レプリカ交換分子動力学シミュレータREMD Toolkitのグリッド上での実行

佐藤　仁 [1)]　　中田秀基 [1) 2)]

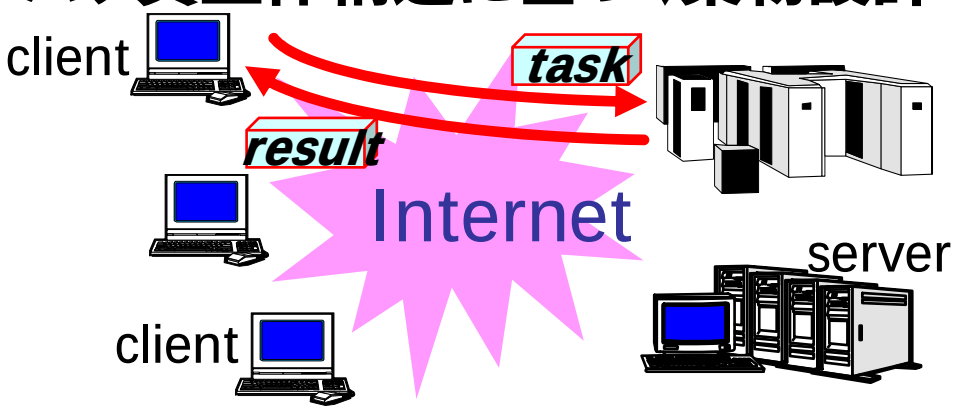
伊藤正勝 [2)]　　松岡　聡 [1) 3)]

1) 東京工業大学
2) 産業技術総合研究所
3) 国立情報学研究所

背景

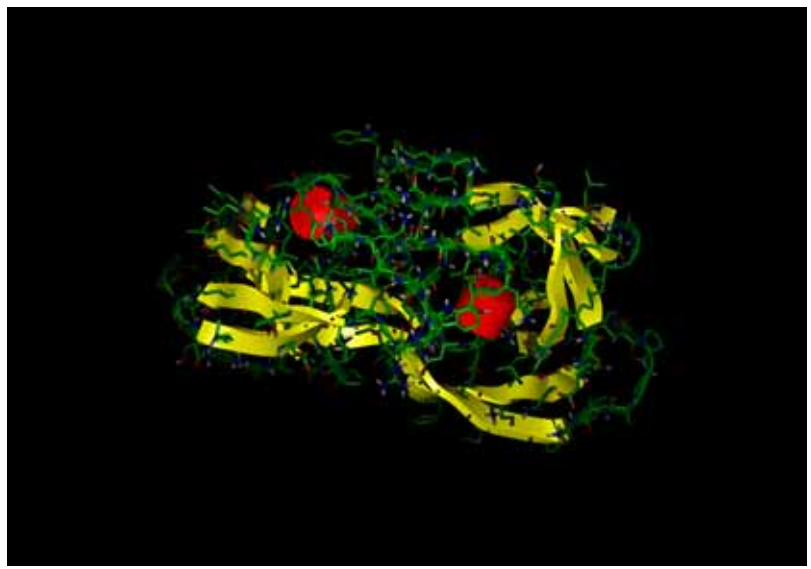
- バイオインフォマティクス等における大規模な科学技術計算の必要性の増加
- 大規模計算環境としてグリッド環境が実用的になりつつある

- グリッド環境向けの実アプリケーション開発
  - タンパク質立体構造に基づく薬物設計　等



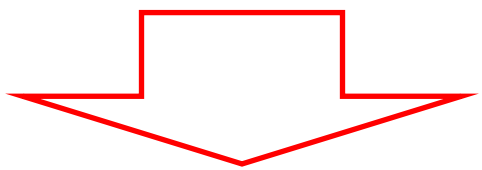
REMD Toolkit
Replica Exchange Molecular Dynamics

- 産総研で開発中
- グリッド環境に向けた実用アプリケーション
- レプリカ交換法を実装
- タンパク質の立体構造に基づく薬物設計
- 最終目標としてHIVタンパク質に作用する薬物の発見



レプリカ交換法　概要

- **タンパク質立体構造解析などで用いられる**
  Simulated Annealing (SA)**に似たアルゴリズム**
- **タンパク質構造のエネルギーを幅広く探索**
- **通信量が少なく粒度が大きいので並列計算向き**



- REMD Toolkit
  - **全ワーカに対する同期を多数含む**
  - **グリッド環境のような性能ヘテロ環境では低速なPEに律速される問題**

本研究の目的と成果

目的

- **レプリカ交換法のアルゴリズムを実装したREMD Toolkitのグリッド環境への対応**

成果

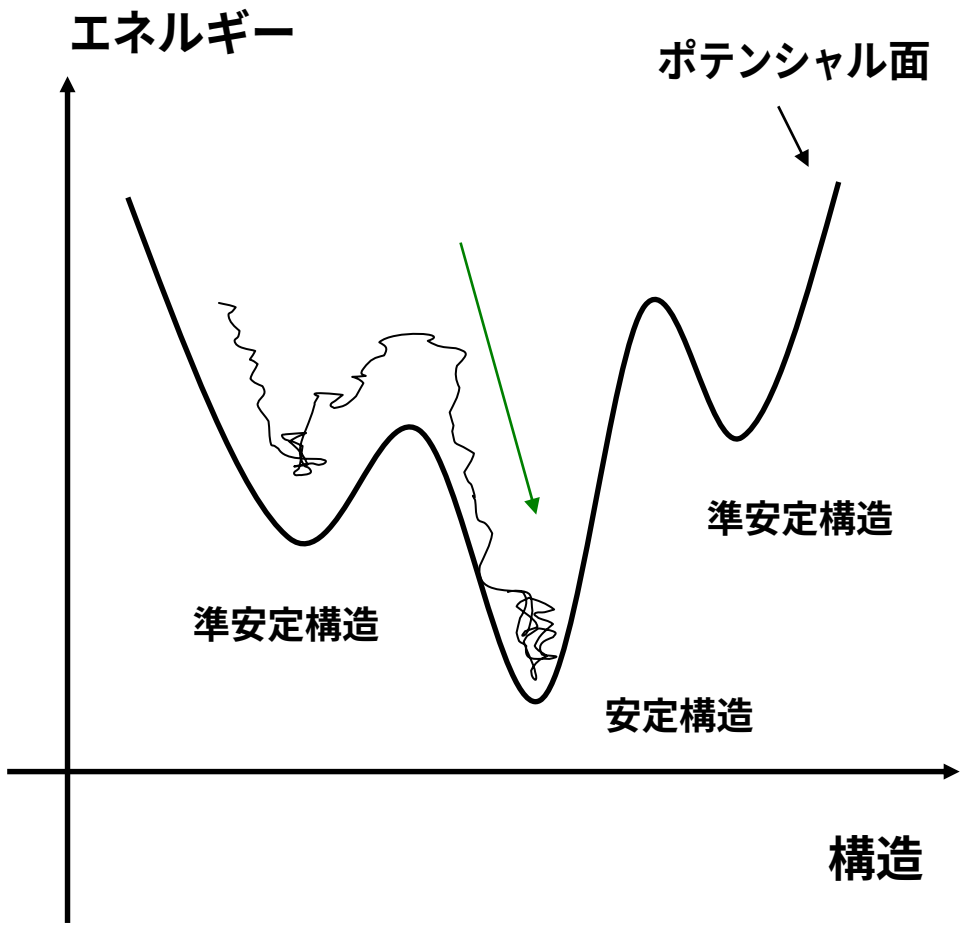
- **性能ヘテロ環境、広域ネットワーク環境での実験より、REMD Toolkitの性能ヘテロ環境への対応の有効性を確認**

レプリカ交換法 (1/3)

- タンパク質構造が保持するポテンシャル面には大小の起伏が多数存在
- タンパク質の構造が最安定構造に至らず準安定構造にトラップされやすい
- 温度を変化させ (エネルギーを与え) 構造を階層的に探索

(準安定構造からの脱出)

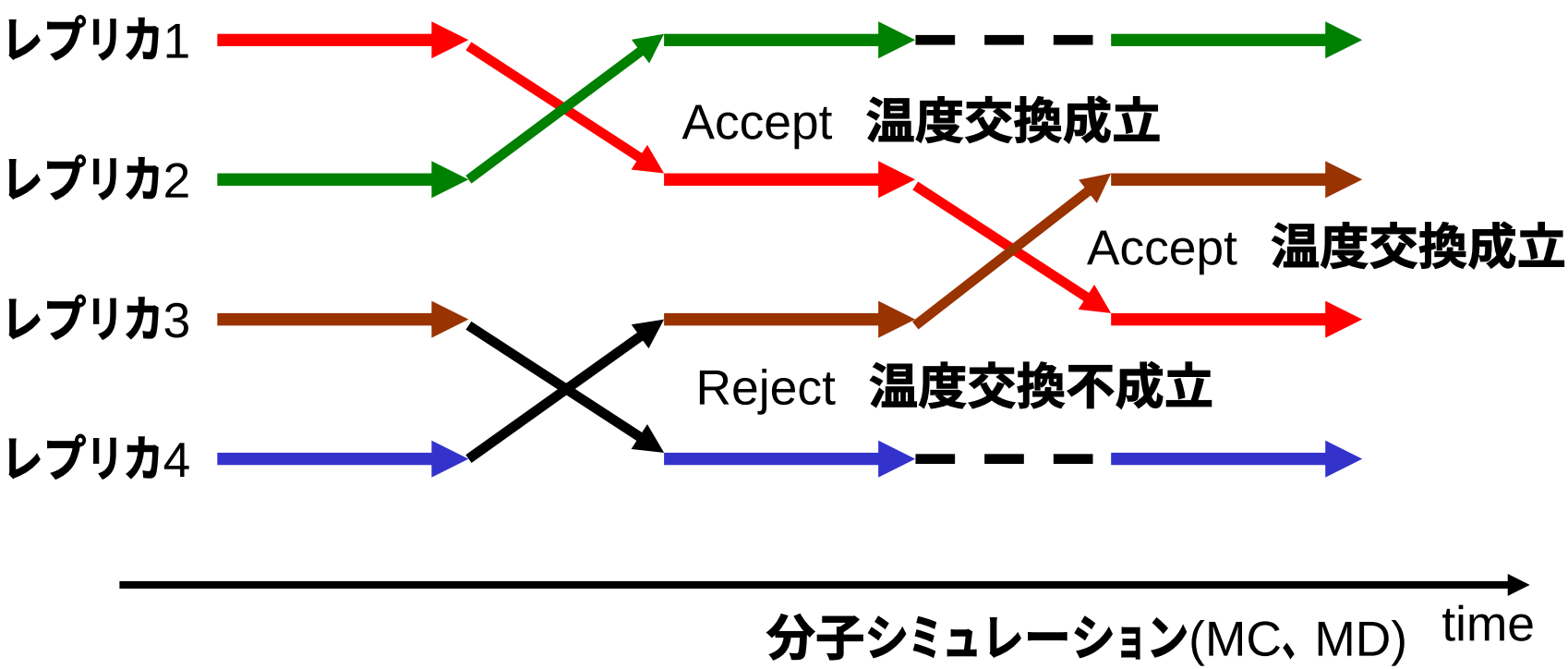
⇒ レプリカ交換法



レプリカ交換法 (2/3)

