計算機アーキテクチャ 122－9
（1997. 2. 4）

超並列計算機RWC-1における光インターコネクション

松岡浩司[†1] 横田隆史[†1] 坂井修一[†2,†3]
西村信治[†4] 吉川隆士[†5]

超並列計算機の実装では，LSI間とボード間接続のピンボトルネックが最も大きな問題となる．RWC-1では，PEノードを複数のLSIから構成し，プロセッサ間結合網としてMDCE(Multi-Dimensional Directed Cycles Ensemble)網を採用することによって，これらのピンボトルネックがもたらす性能上の影響を小さくしている．光インターコネクションの導入により，主に，ボード間接続の実装上の問題を解決するこができ，より簡便に大規模な超並列計算機システムを構築することが可能となる．

Optical Interconnections on Massivelly Parallel Computer RWC-1

HIROSHI MATSUOKA[†1], TAKASHI YOKOTA[†1], SHUICHI SAKAI[†3],
SHINJI NISHIMURA[†4] and TAKASHI YOSHIKAWA[†5]

The most serious problem on implementation of massivelly parallel computer systems is a pin bottleneck one which is observed on inter-connections between LSIs/PCBs. RWC-1 PE(processing element) nodes are consist of several LSI chips to avoid influence of the LSI pin bottleneck problem. And also, RWC-1 adopts MDCE(Multi-Dimensional Directed Cycles Ensemble) network for inter-processor communicatios to decrease influence of the board connector pin bottleneck problem. Adopting optical interconnection technologies, RWC-1 not only breaks through these inplementation limitations and also makes its implementation easy to construct a large scale computers.

1. はじめに

データを共有しながら処理を進める並列処理では，プロセッサ間でいかに効率的にデータをやりとりするかが重要な課題となる．超並列計算機RWC-1はプロセッサアーキテクチャとしてRICA(Reduced Inter-processor Communication Architecture)[1)2)3)]を採用し通信の遅延を隠蔽するだけではなく，プロセッサ間結合網としてMDCE(Multi-Dimensional Directed Cycles Ensemble)網[4)]を採用することにより，高いスループットを維持つつ通信の遅延そのもの小さくしている．

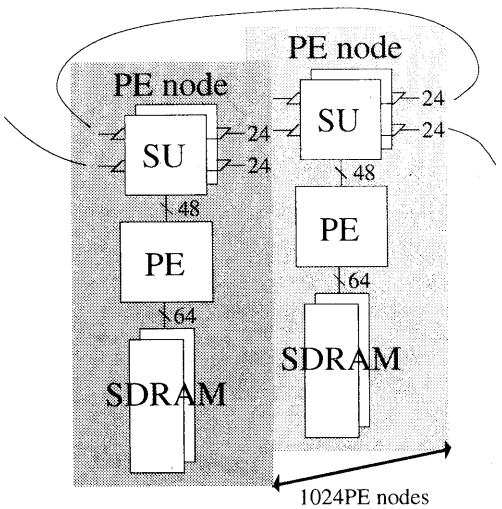
高いスループットを維持しつつ，通信の遅延の小さな超並列処理に適した結合網を構成するためには，実装上さまざまな制約がある．本稿では，まず，RWC-1における結合網の実装を概説し，その問題点について述べる．次に，光インターコネクションの適用することによって，その問題点を解決する方式について述べ，超並列計算機に光インターコネクションを適用するに当たっての課題について論じる．

†1 (技組) 新情報処理開発機構 つくば研究センタ
Tsukuba Research Center, Real World Computing Partnership
†2 筑波大学 電子・情報工学系
Institute of Information Sciences and Electronics, University of Tsukuba
†3 電子技術総合研究所
Electrotechnical Laboratory
†4 (技組) 新情報処理開発機構 光日立研究室
Optoelectronics Hitachi Laboratory, Real World Computing Partnership
†5 (技組) 新情報処理開発機構 光NEC研究室
Optoelectronics NEC Laboratory, Real World Computing Partnership

