計算機アーキテクチャ　95－5
(1992. 8. 19)

トーラス・リング・バス階層アーキテクチャ

伊藤　拓　　小畑　正貴　　宮垣　嘉也
岡山理科大学工学部

数万台規模の並列計算機の実装においては、そのネットワーク量とコネクタなどのピンヘッドの削減がコストと共に実装可否の重要な条件となってくる。これらの問題に対して、接続ネットワークを削減、局所化するクラスタ構造の導入が有効であるが、この手法は一般的にクラスタ内外での通信性能に不均衡を生じる事から通信性能の低下の問題がある。本稿では二次元トーラスとリングを用いたクラスタ構造にさらにバスによる通信径路の補強を行なった階層型ネットワークを提案しその基本的な性能について述べ、またこのネットワークが数万台程度のＰＥ数による近接データを用いた数値シミュレーションに対しても有効性がある事を示す。

THE DESIGN AND EVALUATION OF TORUS-RING-BUS NETWORK

Taku Ito　　Masaki Kohata　　Yoshiya Miyagaki
Okayama University of Science

The increase of the number of interconnection wiers is one of the serious problem in the impremantation of large-scale multi-processor systems. In regard to this problem we propose a new cluster structure network called Torus-Ring-Bus Network(TRB). In a cluster, processors are connected by both of ring and bus, and the clusters are connected by torus network. This paper shows the network construction and evaluates the abirilty.

１．はじめに

近年では、数値シミュレーションなどに代表される大規模な科学技術計算を高速に実行するコンピューターとして数千から数万台の規模のＰＥ数を擁する並列計算機に対する期待も高まりつつある。しかしながら数万台規模の並列計算機ではネットワーク量や基板間の接続コネクタのピンヘッドが、そのコストと合せて実装可否の重要な条件となってくると言える。

このような問題に対してネットワークとハードウェアコストを抑えつつ、多数のＰＥを実装するアプローチとして複数のプロセッサをローカルに接続して１つのネットワークノードとするクラスタ構造の導入が有効であると考えられるが、この手法の問題点としてはクラスタの内外で通信性能に不均衡を生じる点や複数のクラスタ間にまたがる通信を必用とするプロセスの配置に対して柔軟さに乏しい点などが上げられる。

本稿ではリングとバスで接続されるＰＥ群を１クラスタとし、そのクラスタを二次元トーラスで接続した階層型のネットワーク（ＴＲＢ）を提案し、その構成と基本的な性能について考察する。

２．ネットワークの構成と実装

２．１ネットワークの構成

図１にＴＲＢによって結合された並列プロセッサの構成例を、図２にそのクラスタ内部の構成例を示す。

図１ではＮ個のＰＥを双方向のリング、および２本のバスによって接続したものを１クラスタとし、各々のクラスタではバスに接続されているＴＣ(Torus Controler) を介してＮ×Ｎの２次元のトーラスネットワークに結合される。ＴＣは主に２次元トーラスからくる４本のチャネルとクラスタ内部のバス間での通信のルーティング及び、クラスタ単位でＰＥ間の同期を行なう。

各ＰＥの内部は図３に示すような構成となりＲＣ（Ring Controler）を介してクラスタ内のリングに接続される。通常クラスタ内部の通信はＲＣを介して隣接ＰＥ間でリングによって行ない異なるクラスタへの通信は、バスを介してＴＣによってルーティングされる。