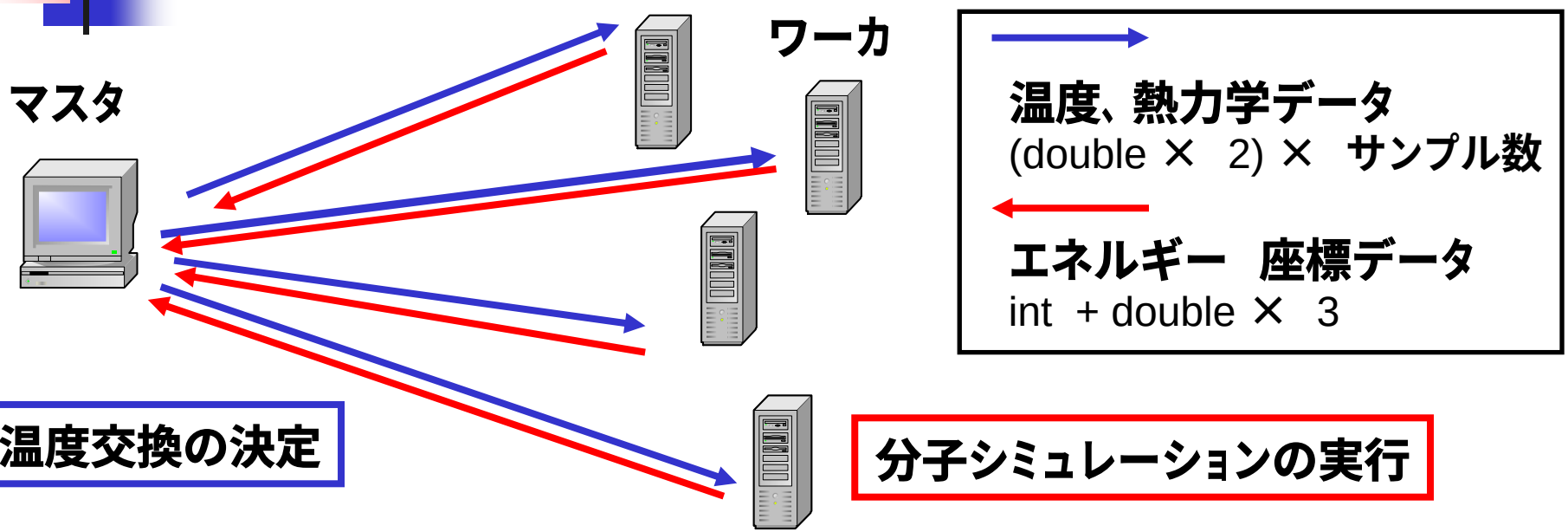
- 互いに相互作用しない異なる温度を持つタンパク質構造のコピー (レプリカ) を複数個用意
- レプリカ交換
  - 各レプリカで独立に分子シミュレーション(MC, MD)を実行
  - 途中で温度値が隣接した2つのレプリカの温度を交換

レプリカ交換法 (3/3)



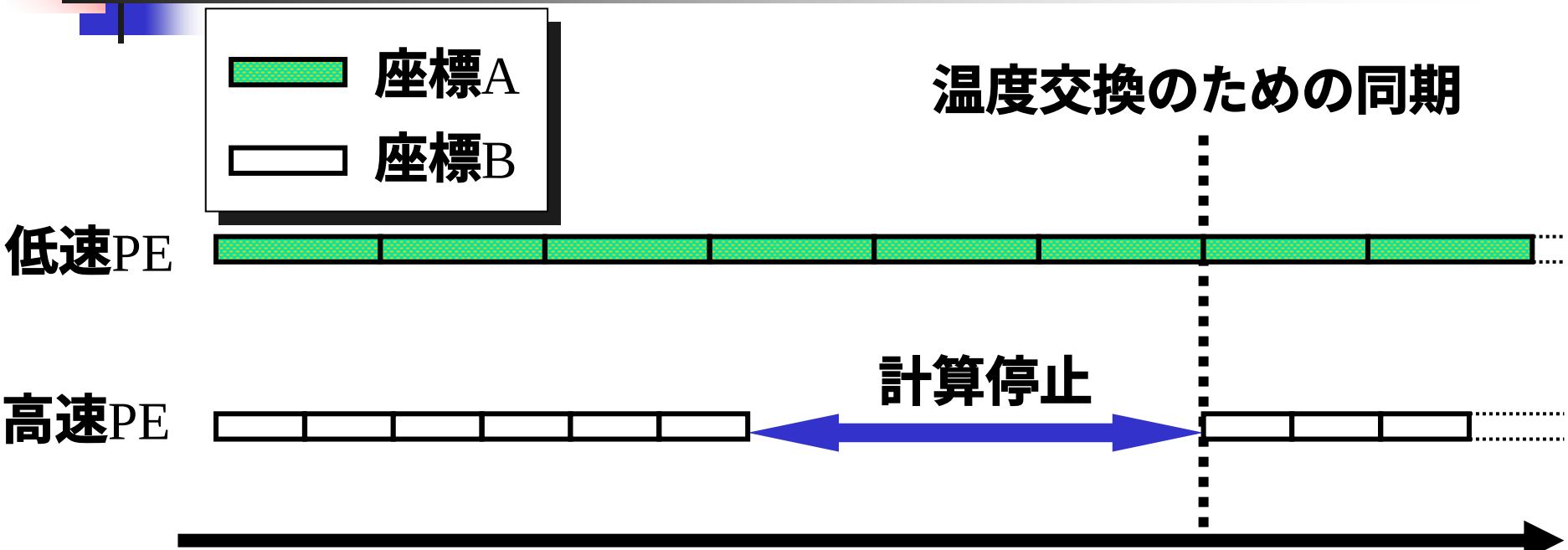
- 温度交換により、タンパク質の準安定構造を解消し、各レプリカで幅広く探索
- 通信量が少なく、粒度が大きいので、並列計算に向く
- 並列計算機上では各レプリカのシミュレーションは各プロセッサが担当

REMD Toolkitの動作
割り当て固定法



- ワーカは分子シミュレーション(MC、MD)を担当し、エネルギーと座標のデータをマスタに送信
- マスタでワーカの温度交換の決定
- マスタは、温度交換決定後、温度と熱力学データをワーカに送信
- 分子シミュレーションと温度交換を複数回繰り返す

割り当て固定法の問題点



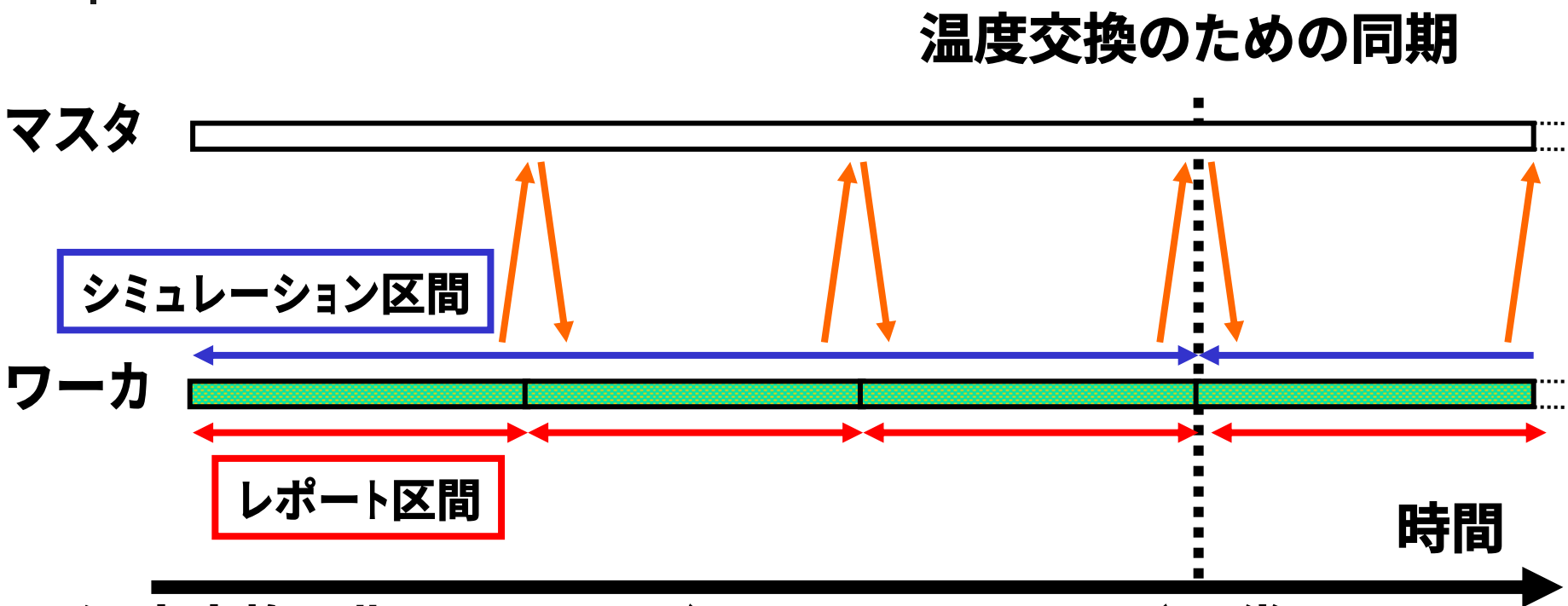
- 温度交換周期毎に全てのワーカが同期
- ワーカ間に性能差がある場合高速PEが低速PEの計算終了を待つ

⇨ 本質的にヘテロなグリッド環境では有効な運用が困難

性能ヘテロ環境対応
(割り当て交換法)

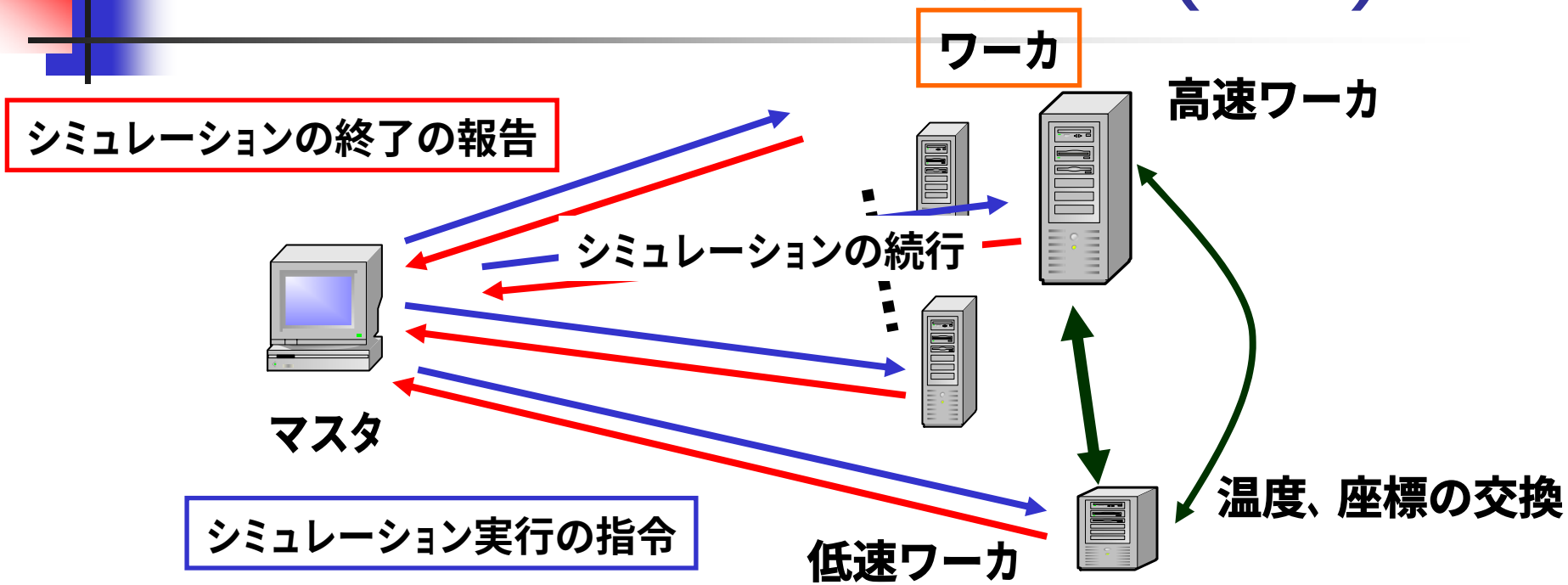
- 温度交換の際、全ワーカとのバリア同期の解消
  - 温度交換するワーカとの1対1の同期
- 各ワーカの進捗状況のモニタ
  - 進捗が遅れているワーカと進んでいるワーカに対して割り当て(温度と座標)を交換

