ハイパフォーマンス
コンピューティング 67－10
（1997．8．19）

超並列計算機 CP-PACS のネットワーク性能評価

松原正純† 板倉憲一† 朴 泰祐†
中村 宏†† 中澤喜三郎†††

超並列計算機 CP-PACS は，プロセッサ間ネットワーク・トポロジとして3次元ハイパクロスバ網を採用し，またメッセージ転送プロトコルとしてリモート DMA 転送を実装している．本研究では，この2つの特徴に焦点を置いて，様々な転送パターンに対する CP-PACS のネットワーク性能を実測により評価する．

測定結果より，ハイパクロスバ網は従来よりシミュレーションなどによって評価されてきたように高い性能を発揮し，またリモート DMA 転送は大規模科学技術計算には最適な転送プロトコルであることが分かった．さらに，実アプリケーションにおいても，CP-PACS のネットワークは高性能を示すことが期待できる．

Performance Evaluation of CP-PACS' Interconnection Network

MASAZUMI MATSUBARA,† KEN'ICHI ITAKURA,†
TAISUKE BOKU,† HIROSHI NAKAMURA††
and KISABURO NAKAZAWA†††

CP-PACS, a massively parallel processor, is equipped with the 3-dimensional Hyper-Crossbar Network as an interconnection network and introduces a fast data transfer protocol named Remote-DMA transfer. In this research, we measure and evaluate the practical network performance using various data transfer patterns on CP-PACS.

As a result, we conclude that the network performance of CP-PACS is very high as confirmed by simulation so far, and it can be expected to solve large scale scientific problems efficiently on CP-PACS.

1. はじめに

CP-PACS（Computational Physics by Parallel Array Computer System）[1]は計算物理学などの大規模科学技術計算を対象とした超並列計算機であり，現在筑波大学の計算物理学研究センターに於いて稼働中である．

この CP-PACS のプロセッサ間相互結合網は，ネットワーク・トポロジとして3次元ハイパクロスバ網，及び高速通信機構としてリモート DMA 転送という性能向上のための2つの大きな特徴を持っている．

これまでも両特徴について理論的に高性能であることが示されてきた[1]．特にハイパクロスバ網については，机上計算やシミュレーションによりその性能が評価されている[2]．

しかし，実機でハイパクロスバ網が採用されたのは CP-PACS が初めてで，また今日までそれまでのシミュレーション結果が正しいか，または理論値のどこまでの性能を達成できるのかなどといったことに対する厳密な実測による評価は行なわれてこなかった．机上計算やシミュレーションなどは，パラメータの一部を理想化していることが多く，実測による評価というのはできる限り行なう必要がある．

そこで本研究では，CP-PACS のネットワークについて実測に基づく基本性能評価を行ない，その有効性について考察する．この研究結果は，後々実アプリケーションのチューニング手法などに反映できるものと考えている．

2. CP-PACS のネットワーク構成

CP-PACS は QCD 問題を解くことを目的としている[1]が，その他多種多様な科学技術計算にも適用する

† 筑波大学 電子・情報工学系
Institute of Information Sciences and Electronics, University of Tsukuba
†† 東京大学 先端科学技術研究センター
Research Center for Advanced Science and Technology, University of Tokyo
††† 電気通信大学 情報工学科
Department of Computer Science, University of Electro-Communications

ことを考慮して開発された．そこでネットワーク・トポロジとしては，広範な転送パターンに柔軟に対応できる3次元ハイパクロスバ網（以下，HXBと略す）が採用されている（図1）．CP-PACSでは，2048台の計算専用ノード以外に，分散磁気ディスク記憶装置などを接続するためのアダプタが備わっている入出力ノードが128台あるので，8×17×16の3次元HXBとなっている．

CP-PACSで用いているHXBのリンク当たりの最大スループットは300MB/secである．このネットワーク上でメッセージパッシングによるノード間でのデータ交換が行なわれる．メッセージはwormholeルーティングによって転送され，この際固定ルーティング方式によりメッセージの通る経路を決定する．

また，高速なメッセージ転送立ち上げ及び転送スループットの向上のために，通常の転送モードの他に，ユーザ・アプリケーションから直接ネットワークへのメッセージ転送起動が行なえる高速転送モードが用意されている．この高速転送モードのことをリモートDMA転送と呼ぶ．

