計算機アーキテクチャ 119－34
（1996. 8. 29）

# RWC-1の入出力機構と基本性能

廣野英雄† 松岡浩司† 岡本一晃†
横田隆史† 坂井修一††

超並列計算機RWC-1の入出力機構は，外部にある入出力装置，PEチップに内蔵された入出力インタフェース，それらを相互に接続する専用の階層化された入出力用結合網から構成されており，これらを用いてRWC-1では演算と干渉することなく入出力動作を行うことが可能となっている．また，入出力用結合網の下位階層であるリングバスについて転送性能を評価したところ，入出力のデータ転送には十分な性能を持つことがわかった．

# I/O system for RWC-1 and its basic performance

HIDEO IIRONO,† HIROSHI MATSUOKA,†
KAZUAKI OKAMOTO,† TAKASHI YOKOTA† and SHUICHI SAKAI††

This paper introduce an I/O system for massive parallel computer RWC-1 and report its basic performance. The I/O system mainly consist of external I/O devices, I/O controllers on a PE chip, and its two-layerd dedicated interconnection network. RWC-1 can transfer I/O data through the I/O system without interfere of instruction executions. And we estimated a performance of ringbus, the lower level I/O interconnection network. The results show ringbus has enough performance for I/O data transfer.

## 1. はじめに

新情報処理開発機構(RWCP)では，実世界における曖昧さ・不完全さ・変容性等を持つ多種多様な情報に対し，これらを許容しつつ妥当な判断や問題解決などを行うことを可能にする「新情報処理」の研究開発を行っている．「新情報処理」を支える計算基盤を実現するため，我々の研究室では，汎用的な超並列システムである超並列計算機RWC-1[8]を研究開発している．

RWC-1は1024台の要素プロセッサ[3](PE)から構成される．それぞれのPEは，超並列向けの新しいアーキテクチャであるRICA(Reduced Interprocessor-Communication Architecture)[5]に基づき，独自に設計・開発したものを用いている．RICAとは，演算と通信をアーキテクチャ的に融合したものであり，並列処理で問題となる通信・同期処理に関するオーバーヘッドを軽減することができる．このことにより，RWC-1はプロセッサの台数効果を活かした高い演算能力を持つことが可能となった．

さて，RWC-1では「新情報処理」の様々な応用が実行される．実世界の多種多様な情報を限られた時間内に処理するためには，高い演算能力はもちろんのこと，システム全体の性能を高めることが求められる．そのためにはPEに外部からのデータを供給し，また演算結果を外部に反映させる入出力機構が重要となる[9]．

本稿では，このRWC-1の入出力機構について，どのように構成され，その上で入出力がどのように行われるかについて述べ，入出力動作に伴うデータ転送の予備評価の結果を報告する．

† 技術研究組合 新情報処理開発機構
Real World Computing Partnership
†† 電子技術総合研究所
Electrotechnical Laboratory

## 2. RWC-1における入出力

### 2.1 並列計算機の入出力

計算機は大きく分けて，演算・制御を行う演算処理部，データを記憶するメモリ，外部とデータのやりとりを行う入出力機構の3つから構成される．これは単一プロセッサ計算機でも並列計算機でも同じである．ただし単一プロセッサ計算機において通信は入出力の1つとされるが，通常並列計算機では演算用結合網で結ばれたPE群を1つの演算処理部と考え，PE間の通信を入出力から除く．すなわち，並列計算機の入出力とはPE群と外部の間のデータのやり取りのみと考える．

### 2.2 RWC-1の入出力機構に求められるもの

RWC-1では，複数のPEにデータが分散して配置さ

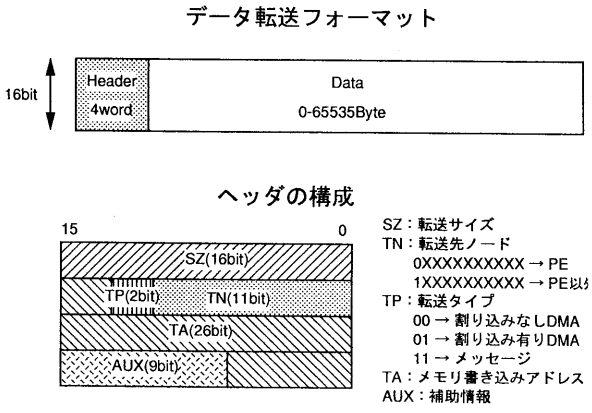


**図 1** 転送データ

れているため，同時に多数の入出力動作が行われる．このため，入出力機構には複数の PE における入出力動作を効率よく実行することが求められ，それぞれの入出力動作が干渉しないように PE 間で調整し，協調動作を行う必要がある．

