Fat-Tree型相互結合網の設計

2L－3

岩本 邦生　高橋 義造
徳島大学工学部知能情報工学科

1．はじめに

これまで、様々な相互結合網が提案されてきたが、我々の研究室では木構造について研究を行なっている。木構造の欠点としてはルート付近で発生するボトルネックがある。ところがFat-Treeではルート付近でのボトルネックの発生を抑えることができ、並列計算機を構成するのに優れている。しかし、Fat-Treeでも、プロセッサエレメント（PE）数が増加すると通信距離が大きくなる。そこで、直径および通信距離が小さく無閉塞である相互結合網について考察する。

2．Fat-Tree

2．1　完全FT

以下に使用する記号を定義する。

N:　1レベル当たりのノード数
P:　リーフ数＝PE数
D:　通信距離
L:　レベル数
d:　次数

Fat-Tree（FT）は原論文[1]によると、木構造をベースとし、リーフからルートに向かってレベルが上がるにつれて、結合網のチャネル数が増えていくものでありPEはリーフにのみ存在すると定義されている。そして、この結果、木構造の欠点であるルート付近での通信のボトルネックが解消される。なお、リーフからルートに向かいチャネル数が指数的に増加するFTを完全FTと呼ぶ（図1）。完全FTの、通信距離は木構造と同じである。

2．2　ラテン方陣FT

1つ上位のレベルのルータを経由して2ステップでメッセージ通信が可能であるNの最大数は

$$N=d*(d-1)+1 \qquad \cdots\cdots\cdots(1)$$

となる[2]。そして、これらのルータ間の接続方法はラテン方陣を用いることにより決定される。以後この接続方法をラテン方陣接続と呼び、この接続方法を用いて構成されるFTをラテン方陣FTと呼ぶ（図2）。このラテン方陣FTは、少ないレベル数でより多くのPEを接続できるため、同じ台数のPEを用いて並列計算機を構成する場合、木構造より、通信距離は小さくなる。ところが、この結合方法では完全FTとは異なり、無閉塞性が保たれない。

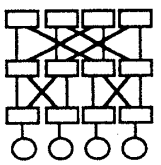

図1．完全FT (d=2,L=3)

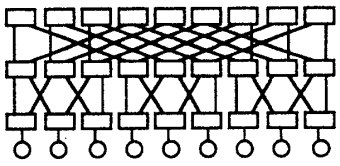

図2．ラテン方陣FT (d=2,L=3)

2．3　混在型FT

ラテン方陣FTは通信距離を小さくすることができるので、これを基にして無閉塞性を保つようにするために、完全FTとラテン方陣FTを交互に用いて相互結合網を構成する方法を考察した。図3のように異なる結合方法を交互に繰り返すことにより、一方の短所をもう一方の長所が補うことができる。その結果、同一プロセッサ台数で完全FTより通信距離が小さくて、ラテン方陣FTにあった閉塞性が解消された。よって、この混在型FTが木構造の中ではより優れた性能であると期待できる。

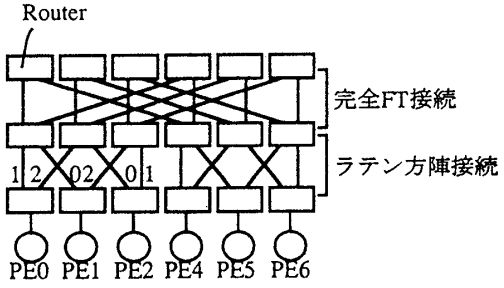

図3．混在型FT（d=2,L=3）

3．評価

3．1　プロセッサ数

完全FTのPE数は

$$P=d^{L} \qquad \cdots\cdots\cdots(2)$$

であり、ラテン方陣FTのPE数は

$$P=(d*(d-1)+1)^{L} \qquad \cdots\cdots\cdots(3)$$

であり、混在型FTのPE数は

$$P=d^{\frac{L}{2}}*(d*(d-1)+1)^{\frac{L}{2}} \qquad \cdots\cdots\cdots(4)$$

である。それぞれ、(2),(3),(4)式より、同一レベルでプロセッサ数を比較するとラテン方陣FTのPE数が多いことが分かる。また、(2)式は$O(d)$であり、

Design of an Interconnection Network Using Fat-Tree.
Kunio IWAMOTO and Yoshizo TAKAHASHI, Department of Information Science and Intelligent System, University of Tokushima.

