

# 広域インターネットを用いた並列計算実験

陣崎 明、水野 裕識、古賀 久志、下國 治

株式会社 富士通研究所 コンピュータシステム研究所  
〒211-8588 川崎市中原区上小田中 4-1-1  
*E-mail: { zinzin , mizuno, koga, osamus }@flab.fujitsu.co.jp*

従来、インターネット通信は通信処理オーバーヘッド、伝送距離による遅延が存在するため、並列計算のように高帯域、低遅延を必要とするアプリケーションには向かないと考えられてきた。しかしながら、インターネット伝送速度が向上し、通信プロトコルの高速処理技術が進展している現在、インターネット技術を基盤とした並列分散システムの構築可能性を検討することは重要である。そこで、ATMによる広域インターネットにおいてNAS Parallel Benchmark(NPB)を実行し、通信パターンの観測と並列計算の性能評価を行った。その結果、1) 並列計算アルゴリズムによっては有効な結果を得る場合があること、2) 広域網の通信品質に問題はないが、性能は帯域や遅延に大きな影響を受けること、3) 現在のNPBをそのまま利用する限り台数効果(スケーラビリティ)が制限されることがわかった。

インターネット、NAS Parallel Benchmark、MPI/LAM、並列分散システム、ATM

## An Experiment of Parallel Computation on The Internet

Akira Jinzaki, Hironori Mizuno, Hisashi Koga, Osamu Shimokuni

Computer Systems Laboratories, Fujitsu Laboratories LTD.  
1-1, Kamikodanaka 4-Chome, Nakahara-ku, Kawasaki, 211-8588 Japan  
*E-mail: { zinzin, mizuno, koga, osamus }@flab.fujitsu.co.jp*

Parallel applications which require high bandwidth and low delay are not suitable because of communication processing overhead and absolute delay of long distance. At the other hand, as internet protocol processing is improved and the transferring speed is higher, it's important to think of the possibility of parallel-distributed systems in wide-area.

From this point of view, to examine the practicality of parallel computing on the Internet, we observe the traffic patterns and make performance evaluation when the NAS parallel benchmarks are run on the wide-area Internet connected by 90km ATM lines. As the results, we obtain the following three results, i.e. 1) wide-area parallel-distributed systems can work effectively depending the parallel algorithm used, 2) the performance is influenced by delay and bandwidth of the wide-area network, though its communication quality is sufficient, and 3) the scalability in terms of the number of processors which participate parallel computing is restricted.

Internet、NAS Parallel Benchmark、MPI/LAM、Parallel-distributed systems、ATM

## 1. はじめに

インターネットの普及に伴って Gigabit Ethernet (GbE)、10 Gigabit Ethernet、SONET など高速かつ低コストな標準ネットワークの実用化が急速に行われており、このような標準ネットワーク技術を用いて計算機システムを構築することが現実的となりつつある。例えば次世代 I/O 技術である InfiniBand[1]では IPv6 アドレスを採用し、ルータを介したインターネットへの直接接続を構想している。OC48c、OC192 などの広域接続技術を用いれば 2.5Gbps ~ 10Gbps という高帯域な I/O 接続が地球規模のインターネットで実現可能となる。このような動きは並列分散システムのプロセス間通信ネットワークにも波及していくと考えられる。

従来、インターネット通信はプロトコルが複雑で、通信処理オーバーヘッドが大きいとされてきた。プロトコル処理を行うために CPU パワーを浪費する上、帯域性能や遅延性能が悪い。また広域システムにおいては伝送距離の長さ起因する絶対的な遅延が存在するため、並列計算のように高帯域、低遅延、低オーバーヘッドを必要とするアプリケーションにインターネットは向かないと考えられている。従来から並列計算アルゴリズムの研究において広域インターネットでの実験を行ったものが散見される(最近では[2])。しかしこれらの研究は小さい通信オーバーヘッドで問題を解くアルゴリズムを開発することを主眼とするもので、「インターネットでも使える」という立場である。

しかしながら、インターネットの物理的な伝送速度が 10Gbps 台となり、通信プロトコルを高速処理する技術の研究開発が進展している現在、定量的な性能評価をもとにインターネット技術を基盤とした並列分散システムの構築可能性を検討することは非常に重要と考えられる。

以上の見地から我々は実際の広域インターネットを用いた並列計算ベンチマークを行った。本報告では我々が構築した実験システムの構成と実験結果を示し、インターネットによる並列分散システムの実現性について考察を行う。

## 2. 実験の目的

従来、シミュレーションにより並列計算ベンチマークにおける通信性能の影響を定量評価した結果が報告されている[3]。この研究では通信帯域と通信遅延の平均

値によってネットワークを表現し、通信性能と並列計算効果について興味深い結果を示している。しかし、この研究ではもともとインターネット環境を想定していないため、通信遅延を 500 $\mu$ s 以下に設定し、帯域は無制限とするなど条件設定が広域インターネットの条件に合致しない面がある。またプロトコルスタックのオーバーヘッドや、通信性能のばらつき(帯域の増減、遅延ジッタ)などは考慮されていない。

