ハイパフォーマンス
コンピューティング 55−12
（1995．3．9）

大規模ワークステーション・クラスタにおける PVM の性能評価

弘中哲夫

広島市立大学 情報科学部 情報工学科

〒731-31 広島市 安佐南区 沼田町 大塚 151-5

E-mail: hironaka@ce.hiroshima-cu.ac.jp

PVM 始めとするメッセージパッシング・ライブラリの登場により安価に大規模な数値計算を行うことが可能になりつつある．本稿では，2次元熱伝導方程式の並列解法を使って大規模ワークステーション・クラスタ上における PVM の性能評価を行った．性能評価は数式による理論的な評価と，実際に130台の SPARCstation LX から構成されるワークステーション・クラスタでベンチマーキングすることで行った．本稿では上記の評価結果から，ワークステーション・クラスタ内のネットワーク・バンド幅の重要性，問題サイズと演算性能の関係などについて検討する．

Performance Evaluation of PVM on Large Scale Workstation Clusters

Tetsuo Hironaka

Department of Computer Engineering
Faculty of Information Sciences
Hiroshima City University

E-mail: hironaka@ce.hiroshima-cu.ac.jp

Parallel programming library, such as PVM, for workstation clusters made it possible to do large scale scientific calculations, with an good cost perfomance. In this paper, we evaluated the performance of PVM working on large scale workstation cluster. The evaluation was done by two method, the theoretical method, and by running an realistic application on the workstation cluster containing 130 workstations. From the evaluation results, the importance of network performance, and the importance to scale the problem size larger, will be shown.

1 はじめに

近年高性能ワークステーションの普及により，大規模な計算をコストパフォーマンス良く安価に実行できるようになってきた．さらに，これらのワークステーションが複数台高速なネットワークで結合されるようになったことにより，個々の独立のワークステーションとしての使用だけでなく，PVM3(Parallel Virtual Machine Ver.3)[1] などのメッセージ・パッシング・ライブラリを用いて並列分散処理も行われるようになってきた．

このように高速なネットワークにつながる複数台のワークステーション（ワークステーション・クラスタ）上で計算を行うことで，ワークステーション1台のみで計算するのと比べ，格段に高い演算性能を達成することができるようになった．そして問題によっては経済性および演算性能でスーパコンピュータに匹敵しうる場合があることがわかってきた [4]．また，クラスタ内に含まれるワークステーション数を大幅に増やすことでより高い演算性能を達成するスケーラビリティーも同時に期待され，その評価も現在いろいろ試みられている [3]．

そして，スケーラビリティーとネットワーク・バンドが複雑関係していることがわかってきた．つまり，実際にワークステーション・クラスタ内のネットワーク・バンド幅の問題により，ワークステーション・クラスタの演算性能を十分に引き出せない場合があることが予想される．

本稿では差分法を用いた2次元熱伝導方程式の並列解法をワークステーション・クラスタ上で解く場合について数式を用いた理論的な評価，および，実際に130台のワークステーションから構成されるワークステーション・クラスタを用いて行った評価の結果を示す．

2 PVM の動作環境

PVM の評価は図1に示すようなネットワーク構成を持つワークステーション・クラスタを用いて行った．

表 1: ワークステーションの仕様

SPARCstation LX	
CPU	microSPARC
クロック周波数	50 MIlz
キャッシュメモリ	6 KB
主記憶容量	32 MB
SPECint92	26.4
SPECfp92	21.0
Mips 値	59.1
MFlops 値	4.6
Ethernet	10BaseT (10Mbps)

表 2: ルータの仕様

Cisco 7000	
CPU	Motorola 68040
インタフェース間データパス	533 Mbps
転送レート	100,000 pps 以上
ルーティング・プロトコル	RIP

図1のワークステーション・クラスタは表1に示すような仕様を持つワークステーション130台から構成される．クラスタ内の各ワークステーションは図1が示すように8～9台ごとに16本のイーサネット・セグメントに分割されて接続されている．16本のイーサネット・セグメントはそれぞれ表2の仕様を持つルータによって相互に接続されており，イーサネット・セグメント間の通信が保証されている．上記の構成をとることによりワークステーション・クラスタは最大7.6Gips，598MFlops のピーク性能を持つ．

