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)]

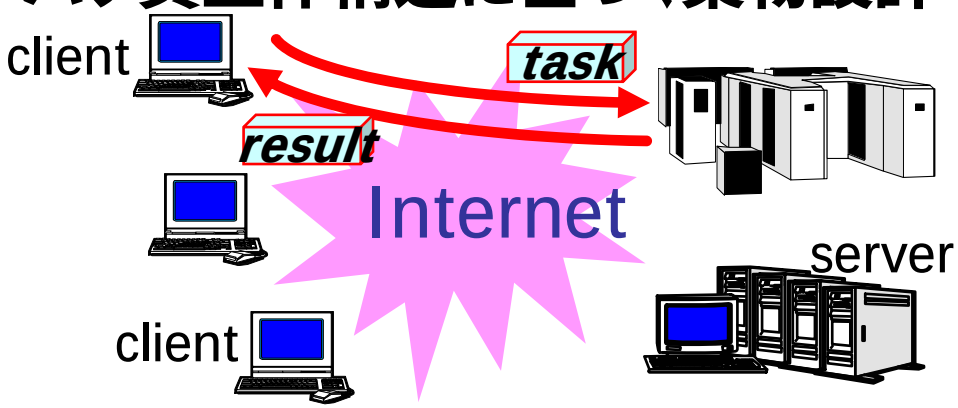
伊藤正勝 [2)]　　松岡　聡 [1) 3)]

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

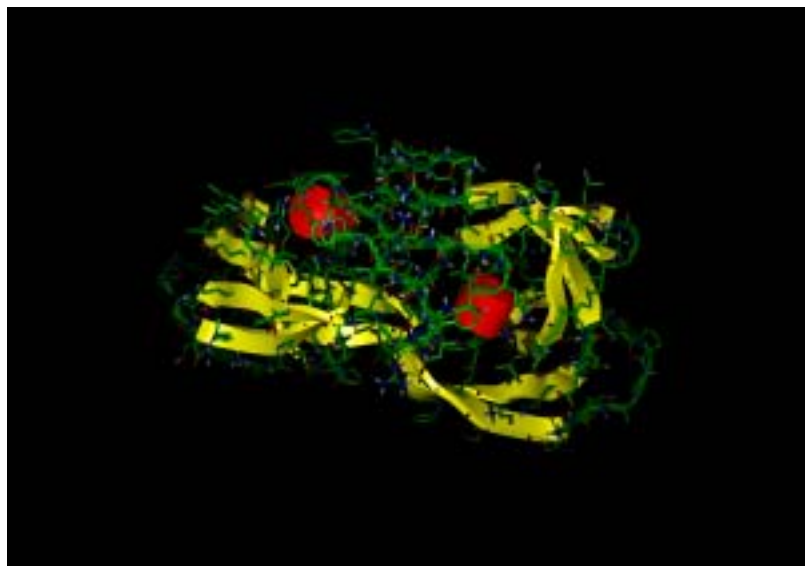
- グリッド環境向けの実アプリケーション開発
  - タンパク質立体構造に基づく薬物設計　等



REMD Toolkit

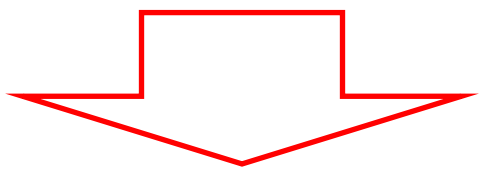
Replica Exchange Molecular Dynamics

- 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



レプリカを増加させた場合のスケーラビリティー
グリッド環境での挙動が不明

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のグリッド環境上で性能の確認**
- **REMD Toolkitの性能ヘテロ環境への対応**

成果

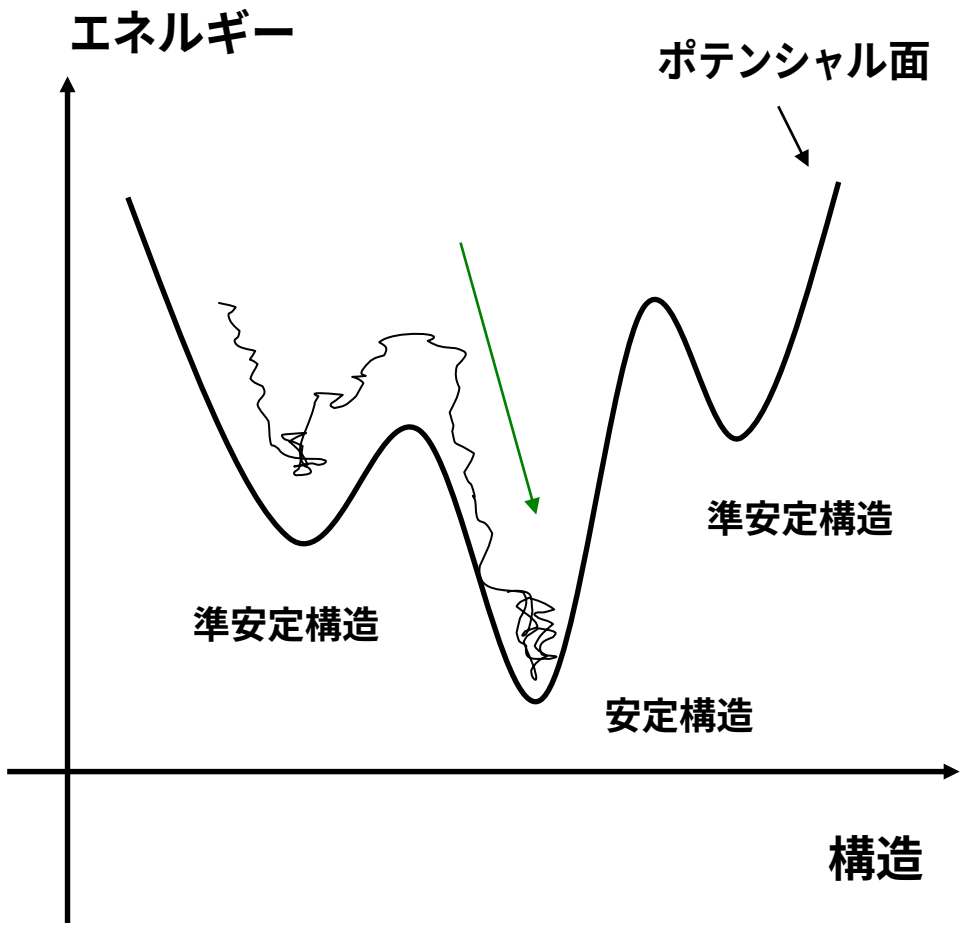
- **REMD Toolkitが100台規模まで十分にスケールすることを確認**
- **REMD Toolkitの性能ヘテロ環境への対応の有効性を確認**

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)

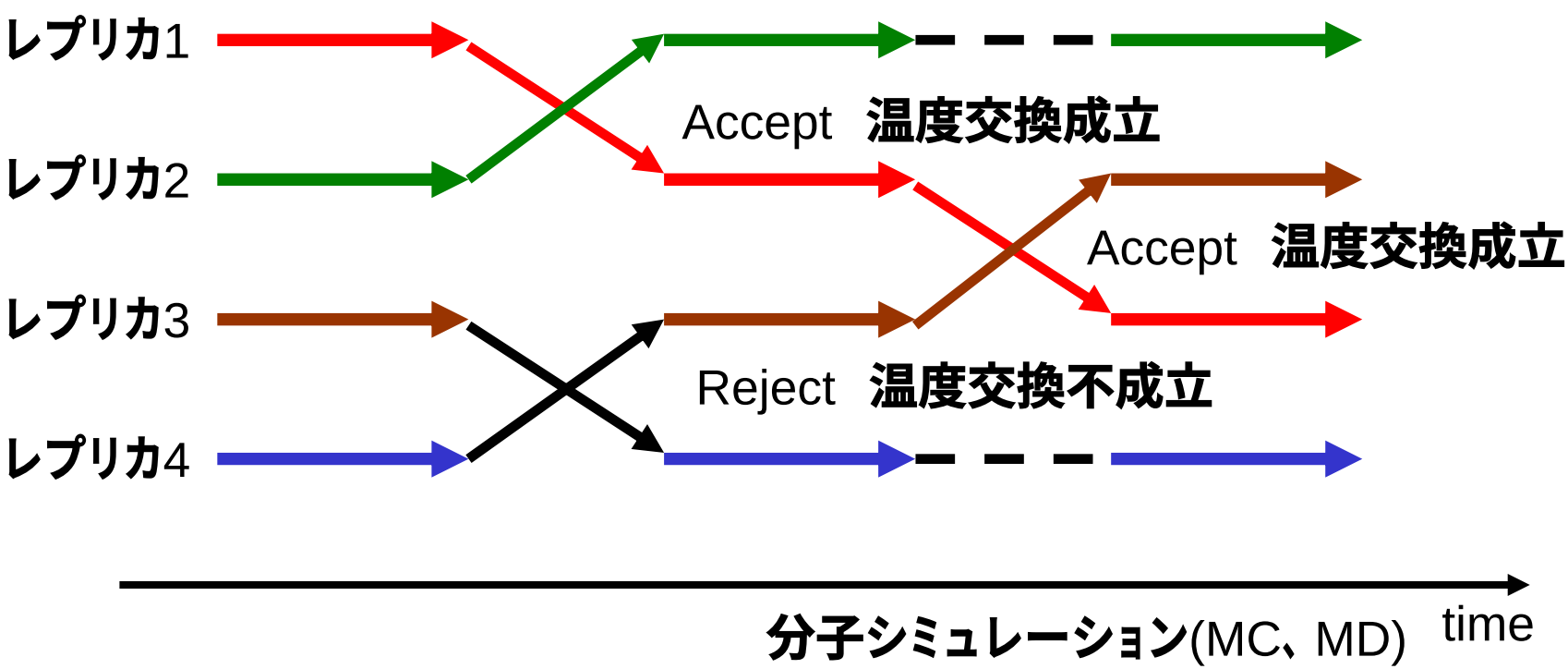
⇒ レプリカ交換法



レプリカ交換法 (2/3)