割り当て交換法の動作 (1/5)



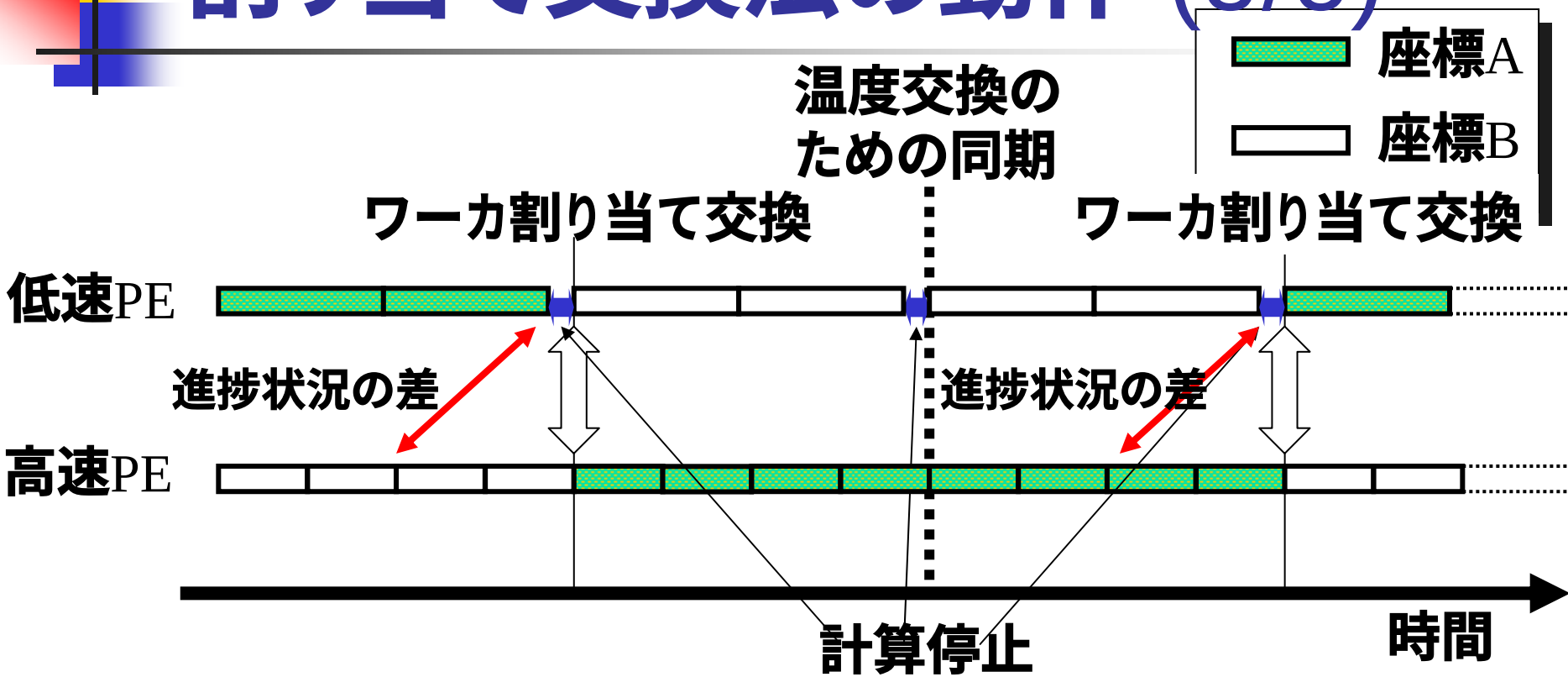
- 温度交換周期までの区間 (シミュレーション区間) を幾つかの区間に分割 (レポート区間)
- ワーカはレポート区間毎にマスタに進捗状況を報告(マスタ報告パラメタ)

割り当て交換法の動作 (2/5)



- ワーカはレポート区間のシミュレーション後、マスタに進捗状況を報告
- マスタはワーカの報告をモニタし、2つのワーカに対して指令
- 進捗状況に差がない場合、シミュレーションの続行
- 進捗状況に差がある場合、ワーカの割り当て(座標と温度)を交換

割り当て交換法の動作 (3/5)



- ワーカ間の性能差による計算停止時間を縮小
- 性能ヘテロな環境での効率の良いシミュレーション

割り当て交換法の動作 (4/5)

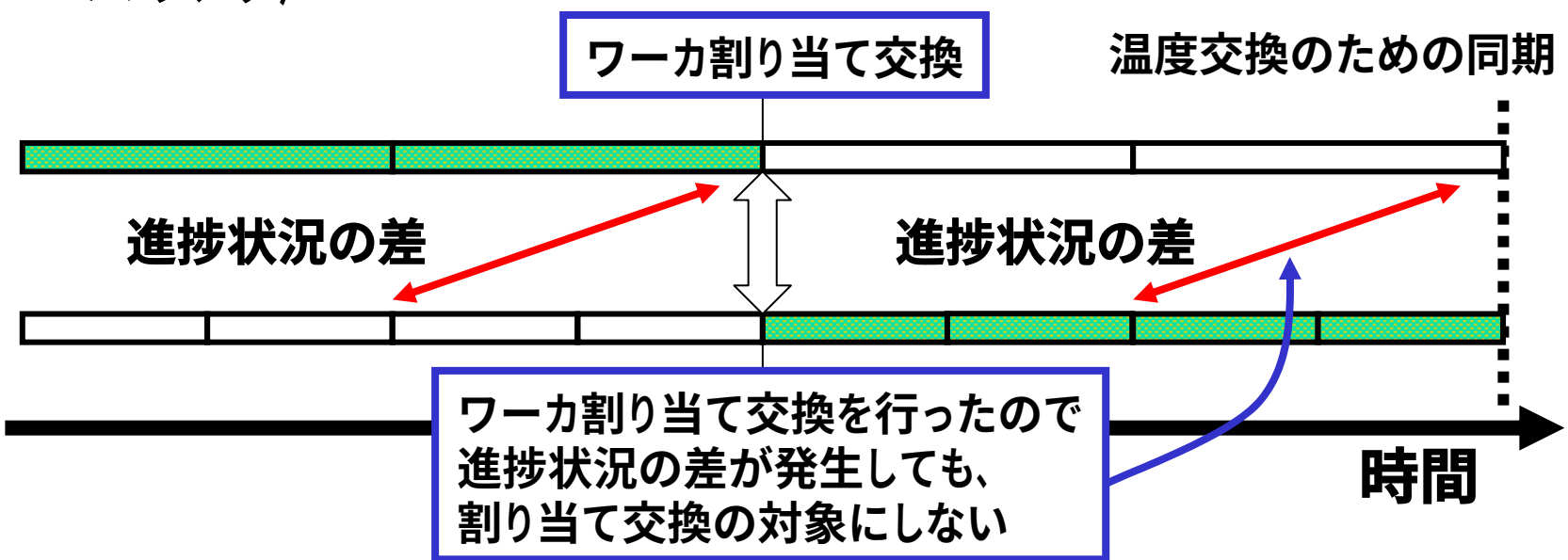
- ワーカ割り当て交換決定のパラメタ（交換決定パラメタ）の設定
  - 最も進捗しているワーカよりも閾値以上遅れている場合に割り当て交換
  - 交換決定パラメタ ＝ （現在のレポート数） / （シミュレーション区間の総レポート数）で決定

温度交換のための同期

- ワーカの行うマスタへの報告（マスタ報告パラメタで設定）

割り当て交換法の動作 (5/5)

- 極端に遅いワーカが存在した場合、そのワーカに対して連続した割り当て交換の発生
  - 割り当て交換を行ったワーカに対しては、一定期間(一定のレポート数)、割り当て交換の対象にしない　(交換非対象パラメタ)



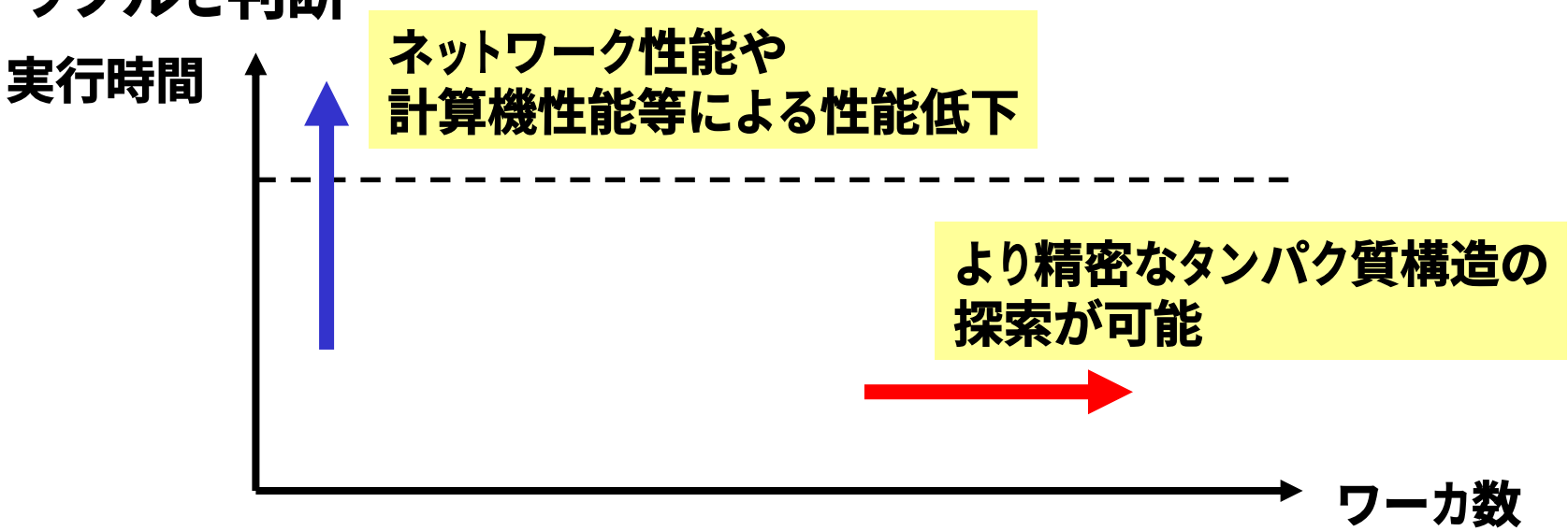
実験 (1/2)