以降では，HXB及びリモートDMA転送について詳述する．

2.1 ハイパクロスバ網

3次元HXBは，全ノードを3次元格子状に配置し，それらをクロスバ（以下，XBと略す）とエクスチェンジャ（以下，EXと略す）で相互結合したものである．XB，EXは共にクロスバ・スイッチによって構成される．EXはルータ機能を持っており，メッセージの送受信ノードのアドレスが2次元方向以上で異なる場合は，各次元方向のXBの乗り換えを，XBの交点にあるEXを介して行なう．

HXBは隣接メッセージ転送を無衝突で行なえるだけでなく，同一サイズ，もしくはサイズが異なる場合でも，総ノード数が等しいかより小さいメッシュ／トーラス・ネットワークもエミュレートすることができる．また，クロスバ・スイッチを多次元化して用いているため，一般的な直接網に比べ，メッセージ衝突が存在する場合におけるバンド幅が非常に高いことが期待される．この点についてはシミュレーションを中心とする評価が既に行なわれている[2]．

2.2 リモートDMA転送

リモートDMA転送（以下，RDMA転送と略す）は，送受信双方のノードで通信データ領域となる物理メモリ領域をユーザプロセスの仮想アドレス空間に固定的に割り付けておき，その間で直接データ転送を行なう高速通信方式である（図2）．各ノードのユーザ仮想アドレス空間にマッピングされた物理メモリ同士で

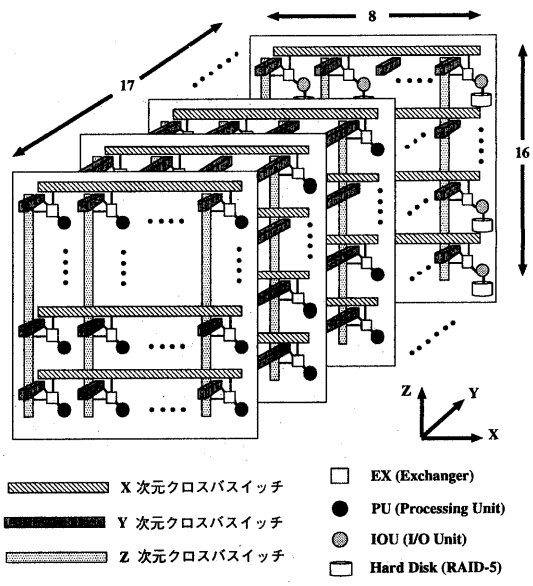

図1　3次元ハイパクロスバ網（8×17×16）

直接データ転送を行なうため，カーネル空間とユーザ空間との間でのデータコピーが発生せず，ネットワークの性能をそのままユーザに提供することが出来る．よって特に大きなメッセージを転送する場合に効率が良い．さらに，RDMA転送は軽いシステムコールによる起動が可能なので，転送立ち上げオーバヘッドが通常転送モードに比べてかなり小さい．このことはメッセージ長が短い場合の転送に非常に有利となる．

また，複数のメッセージを送信する場合の転送立ち上げオーバーヘッドを減らすためにTCWチェインという機能が用意されている．CP-PACSではメッセージ毎にTCW（Transfer Control Words）と呼ばれる転送制御データ構造を作成する．このTCWには，送信するデータのアドレスやサイズ，送信先などといったメッセージ転送に必要な情報が書かれている．そして送信時には，NIA（Network Interface Adaptor）にTCWのアドレスが渡され，転送の単位であるパケットに分割して転送される．CP-PACSでは，連続して複数のメッセージを送信する場合には，各メッセージのTCWを連結して一つにまとめることができる．これをTCWチェインと呼ぶ．TCWチェインによりまとめられたメッセージ送信では，遅延の大きいソフトウェアによる送信起動オーバーヘッドは最初のパケット転送時の一回のみで，後のパケットの転送はNIAによるハードウェア・オーバーヘッドだけで済む．したがってこの機能は，多数の短メッセージを送信する時に非常に有効となる．