3 熱伝導方程式の並列解法を用いた PVM の評価

差分法を用いた2次元熱伝導方程式の並列解法を2節の大規模ワークステーション・クラスタにおいてPVM3.3 で実現し，その演算性能を評価した．なお，実験に使用したプログラムは文献 [5] 中の例題プログラムである「差分法による2次元の熱伝導方程式解法プログラム」を若干改造して用いた．

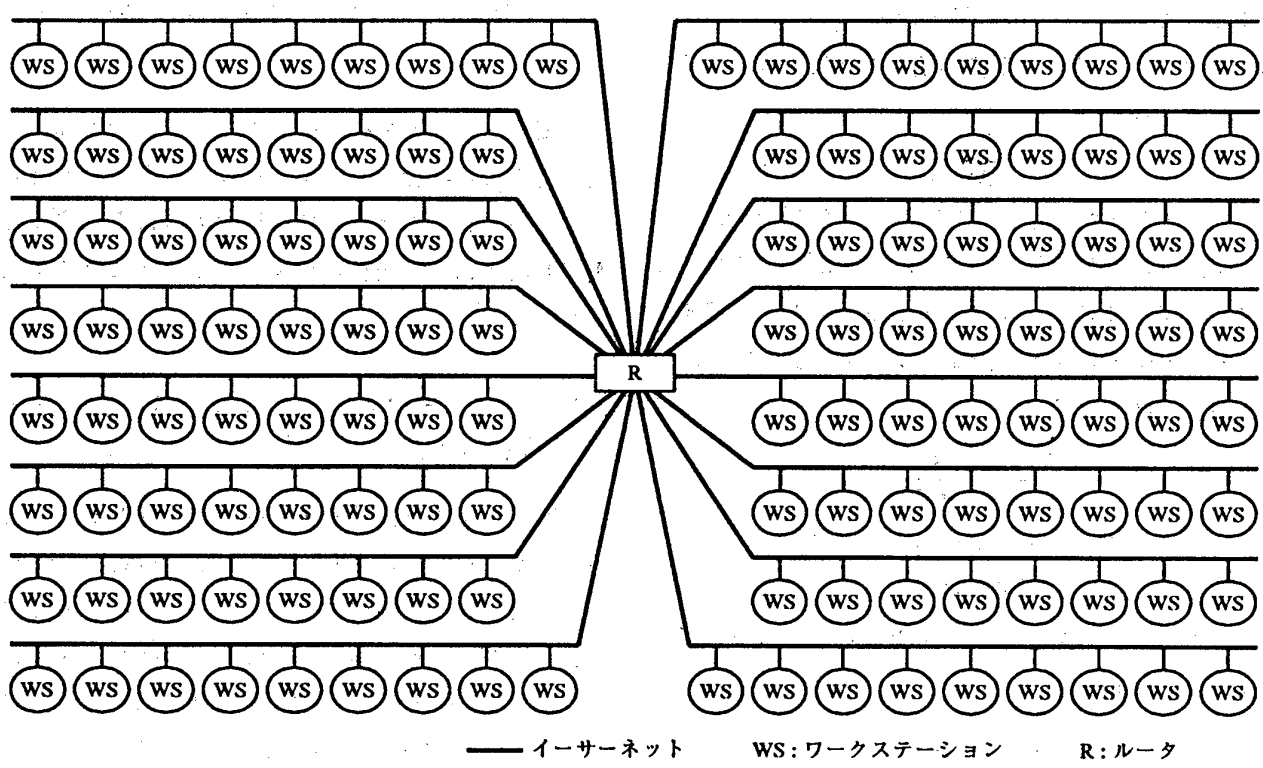

図 1: ワークステーション・クラスタのネットワーク・トポロジー

このプログラムは 2 次元の領域の一部に熱を加え，一定時間後の最高温度を求めるものである．具体的には $N \times N$ の正方サブ領域に分割した 2 次元領域を $M \times M$ 個のタスクに図 2が示すように割り当てて計算を行う．各タスクはタスク間で割り当てられた領域の境界温度をメッセージで交わしながら一定時間後の温度分布を求める．したがって，1 つのタスクが送信するメッセージは上下左右の 4 方向に境界温度を送信するので 1 タスク当たり $4N/M$ 個のメッセージを隣接する領域を持つタスクに送信することになる．

