計算機アーキテクチャ 112－7
（1995．6．23）

トーラス・リング・バス階層型並列マシン

伊藤 拓，小畑 正貴

岡山理科大学 工学部 情報工学科
e-mail: itoh@ice.ous.ac.jp

著者らは、近接通信を用いる並列数値計算を数万台程度の規模で実行する並列システムとして、PEをリングとバスで接続したクラスタを2次元トーラスで階層化するTRB(Torus-Ring-Bus)と呼ぶネットワークを提案している。TRBは3次元トーラスやメッシュ結合のネットワークに対する配線量の削減と数千から数万台規模までの性能の両立を目標とするアーキテクチャである。本稿では48プロセッサで実装を行ったプロトタイプシステムの構成と通信方式について述べ、Red- Black-SORを用いた実測性能から評価を行う。

A Torus-Ring-Bus connected hierarchic parallel machine.

Taku Ito , Masaki Kohata

Faculty of Engineering, Okayama University of Science

The increase of interconnection wires is one of the serious problems at the implementation of parallel processing systems which have over ten thousands of processors. We proposed a cluster architecture called TRB (Torus-Ring-Bus) for these massively parallel systems. This architecture aims at both decrease of interconnection wires and performance in a numerical computation. In this paper, we describe the prototype system which consists of 48 processors, and evaluate the network from performance measurement results in Red-Black-SOR method.

1 はじめに

数千から数万台規模の大規模並列システムの結合ネットワークでは直径、バンド幅、並列度と言った通信性能に加えて、チップや基板間の入出力信号線数や配線数といった実装上の問題が重要な条件となる。

初期の並列計算機で多く用いられたハイパーキューブ結合は、数千ノード以上の接続では結合次数が10以上となりスケーラビリティに問題を生じる。このため、最近の大規模な商用並列システムは、流体や構造解析等の近接通信を用いた科学技術計算の用途に適し、結合次数が固定値であるメッシュ/トーラス構造を基本にするものが多い。

また、最近開発が始まっている超並列システムの接続網としては次の様なものが提案されている。JUMP-1 で採用される RDT(次数 8)[1] は、2 次元トーラスの対角方向に上位トーラスを階層化する構成をとっている。RWC-1 で採用される CCCB(単方向, 次数 6)[2] はサーキュラバンヤン網とハイパーキューブを複合した構成を持っている。また、間接網では CP-PACS でクロスバーネットワークを 3 次元で構成する HXB[3][4] 等が提案されている。これらは、いずれもメッシュ構造のエミュレーションが可能で、かつ科学技術計算の分野以外での使用も目的とする汎用の並列システムを狙ったものである。

一方、ある程度用途を絞ってネットワークの配線量の削減を行いつつ多数のノードを実装するアプローチは、汎用性は多少犠牲とするものの、比較的低コストでの大規模実装も可能と考えられる。著者らはこの様な観点から、近接データを用いた数値計算を主たる応用目的とし、3 次元トーラスの 2 軸方向のネットワークを簡素化した構造を持つ TRB (Torus-Ring-Bus) と呼ぶネットワークを提案している。

本稿では、この TRB のネットワーク構造と特徴、実際に実装を行ったプロトタイプシステムのハードウェア構成と通信機構、通信の制御方式とソフトウェアについて述べる。また非定常流体問題の解析等の並列計算で用いられる Red-Black-SOR 法を例に、実行結果からシステムの性能評価を行う。

2 TRB の構造と特徴

TRB ではリング・バスでローカルに接続された要素プロセッサ (PE) をクラスタとして、クラスタ間を 2D トーラスで接続する 2D トーラス+クラスタの構造を持つ (図 1)。

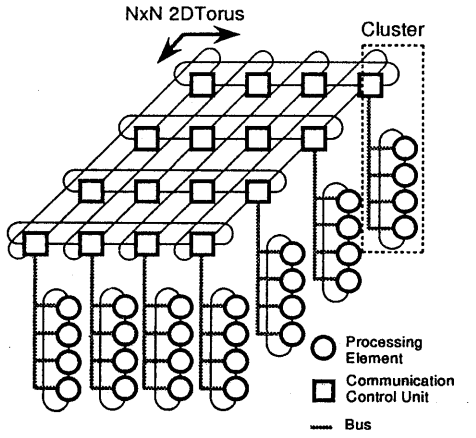


図 1: TRB のネットワーク構成 (64 プロセッサ)

