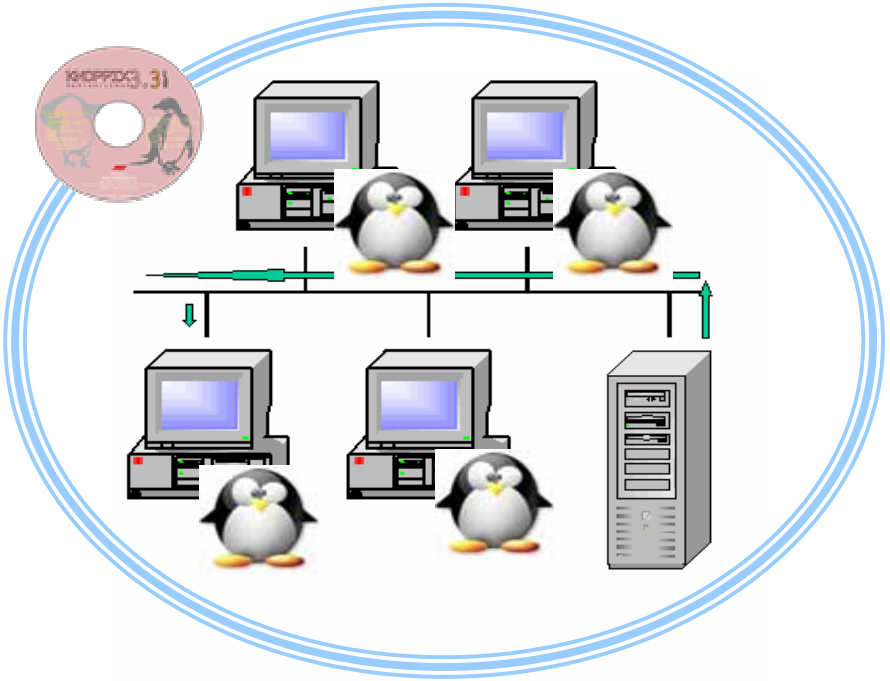
KNOPPIX クラスタ 情報交換会
事例紹介1

広島県立広島国泰寺高校
科学部物理班

3年 箱崎 亮太
2年 浜田 浩二
平野 敬純
岡本 潤一

CDだけでできるMPI並列処理のためのPCクラスタシステムの開発

～普通のパソコンがあっという間にスパコンに！？～



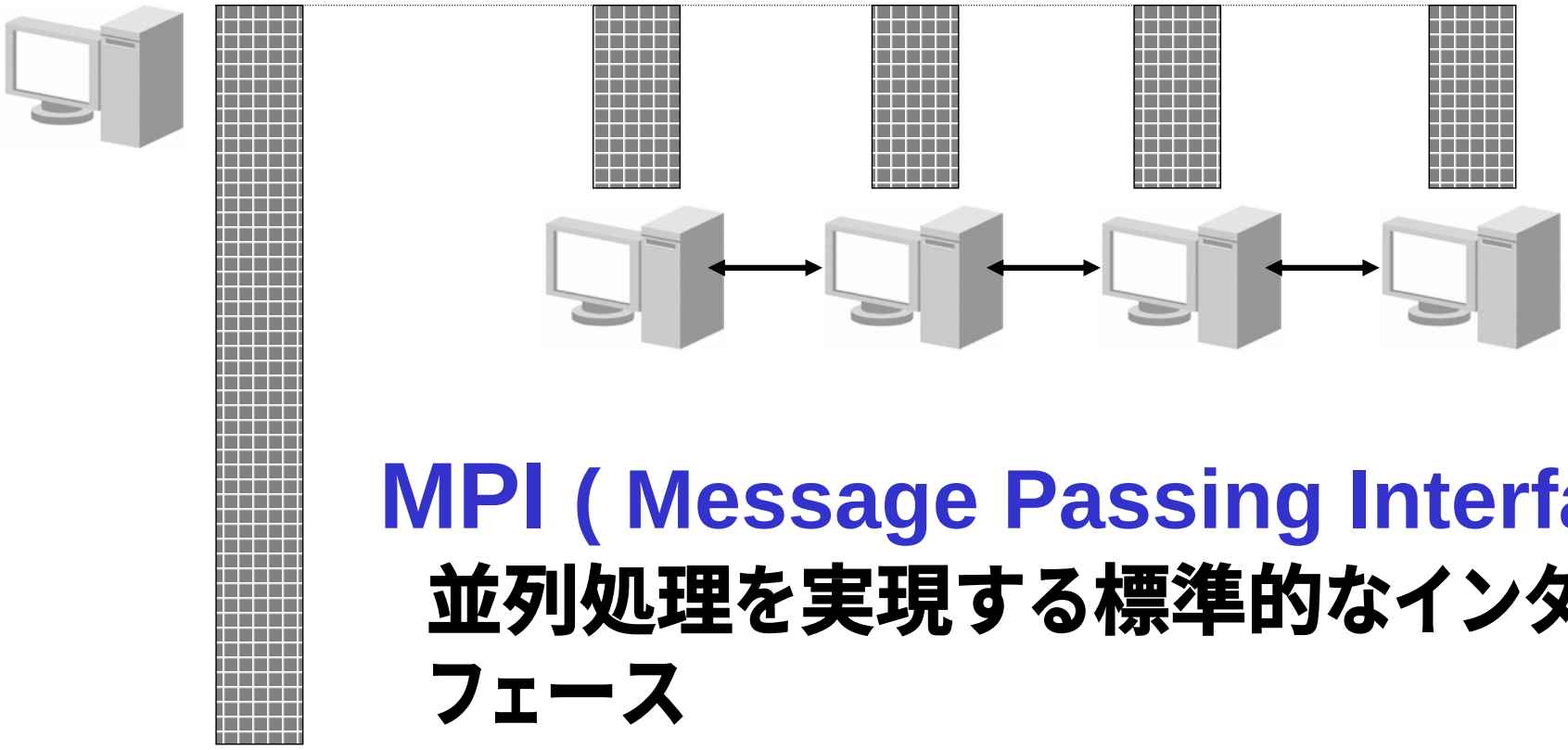
広島県立広島国泰寺高等学校
科学部物理班

3年　箱崎　亮太
3年　作田　晴朗

はじめに

並列処理とは

複数のPCで処理することで計算速度が向上



MPI (Message Passing Interface)