2. 超並列計算機RWC-1

2.1 RWC-1の構成

図1に示すように，RWC-1は1024のPE(Processing Element)ノードから構成される超並列計算機である．各PEノードは命令を実行するプロセッサ(PE)とプロセッサ間通信を制御するスイッチングユニット(SU)[5)]から構成されている．SUはプロセッサ間通信のスループットを高めるために，さらに，データ幅方向にスライス(2分割)されている．各PEノードは32MBのメモリ有し，メモリチップには高いデータ供給能力を持つSynchronous DRAMを採用している．**図2**にボード上におけるPEノードの実装状況を示す．ボード上に



performance@1024nodes:
100-132GIPS(peak)
50-66GFLOPS(peak)
memory 32-64GB(total)

processing element:
RICA: Reduced Inter-Processor Communication Architecture
internal clock:50-66MHz
super scalar execution:100-132MIPS(peak)
internal FPU:50-66Mflops(peak)

inter-connection network:
MDCE: Multi-dimensional Directed Cycles Ensemble
100-132MTrans/sec(300-396MB/sec)/port

図1 RWC-1の構成
Fig. 1 Configuration of RWC-1

は4つのPEノードが実装されており，中央のPEとSDRAM用のSIMコネクタを挟み込むように2つのSUが実装されている．RWC-1では，このようなPEノードがMDCE網と呼ばれる直接網によって接続されている．

2.2 PEノードの動作

超並列計算機では，数多くのPEノードを接続する必要があるため，システムが物理的に大きくなってしまう．したがって，従来の計算機のようにシステム全体を同期して動作させることが難しく，RWC-1では，各PEノードを非同期に動作させている．各PEノードを非同期に動作させるために，プロセッサ間結合網を構成するSUは，複数の入力に到達したそれぞれのデータの位相を内部の動作クロックに同期化する機構を有している．なお，RWC-1では，全てのクロックが1つの基本クロックから生成されるので，厳密な意味では，非同期とは言えない．ここでいう非同期とは，複数のクロック間で位相関係が保証されないという意味である．また，例えば，同一の基板上に実装された複数のプロセッサを敢えて非同期に動作させる必要は無く，これらのプロセッサはほぼ同期して動作する．

2.3 プロセッサ間結合網

データの転送を双方向に行なおうとすると，データの転送方向を変更する時にデータを転送できない時間が生じ，実効的なスループットが低下する．このような問題が生じないようにするため，RWC-1では，データを一方方向にのみ伝搬させている．

図3に示すように，RWC-1でプロセッサ結合網に採用したMDCE網の基本構造はデータが一方方向にのみ伝搬する単方向リングである．2次元方向に見ると，基本構造の単方向リングが，それと交差する別の単方向リングによって接続されている．このような構造の結合網は，特に，circular Banyan網と呼ばれる．さらに，そのcircular Banyan網が3次元方向の単方向リングによって接続されたものがMDCE網となる．この網を3次元方向に見ると，CCC網を構成している．

このようなMDCE網の特徴の1つは結合網の次数，つまり，各PEから入出力しなければならないリンクの数を少なくできる点にある．3次元のMDCE網の次数は3で，例えば，3次元トーラスの場合，その次数は6である．結合網の次数は，直接的に，網を実装するために必要となる信号線の数を反映しており，同一のハードウェアコストをかけた場合，MDCE網の方がデータの幅(つまり，リンク当たりのスループット)を2倍とすることができる．実装面から考えると，次数が小さく，データが目的地に到達するまでに要する時間を表す直径が小さい網ほど，超並列向けの結合網と言える．MDCE網は，従来提案されている直接網の中で最もこの条件を良く満たしている[6]．

2.4 実装上の制約とチップ分割

超並列計算機の実装で最も大きな問題となるのはLSI間およびボード間接続のピンボトルネックである．現状で，LSIからは500，ボードからは2000程度の信号しか取り出すことができない．