クラスタの内部接続であるリング・バス結合については、著者らは以前より幾つかの並列システムを試作し評価を行ってきており [5][6]、隣接プロセッサ間で多重度の高い通信を容易に実現でき、また結合次数が 3 で実装面で有利などの特徴がある。

TRB においてもクラスタ内部で各 PE は両隣に対する 2 組とバスに対する 1 組の計 3 組のリンクを持てばよく、個々のリンクは多ビット並列化が容易である。また、複数 PE を 1 チップ化する場合でもチップから出る信号はリングの両端とバスの 3 組のみとなり、1 クラスタを 1 ボードに実装する場合でも、ボードからはトーラスのための 4 リンクを引き出せばよい。複数クラスタを 1 ボード化する場合はクラスタ数×2 のリンク信号がでることになるが、この場合でも隣接基板間であるので接続は比較的容易と考えられる。

TRB はこの様に実装面で有利な点が多いが、クラスタ間で共有路となるトーラス上で通信の輻輳による遅延が生ずる。これに対しては、ソフトウェアシミュレータ [7] を作成し評価した所、数万台規模までで、以下の様な特性を持つことが判っている。

(1) 同時発生するパケット数が総 PE 数の約 1/10 以下程度の領域では 2D トーラスと 3D トーラスの中間の性能を示す。

(2) 総 PE 数の約半数が同時にパケットを発生しても 2D トーラスと比べて二倍程度の遅延ですむ。

(3) クラスタ内の通信割合が増大する程 3D トーラスの通信性能に近ずく。

3 プロトタイプの構成

3.1 プロトタイプの設計方針

TRB の実験用システムとして、プロトタイプマシン [8] の設計を次の様な方針で行った。

(1) 近接作用の数値計算問題等アプリケーションの移植性を考慮し、システムを仮想的な 1 次元リング、2 次元トーラス、3 次元トーラス結合の計算機として扱えるプログラミング ·モード (仮想ネットワークモード) を持つ。

(2) バスによってフラグ収集やデータ分配を階層的に行うための支援機能を持つ。

(3) 任意の PE 間を結ぶランダムパケットを扱う。

(4) 通信ネットワークとは別個に計算結果を可視化したデータを外部に出力する機構を設ける。

(5) TRB ではクラスタ間の通信量が多くなるとネットワークの遅延が大きくなるため、できる限りこれを回避する工夫を行う。

プロトタイプの設計にあたっては、これらの要求を満たす機構の総てをワイヤードロジックで実現するのは困難であったため、通信性能を幾分犠牲にはするが、通信コントローラ上に CPU を搭載し、ソフトウェアによる柔軟な対応を図ることとした。

3.2 プロトタイプの構成と実装

現在のプロトタイプの構成は図 2の様に 4PE × 12 クラスタの 48PE によるもので、クラスタ毎に 1 台の通信コントローラー (CCU) を設ける。PE は浮動小数点 DSP である TMS320C30[9] と 128KB の SRAM,4MB の DRAM から成り、4PE と 1CCU を 1 基板とする。CCU はフラットケーブルで 2 本、マザーボード側に 2 本の計 4 本の 32bit 幅リンクを用いて、隣接する CCU に接続され、2 次元トーラスを構成する。

CCU は PE と同様の DSP を持っており、クラスタ内外の通信と同期の制御を行う。CCU では、クラスタ内の同期機構、リングを用いた通信を行う場合の DMA、トーラス間の同期機構を、FPGA(Right) と FPGA(Left) の二つの FPGA に振り分けている (図 3)。