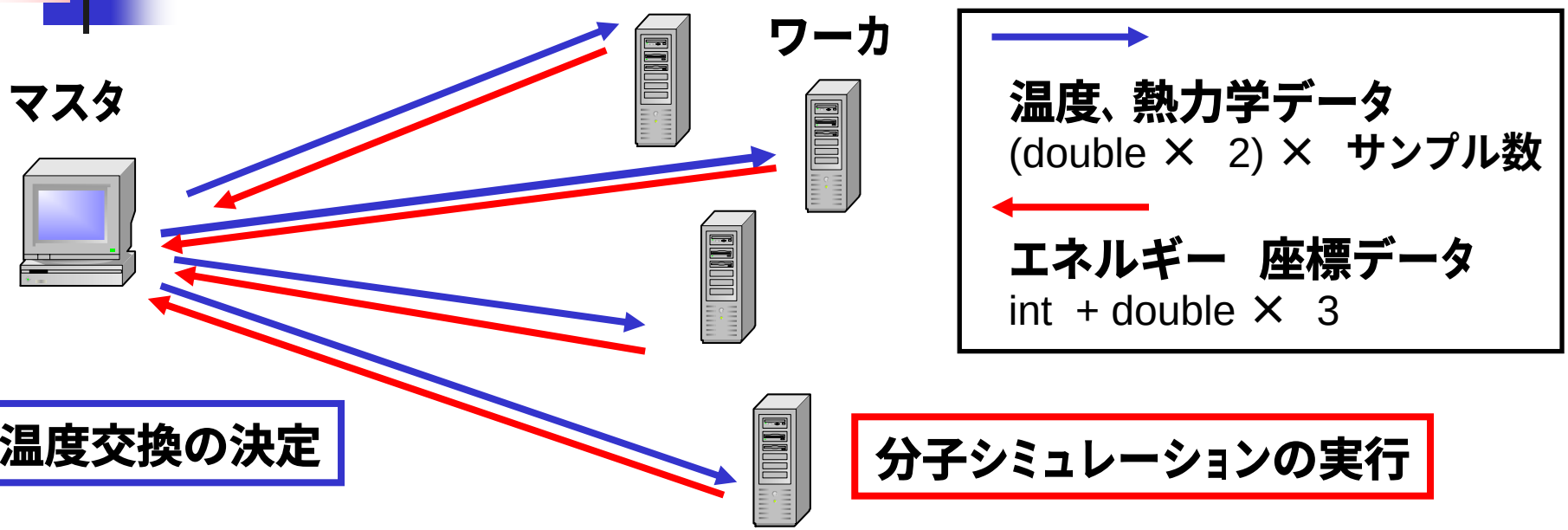
- 互いに相互作用しない異なる温度を持つタンパク質構造のコピー (レプリカ) を複数個用意
- レプリカ交換
  - 各レプリカで独立に分子シミュレーション(MC, MD)を実行
  - 途中で温度値が隣接した2つのレプリカの温度を交換

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

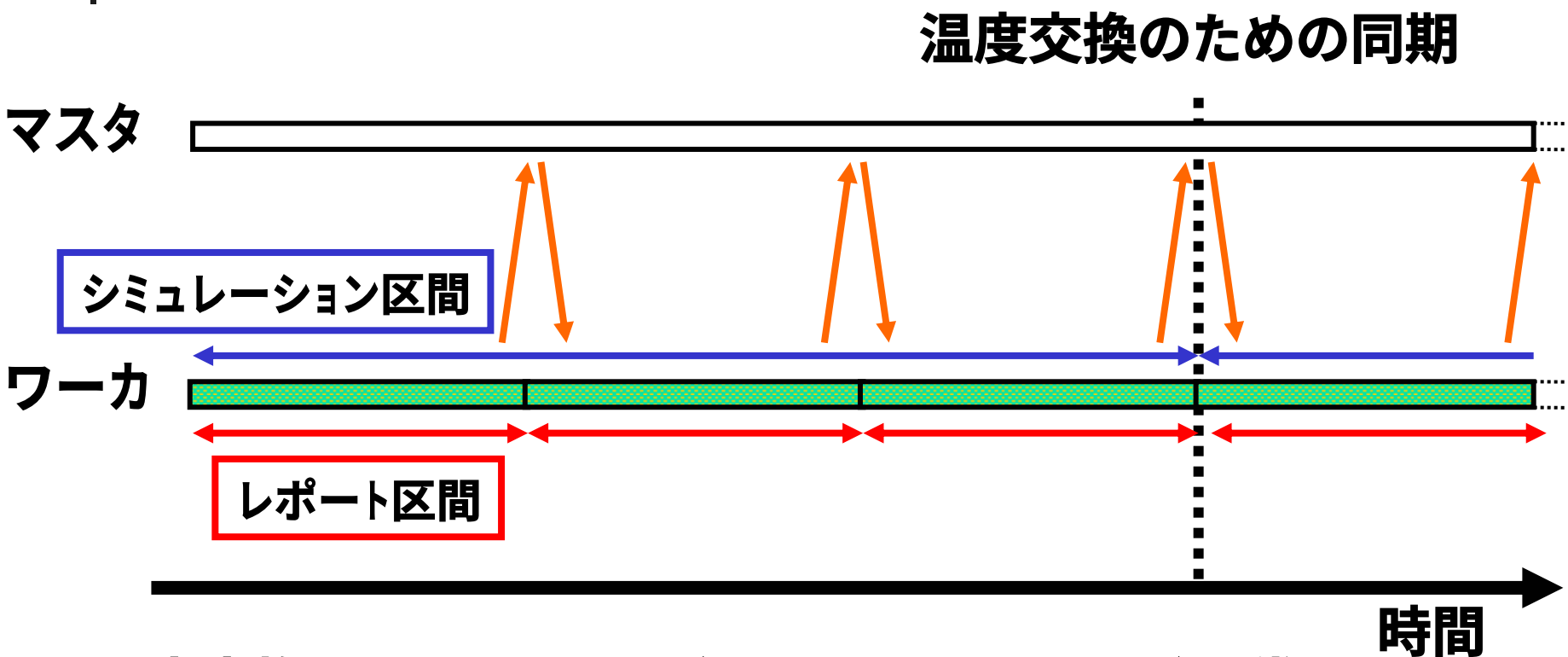
- 温度交換周期毎に全てのワーカが同期
- ワーカの性能差がある場合高速PEが低速PEの計算終了を待つ

⇨ 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)

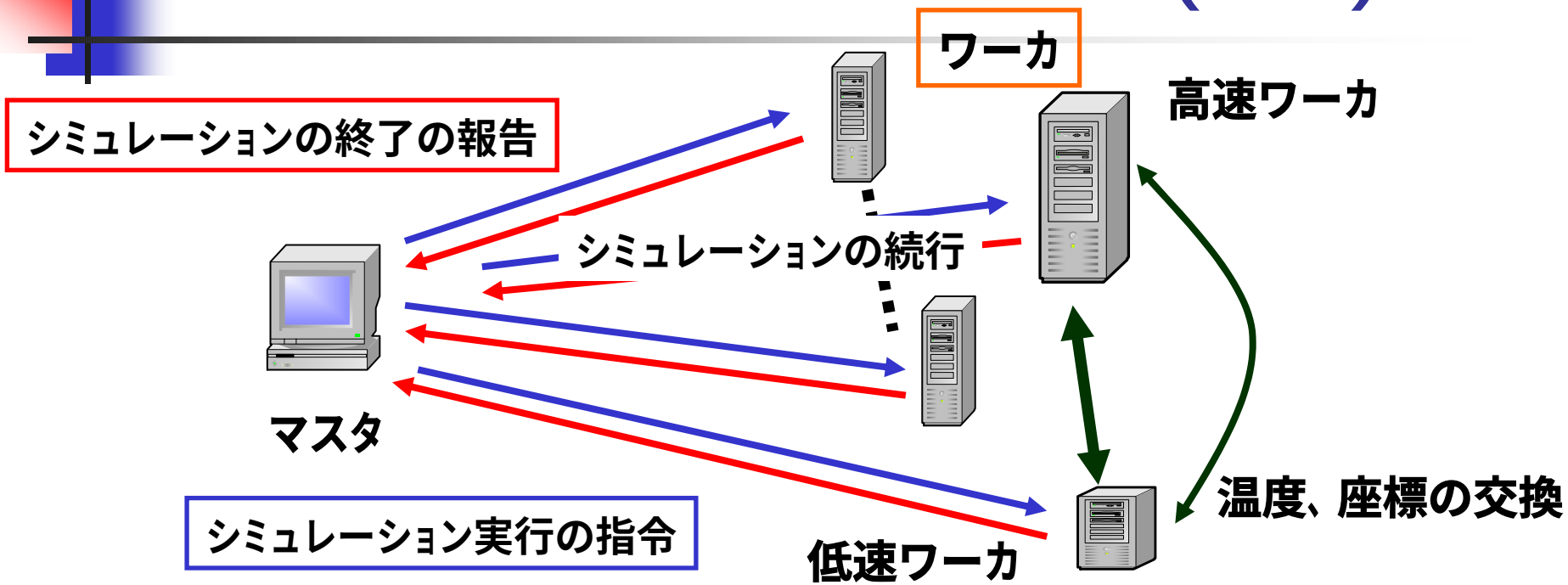
- 温度交換の際、全ワーカとのバリア同期の解消
  - 温度交換するワーカとの1対1の同期
- 各ワーカの進捗状況のモニタ
  - 進捗が遅れているワーカと進んでいるワーカに対して割り当て(温度と座標)を交換