LSIやボードから引き出すことのできる信号の数は用いるコネクタの接点の形状によって決まり，機械的な接点を必要とするので，将来的にも飛躍的に多くなるとは考えられない．したがって，これらのピンボトルネックを回避するためにはアーキテクチャ的な対策を講じることが必須であり，方式として以下のようなものが考えられている．

- 集積化
  複数のLSIによる実装では，LSIを相互に接続する配線が必要である．全ての機能を1つのLSIに集積できれば，LSI間の配線のためにピンを使用する必要がなくなり，ピンボトルネックは緩和される．プロセッサ/メモリ間接続が好例で，例えば，MCM(multi chip module)上にプロセッサとメモリを実装すれば，MCMのピン数を少なくできる．

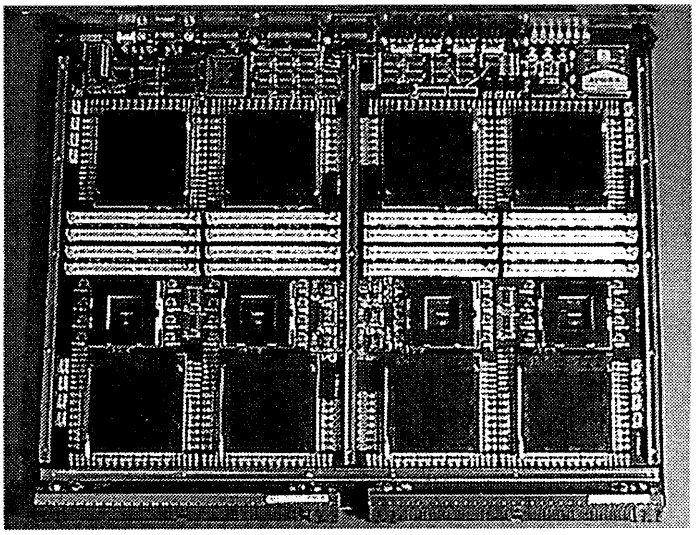


図2 PEノードの実装
Fig. 2 Implementation of PE nodes

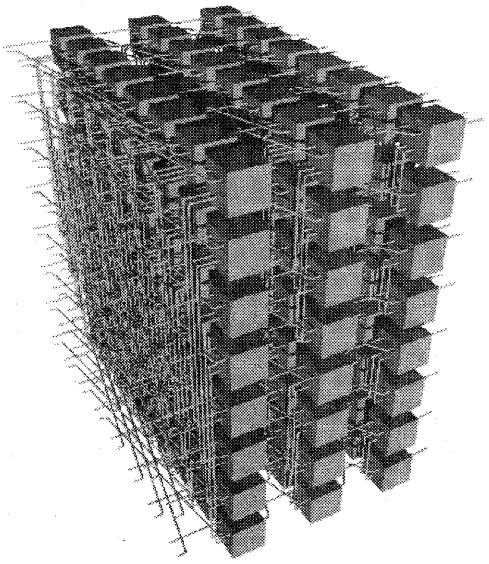


図3 RWC-1のMDCE網
Fig. 3 MDCE Network for RWC-1

つまり，同じ数のピンを使うならば，プロセッサ間通信のスループットをより高くできる．

- データ幅方向の分割(スライス)
  特定の機能を分割しようとすると，分割された部分同士を接続するために多くの配線を必要とする場合が多い．スライスは，データ幅が，例えば，1/2の2つのLSIによって，1つの機能を実現しようとするものである．この場合，個々のLSI上に全く同じ制御回路を必要とするが，2つのLSIを相互に接続する必要は基本的には無い．データ幅を1/2とすることによって，個々のLSIのピン数を1/2程度に削減できる．
- データの多重化
  超並列計算機においてピンボトルネックが問題になるのは，主として，高いプロセッサ間通信のスループットを確保することに原因が求められる．スループットを高めるためには，データ転送の幅を大きくとるか，あるいは，データ転送の周期を短くすれば良い．論理振幅を小さく押えることによってより短い周期のデータ転送を行なおうとする方式がいくつか提案されている．入出力のデータ転送周期を，例えば，1/2とすれば，LSIのピン数を1/2程度に削減できる．ただし，LSI外部の回路を高速で安定動作させためには技術的な課題が多い．

