ハイパフォーマンス
コンピューティング 57－2
（1995．8．24）

大規模ワークステーションクラスタ上での性能評価プログラム設計に関するネットワーク評価

日向寺 祥子，緒方 道郎，星野 辰夫，長嶋 雲兵,† 高橋 隆男，井上 靖

東海大学 電子計算センター
†お茶の水女子大学 理学部 情報科学科

{sachiko,ogata,thosino,takao3,yinoue}@cc.u-tokai.ac.jp
umpei@is.ocha.ac.jp

概要

様々な特徴を有する並列計算機を、ひとつの評価基準でそれに見合った性能評価を行う技術が強く求められている。本研究では、性能評価プログラム設計に向け、イーサネットで結合されたワークステーションクラスタにおける通信のオーバーヘッドの動きを観察するため、PVM ベースで単純化したプログラムを用いて時間を測定した。その結果、イーサネットで結合された環境では、同時に約 4000 バイトのデータ通信がおこなわれるように設計をおこなうと通信の効率がよいことがわかった。

Network Evaluation for Large Scale WS Cluster to Design Benchmark Programs

Sachiko HYUGAJI, Michio OGATA, Tatsuo HOSHINO, Umpei NAGASHIMA,†
Takao TAKAHASHI, Yasushi INOUE

Tokai University, Computer Center
†Ochanomizu University

abstract

Many problems for evaluating parallel systems are found out with development of parallel processing technique. More particularly large scale WS cluster have a serious problem in network overhead, and should be evaluate by different method from massively parallel systems as network communication.

In this study, Network conflict in one segment is observed by simplified PVM programs, then message passing limit is shown.

In the case of Ethernet connection, less than 4000 bytes for total traffic give the best efficiency for total throuput.

1 はじめに

近年のめざましい科学技術の進歩とともに、従来にも増して高い計算能力が求められるようになった。特に、物理、化学の分野では計算を含めた研究がさかんに行われており、現在の計算能力をフルに活用してもはるかに及ばない程の計算能力が求められている。並列計算は理論上は計算速度がプロセッサの台数分だけの性能向上が望めるため、ネットワークやソフトウェアの発達により、大幅な計算能力向上の可能性が大きく、物理や化学の分野でもソフトウェアを改良して実際に並列処理を行う傾向が強い。

応用分野の実用化では計算機の性能を最大限活用するためのプログラム設計は最重要課題の一つでもあり、そのためにシステムの正確な性能評価が望まれる。並列システムの性能評価には、逐次計算機やベクトル計算機とは違ったアルゴリズムでのベンチマークプログラムが必要であり、従来のベンチマークプログラムである Linpack などは適当ではない。また、並列システムも様々なアーキテクチャやインターフェースを持つため、同一のプログラムによりその性能を評価することは容易ではない [1]。

一方、並列処理の手段の一つとして安価に構築が可能なワークステーションクラスタが考案され、クラスタ技術も大きく発展している。メッセージパッシングライブラリを用いてプログラミングが容易におこなえるようになったことも、ワークステーションクラスタ普及の大きな一因となっている。ところが、ワークステーションクラスタを用いた並列分散処理では、汎用並列計算機に比べ優れた専用のネットワークを持たないために、ネットワークの遅さによる通信のオーバーヘッドが大きなネックとなる。そのために、CPU 台数の小さいところで性能向上率が飽和してしまい、ワークステーションクラスタを用いて高い性能を得ることは通信部分が大変小さいプログラムを除いては大変困難である。特に大規模なクラスタでは、そのネットワークの構造が、1 つのセグメント内に多くの計算機を有する可能性が高くなり、同一セグメント内で同時に通信がおこなわれる確率が高くなる。そこで通信の競合が大きくなることは避けられない [2]。ワークステーションクラスタの性能を評価する場合には、そのような通信のオーバーヘッドを充分考慮にいれなければならない。

現段階でパラレル・ベンチマークとして用いられているプログラムの中には例えば NAS Paralell Benchmarks のようなアルゴリズムでのみ与えられているいるようなものが存在する [3]。アルゴリズムのみが与えられている場合には、それぞれのシステムの構成にあわせてベンチマークプログラムを設計することが可能であるため、大規模なワークステーションクラスタの性能を汎用並列計算機と比較評価するには適しているものと思われる。本研究では、正確な大規模ワークステーションクラスタの性能評価に先立ち、NAS Paralell Benchmarks などによるプログラム設計において大きなポイントとなる通信について、とくに通信の競合に着目して評価をおこなった。

