

広域インターネットを用いた 並列計算実験

2000年9月12日

陣崎 明 水野裕識 古賀久志 下國 治
{zinzin, mizuno, koga, osamus}@pds-flab.rwcp.or.jp

富士通研究所

経緯

2000 年春合宿で提案

5 月：実験用計算機、ソフト準備

6 月：ローカル環境での予備実験

7 月：広域環境での実験

8 月：SWoPP2000 で発表

8 月末：Linux での追加実験

本実験を行う動機

高速インターネットを活用できるアプリケーションをさがしたい

並列計算の通信トラフィック特性を把握したい
インターネットの新しい要件をさがしたい

広域高帯域インターネットでの定量評価

実験システム

実験機材（ハード、ソフト）

クラスタ計算機

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

実験ソフトウェア

BSD/OS3.1、MPI/LAM-6.4-a3

RedHat Linux 2.3e、MPI/LAM-6.4-a3

NAS Parallel Benchmark (NPB) 2.3

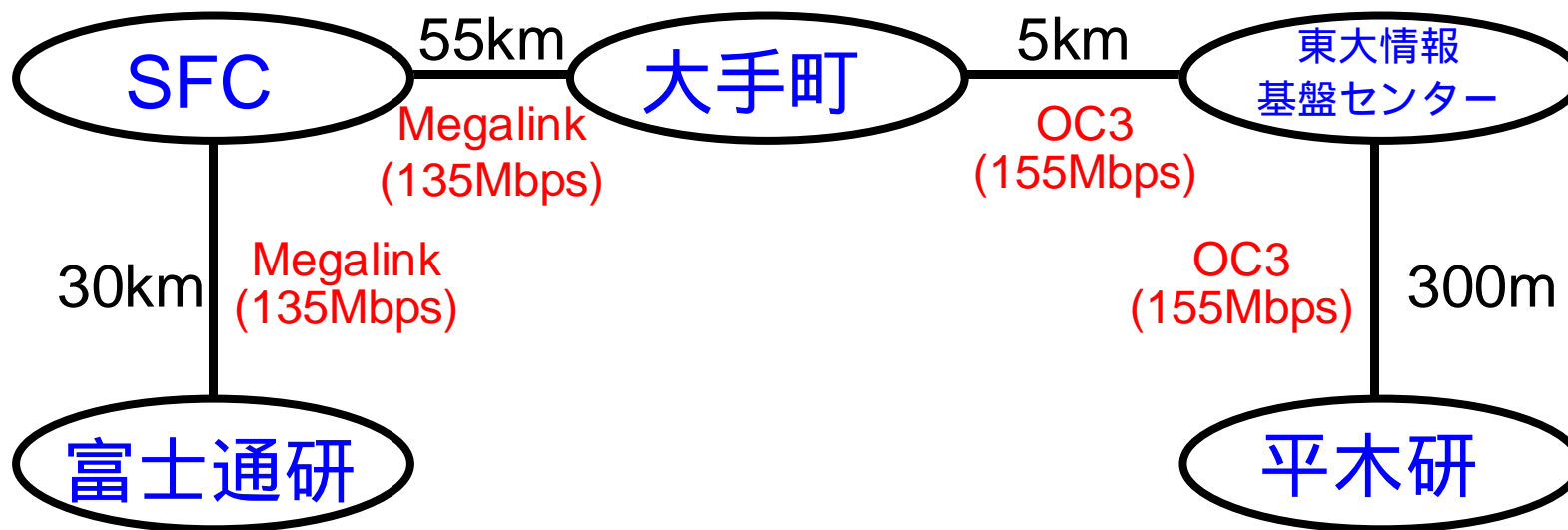
Kernel Benchmarks 5種類 (EP, MG, CG, FT, IS)

Simulated CFD Application 3種類 (LU, SP, BT)

BT は、BSD/OS でシステムエラーで実行できなかった (今回は除外)