集積化を行なえばボード上におけるPEノードの占有面積を小さくできるため，結果が実装上の制限内に収まるならば，集積化によるピンボトルネックの回避が最も有効である．しかし，現状では，LSIのピンボトルネックのためにプロセッサ間通信のスループットを低くせざるを得ず，パスの競合による待ちが生じ，性能上の問題となる可能性がある．超並列計算のためにどの程度のスループットを維持するべきかと言った議論が必要であるが，RWC-1では，PEノードを複数のLSIによって構成することによって，高いプロセッサ間通信のスループットを実現している．

前節でも述べたように，RWC-1では，まず，PE

ノードを演算処理を実行するPEとプロセッサ間通信を制御するSUに分割している．このような機能分割を行なうと，PEにはプロセッサ通信のパスが無くなり，メモリアクセスパスのデータ幅大きくとることができる．一方，SU側には大きなメモリアクセスパスが無いので，LSIのピンの多くをプロセッサ間通信のパスに充当することができる．また，SUはデータ幅方向に2つのLSIに分割されている．このようなデータ幅方向のスライスを行なうことによって，個々のLSIが扱わなければならないデータの幅を小さくすることができる．さらにSUは内部の動作クロックの1/2の周期で入出力を行なうようになっている．このデータの多重化は，LSIのピンボトルネックを回避するだけではなく，ボード間接続のコネクタのピンボトルネックを回避するのにも寄与している．

2.5 電気インターコネクション

データ転送のスループットを高めるためには，データ転送の周期をできるだけ短くしなければならない．RWC-1のプロセッサ間データ転送，具体的には，SU-SU間データ転送の周期は10nsとなっている(データ転送周波数にして100MHz)．これは，開発当初一般的に利用することのできたLSIの高速入出力インターフェースが通信用途に対応して156MHzまでの動作を保証していたためである．通信用途で扱う高速の信号は少なく，これらのLSIも少数の信号ならば156MHzでの転送が可能であるが，RWC-1ではデータ幅が広いため，若干余裕を持たせた設計となっている．

一方，RWC-1では，データ伝送距離が最大で10m程度になる．線路を信号が伝搬する速さは光速の1/2程度であるから，送信点から送出された信号が10mの線路を経由して受信点に到達するまでには数10nsの時間を要する．通常のデータ転送では，1転送周期に1度だけ信号のレベルが変化するが，データ転送周期がデータの伝搬する時間に対し相対的に短いと，線路上で複数回信号のレベルが変動することになる．このようなデータ転送では，データ転送の経路を分布定数線路とし，終端で整合をとり信号の反射を抑制する必要がある．RWC-1では，高速で動作し終端整合をとることのできる入出力インターフェースとしてECLを採用している．また，基板上の配線はストリップラインを形成している．

また，安定したデータ転送を実現するために，RWC-1では，差動データ転送方式を採用している．差動データ転送方式では，1つの信号を運ぶために2本の信号線が必要になるため，一般的なsingle endデータ転送の2倍の信号線が必要となる．しかし，物理的に離れた複数のきょう体間で基準電位を一定に保つことは技術的に難しく，single endデータ転送でも，基準電位用の接続が多く必要で，配線数でのデメリットはあまり大きくない．むしろ，きょう体間の配線は外部の雑音環境に曝されるため，対雑音対策が重要であり，雑音に強い差動データ転送の方がシステムを安定に動作させるために有利である．

