属する複数のフェイズおよびフェイズ $2N+1$ に属する複数のフェイズが一つのチャネルに重なることはないので、考えられる必要チャネル数の最大値は5となる。□

$m=2$ とし、基本モジュール間リンクを全て使用した場合、必要なチャネル数は図4のようになる。この場合、(a)のリンクでフェイズ0、フェイズ $4L-4$ およびフェイズ $2N$ とおよびフェイズ $2N+1$ のうち各一つずつを使用することになる。フェイズ $2N+1$ は2つのチャネルを、他のフェイズでは1つのチャネルを使用するので、合わせて5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。また、(b)の部分でフェイズ0、フェイズ $2N$ およびフェイズ $4L-4$ の、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。

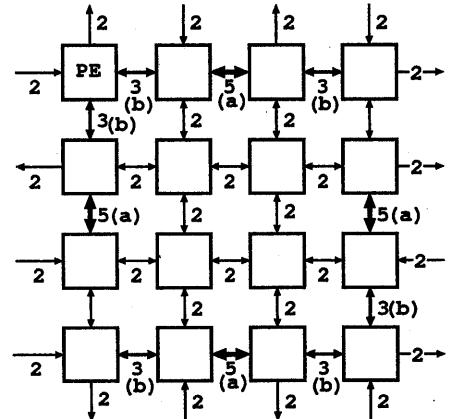


図4: TESH の最大仮想チャネル数

4 シミュレーションによる通信性能

4.1 シミュレーション条件

4096PE からなる TESH(2,3,1) 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5に示すように、各レベルにつき 2 つのリンクを持つ。

シミュレーションは、TESH(2,3,1) およびメッシュ

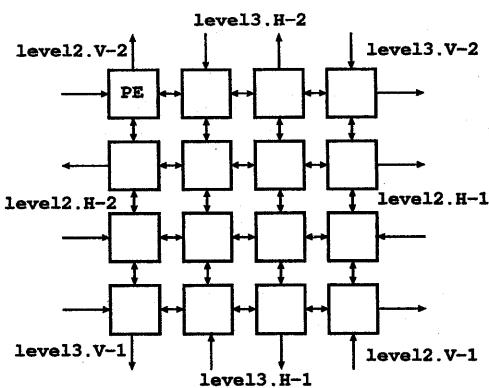


図 5: TESH(2,3,1)におけるリンク配置

結合について行う。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 FFT の 2 種類で行う。

本実験では、メッシュと TESH(2,3,1) の双方とも仮想チャネルの数を 6 としている。TESH(2,3,1) では、必要な仮想チャネル数がリンクの位置により異なるため、最低限必要なチャネル以外はフリーチャネルとし、全フェイズのメッセージが使用できるものとしている。各フェイズごとに固定のチャネルは用意されているので、Duato の条件 [4] により、デッドロックフリーは保証される。パケットの転送方式はワームホールルーティングとし、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、仮想チャネルのアビトレーション法はラウンドロビンとしている。

4.2 ランダム通信

各 PE で、目的地をランダムとしたパケットを連続して送信し、20000 サイクルの間に送信したフリット数を比較する。送信したフリット数が多い程、ネットワークが多数のメッセージを送信する能力を持つということである。なお、パケット長は固定で、4, 16, 64, 256 フリットの時について実験を行っている。

実験結果を図 6 に示す。図中、横軸はパケット長、縦軸は送信したフリット数である。図のように、パケット長がいずれの場合においても TESH(2,3,1) はメッシュより多数のメッセージの転送が可能であることが分かる。TESH は、通信の局所性を利用した

ネットワークであるため、ランダム通信のようなパターンは本来不得意であるが、ネットワークの直径が小さい分メッシュよりもかえって良い結果を示すことになる。

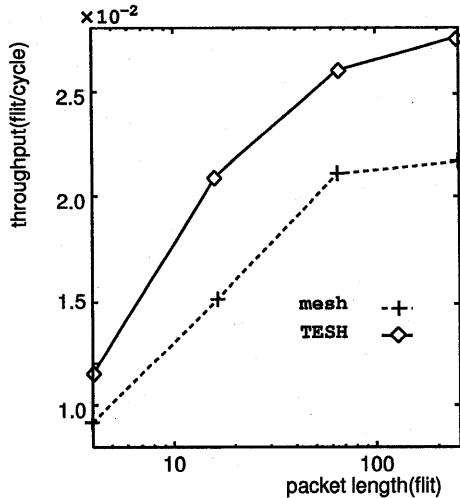


図 6: ランダム通信の実験結果

4.3 FFT

TESH ネットワーク上で FFT を実行する際の通信パターンをシミュレートして性能を評価する。メッシュおよび TESH 上での FFT のアルゴリズムを下記に示す。