実験機材（ネットワーク）

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



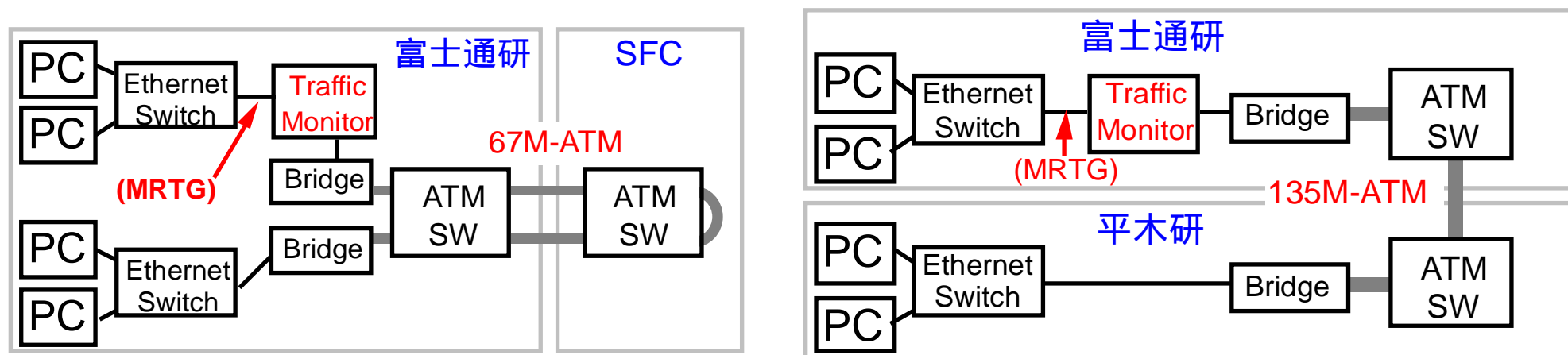
実験形態

ローカル試験

ACE180 Switch に PC を最大 9 台まで 100BT, GbE 接続
性能低下がおきないように、できるだけ少ない Switch で構成

広域試験

PC を 4 台ずつスイッチにつなぎ、ブリッジを経由して、ATM 回線につなぐ。
全体通信をする時、広域ネットワークのトラフィックが最大になる。

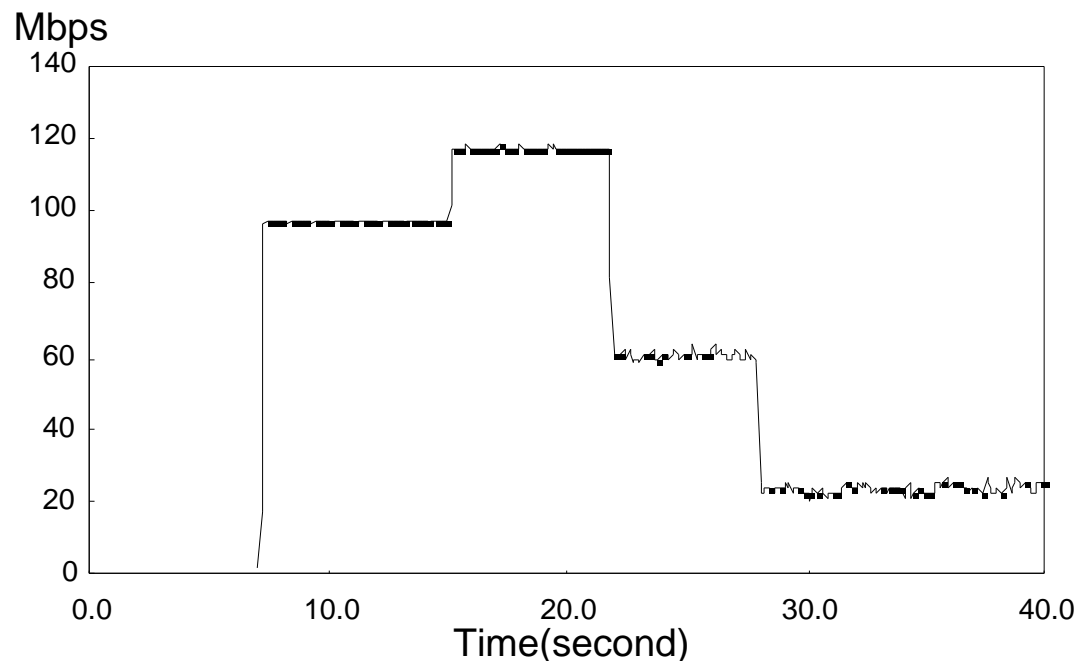


ATM 伝送路の特性

135Mbps メガリンクに UDP の負荷をかけた

120Mbps まで転送可→今回の試験では十分な速度

試験後、ATMSW でのセル廃棄無しを確認



計測内容

NPB による計測（台数効果）：Class A

100M-LAN, 135M-ATM, 67M-ATM と Fast-PC, Slow-PC の組み合わせ

通信パターンの観測：Class W

自作トラフィックモニターによる観測

転送性能は双方向で 300Mbps なので、ATM135Mbps には十分

機能：イーサネットパケットを 1 個ずつ確認して記録

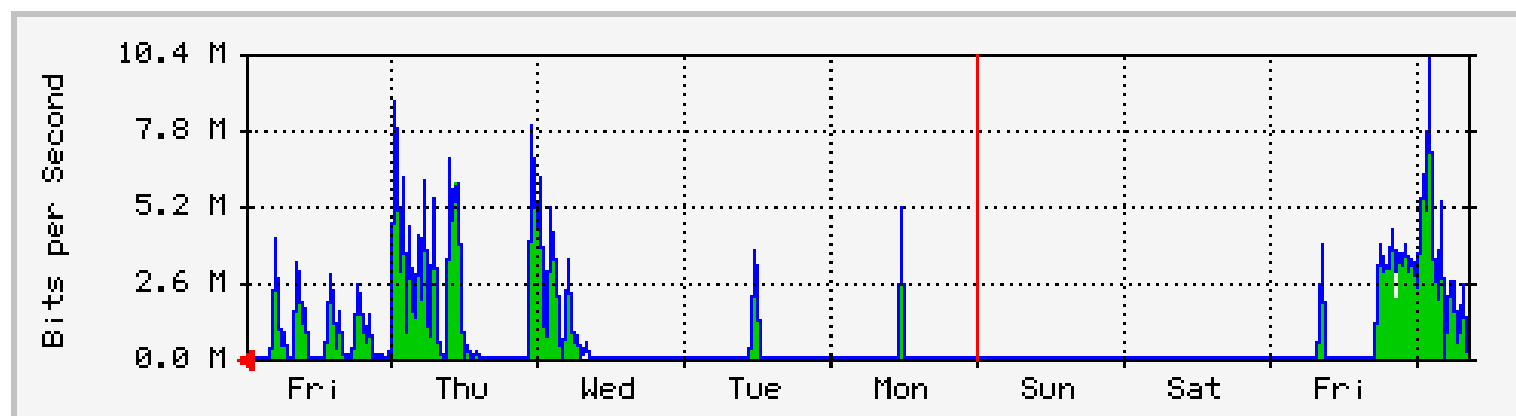
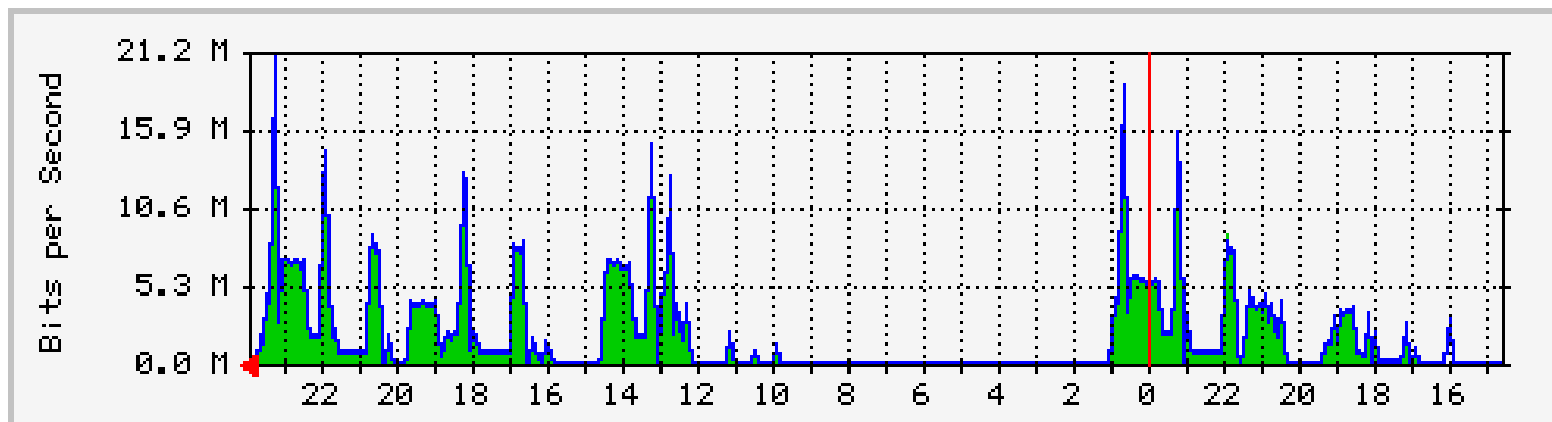
統計情報：パケット数(64B 単位)、0.1 秒ごとに利用帯域を計測