3. 実装上の問題点と光インターコネクション

エッジボードコネクタを用いたボード間接続では，ボードから入出力できる信号線の数はボードの物理的な大きさによって決まり，このボトルネックを回避することは難しい．光インターコネクションを導入することによって，このボード間接続のボトルネックを回避することは可能である．しかし，単純にプロセッサ間通信のスループットを高めようとすると，LSIのピンボトルネックの問題が顕著となりシステムの構築が難しい．具体的には，LSIのピンボトルネックの問題はチップ分割によって対応せざるを得ず，これはPEノードの実装面積を大きくするため，ボード上に実装できるPEノードの数が少なくなり，スループットを高くする必要がなくなるといった矛盾が生じる．また，光インターコネクションでは電気-光インターフェースをボード上に実装する必要があって，これもまた実装面積を大きくする方向に働くことにも留意しなければならない．本節では，RWC-1に光インターコネクションを導入するに当たっての利点を電気インターコネクションの実装の問題点の観点から述べ，光インターコネクションの実装面積上の問題をどのように解決すべきかを論じる．

3.1 エッジボードコネクタのピンボトルネック

1枚のボードから導出できる信号線の数は，ボードのエッジの長さとエッジボードコネクタのピン密度によって決まる．コネクタでは機械的に接点を形成することが必要であり，現在，1.27mmピッチまでのものが実用されている．RWC-1では，後述するきょう体間接続の問題から2mmピッチのコネクタを採用している．

コネクタのピッチが決まると，コネクタのピン密度を上げるためには，ボード垂直方向に並ぶピンの極数を多くとれば良い．ところが，ボード間を接続するマザーボードの立場でこれを見ると，ボードを貫通する穴が多くなり，配線領域が確保できないといった問題が生じる．特に，マザーボード内にストリップラインを形成する必要のある高速データ転送ではこの問題は致命的で，RWC-1では，5極のコネクタを採用するにあたりマザーボード上のコネクタの配置に詳細な検討が必要であった．

マザーボード上の配線を削減するために，大きな効果があったのが，対向コネクタの採用である．通常のコネクタを用いた実装では，きょう体間を跨ぐ配線は，PEボードから一旦エッジボードコネクタを介して取り出され，マザーボード上のきょう体間配線用のコネクタに導かれる．このような実装ではマザーボード上を引き回さなければならない信号線数が多い．対向コネクタは，物理的には，マザーボードに打ち込まれたピンの集合である．PEボードはこのピンをオスコネクタとするエッジボードコネクタで接続され，きょう体間の配線はこのコ

ネクタをオスコネクタするケーブルコネクタで接続される．つまり，ボードときょう体間接続のケーブルを直接接続することができ，マザーボード内の配線の数を削減できる．

以上のような実装上の制約から，RWC-1では，1枚のボードから約2000の信号を入出力している．なお，マザーボードの逆側にもコネクタを実装すれば，倍の信号を入出力することができるが，システムのメインテナンスが困難になるという観点からこのような実装は採用していない．一方，光インターコネクションでは，機械的な接点を形成する必要がないため，原理的に，コネクタの密度を高めることができる．しかし，光の損失を押えるためには，ファイバの位置を高精度で固定する必要があって，加工精度上の問題から多極(例えば，100極)のコネクタはコスト的に問題があると考えられる．RWC-1では，小さな(12極)のコネクタを並べて，ある程度可動できる複数のコネクタを一括して脱着する方式を検討している．通信用途のコネクタを流用しても，電気インターコネクションと同等の線数を扱うことが可能である．さらに，コネクタの2次元方向の高密度実装を検討すれば4倍程度の線数を容易に実現できる．

3.2 コネクタの挿入圧

ボードから入出力する信号線の数を多くするためには，今1つ大きな問題がある．それはボードの挿入圧の問題である．

コネクタは電気的な接続を確実なものとするため，メスコネクタのバネ状の接点がオスコネクタのピンを大きな圧力で挟み込んでいる．ボードの挿抜は挟み込む方向に垂直に行なわれるが，摩擦の問題から，極数が多くなるとボードの挿抜にはかなり大きな力を必要とする．RWC-1では，ボードの挿抜に30Kg程度の力を必要とし，ボード先端に穴を空け，その穴に特殊な治具を挿入し，てこの原理でフレームからボードを引き抜くようになっている．強度的にこの程度が限界であり，これ以上のピン数の増加はボードの挿入圧の問題からも困難である．なお，挿入圧の問題はLSIのソケットと同様なfriction freeのコネクタを採用することで回避できるが，きょう体の奥に位置するコネクタの固定/解放機構をいかに動作させるかなど問題が多くRWC-1では採用していない．

