超並列計算機用結合網 Ring Tree on Mesh の提案

工藤知宏† 福嶋泰仁† 好村公一† 天野英晴‡ 西村克信‡

†東京工科大学情報工学科
‡慶應義塾大学理工学部

本報告では、まず粗さの異なる同一構造の結合網を組み合わせることにより Fat Tree を構築する手法について述べる。この手法は、RDT 結合網において用いられた手法であるが、より一般的に種々の結合網について用いることができる。

次に、この手法を用いた結合網の一つである RTM(Ring Tree on Mesh) を提案する。RTM は、Mesh または Torus 網により結合されたノード間を大きさの異なる複数のリングにより結合した強化メッシュネットワークである。また、RTM のクラスに含まれる HRM(Hypercube Ring on Mesh) とと DHRM(Double Hypercube Ring on Mesh) についても述べる。

超並列計算機、結合網、ファットツリー

Ring Tree on Mesh: An Interconnection Network for Massively Parallel Processors

T. Kudoh† Y.Fukushima† K.Yoshimura† H.Amano‡ K. Nishimura‡

†Tokyo Engineering University
‡Keio University

First, a scheme to construct a *fat tree* by combining different scale networks of the same structure is presented. While this scheme have been used in the RDT network which have proposed earlier, it can be applied for many other network structures. Here, we summarize the scheme.

Next, an interconection network, called RTM is proposed. RTM combines different sized rings which connect nodes of a mesh or torus network. HRM(Hypercube Ring on Mesh) and DHRM(Double Hypercube Ring on Mesh) which belong to the class of RTM are also discussed.

Massively Parallel Processor, Interconnection Network, Fat Tree

1 はじめに

mesh/torus 結合は、大規模並列計算機の結合網としてきわめて有利な特徴を持ち、現在広く用いられている。しかし、ノードの規模が 10^4を越える超並列計算機の結合網としては、その直径の大きさと、遠隔交信能力の不足が問題になっている。しかし、現在 mesh/torus 結合の並列計算機の利用は広く進んでおり、結合網のトポロジを念頭に置いた計算や転送のアルゴリズムも蓄積されている。これらの蓄積を生かした上、遠隔交信能力を与えるため、二次元、三次元メッシュに対してリンクを付加する構造を持つ強化メッシュの研究が盛んである [3] [5][4] [8][6][7]。

一方、大規模並列計算機上で分散共有メモリを実現する場合、キャッシュラインの update や invalidate 情報が、キャッシュブロックを共有しているノードにマルチキャストされる必要がある。大規模な分散共有メモリ型計算機では、キャッシュラインの共有情報 (ディレクトリ) の管理をライン単位で行なうと情報量が多くなり過ぎるため、より大きなページ単位で管理する方法が提案されている [9][10]。この手法では宛先ノード数が多くなるので、木構造をとる結合網を用いて階層的にマルチキャストすることが望ましいと考えられる。さらに、通信の集中を避けるためには、木構造は Fat Tree であることが望ましい。

このような並列計算機に適した強化メッシュ構造は、単に遠隔交信能力を高めるために付加リンクを与えるのではなく、付加リンクにより Fat Tree の階層構造を構成できることが望ましい。さらに、ノードの構造とルーティングを単純化するため、結合網は単方向のリンクから成る単純なものが有利である。本論文では上記の条件を満足する強化メッシュのクラスのひとつである RTM(Ring Tree on Mesh) を提案する。

2 段階的に粗さの異なる結合網の組み合わせによる Fat Tree の実現

粗さの異なる結合網を組み合わせることによって mesh または torus 上に木構造を実現した例として、最も古典的な結合網は、Mesh with Grobal Mesh(MGM)[3] である。MGM では、2 × 2 のネットワークを図 1のように再帰的に組み合わせる。ここで、黒い点で表された部分がプロセッサである。2 × 2 のネットワークを疑似的に 1 つの節と考え、プロセッサ間を結合する斜めの線を枝ととらえると 4 進木が実現している。