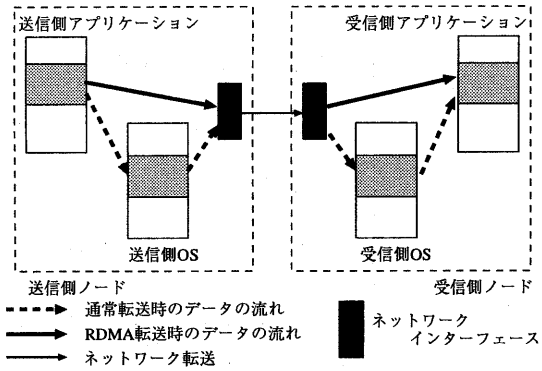

図2　リモートDMA転送による通信

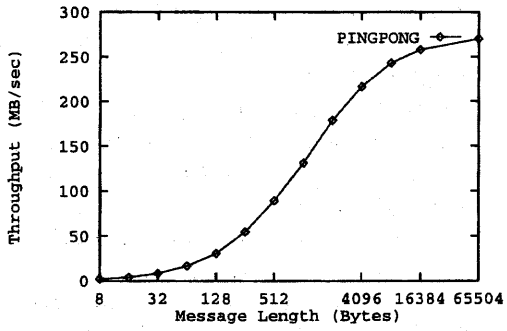

図3　ピンポン転送における実効スループット

以上のように通信を高速に行なえるという長所がある反面，OSの関与を極力減らしたことにより並列プログラミングが複雑になるといった欠点もある．そこで，この欠点を補うために通信ライブラリであるPVMを実装するという研究も行なわれている[4]．

3. 性能評価

本研究では，様々な転送パターンに対して個々にプログラムを作成し，実測によるネットワークの性能評価を行なう．特に基本的な一対一転送性能を求めるための転送（ピンポン転送）や，実アプリケーションで多様される転送パターンを中心に測定を行なう．

本節では，それぞれの転送パターンの測定結果を示し，得られた結果の考察を行なう．

3.1 測定方法

使用するノード数は256（4 × 8 × 8）台である．時間測定には経過時間をクロックサイクル単位で測定できる関数を用いる．また，ピンポン転送以外の測定では，時間測定を開始する前に全ノードでソフトウェアによるバリア同期をとっている．

送信関数には，プリミティブな関数の中でも最も立ち上げオーバーヘッドの小さい関数を用いる．

3.2 一対一転送

まず始めに，基本的な一対一転送であるピンポン転送についての評価を行なう．ピンポン転送とは，2つのノード間で交互にメッセージの送受信を行なう転送のことである．測定結果を図3に示す．

スループットは最大で約270MB/secとピーク性能の9割を達成できている．これは，システム領域でのデータのバッファリングを排除したことが大きな要因と考えられる．また，通信オーバーヘッドは約3.5 μ sec，$N_{\frac{1}{2}}$（ネットワークのピーク性能の半分を達成しているメッセージ長）は約1KBと極めて小さい．このように低オーバーヘッドで済んでいるのは，RDMA転送が軽いシステムコールで起動できることに因る．

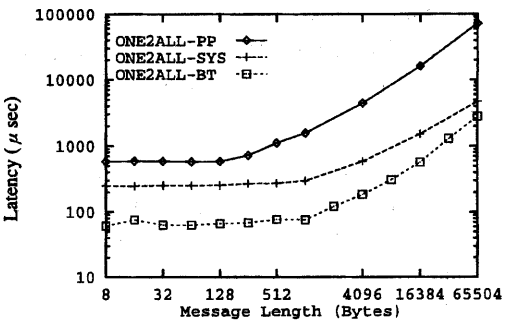

図4　1対全broadcastにおけるレーテンシ

以上より，CP-PACSに実装されている高速通信機構は効率良くハードウェアの性能を引き出していることが分かる．

3.3 1対全broadcast

1対全broadcastについて評価する．1対全broadcastとは，あるノードが持っているデータを全てのノードに持たせる転送のことである．

転送アルゴリズムとしては，システム関数を用いる方法（ONE2ALL-SYS），point-to-point転送によって送信ノードがその他全ての受信ノードにメッセージを送信する方法（ONE2ALL-PP），2進木状にメッセージを送信する方法（ONE2ALL-BT）の3種類について測定する．なお，ONE2ALL-BTでは，各ノードが送信するメッセージ全てをTCWチェインにより一回の送信に統合している．

測定結果を図4に示す．ONE2ALL-BTがどのメッセージ長においても最短時間で転送を終了している．これは，転送回数が$log_2 P$（Pはノード数）で済むこ

とが最大の要因である．さらに，TCW チェインによる送信起動オーバーヘッド低減もかなり効いている．このTCW チェインの効果については，後述の行列の転置のところで考察する．