並列処理を実現する標準的なインターフェース

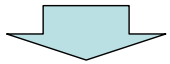
目的と問題点

目的

高校生でも利用できる高性能なMPI並列処理のための環境をつくる

問題点

- 専門的知識と多大な労力が必要
- 専用のハードウェア環境が必要



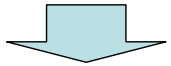
気軽に使えない！

方法と結果

方法

ハードウェア：　学校の情報教室にあるPCを利用

ソフトウェア：　フリーウェアを組み合わせて利用



オリジナルのPCクラスタシステムの開発

結果

誰でも簡単に

どこでも自由に

スパコン並の性能を実現！

システムの特徴

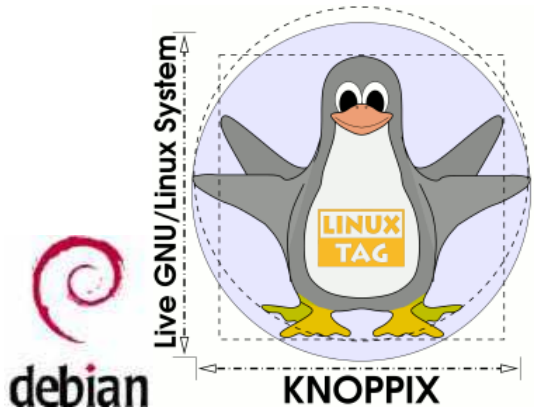
	国泰寺高校	通常のＰＣクラスタ
利　用　形　態	一時的	恒常的
起　動　方　法	ＣＤ	HD・ネットワーク
設　　定	簡単	難しい
可　搬　性	◎	×
規　　模	小規模向き	大規模も可
コ　　ス　　ト	ＣＤメディア代	数百万円

システムの概要 (1)

KNOPPIX をカスタマイズ

KNOPPIX：CDから起動する Linux

debian パッケージを利用可



MPI の実装として LAM を使用

	LAM	MPICH
移植性	○	◎
利用者数	○	◎
実行ファイル配布	ファイル転送機能	要 NFS,NIS

システムの概要 (2)

ホスト名の解決のために IP エイリアスを利用

DHCP **サーバー**：IP **アドレスの割り当ては動的**

CD-ROM：**ホスト名のテーブル は 固定**

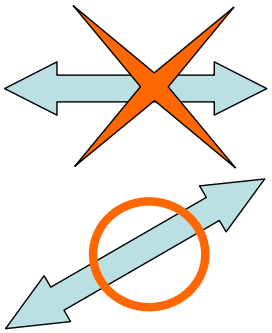
pcl01

DHCP**からの**IP**アドレス**

192.168.0.?

IP**エイリアス**

10.0.1.1



ホスト名のテーブル

pcl01	10.0.1.1
pcl02	10.0.1.2
……	
……	
……	

実行手順

```
$ pcl 1
$ lam 4
$ lamboot –v lamhosts
$ mpicc hogehoge.c
$ mpirun –np 4 –s n0 a.out
$ xmpi
$ lamclean
$ wipe –v lamhosts
```