以下の環境でMPIによって並列化された
REMD Toolkit を実行し、プログラムの実行時間を計測

- **性能ホモな環境**
  - **PCクラスタ上での実行**
- **性能ヘテロな環境**
  - **PCクラスタ上でワーカの一部に負荷をかけて実行**
- **広域ネットワーク環境**
  - **Globusを使用したMPI実装であるMPICH-G2を用いた実行**

実験 (2/2)

- シミュレーション時間 (分子シミュレーション 200 steps、温度交換1000 times) を固定して実験
- ワーカ数増加により、より精密なタンパク質構造の探索が可能
- ワーカ数増加において実行時間の増加が抑えられていればスケーラブルと判断



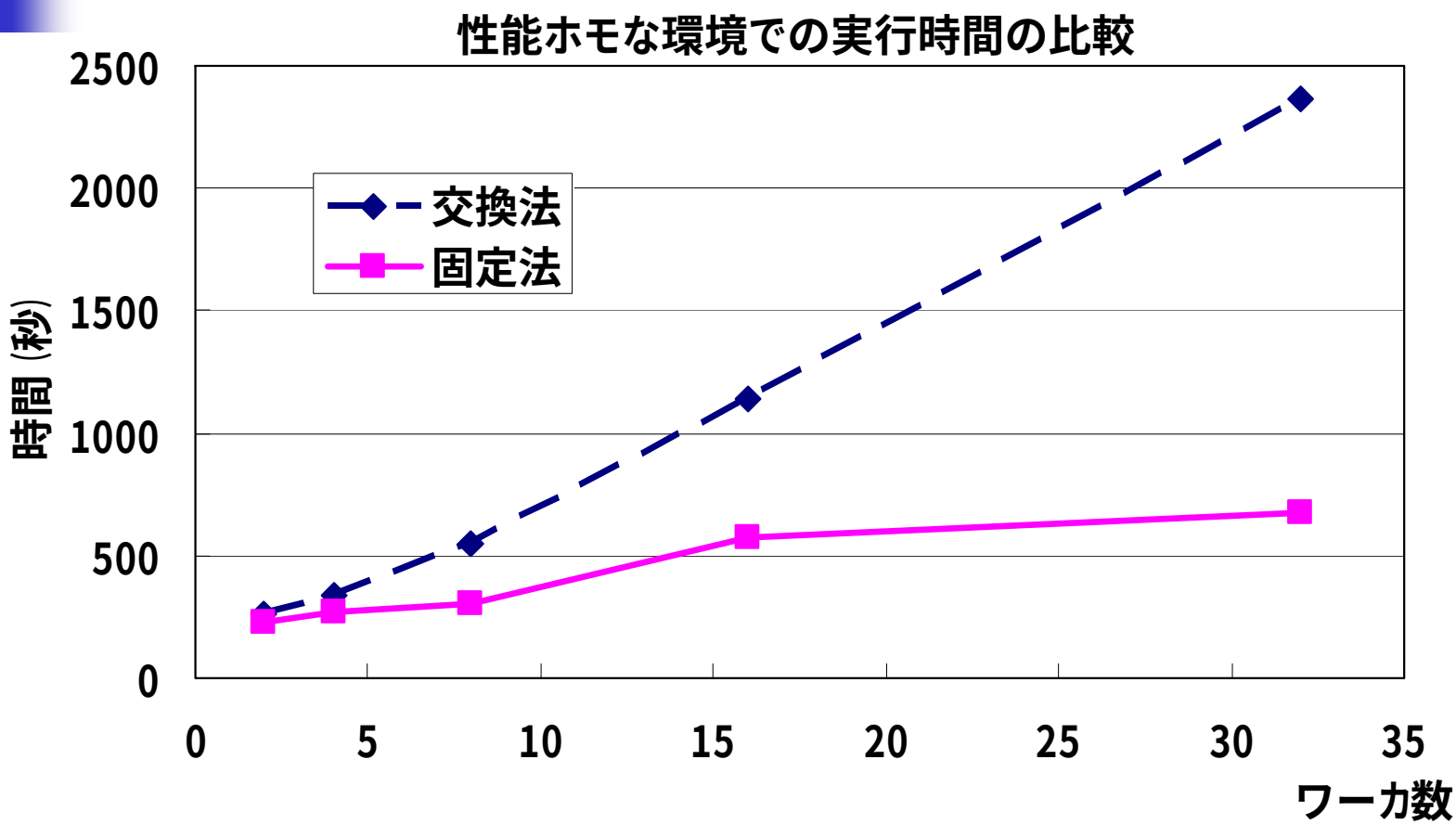
性能ホモな環境での実験(1/7)

以下の3点に関して実験

- ワーカ数を増加させた場合のスケーラビリティー
- マスタ、ワーカ間の通信のコスト
- ワーカ間の割り当て交換の際の通信のコスト

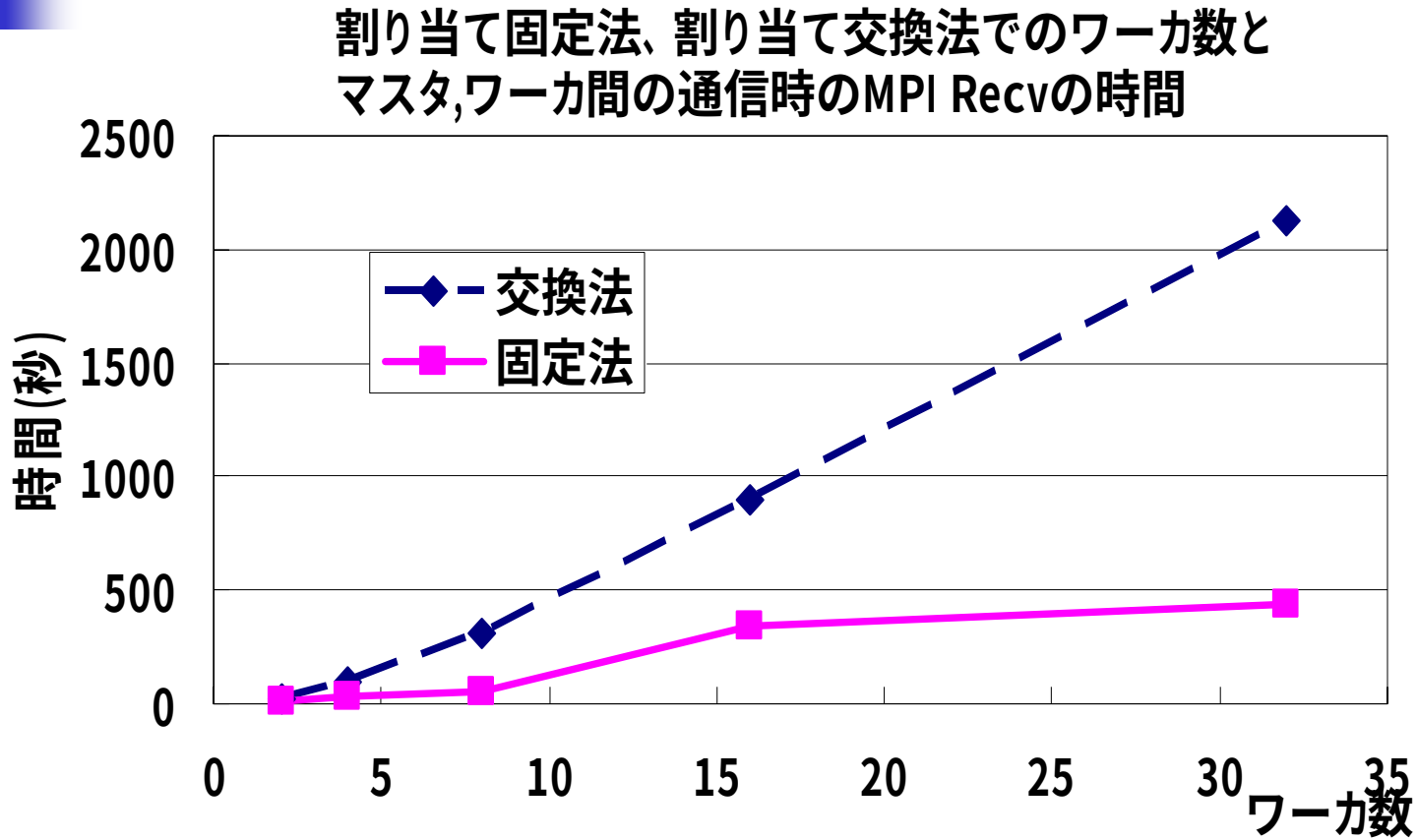
- 実行環境　東工大松岡研Presto III
  - CPU: Athlon MP 1900+ Memory: 768MB
  - Network: 100base-T OS: Linux 2.4.18
  - MPI：MPICH-1.2.5

性能ホモな環境での実験(2/7)



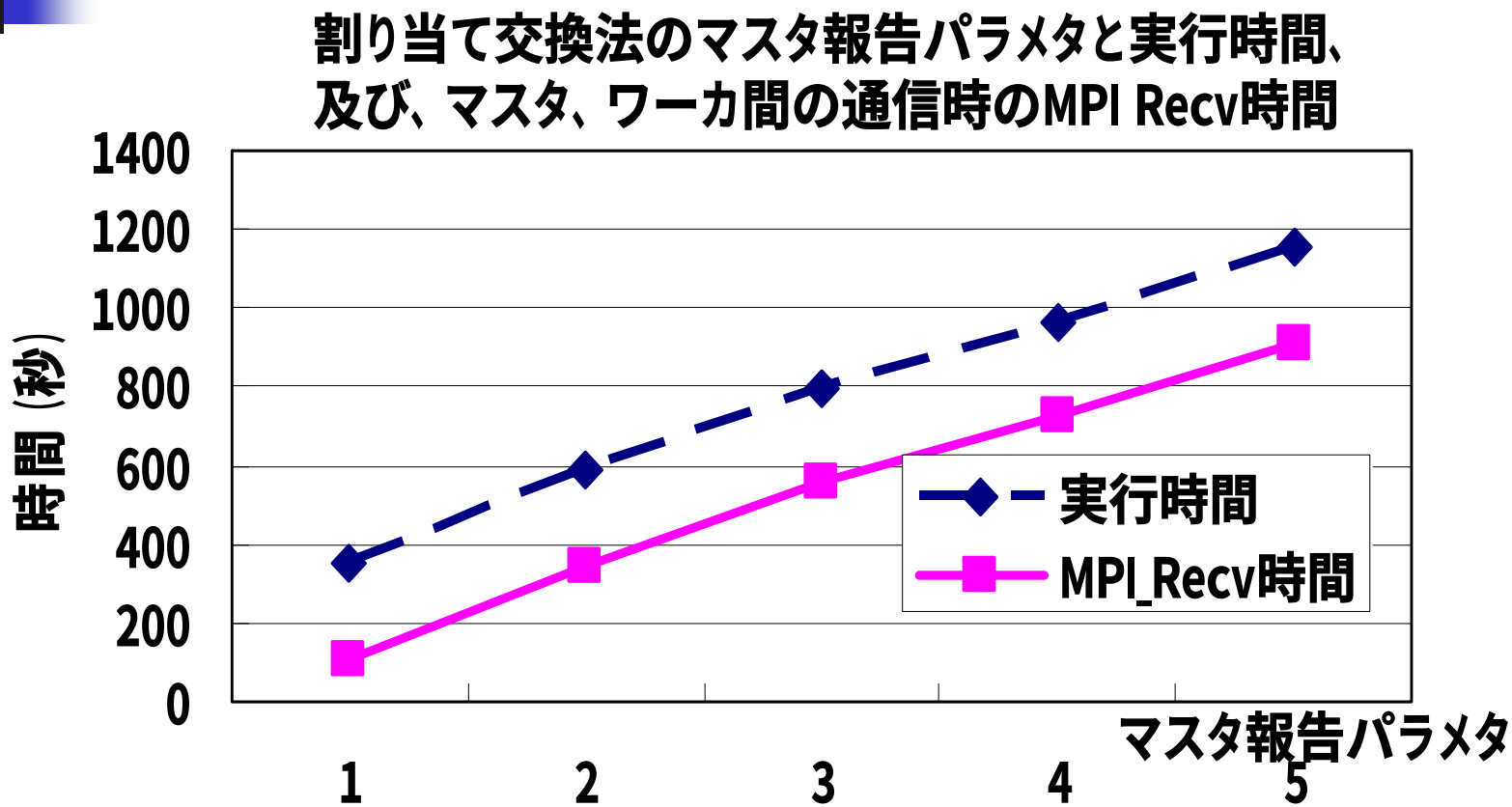
- 割り当て固定法 … ワーカ数の増加に対し、実行時間の増加が抑えられている
- 割り当て交換法 … ワーカ数の増加に応じた実行時間の増加