ONE2ALL-SYS は，同一次元方向に並んだノードに対してTCW チェインを用いて連続的に転送を行なうため，合計3 回の転送で済む．しかし，OS による割り込み処理が介在するため ONE2ALL-BT よりも遅くなっている．但し，ONE2ALL-BT では転送のために全てのノードが通信に参加しなくてはならないのに対し，ONE2ALL-SYS はユーザの関与を最小限に抑えているので，プロセッサによる内部処理と並行して進めることができる．

また，ONE2ALL-BT は 2 進木に沿って転送を行なうため，ノード数が2 の冪乗でなくてはならないという制約がある．したがって，ライブラリなどの形で1 対全 broadcast を関数内に閉じ込めてしまう場合には，システム関数つまり ONE2ALL-SYS を用いた方が良い．

以上より，1 対全 broadcast を行なう場合は，対象となるアプリケーションの性質を考慮した上で ONE2ALL-SYS，ONE2ALL-BT を使い分ける必要があると言える．

3.4 Scatter & Gather

まず，ノード 0 が持っているデータを P 等分（P はノード数）して，それを他のノードへ分配する（Scatter）．各ノードは受けとったデータに対して何らかの処理を施し，その後ノード 0 にデータを送り返す（Gather）．このような処理のことを Scatter & Gather と呼ぶことにする．

測定では，プロセッサ内部処理時間を変えたときに全実行時間がどのようになるかを見てみる．さらに，各ノードに分配するデータ長を 16, 32, 48, 64KB と変えた場合について測定もする．なお，ノード 0 からその他のノードへの送信は，全てTCW チェインにより統合してある．測定結果を図5に示す．

スループットを算出すると約430MB/sec となった．Scatter の際にデータを分配するノード 0 が，まだ全てのノードへのデータ分配を終えないうちに，その他のノードがプロセッサ内部での処理を終えてデータを送り返してくることがある．この場合，NIA は送受信双方向の同時処理に対応しており，またメモリはそれに耐えうるだけのバンド幅を備えているので，最大で2 倍のスループットを実現できる．したがって，このような高いスループットとなった．

3.5 全対全 broadcast

バタフライ・コレクション・アルゴリズム（図 6）

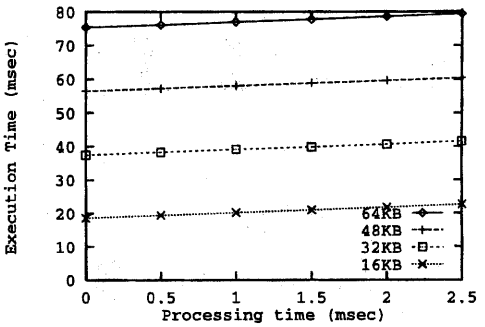

図5　Scatter & Gather におけるレーテンシ

を用いて全対全 broadcast の評価を行なう．全対全 broadcast とは，各ノードが持っていたデータを全員が全員分持つようにする転送のことである．また，バタフライ・コレクション・アルゴリズムとは各ノードが隣のノードとのデータ交換からスタートして，送信距離，転送量ともに倍々と増やしていくアルゴリズムである．

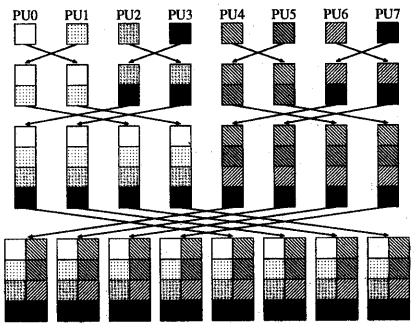

図6　バタフライ・アルゴリズムによる全対全 broadcast

測定結果を図7に示す．全ノードがその他のノードに対して point-to-point 転送でメッセージを送信した場合と比べて，バタフライ・コレクション・アルゴリズムは，総メッセージ転送量は変らないものの，各ノードがメッセージを送信する回数が（$P-1$）回から $log_2 P$回（Pはノード数，ここでは $P=8$）に減る，効率の良いアルゴリズムである．また，HXB では，バタフライ・コレクション・アルゴリズムによる全対全 broadcast を無衝突で行なうことができる．

図7では，メッセージ長が短いときは全対全 broadcast がピンポン転送の性能を越えているように見える．これは，送信一回当たりのメッセージ長がステップをふむ毎に倍々と大きくなるからである．

