計算機アーキテクチャ 107－7
（1994. 7. 21）

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-

伊藤 拓　小畑 正貴　宮垣 嘉也
岡山理科大学工学部 情報工学科

概要

数万台以上の規模の超並列システムでは、チップや基板間の入出力信号線や配線数の削減は現在でも無視する事はできない問題であると考えられる。我々はネットワーク配線量の削減を中心に、PEをリングとバスで接続したクラスタ単位を2次元トーラスで接続する疑似的な3次元ネットワークTRB(Torus-Ring-Bus)を提案しその性能見積りを行っている。本稿では、64PEで実装を行ったTRBプロトタイプシステムの構成と通信形態、及び通信コントローラの機能とその性能について述べる。

Torus-Ring-Bus connected parallel machine
- Prototype system and its evaluation -

Taku Ito　Masaki Kohata　Yoshiya Miyagaki
Department of Information & Computer Engineering
Okayama University of Science

Abstract

The increase of interconnection wires is one of the serious problems at the implementation of parallel processing systems which have over ten thousands processors. We proposed a cluster architecture named TRB (Torus-Ring-Bus) for these massively parallel systems. The processing elements are connected by both of a ring and a bus in a cluster, and a CCU (Communication Control Unit) is on a edge of the bus. Then the clusters are connected each other as 2D torus via the CCUs. In this paper, we describe the prototype system which we have build with 64 processors, and evaluate the communication performance.

1．はじめに

数万台以上の規模の超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、実装上の問題（チップや基板間の入出力信号線数や配線数）も重要な条件となってくる。この様な超並列システムとして最近では大規模な3次元接続のネットワークを有する物も現われてきている。しかし、これらはまだまだ特殊な実装技術を要するものが多く、配線数の削減は現在でも無視する事はできない問題であると考えられる。

我々はこの様な観点から、ネットワークの配線量の削減を中心にリングとバスでローカルに接続されたプロセッシングエレメント(PE)を1クラスタとし、クラスタ間を2次元トーラスで接続する疑似的な3次元ネットワークTRB(Torus-Ring-Bus)を提案している。これまでに、このネットワークに対して基本的な通信方式と近接データを用いた数値シミュレーションでの性能見積りを行い、数万台程度までの性能向上が見込める見通しを得ている[1]。本稿では、64PEで実装を行ったTRBプロトタイプシステムの構成と通信形態、及び通信コントローラの機能とその性能について述べる。

2．TRBの特徴とプロトタイプの構成

2.1 TRBの基本構成と特徴

TRBにおける計算モデルはデータパラレルの処理を中心とし、3次元流体問題等の数値シミュレーションとその可視化を主たる応用目的とする。TRBの全体構成とクラスタの構成を図1(a),(b)に示す。TRBの各PEは両隣のPEに対する2組とバスに接続する為の計3組のリンクのみを持てばよく、個々のリンクは多ビット並列による高速化が可能である。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組だけでよい。1クラスタを1ボードに実装する場合は、ボードからはトーラスの為の4リンクを引き出せばよい。また、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事になるが、隣接基板間であるので接続は比較的容易と考えられる。以上の様にTRBは実装面で有利な点が多い。

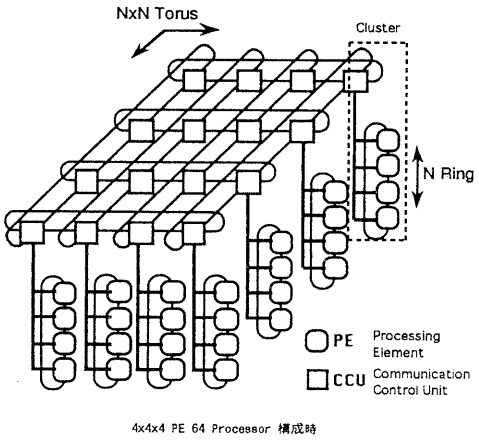
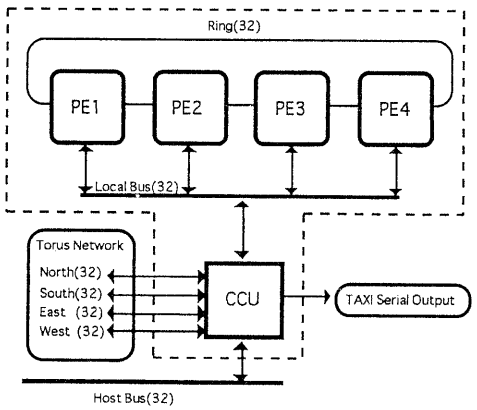