また，RWC-1 で行なわれる入出力には，単一プロセッサ計算機と同様にデータを保存するためのディスク入出力，視覚情報をやりとりする画像入出力などが含まれる．これらの入出力は

- 転送サイズが大きい
- 時間依存性が高い

という特徴を持つ．したがって，それを実現する入出力機構には

- 長時間特定の経路を占有する通信が行なえること
- 十分なバンド幅が確保できること
- 転送時間が保証できること

が求められている．

## 3. RWC-1 入出力機構の実現

### 3.1 全体構成と動作

RWC-1 の入出力機構は，ディスク・画像ボードなどの入出力装置，PE の入出力インタフェース，それらを相互に接続する入出力用結合網からなる．この入出力機構上を入出力のためのデータ (転送データ) が流れることによって入出力が行なわれる．転送データは宛先等が格納されたヘッダと最大 64K バイト可変長のデータ本体から構成される (図 1)．

転送データのサイズが比較的大きいため，RWC-1 では一般の計算機と同様に，データ本体をメモリのイメージとして扱う．すなわち，送信時にはメモリの指定された番地から順番に読みだされた値をデータ本体として入出力用結合網に送り出し，受信時には受けとったデータ本体を指定された番地のメモリに順番に書き込む．このような実装により，PE における入出力の取り扱いに関するオーバヘッドを小さくすることができた．

また，サイズが小さい転送データ (64bit × 4word) については，ヘッダの内容により，メモリではなくレジスタ上にデータ本体を格納する機能を持たせた．この転送データのことをメッセージと呼ぶ．メッセージは，予めメモリ上にデータ本体の格納場所を確保する必要がないため取り扱いが簡易である．

RWC-1 ではその性質を活かして，メッセージを要求・返答などの PE と入出力装置間の動作フローの制御に用いている．そして，DMA によるデータ転送と組合せることにより入出力動作のための通信プロトコルを実現している．これらすべての入出力動作は，OS の管理下で行なわれる．そのため，必要であれば OS のレベルで通信をおこない，他の PE と協調した入出力動作を行うことも可能となっている．

以下では，入出力機構を構成する入出力装置・PE の入出力インタフェース・入出力用結合網それぞれを，RWC-1 ではどのように実現したかについて述べる．

### 3.2 入出力装置

RWC-1 の入出力装置は PE とは直接つながっておらず，入出力用結合網を介して PE と接続されている．このため，入出力装置には入出力用結合網のインタフェースや，入出力動作のための通信プロトコルを実現するコントローラが内蔵され，高機能なものとなっている．

ディスク入出力装置としては，複数のディスクノードから構成されるディスクシステムが実装される．このディスクシステムは並列アクセスが可能であり，また動的に負荷分散が行われる[7]．画像入出力装置としては，NTSC レベルの画像データを複数の PE に分割して送信，また PE から送られてきたデータを 1 枚の画像に統合することが可能なものが実装される．このほか，ユーザインタフェースとなるホストコンピュータや他の計算機，ロボットのセンサ・コントローラ等も接続される予定である．

### 3.3 PE の入出力インタフェース

RWC-1 の PE チップは独自に設計したため，入出力のための様々な機能を PE の入出力インタフェース (RBC) として内蔵することが出来た (図 2)．PE チッ