67M-ATM の 8 ノード（SP のみ 4 ノード）の計算を行った時に使用

MRTG による観測

MRTG ツール(1 分、30 分平均の帯域をグラフ表示)

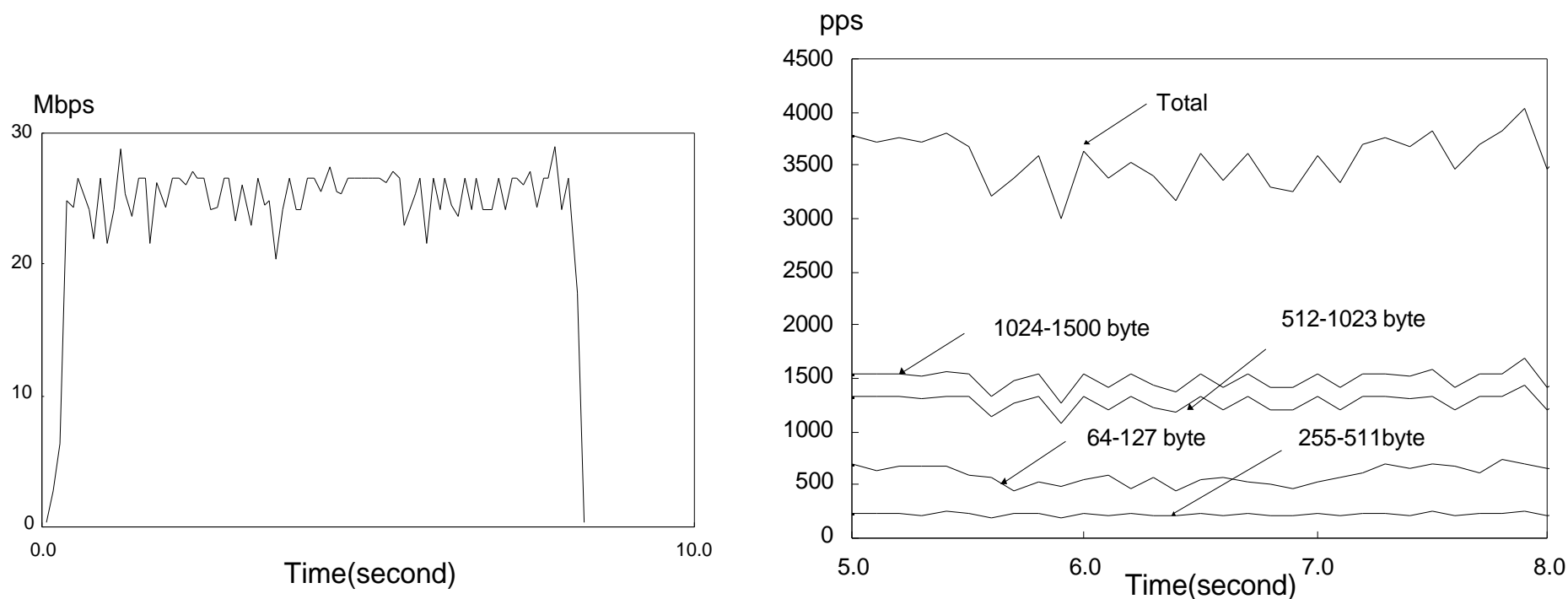
通信帯域の観測(MRTG)



通信状況の観測(トラフィックモニター)

CG : 定常的に 25Mbps 程度の帯域を消費 (下図左)

5-8 秒を拡大して、パケット長ごとにパケット転送数を表示 (下図右)