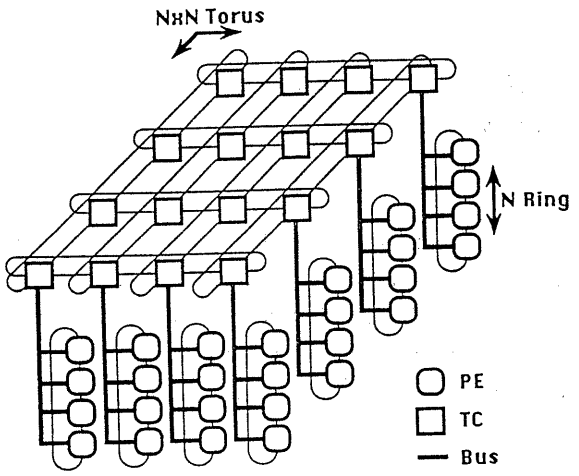


図１．ＴＲＢの構成

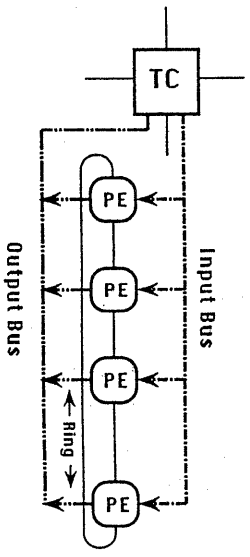


図２．クラスタ構成

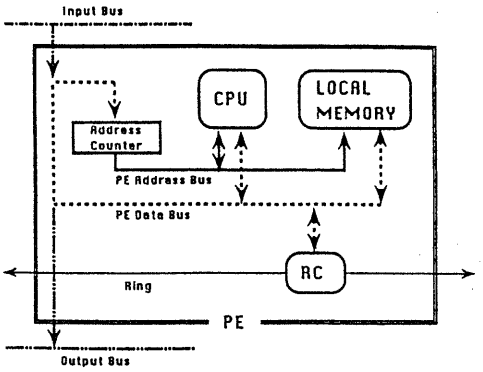


図３．ＰＥの内部構成

ＰＥ台数	Torus2D	Torus3D	ＴＲＢ	
()TRB一辺	Link	Link	Link	Bus
64 (4)	128	192	96	32
729 (9)	1458	2187	891	162
4096 (16)	8192	12288	4608	512
15625 (25)	31250	46875	16875	1250
46656 (36)	93312	139968	49248	2592

表１．通信ネットワーク量の比較

図１、図２は共にクラスタ内ＰＥ数と、トーラスの１辺を４とおいた６４プロセッサの構成例であるが、これ以外でも様々な組合せがありえる。また、トーラスの通信路が４本である事からクラスタ内部のバス数をそれに合わせる構成も考えられる。これについては、実装上でＴＣのハードウェア量と配線コストの増大を招く為、実装アプリケーションの性能向上とコストの両面で検討する必用がある。

２．２ハードウェア量と実装

ＴＲＢでは総ＰＥ数をＮとした場合の各ＰＥを接続するリンク数は$2N^{2/3}+N$となり、この時のバスは$2N^{2/3}$本となる。総ＰＥ数Ｎを等しくおいた場合の通信ネットワーク量をＴＲＢと２次元トーラスの２Ｎ及び、３次元トーラスの３Ｎとで比較し表１に示す。

実際の実装に関してはＴＲＢではバスによるクラスタ内結合の限界と、ランダムトラフィックで密な通信を行なった場合、通信路の共有による性能低下が顕著になってくると考えられる事から無制限にＰＥ数を増やす事は無理である。しかしクラスタ化によって配線が局所化され、リンク数も少ない事からコネクタピンヘッドや配線面では有利であり、クラスタ数とトーラスの１辺が等しくｎの場合の実装ＰＥ数n^3 よりｎ＝１６で４０９６台，ｎ＝３２で３２７６７台程度のＰＥ数での実装が可能であると考えられる。

３．ネットワークの通信特性

３．１ランダムトラフィックにおける通信性能

ＴＲＢでは基本的なポイントツーポイントでの平均的な通信距離はＰＥの総数をＮとして、クラスタ内をD_R とクラスタ外D_T をとおくと、$D_R \simeq N^{1/3}/2$，$D_T \simeq N^{1/3}/2+2$。また、システム全体ではクラスタ内外の通信比率、$N^{1/3} : N-N^{1/3}$によって平均の通信距離Ｄは次の様になる。

$$D \simeq \frac{N^{1/3}}{2}+2-\frac{2}{N^{2/3}} \simeq \frac{N^{1/3}}{2}+2$$

ランダムトラフィックな通信を行なう場合、複数の通信の競合が通信性能に及ぼす影響が深刻である為、これを調べる目的でＣ言語による通信シミュレータを作成し実験を行なった。