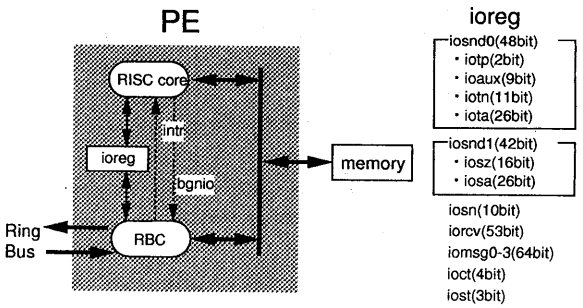


**図 2** PE の入出力インタフェース

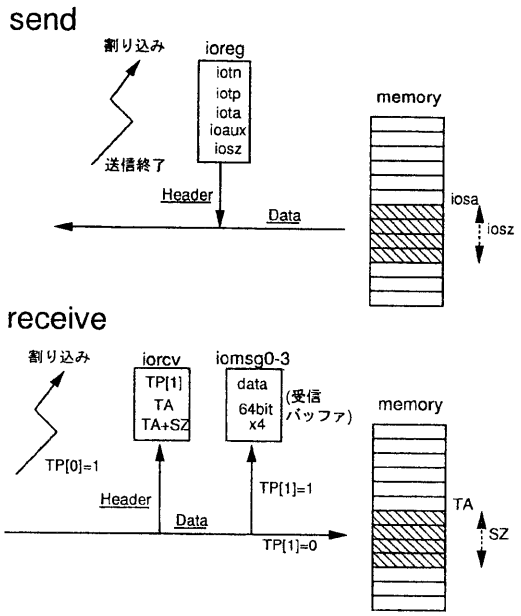


図3 PE の入出力動作

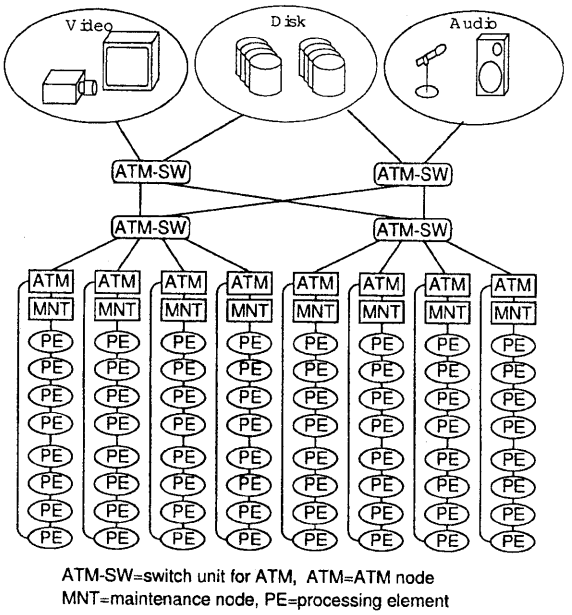


図4 RWC-1 の入出力用結合網

プそのものが入出力機構の一部となっていることで，演算と入出力の並列動作を効率的に行なうことが可能となった．また実装面でも，結合網のインタフェースを内蔵したことにより，必要な周辺回路が少なくなる利点があった．

入出力インタフェースは送信・受信時にそれぞれ次のような入出力動作を行なう (図3)．

送信時

(1) 演算処理部における I/O 命令 (bgnio)[4] の実行で起動
(2) I/O レジスタの値から自動的にヘッダを生成
(3) メモリからデータ本体を自動的に読みだす
(4) ヘッダ・データ本体を転送データとして結合網に送出
(5) 終了時に割り込みを発生

受信時

(1) 結合網から転送データの到着で起動
(2) ヘッダの内容により格納場所，書き込み番地を設定
(3) データ本体をメモリまたはレジスタに自動的に格納
(4) 終了後に割り込みを発生

以上のように，入出力インタフェースは演算処理部と直接やりとりをしつつ，入出力動作を行なっている．このため RWC-1 では素早い入出力の取り扱いが可能となり，ソフトウェアの作成も容易となっている．また入出力インタフェースそのものは，演算処理部とは独立しているため，演算と入出力は並列に動作する．

### 3.4 入出力用結合網

RWC-1 では，入出力のために使用される結合網を演算用結合網[6] と独立させて実装した．これは，演算と入出力がそれぞれ独立した動作をするため，結合網を共通にすると入出力データの転送時間が保証出来ず，時間依存性が高いデータの処理などのスケジューリングが困難になるからである．

RWC-1 の入出力用結合網は，局所性の利用と実装上の制約から 2 レベルに階層化されている (図4)．8 つの PE を接続する下位階層には，

- 制御が比較的簡単
- 必要な転送容量を実現するための信号線数が最も少ない
- 入力と出力が 1 対 1 に対応するために高速動作が可能

の特徴をもつリングバスを用いた．リングバスを採用することにより，宛先を外部入出力装置から PE に変更するだけで，入出力機構を用いた PE から PE へのメモリ間データ転送を簡単に実現することが可能となった．

RWC-1 実機に先立ち，小規模な並列計算機であるテストベット 1 を試作し，入出力に関する実験も行った．テストベッド 1 では，データの転送には固定長のパケットを用い，パケットそのものに制御用・調停用のフラグを持たせ，リング上のマスタノードで調停を行なうようにした．[1] しかし，この実装では調停ノードが複雑になること，転送効率が悪いことが問題となった．