また全 CCU はトーラス以外にホストに対してバスで接続され、ホストから CCU に対しては 1 対 1 のリード/ライトの他にブロードキャストもサポートする。

実装は VME のダブルハイト型ラックで行い、電源と共に 50cm 立方の 1 框体に納めている。現在の

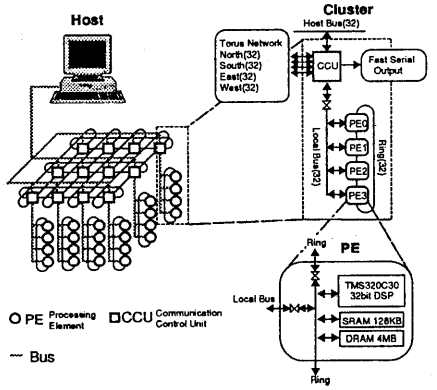


図 2: プロトタイプの構成

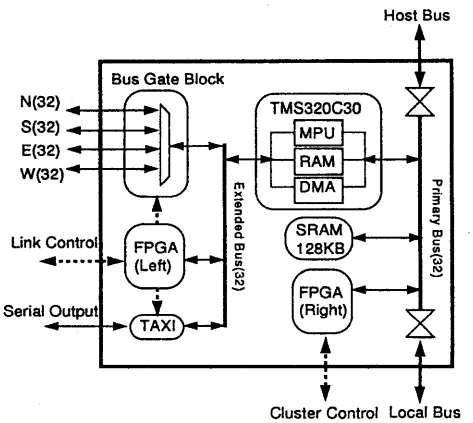


図 3: CCU の内部構成

プロトタイプシステムの外観を図 4に、諸元を表 1 にまとめる。

4 通信機構とソフトウェア

4.1 CCU の構成と通信機構

CCU 内部の FPGA(Left) では、トーラスに対して図5に示す様に 1 リンク毎に 7 本の制御線を持ち、アービタによるリンクの確立と排他制御、ハードウェアハンドシェイクによるデータ転送を行う。

FPGA(Right) ではクラスタ内における PE 間の同期と内部に設けた独立の DMA によってクラスタ

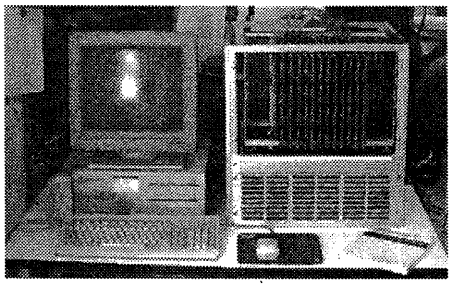


図 4: プロトタイプの外観

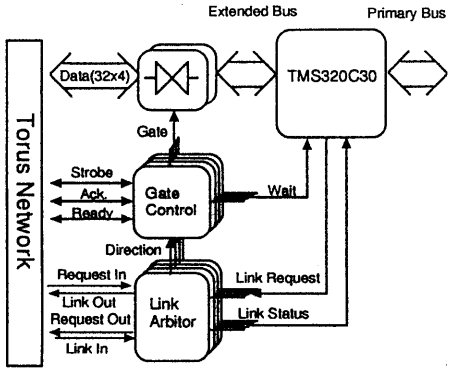


図 5: トーラス側の通信機構

PE	
プロセッサ	TMS320C30(32bit 浮動小数点 DSP)
サイクルタイム	60(ns)
計算性能	16MIPS / 32MFLOPS
メモリ	4.12MB
FPGA	3200Gate-100MHz
CCU	
プロセッサ	PE と同じ
メモリ	128KB
FPGA	4200Gate-100MHz × 2
シリアル出力	TAXI Am7968(125Mbps)
全体	
プロセッサ数	48 + 12
メモリ	199.7MB
結合方式	2D Torus + Ring + Bus

表 1: プロトタイプの諸元

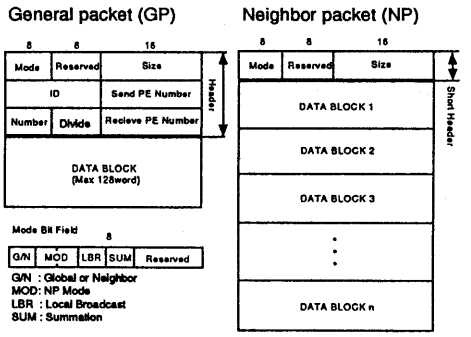


図 6: パケットフォーマット

内でブロードキャスト、リングを介したデータの回転などを行う。

シリアルトランシーバである TAXI チップ[10] は、3 章 1 節 (4) の PE 上で分散計算された可視化のためのデータを、外部に出力する目的で設けている [11]。125MHz の TAXI チップが 12.5Mbyte/s 程度の転送能力を持ち、これはアプリケーションからの要請で CCU 上の DSP で駆動する。