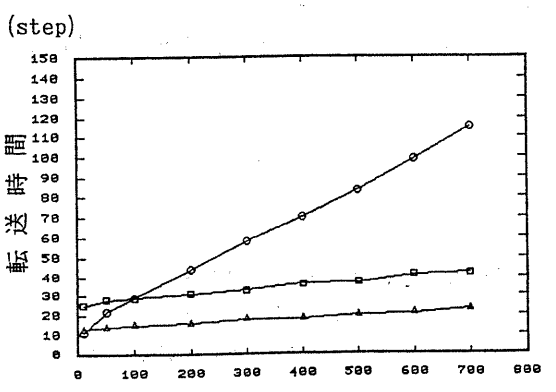
通信シミュレータでは以下の条件でのランダムなポイントツーポイントの通信を行なう。

- ●ネットワーク上の各ノード（ＴＣを含む）は１単位時間当り１パケットの送信と受信が可能である。
- ●通信の形態は不規則で局所性を持たない通信を基本とする。
- ●送信ＰＥと受信ＰＥが同一のクラスタである場合のルーティングはリングを介してのみ行なわれ、また目的ＰＥに対して常に近ずく方向にルーティングされる。
- ●ＴＣによるルーティングでは目的クラスタに常に近ずく方向でルーティングされ、また近ずく径路が複数存在し一方が使用中であれば他方に対してルーティングされる。
- ●送信径路が全て使用中であるパケットはＴＣ、ＲＣ共にブロックされる。

以上の条件のもとにある時点で複数のＰＥから同時に発生したパケットが全て目的のＰＥまで到達するまでの通信時間をパケット収拾時間とし、異なるネットワークでの通信性能の指標として比較を行なった。７２９台のＰＥでのパケット収拾時間を図４に示す。またクラスタ内外での通信比率が変化した場合のシミュレーション結果を図５に示す。なお、これらのグラフはＰＥ台数が変化してもパケット量とＰＥ台数の比によって同様な形状の特性を示した。

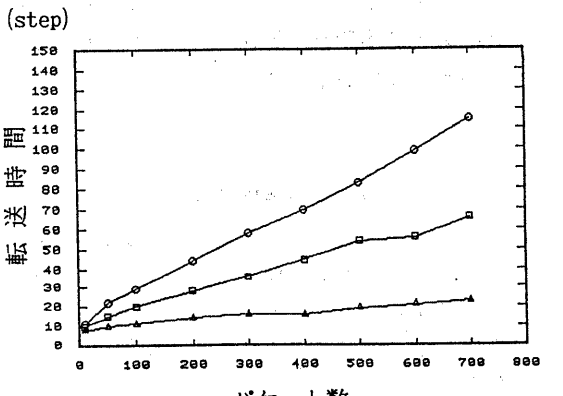
図４から、ＴＲＢが１００パケット以下の領域で２次元トーラスに対して優位である事がわかる。これはネットワーク上でのパケットの競合が少ない場合にＰＥ台数を同じにとるとＴＲＢが２次元トーラスより平均的な通信距離が短い事によるものである。しかし、ランダムトラフィックにおいてはパケットが密になってくるとＴＲＢはトーラスの場合と異なり通信性能の低下が著しい。

図５ではクラスタ内外における通信比率をそれぞれ、内：外で１：９，５：５，９：１とした場合の通信性能の変化を調べたものである。



○ TRB ，□ Torus-2D ，△ Torus-3D

図４．同時発生パケットと収拾時間
(729 Processor)



In Out In Out In Out

○ １：９，□ ５：５，△ ９：１

図５．クラスタ内外での通信比率による収拾時間の短縮 (729 Processor)

図より、クラスタ内の通信割合いが大きくなるほどパケット収拾時間が減少しており、通信性能が向上する事がわかる。これはクラスタ内の通信割合を増加し、クラスタ外への通信が減少させる事によって、クラスタ外ネットワークの競合が減少し、クラスタ単位で通信多重度が上昇する事からも明らかである。この特性は一般にクラスタ構造によって複数プロセッサをローカルに接続し、グローバルなネットワーク量を削減したネットワークでは共通するものである。

３．２近接通信における通信性能

ＴＲＢは基本的に１辺をｎとする３次元トーラスの２軸方向のネットワークを１／ｎに削減しバスで括った構成となる。この構成での近接通信におけるＴＲＢの性能を２次元、３次元の近接通信をそれぞれ例にとって考える。

