計算機アーキテクチャ 113−16
（1995. 8. 24）

RWC-1の入出力リングバス
− テストベッド上での実装 −

廣野 英雄† 松岡 浩司† 岡本 一晃† 横田 隆史† 坂井 修一†

†技術組合 新情報処理開発機構 つくば研究センタ

概要超並列計算機RWC-1の入出力機構について報告する。RWC-1の入出力系は、(1) 二階層の専用結合網、(2) 大容量二次記憶系、(3) 画像・音声インタフェース、(4) 外界とのインタフェース、などから成る。ここでは、下位階層の入出力系に関して、結合網とプロセッサ内部の入出力処理部について述べる。

下位の入出力系の特色として、(a) 演算処理からの高い独立性、(b) リングバスによる高スループットのデータ転送、(c) 優先度を用いた高度な制御、などがあげられる。本入出力系は、最初の版のRWC-1プロセッサチップおよびこれを用いたテストベッドに実装され、現在試験中である。

I/O Ringbus for the Massively Parallel Computer RWC-1

Hideo HIRONO† Hiroshi MATSUOKA† Kazuaki OKAMOTO† Takashi YOKOTA†
Shuichi SAKAI†

†Tsukuba Research Center, Real World Computing Partnership
Tsukuba Mitsui Building 16F, 1-6-1 Takezono,
Tsukuba-shi, Ibaraki 305, Japan

Abstract This paper presents the I/O mechanisms of RWC-1. The RWC-1 I/O system mainly consists of (1)two-layered dedicated interconnection network, (2)mass storage disk system, (3) audio/visual interface and (4) interface from/to the outer world. Here we present the lower level I/O system, especially its interconnection network and I/O controller on a PE chip.

The features of the lower level I/O system are (a)independence of the instruction execution, (b) high throughput data transfer by a ringbus, and (c)precise control by priorities. The I/O system has been implemented on the first version RWC-1 processor chip and on the RWC-1 testbed 1 system.

1 はじめに

実世界 (Real World) の現象を扱うには、時々刻々と変化し、不完全でかつ矛盾を含み複雑に関連しあった情報を処理しなければならい。新情報処理開発機構 (RWC) では、「柔らかな情報処理」というキーコンセプトのもと、実世界にも対応できる統合・学習型のあたらしい情報処理を目指し、その計算基盤として 1,000 台規模の要素プロセッサ (PE) からなる超並列計算機 RWC-1 の研究・開発を行なっている。

RWC で研究が進めれている新しい情報処理には、画像理解・音声認識などのパターンと記号を融合した処理、ロボットの柔軟な自律制御、ニューラル処理など、さまざまな様式のアプリケーションが存在する。こうしたさまざまな問題に柔軟に適応するために、超並列計算機に求められる機能としては、1) 問題に内在する並列性を最大限に抽出し、計算機の物理的並列性へのマッピングを最適化すること、2) 大量のデータを限られた時間内に転送できること、などがあげられる。

RWC-1 では、演算と通信を融合した並列アーキテクチャ RICA(Reduced Interprocessor-Communication Architecture)[2] を採用することで通信・同期にかかるオーバヘッドを削減し、細粒度の並列処理を効率良く実行することを可能にしている。一方、RWC-1 は演算処理と独立した入出力機構を備えており、大量のデータ転送に耐え得る高速性と、さまざまな形態の問題に対応できる柔軟性を実現している。

本稿では、まず超並列計算機における入出力機構の要件を述べ、次に RWC-1 の入出力機構でこれをいかに実現するか述べ、最後にテストベッド上に実装した入出力リングバスについて報告する。

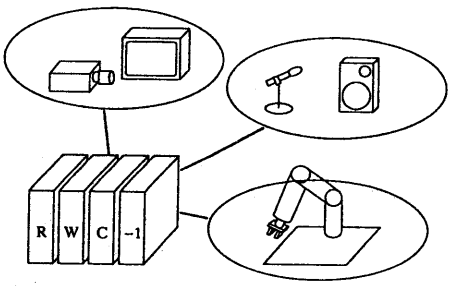

図 1: RWC-1 とアプリケーション

2 超並列計算機の入出力機構

入出力機構とは、演算を行なうプロセッサに対して、外界とのデータのやりとりを行なう入出力装置とそのインタフェースの部分である。この章では、超並列計算機の入出力機構の要件について、結合網と要素プロセッサの面から考察する。