転送パケット統計情報をグラフ化

結果 1

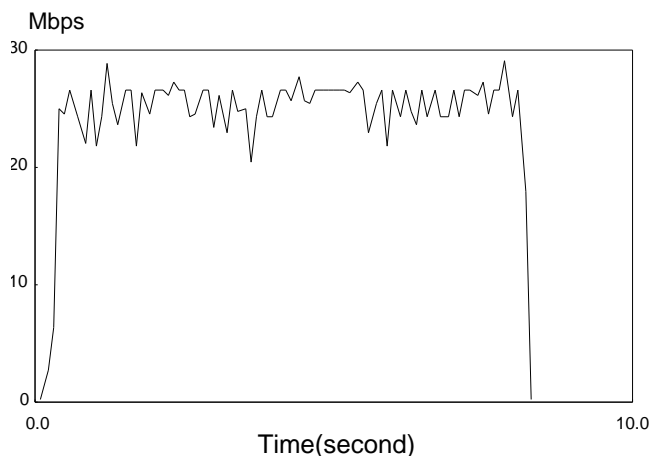
スケーラビリティ（台数効果）

8 台 Mops 値 / 1 台 Mops 値（直線の傾き）

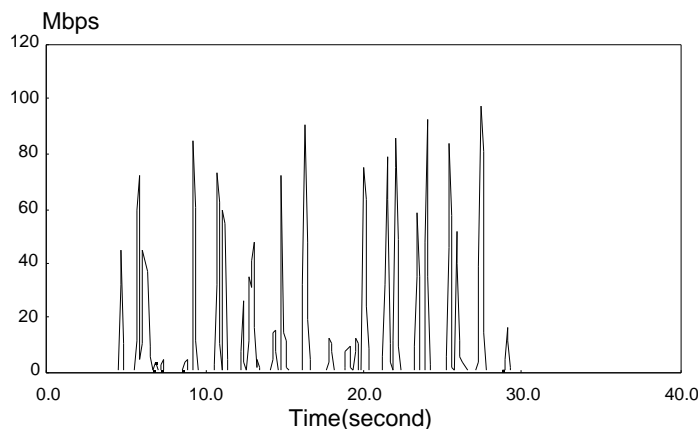
135M-ATM : CG,EP,FT,LU,MG 効果を確認

	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

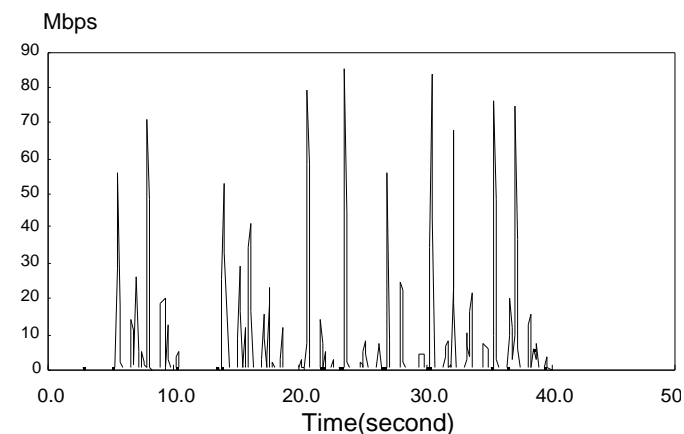
双方向トラフィックパターン(67M-ATM, 8台、双方向)



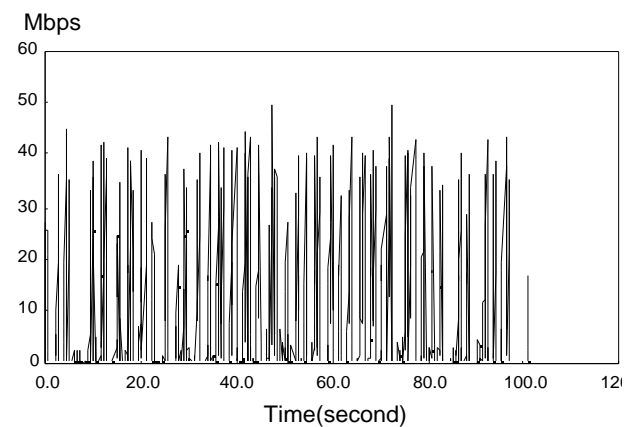
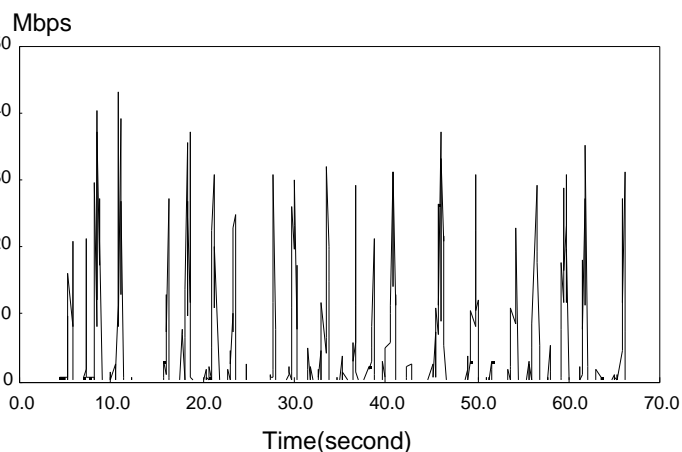
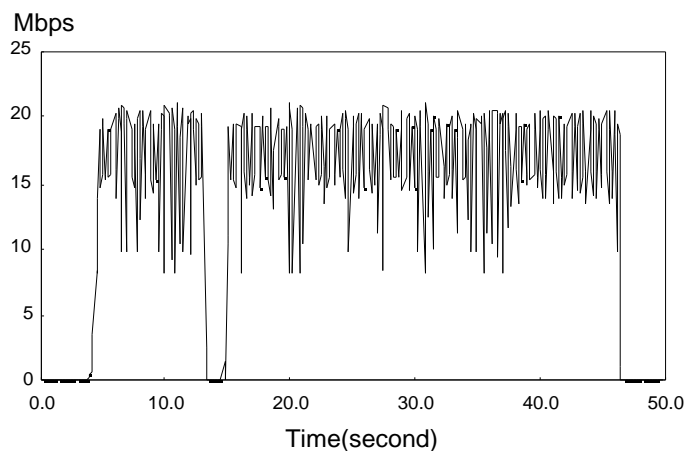
CG LU



FT MG



IS SP



結果 2

通信品質の影響

実験中に伝送エラー：無

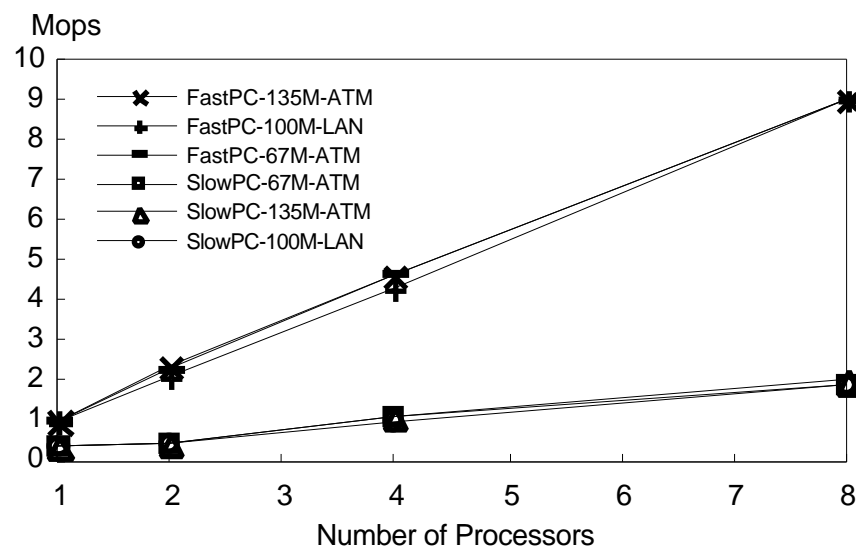
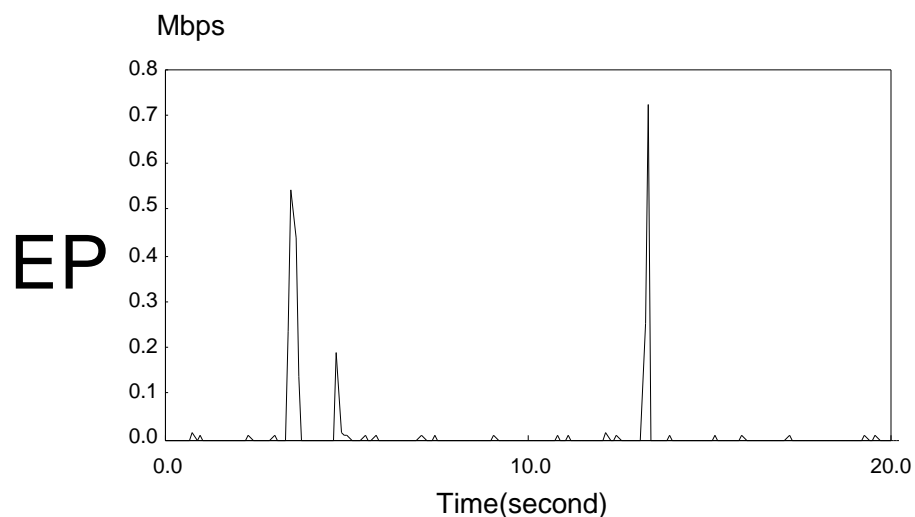
最小 0.8Mbps(EP)、最大 100Mbps(FT)