割り当て交換法の動作 (1/5)



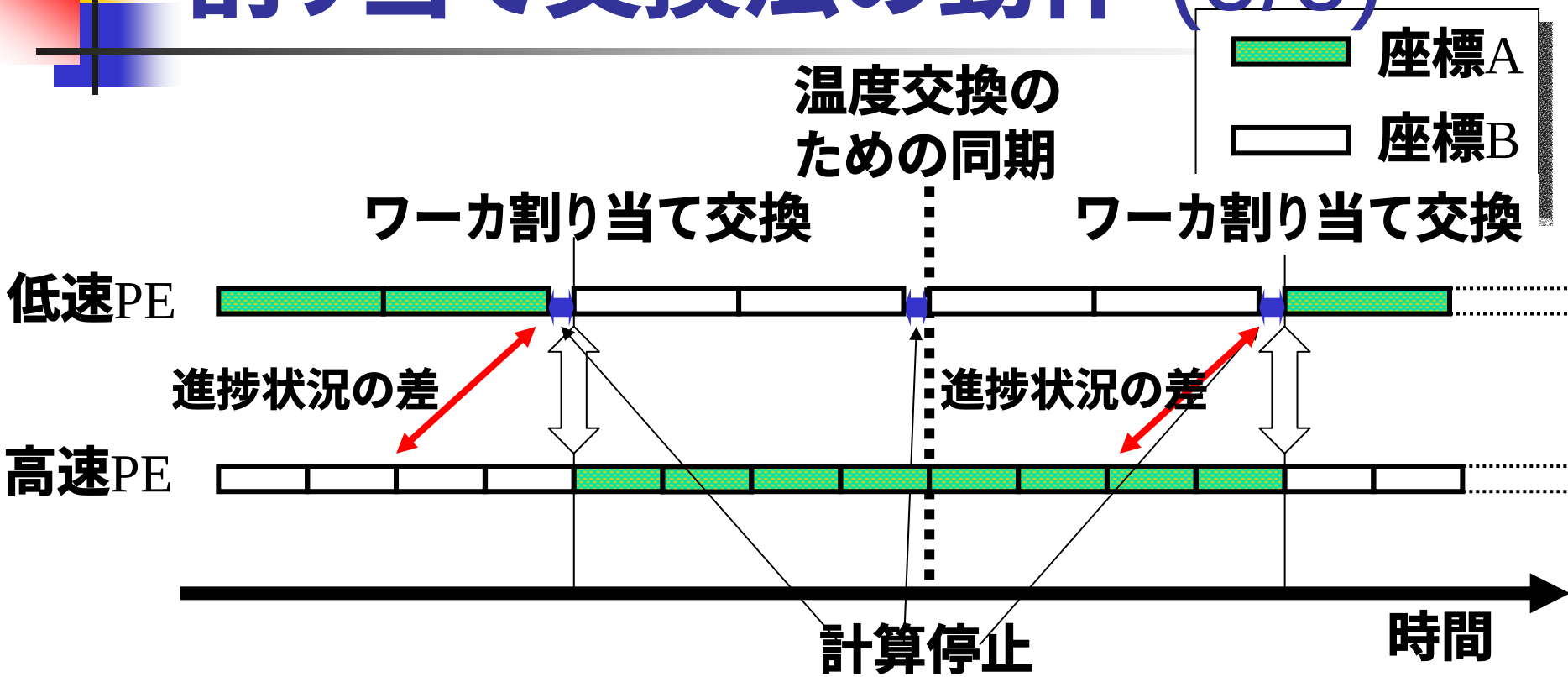
- 温度交換周期までの区間 (シミュレーション区間) を幾つかの区間に分割 (レポート区間)
- ワーカはレポート区間毎にマスタに進捗状況を報告

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

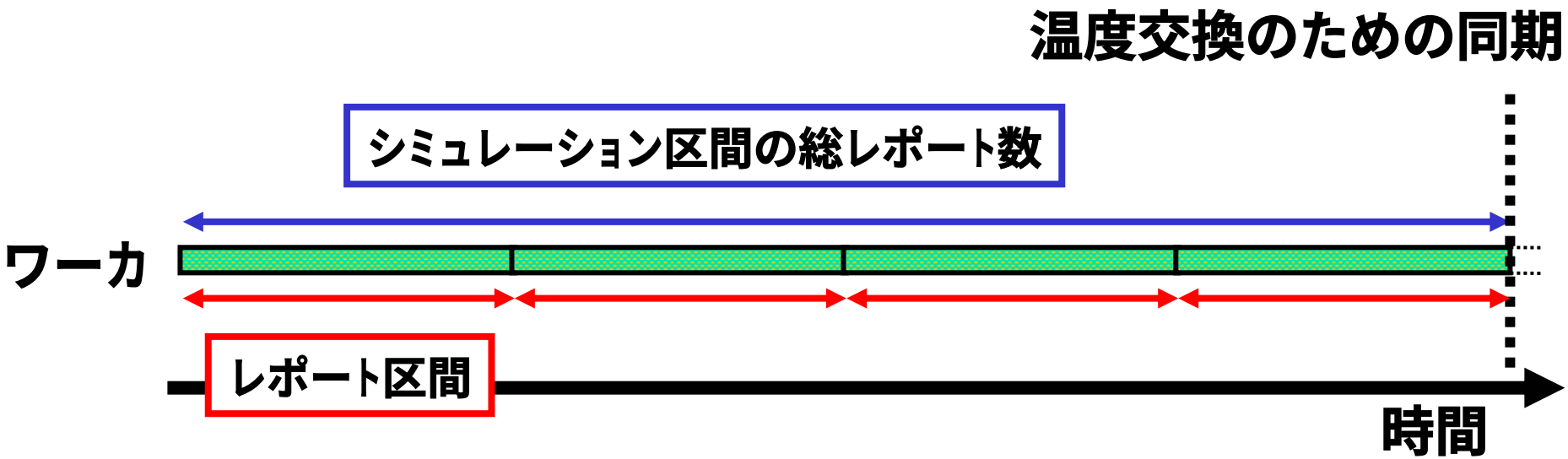
割り当て交換法の動作 (3/5)