2.1 結合網の要件

並列計算機では入出力装置は共有資源であるため、PE 間の協調動作の必要性から、入出力データはすべての PE で利用できなければならない。このため、入出力データは結合網を通した通信によって入出力装置と PE の間でやりとりされる。本稿では、この入出力データのための結合網のことを入出力用結合網と呼ぶ。

入出力のための通信の特徴は、1 回に転送するデータ量大きいことである。したがって入出力用結合網は、長時間特定の通信路を占有する通信が行なえることが必要となる。

また、入出力データには画像・音声のような時間依存性の高いデータがある。これらのデータを用いて実時間処理を行なおうとする場合、サンプリング周期よりも短い時間内に、PE と入出力装置の間でデータの転送を完了させなければならない。このため、入出力用結合網には転送時間の最悪値を保証できることが必要になる。

さらに超並列計算機において効率的な運用を考えると、計算機を空間分割しグループごとに独立して動作させることが必要である。空間分割されたそれぞれのグループが別々に実時間処理を行なうことを考えると、グループ同士が互いに干渉しないようにすることが必要となる。このため入出力用結合網は、グループ内の通信が他のグループを通らないなど、空間分割に対応したものでなくてはならない。

2.2 要素プロセッサの要件

入出力データは大容量のため、入出力の動作は入出力用結合網と直接メモリとの間のデータのやりとりとなる。この入出力動作は、演算処理部においてソフトウェア的に実行することも可能である。しかしこの部分をハードウェア化し演算処理部と別に用意すれば、入出力動作を演算と並列に動作させることが可能となり、計算性能を落とすことなく入出力を行なうことが出来る。すなわち単一プロセッサにおいて DMA コントローラによって高効率な入出力が可能となったように、超並列計算機においても

演算処理部と独立したメモリアクセス機構を実装すれば、高効率な入出力を実現することができる。

また実装規模が大きいことから、超並列計算機では高密度の実装が必要とされる。このため要素プロセッサにおける入出力機構は、プロセッサチップの外部に必要なハードウェアをできるだけ少なくすることが必要である。

3 RWC-1 の入出力リングバス

前章の考察を踏まえ検討した RWC-1 の入出力リングバスは、以下のような特徴を持つ。

3.1 入出力用結合網

超並列計算機 RWC-1 は、細粒度並列処理を主眼にしており、演算と通信とが融合された RICA を採用している。RICA において通信は、ハードウェア化された通信・同期機構によって、2 つの PE 間のレジスタからレジスタへのデータ転送として実装されている。よって演算処理の通信は、転送サイズが小さく、比較的ランダムに行なわれる。このため、RWC-1 の演算用結合網には少ないハードウェア量で低レイテンシ・高スループットの高速な通信が可能でありかつランダムな通信にも高い性能を示す超並列向けの新しいトポロジ、MDCE(Multi-dimensional Directed Cycles Ensemble) を採用している [3]。

このように優れた特徴をもつ演算用結合網であるが、これを演算だけでなく入出力の処理にも併用すると、演算結果と入出力という性質の異なる通信が共存するため、次のような不都合な点が生じる。

1. 入出力のための通信が、長時間、局所的に経路を占有するため演算処理の通信が阻害される。
2. 演算処理の通信が入出力の通信を阻害するため、入出力の転送時間の最悪値の保証が出来ない。

このような状況を回避するために、RWC-1 では演算用結合網とは独立した入出力用の結合網を用意する [6]。

3.2 入出力用結合網のトポロジ

RWC-1 では空間分割を行なって運用することを考えており [1]、それに対応するため入出力用結合網は階層型結合網とした。階層型結合網では下位階層をグループの単位と一致させることにより、最小グループ内の通信を上位階層を介さずに行なうことができる。また上位階層のトポロジによっては任意のグループに分割しても、他のグループと干渉することなく通信を行なうことが可能である。