光インターコネクションにおけるファイバの接続では，ファイバの心線の位置関係のみが問題であり，原理的に，接続するために圧力をかける必要はない．このため，接続すべき光ファイバの数が多くなっても，電気インターコネクションにおけるコネクタ挿入圧のような問題は生じない．

3.3 長距離データ転送

きょう体間を結ぶ長距離データ転送では，直流損失などが問題となり，10mを越えるようなデータ転送ができない．光インターコネクションでは，この程度の距離での損失はほとんど問題とならず，より長い距離でのデータ転送が可能である．

3.4 終端抵抗

RWC-1で採用したSUチップのパッケージは，基本的にはPGAであるが，高周波的な信号伝搬特性を劣化させないように100milピッチの格子の外周部にのみピンが配置されている．このためSUチップの実装面積は約10cm□と大きく，RWC-1では，ボード上に4つのPEしか搭載できなかった(図2参照)．SUチップの実装面積を削減するためには，信号伝搬特性が許す限り密な信号ピン配置を取れば良い．例えば，SUは409ピンのパッケージに実装されているが，ピンを千鳥格子上に配置すれば7cm□各程度に小さくできる．しかし，このように密に配置されたピンから外部に信号を引き出すことが実際には困難である．

LSIからの配線の引き出しの問題では，終端抵抗の存在が致命的に大きい．終端抵抗は信号の受信端の直近に配置しなければならないが，受信端はLSIであり，大きな発熱がある終端抵抗をLSI上に形成することはできないので，通常，LSIパッケージの近傍に終端抵抗を配置する．つまり，信号は到達すべきピンに配線され，さらに，そこから延長され終端抵抗に接続される．終端抵抗の無い場合と比較してLSI近傍の配線数は2倍になるため，ピンの密度を上げることが難しい．

光インターコネクションを採用することによって，物理的に離れたPEノード間のデータ転送を行なう必要がないので，終端を必要とするデータ転送を行なう必要がない．このため，より小さなパッケージの採用が可能である．それだけではなく，終端抵抗用のチップ抵抗が不用になるため，さらに面積が削減できるだけではなく，システムの製作が容易になる．

さらに，データ転送の距離が短ければ，差動データ転送を行なう必要がなく，電気によるデータ転送は簡素化できる．これによって，SUのピン数を1/2程度に削減できSUの実装面積のさらなる削減が可能である．

4. 光版RWC-1

光インターコネクションの有効性を実装するためには，光インターコネクションの持つ性能を最大限に発揮できる構成のシステムを構築しなけらばならない．このようなシステムを構築すれば，少なくともプロセッサ間通信のスループットにおいては，1桁程度の高い性能を持つシステムを構築することが可能だと考えている．しかし，このような光インターコネクションが中心となる超並列計算機システムの開発は，例えば，開発期間が長くなるといった問題が想定され，現実的ではない．そこで，我々は，電気インターコネクションを有するシステムのデータ転送経路を光インターコネクションに置き換えるアプローチを採用している．このようなアプローチでは，電気インターコネクションによって実現できる性能しか得られないが，例えば，同じ性能を得るための実

装の効率を高めることができる．また，実際にシステムに光インターコネクションを組み込むことによって，光インターコネクションを組み込むに当たっての問題点を明確にし，今後の研究の指針を検討することができる