図1(a) TRBプロトタイプの全体構成



(b) クラスタの構成

2.2 プロトタイプの設計方針

TRBのプロトタイプの設計には、以下の様な方針が科せられている。

(1) 現行の一般的な実装技術のみを用いる。
(2) 任意のPE間で一般的なパケット通信を行う。
(3) 近接作用問題等の数値シミュレーションに対して仮想的な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・モードを持つ。
(4) データの収集、分配に対する支援機構を持つ。
(5) 通信ネットワークとは別個に計算結果の可視化の目的でPEのローカルメモリの一部を分散された仮想的なVRAMと見立てて外部に出力する機構を設ける。
(6) 複数ユーザーによる複数パーティションでの利用を可能とする。
(7) TRBの一般的な通信特性はシミュレーション結果[1]からパケット量が少ない場合は3次元トーラスに近い性能を発揮できるがクラスタ間のネットワークのトラフィックが大きくなると複数のPEで共有される経路での性能低下が著しいので、できるだけこれを回避する。

これらの要求は個々の処理がそれぞれある程度のボリュームを持っている事から、通信コントローラにCPUを搭載してソフトウェアによる柔軟な対応を図る。

2.3 プロトタイプの構成

プロトタイプは図1(a)に示した構成の *4PE×16クラスタ* の計64PE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。CCUとPEは独立した基板を接続して1クラスタとなる様に構成している（図2）。PEはTI社の32bit浮動小数点DSPであるTMS320C30[2]と128KbyteのSRAM,4MbyteのDRAMから成り、4PEを1基板上に実装する。CCU基板からは4方向の隣接するCCUに対して各々独立した32bit幅のリンクで2次元トーラス状に接続される。CCUにはPEと同様のDSPが1台搭載され、クラスタ内外の通信と同期の制御を行う。CCUはこの他に画像出力目的で設けた高速シリアルトランシーバ（TAXIチップ：AMD社)[3]の制御も行うが、これらの処理の総てをDSP単体で行うのは負荷が過大である為、クラスタ内の通信に用いるDMAや同期機構、トーラス間での同期機構は二つのFPGA[4]に振り分ける（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をサポートする。現在のプロトタイプの外観を図3に、諸元を表1にまとめる。