2 ワークステーションクラスタの構成

東海大学電子計算センターの教育用端末である、NEC EWS 4800/320VX 128 台を用いてワークステーションクラスタを構成した。用いたメッセージパッシングライブラリは PVM3.3.7 である [4]。用いた 128 台のワークステーションは、約 30 台ずつサブネットを構成しており、128 台全てが NFS でファイル共有がなされている。ネットワークはイーサネット (10Mbps) である。

ワークステーションクラスタの構成を図 1に示す。

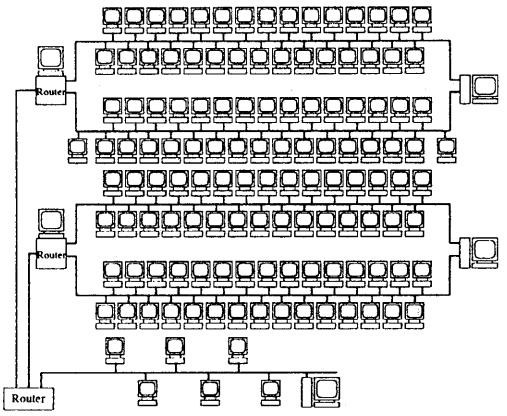

図 1: ワークステーションクラスタの構成

また、使用したワークステーションの単体性能を表 1にまとめる。

CPU	R4400SC (200MHz)
MIPS 値	198
SPECint92	108
SPECfp92	119
キャッシュメモリ	1 次:16K+16K バイト 2 次:1M バイト
メモリ	48M バイト

表 1: NEC EWS4800/320VX の単体性能

プログラムは NFS により全台で共有したが、PVM は NFS により多くのプロセッサが一度にアクセスすると致命的な速度の遅延となる恐れがあるため、各計算機上にローカルにおくこととした。

3 測定方法および結果

NAS Paralell Benchmarks などのプログラムを設計する上で実際に使用される頻度が高いと思われる通信方法を 3 種類取り上げ、多数の計算機からなる同セグメント内での通信の競合を見ることとした。3 種類の通信方法としては、

- 1 対 1 通信
- 1 対多通信 (1 対 5 通信)
- 多対多通信

を取り上げ、これらの通信を同セグメント内で数組同時におこない、その経過時間を測定した。それぞれの通信の測定方法および結果は以下の通りである。

3.1 1 対 1 通信

測定方法

通信のコストとしては、下図に示すように pvm_pack および pvm_unpack を含めたセンド-レシーブの往復の経過時間を測定した。データサイズは 128 バイト から 51200 バイト まで変化させ、同時に 1 組から 15 組の 1 対 1 センド-レシーブをおこなった。25 回から 50 回のセンド-レシーブの繰り返しを各 PU ごとに測定した経過時間より、1 回の往復の通信時間を算出し、その最大経過時間を全体の経過時間とした。

```
gettime by gettimeofday
for(i=0; i<50 (or 25); i++)
{
pvm_initsend
pvm_pkdouble
pvm_send      --> pvm_recv
                  pvm_upkdouble

                  pvm_initsend
                  pvm_pkdouble
pvm_recv      <-- pvm_send(tid[0])
pvm_upkdouble
}
gettime by gettimeofday
```