共用の試験：今回、無し

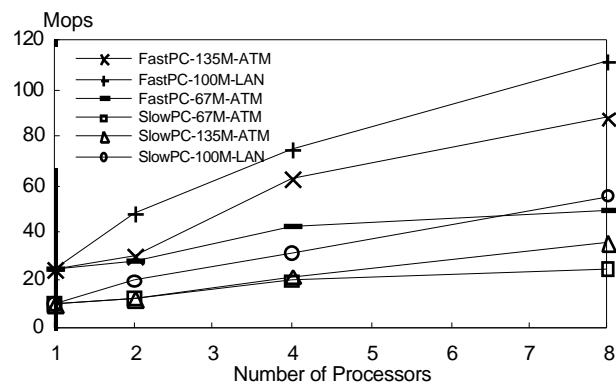
双方向 100Mbps 位の通信までは問題無し。

結果 3

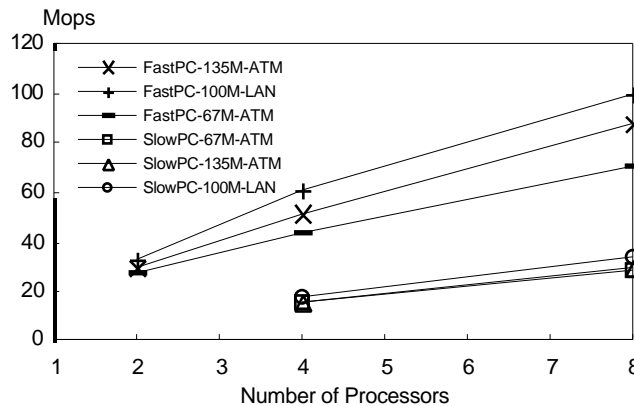
スケーラビリティ 傾きは、8 台で低下
今後 WDM など高帯域ネットワークを利用して
も、10-20 台程度のクラスタしか有効でない。
EP は良好にスケール（広域向き）



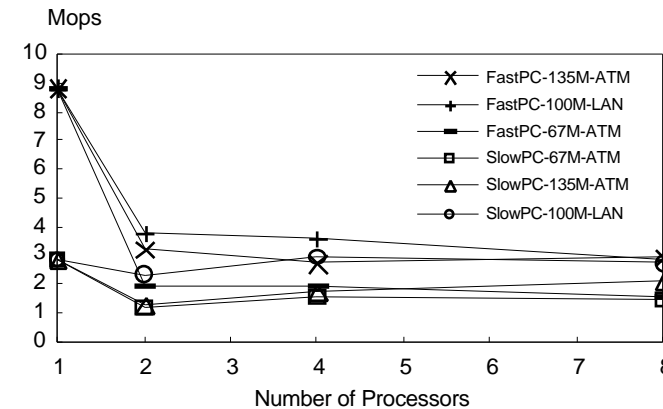
スケーラビリティ (67M-ATM、8 台、双方向)



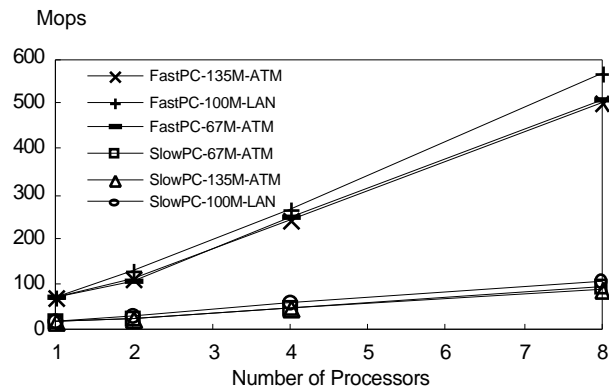
CG ()



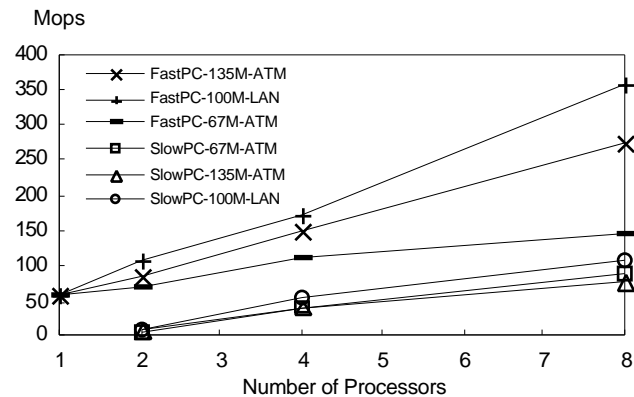
FT ()



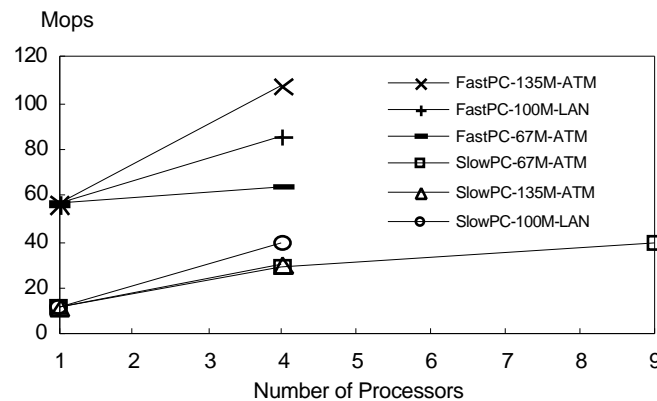
IS (x)



LU ()



MG ()



SP (x)

