# 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

三浦康之[†] 堀口進[†] Vijay K. Jain[††]

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 を用いたマルチプロセッサシステムでワームホールルーティングを行うには仮想チャネルが必要である．このとき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により異なる．したがって，少ない仮想チャネルでワームホールルーティングを実現するには，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．

# Deadlock-free Routing for Hierarchical Interconnection Network: TESH

Yasuyuki Miura,[†] Susumu Horiguchi[†] and Vijay K. Jain[††]

A hierarchical interconnection network: TESH (Tori connected mESHes) consists of torus interconnection between meshes as basic module (BM), and can utilize communication 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 of inter-BM links. 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 の大規模化や，大規模 VLSI の再構成技術の発達などにともない，ウェーハスタック構造による超並列計算機が実現可能となりつつある．ウェーハスタック構造では，1 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て実装する．ウェーハスタックにより超並列計算機を構築する際問題となるのは，ウェーハ間の結線数である．ウェーハ間の配線は，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，いかに配線の数を少なくするかという点が問題となる．ウェーハ間の配線のコストを抑えるには，階層型相互結合網が有効である．なかでも，階層型相互結合網 TESH（Tori connected mESHes）[1)~3)]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ものである．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[4)]を複数付加する必要がある．このとき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，TESH 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法を提案する．また，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し，固定ルーティングによる TESH の通信距離を算出する．さらに，ランダ

† 北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science Technology
†† 南フロリダ大学
University of South Florida

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

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[1)]．本稿では主に $4\times4$ 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 1** に，2 階層の TESH を構成した例を示す．図 1 では，BM 1 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．3 階層の TESH では，さらに 4 本のリンクを使用して 2 階層の TESH どうしを結合する．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[1)]．

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．各レベルにつき $2^q$ 組（つまり $4\times2^q$ 本）の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[1)]．

TESH$(m,L,q)$ の PE は，式 (1) に示す $2^m$ 進数でアドレス付けされる．

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

式 (1) 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．たとえば，$m=2$ で 3 階層 TESH（4096 PE）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 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 が次の条件を満たしたとき，結合リンクを有する．ただし，$i,j\geq2$ とする．

$$\begin{aligned} &\exists i\{n^1_i=(n^2_i\pm1)\bmod 4 \\ &\wedge\forall j(j\neq i \ \rightarrow \ n^1_j=n^2_j)\} \end{aligned} \quad (2)$$

すなわち，$n^1$ と $n^2$ を含む BM アドレスの各桁を比較したとき，差が $\pm1$ または $\pm3$ となる桁が 1 つ存在して残りの桁は同一の値を持つとき，双方の BM は結合リンクを持つ．たとえば，図 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．

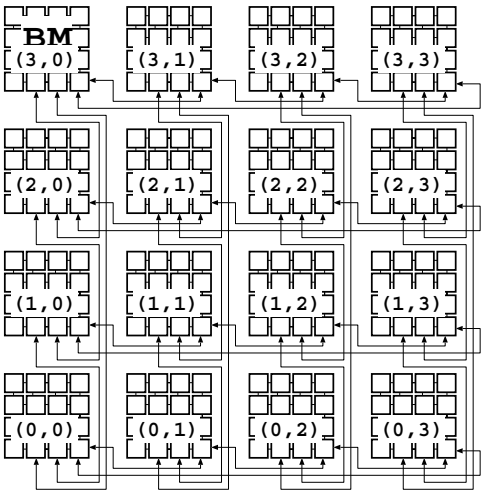



**図 1** TESH(2,2,0) の構成例
Fig. 1 A hierarchical interconnection of TESH(2,2,0).

## 3. ルーティング

### 3.1 基本モジュール間のリンク配置

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 2 本のリンクは同一レベルの 1 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 を使用するときは，ホップ数を少なく保つため隣りどうしの PE から出ているリンクどうしを 1 組とする．

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 2 つの PE から出ているリンクを同一レベルの 1 組のリンクとして使用する．さらに，**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