名前を登録（全PCで実行）

lamhosts にPCを登録

LAMを起動

hogehoge.cをコンパイル

a.outを4台で実行

xmpi起動

終了処理

LAMを終了

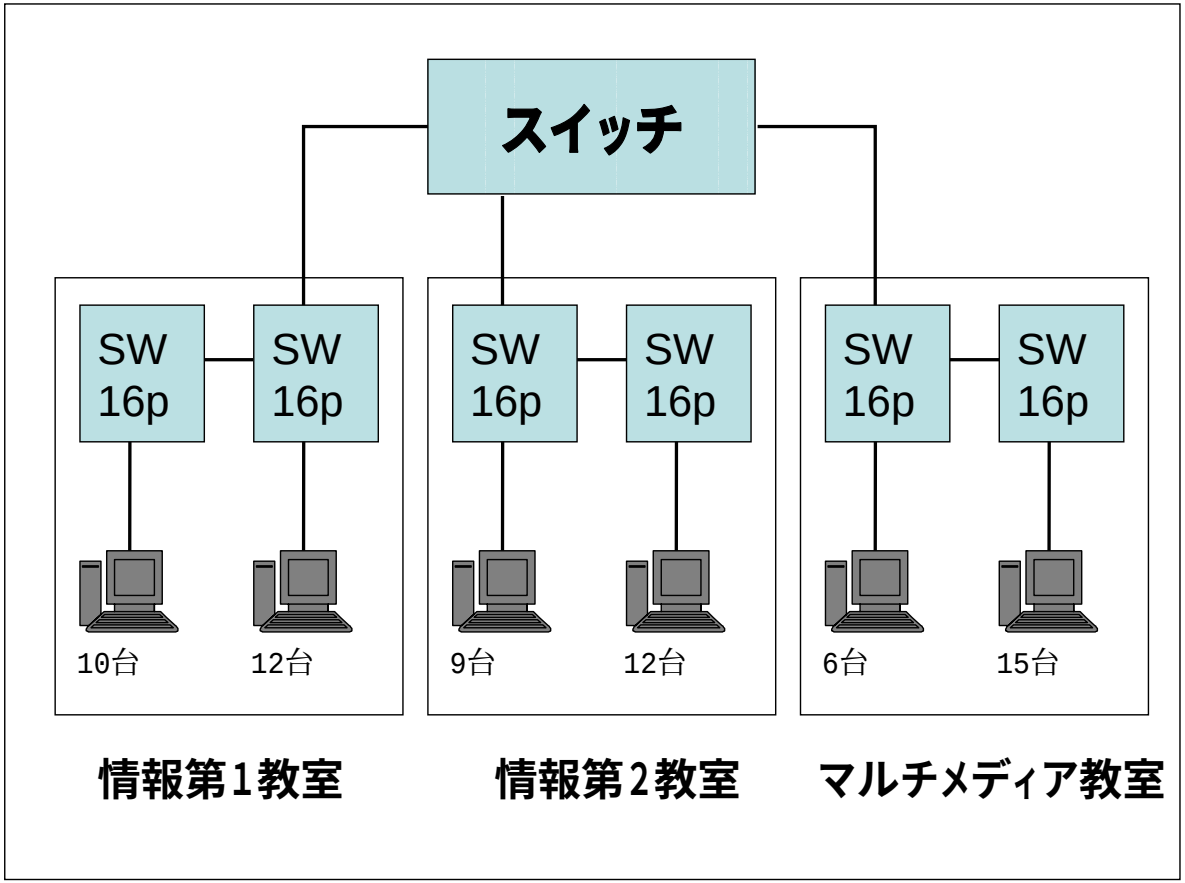
国泰寺高校のPC環境

- PC 64台
  Pentium(R)4　2.4GHz　512MB RAM 内蔵NIC
- スイッチ 7台

通信性能：100BASE-TX

＊専用のＰＣクラスタでは

Myrinet や GigaBASE

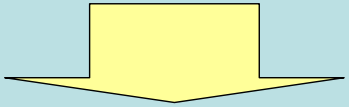