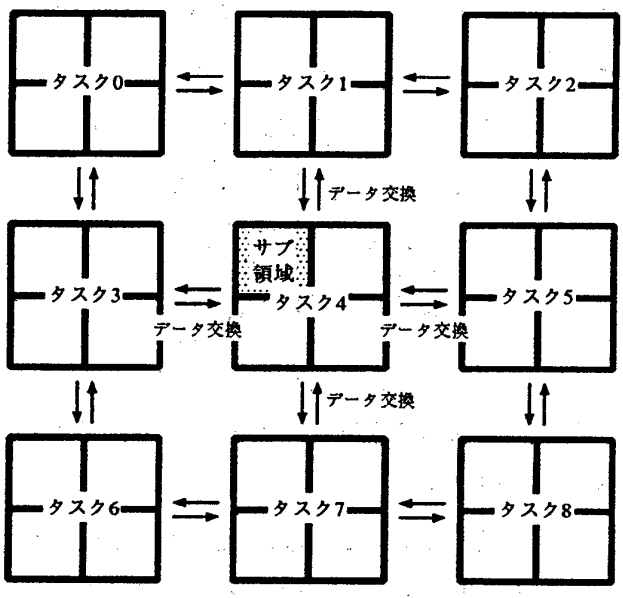

図 2: 各タスクへのサブ領域の割り付け

3.1 大規模ワークステーションクラスタに対する差分法の適合性

実際に実験結果を示す前に大規模ワークステーションクラスタに対する差分法の適合性を数式を用いて簡単に評価する．まず，$N \times N$ の正方サブ領域に分割した 2 次元領域を $M \times M$ 個のタスクに均等に割り当てる．このような場合，差分方程式を 1 ステップ解くのに必要な演算の数を C と置くと，1 タスク当たりの演算数は

$$\frac{CN^2}{M^2}$$

となる．また，1ステップ解いた後，1タスク当たり周囲のタスクに送信する境界領域の温度を知らせるメッセージ数は，

$$\frac{4N}{M}$$

である．タスクが実行する1ステップでは演算とメッセージ送信がそれぞれ1回づつ行われる[1]．ここで，タスクを実行するワークステーションの演算性能をF，ネットワークのバンド幅をBとすると，1ステップの実行時間は，

$$\frac{CN^2}{FM^2}+\frac{4N}{BM}$$

となる．ここで各ステップにおいて，すべてのタスクが同時に始まり，同時に終ると仮定すればワークステーション・クラスタ全体で得られる単位時間当たりの演算性能は

$$\frac{CN^2}{\frac{CN^2}{FM^2}+\frac{4N}{BM}}$$

となる．この式を整理すると，

$$\frac{M^2F}{1+\frac{4FM}{CBN}}$$

上記の式に対して変数N,F,B,Cが演算性能に与える影響をまとめたものが次の表3である．

表3: 各パラメータが演算性能に与える影響

パラメータ	演算性能
$N\to\infty$	M^2F
$F\to\infty$	$\frac{CBNM}{4}$
$B\to\infty$	M^2F
$C\to\infty$	M^2F

表3より次のような事柄がわかる．並列計算で性能を引き出すには，

- Nを大きくする：2次元平面を分割する正方サブ領域数を大きくして，ワークステーション1台当りで行う計算を大きくする，
- Bを大きくする：なるべく高いバンド幅のネットワークを使用する，
- Cを大きくする：なるべく演算数の多い差分方程式を用いる，

が重要であることがわかる．このうちCの増加は単純に適用した場合，それにともない一定時間の計算をするのに必要ステップ数が減少するような演算を選択しなければ，無駄に計算を行うことになり無意味となる．