性能ホモな環境での実験(3/7)



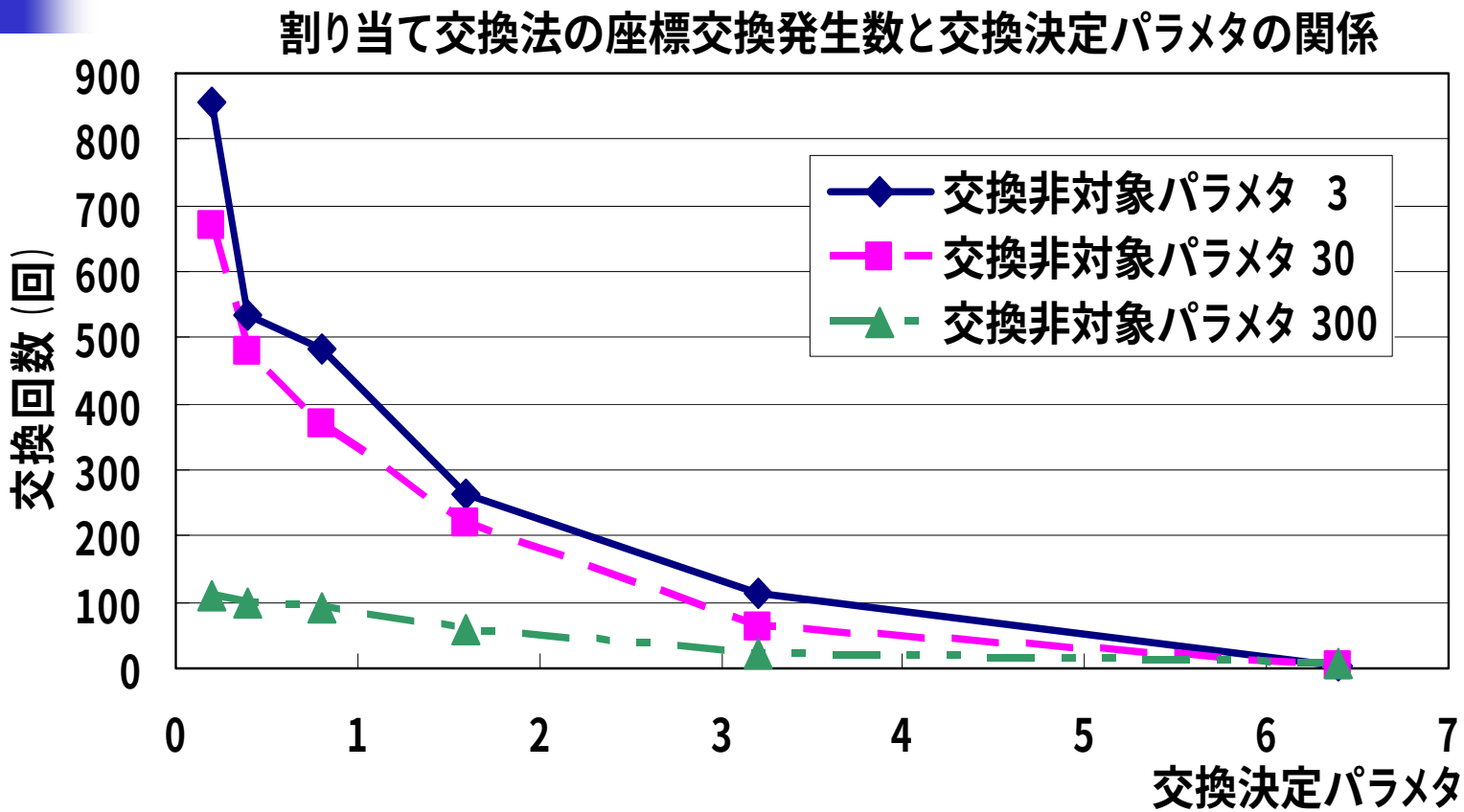
- 性能低下の要因　… MPI Recvに費やされる時間の増加
- 割り当て交換法　… ワーカ数の増加に応じたMPI Recv時間の増加

性能ホモな環境での実験(4/7)



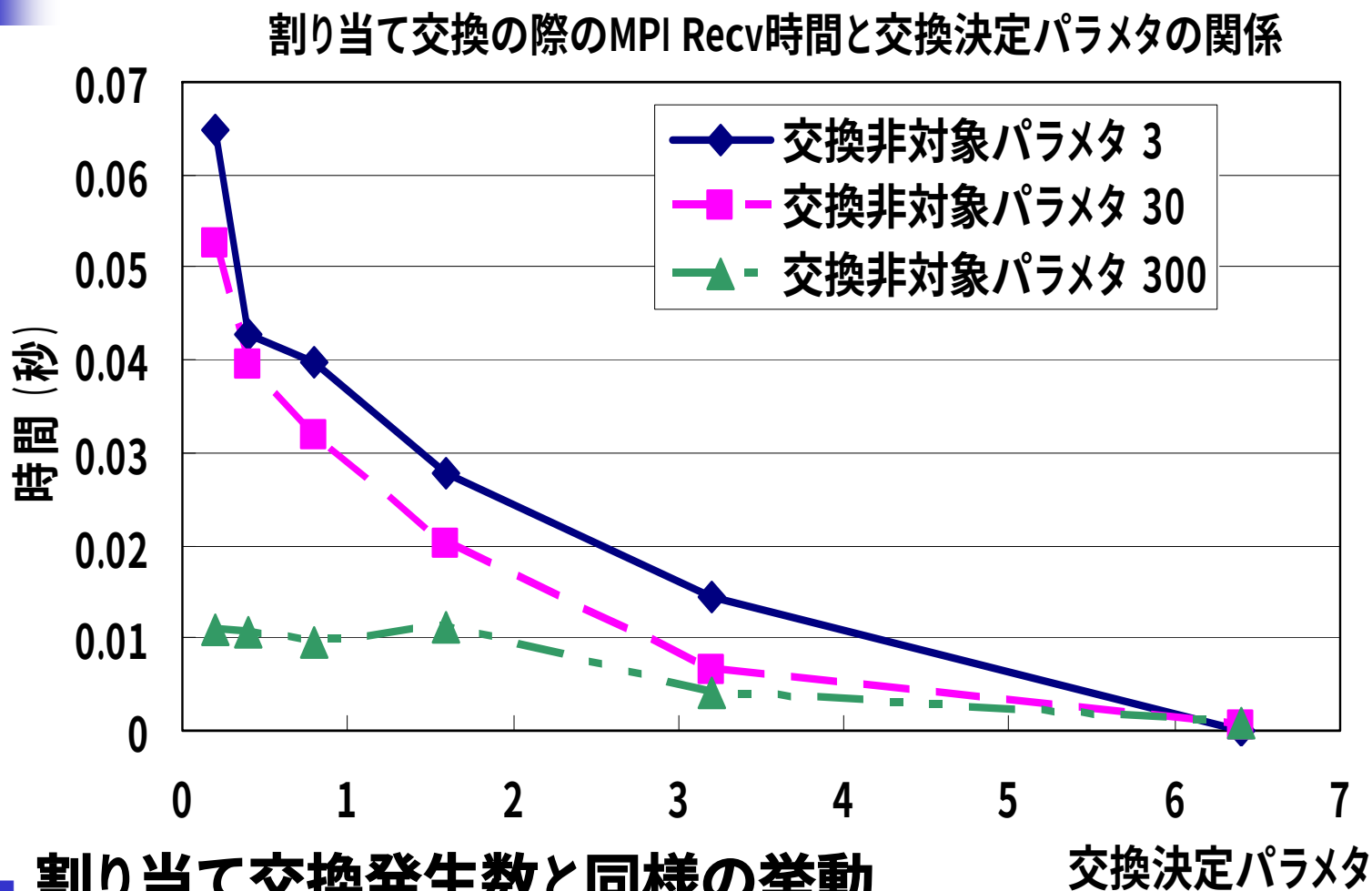
- ワーカのマスタへの報告数の増加に応じた実行時間の増加
- MPI Recvに費やされる時間の増加が要因

性能ホモな環境での実験(5/7)