- 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

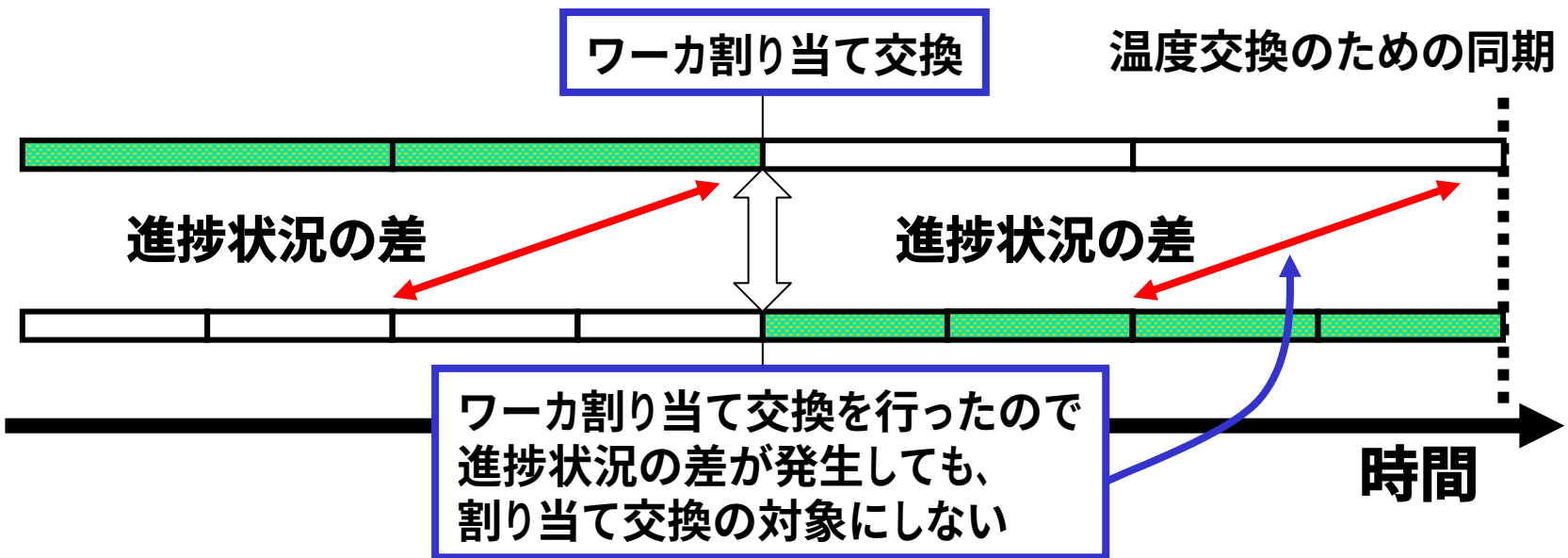
割り当て交換法の動作 (4/5)

- 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
  - T = (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



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、割り当て交換の対象にしない



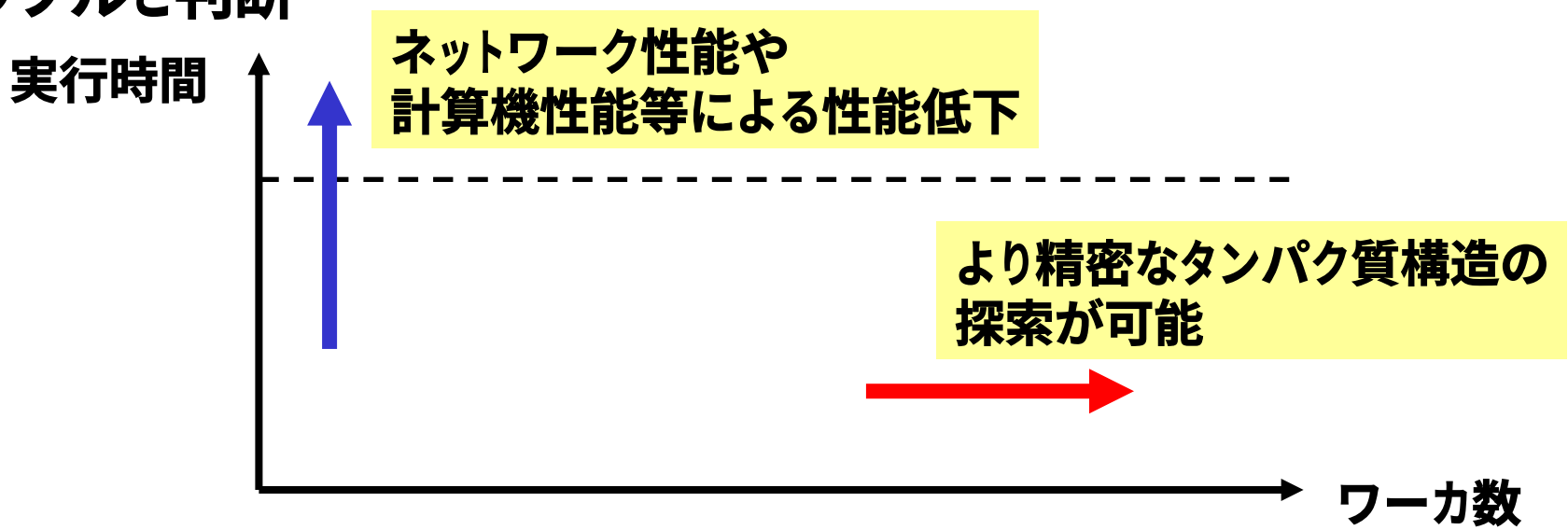
実験 (1/2)

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

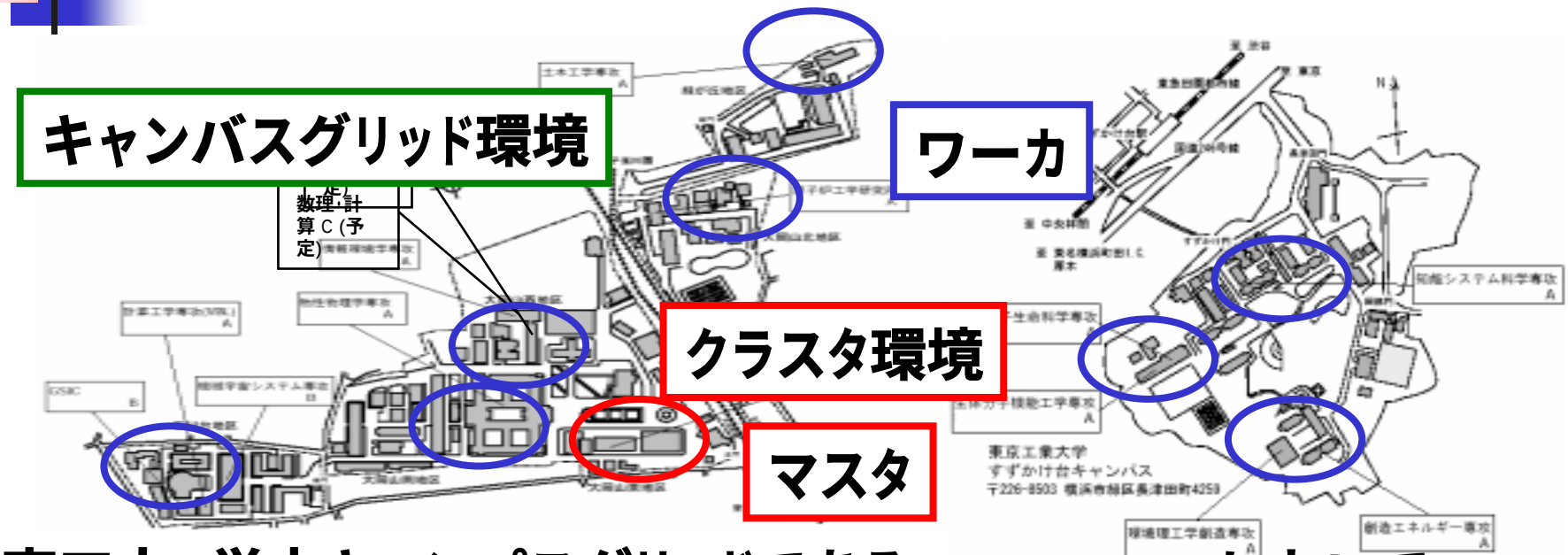
- ネットワーク性能ヘテロな環境
  - 東工大キャンパスグリッド上での実行
- 計算機性能ヘテロな環境
  - PCクラスタ上でワーカの一部に負荷をかけて実行
- グリッド環境
  - Globusを使用したMPI実装であるMPICH-G2を用いた実行

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低下が抑えられていればスケーラブルと判断



ネットワーク性能ヘテロである環境での評価 (1/5)



東工大　学内キャンパスグリッドである Titech Grid において

- 同一サイトのクラスタにマスタとワーカを割り当て実行 (クラスタ環境)
- 異なるサイトのクラスタにマスタとワーカを割り当て実行 (キャンパスグリッド環境)