3.2 大規模ワークステーションクラスタにおける差分法の実際

実際に熱伝導方程式の並列解法のプログラムを用いて実際にPVM上で評価を行った．2次元平面のサブ領域数$N\times N$は$M=11, N=330,660,1320,2640$とおいて問題を解く場合に必要な実行時間を評価した．この評価結果が図3である．なお，この実験では1台のワークステーションには1つタスクのみを割り当て，合計121台のワークステーションを用いた．評価結果の図3から次のような事柄がわかる．$N=2640$で計算した場合にくらべ$N=330,660,1320$で計算した場合は，計算ステップ数の増加に対する実行時間の伸びが非常に小さい．これは通信時間が実行時間の大部分を占めているためである．このことは各条件における演算性能を表した図4を見ても非常に明確である．

ここで得られた演算性能がどの程度のものか比較するため，同じプログラムを64セル構成のAP1000上実行した結果を評価している文献[2]の評価結果と比較する．文献[2]の評価結果ではPVMを移植した64セル構成のAP1000上で$N=320$，タスク数64の演算性能は約44MFlopsである．それに対して本ワークステーション・クラスタで$N=330$，タスク数121の演算性能は約26MFlopsである．つまり，ワークステーション・クラスタでは約2倍のプロセッサ数を用いてAP1000の半分の性能にしか達していない．これは両者にあるネットワーク・バンド幅の大きな違いからくると考えられる．