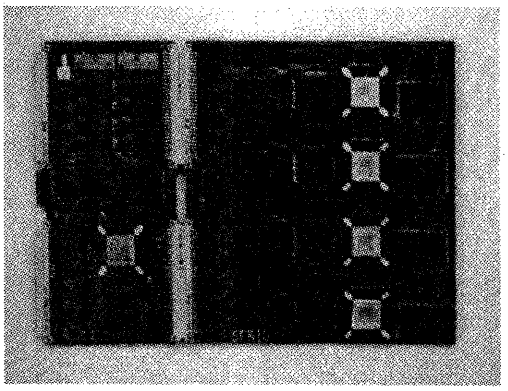


図2．クラスタ基盤（CCU:左, PE:右）

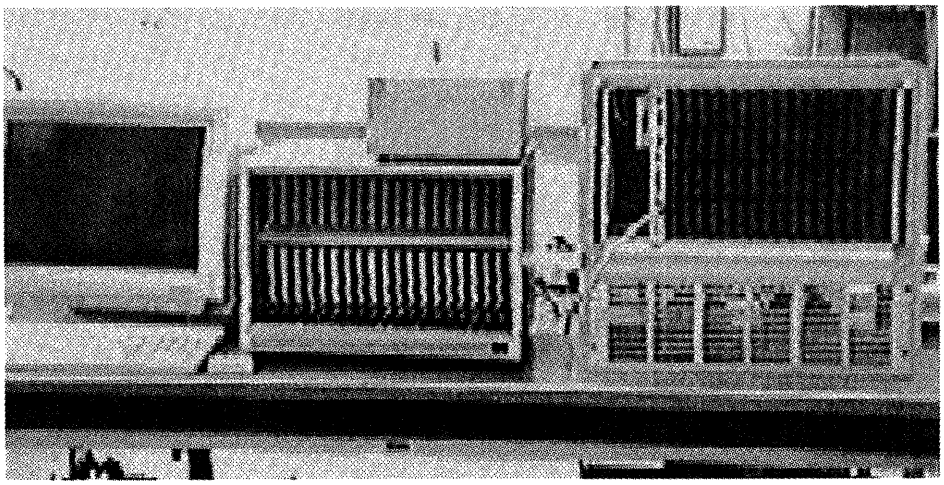


図３．プロトタイプシステムの外観

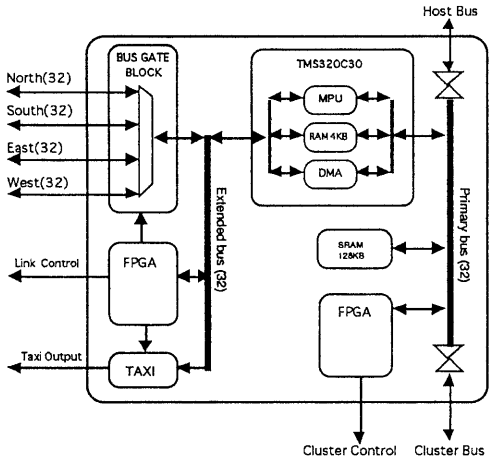


図４．CCUの内部構成

PE	
プロセッサ	32bit 浮動小数点DSP
サイクルタイム	60ns
計算性能	16.7MIPS / 33.3MFLOPS
メモリ	SRAM 128KB,DRAM 4MB
CCU	
プロセッサ	PEと同じ
メモリ	SRAM 128KB
FPGA	4200Gates-100MHz × 2
画像シリアル出力	TAXI Am7968 (100Mbps)
全体	
プロセッサ数	64 + 16
メモリ	264.2MB
結合方式	2D Torus,Ring,Bus
画像シリアル出力	100Mbps × 16

表１．プロトタイプシステムの諸元

３．通信コントローラとソフトウェア

３．１ CCUの構成と通信機構

CCUはホストとの通信以外のTRBにおけるデータ通信を総てコントロールする。トーラス側のFPGAでは、４本のリンクの各々に対して８本の制御線を持ち(図５)、リンクの確立と排他制御、デッドロックの回避を行う。リンク確立後はハードウェアハンドシェイクを行いBUS GATE BLOCK部の制御までを受け持ち、実際のデータの読み出しと書き込みはDSP内部のMPUかDMA

で行う。クラスタ側のFPGAは内部にDMAを設けクラスタ内のPE間の同期とデータ転送を受け持つ。シリアル出力であるTAXIチップは、2.2(5)で述べたPEのローカルメモリの一部をHDTVの分散VRAMに見立てて出力する目的で搭載しており、100MHzのTAXIチップが12.5Mbyte/sの転送能力を持つ事から16台のCCUで180～200Mbyte/s程度の転送能力となる。これはアプリケーションからの要請でCCU上のDSP内部タイマーによる割り込み処理で駆動する。