4.2 データパケットと通信方式

プロトタイプのトーラス上で通信を行うためのパケットフォーマットは、3 章 1 節 (3) の一般通信をサポートするパケットと、3 章 1 節 (1) で述べた仮想ネットワークモードでの通信をサポートするパケットの 2 種類の物理的フォーマットが存在する。前者を General Packet(GP) と呼び、後者を Neighbor Packet (NP) と呼ぶ (図 6)。GP は任意の PE 間を結ぶ一般的なパケットで、1 つのデータブロックとルーティング情報を持ったヘッダからなる。NP は隣接するクラスタ間での通信のみに的を絞ったフォーマットを持ち、クラスタ単位でまとめられた複数のデータブロックと 1 ワードのショートヘッダを持っている。

GP のヘッダは 3 ワードで、受信 PE のアドレス、パケットの ID、分割個数である Divide, 転送順位である Number、及びパケットの Size からなる。パケットの Size は現在バッファ容量の関係上 1 ～128 ワードに制限してある。

GP を用いたクラスタ間の通信方式は基本的に Store and forward で行い、この場合の通信デッドロックの回避は、宛先 CCU までの距離による螺旋バッファを構成して、ループ経路の排除を行っている。

NP は、数値シミュレーションの問題を扱う場合によく現れる近接通信に対して、図 7 に示す様に、隣

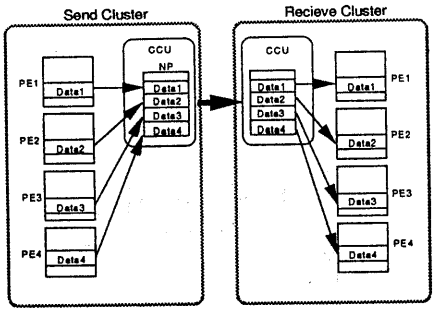


図 7: NP による隣接クラスタ間通信

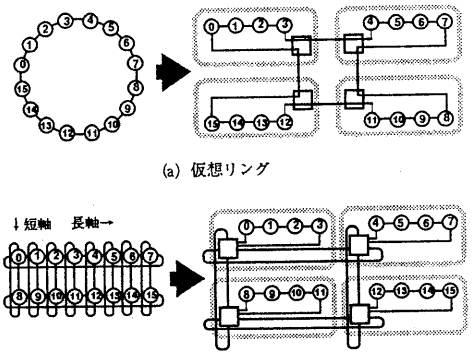


図 8: 仮想ネットワークの PE 割り付け

接クラスタにあらかじめルーティングされることが明らかな複数のパケットを CCU が取りまとめた後、ショートヘッダを付加した1パケットとする。これによって、トーラス上における不要なリンク確立やルーティング処理等のオーバーヘッドを削減する効果を狙っている。

この場合の通信処理は、図 8に示す様な論理的に番号を割り振った PE に対して、各仮想ネットワークの通信をシステムがエミュレーションすることとなる。そのエミュレーション方法には以下の三つのパターンがある。

(1) 図 8(a) の仮想リングモードや図 8(b) の仮想 2 D トーラスモードの長軸方向の通信では、クラスタ内で一旦同期を取り、リングを介してデータを回転させた後、クラスタ間の境界に位置する PE 分のデータをデータブロック数 1 の NP を用いて通信する。

(2) 仮想 2D トーラスモードの短軸方向、図 8(c) 仮想 3D トーラスモードの NS,EW 方向では図 7 に示した複数データブロックを持つ形式の NP で通信を行う。

(3) 図 8(c) の UL 方向の通信はクラスタ内部のリングによるデータの回転のみで通信が終了する。

また、トーラス上で用いられる NP では GP の場合と異なり、リンクが確定した時点で隣接クラスタにルーティングされることがあらかじめ判っているので、CCU 上でのバッファリング処理も割愛する事ができる。

4.3 ソフトウェア

ソフトウェア開発は通信関数を付加した C 言語で行う。アプリケーションに供給される基本的な通信関数を表2にまとめる。一般関数の通信は GP によって行われる。仮想リング, 仮想 2D トーラス, 仮想 3D トーラスの各モードの通信関数を用いた通信はクラスタ内部のリングでのデータの回転等と共に、NP を用いて行われる。また、仮想ネットワークに関してはコンパイル時に静的に決定される形式をとっている。