３次元の近接通信において、３次元トーラスの一辺とＴＲＢのクラスタ内のＰＥ数を共にｎとした時、３次元トーラスにおける東、西、南、北、上、下の６リンクからのデータ通信ステップを６ステップとすると、同等な通信条件でＴＲＢでは２本のバスを用いて４ｎ＋２ステップとなり、３次元トーラスに対して３／（２ｎ＋１）倍の性能になる。また２次元の近接通信においては、２次元トーラスの４リンクに対して東、西、南、北に同様な通信を仮定した場合、２次元トーラスの４ステップに対して、ＴＲＢでは２本のバスとリングの併用によって図６のように１軸方向のリングを複数クラスタ間に跨って構成した場合でｎ＋２ステップとなる事から２次元トーラスに対するＴＲＢ通信性能は、２／（ｎ＋１）倍となる。この場合のＴＲＢは図７に示す様なｎ²×ｎ の長方形型の２次元トーラスを仮想している。

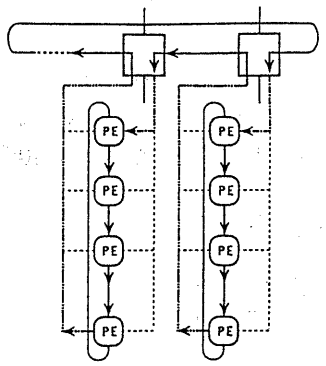


図６．複数クラスタ間に跨るリングの構成

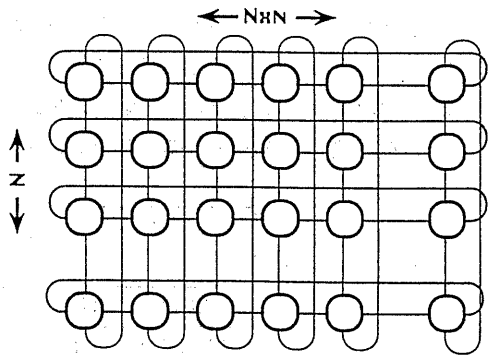


図７．ＴＲＢで構成される仮想的なｎ²×ｎの二次元トーラス

４．応用プログラムでの性能評価

４．１システムの応用分野

一般に並列プロセッサの応用範囲としては、数値シミュレーション、グラフィクス、人工知能、データベースと言ったように様々な分野が期待されている。なかでも数値シミュレーションは静的な負荷分散が比較的容易で、データ並列性が高いと言った利点から並列処理の応用分野としての期待が高まっている分野である。

ここでは数値シミュレーションの中でも最も多用され、また応用範囲も広いと考えられる偏微分方程式の数値的求解を基本とする近接作用問題を例にとって、問題の処理粒度、負荷分散、ＴＲＢ

上での通信性能等を考慮して実装時の台数効果を見積る。

近接作用問題とは連続空間に分布する温度や圧力といった物理量が、場を支配する偏微分方程式の形で表わす事が可能である場合に、空間を格子点状に分割し偏微分方程式を差分近似等で全ての格子点間の連立１次方程式の形に変形した後、この連立方程式を数値的に解く事で新たな分布を求めると言う問題である。この問題は各格子点の物理量がその微係数と近接する格子点の物理量によって計算できると言う特徴を持っている。

この様な問題の計算処理はその主体が一般的に対角優位な疎行列となる連立１次方程式の求解である事から SOR法などに代表される緩和解法によって収束解を求める手法に適している。しかし並列処理化に関しては緩和段階のループ処理にデータ依存関係がある事から、依存関係の無い奇数点の点緩和と偶数点の点緩和をそれぞれブロックにまとめて並列化する Red & Black SOR法やSLOR法などが用いられる場合が多い。

４．２近接作用問題での性能見積り

前章ではＴＲＢとトーラスの通信性能を近接通信で比較したが、実際に応用問題を実装して処理する場合に、その実行性能におよぼす影響は、各ＰＥにおける処理の粒度とばらつき、通信量、通信速度、問題の規模、ＰＥ台数と言った種々の条件の総合によって決ってくるものである。したがってこれらを定量化する事が可能であれば台数効果を見積る事ができる。以下にそれぞれの条件の見積り方法とその値を示す。