また階層型結合網には、局所通信に強く、全体のハードウェア量を減らすという利点がある。

入出力用結合網の上位階層には、汎用性にすぐれ長距離通信にも強く、十分大きなバンド幅をもつ ATM ネットワークを採用し、下位階層は以下の理由からリングバスとした。

- 入出力の通信では 2 ノード間のレイテンシより、バンド幅が優先される
- 入力と出力が 1 対 1 に対応し、ドライバやスキューなど設計上の制約が少ない
- ルーティングが不要であるなど制御が簡単である

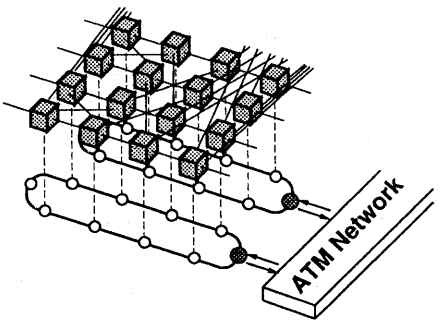

図 2: MDCE(上) と入出力リングバス (下)

3.3 調停方式

転送時間の最悪値を保証するためには、次の 2 つの方法がある。

1. 静的なスケジューリングにより、転送バンド幅を確保する
2. 通信に優先度を持たせ高優先度のものに転送を割り当てる。

1 は、同期的な入出力要求に対しては実現が容易だが、動的に入出力が発生する場合、時分割的な割り当てに限界がある。このため RWC-1 の入出力機構では 2 を実装した。画像・音声データなどの通信には高い優先度を与えることで、ファイルアクセスなどの低い優先度の通信で結合網が混雑していて

も、転送時間の最悪値を保証することが可能となっている。また同一の優先度の要求があった場合、ノード間で偏りがないようラウンドロビンにより、平等にスケジューリングされるようになっている。

3.4 プロセッサの入出力機構

RWC-1 のプロセッサには、入出力用メモリアクセス機構が内蔵されている。この入出力用メモリアクセス機構は演算処理部とは独立しているため、結合網からのデータを自動的に指定されたメモリ領域に書き込む、あるいは指定されたメモリ領域から結合網に自動的にデータを送り出す動作を演算と並行して行なうことができる。

演算処理部からはこれらの入出力動作が、I/O 用命令、制御レジスタ、内部割り込みという形で扱える [5]。これにより入出力のためのソフトウェアが作りやすくなっている。入出力のメモリアクセスは、演算処理のメモリアクセスと一緒にラウンドロビンによって調停されている。

またプロトコル制御のためのハードウェアを含んだ入出力結合網のインタフェースを内蔵しているため、少ない周辺回路で入出力用結合網が実現できる。このため、計算機自体をコンパクトに実装をすることが可能となっている。

4 RWC-1 テストベッドへの実装

4.1 RWC-1 テストベッドの全体の構成

RWC-1 テストベッドは RWC-1 プロセッサの動作試験・機能検証を行なうための小規模なテストベッドシステムであり、1 枚のボード上に 2 つのプロセッサチップ [4] が実装されている。また RWC-1 テストベッドでは、下位階層の入出力用結合網であるリングバスの機能検証も行なう。このため、ボード上にテストベッドのホストとなる IBM PC とのインタフェースを行なう PC ノード、リングバス上のデータ転送を管理するマスタノードを CPLD を用いて実装した。これら 4 つのノードは 8bit 幅で互いの入力と出力が結ばれている。

なお RWC-1 テストベッドでは、上位階層の結合網や入出力装置は実装していないが、これらへの通信は、PC ノードを介して IBM PC でエミュレートすることで行なう。

4.2 入出力の全体動作

RWC-1 テストベッドでは、リングバス上の通信は、8 バイトのヘッダと 32 バイトのデータからなる固定長のパケット (I/O パケット) をノードからノードへ転送することによって行なわれる。これらの転送の制御はマスタノードにより集中的に管理される。マスタノードはリングバス上の各ノードからの送信要求を受け付け、それらの中から公平に 1 つのノードを選び、そのノードに送信許可を与える。送信要求を出しているノードは、マスタノードからの送信許可が与えられてからはじめて送信を行なうことが出来る。