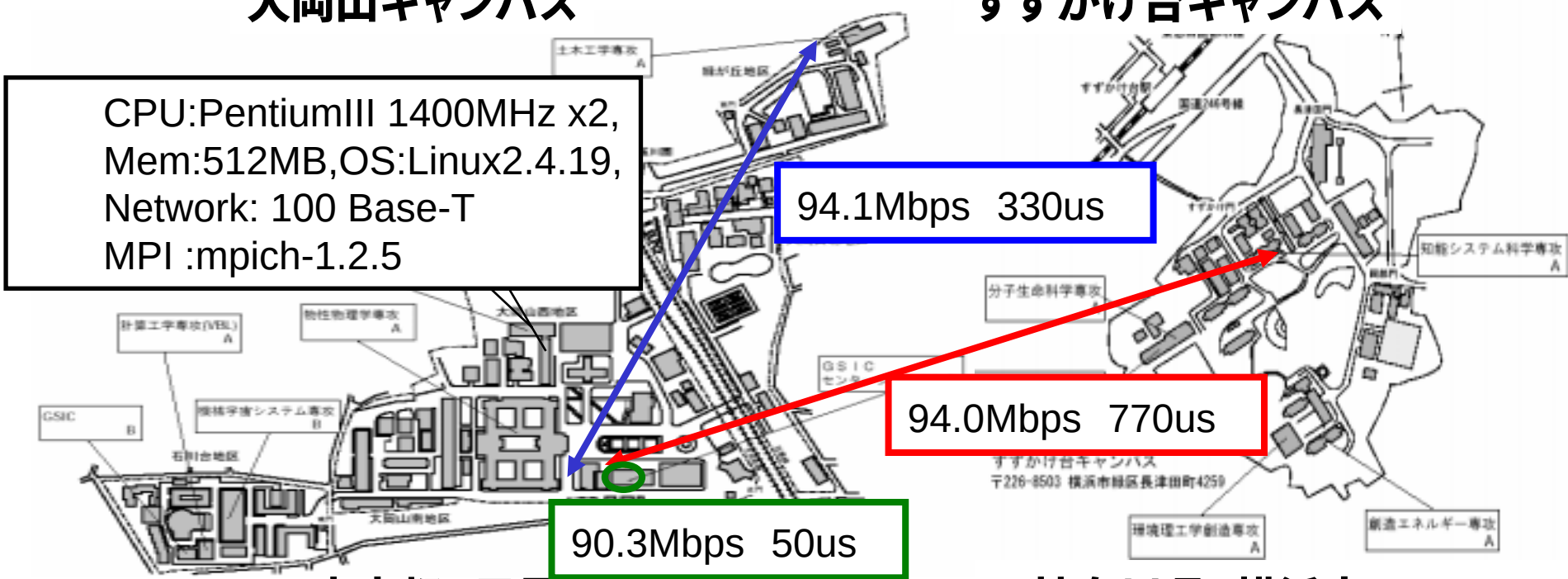
ネットワーク性能ヘテロである環境での評価 (2/5)

Network: クラスタ間 ↔ 90.3Mbps サイト間 ↔ 94.1Mbps ↔ 94.0Mbps
ping latency: クラスタ間 ↔ 50us サイト間 ↔ 330us ↔ 770us

大岡山キャンパス

すずかけ台キャンパス

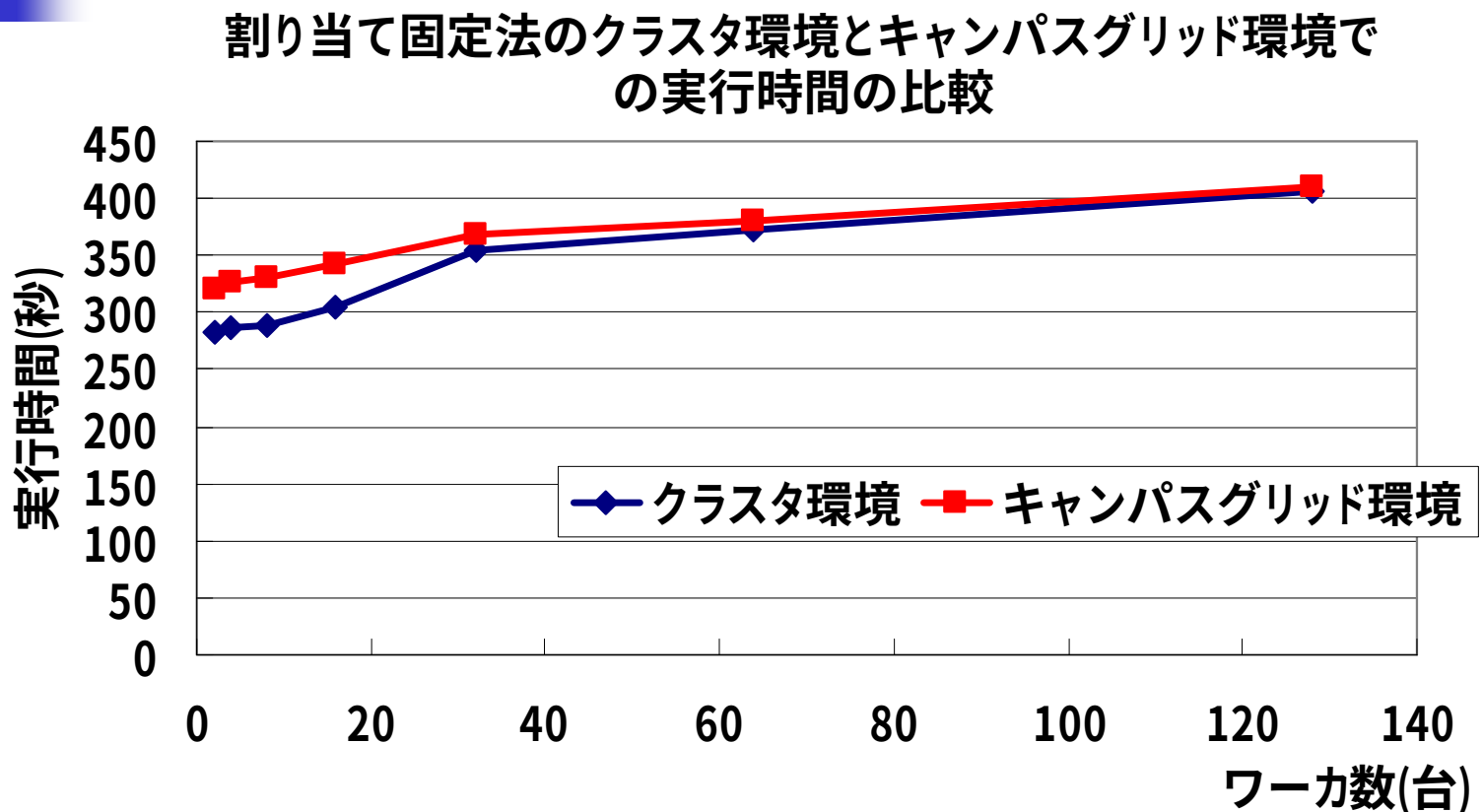
CPU:PentiumIII 1400MHz x2,
Mem:512MB,OS:Linux2.4.19,
Network: 100 Base-T
MPI :mpich-1.2.5



東京都 目黒区
(10 **サイト**)

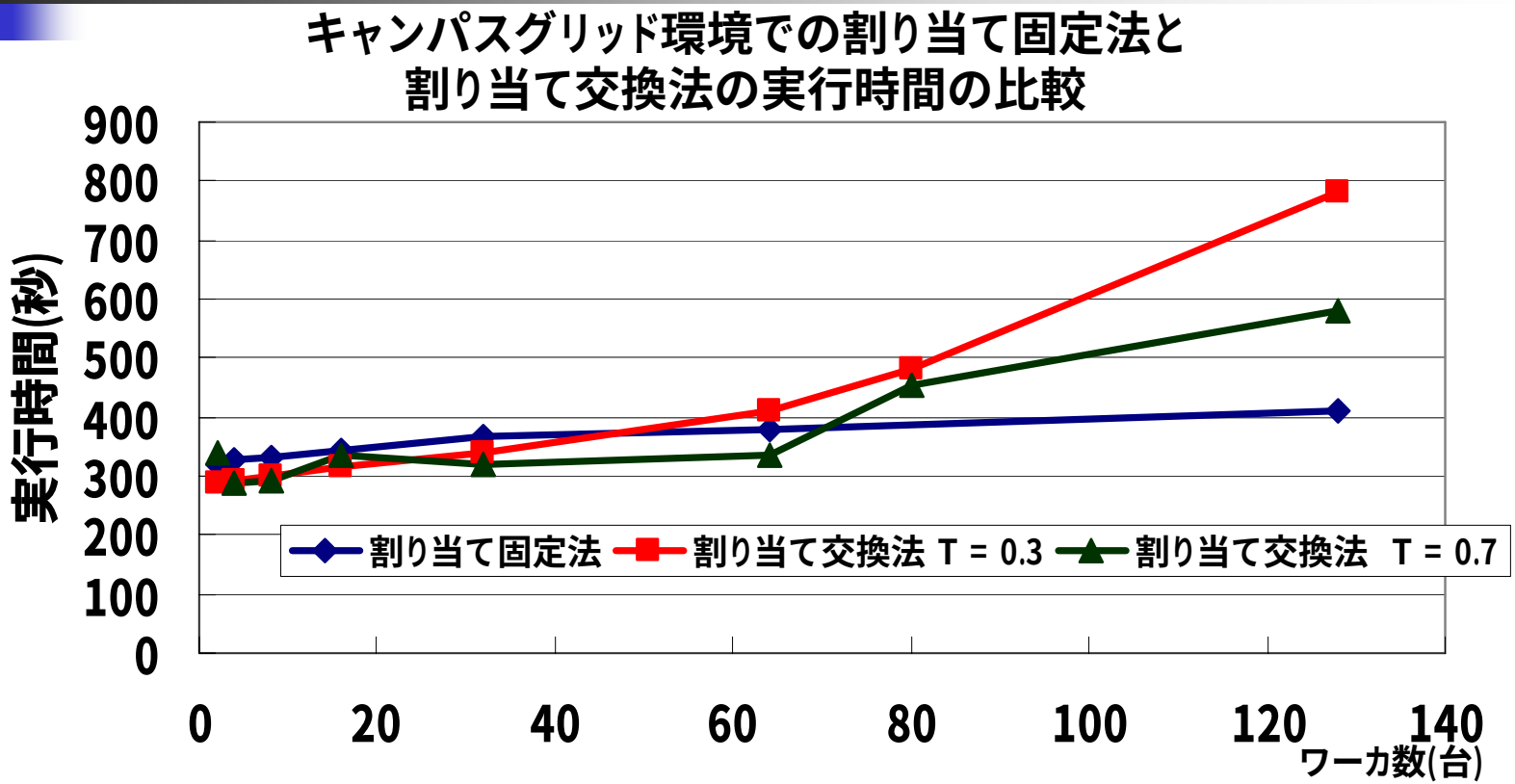
神奈川県 横浜市
(5 **サイト**)