従来の結果を踏まえ、我々は実際のインターネットを用いた実験を行い、詳細な通信特性を調査することで、今後の広域インターネットでの並列計算の可能性を調べることにした。具体的に以下のような項目を明らかにすることを最終的な目的とする。

### • 通信帯域の影響

インターネットの通信帯域は Gbps を超えており、今後も WDM 技術等によって向上していくと考えられる。問題は共用ネットワークの輻輳による帯域低下である。インターネットでは Diffserv などの優先制御が行われているが、このような技術の効果も興味深い。

### • 通信遅延の影響

RTT (Round Trip Time) は信頼性のある通信を行うために必要な acknowledging の性能に大きく影響する。遅延には伝送距離による遅延とルータなど通信装置における遅延がある。伝送遅延は 5 $\mu$ s/km、5ms/1000km 程度なのに対して通信装置の遅延は装置数に比例するが、高速化が行われているので伝送遅延レベルとなる可能性がある。そのような動向を含めて広域インターネットの遅延が並列計算にどのような影響を及ぼすかは興味のある問題である。

### • 通信品質の影響

今日の高速インターネットは光通信技術を用いており、いわゆる伝送エラーは非常に小さい。しかし、ATM 網におけるセル廃棄やルータやホストシステムのバッファ枯渇によるパケット廃棄は常にありうる。一度、パケットが廃棄されると再送のためにトラフィックが増加し、さらにパケット廃棄を誘う可能性もある。特にインターネットのような共用ネットワークでは予測できない状況で輻輳状態が発生する可能性があるため、安定した通信特性を必要とする並列計算に与える影響は大きいとみられる。

- スケーラビリティ

インターネットを用いて並列計算をする意味は一義的に「多くの計算機資源を利用できる」という点にある。これからのインターネットが提供可能な実効通信容量でどのような種類、規模の並列計算が可能かは興味深い問題である。

もとより広域インターネットは複雑なシステムで、上記の項目を明らかにすることは単純な作業ではないし、一通りの実験で結果が得られるものではない。今回の実験は第一ステップと位置付けている。

### 3. 実験システム

#### 3.1. 計算機ハードウェア

実験システムは PC 互換機をインターネットを用いてクラスタ化したいわゆる PC Cluster である (図 1)。

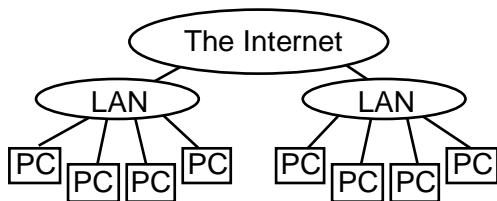


図 1 実験システム

ネットワークを用いた並列計算では CPU 性能とネットワーク性能のバランスが重要である。並列計算のための通信オーバーヘッドが CPU 性能に対して相対的に小さければ大きな台数効果が得られ、逆に大きければ台数効果は低下する。そこでクラスタの要素計算機として高速マシン (Fast-PC)、低速マシン (Slow-PC) の二種類を準備し (表 1)、Fast-PC は最大 8 台、Slow-PC は最大 9 台からなるクラスタを構成した。

表 1 クラスタ要素計算機

	CPU	メモリ
Fast-PC	Xeon 733MHz	512MB RDRAM
Slow-PC	Pentium Pro 200MHz	256MB DRAM

#### 3.2. ソフトウェア

現在 PC Cluster では標準的に Linux が用いられているが、本実験システムでは OS として BSD/OS3.1[4]を用いた。BSD/OS の Internet Protocol Suits は多くの IP ソフトウェアのベースとなっている BSD UNIX のもの [5]を採用しており、伝統的な標準と見なすことができる。新しく開発された Linux のほうが一般に通信性能が良いと考えられるが、ここではインターネットの性

能評価を主眼として敢えて標準的なプロトコルスタックを採用した。

並列計算ベンチマークとしては NAS Parallel Benchmark2.3 (NPB) [6]を使用した。NPB を動作させるために MPI/LAM-6.4-a3[7]を BSD/OS に移植した。NPB は BT 以外のベンチマークを Class S, W, A で実行した。BT は BSD/OS ではシステムエラーにより実行できなかったため今回の実験では除外した。

#### 3.3. ネットワーク

表 2 に示すネットワークを準備した。RTT は "ping -s 64" の結果である。

表 2 ネットワーク

	ネットワーク構成	RTT
100M-LAN	100Base-T、Switch 接続	226 $\mu$ s
67M-ATM	135M Megalink Wrap、60Km	12ms
135M-ATM	135M Megalink、90Km	12ms
(1G-LAN)	1000Base-SX、Switch 接続	331 $\mu$ s

100M-LAN は PC が標準装備している 100Base-T アダプタを用いたもので、ドライバは BSD/OS 標準である。100Base-T コントローラとしては、Fast-PC、では Intel EEPRO10/100 を、Slow-PC では DEC21140-AC を用いている。

ATM 接続は WIDE Project[8]が維持運用する WIDE Backbone を使用し、川崎市の富士通研究所から藤沢市の慶應義塾大学湘南藤沢キャンパス (SFC)、大手町 NOC を経由して文京区本郷の東京大学までを専用の VC で接続した (図 2)。ネットワーク経路は全て ATM スイッチで構成され、ルータは介在しない。このため通信遅延はほぼ同じである。