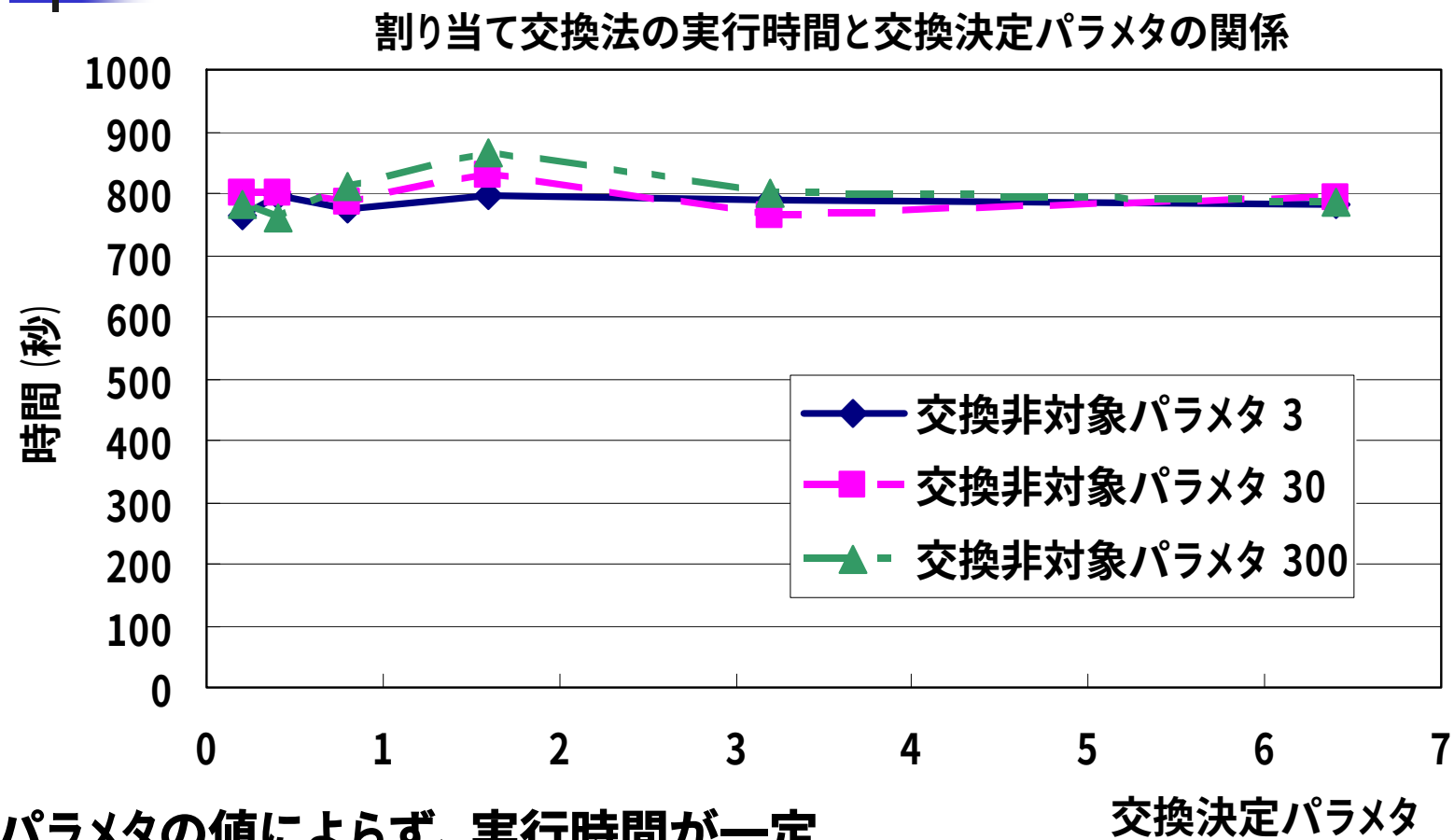
- 交換決定パラメタ値が小さい場合、大量の座標割り当て交換の発生

性能ホモな環境での実験(6/7)



- 割り当て交換発生数と同様の挙動

性能ホモな環境での実験(7/7)



- パラメタの値によらず、実行時間が一定
- ワーカ間の割り当て交換の際の時間が全体の実行と比較して極めて小さい

議論
性能ホモ環境での実験

- 割り当て固定法
  - ワーカ数の増加に対し、実行時間の増加が抑えられている
- 割り当て交換法
  - ワーカ数の増加に応じて、実行時間の増加
  - ワーカがマスタへ行うシミュレーションの進捗状況報告の通信にかかる時間が増加

性能ヘテロである環境での実験 (1/5)

以下の2点に関して実験

- **負荷の大きさと実行時間**
- **負荷の大きさと割り当て交換発生回数**

- **実行環境　東工大松岡研**Presto III
  - CPU: Athlon MP 1900+ Memory: 768MB
  - Network: 100base-T OS: Linux 2.4.18
  - MPI：MPICH-1.2.5

性能ヘテロである環境での実験 (1/5)

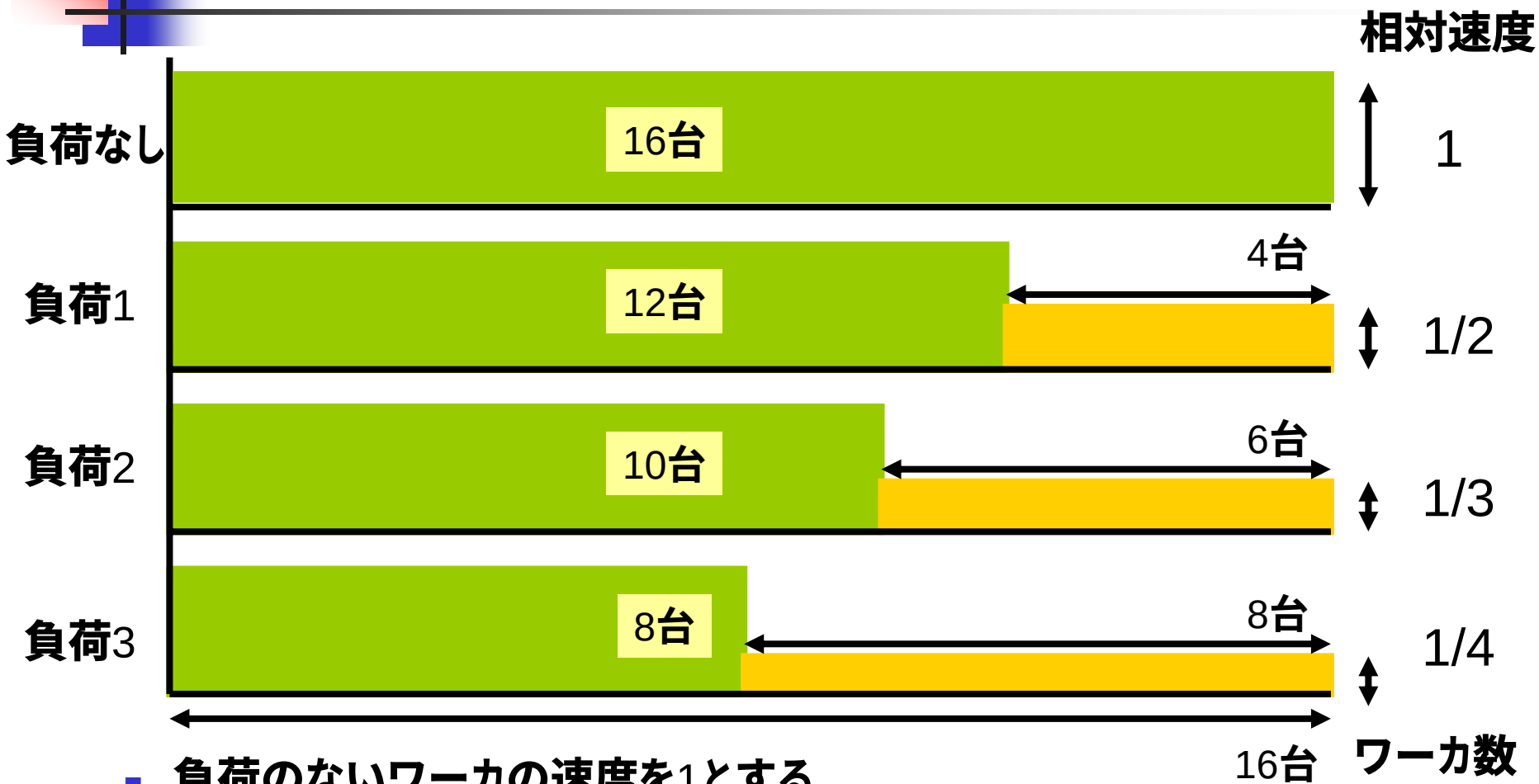
- 一部のPEに複数のワーカを割り当て、擬似的に計算機性能が低下したのと同等な環境を設定

	高速ワーカ数	低速ワーカ数	低速ワーカ相対速度
負荷なし	16	0	1
負荷1	12	4	1/2
負荷2	10	6	1/3
負荷3	8	8	1/4

ワーカ数16

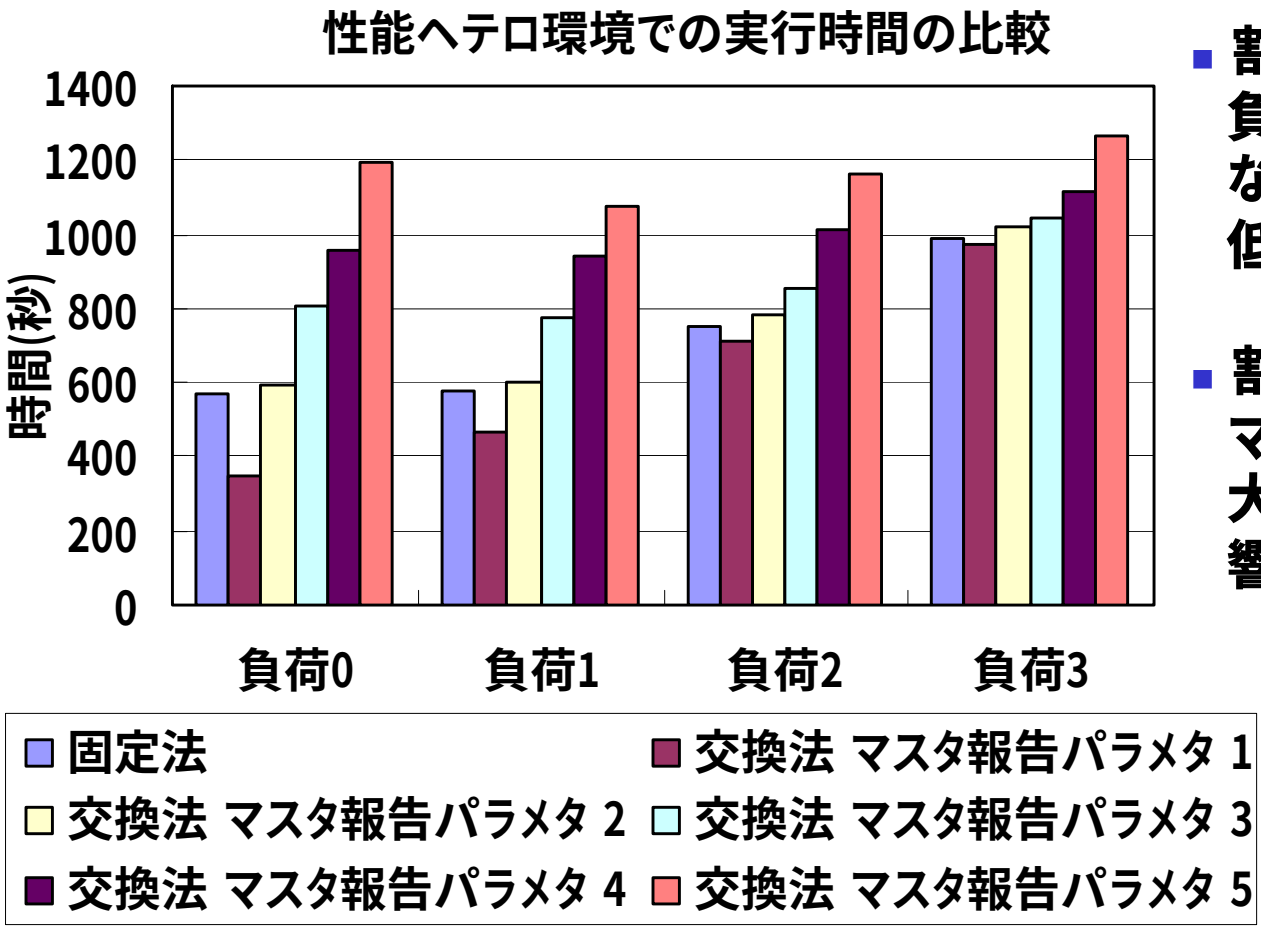
- ワーカ数　16台に固定
- 交換決定パラメタ 3
- マスタ報告パラメタ　1から5まで変化

性能ヘテロである環境での実験 (2/5)