①処理粒度：境界条件を考慮にいれた緩和法の差分式１ループに要するＣＰＵの処理時間をＣ言語でコーディングし実測して用いた。対象ＣＰＵは、ＲＩＳＣチップとしてはＳＰＡＲＣ２５ＭＨｚ，ＤＳＰ ＴＭＳ３２０Ｃ３０ ４０ＭＨｚ、汎用ＣＰＵでは８０３８６＋３８７ ２０ＭＨｚの三つで、測定結果を表２に示す。ただしＳＰＡＲＣに関しては割り込みを止めていない為、やや厳密性に問題があるが、最近のＲＩＳＣチップの処理速度の指標とした。

ＣＰＵ	処理時間(μｓ)	
	２次元	３次元
SPARC 40MHz	7.0	9.0
TMS320C30 40MHz	13.8	15.9
80386+387 20MHz	76.0	89.2

表２．緩和法１ループの処理時間

②処理のばらつき：境界条件の定義のしかたとコーディング方法によってやや変化するかも知れないがほとんど無視できると考える。

③通信量：一般的には２次元問題で自分自身を含む近接の５点、３次元問題では７点の格子点値である為、１ループ、１格子点毎にそれぞれ４ﾜｰﾄﾞ、６ﾜｰﾄﾞの通信量を必用とする。

④通信速度：現在６４プロセッサによるＴＲＢのプロトタイプの設計を行なっており、１リンクあたりに通信速度は５～１０M(ﾜｰﾄﾞ/sec.)程度となる予定である。また１通信あたりの同期処理とＤＭＡの駆動の為のオーバーヘッドとして２(μｓ)程度を想定している。

⑤負荷分散：格子点をブロック単位でＰＥに分割し隣接する格子点をできるだけＰＥ内部に集める。この分割によって通信が必用となるのはブロック分割の境界面に位置する格子点のみとなる。

⑥問題の規模：流体計算などの分野で慣性項が大きい計算では少なくとも３次元で８００万格子点、２次元で８万格子点程度の規模が必用と言われている事から[4]この値を問題の規模とする。しかし２次元問題の場合は数万台で実行される場合、各ＰＥの負荷が小さくなり過ぎる為、⑤の通信量の削減が殆ど行なえなくなる事から通信時間の割合が増大してくる。したがって、この規模ではＴＲＢによる数万台での実行には不向きである。

以上の条件から３次元問題に対して台数効果の見積りを行なった。この時の格子点数は８００万であり、また並列化を行なっても点緩和の収束回数は殆ど変動しないものと仮定している。ＣＰＵがＳＰＡＲＣ、ＤＳＰ、８０３８６のそれぞれの場合で通信速度が５Ｍ(w/sec.)，１０Ｍ(w/sec.)の二通りについて求めてみた結果をそれぞれ図８、図９、図１０に示す。ＳＰＡＲＣでは3万台付近で台数効果はやや低下し５Ｍ(w/sec.)の場合で約６３(%)、１０Ｍ(w/sec.)で約７４(%)程度の台数効果である。しかし、いずれの場合でもこの問題では３万台程度までのＰＥ台数において比較的高い台数効果が得られている事がわかる。これより、３次元の近接作用問題に対しては、ＴＲＢのようにコストを抑えたネットワークでも数万台規模までであれば十分な有効性が得られる見通しが得られた。