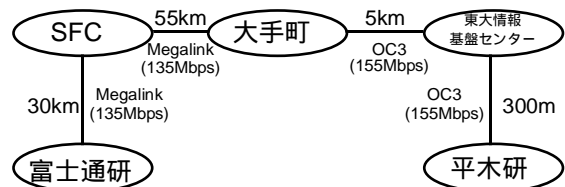


図 2 広域ネットワーク

135M-ATM におけるエンド・エンドの帯域は 135Mbps (双方向で 270Mbps) である。

67M-ATM は 135M-ATM を川崎から 30Km の SFC で折り返して使用したもので、伝送路としては特殊である。片方向だけならば 135Mbps の通信が可能だが、

双方向では半分の 67Mbps となる上、これを越えると上りと下りがセル廃棄を誘発して効率が低下する。

本実験で準備した広域ネットワークは最高 135Mbps (双方向で 270Mbps) の帯域をもつので、ローカルエリアでもこの帯域を上回る性能で接続する必要がある。そこで ATM 接続時の LAN は GbE で構成した。1G-LAN は我々が開発した Comet アダプタ[9]に GbE を搭載したものである。PCI が 32bit、33MHz のため、実効性能は単方向で 30MB/sec、双方向で 34MB/sec 程度であるが、実験ネットワークに対して LAN がボトルネックとならず、かつ ATM との速度差が極端に大きくない。ATM-GbE ブリッジにも Comet アダプタを用いている。

### 3.4. その他のツール

並列計算を行っている時の通信の様子を観測するためのモニタリングツールとして Traffic Monitor (Comet TM) を準備した。Comet TM は GbE を 2 個搭載した Comet アダプタであり、GbE ブリッジとして動作しつつフォワードしたパケットに関する統計情報を記録するものである。Comet TM は 1G-LAN のアダプタと同じハードウェアを用いているが、最高転送性能は 300Mbps と、135M-ATM に対してボトルネックとならない。実験ではパケット長 (64B 単位) 毎のパケット数と転送量を 0.1 秒毎に計測することで、並列計算実行時の通信状況を実時間観測した。

さらに、Switch などネットワーク装置における比較的マクロなトラフィック観測のために MRTG[10]を用いた。MRTG はインターネット管理でよく用いられるツールで、1 分平均、30 分平均といった大掴みなレベルのトラフィックを記録し、グラフ化する機能があり、Web ブラウザを用いて手軽に結果を参照できる。

## 4. 実験形態

### 4.1. LAN 接続実験

LAN 接続実験の形態を図 3 に示す。

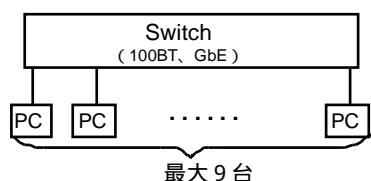


図 3 LAN 接続実験形態

Switch を中心に最大 9 台の PC Cluster を構成した。こ

の形態では経由する Switch の数が性能に影響を及ぼすので、可能な限り少ない Switch で接続するようにした。Comet TM も除いた。性能のばらつきが Switch の機種に依存する傾向もみられたが、Class A では比較的安定した。

### 4.2. 広域接続実験

広域接続試験は 67M-ATM と 135M-ATM を用い、PC Cluster を 4 台と 4 台に振り分ける形で行った。PC 同士が完全グラフ的な通信をする場合、広域ネットワークのトラフィックが最大となる。

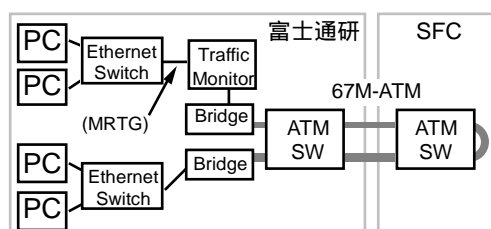


図 4 広域折り返し実験形態 (67M-ATM)

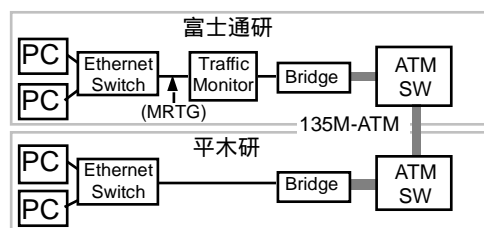


図 5 広域実験形態 (135M-ATM)

ATM においてセル廃棄の影響は重要である。帯域まではエラーなく通信できるが、少しでも帯域を越えるとセルロスが発生する。図 6 は 135Mbps Megalink において帯域超過通信を行った時の通信性能を示す。120Mbps 程度で定常的にエラーのない通信ができていた状態でさらに負荷を投入するとランダムなセルロスが発生し、パケット廃棄がおこる (22 秒 ~ 28 秒)。その上に負荷を投入するとさらにパケット廃棄が増大する (28 秒 ~ )。