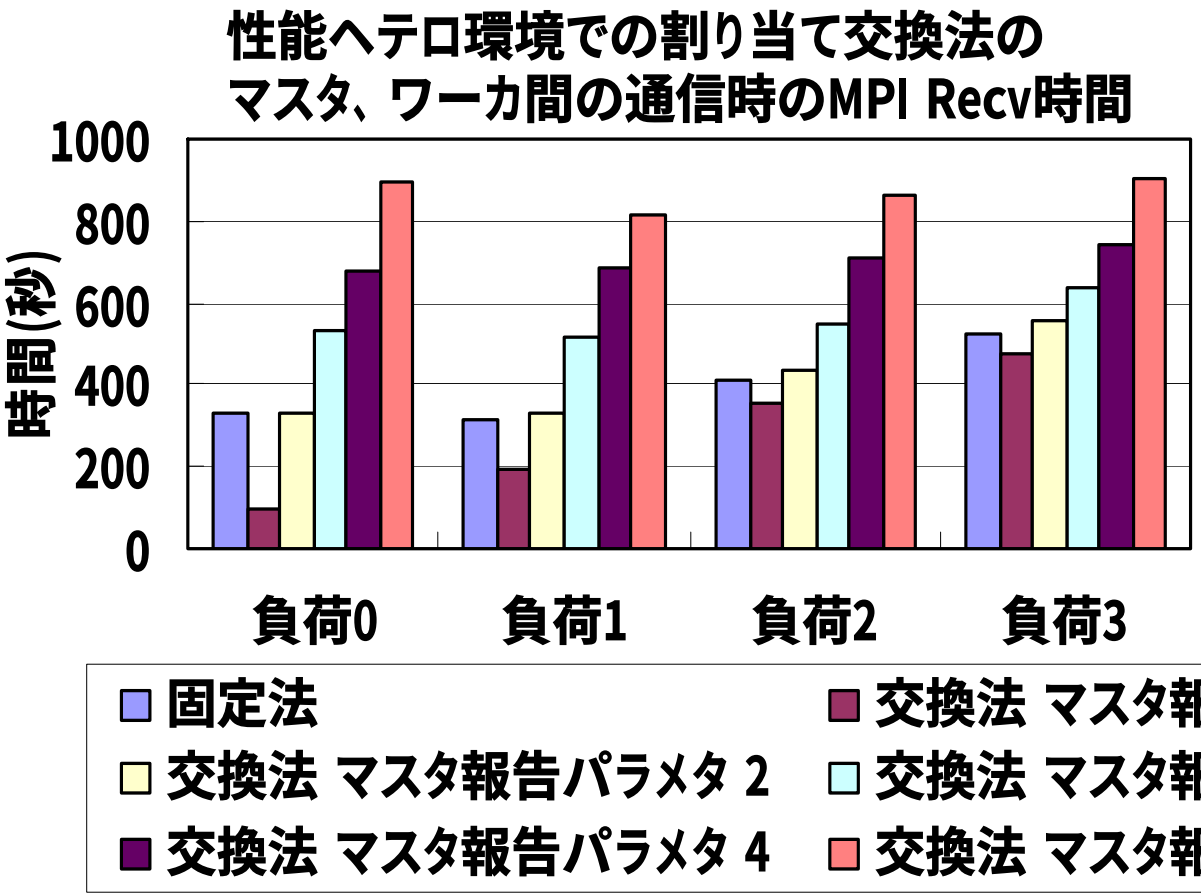
- 負荷のないワーカの速度を1とする
- 負荷1 2 3 では高速ワーカと低速ワーカを混合させて実行

性能ヘテロである環境での実験 (3/5)



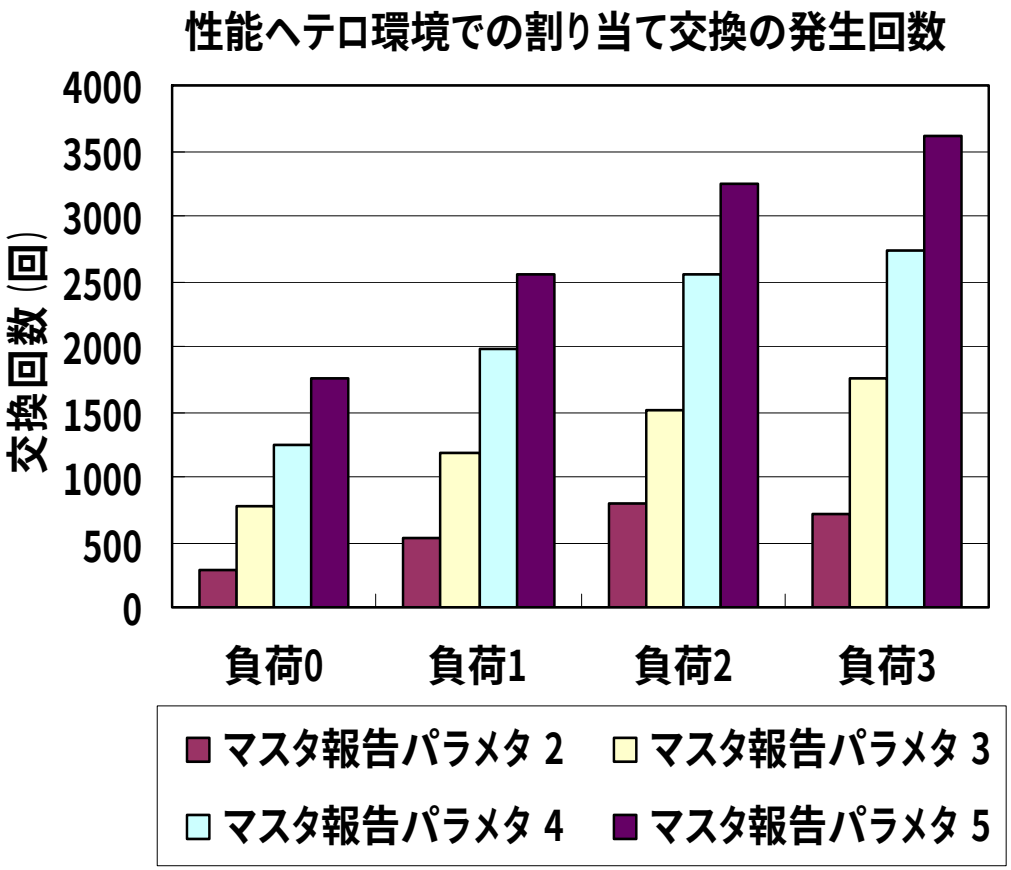
- 割り当て固定法
  負荷の増加につれ、低速なPEの性能に応じて性能低下
- 割り当て交換法
  マスタ報告パラメタの値が大きい場合、低速PEの影響を緩和

性能ヘテロである環境での実験 (4/5)



- 割り当て固定法
  負荷の増加につれ、線形に低下
- 割り当て交換法
  負荷の増加に対し、増加の割合が抑えられている

性能ヘテロである環境での実験 (5/5)



- マスタ報告パラメタが大きい場合負荷の増加につれ、交換回数の増加
- マスタ報告パラメタが小さい場合負荷の増加に対し、交換回数が少なく抑えられている

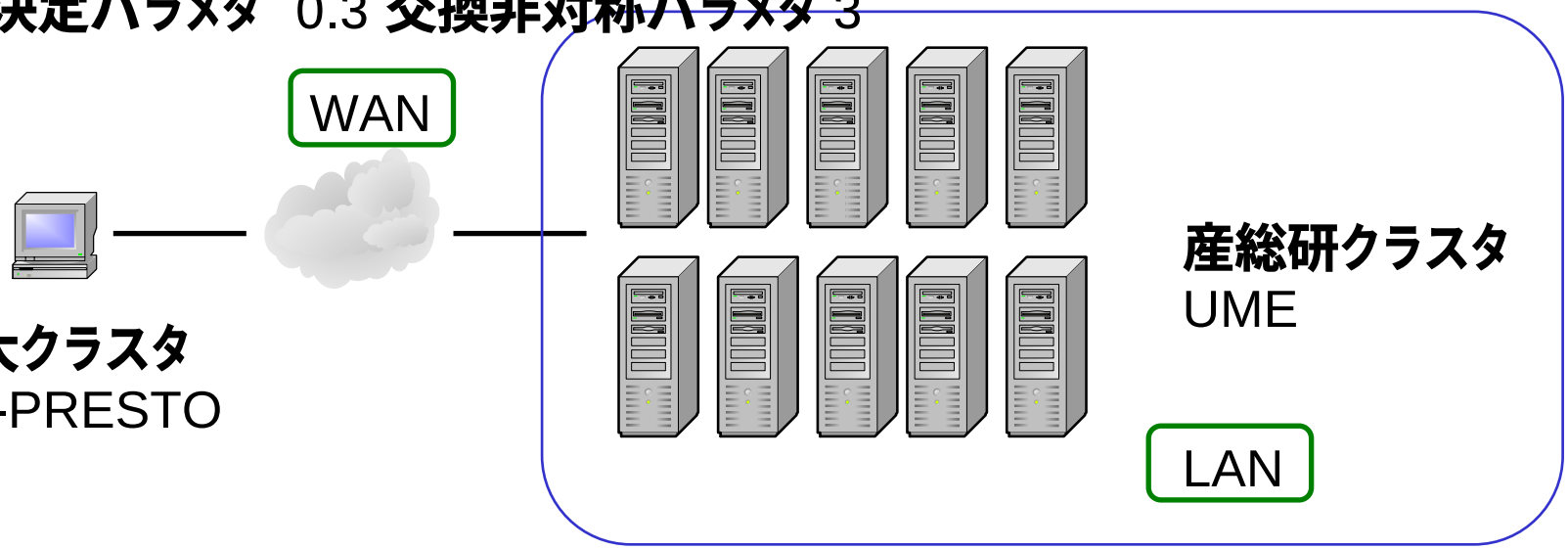
議論
性能ヘテロ環境での実験

- 割り当て固定法
  - 負荷の増加につれ、低速なワーカに律速
  - 実行時間の線形な増加
- 割り当て固定法
  - マスタ報告パラメタの値が大きい場合
    - 実行時間増加の抑制
    - 大量の割り当て交換の発生

広域ネットワーク環境での実験 (1/4)

- **マスタにGSIC-PRESTO、ワーカにUMEを割り当て実行 (実験1　広域性能ホモ環境)**
- **マスタにUME1台、ワーカにGSIC-PRESTO1台 (低速ワーカ)、UME15台(高速ワーカ)を割り当て実行 (実験2　広域性能ヘテロ環境)**