[1] 簡単のため近隣のタスクからのメッセージの受信は送信と同時に行われると仮定する．

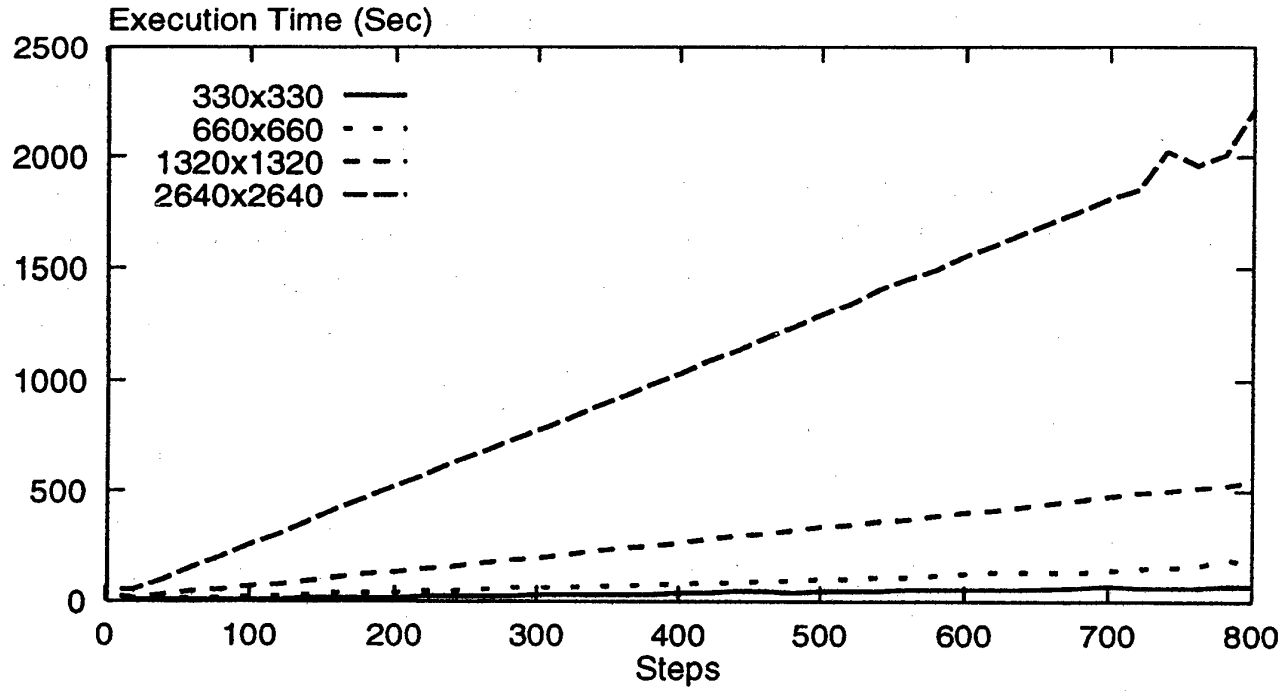

図 3: 差分法を用いた 2 次元熱伝導方程式の並列解法の実行時間

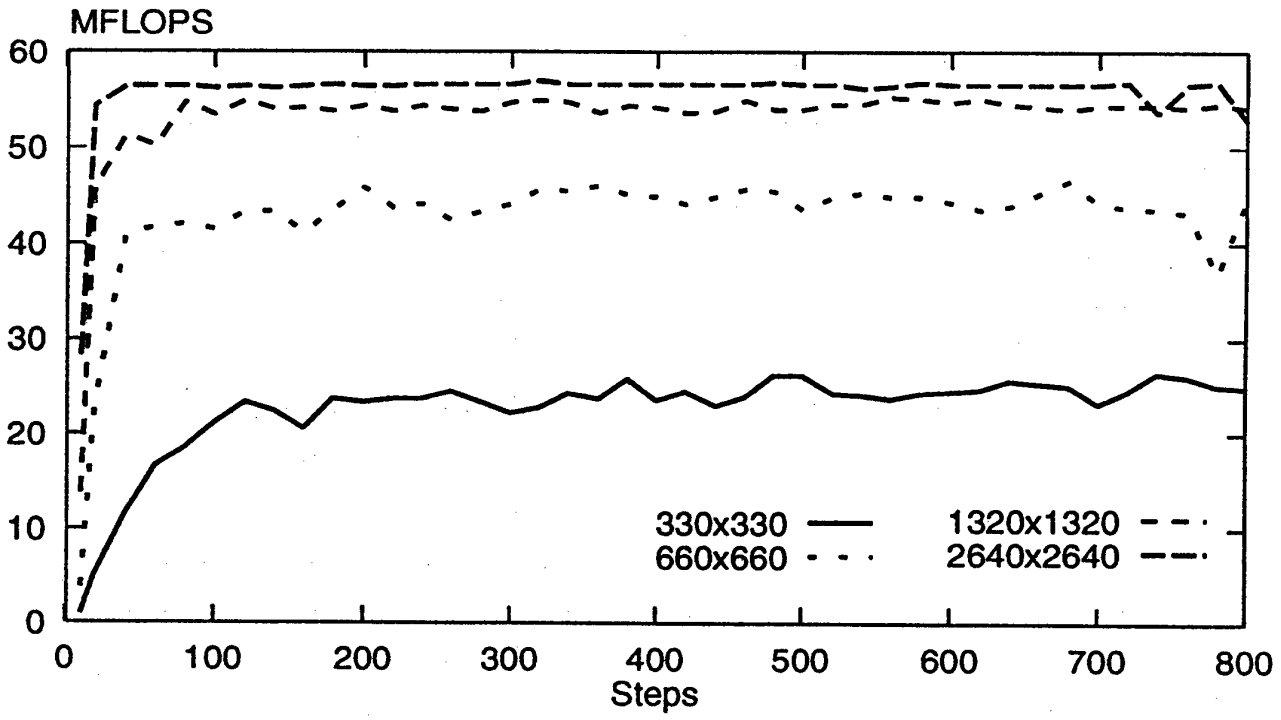

図 4: 差分法を用いた 2 次元熱伝導方程式の並列解法で得られる演算性能

4 大規模ワークステーション・クラスタにおける性能低下要因

3節の結果から，イーサネットを中心としたネットワーク構成で大規模ワークステーション・クラスタを構築した場合，ネットワークのバンド幅 B が大きな問題になると思われる．そこで今回実験を行った環境でネットワークのバンド幅を低下させる原因とな

りそうな事柄をいくつか挙げてみた．

1. アクセス競合
   イーサネットはバス型のネットワークであるため，ネットワークにつながる複数台のワークステーションの内，1時には1台しか通信できない．そのため，1本のイーサネットに多くのワークステーションを接続し，同時に通信すると通信時間に加えワークステーション間のアクセス競合をアービトレーションするのに多くの時間を費やすこととなる．また，1つのバスを複数台のワークステーションで時分割して使用するため，1台当たりの平均通信スループットは単独でネットワークを使用する場合に比べて小さくなる．