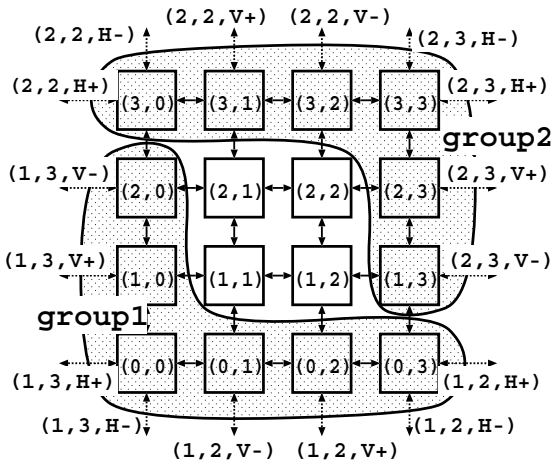



**図 2** BM 間リンクの配置
Fig. 2 A link allocation method at BM.

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 2 に，レベル 3 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うえに，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.

### 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．最後に，目的の BM に到着したとき，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 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 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 1 引数から順に $g$，$l$，$d$，$\delta$ を使用する.

図 3 のアルゴリズムによる転送例を**図 4** に示す．図 2 のようにリンクを配置すると，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 3，図 4 の (a) の，最初のループにおける BM 内転送）と受信先 PE までの最後の BM 内転送（図 3，図 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，仮想チャネル割当ての都合上，これ以降はこれらを分けて

☆ 使用する BM 間リンクが 8 本以内の場合，四隅の PE から出ているリンクのみを使用する方法も考えられるが，本稿では考えない.

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

Routing(s,d);
source;s={s2L-1,s2L-2,...,s0};  destination;d={d2L-1,d2L-2,...,d0};
tag;t2L-1,t2L-2,...,t0;     group;g ;

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

    g = get_group_number(s,d,routedir);

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

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

      ti= ti- 1;
    endwhile;
  endfor;

    BM_routing(d1,d0);
end.

BM_routing(dx, dy);
source;sx,sy; destination;dx,dy;
tag;tx,ty;