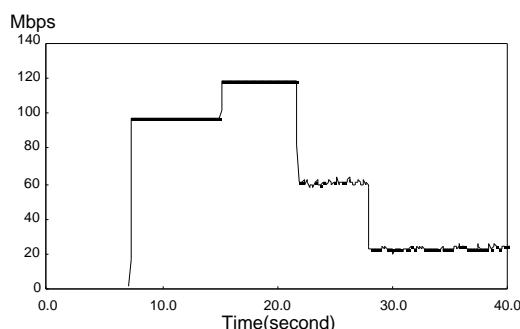


図 6 ATM の伝送路特性

本実験では共用の実験ネットワークである WIDE Backbone を用いている。このため、他トラフィックが増大するとセル廃棄が発生する可能性がある。本来の目的はこのような輻輳の影響を調べることであるが、今回は基本性能を測定するため、他実験者と実験期間を調整し、他トラフィックの影響を押さえるよう配慮した。また、実験後に ATM スイッチのセル廃棄状況を調べ、実験期間にセル廃棄が発生していないことを確認した。

図 7 に実験を行っていた期間中に ATM を流れた通信量の推移 (MRTG による) を示す。表示値は 1 分平均、30 分平均でピーク性能ではない。平均すると帯域に余裕があることがわかる。

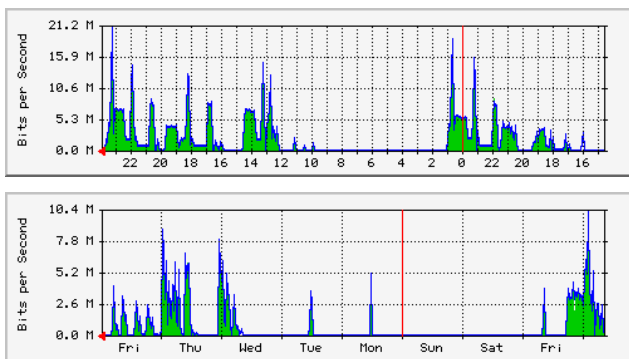


図 7 実験中のトラフィック (MRTG)

## 5. 実験結果

本章では NPB の各ベンチマークを実行した結果を示す。通信パターンは全て Class W、67M-ATM、8 ノード (SP のみ 4 ノード) の実行開始から完了までを観測したものである。転送量は Ether フレームヘッダ、IP ヘッダなどを全て含む双方向の総転送量を Mbps に換算して表示している。パケット数は 0.1 秒毎の双方向のパケット長別の総転送パケット数を pps (packet per second) に換算し、時間軸を拡大して一部分を表示している。台数効果は全て Class A の結果である。100M-LAN、135M-ATM、67M-ATM と Fast-PC、Slow-PC の組み合わせで測定した。

### 5.1. NPB CG

CG は一定量の通信が絶えまなしに継続するため、ネットワーク性能の差がそのまま台数効果に表れる。転送パケット長についても同程度のパケットが継続して流れている。遅延、帯域ともに影響があるが、100M-LAN のほうがよいことから遅延の影響が大きいと考えられる。

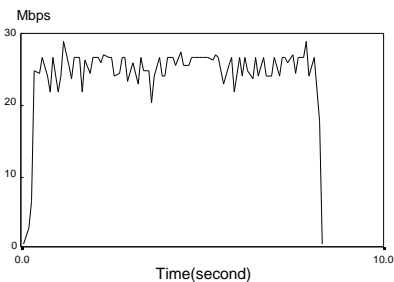


図 8 CG 通信パターン (転送量)

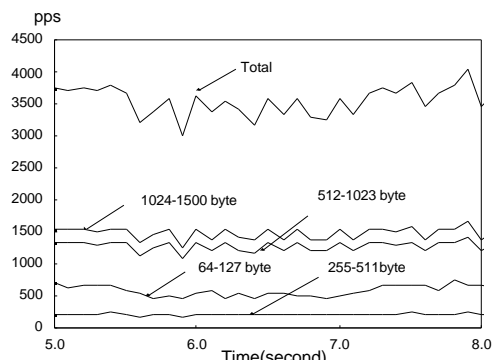


図 9 CG 通信パターン (パケット数)

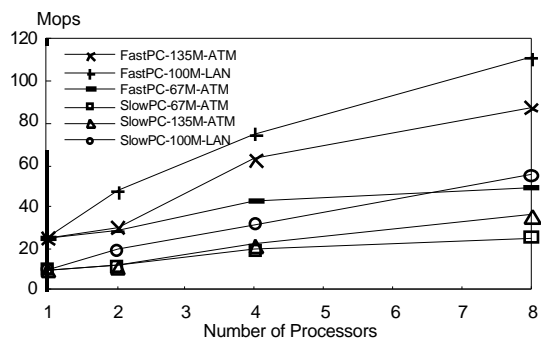


図 10 CG 台数効果

### 5.2. NPB EP

EP はほとんど通信が行われないため、広域ネットワーク向きである。パケット数は省略した。

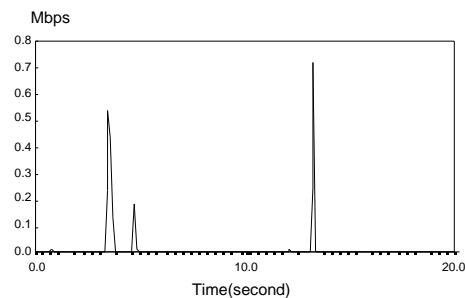


図 11 EP 通信パターン (転送量)