結果 4

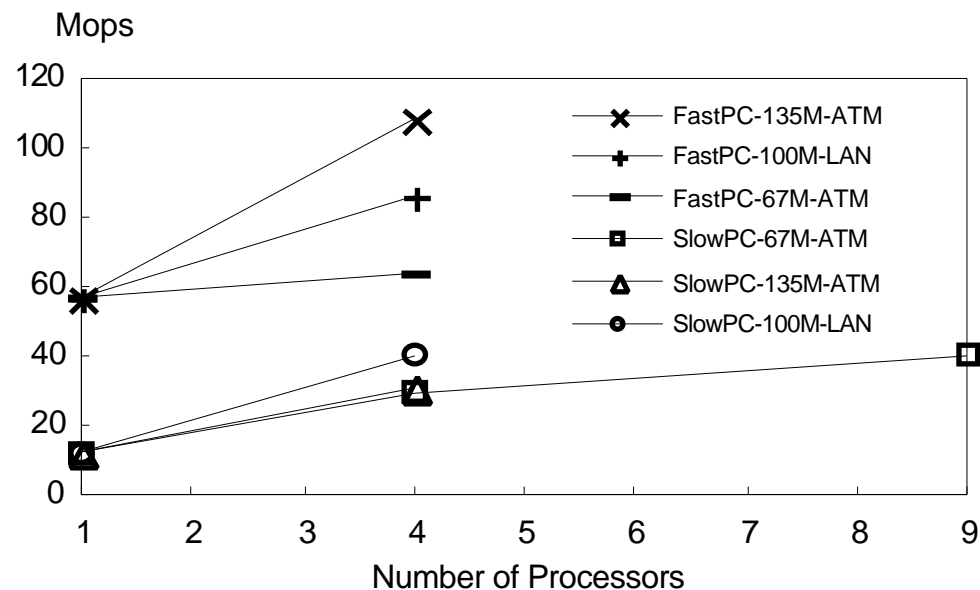
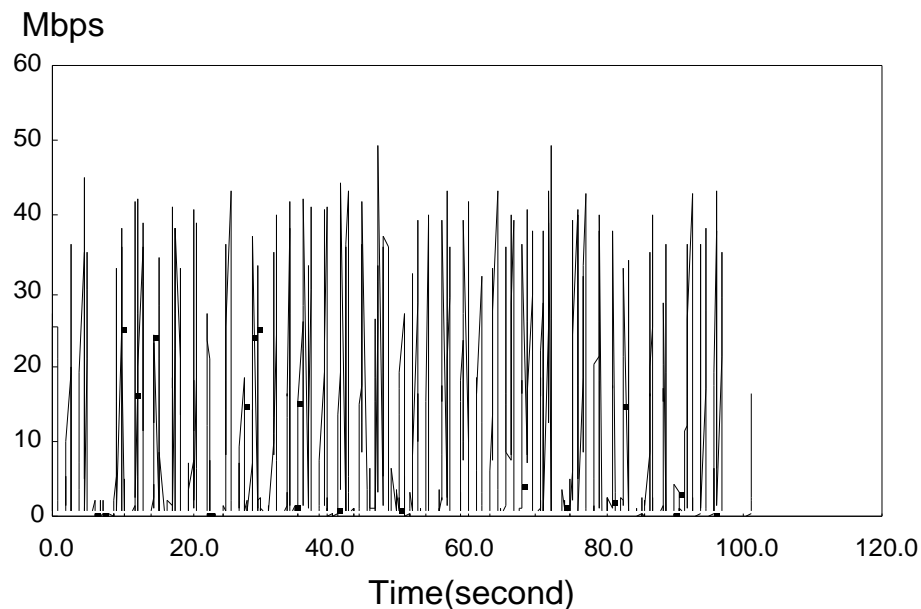
プロセッサの違いによる帯域、遅延の影響

Fast-PC : 帯域よりも遅延の影響 : 大
(100M-LAN)、(135M-ATM)、(67M-ATM)の順

Slow-PC: ネットワークの違いの影響 : 小

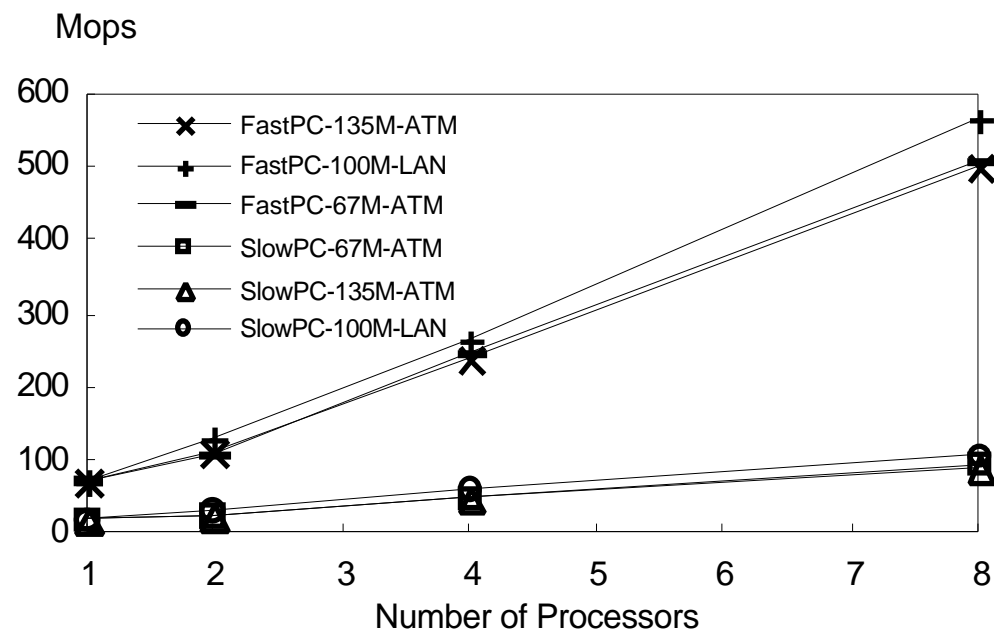
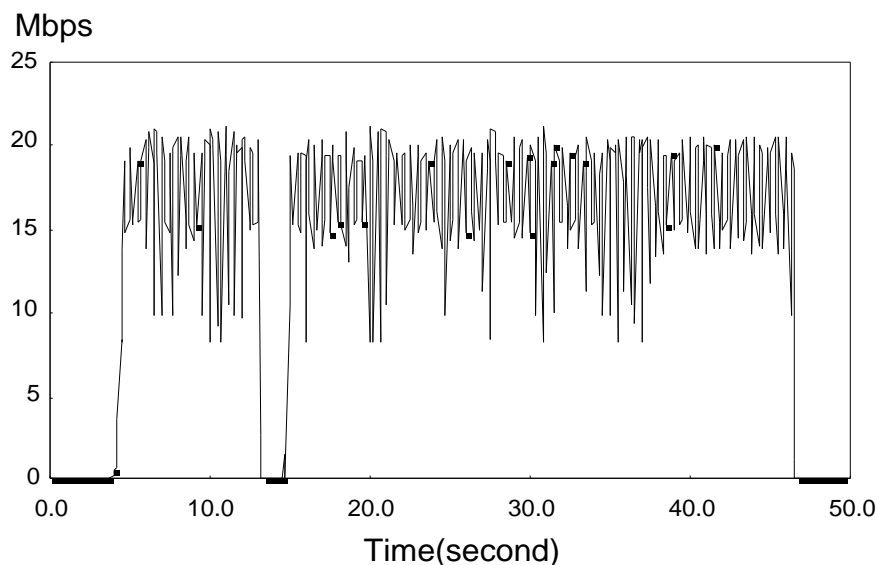
結果 5—1 (SP)