  tx= dx- sx;
  ty= dy- sy;
  while(ty != 0) do
    if ty> 0 then move packet to upper node; ty= ty- 1; endif;
    if ty< 0 then move packet to lower node; ty= ty+ 1; endif;
  endwhile;
  while(tx!= 0) do
    if tx> 0 move packet to right node; tx= tx- 1; endif;
    if tx< 0 move packet to left node; tx= tx+ 1; endif;
  endwhile;
end.
```

(a) (b) (c)

**図 3** TESH のルーティングアルゴリズム
Fig. 3 Routing algorithm for TESH.

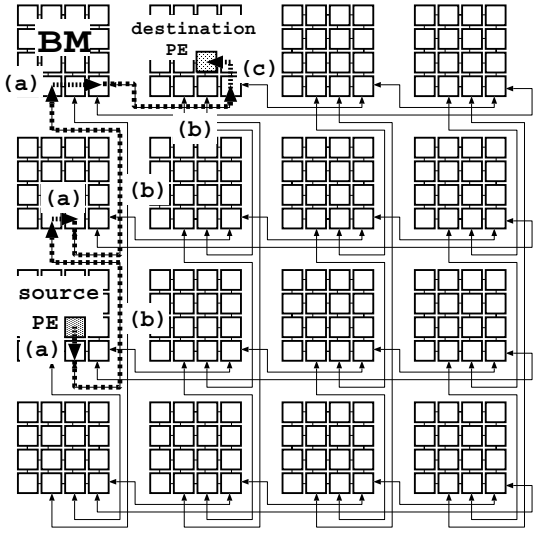



**図 4** TESH の転送例
Fig. 4 An example of routing.

考える．

### 3.3 デッドロックフリー

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り，ネットワーク性能の低下の原因となる．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[5),6)]．本稿ではルーティングに制約を与えない方法として，物理リンクに複数の仮想チャネルを付加する方法[7)~9)]を適用する．

以下，3.2 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当ての都合上，3.2 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 間リンクへ向かうまでの BM 内転送（図 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 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 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 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 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 つのチャネルでよい．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 1 が成り立つ．

**定理 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 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 桁を表す．つまり，あるチャネルが $\mathrm{PE}n^s$ から $\mathrm{PE}n^d$ 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ほか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）のほかに異なるレベル・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

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

$c_h$ =（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，ルーティングに従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換えた新アドレス $\nu$ を導入する．ここで用いられるアドレス $\nu$ 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 $\nu$ 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換わる．

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 5** のようになる．図 5 に示される数字が，必要な仮想チャネルの数である．図 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 1 つの物理リンクを共有することになる．

## 4. 最大ホップ数

3.2 節で示したルーティングを行ったときの 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ときの 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．これらは一列配置

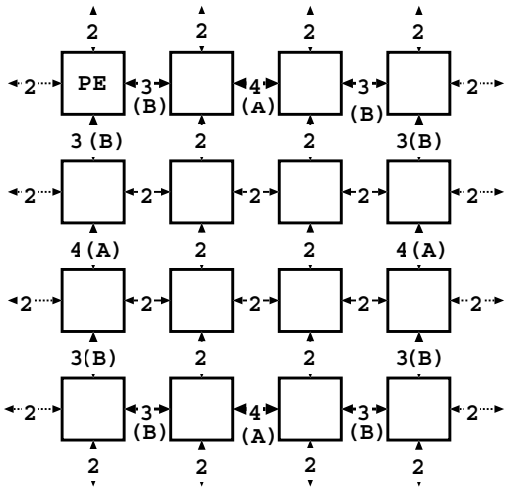



**図 5** TESH の最大仮想チャネル数
Fig. 5 Number of required virtual channels for TESH.

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 3 (a) の最初のループに相当）の転送回数 $D_1$ は

$$D_1 = \begin{cases} 5, & \text{for } q=0, \\ 3, & \text{for } q=1, \\ 1, & \text{for } q=2, \end{cases} \quad (3)$$

となる．

(2) 上位階層は $4\times4$ 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 1 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，

$$D_2 = 2+2+1 = 5, \quad (4)$$

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 2 回ずつとなるので

$$D_3 = 2, \quad (5)$$

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

$$D_4 = 6, \quad (6)$$

となる．

**表 1** TESH の最大ホップ数
Table 1 Muximum number of hops of TESH.

| PE 数 | 結合網 | 格子サイズ | ホップ数 | 次数 |
|---|---|---|---|---|