メッシュの PE 番号を、 x 座標と y 座標により $PE(x, y)$ と表現する。FFT のデータ番号を 8 進数で $O_n O_{n-1} \dots O_3 O_2 O_1 O_0$ と表現し、各 FFT データを $PE(O_3 O_2, O_1 O_0)$ に配置する。 $PE(0,0)$ には、データ $4096 \times i (i = 0, 1, 2, \dots)$ 、 $PE(0,1)$ にはデータ $4096 \times i + 1 (i = 0, 1, 2, \dots)$ がマッピングされる。メッシュ上にこのようにデータ配置されると、データ数に関係なく通信は一直線上に全部で 12 回行なわれることになる。

各 FFT データを TESH(2,3,1) 上に配置するには、各 FFT データを 4 進数で $B_n B_{n-1} \dots B_3 B_2 B_1 B_0$ と表現し、それぞれを $PE(B_5, B_4)(B_3, B_2)(B_1, B_0)$ に割り当てる。すると、 $PE(0,0)(0,0)(0,0)$ にはデータ $4096 \times i (i = 0, 1, 2, \dots)$ 、 $PE(0,0)(0,0)(0,1)$ にはデータ $4096 \times i + 1 (i = 0, 1, 2, \dots)$ 、という具合にマッピングされる。

このように配置すると、全 12 回の通信のうち 4 回は基本モジュール内のみでの通信となる。また、残り 8 回の通信もレベル 3 またはレベル 2 リンクのみを使用しない通信となるため、通信距離は短くなる。ここでは、FFT データの数が 4096,16384 の時(つまり、各 PE に 1 つまたは 4 つのデータが存在する時)について実験を行った。実験結果を図 7 に示す。図中、横軸はデータ数、縦軸は実行時間である。図 7 のように、TESH(2,3,1) における性能向上は、メッシュの場合と比較して 2 倍以上となっている。FFT の場合、メッシュ上では一つのステージで最大 64 個のパケットが一つのリンクを取り合うことになるが、TESSH(2,3,1) では最大でも 32 個であり、通信距離もメッシュで最大ホップ数が 32 に対して TESSH(2,3,1) では最大 10 となる。このように、リンクの混雑の度合と通信距離がともに少なくなるため、FFT を TESH 上で実行する場合はメッシュよりも性能が良くなる。

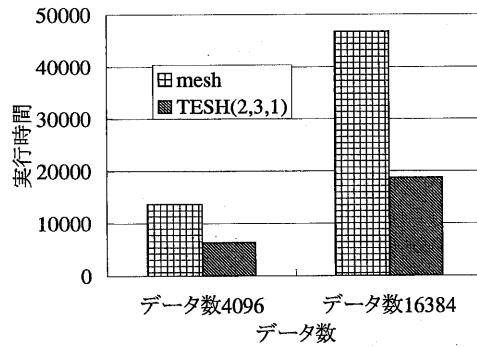


図 7: TESH とメッシュ上での FFT の実行時間

5まとめ

階層型相互結合網 TESH におけるデッドロックフリールーティングを考察し、ホップ数および仮想チャネル数を抑えるための通信アーキテクチャについて検討した。TESHにおいて基本モジュール間のリンクを一列に並べる方法により必要な仮想チャネル数が 5 であることを証明した。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

信性能を示すことを示した。

今後は、仮想チャネル、クロスバスイッチ、ネットワークインターフェイスなどに対し、占有法 [5] などの優先順位決定法を適用して性能比較を行う一方、TESH の適応ルーティングに関して検討する予定である。

謝辞

本研究の一部は、文部省科学研究助成：基盤研究(B) を用いて行われた。関係各位に感謝する。

参考文献

- [1] V.K.Jain,T.Ghirmai,S.Horiguchi "TESH:A New Hierarchical Interconnection Network for Massively Parallel Computing", IEICE Transactions, vol.E80-D, No.9, pp.837-846, 1997
- [2] V.K.Jain, T.Ghirmai and S.Horiguchi "Reconfiguration and Yield for TESH: A New Hierarchical Interconnection Network for 3-D Integration", IEEE Proceeding of International Conference Wafer Scale Integration, pp.288-297, 1996
- [3] 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) Systems, vol.6, No.3, pp.346-353, 1998
- [4] 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
- [5] 三浦, 阿部, 堀口 "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, 1998