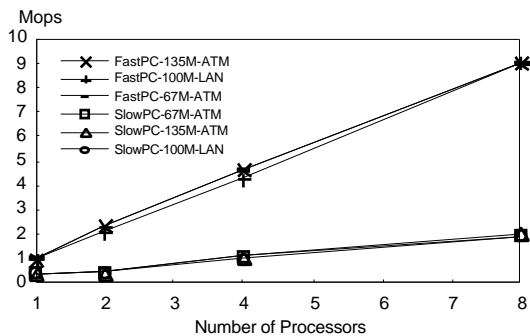


図 12 EP 台数効果

### 5.3. NPB FT

FT は比較的ばらついた通信を行うため通信性能の影響を受けやすいと考えられたが、比較的素直な台数効果が得られた。Slow-PC ではネットワーク差が非常に小さい。

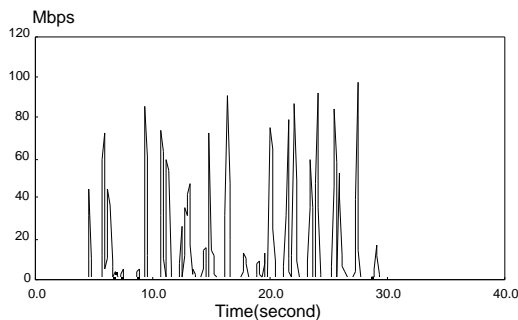


図 13 FT 通信パターン (転送量)

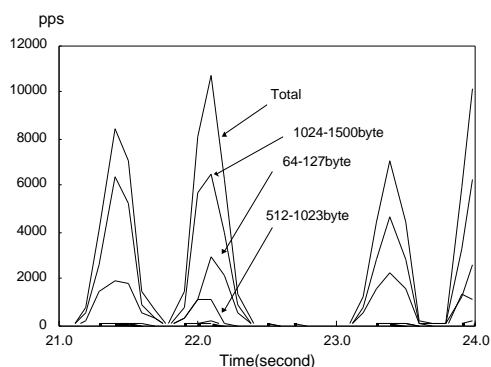


図 14 FT 通信パターン (パケット数)

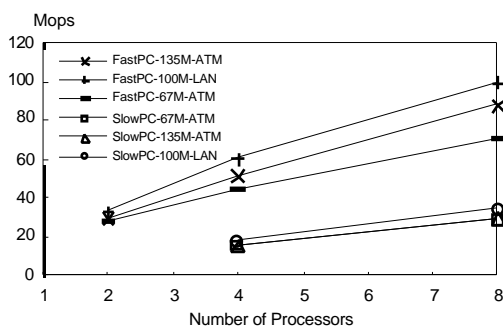


図 15 FT 台数効果

### 5.4. NPB IS

IS は台数効果が得られず、結果もかなりばらついた。通信パターンは FT と類似しているが、ストリームの発生する間隔が比較的広いかわりにストリームの立ち上がりが急峻な点が異なる。FT より瞬発力のあるデータ転送を要求していると考えられる。Linux Redhat6.2 で同様の試験を行ったところ、台数効果はなかったが 2 台以降の性能低下はなかった。この結果からみて、BSD/OS の通信特性に問題があるものと推察されるが、詳細な説明はできていない。

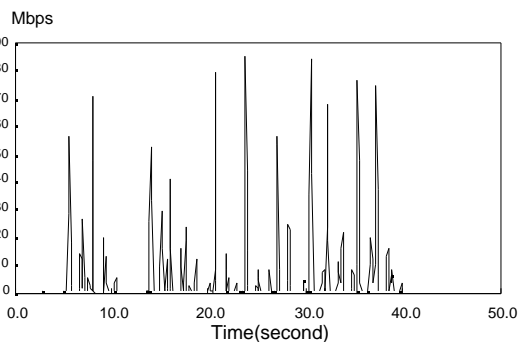


図 16 IS 通信パターン (転送量)

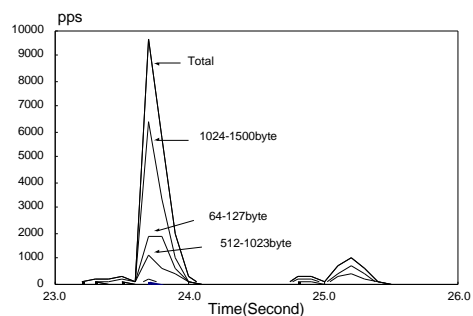


図 17 IS 通信パターン (パケット数)