| 256 | 2D メッシュ | $16\times16$ | 30 | 4 |
| | 2D トーラス | $16\times16$ | 16 | 4 |
| | 3D メッシュ | $8\times8\times4$ | 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 | $64\times64$ | 126 | 4 |
| | 2D トーラス | $64\times64$ | 64 | 4 |
| | 3D メッシュ | $16\times16\times16$ | 45 | 6 |
| | ハイパーキューブ | $2^{12}$ | 12 | 12 |
| | TESH(2,3,1) | | 25 | 4 |

$L$ 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

$$D_{\mathrm{TESH}(m,L,q)} = D_1 + D_2\times(L-1) + D_3\times(2L-3) + D_4 \quad (7)$$

となる．

**表 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☆．表 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．

## 5. シミュレーションによる通信性能評価

### 5.1 シミュレーション条件

4096 PE 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 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，TESH(2,3,1) 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次元順ルーティング[10]によりデッドロックを回避する．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 FFT および最大値問題の 3 種類で行う．

本実験では，TESH(2,3,1) を使用するため，最大仮想チャネル数は 4 となる．メッシュの仮想チャネル数は 1 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[10]とし，サイズの大きなメッセー

☆ 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[10]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．

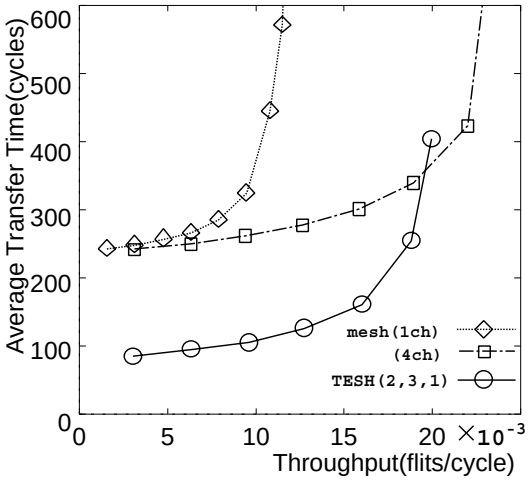



**図 6** ランダム通信の平均転送時間
Fig. 6 Average transfer times for random communications.

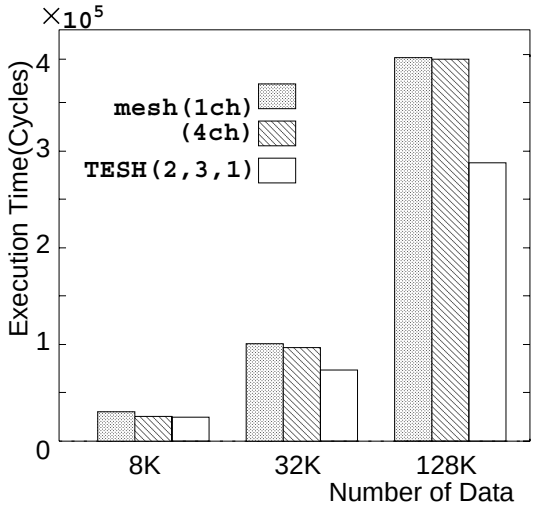



**図 7** TESH とメッシュ上での FFT の実行時間
Fig. 7 Execution times of TESH and meshes for FFT.

ジなども 1 つのパケットで転送できるものとしている．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．

### 5.2 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 をランダムとしたパケットを送信したときの平均転送時間を**図 6** に示す．転送時間は，パケットの先頭が送信元 PE を出発してから最後尾が受信先 PE に到着するまでの時間である．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．

図 6 で，横軸はスループット，縦軸は転送時間である．図 6 より，低スループットの部分では TESH(2,3,1) の平均転送時間はメッシュの半分以下となる．これは，TESH(2,3,1) のホップ数がメッシュに比べて短いためである．また，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．

### 5.3 FFT

高速フーリエ変換（FFT）の持つ通信パターンをシミュレータ上に再現して，実行時間をシミュレーションにより測定する．FFT は，データ数 $d$ に対して $\log d$ 回のバタフライ演算を行う．その際，$k$ 回目の演算は $2^{k-1}$ 離れたデータとの間で行う．そこで，$N$ 個 ($N < d$) の PE に，$2^{p-1}$ ($p > \log N$) 個離れたデータどうしが同じ 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，通信のパターンが局所性を持つ．なお，本実験では，バタフライ演算に要する時間および通信の前後処理に要する時間を 220 サイクル，通信先 PE の計算および通信の前後処理に要する時間を 240 サイクル，1 つの FFT データにつき 64 バイトの長さを持つものとして実験を行った．

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 7** に示す．図 7 より，TESH(2,3,1) における実行時間は，メッシュの場合と比較して短縮されている．チャネル数 1 のメッシュとチャネル数 4 のメッシュを比較すると後者はブロックを回避できる分実行時間も短くなる．ランダム通信では，高スループットの領域でチャネル数 4 のメッシュに比べて通信性能がやや悪くなるが，FFT はチャネル数 4 のメッシュと比較してもデータ数にかかわらず実行時間が短くなる．FFT では，メッシュ上では 1 つのステージで最大 64 個のパケットが 1 つのリンクを取り合うことになるが，TESH(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 TESH(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いと通信距離がともに少なくなるため，TESH 上で FFT を実行した場合メッシュよりも高速に実行できる．

### 5.4 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 どうしで行うと仮定し，4096 PE

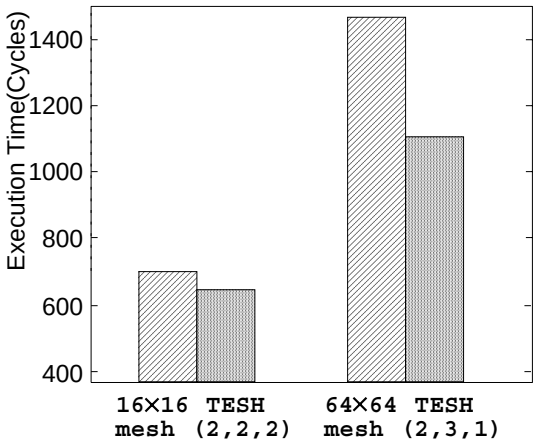



**図 8** 最大値問題の実行時間
Fig. 8 Execution time for max-min.