5 性能評価

プロトタイプシステムの性能評価として、各通信関数における通信性能の測定と、連立方程式の並列緩和求解法の一種である Red-Black-SOR 法で性能測定を行った。

5.1 プロトタイプの通信性能

プロトタイプのハードウエアの通信性能は、ネットワーク上にパケットが存在しない条件で、トーラス上に通信要求が発生して第一ワードが受信側の CCU で読み出されるまでの平均的な遅延が約 2.7 μ s、データの転送レートは 8〜9MB/s 程度である。また PE-CCU 間の転送レートは 12MB/s 程度、クラスタの内部におけるリングでの DMA 転送による回転では 32MB/s となる。

一般通信	Send() RBalloc() RBfree() RBsense() RBflush()	パケット送信 受信バッファ取得 受信バッファ解放 受信バッファステータス取得 受信バッファのフラッシュ
仮想 Ring	Left() Right()	Ring モードでの左右隣接通信
仮想 2D Torus	North() South() East() West()	仮想 2D Torus モードでの 4 方向隣接通信
仮想 3D Torus	North() South() East() West() Upper() Lower()	3D Torus モードでの 6 方向隣接通信

表 2: 基本的なプロセス間通信関数

一般に輻輳による遅延を生ずるネットワークの通信性能は、上述したハードウェア性能に加えて、アプリケーションの転送データ長、メッセージの発生間隔、リンク確立の失敗によるリトライ処理に要する時間等によっても影響を受ける。このため、表2に示した各通信関数による実転送時間を、実際のアプリケーション上から見た通信性能として測定した。測定値はいずれも、各通信モードの関数を用いて1万回連続で100ワードの通信を行った場合の平均である。同一の通信パターンを Send() 関数を用いてGP で行った場合の結果と合わせて図9に示す。

図9の NP において (1) (2) (3) の表記は、それぞれ4章2節の (1) (2) (3) の通信パターンに相当する。仮想ネットワークモードではモード毎にネットワークの使用形態が異なるので、同一名の通信関数であってもリング、2次元トーラス、3次元トーラスモードで異なった性能となる。また、(3) の通信パターンである3次元トーラスモードにおける Upper() 関数、Lower() 関数が飛び抜けて高速であるのは、クラスタ内部のリングを介する DMA 転送のみで処理が終了し、トーラス上の通信を全く必要としないためである。

これらの通信パターンでは、一般通信を行う GP と比べていずれの場合も NP による通信速度向上の効果が現れている。

また、(1) (2) (3) の各隣接通信のパターンにおける通信時間は、転送ワード数を変化させて測定を行ったところ、通信時点の CCU の状態によって多少のばらつきを生じるものの、平均的に見てワード数に

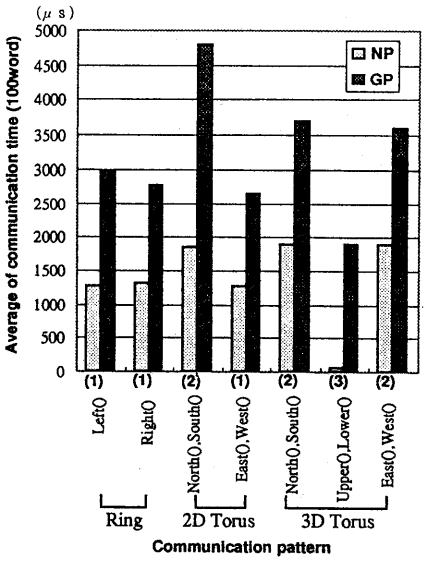


図 9: 各通信パターンでの GP/NP 通信速度比較

通信形式	セットアップ時間 (μs)	転送時間 (μs)
(1)	750	11.0 × ワード数
(2)	680	6.5 × ワード数
(3)	28.4	0.275 × ワード数

表 3: 各通信パターンの平均通信性能

対してほぼ直線的な特性を示した。ここでは、結果のみを表3に示す。表中の (1) (2) のパターンにおいてはセットアップ時間が 750 μ s や 680 μ s と大きな値となっている。この主な原因には、通信パターンが単純な送信のみではなく送信と受信の両方を持っており、トーラス上における2回分のセットアップ時間を含んでいる点、CCU の通信処理の大部分がソフトウェアコントロールである点、があげられる。このソフトウェアコントロールの部分については、通信機構のハードウェア化をさらに進める事で今後改善される見込みがある。