ネットワーク性能ヘテロである環境での評価 (3/5)



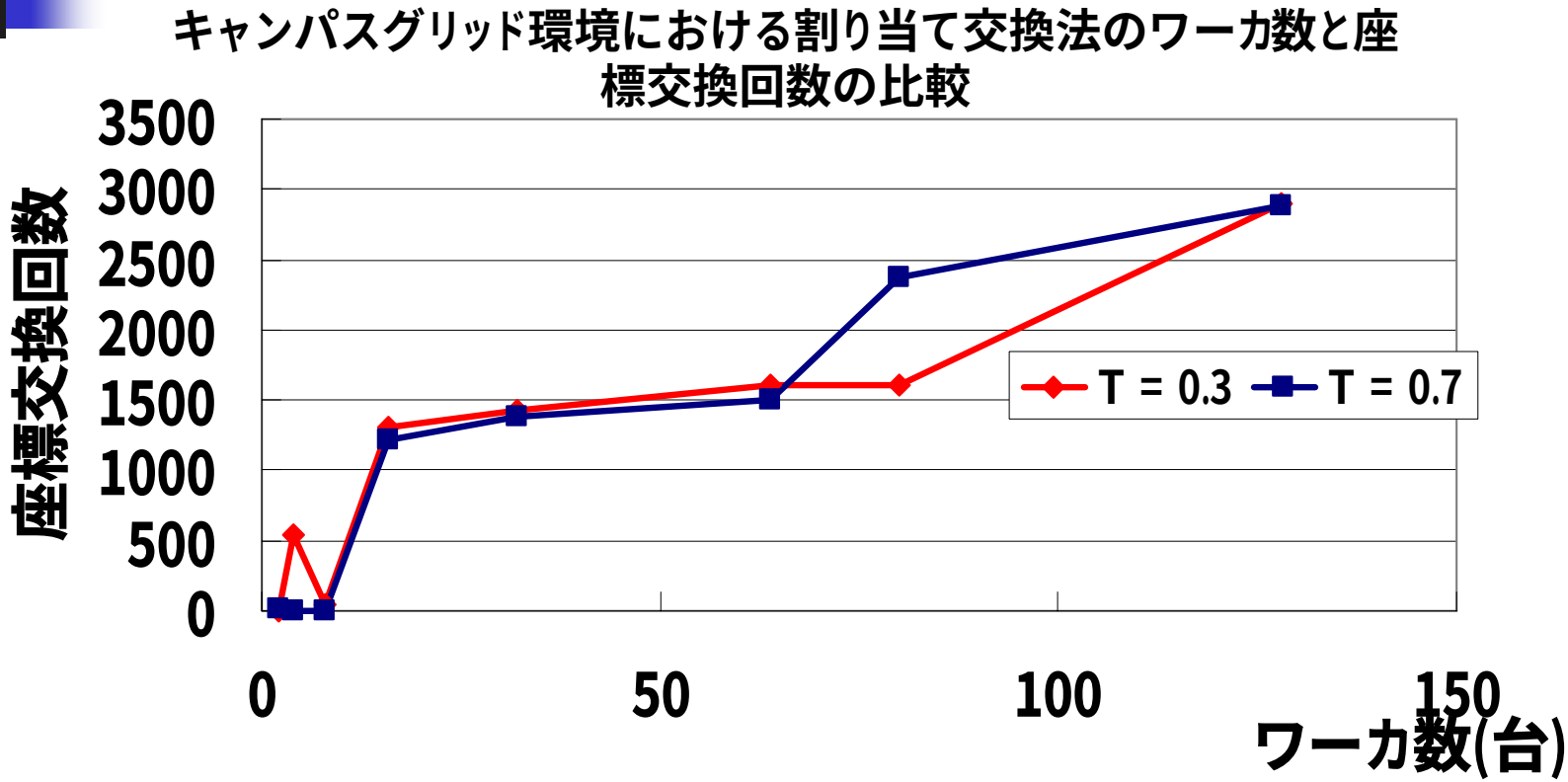
- ワーカ数を増加させた場合、実行時間の増加がみられるが、ネットワークによるボトルネックは小さい

ネットワーク性能ヘテロである環境での評価 (4/5)



- 割り当て交換法では、ワーカ数が増加するにつれワーカ間の座標交換回数が増加するため、実行時間の低下がみられる
- ここで、Tはワーカ割り当ての交換を決定するパラメタ
- T = (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)

ネットワーク性能ヘテロである環境での評価 (5/5)



- ワーカ数が増加するにつれ座標交換回数が大幅に増加する
- ここで Tはワーカ割り当ての交換を決定するパラメタ
- T = (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)

議論
ネットワーク性能ヘテロである環境での評価

- 割り当て固定法によるREMD Toolkitは100台規模までスケーラブルである
  - ワーカ数の増加に比べ実行時間の低下が抑えられている
- ワーカ数が増加するにつれ、割り当て交換法の性能が割り当て固定法に比べて大幅に低下
  - 割り当て交換を行うためのアルゴリズムやパラメータが未成熟
  - 不必要な座標交換の発生

計算機性能ヘテロである環境での評価 (1/2)

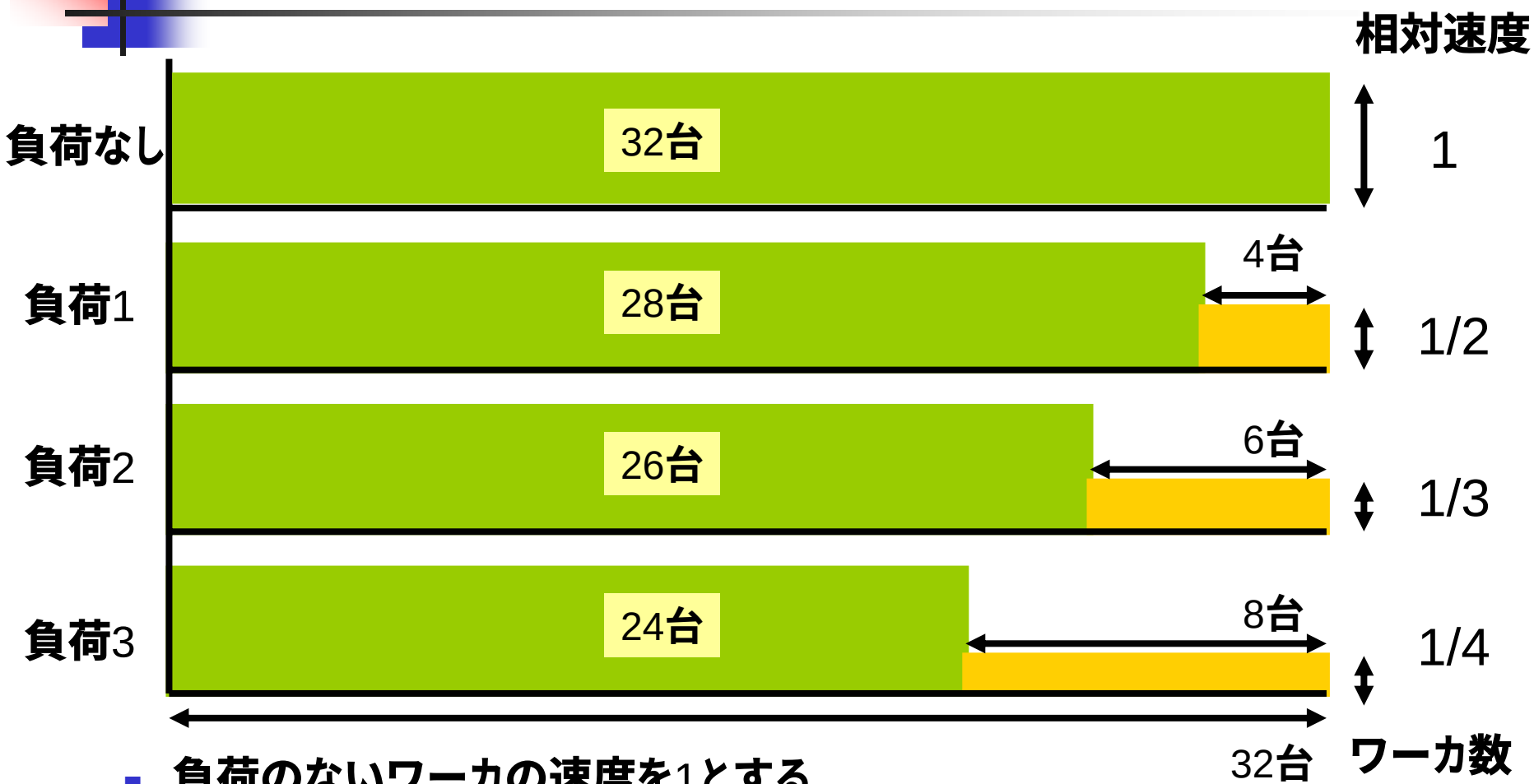
- 一部のPEに複数のワーカを割り当て、擬似的に計算機性能が低下したのと同等な環境を設定