RWC-1 本体では，リング上ではなく外部に調停回路 (アービタ) を設け，すべての制御信号が直接 PE チップとアービタ間に接続されるオーソドックスなものとし

た．制御信号には以下の4種類がある．

- req(4bit): 送信要求 (PE →アービタ)
- ack(1bit): 送信返答 (アービタ→ PE)
- ena(1bit): 受信許可 (アービタ→ PE)
- rdy(1bit): 受信可能 (PE →アービタ)

req は転送先の情報を含み，アービタではこれをデコードし転送先ノードを決定する．もし転送先ノードが受信可能，かつ，転送元ノードから転送先ノードまでの経路で入出力動作中のノードがなければ，転送元ノードに ack，転送先ノードに ena を返し，新たなデータ転送を確立する．この方法を取ることにより，転送経路が重ならなければリングバス上で複数のデータ転送を同時に行うことができ，転送容量を増やすことができた．また，固定長のパケットも改め，転送データは可変長のひと塊のデータとしてバースト転送される．

リングバスによって1つにまとめられた PE クラスタと入出力装置の接続には，

- 転送容量が大きい
- 標準化されているため汎用性・接続性に優れている
- ノード間の接続に光ファイバを用いるため装置の設置が容易

の理由により複数の ATM スイッチからなる間接網 (ATM 網) を用いた．ATM 網では転送データは ATM プロトコルに従い，多数の ATM セルという形で転送される．

さて，入出力用結合網上の上位と下位とではデータ転送プロトコルが異なるため，その変換を行なう必要がある．上位・下位の接続部にある ATM ノードは，転送データをバッファリングし，下位から上位への転送では転送データを分解し，上位から下位への転送では転送データを再構成することにより，データ転送プロトコルの違いを吸収する[2](図5)．

## 4. 性能評価

RWC-1 実装に先立ち，その入出力機構のうち，PE の入出力インタフェース及び入出力用結合網の下位階層であるリングバスについて，そのデータ転送能力の予備評価を行った．モデルにはハードウェア記述言語で記述された PE・ATM ノード・アービタを相互に結合したものを用い，シミュレーションにより PE のシステムクロック単位の精度でデータを採取した．

### 4.1 転送サイズによる評価

PE と ATM ノード間で，データを転送する実験を行った．実験には 32K，8K，2K，512，128，32 バイトの6種類の異なるデータ長の転送データを用い，合計 256K バイトの転送が終了するまでの時間を以下の4つの場合について測定した．

(1) ATM ノードから1つの PE への転送

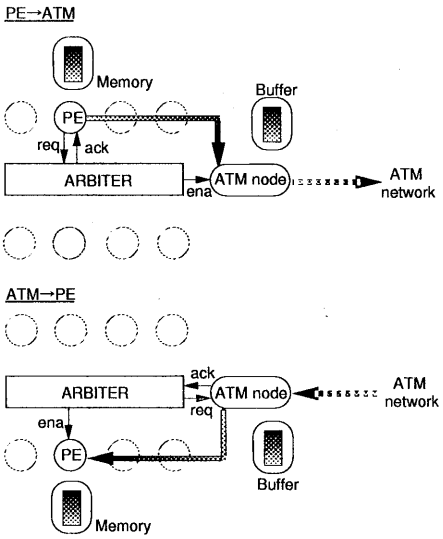


図5 リングバスと ATM ノード

(2) ATM ノードから8つの PE への転送

(3) 1つの PE から ATM ノードへの転送

(4) 8つの PE から ATM ノードへの転送

実験結果は表1の通りで，これをグラフにしたのが図6である．

これらの結果から，1秒当り何バイトのデータを転送できるかを表す転送速度は

- データ長が大きいほど大きい
- PE が送信側の場合には8PE に分散した方が大きい

ことがわかる．

データ長が大きいほど転送速度が大きくなるのは，転送データに占めるヘッダの割合が小さくなることと，転送回数が少ないため必要なオーバヘッドの合計が小さくなるからである．1回の転送に必要なオーバヘッドには，