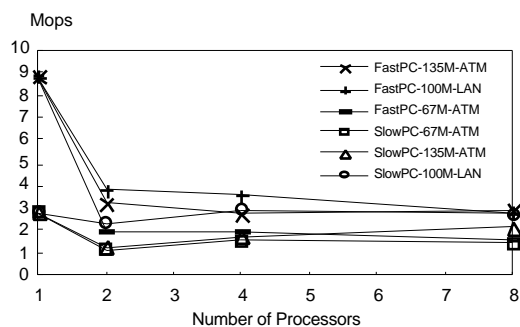


図 18 IS 台数効果

### 5.5. NPB LU

LU は CG について定常的な通信を継続している。256 ~ 511 バイトのパケットと 512 バイト以上のパケットが交互に流れ、帯域的には余裕がある。

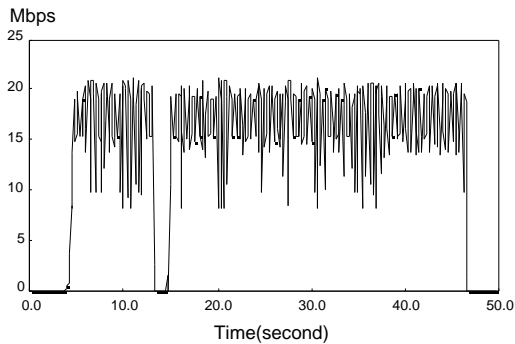


図 19 LU 通信パターン (転送量)

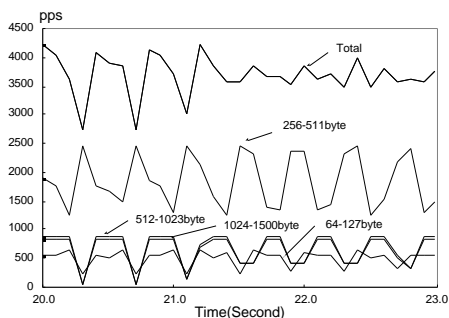


図 20 LU 通信パターン (パケット数)

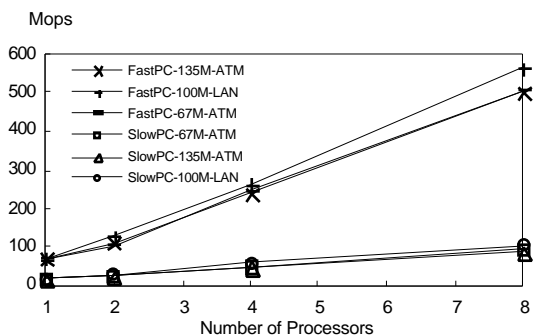


図 21 LU 台数効果

## 5.6. NPB MG

MG の通信パターンは FT よりもまばらでピークが低いが一回の経過時間が長い。Fast-PC では遅延と帯域の両方の影響をかなり受けているが、Slow-PC では影響は小さい。なお MG は Slow-PC 1 台では動作しなかった。

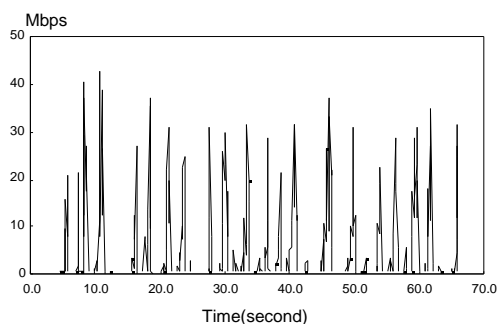


図 22 MG 通信パターン (転送量)

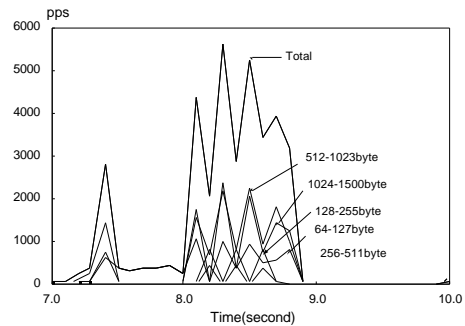


図 23 MG 通信パターン (パケット数)

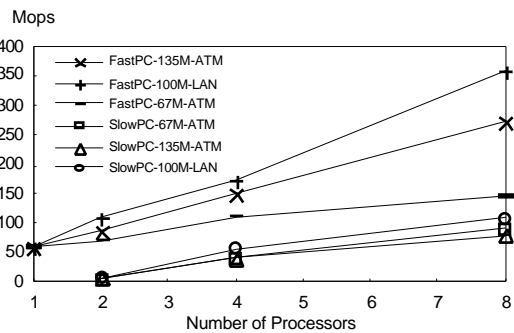


図 24 MG 台数効果

## 5.7. NPB SP

SP は機材の準備の関係で、Slow-PC の 67M-ATM のみ 9 台で、あとは 4 台で測定した。通信パターンは FT と MG の中間のような特性となっている。Fast-PC ではネットワーク性能の差が台数効果に表れている。

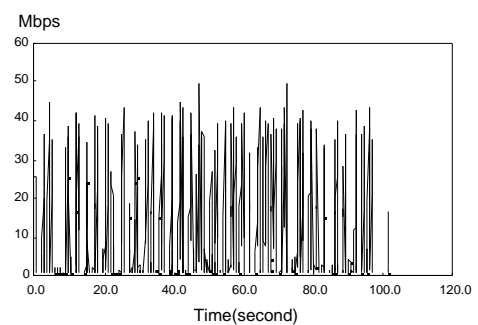


図 25 SP 通信パターン (転送量)