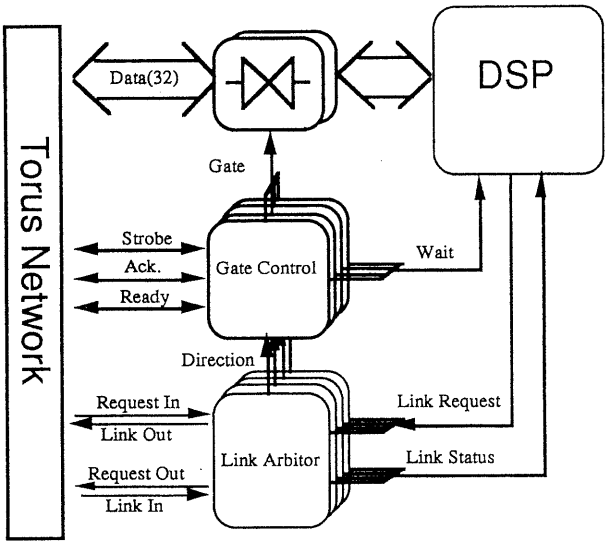


図5. トーラスのリンク通信機構

3.2 通信方式とデータパケット

プロトタイプの通信方式の基本はストアアンドフォワードであるが、クラスタ内ではこれ以外にブロードキャストとリングを介したデータの回転をサポートする。

現在のプロトタイプ上でのパケットフォーマットとしては、2.2(2)の一般通信をサポートするパケットと2.2(3)の仮想的なネットワークをサポートするパケットの2種類の物理的なフォーマットが存在する（図6）。前者は任意のPE間を結ぶ一般的なパケット(General Packet:GP)と呼び、後者は隣接するクラスタ間でのみ通信を行う場合に用いるパケット(Neighbor Packet:NP)と呼ぶ。GPのヘッダ3ワードは、受信PEのアドレス、パケットのID、分割個数である Divide, と転送順位である Number 、及びパケットのSizeからなる。Size のビットフィールドは16bitの幅があるが現在の1パケットの長さはバッファ容量の関係上1～128ワードに制限してある為、8～15bitは実質的にリザーブの状態である。TRBの一般的な通信処理はGPによって行う。NPは2.2(3)の様な数値シミュレーションの問題を扱う場合に、図7に示す様に隣接クラスタにあらかじめルーティングされる事が明らかな複数パケットをクラスタのCCUが取りまとめてショートヘッダを付加し不要なリンク確立やルーティング処理のオーバーヘッドを削減する。