アムダールの法則

並列処理では並列化可能な部分と不可能な部分がある。

1台での処理時間　T_1

p台での処理時間　$T_p = r \cdot T_1/p + (1-r)T_1$



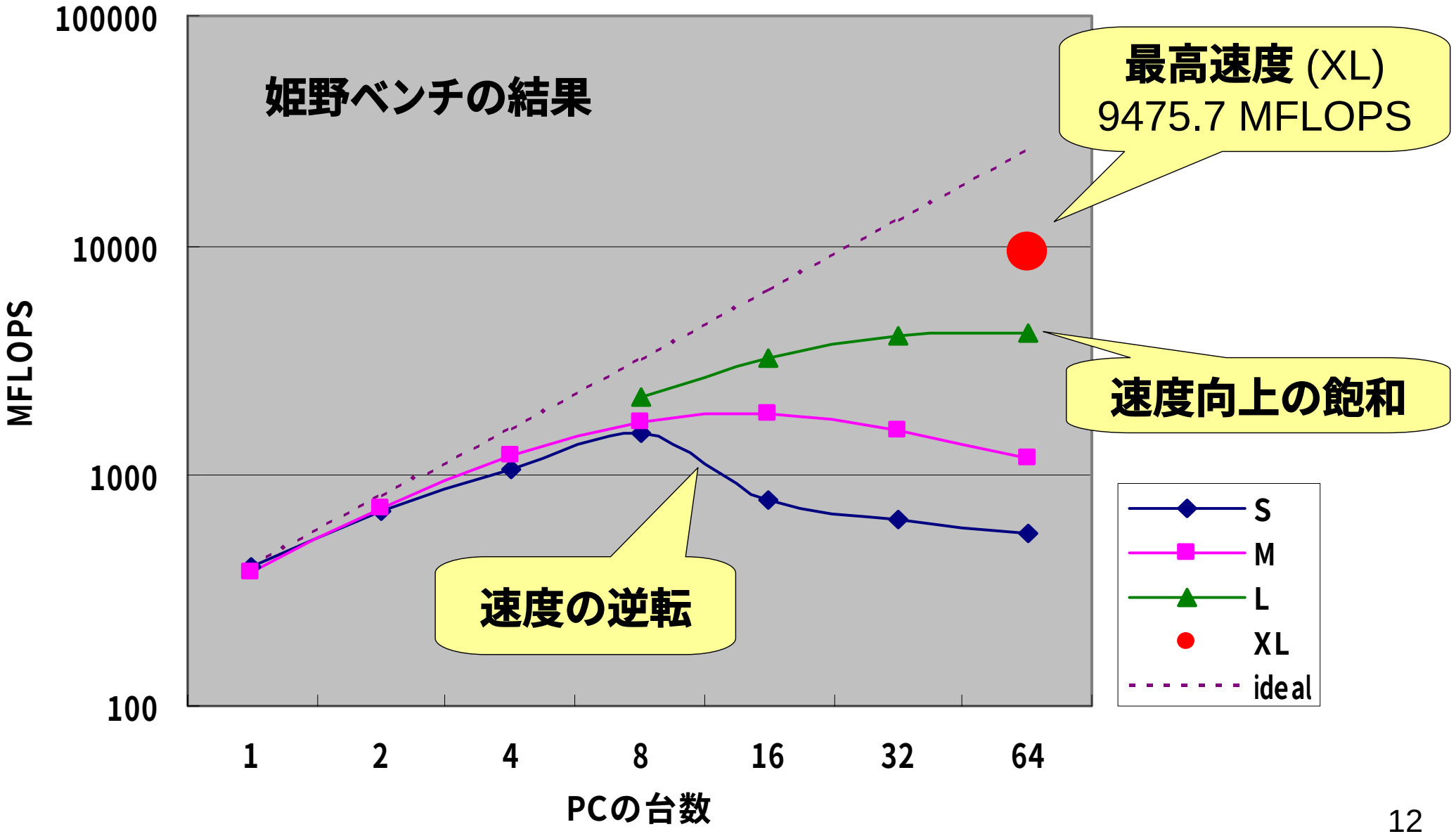
r:並列化率

速度向上　$S(p) = T_1/T_p = 1/\{r/p + (1-r)\}$

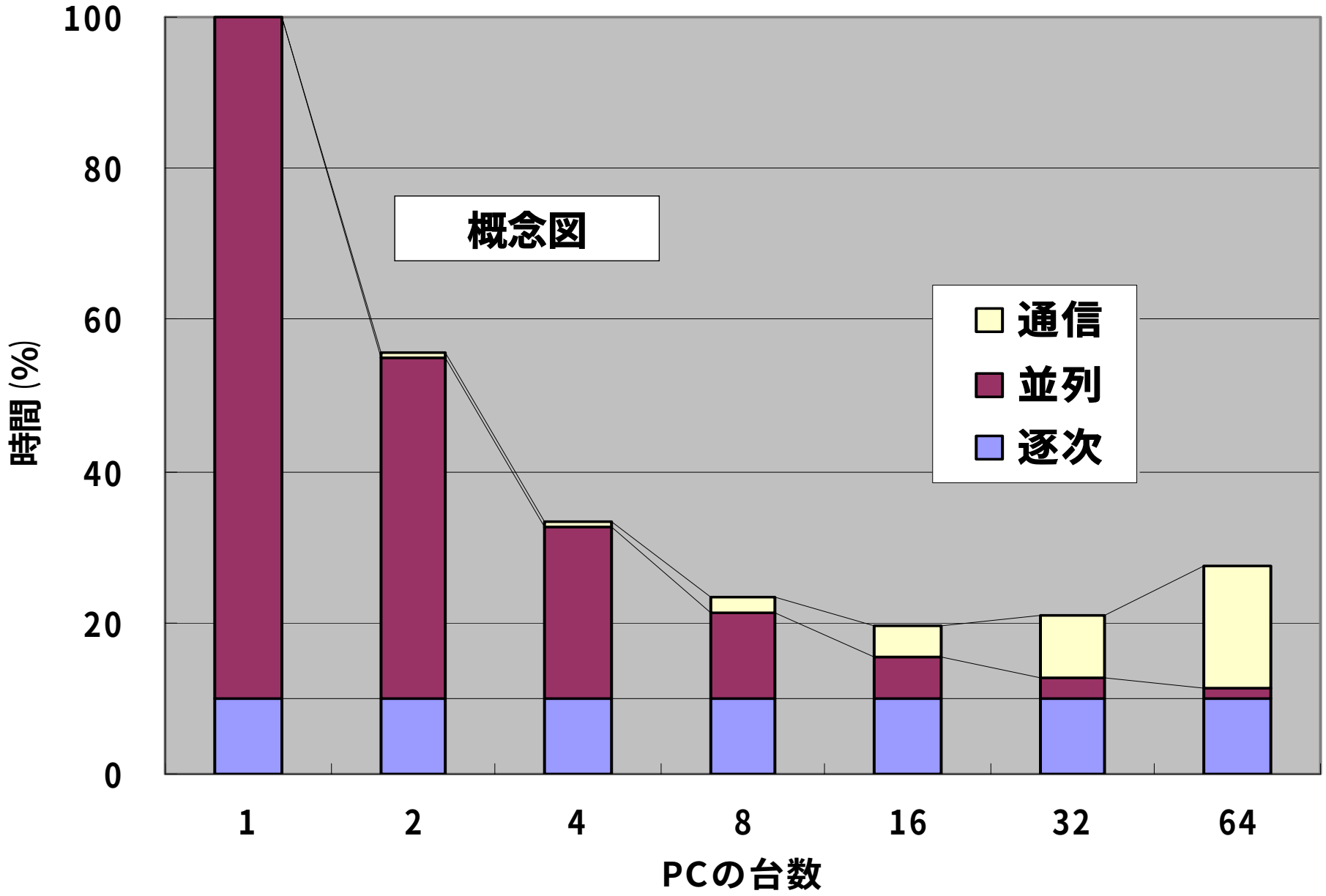
速度向上の最大値　$S(\infty)=1/(1-r)$

速度向上の限界

性能テスト