2. ルータのオーバヘッド
   1で述べたような問題の影響を小さくするため，通常をイーサネットを小さくセグメント化し，これらのセグメントをルータにより1つのネットワークに結合する．ルータは自分に接続しているすべてのイーサネット上のパケットを監視し，当該セグメント外への通信であれば，ディスティネーション・セグメントへパケットを中継する．この中継はルータに接続するすべてイーサネット・セグメントのパケット解析し，パケット内のディスティネーション・アドレスを解読して行う．もし，ルータの能力が不十分であれば，イーサネットのセグメント間通信スループットが低下する．

3. アプリケーションの通信パターンと実際のネットワーク・トポロジーとの違い
   ワークステーション・クラスタ内のネットワーク・トポロジーと実行するアプリケーションの通信パターンが大幅に違う場合，ネットワークにホットスポットが生じ，オーバヘッドとなる．

次の節では，これら問題がPVMの動作にどの程度の影響を与えているのかを調べて見る．

5 通信性能の評価

2節で示したような動作環境を用いてPVMのスレーブ・タスクを130台のワークステーションすべてに1づつ生成し，PVMで作成した測定用プログラムを用いてワークステーション間通信速度を測定した．使用した測定プログラムは任意の2つのスレーブタスク間でメッセージを10往復させ，その経過時間を計ることで通信性能を計測する．

5.1 イーサネット・セグメント内の通信

4節で述べたように1本のイーサネットに複数台のワークステーションを接続した場合，複数台が同時に通信をすることでアクセス競合によるオーバヘッドが生じる可能性がある．図5は8台のワークステーションを使って同時に最大4組の独立した1対1通信を行った時の通信時間を測定した結果である．図5の結果より次のようなことがわかる．

1. 1000Byte以下のメッセージを送信する場合，どのようなサイズのメッセージを送受信しても通信時間はほとんど変わらない．

2. 1000Byte以上のメッセージを送信する場合，サイズに比例してメッセージを送受信するのに必要な通信時間が増加する．

3. 1000Byte以下では通信競合の影響はほとんど見られないが，1000Byte以上では通信競合の影響が明確に現れる．

このように1000byteを堺にさまざまな影響が現れるのは，1000byte以下ではメッセージを実際に送信している時間に比べ，メッセージ長と無関係に必要なメッセージ送信のためのセットアップ時間が長いためである．セットアップ時間の大部分はOSによる処理が行われる時間と思われる．