SP : (遅延よりも) 帯域の影響 : 大
(135M-ATM)、(100M-LAN)、(67M-ATM)の順



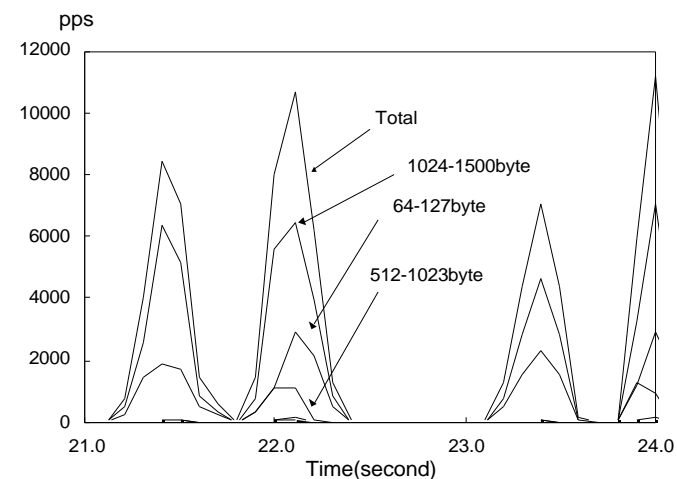
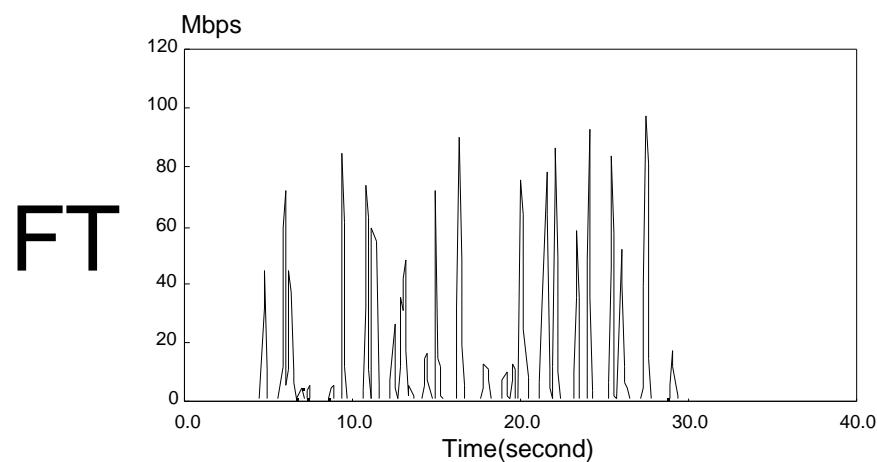
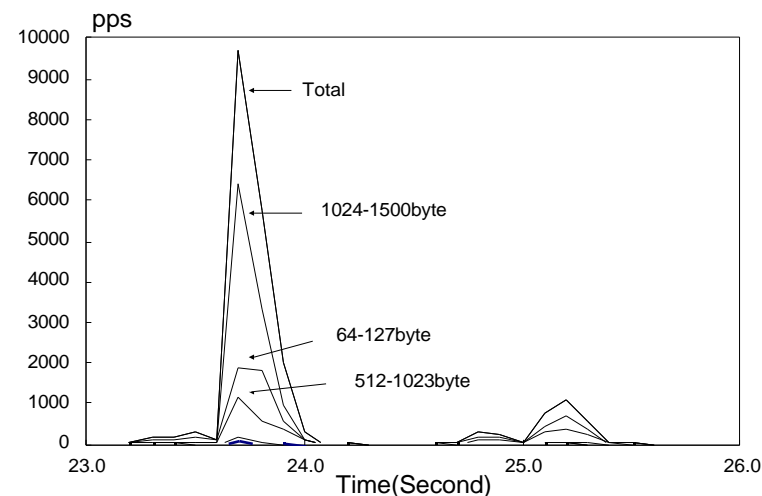
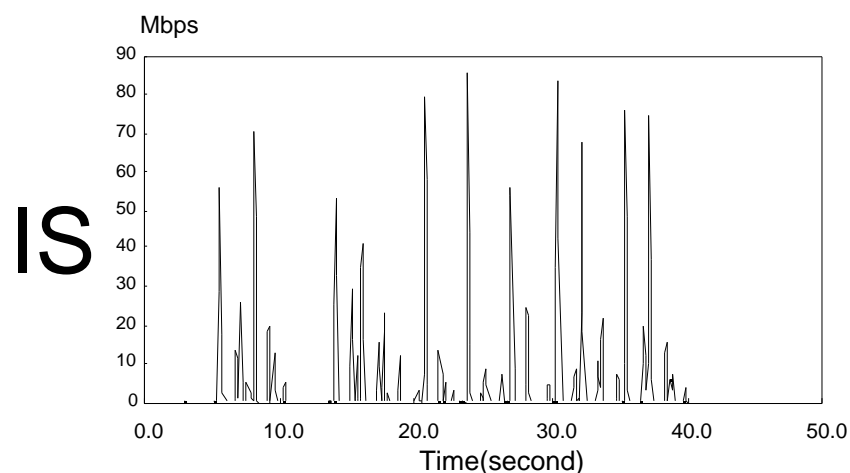
結果 5—2 (LU)