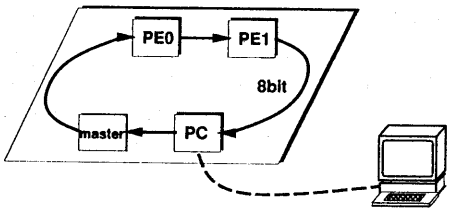

図 3: RWC-1 テストベッド

送信は、宛先が含まれたヘッダとデータからなる I/O パケットをリングバス上に出力することで行なわれる。I/O パケットはリングバス上のノードを順番に転送されていくが、各ノードは自分が宛先で指定されたノードであれば、I/O パケットを取り込む。このように受信は送信のように集中的に管理されず、各ノードに分散して行なわれる。

上記の動作を繰り返すことで、リングバス上でのデータ転送が実現されている。なお、制御を簡単にするために 1 つのリングバス上には同時に 1 つの I/O パケットのみが存在するという制約を課している。

4.3 I/O パケットの形式と種類

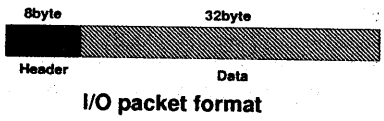

図 4: I/O パケットのフォーマット

図 4、図 5に I/O パケットとそのヘッダのフォーマットを示す。

PS ビットは I/O パケットが送信ノードから出力されて、宛先のノードで受け取られるまでの間セットされ、I/O パケットが有効であるかどうかを示す。ND は I/O パケットの宛先を示す。ND は 11 ビットあり、すべての PE と入出力装置に番地を割り振ることができる。MD は I/O パケットの種類

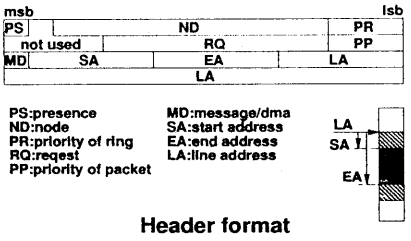

図 5: I/O パケットのヘッダのフォーマット

を示す。I/O パケットには、DMA パケットとメッセージパケットの 2 種類があり、リングバスから受信されたデータパケットは、PE 内部でそれぞれ別の方法で処理される。

DMA パケットはノード間のメモリからメモリへのデータ転送を実現するために用いられる。DMA パケットを受信したノードでは、データをヘッダの LA、SA、EA から計算されるメモリ領域に自動的に書き込む。この動作は、演算処理部とは独立に行なわれる。

メッセージパケットは、ノード間で演算処理部から演算処理部への通信を行なうために用いられる。受信されたメッセージパケットのデータ部は、自動的にメモリ上のメッセージキューに挿入され、このとき受信側の演算処理部には割り込みが生じる。その後は、キューからデータを取りだし、ソフトウェア的な処理を行なう。

PR と RQ と PP はプロトコルを実現するために用いられる。これらの詳細は次章で述べる。

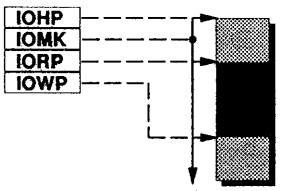

図 6: メッセージキュー

4.4 通信プロトコル

RWC-1 の入出力の通信プロトコルは、マスタノードの集中管理の元で 1 つ 1 つの I/O パケットが、ノードからノードにどのように転送されるかを規定する下位のレイヤと、複数の I/O パケットを用いて異なるノードのメモリからメモリへのブロック転送がどのように行なわれるかを規定する上位のレイヤの 2 つからなる。下位のレイヤの実現のためにはパケットのヘッダが用いられ、すべての動作はハードウェアで自動的に行なわれる。上位のレイヤの実現にはメッセージパケットが用いられ、ソフト的に行なわれる。

1 つ 1 つの I/O パケットの転送は基本的には (1) 送信要求の集計、(2) マスタノードでの調停、(3) 送信許可、(4) データ転送を繰り返すことで実現されている。送信要求の集計は、マスタノードから送り出された I/O パケットが、リングバスの特性によりすべてのスレーブノードを通りマスタノードに戻ってくることを利用して実現している。I/O パケットのヘッダにはリング上の各ノードに対応したビットからなる RQ フィールドがあり、I/O パケットが到着したノードは送信要求のあるなしで、自分のビットを設定する。すなわち I/O パケットがマスタノードに到着したときには、RQ フィールドは送信要求のあるノードに対応するビットが立っていることになる。