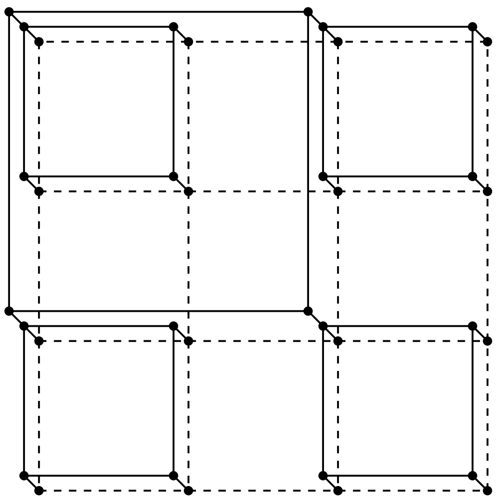

図 1: Mesh with Grobal Mesh

Carlson らは、MGM 上で Prefix Calculation の効率的な実行等のアルゴリズムを検討した。これらのアルゴリズムは他の階層構造を持つ結合網でもある程度応用することが可能であるが、MGM 自体は、Fat ではなく単一の tree である、中間ノードの利用が困難、不平衡型である等の問題点を持っている。

これに対し、Recursive Diagonal Torus(RDT)[4] は、Fat Tree の実現が可能で、実際に文部省重点領域研究の一環として実装が進んでいる JUMP-1[1][2] で利用されている結合網である。

RDT は、目の細かい torus の上に目の粗い torus を 45 度傾けて再帰的に重ねていくことによって構成される。図 2に RDT を示す。この図では、2 種類の粗さの torus が示されている。この図に示した斜め方向のトーラスの上にさらにより粗いトーラスが再帰的に構成される。粗い torus を構成する各ノードが、図に網掛けで示した亀甲形の領域にマルチキャストすることにより疑似的に 8 進木 (送信元を含む) を構成していると考えることができる。図では粗いトーラスを一つしか示していないが、完全 RDT と呼ばれる構成では、全てのノードが図に示したトーラスを平行移動したトーラスのノードとなる。

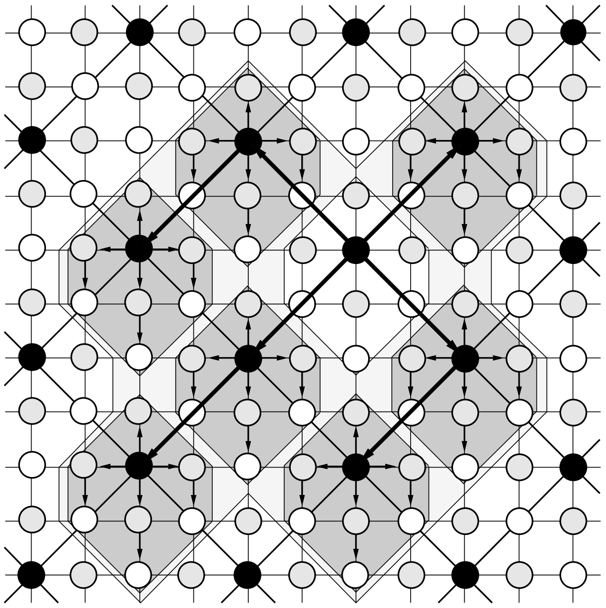

図 2: RDT

このような構成をとると、Fat Tree が構成されることになり、しかも全てのノードが一番粗い torus のリンクを持っているから、どのノードもいずれかの木のルートとなっており、マルチキャストを簡単に行なえる。この手法は、torus/mesh に限らず、様々な形態の結合網にも同様に適用できる。木を遡る必要がないため、マルチキャスト時の平均転送距離が短くなり、混雑が木の根近くに集中することもない。

RDT は、torus/mesh に加えて Fat tree に相当する階層構造を持つ点で優れているが、斜めのリンクを用いるためマルチキャストの領域が複雑であり、また、degree が 8 でやや大きいという問題点がある。秋山らは、前者の問題を解決した結合網を [8] 提案し、岩崎らは、能力を落さずリンク数を削減する結合網 [6] を提案している。また、広く一般化した結合網として定義する試み [7] もある。

3 RTM

3.1 完全形

我々は、mesh に対して付加リンクを与える方法として、階層的なリング構造に着目した。リングは組合せにより mesh 自体をも構成することができる。すなわち、RTM は、mesh または torus 網により結合されたノード間を大きさの異なる複数のリングにより結合した結合網のクラスとして定義される。