	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	32	0	1
負荷1	28	4	1/2
負荷2	26	6	1/3
負荷3	24	8	1/4

ワーカ数32の場合

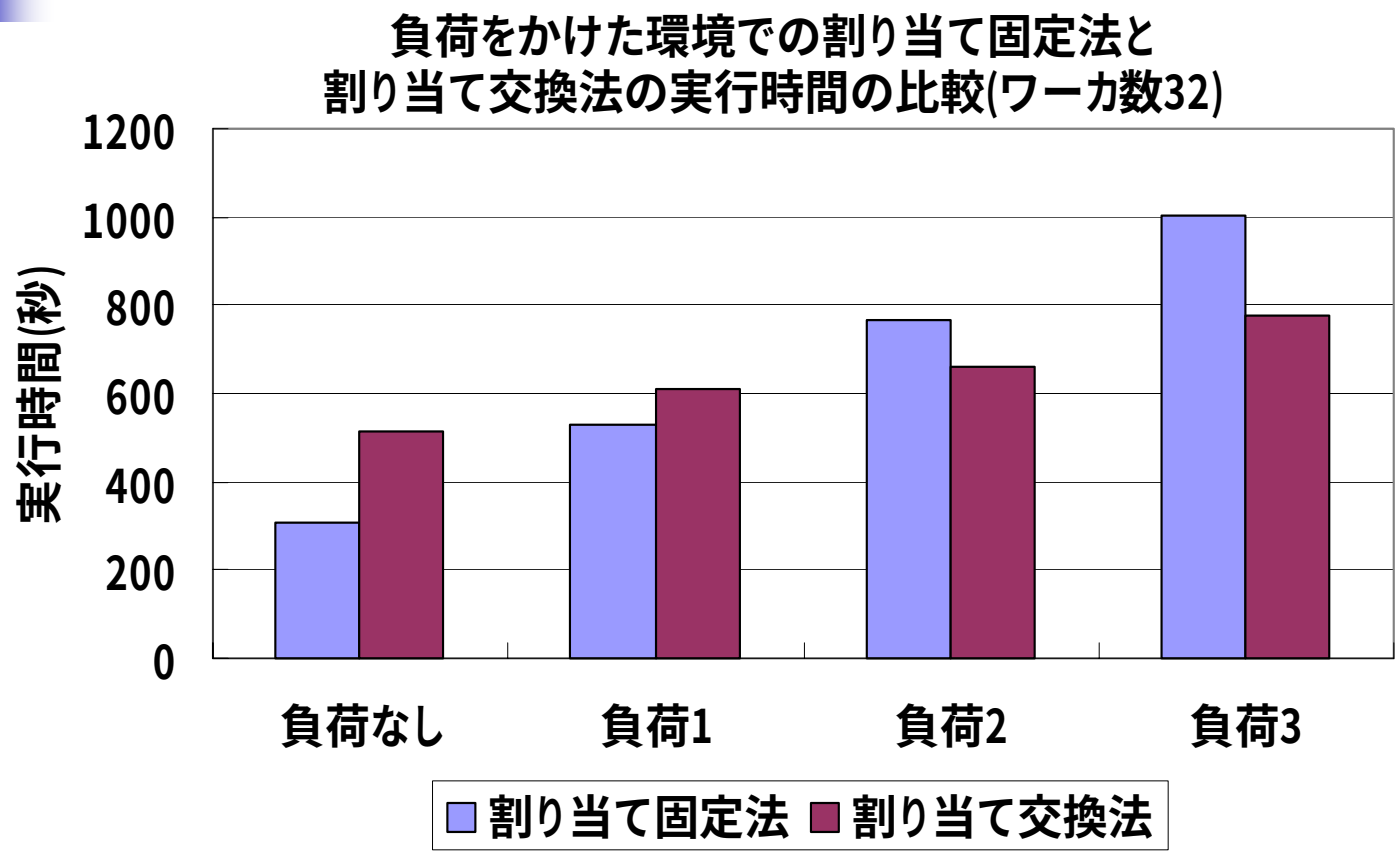
- 実行環境　東工大松岡研Presto III
  - CPU: Athlon MP 1900+ Memory: 768MB
  - Network: 100base-T OS: Linux 2.4.18
  - MPI：MPICH-1.2.5

計算機性能ヘテロである環境での評価 (2/4)



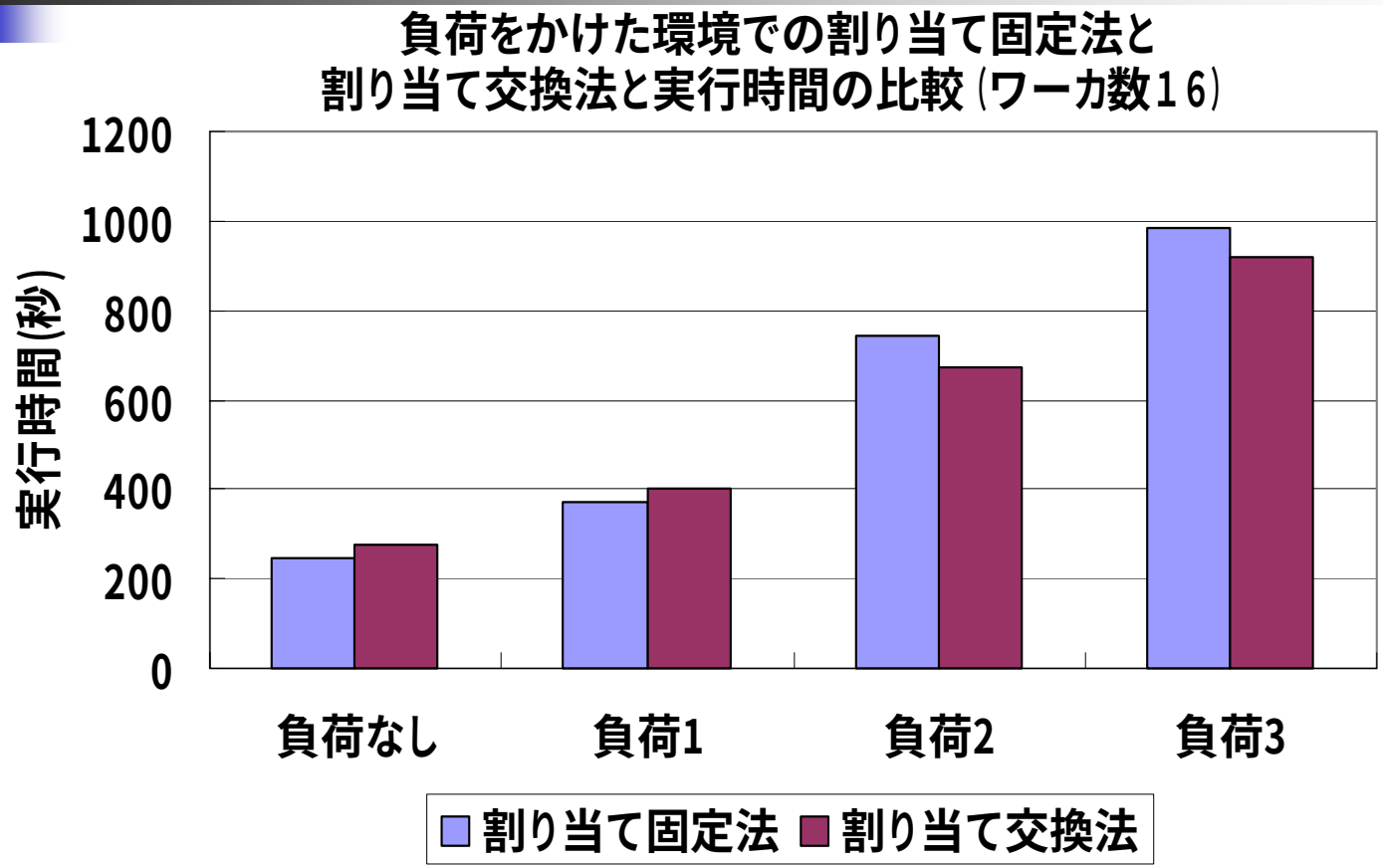
- 負荷のないワーカの速度を1とする
- 負荷1 2 3 では高速ワーカと低速ワーカを混合させて実行

計算機性能ヘテロである環境での評価 (3/4)



- 負荷の増加につれて割り当て交換法の性能低下が低く抑えられる

計算機性能ヘテロである環境での評価 (4/4)