以上の結果からPVM3.3ではメッセージ1つ当たりの送信単位を1000Byte以上にする方がメッセージを送信する上で効率がよいことがわかる．

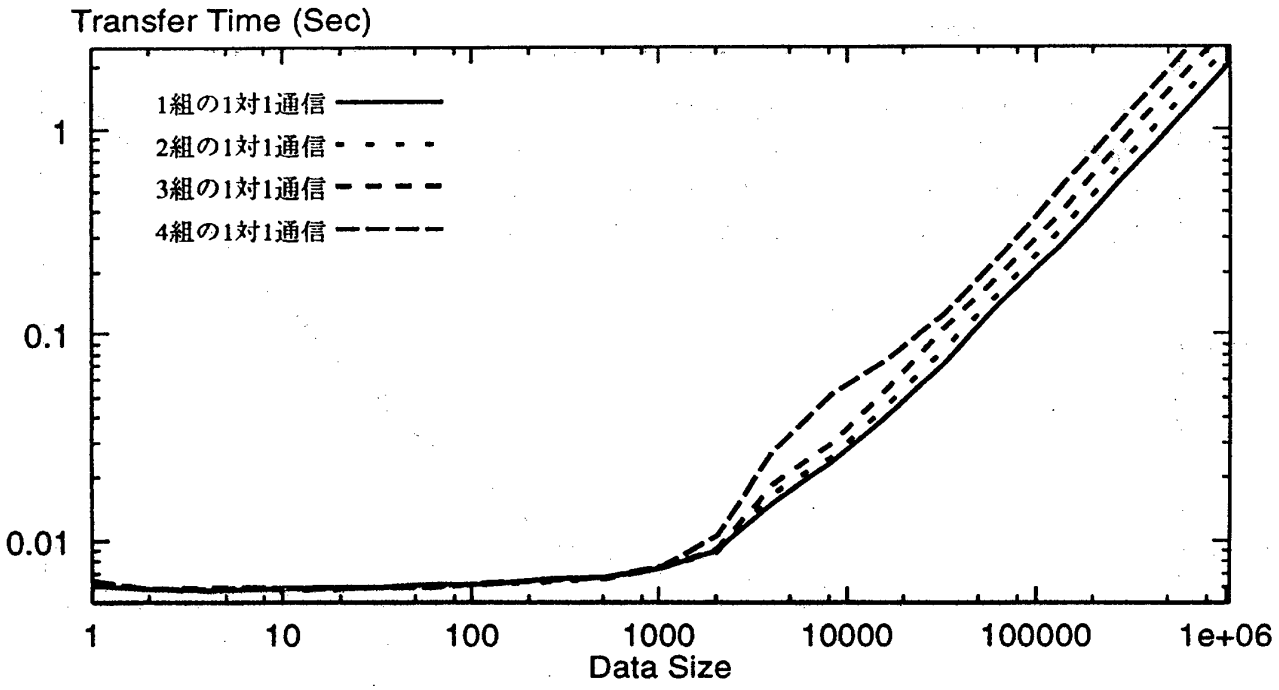

図 5: イーサネット・セグメント内の通信能力

5.2 イーサネット・セグメント間の通信

次にイーサネット・セグメント間の通信性能を評価してみた．本測定環境では 2 つにイーサネット・セグメント間を跨ぐ通信はすべて表 2のルータを介して行われる．図 6にイーサネット・セグメント間の通信を最大 8 組実現し，イーサネット・セグメント間の通信が複数生じることにより性能に及ぼす影響を評価した．

図 6より同時に発生するイーサネット・セグメント間の通信が増加しても，通信性能がほとんど変わらないことを除き，イーサネット・セグメント内の通信時に見られる特性とほぼ同じ特性がみられた．これにより，本測定環境ではイーサネット・セグメント内の通信とイーサネット・セグメント間の通信には特性状大きな差が見られないことわかる．

5.3 ネットワーク・トポロジーの問題点

5.1節，5.2節の評価より，本実験で使用したワークステーション・クラスタは仮想的にクロスバーで相互接続された 16 本バスといった形態をとったネットワークとみることができる．それに対して今回のワークステーション・クラスタで実行したアプリケーションは 2 次元メッシュのタスク間通信パターンを持っている．そのため，PVM が自動的に行ったタスク割当が適切でない可能性がある．つまり，PVM がタスクを自動配置したやり方次第では特定のイーサネットセグメントに極端にひどいホットスポットを作ってしまう可能性がある．なお，今回の評価ではタスクとイーサネット・セグメント間の割り当てに関しては PVM にまかせ，なにも制御を行っていない．

6 まとめ

本稿では 130 台の SPARCstation LX から構成されるワークステーション・クラスタで PVM を使用した場合についての評価を行った．評価に用いたプログラムは差分法を用いた 2 次元熱伝導方程式の並列解法である．評価は本プログラムを実行した場合，2 次元平面の分割数，ネットワークのバンド幅，各ワークステーションの演算性能などの各パラメータがどのようにワークステーション・クラスタ全体の演算性