1 対 1 通信の場合の基本プログラム

測定結果

1 組から 15 組までの 1 対 1 通信の結果を図 2 に示す。

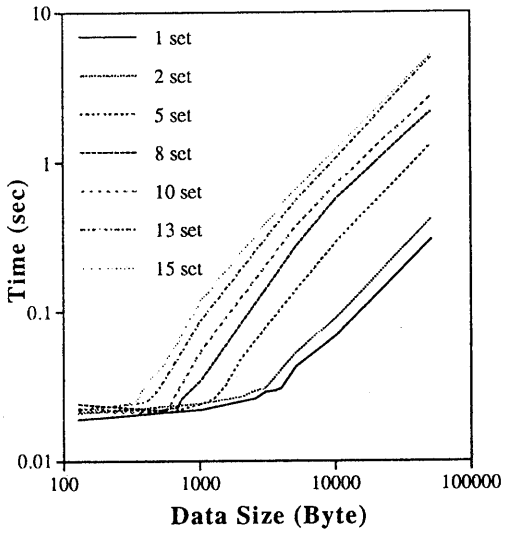

図 2: 1 対 1 通信の場合の競合

測定結果より、1 組の 1 対 1 通信では約 4000 バイトで通信が飽和し競合が見え始めるが、同時に数組の 1 対 1 通信をおこなうと、データサイズの小さいところで飽和し通信の競合が明確に見え始める。これまでの研究では、このようなスワップサイズの飽和は明確には報告されていないが、遅いネットワークを用いる際には、通信速度の飽和するデータサイズの見積りは実行速度をあげるうえで重要なポイントである。表 2 に通信速度が飽和するデータサイズを示した。

表 2 より、1 本のイーサネット上を pvm_send により通信がおこなわれる全データのサイズが約 4000 バイトを越えると、飽和して通信の競合がおこることがわかる。同時に通信をおこなう組の数が増えると全通信データサイズが増すように見えるのは、各 PU 間でおこなわれる通信の時間が多少ずれているためであると考えられる。このずれを考慮にいれなければ同一セグメント上で通信をおこなう場合には、全通信データサイズが 4000 から 5000 バイトを越えるようにプログラムを設計するとネットワーク性能の最大を利用するといえる。

組数	データサイズ	全通信データサイズ
1 組	4000 バイト	4000 バイト
2 組	2500 バイト	5000 バイト
5 組	1200 バイト	6000 バイト
8 組	700 バイト	5600 バイト
10 組	600 バイト	6000 バイト
13 組	500 バイト	6500 バイト
15 組	350 バイト	5250 バイト

表 2: 1 対 1 通信で競合がおこるデータサイズ

3.2 1 対多通信

測定方法

通信のコストとしては、1 対 1 通信の時と同じく pvm_pack および pvm_unpack を含めたマルチキャスト-レシーブおよびセンド-レシーブの往復の経過時間を測定した。1 組は 6 台で構成し、1 台が pvm_mcast および最後の収集をおこない、残り 5 台が pvm_recieve に続き pvm_send をおこなう。データサイズは 32 バイト から 51200 バイト まで変化させ、同時に 1 組から 5 組のマルチキャスト-レシーブをおこなった。1 対 1 通信の場合と同じく、25 回から 50 回のセンド-レシーブの繰り返しを各 PU ごとに測定した経過時間より算出し、その最大経過時間を全体の経過時間とした。

測定結果

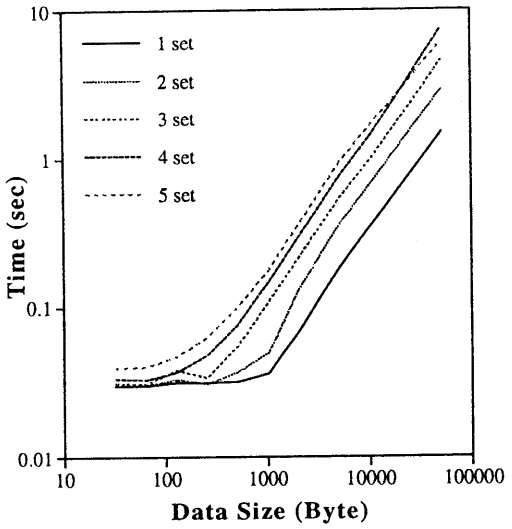

図 3: 1 対多通信の場合の競合

1 対多通信の場合の結果を図 3 に示す。1 組の 1 対 5 通信の競合が始まるデータサイズは約 1000 バイトである。つまり 1 組の 1 対 5 マルチキャスト-レシーブでは、1 組の 1 対 1 のセンド-レシーブと比較して約 4 倍のデータが同時にイーサネット上に流れていると推測できる。これは、1 対 1 通信に比較して、1 対 5 通信がより効率良くおこなわれていることを示している。実際に 1 対 5 のマルチキャストでおこなわれるデータ量は 5000 バイトであるから、pvm_mcast では約 80 %の通信が並列におこなわれていることを示唆している。