交換決定パラメタ　0.3 交換非対称パラメタ 3

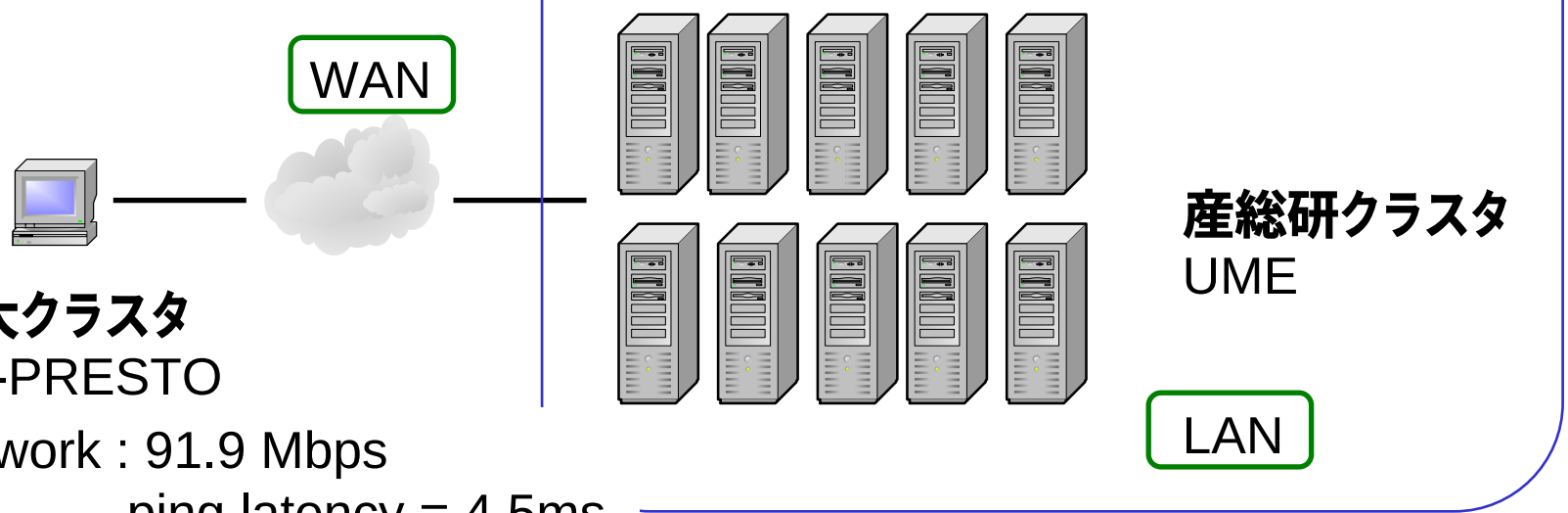


広域ネットワーク環境での実験 (2/4)

- **実行環境**

CPU:PentiumIII 1400MHz ×2,
Mem:2.3GB,OS:Linux2.4.18,
Network : 1000Base-T

CPU:PentiumIII 800MHz ×2
Mem:512MB,OS:Linux 2.2.20
Network : 100Base-T
MPI: MPICHG2-1.2.5

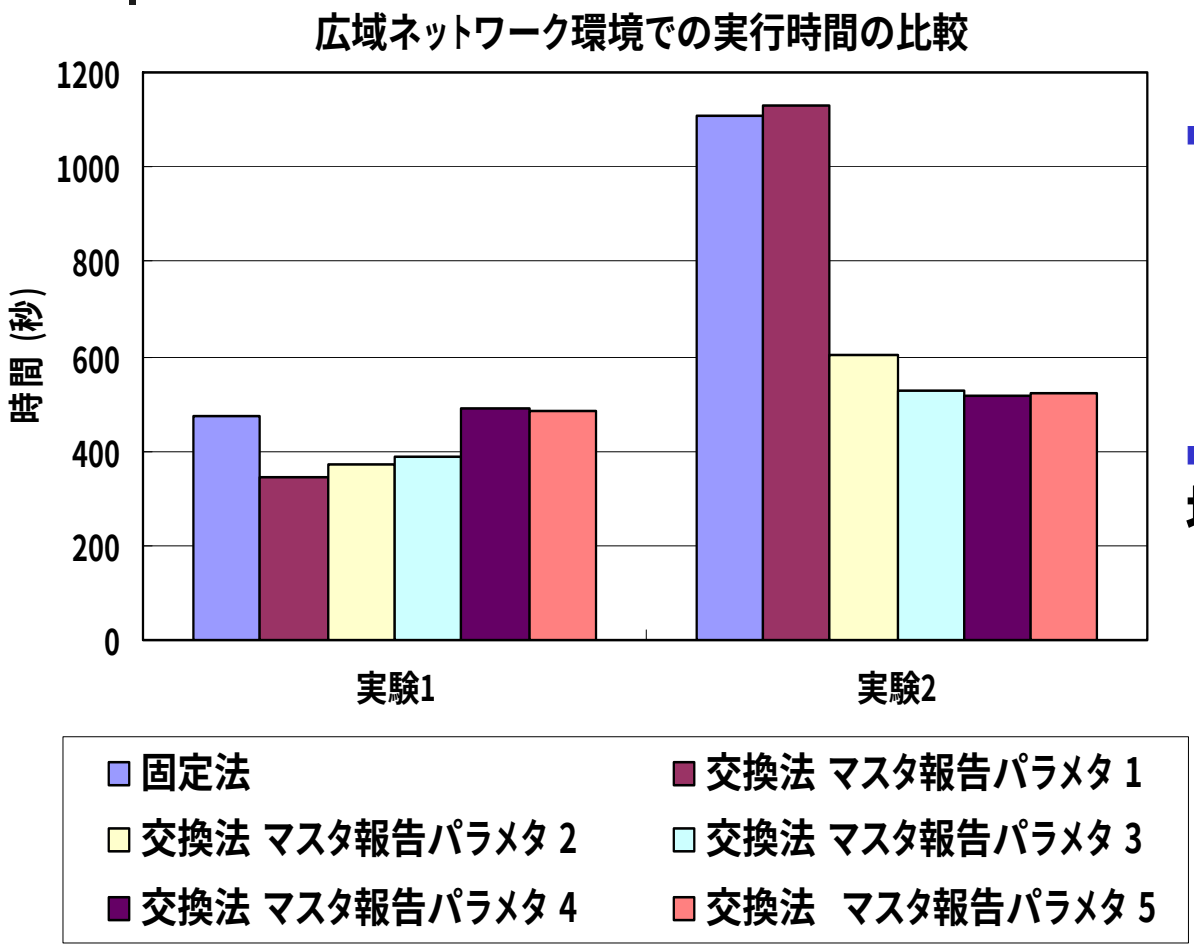


東工大クラスタ
GSIC-PRESTO

産総研クラスタ
UME

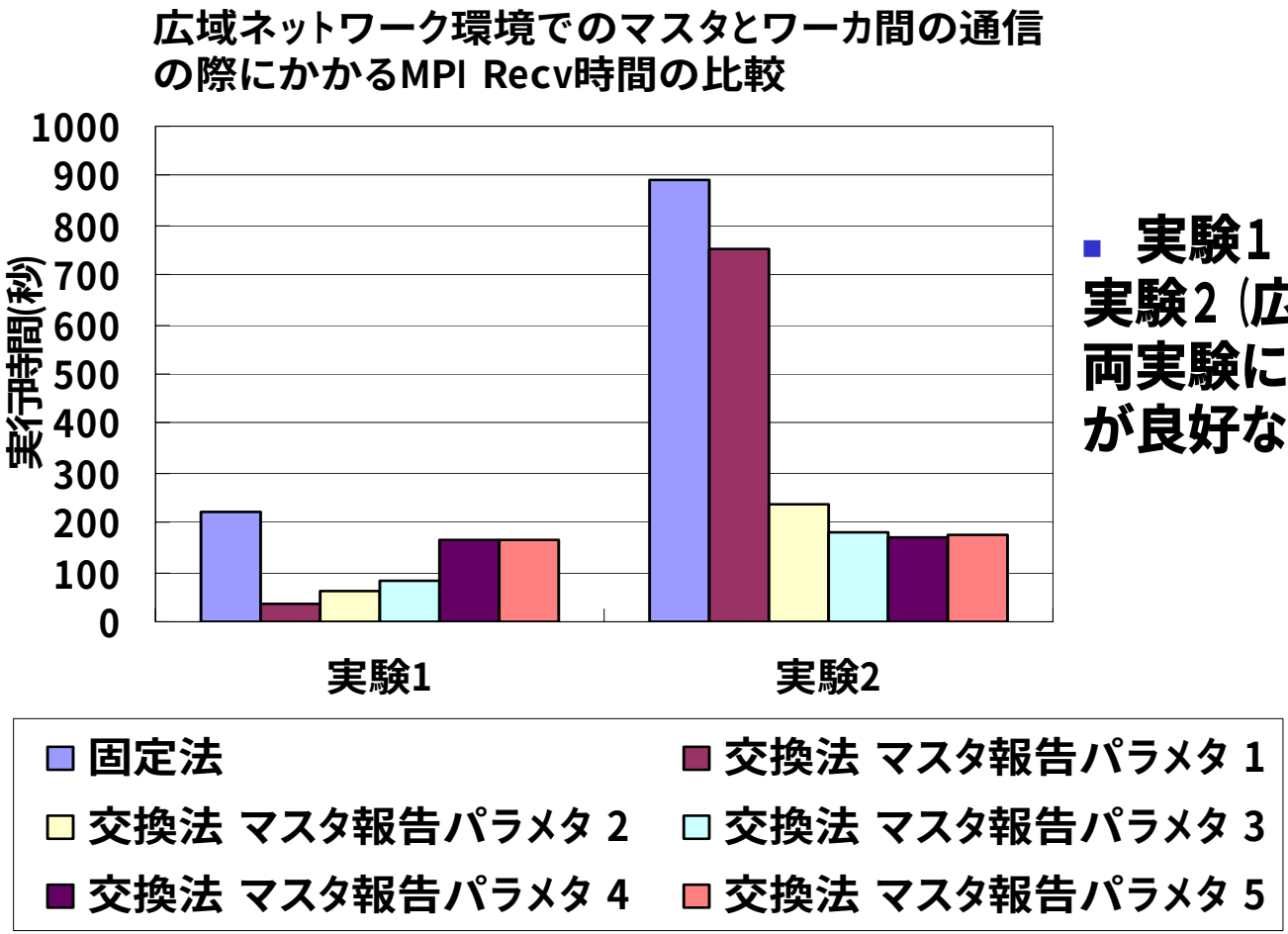
Network : 91.9 Mbps
ping latency = 4.5ms

広域ネットワーク環境での評価 (3/4)



- 実験1 （広域性能ホモ環境）
  実行時間の増加がみられるが、割り当て交換法が高速に動作
- 実験2 （広域性能ヘテロ環境）
  マスタ報告パラメタの値が小さい場合に性能低下がみられるが、割り当て交換法が高速に動作

性能ヘテロである環境での評価 (4/4)



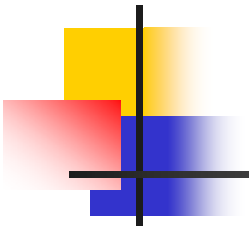
- 実験1（広域性能ホモ環境）、実験2（広域性能ヘテロ環境）の両実験において割り当て交換法が良好な結果を示す

議論
広域ネットワーク環境での実験

- 広域性能ホモ環境、広域性能ヘテロ環境の実験より割り当て交換法の有効性を確認
- 割り当て交換法によりマスタ、ワーカ間のバリア同期のための計算停止時間(MPI Recv時間)の減少

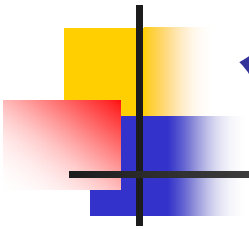
まとめ(1/2)

- REMD Toolkitの性能ヘテロ環境へ対応するための割り当て交換法の提案、実装
- 以下の環境での性能評価
  - 性能ホモ環境
  - 性能ヘテロ環境
  - 広域ネットワーク環境



まとめ(2/2)

- いくつかの環境では性能低下
  - 性能ホモ環境
    - MPIの実装の問題で通信がバッファリングされている可能性
    - 別のMPIの実装の場合に関しても調査、検討
- 性能ヘテロ環境では有効
  - 性能ヘテロ環境
  - 広域ネットワーク環境



今後の課題

- 割り当て交換法のアルゴリズム、パラメータの改良
- 高レイテンシなネットワークでの性能測定