5.2 応用問題による性能評価

応用問題における性能評価として、非圧縮 Navier-Stokes 方程式の一解法である MAC(Maker-and-Cell) 法の中心処理として用いられてる SOR 法の計算部分を、Red-Black-SOR 法に変形して2次元、

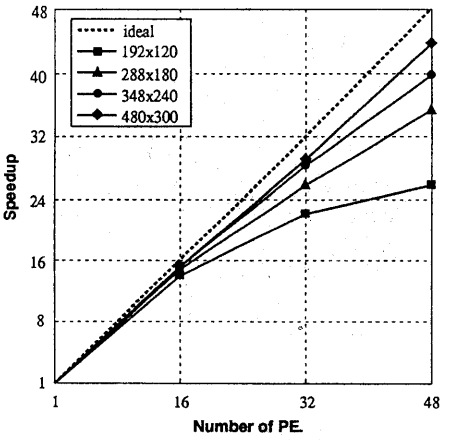
3 次元の領域分割で実行した。Red- Black-SOR 法は、SOR 法の点緩和ループの変数依存性を排除して並列化する手法であり、特に MAC 法による大規模な非定常流体の解析ではこの計算過程が繰り返して行われ、その処理時間の大部分を占めている。

ここでは、2 次元問題を仮想 2D トーラスモードによって並列実行した場合の性能測定結果を図 10(a) に示し、3 次元問題を仮想 3D トーラスモードによる 3 次元領域分割によって並列実行した場合の性能測定結果を図 10(b) に示す。図中の 192 × 120 や 48 × 30 × 20 等は領域分割した問題の総格子点数を表し、速度向上では 2 次元分割の 480 × 300 の問題規模において 48 プロセッサで 43.7 倍の性能が得られた。

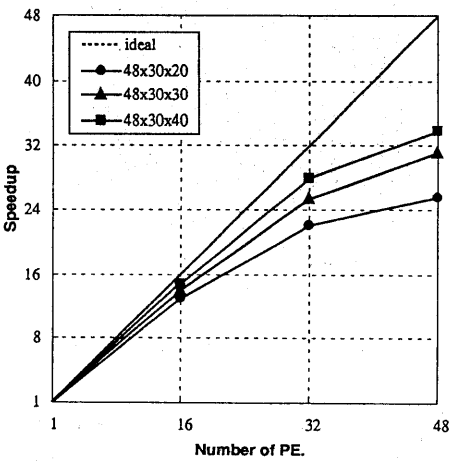
3 次元分割では 48 × 30 × 40 の問題規模までの結果を示す。この場合の速度向上は 48 プロセッサで 33.7 倍程度であるが、これはプロタイプのメモリー容量の制約から PE 単体の場合にこれ以上大きな問題規模の実装が困難であったことによる。

これらの結果と現在のプロトタイプシステムの通信性能をもとに、2 次元、3 次元問題において問題規模、実装 PE 数を変化させた場合の性能見積もり結果を図 11(a),(b) に示す。この図では、性能を実装 PE 数 n で n 倍の計算性能が得られた場合を効率 100 %として表している。

図 11(a) に示す 2 次元問題では 2000 × 2000 の問題規模で、現在のプロトタイプシステムでは 2744 プロセッサで約 75 %、4096 プロセッサでは 65 %と数千台程度で性能がやや落ち込んでくる。一方、図 11(b) の 3 次元問題では、40 × 40 × 40 や 80 × 80 × 80 といった比較的小さな問題規模において、1000 台以下の実装規模で 50 %程度かそれ以下となり、高い並列処理性能が得にくくなっている。しかし、より大きな問題領域においては、本ネットワークによっても数千から数万台程度の実装規模で高い並列化効果を得られることが判る。