(3)式は$O(d^2)$であるので次数が大きいほどラテン方陣結合の方がより多くのPEを接続できる。

3．2　レベル数と通信距離

完全FTの通信距離は

$$D=\frac{\sum_{i=1}^{L}\left(2*i*d^{i}*\frac{3}{4}\right)}{P_L-1} \quad \cdots\cdots(5)$$

であり、ラテン方陣FTの通信距離は

$$D=\frac{\sum_{i=2}^{L}\left(2*i*P_i*\frac{\left(d^2-d\right)}{\left(d^2-d+1\right)}\right)}{P_L-1} \quad \cdots\cdots(6)$$

であり、混在型FTの通信距離は

$$D=\frac{\sum_{i=2}^{L}\left(((i+1)\%2)*\left(2*i*\left(d^2-d\right)*P_{i-1}\right)+(i\%2)*(2*i*(d-1)*P_{i-1})\right)}{P-1} \quad \cdots\cdots(7)$$

である。木構造の結合網の場合、同じレベル数ならば、その相互結合網の平均通信距離には、あまり差はない。なぜなら、レベルが一つ上がるにつれて増加するPE数の全体に対する割合がほぼ同じだからである。

3．3　プロセッサ数と通信距離

ラテン方陣FTは、完全FTに比べて小さなレベルで多数のPEで接続できる。レベル数が小さいということは、通信距離が小さいということなので、ラテン方陣FTはPEの数に対して通信距離が小さい。

表1．ノード数と通信距離の比較

	ノード数	通信距離
完全FT	256	14.063
	1024	18.020
ラテン方陣FT	243	11.041
	2187	15.006
混在型FT	216	12.456
	1296	16.412
トーラス結合	256	8.000
	1296	18.000

3．4　他の相互結語網との比較

木構造以外の相互結合網として、k-ary n-cube(n=2)のトーラス結合と比較すると、1000台規模までだと、トーラス結合の方がPE間の平均距離は短いのだが、これを越える場合、トーラス結合よりも通信距離は小さくなっており、将来望まれる大規模並列計算機を実現するに非常に有効な相互結合網となっている。

4．ルーティング法

完全FTのルーティング法は送信先をサブトリーに含まれるレベルまでメッセージを送信し、その後は送信先のPEに送信する。メッセージを上位レベルに送信する際には使用されていない経路を選択すれば良い。

ラテン方陣FTのルーティング法は、送信元と送信先のPE間にはただ一つのみの経路しか存在しないので、経路の選択の余地がない。

混在型FTのルーティング法は完全FTとラテン方陣FTのルーティング法の両方を使用する。完全FT接続されているところでは、経路の選択は使用されていないところで良いのだが、ラテン方陣FT接続されているレベル間では、送信先によって、図1のように使用する経路は決定されているので、その経路を必ず使用しなければならない。

そこで、ルーティングの際、経路を決定しやすくするためにPEのアドレスは図4のように付ける。

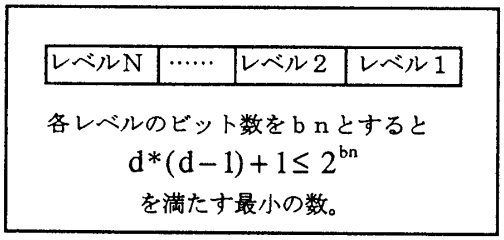

図4．PEのアドレス

これらより、混在型FTは完全FTよりPEの通信の際に使用できる経路が少ないのであるが、無閉塞であるので相互結合網として有効である。

5．おわりに

今回提案した相互結合網は将来大規模並列計算機を実現するにあたりPE数が多くなればなるほど有効なものとなっている。ただし、これを実現するに当たり、2種類のルータが必要になるが、これらのルータを共用できるものを開発することが必要である。

参考文献

[1]. C. E. Leiserson: "Fat-Tree: Universal Network for Hardware-Efficient Supercomputing ", IEEE Trans. on Computer, Vol. C-34, No.10, pp.892-901 (1985).

[2]. M.Valerio, L.E.Moser, P.M. Melliar-Smith:"Using Fat-Trees to Maximize the Number of Processors in a Massively Parallel Computer", ICPADS'93, pp.128-135 (1993).