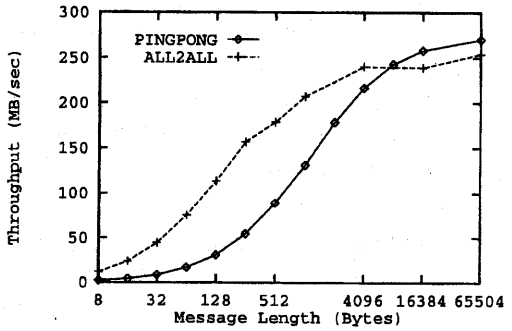

図7　全対全broadcastにおける実効スループット

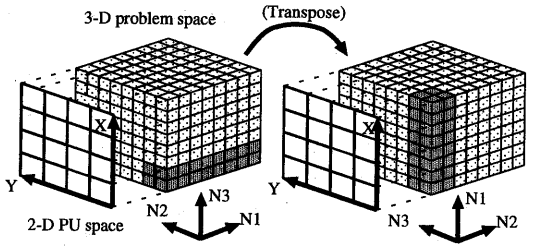

図8　2次元PU空間にマッピングされた3次元配列の転置

3.6　行列の転置

行列の転置を行なう転送について測定を行なう．行列の転置は多次元の高速フーリエ変換などに使用される．具体例を図8に示す．この場合，4×4の2次元PU空間に3次元配列をマッピングしているので，各PUは一つの次元方向（図中ではN1）に対して連続したデータを持つ．しかし，転置を行なうことによって今度はN2方向に連続したデータを持つことになる．したがって，転送前に持っていたデータは，分割して他のPUに送られる．この際，転送するデータは転送先によって異なり（個別転送），また分割するPU数が多いほどメッセージ長は短くなる．

本測定では，3次元PUに3次元配列をマッピングしたときの転置について測定する．したがって，上例よりもさらに複雑な転送を要する．この際必要となるメッセージ転送について，メッセージ数（#MSG），メッセージ当たりのデータ量（data/MSG）を表1に示す．これより，各PUに512KBのデータを持たせた場合でもメッセージ長は128Bと$N_{\frac{1}{2}}$を大幅に下回っているのが分かる．また，転送回数も多い．

このような転送に対し，各転送を順次転送する方法（basic）と，TCWチェインにより各ノードが送信するメッセージを全てまとめて一つのメッセージとする

表1　行列の転置に伴う転送

data/PU (Bytes)	data/MSG (Bytes)	#MSG (個)
128	8	16
256	8	32
512	16	32
1024	16	64
2048	16	128
4096	32	128
8192	32	256
16384	32	512
32768	64	512
65536	64	1024
131072	64	2048
262144	128	2048
524288	128	4096

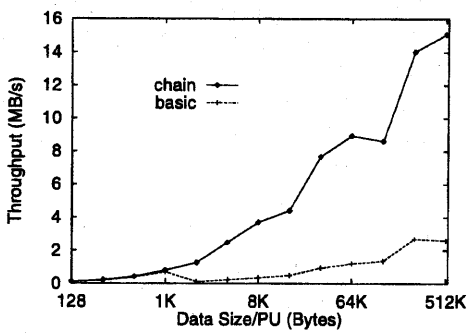

図9　転置転送における実効スループット

方法（chain）の2種類について測定を行なう．

測定結果を図9に示す．basicにおいてN回のメッセージ転送を必要とした場合，chainでは，TCWチェインにより1回の送信で済むため$(N-1)$回分のソフトウェアによる送信起動オーバーヘッドを省くことができる．測定結果を見ても，chainはbasicに比べて非常に高い結果が得られている．

今回のように多数の短メッセージ転送を必要とする場合でも，CP-PACSではTCWチェインを用いることによりある程度性能低下を防ぐことができる．

3.7　ランダム転送

これまでは，転送相手が規則正しい転送についての評価を行なった．しかし，実アプリケーションによっては不規則な転送を行なうものもある．そこで，そのような不規則な転送の代表として，ランダム転送の評価を行なう．ランダム転送とは，送信先をランダムに決定する転送である．また，今回の測定ではメッセージ生成率を0.1～1.0に変えて測定する．