RWC-1 の入出力リングバスでは、この仕組みにさらに優先度を組み合わせている。ヘッダの PR フィールドは、送信要求を行なったノードによって書き換えられ、送信要求を行なったノードのうち最も高い優先度が格納されている。各スレーブノードでは、この PR とノード自身の優先度との関係によって送信要求ビットの設定が以下のように行なわれる。

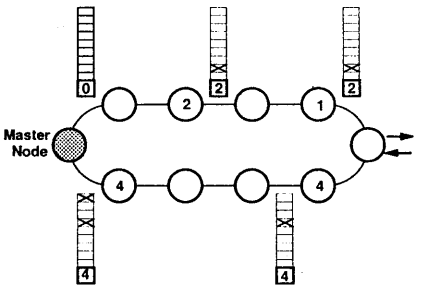

図 7: 優先度つきアービトレーション

a) ヘッダの優先度が高い
　送信要求ビットはセットしない

b) ヘッダの優先度と同じ
　送信要求ビットをセットする

c) ヘッダの優先度が低い
　送信要求ビットをセットし、さらに自ノード以外のビットをリセットする

このようにして、もっとも優先度の高い送信要求のみがマスタノードに伝えられ、その中から送信

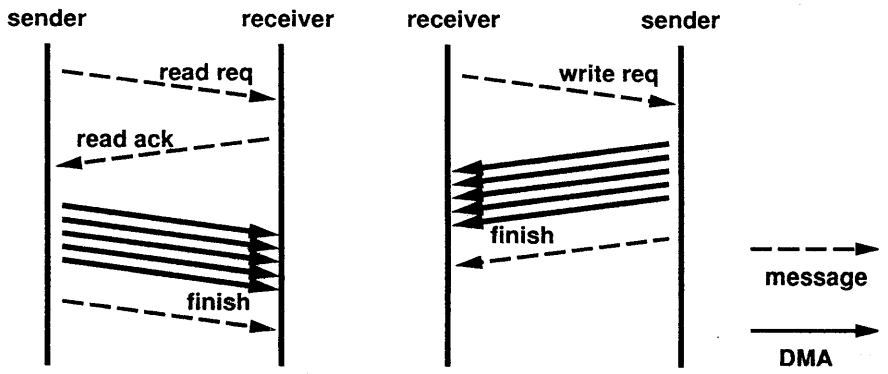

図 8: 入出力用通信プロトコル

許可が与えられるノードが選ばれる。これらの要求と調停のための動作は、I/O パケットの転送とパイプライン的に行なわれるため、I/O パケットは連続して送信することが出来る。

なお PP には送信したノードの優先度が書き込まれ、上位階層からリングバスに送信される際の優先度に用いられる。

次に上位レイヤのブロック転送であるが、前述のようにメッセージパケットを用いると入出力用結合網を通して、ノード間でソフトウェア的にネゴシエーションを行なうことができる。図 8に、送信側主体のデータ転送と受信側主体のデータ転送のプロトコルを示す。ブロック転送のプロトコルはソフトウェアによる処理のため自由度が大きい。