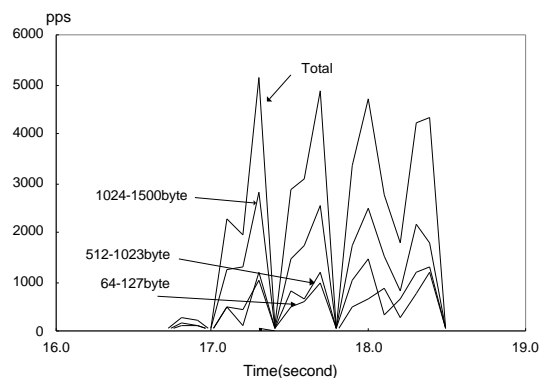


図 26 SP 通信パターン (パケット数)

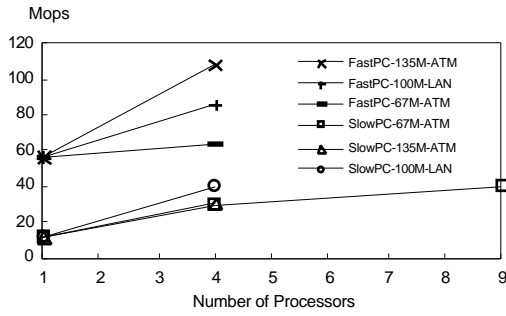


図 27 SP 台数効果

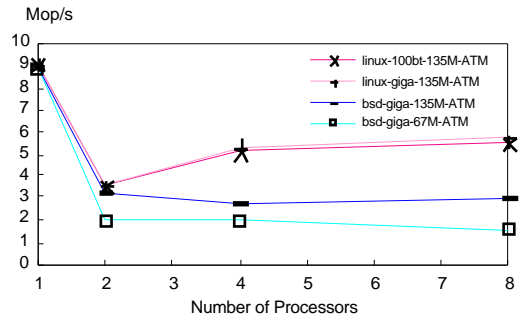


図 31 IS 台数効果

## 6. プロトコルスタックの影響

BSDI 社の BSD/OS 3.1 と RedHat Linux 6.2e とで NPB 性能の違いを調べた。全般に Linux のほうが高い性能が得られることがわかった。