メッセージ生成率を上げるにつれて，ネットワーク

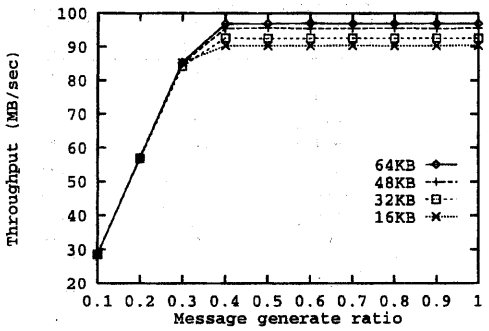

図10　ランダム転送における実効スループット

中のメッセージ数も多くなり，メッセージ同士の衝突が増える．ある一定以上のメッセージ生成率になるとスループットは飽和状態になるが，それはピーク性能の約1/3のところであった（図10）．

これまで我々が行なってきたシミュレーションによるランダム転送の実験でも，3次元HXBの場合はピーク性能の約1/3と非常に近い値が出ている．またこのシミュレーションでは，HXBが他のメッシュ／トーラスといった他のネットワーク・トポロジと比べて非常に高いという結果を示している．

4. 実アプリケーションへの応用

以上の測定結果より，HXBは様々な転送パターンを効率的にこなせることが分かった．そこで次に，実アプリケーションに対してどれだけHXBが有効かを考察する．

実アプリケーションでは，必ずしも一種類の転送パターンしか用いないわけではない．例えばNAS並列ベンチマーク[5]に出てくるCG（Conjugate Gradient）法を例にとると，shuffle転送，butterfly collection，butterfly summationなどといった複数の転送が必要となる．

メッシュやトーラスといったその他のネットワーク・トポロジでも，それぞれの転送パターンに適した各ノードへのデータ割り付けを行なうことにより無衝突もしくはそれなりに効率良くデータ転送を行なうことができるかもしれない．しかしCG法のようにプログラム中に複数の転送パターンがある場合，どれか一つの転送パターンを効率良く行なえるデータ割り付けであったとしても，その他の転送パターンではうまくいかないことがある．

それに対し，HXBでは不規則な転送を含めた各種転送パターンに強いので，どのようなデータ割り付けであってもそれぞれの転送を効率良くこなせる．したがって，対象とするプログラムに最も適したノードへのデータ割り付けを行なった後に，適当な転送パターンを用いてノード間のデータの交換を行なうといったこともできる．

よって，各種転送パターンだけではなく，実アプリケーションにおいてこそ更にHXBの有効性が示されるものと推測される．

5. ま と め

本研究では，CP-PACSのネットワークについて様々な転送パターンに対する性能評価を行なった．その結果，HXBというネットワーク・トポロジは様々な転送パターンに対して柔軟に対応でき，また，CP-PACSに実装されている高速通信機構はハードウェアの性能を効率良く引き出せていることが分かった．実アプリケーションにおいては，更にその有効性が確かめられるものと考えられる．

我々は，CP-PACSの単体プロセッサの厳密な性能評価も行なっている[3]．今後は，このプロセッサの性能評価及び今回行なったネットワークの性能評価の結果をもとに，行列計算，分子動力学法などの実アプリケーションのチューニングを行ないたい．

謝辞　CP-PACSを利用する機会を与えて下さった筑波大学計算物理学研究センターの関係者各位に感謝します．また，本研究に関して貴重な御意見を頂いたアーキテクチャ研究室の諸氏にも感謝します．なお，本研究の一部は創成的基礎研究費（08NP0401）の補助によるものである．

参 考 文 献

1) 岩崎 洋一, 中澤 喜三郎ほか：計算物理学と超並列計算機 — CP-PACS計画—, 情報処理, vol.37, No.1, pp.10-42(1996).
2) 朴 泰祐ほか：ハイパクロスバ・ネットワークの性能評価, 電子情報通信学会技術研究報告,pp41-48 (1993).CPSY93-40.
3) 板倉 憲一ほか：超並列計算機CP-PACSの基本性能評価, 情報処理学会研究報告, ARC123-4, pp.19-24(1997).
4) 松原 正純ほか:超並列計算機CP-PACSにおけるPVMの実装, 情報処理学会研究報告,ARC96-119, pp13-18(1996).
5) Bailey D. et al.：THE NAS PARALLEL BENCHMARKS,RNR Technical Report RNR-94-007 (1994)
6) 田中 良夫ほか：並列アルゴリズムにおけるCollective通信の性能比較, 情報処理学会研究報告, HPC62-4, pp.19-26(1996).