表1 転送サイズと転送性能

| len | ATM→ 1PE | ATM→ 8PE | 1PE→ ATM | 8PE→ ATM |
|---|---|---|---|---|
| 32K | 5.246 | 5.246 | 5.248 | 5.247 |
| | (49.97) | (49.97) | (49.95) | (49.96) |
| 8K | 5.257 | 5.255 | 5.260 | 5.256 |
| | (49.87) | (49.89) | (49.83) | (49.88) |
| 2K | 5.299 | 5.292 | 5.310 | 5.292 |
| | (49.47) | (49.53) | (49.37) | (49.53) |
| 512 | 5.468 | 5.440 | 5.510 | 5.438 |
| | (47.94) | (48.19) | (47.58) | (48.20) |
| 128 | 6.144 | 6.032 | 6.309 | 6.022 |
| | (42.67) | (43.46) | (41.55) | (43.53) |
| 32 | 8.192 | 8.397 | 9.504 | 8.357 |
| | (32.00) | (31.22) | (27.58) | (31.60) |

上段: 要した時間 (ms)，下段: 転送速度 (MB/s)

アービタによる調停にかかるものと，1つのPEが送信から次の送信，または受信から次の受信に移る際に必要な回復時間がある．受信時の回復時間は，最後のデータを受取ってからバッファの内容をメモリに書き終わるまでであり，比較的小さい値である．これに対して送信の場合，送信終了割り込みが起きてからソフトウェアで次の送信条件を設定し，送信開始しなければならず，メモリから最初のデータをバッファに読み出す時間を加えた回復時間は大きくなってしまう．PEが送信側の場合に1PEに対する転送速度と8 PEに対する転送速度が異なるのは，1PEでは隠蔽できない回復作業を8PEでは他のPEの転送動作時に行うことができ，これを隠蔽することが可能だからである．

次にこの結果から画像データとディスク操作時におけるファイルデータの転送性能を考察する．

640 × 480の24bitフルカラーの画像データのサイズは1Mバイト弱であるが，これを30分の1秒周期で転送ことを考えると，30Mバイト / 秒の転送速度が必要である．1つの転送データのデータ長を32Kバイトとすると，画像の入力 (ATM → PE) についても出力 (PE → ATM) についても，あるいは1対1であっても8つのPEに振り分けても，十分余裕をもって転送することが可能であることがわかる．ただし，ハイビジョンなど50Mバイト / 秒を越える転送速度が必要とされるデータの転送には，複数のクラスタに分割することが必要となる．

ディスク操作に用いられる転送データのデータ長は，ブロックサイズと同じであるため，通常512から2Kバイトである．実験結果より，この程度のデータ長であれは転送速度の低下はほとんどないことがわかる．

次に送信間隔の限界を考える．2Kバイトの転送データを送出するには約40マイクロ秒が必要であるため，1つのノードから可能な送出回数は1秒間に約25,000回である．ATMノードは送信と受信を同時に行えるため，転送経路が重ならなければ，PEからATMノード，ATMノードからPEへの転送をそれぞれ約25,000回実行できる．実際には重ならない組合わせは28/64通りであり，また8PEに対して転送を行っているため，1PEあたりの送信・受信回数の限界を計算すると，それぞれ1秒間に約2,200回となる．これをクロックに換算すると約22,000クロックに1回の転送となる．同様に512バイトでは約8,900回 (5,600クロック) が限界となる．このときの転送速度はどちらも送受信共に約4.4Mバイト / 秒である．これはディスクの転送速度が数Mバイト / 秒であること，常にすべてのPEがディスク操作を行っているわけではないことを考えると十分なものといえる．