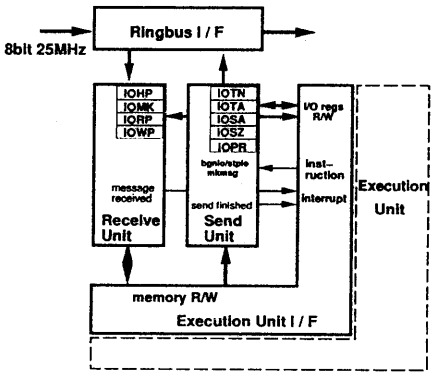

図 9: プロセッサの入出力機構

4.5 I/O 命令と制御レジスタ

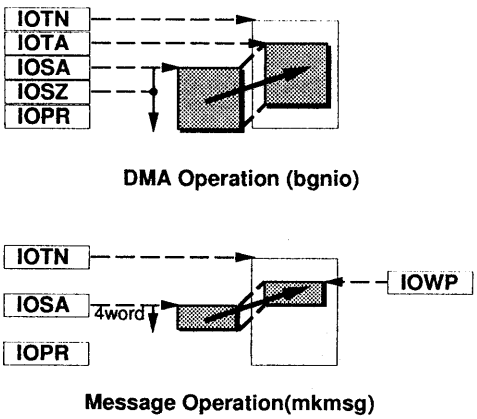

図 10: DMA とメッセージ転送動作

RWC-1 には I/O 用にいくつかの命令と制御レジスタが用意されている。I/O 命令はデータ転送の制御を行なうために用いられ、以下の 3 つがある。

bgnio(begin I/O) DMA 転送を開始する

stpio(stop I/O) DMA 転送を中断する

mkmsg(make message) メッセージを転送する

bgnio が実行されると、演算処理部とは独立に入出力機構が起動される。そして、制御レジスタの内容にしたがって自動的に自ノードのメモリ領域からデータを読みだし、ヘッダを付加して DMA パケットを生成し、順番に送信する。stpio が実行されるとパケット単位でデータ転送が中断される。mkmsg は bgnio と同様な動作を行なうが、DMA パケット

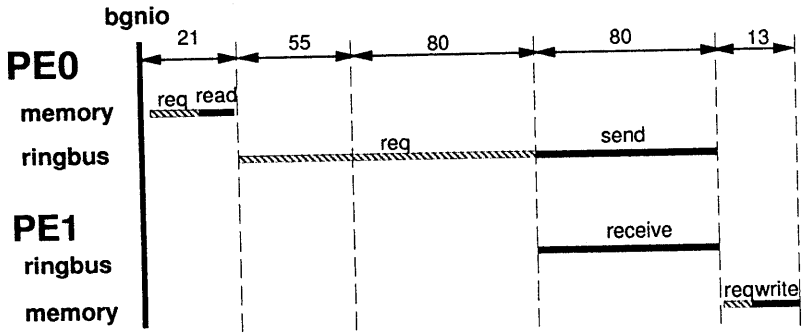

図 11: PE 間のデータ転送

の代わりにメッセージパケットを 1 個だけ生成し、送信する。bgnio、mkmsg とも終了時に演算処理部に対して割り込みが発生する。

I/O 用制御レジスタは、大きく送信用と受信用の 2 種類に分かれる。送信用には以下のものがあり、I/O 命令により起動されるデータ転送の条件となる。

IOTN：転送先ノード
IOTA：書き込み開始番地
IOSA：読みだし開始番地
IOSZ：転送サイズ
IOPR：優先度

受信用には以下のものがあり、メッセージキューの管理用に用いる。

IOHP：キューの先頭番地
IOMK：サイズ
IORP：読み出し番地
IOWP：書き込み番地

図 10に I/O 命令実行時の各制御レジスタの役割を示す。

4.6 データの転送例

隣合った PE0 から PE1 への 1 パケット分 32 バイトのデータ転送を行なった (図 11)。転送に伴う各動作にかかったクロック数は以下の通り。

bgnio 実行から送信要求状態になるまで	21
リングバスに送信要求を出力するまで	55
送信許可を受け最初のデータを送信するまで	80
最初のデータの入力から最後のデータの入力まで	80
最後のデータが入力から書き込みが終了するまで	13
合計	249

各動作についてかかった時間を詳しく見ていくと、送信側 PE0 で、bgnio 実行から PE が送信要求状態になるまでには、1 度メモリの調停が入っている。この例では命令キャッシュのフィル動作とぶつかっているため、6 クロック余分にかかっている。競合するものがなければ、この動作の最小時間は 15 クロックである。

次に、リングバスを介して送信要求を出力するのは、PE0 に I/O パケットのがやってきた時である。したがって送信要求状態になってから、送信要求を出力するのは最大で 1 パケットサイズ分の 80 クロック後である。送信を開始できるのも次の I/O パケットが到着した時であるので、同様に 80 クロック後となる。