姫野ベンチの結果
最高速度 (XL)
9475.7 MFLOPS
速度向上の飽和
速度の逆転
100000
10000
1000
100
MFLOPS
1
2
4
8
16
32
64
PCの台数
S
M
L
XL
ideal

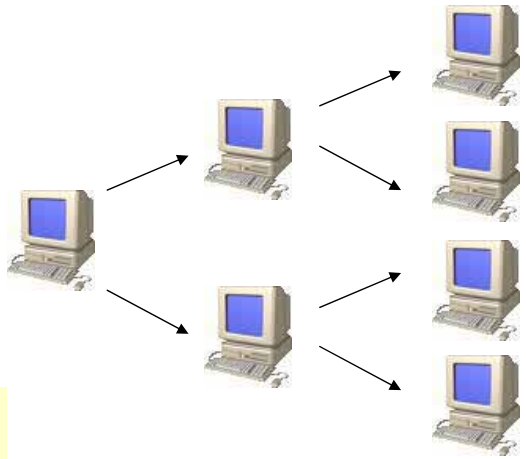
逆転現象の起こる原因



通信コストの理論化

アムダールの法則＋通信コストを考慮

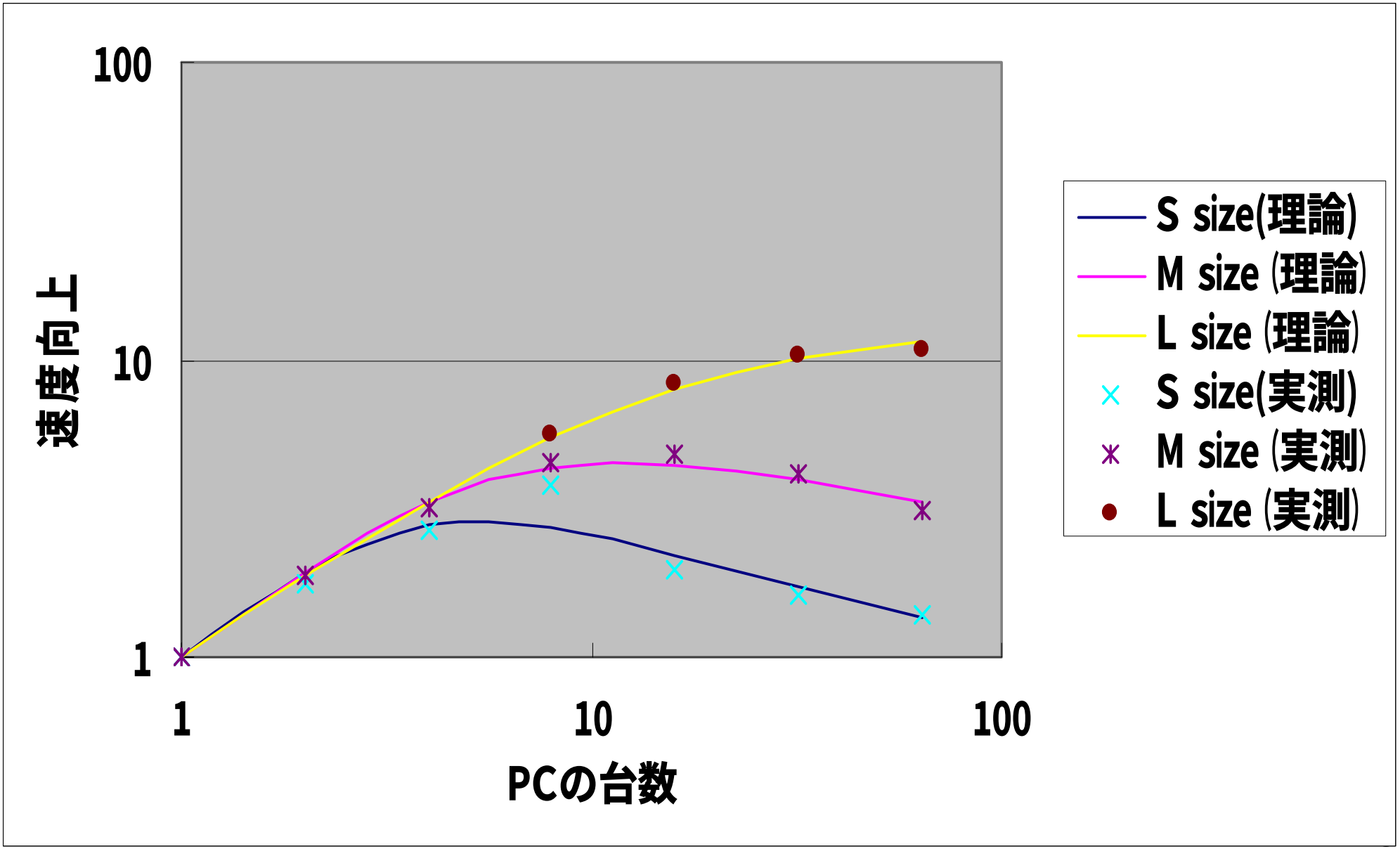
$S(p)=1/\{r \cdot 1/p+(1-r)+C \cdot \log_2 p\}$