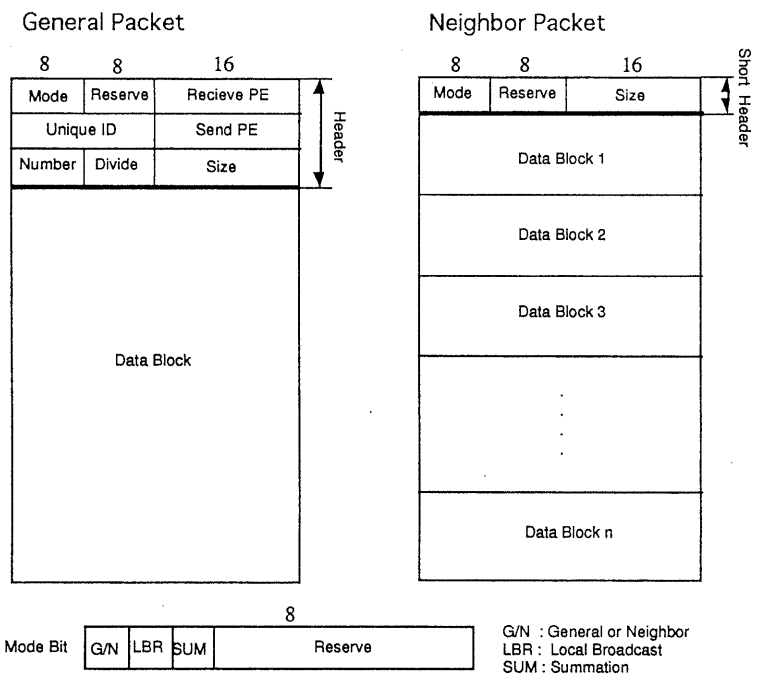


図6. パケットフォーマット

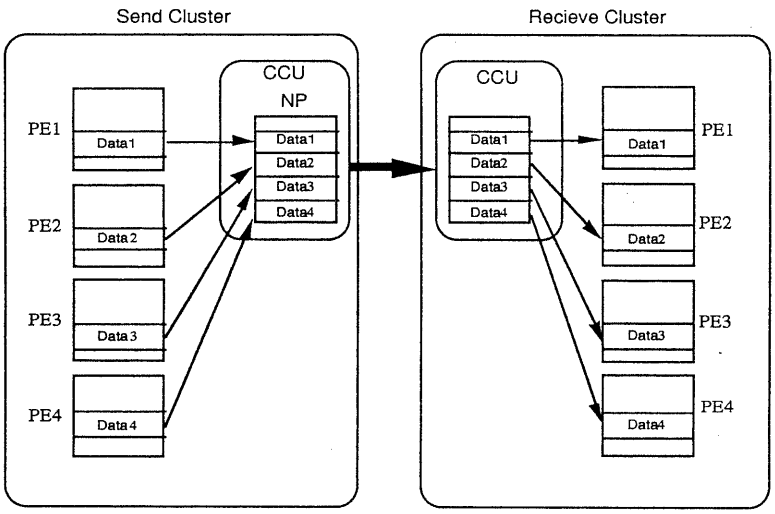


図7. NPによる隣接クラスタ間の通信

3.3 ソフトウェア

TRB上でのソフトウェア開発は通信関数を付加したC言語で行う。アプリケーションレベルに供給される基本的な通信関数を表２にまとめる。仮想Ring,仮想2D Torus,仮想3D Torusの関数でのクラスタ間の通信はNPによって行われる。仮想ネットワークはパーティションによる複数ユーザの利用を実現する関係上CCUがパーティション境界を認識しておく必要がある事から、コンパイル時に静的に決定される形式をとっている。

一般通信	Send() RBalloc() RBfree() RBsense() RBflush()	パケット送信 Recieveバッファの取得とCCUへの通知 Recieveバッファの解放とCCUへの通知 Recieveバッファのステータス Recieveバッファのフラッシュ と再生
仮想 Ring	Left() Right()	仮想リング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()	仮想3D Torusモードでの６方向隣接通信
仮想 VRAM	VBalloc() VBfree() VBmode()	仮想 VRAM 領域の取得とCCUへの通知 仮想 VRAM 領域の解放とCCUへの通知 仮想 VRAM のモード設定

表2. 通信関数と機能概要

4．通信性能

プロトタイプ上で、以下の二点を実際に測定しクラスタ間の通信性能とする（表３）。

- トーラスにリンクリクエストが発生してから第一ワードが受信側に到達するまでの時間をリンクセットアップ時間とする。
- DMAでハンドシェイクによる転送レートの測定を行う。この時バッファにはDSPの内部メモリを利用した。

リンクセットアップ時間	2.72 (μ s)
１ワード転送時間	320(ns)
クラスタ間転送レート	12.5Mbyte/s

表３．クラスタ間の通信性能

DSP内臓のDMAの最高転送レートは内部メモリを用いた場合で約22.2MHzであるが、測定値

ではそれの５６％程度となってしまっている。この主な原因はハンドシェイクを行うFPGAの信号遅延によるもので、送信側のストローブとそれに応答するアクノレッジが共に40ns程度の遅れを持ち、それらが双方合わさってCPUに１Wait発生させてしまうものであった。

CCUとPEのメモリに対するデータ転送はソフトウェア転送で行われる。現状での速度は。約1.7μs/wordであるがこれは現在通信関数が実装途中であり、動作検証優先の為にC言語で特にチューニングを行わずに記述してある為である。これらは随時アセンブラで書き下ろして行く予定である。

5. 現状と今後

TRBプロトタイプ全体の状況はハードウェアの実装作業がほぼ終了し、現在一般通信用のプログラムのコーディングとデバッグを行っている段階である。また、現時点ではデバック環境の関係からホストコンピュータにパーソナルコンピュータを使用しているが、今後進めて行く仮想ネットワーク、複数パーティションへの対応と共に、１ボードのUNIXマシンをホストとして同一の筐体に組み込みスタンドアロンでネットワークに接続可能な形式とする予定である。さらに今後は応用アプリケーションでの性能検証も行っていく必要がある。

＜参考文献＞

1)伊藤,小畑,宮垣："トーラス・リングバス階層アーキテクチャ",情報処理学会,計算機アーキテクチャ研究会 ARC95-5 1992
2)日本 TEXAS INSTRUMENTS 株式会社: "TMS320C30 ユーザーズマニュアル", 1992
3)日本 AMD 株式会社: "Am7968/Am7969 TAXI chip set テクニカルマニュアル", 1989
4)XILINX:"Programable Gate Array Data Book",1989

本研究は文部省科学研究費補助重点領域研究(05219212)の補助を受けています。
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(株)に感謝致します。