の場合 12 回，256 PE の場合 8 回の転送を行っている．各回の通信先のアドレスは FFT の場合と同じになるが，最大値問題の場合，大きい方の値のみを転送すれば十分なので，パケットどうしのブロッキングは少なくなる．なお，本実験では，パケット長を 18 フリット，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 8** に示す．図 8 より，TESH とメッシュの実行時間に差が見られる．最大値問題の場合も FFT の場合と同様に，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，256 PE を持つ $16 \times 16$ メッシュと TESH(2,2,2) の実行時間の差と，4096 PE を持つ $64 \times 64$ 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．

## 6. 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提案し，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 4 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，チャネルの割当て方についてさらに検討するとともに，TESH の適応ルーティング法について検討する．

**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) Jain, V.K., Ghirmai, T. and Horiguchi, S.: TESH: A New Hierarchical Interconnection Network for Massively Parallel Computing, *IEICE Trans.*, Vol.E80-D, No.9, pp.837–846 (1997).
2) Jain, V.K., Ghirmai, T. and Horiguchi, S.: Reconfiguration and Yield for TESH: A New Hierarchical Interconnection Network for 3-D Integration, *IEEE Proc. International Conference Wafer Scale Integration*, pp.288–297 (1996).
3) Jain, V.K. and Horiguchi, S.: VLSI Considerations for TESH: A New Hierarchical Interconnection Network for 3-D Integration, *IEEE Trans. Very Large Scale Integration* (*VLSI*), Syatems, Vol.6, No.3, pp.346–353 (1998).
4) Dally, W.J.: Virtual-Channel Flow Control, *IEEE Trans. Parallel and Destributed Systems*, Vol.3, No.2 (1992).
5) Duato, J.: A New Theory of Deadlock-Free Adaptive Routing in Wormhole Networks, *IEEE Trans. Parallel and Distributed Systems*, Vol.4, No.12, pp.1320–1331 (1993).
6) Glass, C.J. and Ni, L.M.: Maximally Fully Adaptive Routing in 2D Meshes, *ISCA92*, pp.278–287 (1992).
7) Merlin, M.P. and Schweitzer, J.P.: Deadlock Avoidance in Store-and-Forward Networks - 1: Store and Forward Deadlock, *IEEE Trans. Comm.*, Vol.COM-28, No.3, pp.345–354 (1980).
8) Linder, D.H. and Harden, J.C.: An adaptice and fault tolerant wormhole routing strategy for k-ary n-cubes, *IEEE Trans. Computers*, Vol.C-40, No.1, pp.2–12 (1991).
9) Dally, W.J. and Seitz, C.L.: Deadlock-Free Message Routing in Multiprocessor interconnection Networks, *IEEE Trans. Computers*, Vol.C-36, No.5, pp.547–553 (1987).
10) Ni, L.M. and McKinley, P.K.: A Survey of Wormhole Routing Techniques in Direct Networks, *Proc. IEEE*, Vol.81, No.2, pp.62–76 (1993).

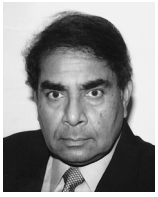
(平成 11 年 9 月 1 日受付)
(平成 12 年 2 月 4 日採録)

**三浦　康之**（学生会員）

昭和 48 年生．平成 9 年東北大学工学部機械知能工学科卒業．平成 11 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．現在同大学院情報科学研究科博士後期課程在学中．並列システムに関する研究に従事．

**堀口　　進**（正会員）

昭和 51 年東北大学工学部通信工学科卒業．昭和 56 年同大学院博士課程修了．昭和 57 年同情報工学科助手．昭和 60〜61 年 IBM ワトソン研究所客員研究員．昭和 64 年同助教授．平成 4 年北陸先端科学技術大学院大学情報科学研究科教授．この間，並列処理，超並列システム，ウェーハ規模集積システム，並列アルゴリズム，マルチメディア統合システムに関する研究を行う．IEEE シニア会員，電子情報通信学会各会員．



**Vijay K. Jain**

1966 年ミシガン州立大学 Ph.D. 現在，南フロリダ大学電気工学科教授．南フロリダ大における DARPA プロジェクトのアーキテクチャ，アプリケーション，および設計グループのリーダー．並列処理システム，相互結合網，VLSI/WSI 設計，高速信号・画像処理，デジタル通信に関する研究を行う．IEEE シニア会員．