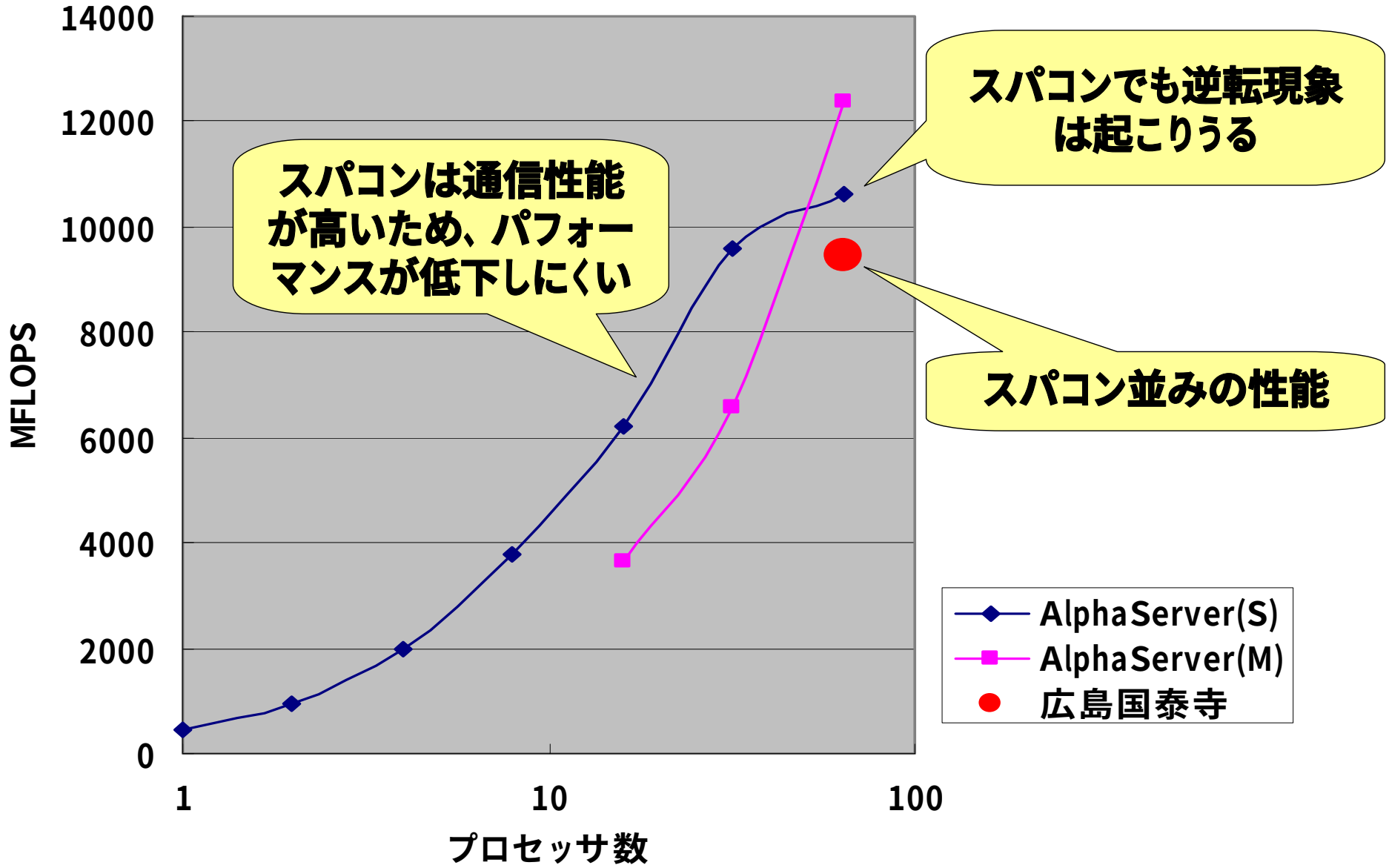
フィッティング

	r	-c
S size	1.304683	0.169900
M size	1.099709	0.063845
L size	0.941276	0.002196

フィッティング

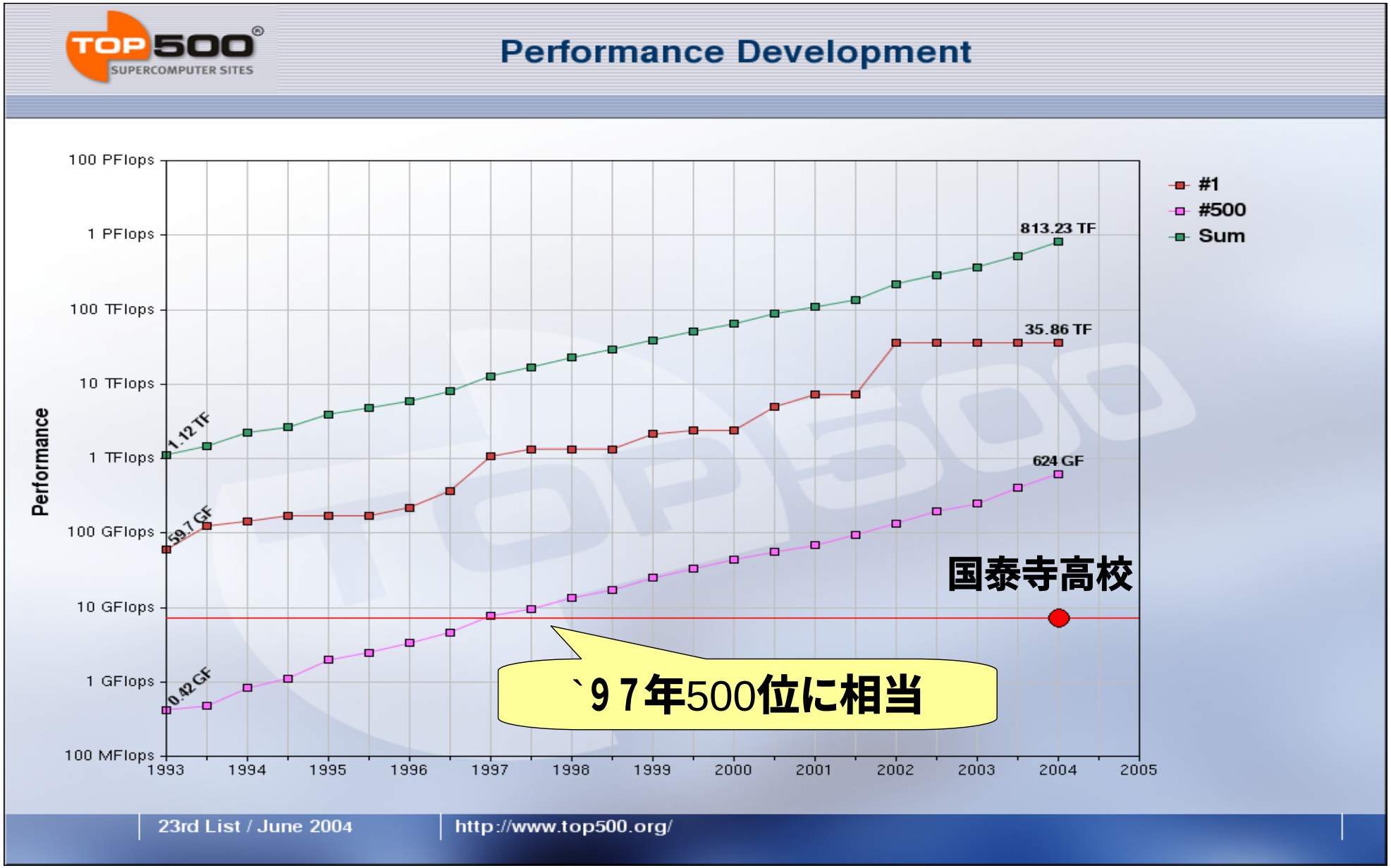

100
10
1
速度向上
1
10
100
PCの台数
S size(理論)
M size (理論)
L size (理論)
S size(実測)
M size (実測)
L size (実測)

スパコンとの比較



※データは理化学研究所姫野ベンチウェブサイトhttp://accc.riken.jp/HPC/HimenoBMT/より

世界ランキングとの比較



まとめと課題

まとめ

CDメディア代のみで、**どこでも簡単に** スパコン並みの性能のPCクラスタを構築できるシステムを完成

ベンチマークの結果から速度向上に逆転現象を発見
並列処理には **最適な並列台数で** の実行が必要

今後の課題

ネットワークから起動するシステムの開発
通信性能の解析と理論化

CD1枚で起動するPCクラスタの開発

広島県立広島国泰寺高等学校
科学部物理班
2年　平野　敬純

現在のシステム

全ノードがCDから起動

↓

64台起動時にはCDが64枚必要

各PCを一台づつ起動して

PCLコマンドで名前をつける

開発中のシステム

- **マスターノード**
  CDから起動

- **スレーブノード**
  PXEを利用したネットワーク起動
  マスターノードからWake on LANでスレーブノードを起動
  CDはマスター用1枚のみ

  ※　全ての操作はマスターノードのみで行う

PXEとは

- PXE：Intelの開発したネットワークブート規格

DHCPサーバ
TFTPサーバ
NFSサーバ

組み合わせて利用

設定が面倒

KNOPPIXでは
Terminal-Serverを利用すれば簡単に実行可能

未解決の問題

- 校内のDHCPサーバとの競合
  距離が近いためか、うまく動いているように見えるが・・・
- スレーブノード起動時
  マスターノードにアクセスが集中
  NFSマウントに失敗することがある
  (2回しかトライしていない？)