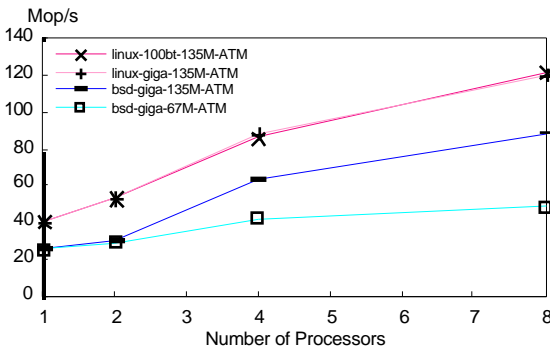


図 28 CG 台数効果

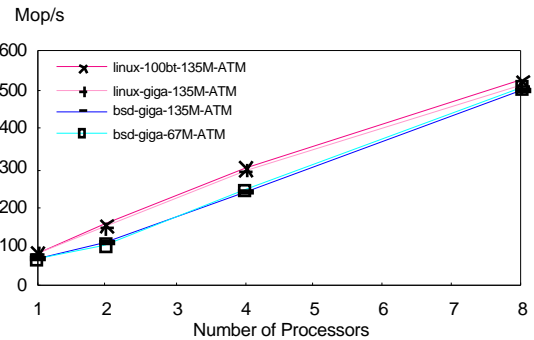


図 32 LU 台数効果

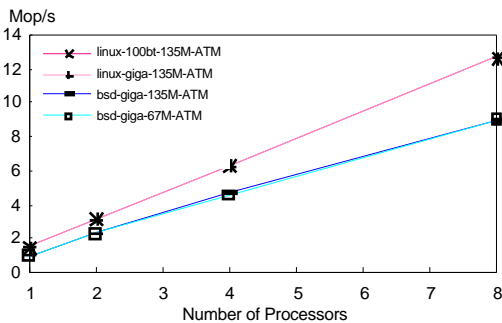


図 29 EP 台数効果

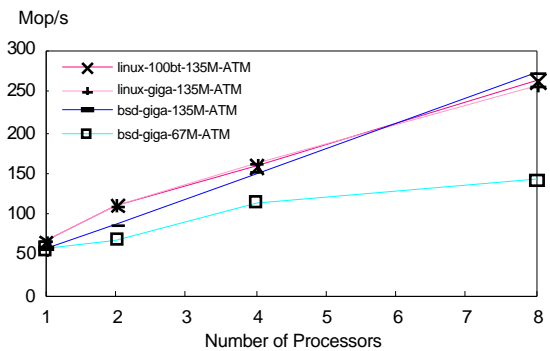


図 33 MG 台数効果

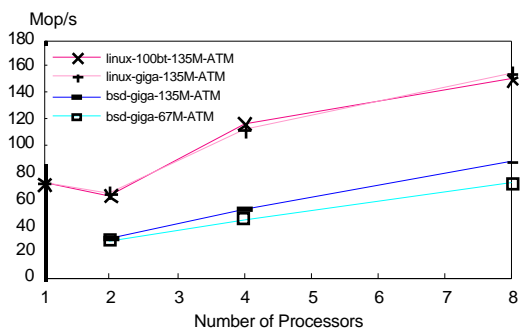


図 30 FT 台数効果

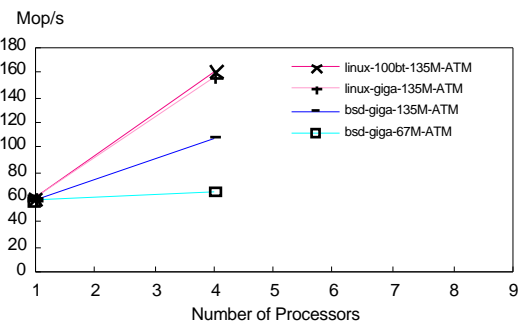


図 34 SP 台数効果

プロトコルスタック実装の違いの解析については今後の課題であるが、一つの要因として Linux では BSD/OS の mbuf のようにパケットを細切れにしないため、バッファ管理の手間が小さいことが考えられる。

## 7. 実験結果のまとめ

### 7.1. 台数効果

表 3に実験結果と同じプロセッサ数の既存システム (IBM<sup>[11]</sup>、Sun<sup>[11]</sup>) との台数効果の比較を示す。

表 3 台数効果の比較

	Fast-PC		Slow-PC		IBM SP66	SUN
	100M-LAN	135M-ATM	100M-LAN	135M-ATM		
CG	4.51	5.50	5.50	3.57	4.27	-
EP	8.81	8.94	6.03	6.28	8.07	7.73
FT	5.07	4.73	6.06	6.83	-	5.7
IS	0.32	0.33	0.97	0.75	5.52	6.88
LU	8.41	7.47	6.57	5.53	7.17	8.4
MG	6.11	4.63	5.23	6.66	7	7.07
SP	1.50	1.90	3.51	2.63	3.6	3.68

135M-ATM でも CG、EP、FT、LU、MG に関して比較できるレベルの台数効果が得られている。

### 7.2. 帯域、遅延の影響

Fast-PC では 100M-LAN がよいことから遅延の影響が大きいと考えられる。Slow-PC では相対的に通信性能の違いによる影響が小さい。SP は遅延より帯域の影響が大きい。また LU は帯域の差が小さい。

### 7.3. 通信品質の影響

伝送路の通信品質に関しては、実験ネットワークの構成では伝送エラーはなく、その意味で広域 ATM 網は高信頼通信に耐えることがわかった。ネットワーク共用の影響については今回は評価しなかったが、通信量の観察結果からみて利用率が 3~5 割程度の ATM 網ならば並列計算に利用可能との感触をもっている。

### 7.4. スケーラビリティ

台数効果の結果をみると 8 台ですでに台数効果の低下がみられる。伝送路として OC-12 (600Mbps)、GbE、OC-48c などを利用すれば実験システムの 10 倍以上の帯域を実現可能であるが、現在のままで伝送路だけ高速化しても 10~20 台程度のクラスタしか有効でないように思われる。遅延を想定した NPB のチューニングを考える必要がある。

## 8. おわりに

インターネットを用いた並列計算の実現性を検討するための基礎検討として、90Km の ATM による広域インターネットで接続した PC Cluster で NPB を実行し、通信パターンの観測と並列計算の性能評価を行った。

この結果、並列計算アルゴリズムによっては有効な結果を得る場合があること、伝送品質については問題がないこと、現在の NPB をそのまま使用する限りスケーラビリティが制限されることがわかった。

またプロトコルスタックの実装に依存して目に見える程度の性能差が生じることから、プロトコル処理ソフトウェアレベルでの性能改善に意味があることがわかった。

今後は帯域、遅延による影響の度合いを定量的に観測するツールを準備し、NPB のチューニングを行いつつ、ルータを含むネットワークなど多様なインターネットを用いて、詳細な観測を行っていく予定である。

## 謝辞

実験ネットワークの構築ならびに実験に関して協力していただいた東京大学理学部平木教授ならびに東京大学情報基盤センター加藤助手を始めとする WIDE Project の皆さんに感謝します。

## 参考文献

- [1] <http://www.infinibandta.org/events/index.html>
- [2] 梅本他、PVM による SAT 並列局所探索プログラム、情処 HPC Vol. 80, No. 17, pp.95--100, 2000.
- [3] 久保田他、"高精度大規模並列プログラムシミュレーション環境による NPB の挙動解析"、情処 HPC、Vol. 98, No. 72, pp.7--12, 1998.
- [4] BSDI: <http://www.bsdi.com>
- [5] G.R.Wright, W.R.Stevens, TCP/IP Illustrated I, II, III, Addison Wesley
- [6] NPB2.3: <http://www.nas.nasa.gov/Software/NPB/>
- [7] MPI LAM: <http://www.mpi.nd.edu/lam/mpi/>
- [8] WIDE: <http://www.wide.ad.jp>
- [9] 小林、陣崎、Comet VIA の評価、SWoPP2000, CPSY, 2000
- [10] <http://ee-staff.ethz.ch/~oetiker/webtools/mrtg/>
- [11] <http://www.nas.nasa.gov/Software/NPB/NPB2Results>