3.3 多対多通信

測定方法

多対多通信にはいろいろな形が考えられるが、今回は PU を 2 次元に並べて考えた時に上下左右それぞれの PU とセンド-レシーブをおこなう構成を取り

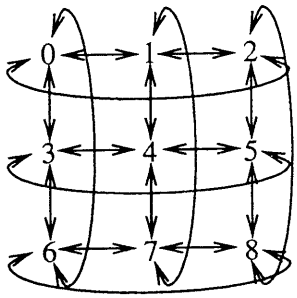

図 4: 多対多通信の構造

上げてた。

通信のコストは、1 対 1 通信の時と同じく pvm_pack および pvm_unpack を含めたセンド-レシーブの経過時間を測定した。全 PU が同時に pvm_send を始め、送られてきた順に pvm_recv をおこなう。データサイズは 128 バイト から 51200 バイト まで変化させ、25 回から 50 回の繰り返しを各 PU ごとに測定した。その経過時間より 1 回経過時間を算出し、その最大経過時間を全体の経過時間とした。

測定結果

多対多通信の場合の通信の競合がおこるデータサイズを表 3 に示す。

PU 数	データサイズ
3 × 3 (9 PU)	1000 バイト
4 × 4 (16 PU)	500 バイト
5 × 5 (25 PU)	300 バイト

表 3: 多対多通信で競合がおこるデータサイズ

通信の競合がおこりはじめるデータサイズが PU の数にほぼ反比例した大きさのところに明確に現れており、ここでもほぼ一定のデータサイズでネット

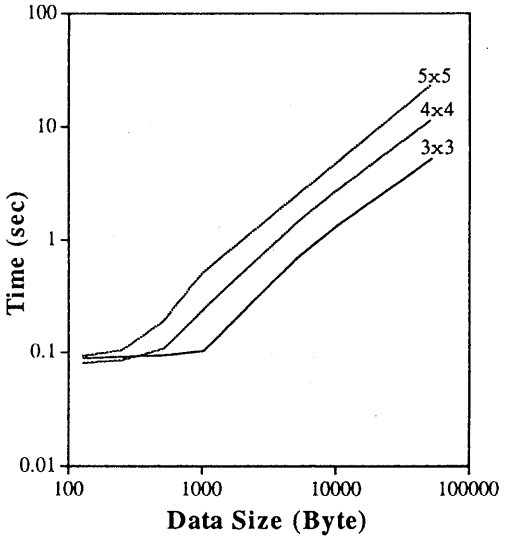

図 5: 多対多通信の場合の競合

ワークが飽和していることがわかる。

4 まとめ

本研究では、大規模ワークステーションクラスタ上での並列プログラム設計に先立ち、同一セグメント上に多くの計算機が存在する環境下での通信の競合を観察した。同一セグメントのイーサネット上で約 4000 バイトのデータが同時に通信された場合にネットワークの競合がおこることが分かった。また、マルチキャスト通信の場合には、1 組の 1 対 5 通信で約 80 %の通信が並列におこなわれていることが分かった。

今後の課題としては、大規模ワークステーションクラスタというシステムに適した性能評価プログラムを設計し、実際に性能評価ををおこなうこと。また、その結果を汎用並列計算機と比較することが挙げられる。

謝辞

本研究をおこなうにあたり、NAS Paralell BenchmarksProgram をご提供くださった筑波大学の板倉憲一氏に深く感謝致します。また、ご助言をいただきました電子技術総合研究所の関口智嗣氏、佐藤三久氏、筑波大学の朴泰祐先生、中村宏先生にも感謝致します。

参考文献

[1] 日本機械工業連合会, 日本応用数理学会, “平成 3 年度 次世代スーパーコンピュータ性能評価技術 調査研究事業報告書”,

[2] 弘中哲夫, “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”, *IPSJ SIG Notes*, 95-HPC-55, 1995, pp. 89–96.;

[3] D. Bailey, E. Barszcz, J. Barton, D. Browning, R. Carter, L. Dagum, R. Fatoohi, S. Fineberg, P. Frederickson, T. Lasinski, R. Schreiber, H. Simon, V. Venkatakrishnan and S. Weeratunga, “THE NAS PARALLEL BENCHMARKS”, *RNR Technical Report*, RNR-94-007, March 1994.;

[4] Al Geist, Adam Beguelin, Jack Dongarra, Weicheng Jiang, Robert Mancheck and Vaidy Sunderam, “PVM:Parallel Virtual Machine A User’s Guide and Tutorial for Networked Parallel Computing”, *MIT Press*, 1994.;