2 つのノードは隣合っているため、PE0 の出力ポートと PE1 の入力ポートは直接つながっている。したがって、PE0 で最初のデータを出力してから PE1 でデータを入力するのには時間がかからない。もし、2 つのノードが隣ではなく離れていれば、1 つのノード毎に 4 クロック余分にかかる。

受信側の PE1 で、最後のデータが入ってから書き込みが終了するまでの時間には、1 度メモリの調停が入っているが、他に競合するものがなく最小の時間となっている。

1 パケットだけの転送では 243 クロックのうちデータの転送に実際にかかったのは 80 クロックだけであり、残りはメモリアクセスにかかった時間と送信するまでの待ち時間である。2 番目のパケット以降は調停のための動作が送信動作と完全にパイプライン化されるため、データ転送に必要な 80 クロック分が加算されるだけである。

5 まとめ

本稿では、超並列計算機 RWC-1 の入出力機構について、入出力用結合網、要素プロセッサの入出力機構、入出力動作の面から報告した。超並列処理では、演算とそれに伴う通信をいかに効率良く行なうかが重要であり、RWC では新しいアーキテクチャである RICA、超並列向け結合網 MDCE を提案してきた。さらに RWC-1 では演算処理を妨げることなく入出力を行なうために独立した入出力用結合網を設けている。これにより性質の異なる入出力データが演算用結合網を流れないようにすると同時に、PE チップに内蔵された入出力制御機構が命令実行部と独立して、入出力結合網を介したデータ転送と PE のメモリへのアクセスを行なうようにしている。演算・通信と入出力を並列に動作させるこれらの入出力機構により、システムとして高効率な超並列処理を行なうことが出来る。

現在我々は入出力結合網の上位階層である ATM 網の特長を生かすために、リングバスと ATM 網の中継を行なう ATM ノードの機能を強化し、バッファリングや再送機能などを持たせることを検討中である。ATM ノードの機能強化に伴って PE 内部の入出力制御機構もバースト転送を基本とするなどの改良を行なっており、今年度中には新しい入出力機構を含んだ PE 数 64 台規模の RWC-1 プロトタイプが完成予定である。プロトタイプでの実験、評価をもとに最終的に 1,000 規模の RWC-1 を製作する予定である。

謝辞

本研究を遂行するにあたり、有益な御指導、御討論をいただいた島田研究所長、石川超並列ソフトウェア研究室長、超並列ソフトウェア研究室員の諸氏、ならびに超並列アーキテクチャワーキンググループの諸氏に感謝いたします。

参考文献

[1] 坂井修一, 岡本一晃, 松岡浩司, 廣野英雄, 児玉祐悦, 佐藤三久, 横田隆史. 超並列計算機 RWC-1 の基本構想. 並列処理シンポジウム JSPP'93, pp. 87–94, 1993.

[2] Shuichi Sakai, Yuestu Kodama, Mitsuhisa Sato, Andrew Shaw, Hiroshi Matsuoka, Hideo Hirono, Kazuaki Okamoto, and Takashi Yokota. Reduceed interprocessor-communication architecture and its implementation on em-4. *Parallel Computing*, Vol. 21, pp. 753–769, 1995.

[3] 横田隆史, 松岡浩司, 岡本一晃, 廣野英雄, 坂井修一. 超並列向け相互結合網 mdce の提案と評価. 並列処理シンポジウム JSPP'94 論文集, pp. 121–128, May 1994.

[4] 松岡浩司, 岡本一晃, 廣野英雄, 横田隆史, 坂井修一. 超並列計算機 RWC-1 用プロセッサチップの設計. 信学技報, CPSY95-18, pp. 55–62, April 1995.

[5] 岡本一晃, 松岡浩司, 廣野英雄, 横田隆史, 坂井修一. 超並列計算機 RWC-1 の命令セット・アーキテクチャ. 信学技報, CPSY95-36, June 1995.

[6] 廣野英雄, 松岡浩司, 岡本一晃, 横田隆史, 堀敦史, 児玉祐悦, 佐藤三久, 坂井修一. 超並列計算機 RWC-1 における入出力機構. 情報処理学会研究報告 ARC, pp. 33–40, August 1993.