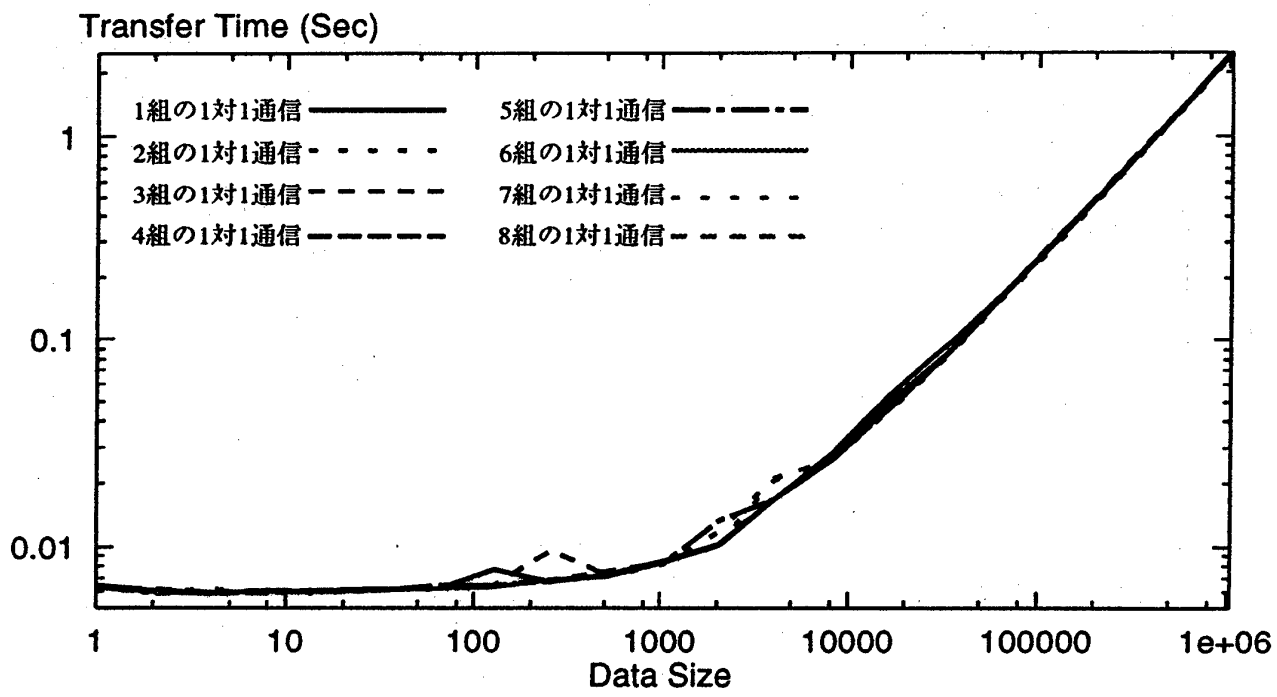

図 6: イーサネット・セグメント間の通信能力

能に反映されるかを評価した．その結果，個々のワークステーションの演算性能よりネットワークのバンド幅が重要であること，2次元平面の分割数を増やし，問題の規模を大きくした方がより最大演算性能に近い性能を達成できることがわかった．また，実際にワークステーション・クラスタ上でプログラムを実行した結果最大約 56MFlops の演算性能を達成することができた．

謝辞

日頃よりご討論戴き，多くの有用な御意見を戴く広島市立大学情報科学部情報工学科の藤野清次 教授，竹内敏己 助手，ならびに，PVM プログラミングに関してさまざまな助言を戴いた九州大学大学院総合理工学研究科の岩下茂信氏に深く感謝します．

参考文献

[1] Al Geist, Adam Beguelin, Jack Dongarra, Weicheng Jiang, Robert Manchek, Vaidy Sunderam, *PVM: Parallel Virtual Machine A User's Guide and Tutorial for Networked Parallel Computing*, MIT Press, 1994.

[2] 岩下茂信, 村上和彰, "KU PVM3/AP1000 の性能評価," 情処研報, HPC-52-16, 1994 年 7 月.

[3] 佐藤三久, 関口智嗣, 長嶋雲兵, "スケーラビリティによる並列計算機システムの比較," 情処研報, HPC-52-22, 1994 年 7 月.

[4] 長谷部晴美, 福井義成, "PVM のアプリケーションへの適用効果 – モンテカルロ法プログラム MCNP の場合 –," 情処研報, HPC-52-17, 1994 年 7 月.

[5] 富士通研究所, AP1000 プログラム開発手引書 (I)C 言語インタフェース, 第 2 版, 1992 年 2 月.