RWC-1における電気インターコネクションを光インターコネクションに置き換えたシステムを我々は光版RWC-1と呼んでいる．RWC-1と光版RWC-1の主な相違点は，ECL作動インターフェースを有するスイッチングユニット(SU)を小振幅インターフェースを有するスイッチングユニットに置き換えた点である．信号インターフェースの変更によりLSIから引き出さなければならない信号線数を削減でき，より小さなパッケージを採用できる．また，終端抵抗を無くすことができるので，電気-光変換モジュールを実装するための面積を確保することができ，実装上の問題が生じないと考えている．

5. まとめ

本稿では，超並列計算機RWC-1の実装とRWC-1に対する光インターコネクションについて述べた．RWC-1では，LSIのピンボトルネックを回避するために，PEノードを複数のLSIから構成している．これによって，きょう体間でも安定した通信が可能なECL差動データ転送を採用することが可能となった．しかし，ECL差動データ転送では，終端抵抗を必要とするなど実装上の問題も多い．これらの実装上の問題は，光インターコネクションを採用することによって，回避することができる．特に大きな問題であるエッジボードコネクタのピンボトルネックの問題を容易に回避できるだけではなく，より簡便に大規模な計算機システムを構築することが可能となる．

RWC-1は電気インターコネクションを適用することを前提にシステムが設計されている．光インターコネクションを導入するにあたって，機能的な拡張が必要であるが，多くは，電気インターコネクションで必要とした機能を流用することが可能である．例えば，SUが有する同期化機構は，光インターコネクションでデータ多重を行なう場合，そのバッファ上で実現することができる．今後の課題としては，システムの機能分割を再検討し，光インターコネクションの可能性を最大限に引き出すことができるシステムを構築する必要があると考えている．また，光インターコネクションを用いれば，従来特殊な実装技術を駆使して始めて得られた性能を容易に手に入れることができる．この意味では，より汎用的な光インターコネクション技術の開発が不可欠であり，RWC-1での実装をふまえ，(超)並列計算機用の光インターコネクションモジュールに関する提案を行なっていく予定である．

謝辞 本研究を遂行するにあたり，有益な御指導，御討論をいただきました島田つくば研究所長，超並列アーキテクチャ研究室の諸氏，光日立研究室，光NEC研究室の担当諸氏に感謝いたします．

参考文献

1) Sakai, S., Matsuoka, H., Okamoto, K., Yokota, T., Hirono, H., Kodama, Y. and Sato, M.: RWC-1 Massively Parallel Architecture, *Proc. High Performance Computing Conference '94*, pp. 33–38 (1994).

2) Sakai, S., Kodama, Y., Sato, M., Shaw, A., Matsuoka, H., Hirono, H., Okamoto, K. and Yokota, T.: Reduced Interprocessor-communication Architecture and its Implementation on EM-4, *Parallel Computing*, Vol. 21, pp. 753–769 (1995).

3) Okamoto, K., Sakai, S., Matsuoka, H., Yokota, T. and Hirono, H.: Multithread Execution Mechanisms on RICA-1 for Massively Parallel Computation, *Proc. Parallel Architectures and Compilation Techniques (PACT'96)*, pp. 116–121 (1996).

4) Yokota, T., Matsuoka, H., Okamoto, K., Hirono, H. and Sakai, S.: A High-Performance Router Design for the Massively Parallel Computer RWC-1, *Proc. Hot Interconnects IV Symposium*, Stanford University, Palo Alto, CA, pp. 1–12 (1996).

5) Yokota, T., Matsuoka, H., Okamoto, K., Hirono, H., Hori, A. and Sakai, S.: A Prototype Router for the Massively Parallel Computer RWC-1, *Proc. Int. Conf. on Computer Design (ICCD'95)*, Austin, Texas, pp. 279–284 (1995).

6) Yokota, T., Matsuoka, H., Okamoto, K., Hirono, H., Hori, A. and Sakai, S.: The Multidimensional Directed Cycles Ensemble Networks for a Multithreaded Architecture, *Proc. Int. Conf. on High Performance Computing (HiPC'95)*, New Delhi, India, pp. 355–360 (1995).