MPI通信速度の実験

広島県立広島国泰寺高等学校
科学部物理班
2年　浜田　浩二

目的

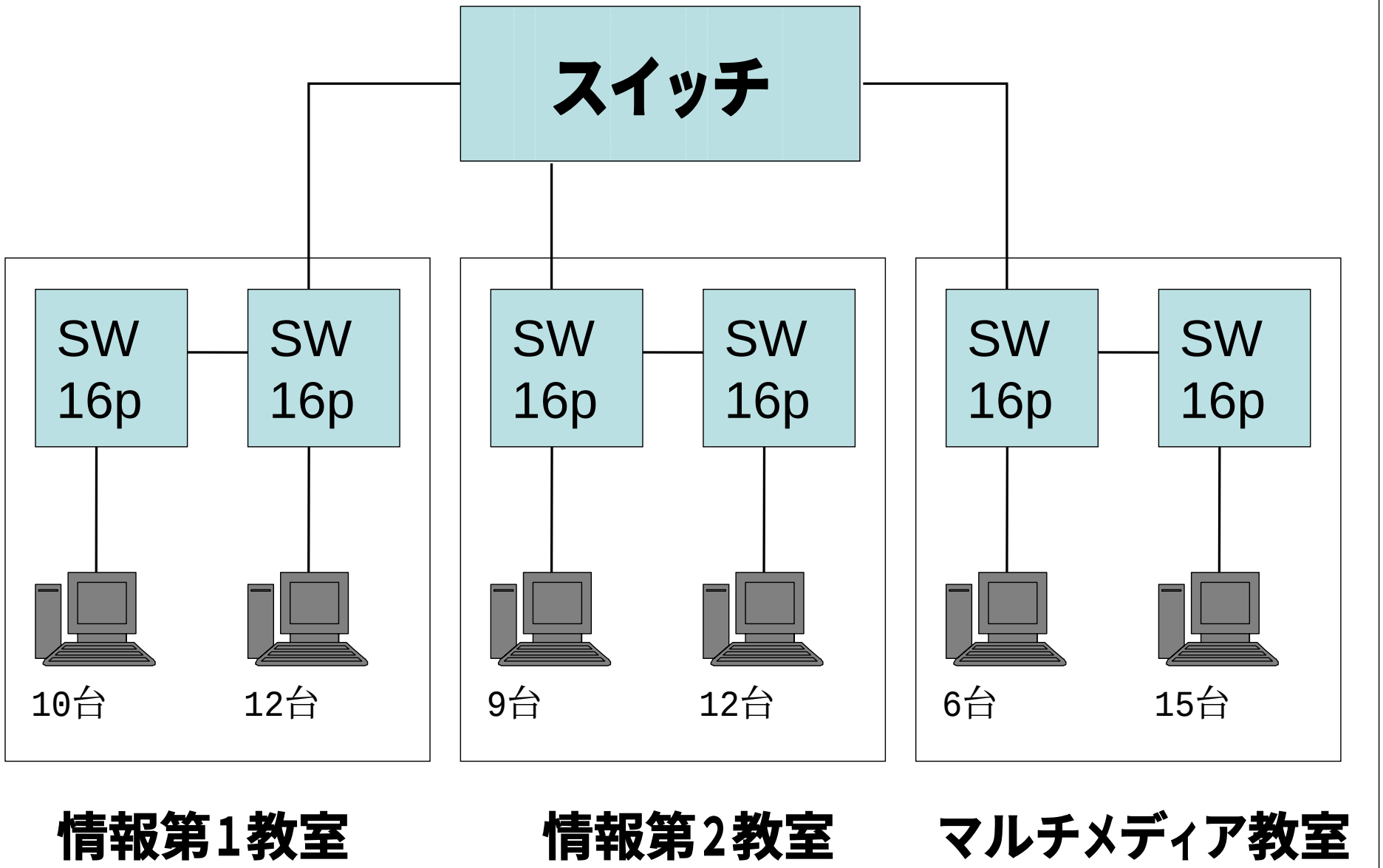
PCクラスタのノード間でMPI通信時間を測定し、通信性能を調べる。

仮説

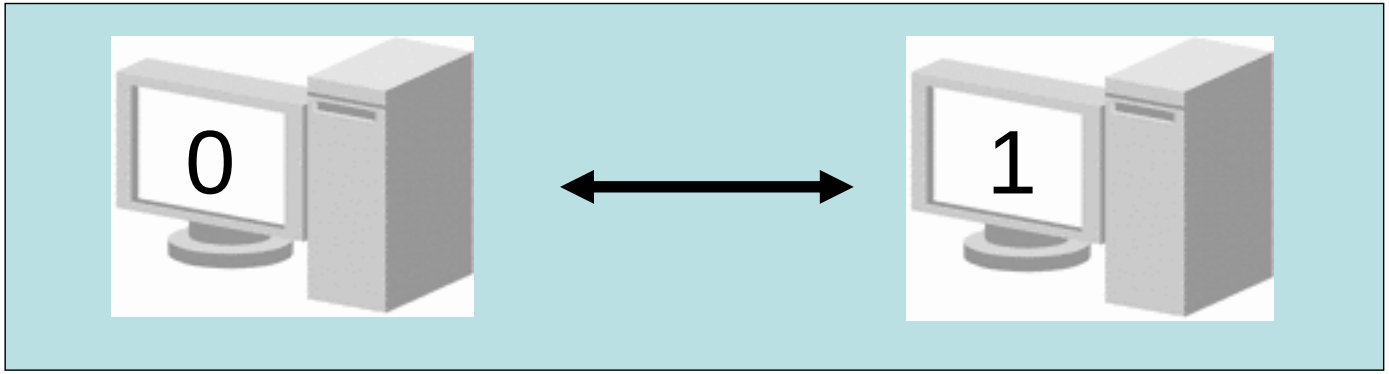
ノード間に存在するスイッチが、校内のPCクラスタの性能のボトルネックとなっている

方法

ノード間でデータ量を変えて、ＭＰＩ通信を行い、時間を計測する。


スイッチ
SW
16p
SW
16p
SW
16p
SW
16p
SW
16p
SW
16p
10台
12台
9台
12台
6台
15台
情報第1教室
情報第2教室
マルチメディア教室