ここで、mesh は一辺が 2^n(n は自然数) ノードの正方形とし、各ノードは 2^nを法とする 2 次元座標 (x, y) で表されるものとする。

今、リング $R+(l, x, y)$ を

$$(x, y) \to (x, y+2^l) \to (x+2^l, y+2^l)$$
$$\to (x+2^l, y) \to (x, y)$$

の順に接続する単方向のリング、リング $R-(l, x, y)$ を

$$(x, y) \to (x+2^l, y) \to (x+2^l, y+2^l)$$
$$\to (x, y+2^l) \to (x, y)$$

の順に接続する単方向のリングであるとする。ここで l をレベルと呼ぶ。

全ての x,y および $0 \leq l < n$ である全ての l に対して $R+(l, x, y)$ と $R-(l, x, y)$ が存在する結合を完全 RTM と呼ぶ。図 3に中央の黒く塗りつぶしたノードに対して、完全 RTM を構成するのに必要なレベル 0 および 1 のリングを示す。ここでは、同一パラメタの $R+$と $R-$を一本の線で表している。完全 RTM では全てのノードに対して同様にリングが結合されるので、リングの重なりによって構成される粗さの異なる torus が多数形成される。したがって、degree は $4n$ となり、リンク数は非常に多い。

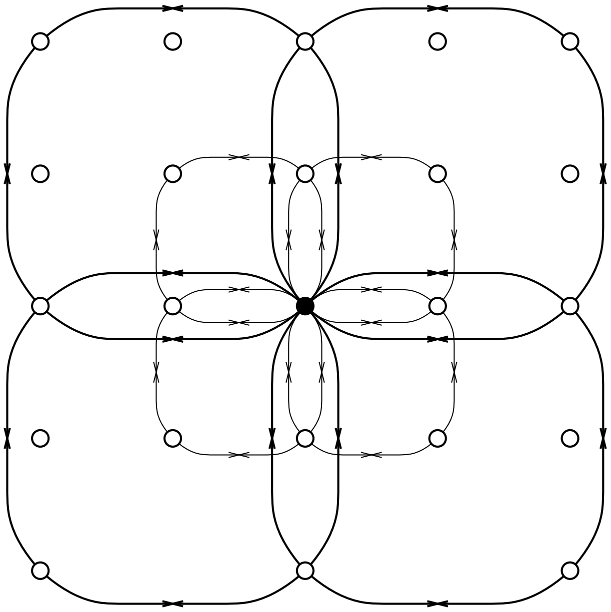

図 3: RTM の基本形において、あるノードが持つレベル 0 と 1 のリンク

完全 RTM は、結合が非常に密であり、様々な結合網を内蔵している。この完全形から一定の規則にしたがってリングを削減しても、なお結合網として十分な性質を持つようにすることができる。例えば、各ノードが、レベル 0 の R+リングと R-リングのいずれかもしくは両方を持つ結合網は 2 次元メッシュである。

3.2 Hypercube Ring on mesh

RTM はハイパーキューブも含んでいる。

レベルlについて、$(x \bmod 2^{l+1}) < 2^l$かつ$(y \bmod 2^{l+1}) < 2^l$ を満たす x,y に対するリング $R+(l,x,y)$ と $R-(l,x,y)$ のいずれかもしくは両方を持つ RTM のリング結合は、ハイパーキューブを構成する特殊例である。ここではこの結合を HR(Hypercube Ring) と呼ぶ。

基本となる mesh/torus 結合を残すために、レベル 0 についてのみ完全 RTM の全てのリングを持ち、レベル 1 以上は上記の条件を見たすリングのみを持つ結合網を Hypercube Ring on mesh(HRM) と呼ぶ。ノード数 4×4 の HRM を図 4に示す。このリングは 4 次元のハイパーキューブとなる。一般に HRM は偶数次元のハイパーキューブを持つ。

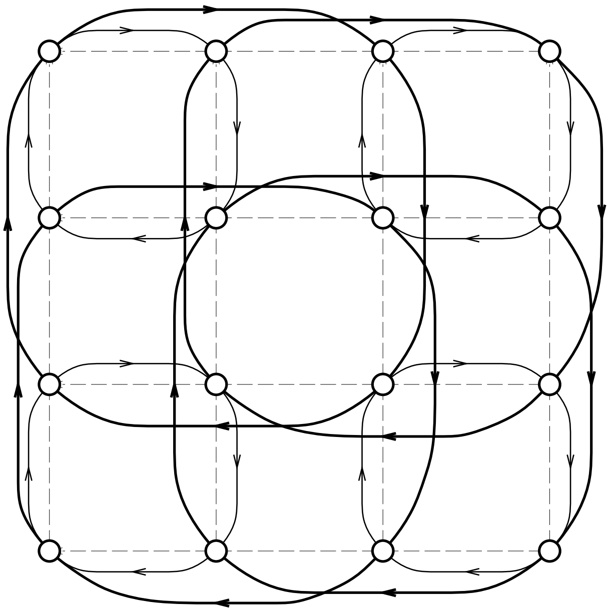

図 4: 4×4 の HRM

HRM のレベル 1 リング結合部はハイパーキューブであるので、近接ノードを除いて、ルーティングはハイパーキューブのルーティングを用いることができる。また、ひとつのリングを疑似的に一つの節と考えると、HRM は fat tree を構成している。