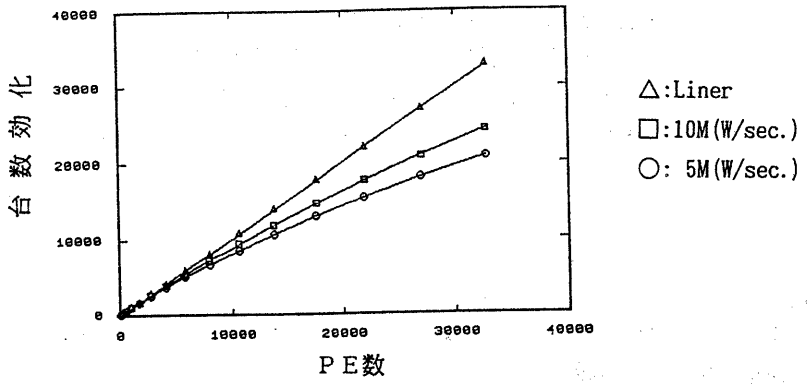


図8．ＳＰＡＲＣでの台数効果

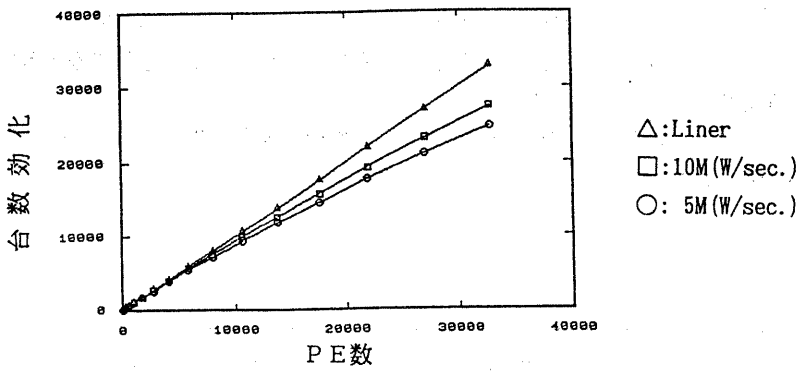


図9．ＴＭＳ３２０Ｃ３０での台数効果

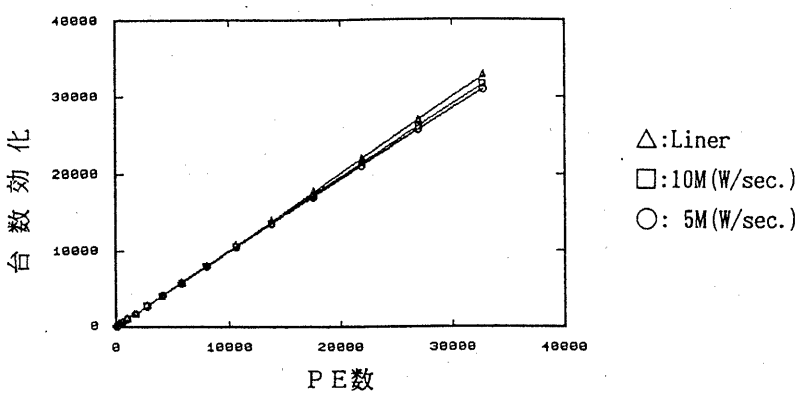


図１０．８０３８６＋３８７での台数効果

５．おわりに

クラスタ化による階層構造の採用によってネットワークのハードウェア量を削減し大規模なＰＥ数での実装を志向するアーキテクチャとして二次元トーラス、リング、及びバスを用いた階層型のネットワークを提案し、その基本的な構成と通信性能について述べた。また比較的小さな処理粒度で頻繁に通信を行なう必要がある大規模な３次元の近接作用問題に対して、このネットワークを用いて数万台規模までは十分に性能向上を見込める事を示した。

ＴＲＢによる並列システムの開発は現在ＣＰＵに３２ビット浮動小数点ＤＳＰ(ＴＭＳ３２０Ｃ３０)*を用い６４プロセッサにからなるプロトタイプの実装を目指して、ＰＥ基板のパターン設計とルータなどの詳細な設計を平行に行なっている段階である。これらについてはまた別の機会に報告する予定である。

＜参考文献＞

１）冨田、末吉：並列処理マシン，オーム社

２）戸川：数値計算，岩波書店

３）大宮司、他：流れの数値シミュレーション，コロナ社

４）桑原：スーパーコンピュータによる流れのシミュレーションとその可視化，JSPP'92招待講演

５）小畑：２５６台ＤＳＰシステムによる並列計算機の実装と評価，電子情報通信学会，コンピュータシステム研究会，CPSY90-29 1990

６）小畑：シグナルプロセッサアレイーＳＰＡー，電子情報通信学会，コンピュータシステム研究会，CPSY79-13 1979

---

*ＤＳＰを提供して頂いている日本テキサス・インスツルメンツ㈱に感謝します。

本研究は平成４年度科学研究費補助重点領域研究(04235210)の補助を受けています。