ピンポンベンチマーク



MPI_Barrier**バリア同期**

時間計測 t_1

MPI_Send → MPI_Recv

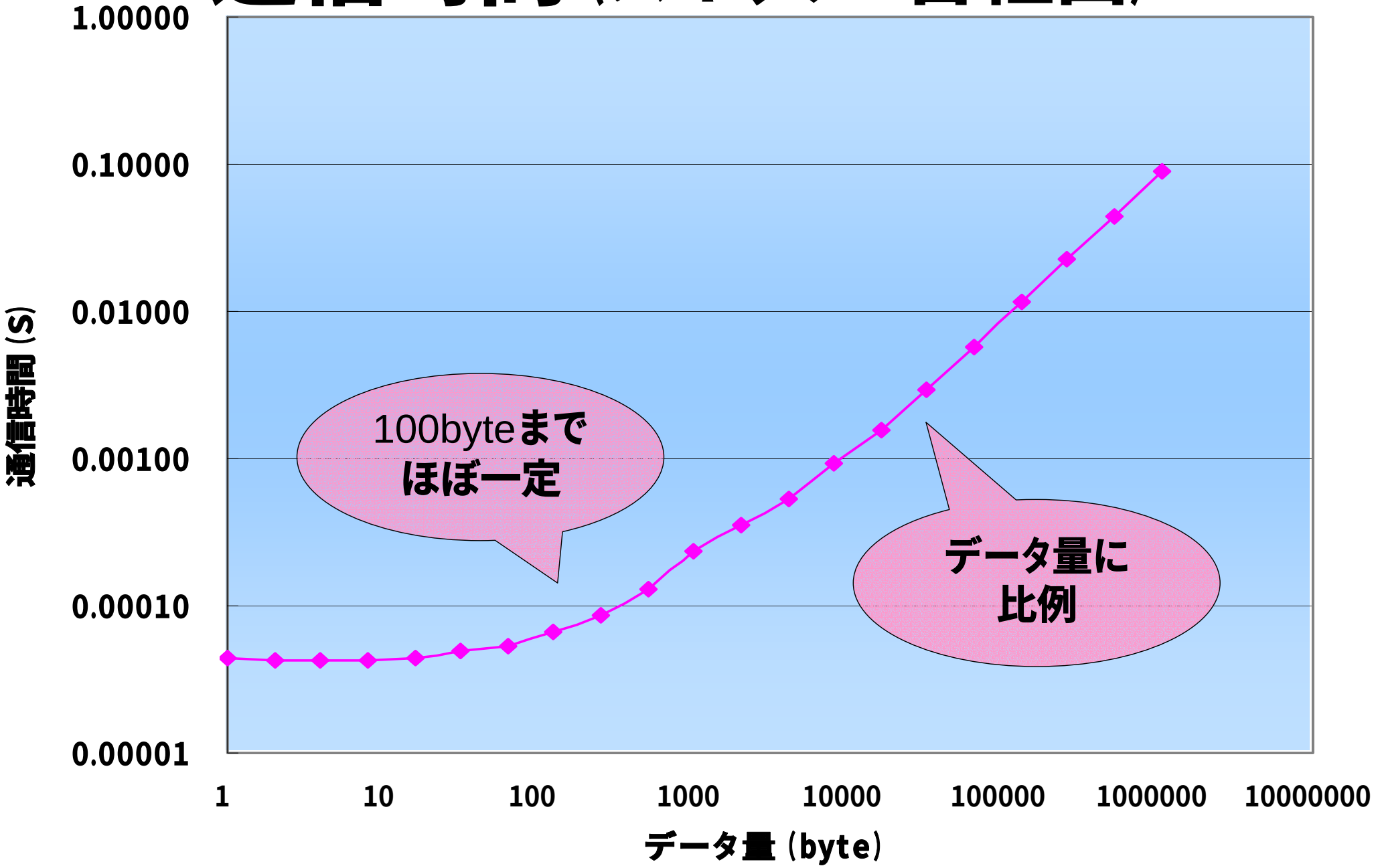
MPI_Recv ← MPI_Send

時間計測 t_2

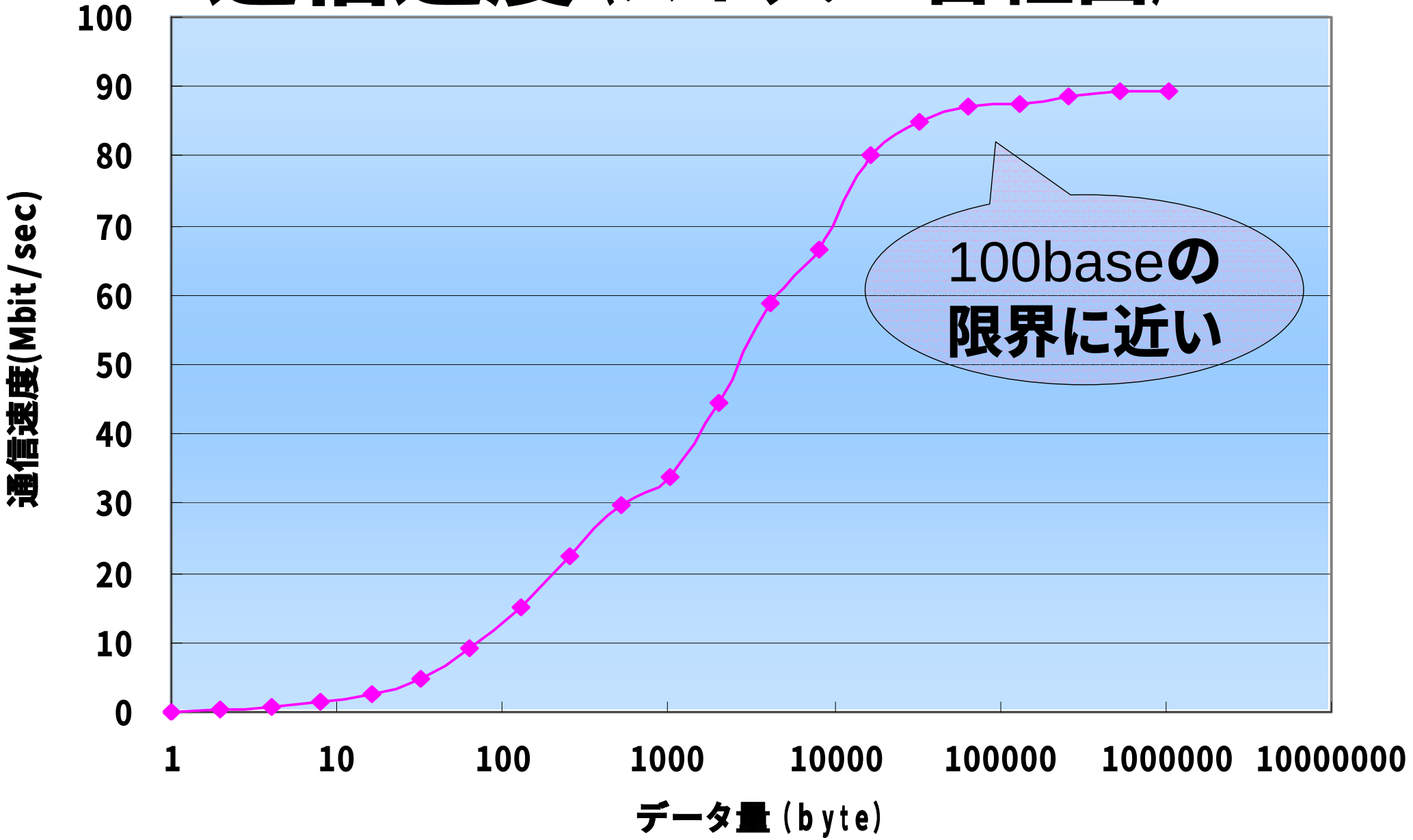
通信時間 $T = (t_1 - t_2) / 2$

１００回の平均を取る

通信時間（スイッチ1台経由）



通信速度（スイッチ1台経由）



今後の課題

- 経由するスイッチの台数を変える。
- 10BASEやGigaBASEに変える。

ご静聴ありがとうございました