LU : 帯域の差 : 他に比較して少ない
(100M-LAN)、(135M-ATM)、(67M-ATM)の差が
小さい

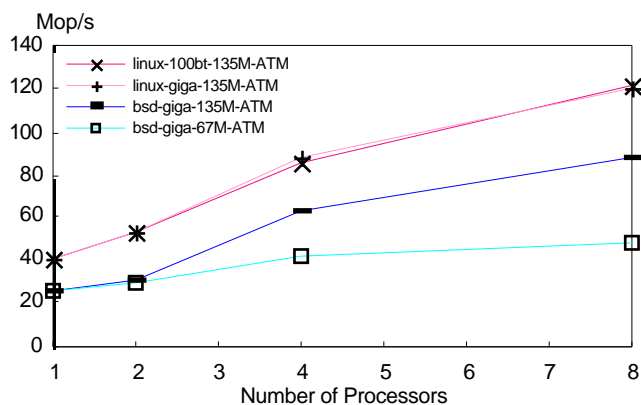


結果 5 — 3 (IS)

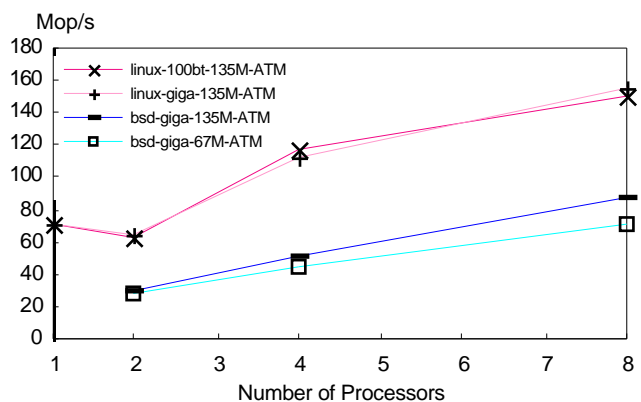
IS, FT :IS の急激な転送、通信の分散、性能の違い \ (Linux > BSD / OS)



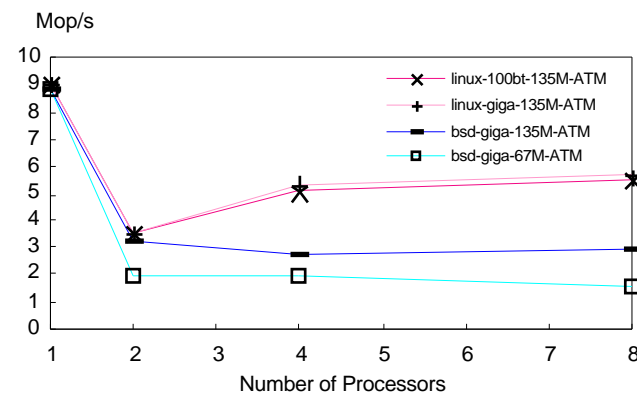
結果 6 プロトコルスタックの影響



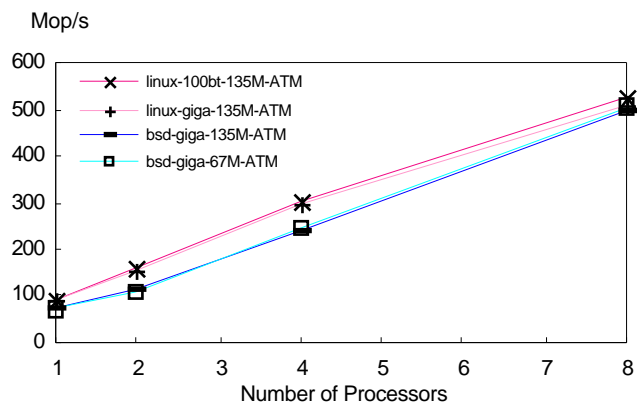
CG



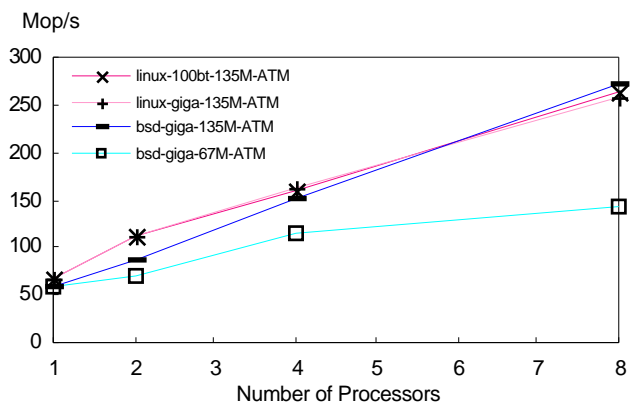
FT



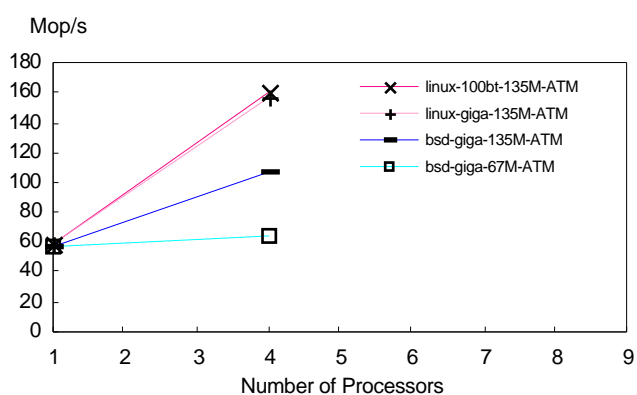
IS



LU



MG



SP

並列処理研究者の反応@SWoPP2000

興味深い実験

さらに調べたい項目

もっと細かいトラフィックモニタリング

UDP (要 acknowledging) での結果

GbE など非 ATM ネットワークでの結果

伝送遅延の影響 (伝送距離)

ルータ等の影響 (パケットロス、ジッタ)

他トラフィックの影響

並列アルゴリズム、実装の改善につながる

今後

トラフィック計測ツールの強化

ネットワーク共用状態での実験

並列計算への影響、他利用者への影響

ルータ接続ネットワークでの実験

バッファ溢れによるパケット廃棄やジッタの影響

ソフト側のチューニング

TCP バッファを増やす、NPB の通信処理を工夫