(a) 2次元問題（5点差分）



(b) 3次元問題（7点差分）

図 10: Red-Black-SOR での速度向上

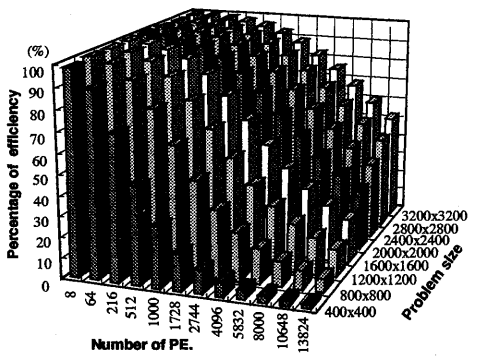
6 おわりに

本稿ではネットワーク数の削減と実効性能の両立を目標とする並列計算機アーキテクチャを提案し、実際にプロトタイプシステムを製作して、その構成、通信方式について述べた。また、Red-Black-SOR 法による近接通信を用いた数値計算で、ある程度以上の大きな問題規模を取り扱う領域において、本ネットワークの様にリンクの数を削減する指向の並列システムでも高い並列化効果を得られることを実測によって示した。しかし、プロトタイプシステムにおいては、通信機構のハードウェア化をさらに進めるなどの改善の余地もある。

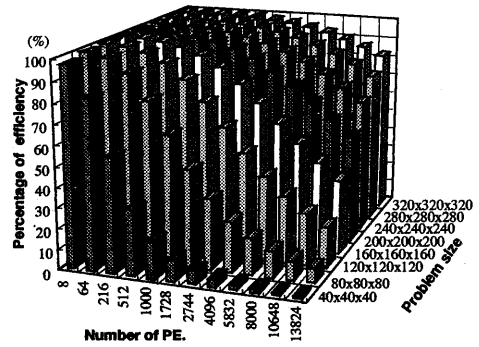
本システムでは、今後、計算結果の可視化等を含めた応用問題において処理手法や性能の評価を進めて行きたいと考えている。

謝辞

本研究は文部省科学研究費補助重点領域研究



(a) 2次元問題 (5点差分)



(b) 3次元問題 (7点差分)

図 11: PE 数と問題規模の変化に対する速度向上率

(05219212) による。また、DSP を御提供頂いた日本テキサス ・インスツルメンツ (株) に深謝致します。

参考文献

[1] 楊愚魯, 天野英晴, 柴村英智, 末吉敏則 : "超並列計算機に向き結合網 RDT", 信学論 (D-I),J78-D-I, 2,pp.118-128,(1995-02)

[2] 横田隆史, 松岡浩司, 岡本一晃, 廣野英雄, 堀敦史, 児玉祐悦, 佐藤三久, 坂井修一 : "超並列計算機列計算機 RWC-1 の相互結合網", 情処研報,93,ARC-101, pp.25-32,(1993-08)

[3] 朴泰祐, 中根猛, 三島健, 板倉憲一, 中澤喜三郎, 中村宏: "ハイパークロスバ網における適応ルーチングの導入とその評価", 信学論 (D-I),J78-D-I,2, pp.108-117,(1995-02)

[4] 中澤喜三郎, 朴泰祐, 中村宏, 中田育男, 山下義行, 岩崎洋一 : "CP-PACS のアーキテクチャの概要", 情処研報, 94,ARC-108, pp.57-64,(1994-10)

[5] 小畑正貴 : 浮動小数点DSP による高並列アレイプロセッサシステム", 情処理学会論文誌,Vol.32, 9,pp.1142-1148,(1991-09)

[6] 小畑正貴, 伊藤拓 : "1 次元 DSP アレイ上での光線追跡法の並列計算", 情処理学会論文誌,Vol.34,5,pp.934-940,(1993-05)

[7] 伊藤拓, 小畑正貴, 宮垣嘉也 : "トーラス ・リング ・バス階層アーキテクチャ", 情処研報,ARC-95-5,(1992-08)

[8] 伊藤拓, 小畑正貴, 宮垣嘉也 : "トーラス ・リング ・バス結合並列マシン-プロトタイプの構成と性能評価-", 情処研報,ARC-107-7, (1994-07)

[9] 日本 TEXAS INSTRUMENTS 株式会社: "TMS320C30 ユーザーズマニュアル",(1992)

[10] 日本 AMD 株式会社: "Am7968/Am7969 TAXI chip set テクニカルマニュアル",(1989)

[11] 小畑正貴, 中條拓伯 : 超並列計算機 JUMP-1 におけるハイビジョン画像表示システム", 情処研報, ARC-108-3, (1994-10)