3.2.1 直径と平均距離

HRM は hypercube を持ち、リング結合部のみの直径と平均距離は以下のようになる。

- $R+$,$R-$のいずれかのみのリングから構成される時、$2^n \times 2^n$プロセッサからなる HRM の任意の 2 点間の平均距離は $1.5n$、直径は $3n$ である。
- $R+$,$R-$の双方のリングから構成される時、$2^n \times 2^n$プロセッサからなる HRM の任意の 2 点間の平均距離は n、直径は $2n$ である。

3.3 クラスタ化による degree の低減

完全 RTM では degree が $4n$(同一ノードとの結合は一つのリンクと考える) と非現実的に大きく、リングを削除した HRM でもなお degree が大き過ぎる。そこで複数のノードを組み合わせてクラスタとし、クラスタを基本形の一つのノードと考える手法が考えられる。クラスタ内の各ノードは、分担してそれぞれいくつかのレベルのリンクを持つ。この時、mesh 結合を残すため、レベル 0 のリングは残すことを前提とし、クラスタの形状も mesh を基本とする。ここでは上下左右に隣接する 4 つのノードでクラスタを構成する場合について考える。

図 5に HRM をクラスタ化した例を示す。ここでは黒く塗りつぶした 4 つのノードが一つのクラスタを構成しており、このクラスタに接続されるリング以外は省略している。また、一つのクラスタを構成する 4 つのノードもリングによって結合している。この構成では、mesh 結合部も含めて degree は 6 である。

この構成では、リング $R+(l,x,y)$ は

$l = 0$: $(x \bmod 2) = 0$ かつ $(y \bmod 2) = 0$ ならば

$$(x,y) \rightarrow (x,y+1) \rightarrow (x+1,y+1)$$

$$\rightarrow (x+1,y) \rightarrow (x,y)$$

$l \geq 1$: $(x/2 \bmod 2^{l+1}) < 2^l$
かつ $(y/2 \bmod 2^{l+1}) < 2^l$

かつ $x \bmod 2 = \lfloor ((l-1) \bmod 4)/2 \rfloor$
かつ $y \bmod 2 = \lfloor (l \bmod 4)/2 \rfloor$ ならば

$$(x,y) \to (x, y+2^l) \to (x+2^l, y+2^l)$$

$$\to (x+2^l, y) \to (x,y)$$

によって与えられ、$R-(l,x,y)$ は

$l = 0$: $(x \bmod 2) = 0$ かつ $(y \bmod 2) = 0$ ならば

$$(x,y) \to (x+1, y) \to (x+1, y+1)$$

$$\to (x, y+1) \to (x,y)$$

$l \geq 1$: $(x \bmod 2^{l+1}) < 2^l$
かつ $(y \bmod 2^{l+1}) < 2^l$
かつ $x \bmod 2 = \lfloor (l \bmod 4)/2 \rfloor$
かつ $y \bmod 2 = \lfloor ((l-1) \bmod 4)/2 \rfloor$ ならば

$$(x,y) \to (x+2^l, y) \to (x+2^l, y+2^l)$$

$$\to (x, y+2^l) \to (x,y)$$

で与えられる。

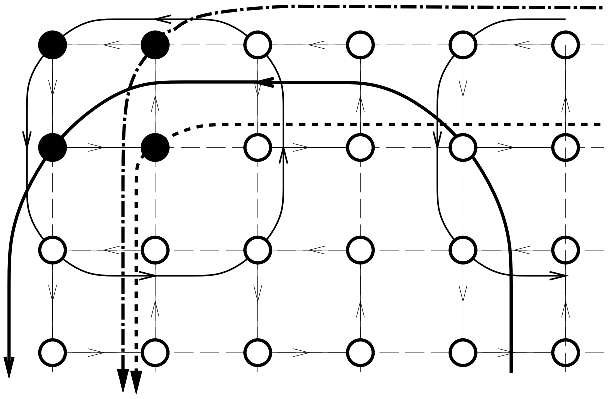

図 5: クラスタ化した HRM

一つのノードで 2 つのレベルを担当すれば degree は 8 となる。この時、4 ノードで 1 クラスタの構成で 8 レベルを持つことができ、最大クラスタ数は $(2^8)^2 = 65,536$、ノード数は 262,144 となり、超並列計算機用結合網として十分な規模を持つと考えられる。