### 4.2 送信間隔による評価

RWC-1では宛先を入出力装置ではなくPEにすることにより，入出力用結合網を通したPE同士の通信も可能である．その場合，送信間隔がどの程度までなら十分

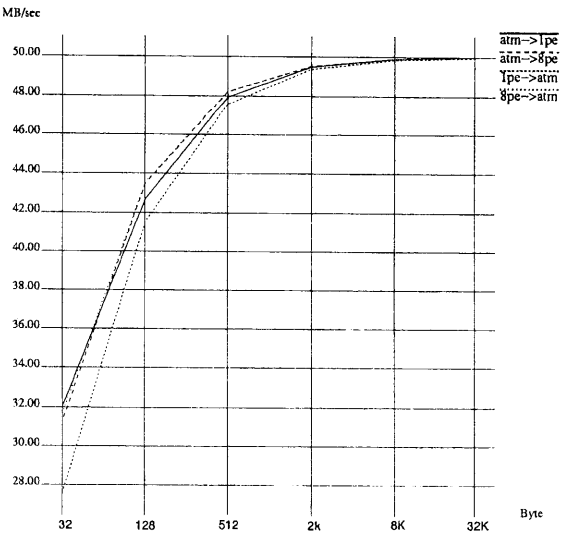


図6 転送サイズと転送速度

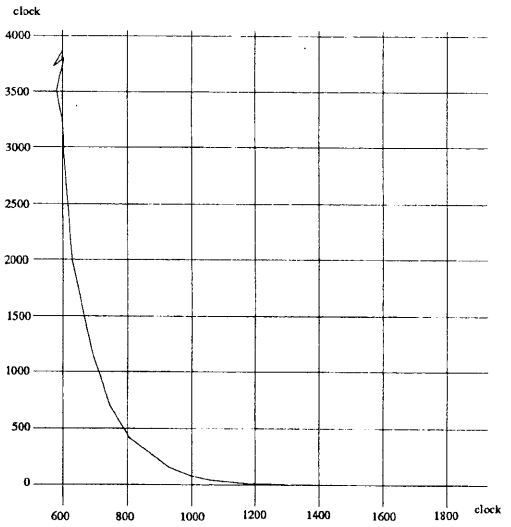


図7 送信間隔と遅延時間

な性能が得られるのかを調べる実験を行った．実験では転送はクラスタ内のPEからPEへのものについて行い，宛先をランダムに，データ長は32バイトから2Kバイトまでの範囲でランダムに設定した．その上で，送信要求から次の送信要求が起こるまでの時間である送信間隔を変更して，送信要求から送信返答までの時間である遅延時間がどのように変化するかを測定した．

図7に測定結果を示す．送信間隔が1300クロック辺りまでは，遅延時間はほとんど発生しない．その後遅延時間は増加し，600クロック辺りで壁となる．これ以下の送信間隔では，リングバスが飽和してデータを送出できない．逆にある程度送信間隔が小さければ，リングバスはPE間のランダムな通信にも使えることがわかった．

この入出力機構を用いたPE間のデータ転送は，(1)転送先のメモリに直接書き込む，(2)演算と並行して動作する，などの利点がある．このため，PE間でのデータのスワップやマイグレーションなど，ある程度サイズが大きいデータのメモリからメモリへの転送を効率よく実行することが可能であると考えている．

## 5. おわりに

我々は，超並列計算機においてシステム全体の性能を向上させるためには，演算と入出力がお互いに干渉することなく並列動作することが重要であると考え，RWC-1の入出力機構においてこれを実証中である．そのために，演算用とは独立したリングバスとATM網からなる入出力用結合網を設け，またPEチップにはメモリアクセス機構などの入出力用インタフェースを内蔵するなどの実装を行った．リングバスは，小さなハードウェアコストで実装したにも関わらず，入出力専用としたために入出力のためのデータの性質に合った十分な性能をもつことが確認できた．

RWC-1は現在，改良版PEチップを開発中であり，これが出来上がると，すでに開発が終り現在調整中の画像入出力装置及びATMノードとともに，128台のPEからなるRWC-1基本部が完成する．この上で実験・評価を行い，その結果をもとに最終的に1,024台に拡張する予定である．

**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，TRC超並列研究部各研究室の室長及び室員の諸氏，ならびに超並列三洋分散研の諸氏に感謝いたします．

## 参考文献

1) 廣野英雄, 松岡浩司, 岡本一晃, 横田隆史, 坂井修一. RWC-1の入出力リングバス. 情処学研報, ARC-113-16, pp. 121–128, Aug. 1995.
2) 廣野英雄, 松岡浩司, 岡本一晃, 横田隆史, 坂井修一. RWC-1の入出力用ATMノード. 第52回情処全大, pp. (6)153–154, 1996.
3) 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.
4) 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.
5) S.Sakai, K.Okamoto, Y.Kodama, and M.Sato. Reduced interprocessor-communication architecture for supporting programming models. In *Proc. of Programming Models for Massively Parallel Computers 1993*, pp. 134–143, 1993.
6) 横田隆史, 松岡浩司, 岡本一晃, 廣野英雄, 坂井修一. RWC-1の階層型mdce網. 情処学研報, ARC-113-11, pp. 81–88, Aug. 1995.
7) Y.Oue, T.Kitamura, K.Ohnishi, and M.Shimizu. Parallel file access for dynamic load balancing on the massively parallel computer. In *Proc. of Int'l. Symp. on Parallel and Distributed Supercomputing*, pp. 179–187, 1995.
8) 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.
9) 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.