3.4 Double Hypercube Ring on mesh

先の節に示した HRM およびクラスタ化した HRM は、mesh 結合を維持するためにレベル 0 のリングを特殊なものとして扱う必要があった。そこで、2 つのクラスタ化した HR をノード座標で 1 ずつずらした位置に配置した結合網を考える。図 6に、クラスタ化した HR で $R+$方向のリングを持つものと、$R-$方向を持つものをずらした位置に置いた例を示す。ここでは、黒く塗りつぶしたノードおよび網掛けで示したノードが持つレベル 3 までのリングのみを示している。

この結合では、あるノードは、隣合う逆方向のリングの両方に属することになり、単方向の mesh と単方向の Hypercube が内蔵される。この結合を Double Hypercube Ring on mesh(DHRM) と呼ぶ。

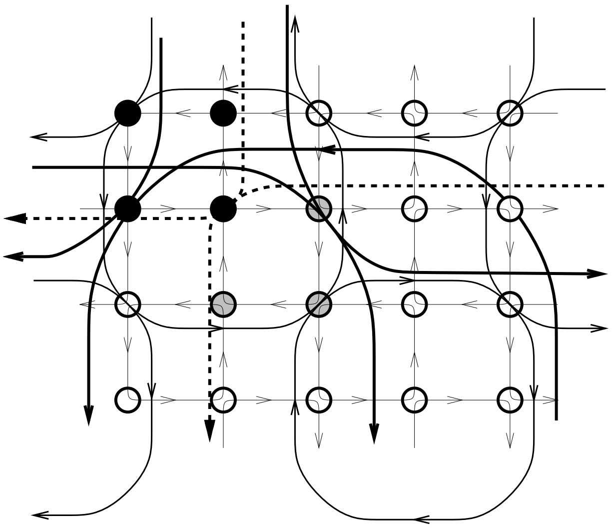

図 6: DHRM

DHRM は mesh 状の結合を含むため、mesh 結合をリング結合と分けて考えなくとも mesh に適したアルゴリズムを持つ問題を効率的に扱え、ルーティングや virtual channel の構成の面で有利である。一方で l が 4 を超えると degree が 12 となる欠点を持っている。

4 おわりに

本報告では、粗さの異なる同一構造の結合網を組み合わせることにより Fat Tree を構築する手法について述べた。

Mesh または Torus 網により結合されたノード間を大きさの異なる複数のリングにより結合した強化メッシュネットワークである RTM を提案し、その一種である HRM と DHRM についても触れた。DHRM はリング状のルーティングの組合せで mesh/torus に適した問題を扱えるという性質を持っている。

RTM は、リングにより構成されるためデッドロックフリールーティングの実現が容易であると考えられる。HRM はリング部がハイパーキューブ構造を持つためルーティングアルゴリズムもハイパーキューブに従うが、DHRM などのそのほかの RTM のルーティングアルゴリズムの開発は今後の課題である。

謝辞

有意義なコメントを頂いた慶應義塾大学理工学部天野研究室の揚愚魯氏に感謝する。

参考文献

[1] 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 2 回シンポジウム予稿集. 1993.

[2] 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 4 回シンポジウム予稿集. 1994.

[3] D.A.Carlson. The mesh with a global mesh. In *1st SCS*, 1985.

[4] Y. Yang, H. Amano, H. Shibamura, and T.Sueyoshi. Recursive diagonal torus: An interconnection network for massively parallel computers. In *Proc. of 1993 IEEE Symposium on Parallel and Distributed Processing*, 1993.

[5] 岩崎一彦, イゼリクリスチャン, 佐藤裕二. 超並列計算機用 vlsi に適した結合網の一提案. 信学論, Vol.J75-DI No.8, 1992.

[6] 岩崎一彦, 志村信幸. 並列計算機に適した格子らせんネットワークの提案. 信学技報, FTS94-4, 1994.

[7] 吉松敬仁郎, 安浦寛人. 超並列計算機のための多重階層メッシュネットワーク. 信学技報, CPSY 94-48, 1994.

[8] 秋山知之, 小池汎平, 田中英彦. 分散共有メモリ型超並列計算機におけるディレクトリ方式と相互結合網について. 信学技報, CPSY 94-49, 1994.

[9] 松本尚, 平木敬. 超並列計算機上の共有メモリアーキテクチャ. 信学技報, CPSY 92-36, 1992.

[10] 松本尚, 平木敬. Memory-based processor による分散共有メモリ. 並列処理シンポジウム JSPP'93 論文集, 1993.