- ワーカ数 32 の場合と同様の結果

議論
計算機性能ヘテロである環境での評価

- 負荷なし、負荷１では割り当て交換法は割り当て固定法に比べ性能低下
  - 割り当て交換を行うためのアルゴリズムやパラメータが未成熟であることによる不必要な座標交換
- 負荷2、負荷3のようなヘテロ性が高い場合では割り当て交換法は高速に動作
  - 割り当て交換法自体は有効

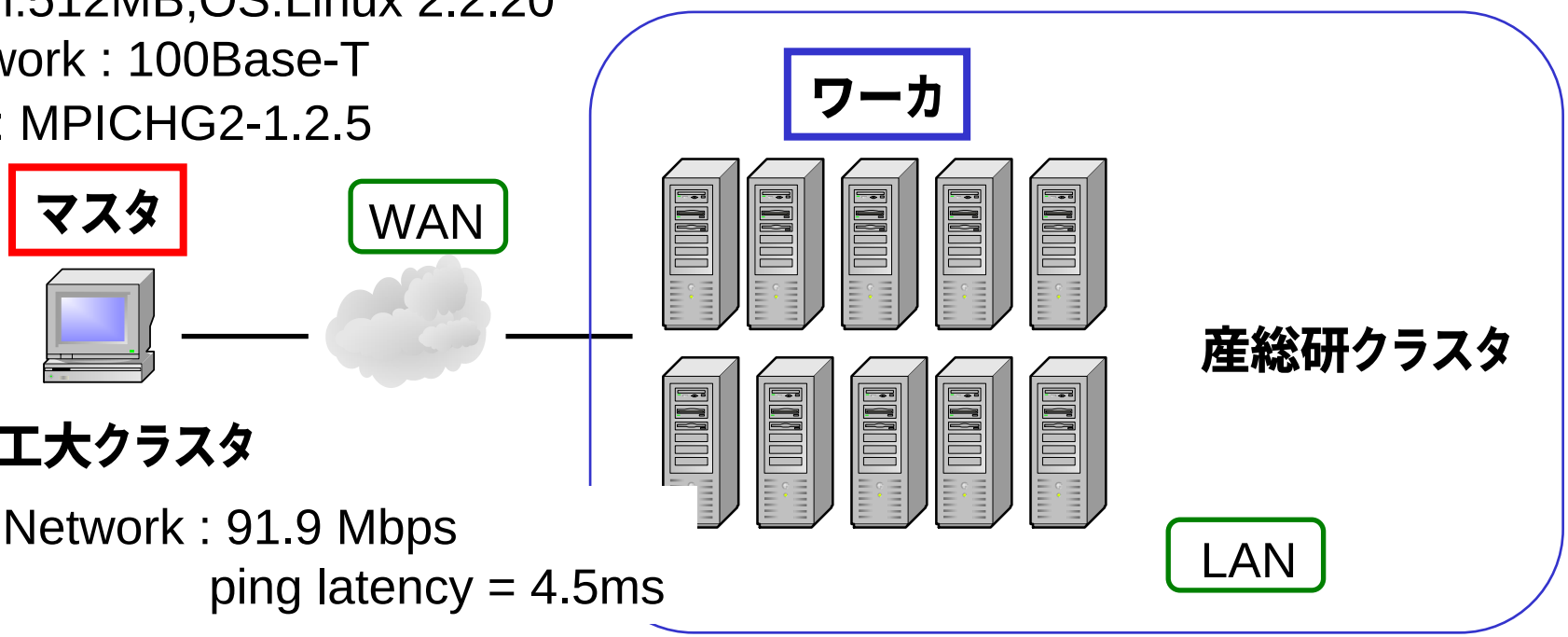
⇨ ワーカ交換のアルゴリズム　パラメータの改善

グリッド環境における実験（1）

- **実行環境**

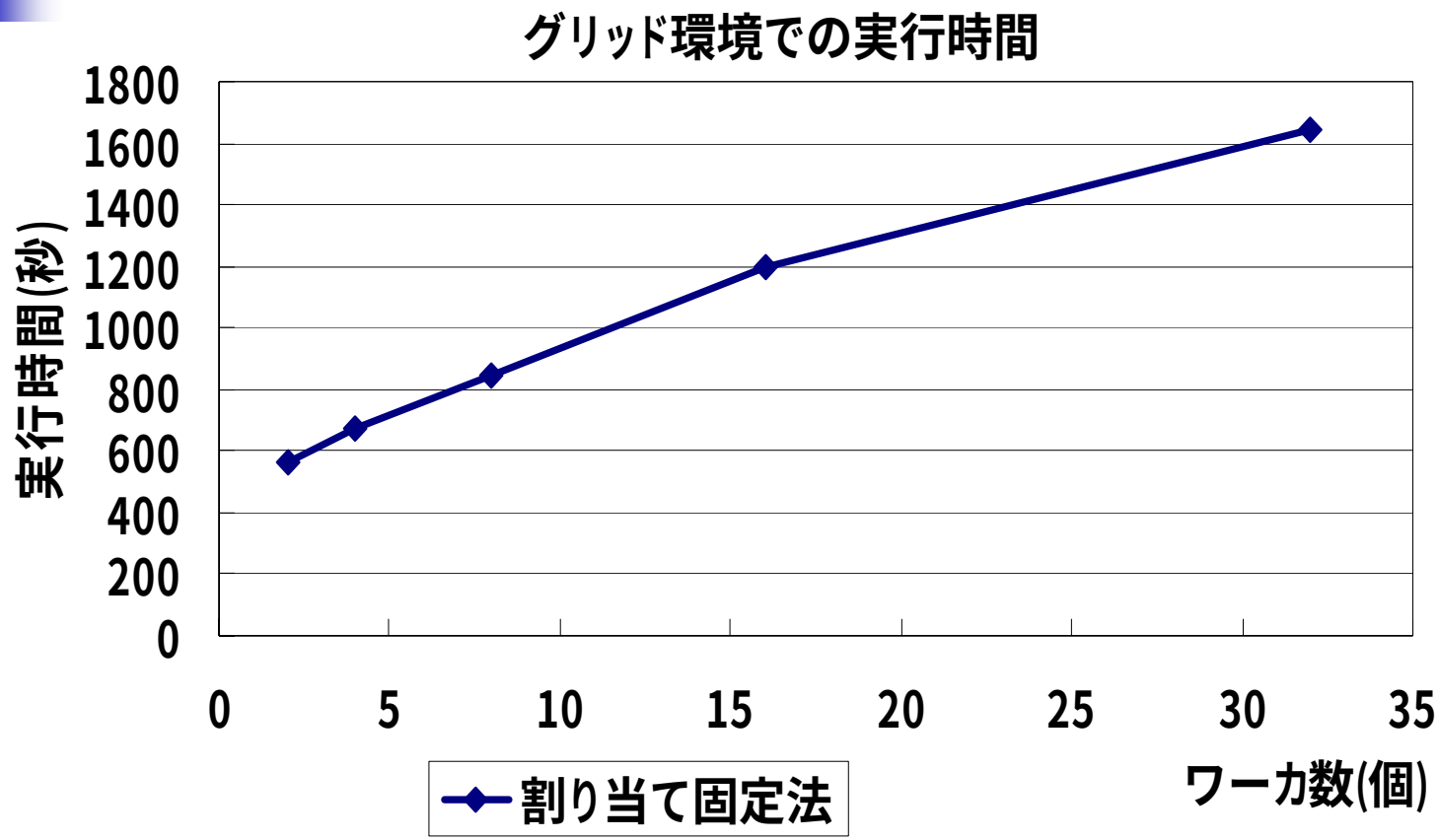
CPU:PentiumIII 1400MHz x2,
Mem:2.3GB,OS:Linux2.4.18,
Network : 1000Base-T

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



Network : 91.9 Mbps
ping latency = 4.5ms

グリッド環境での実験 （2）



- ワーカ数が増加するとともに実行時間が大幅に増加

議論
グリッド環境の実験

- 台数の増加に応じた実行速度の低下を観測
  - クラスタを占有できないため他のプロセスが走っていた可能性
  - 割り当て固定法のバリア同期により低速なプロセスに律速される

速度低下がグリッド環境のレイテンシだけとはいえない

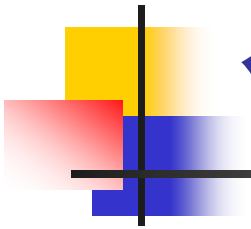
⇨ より精密でコントロールされた実験が必要

まとめ (1/2)

- 割り当て固定法によるREMD Toolkitは100台規模まで十分にスケーラブルである
- 割り当て固定法のグリッド環境の実験では台数の増加に応じた性能低下を確認
  - クラスタを占有できないため、他のプロセスが走っていた可能性
  - 割り当て固定法のバリア同期により低速なプロセスに律速される
  - よりコントロールされた実験が必要

まとめ (1/2)

- 割り当て交換法はいくつかの環境では性能が低下したものの、性能ヘテロな環境では効果がある
  - 負荷の増加につれて割り当て交換法が高速に動作することを確認
  - ヘテロ性の低い場合では、割り当て固定法が高速に動作
  - 割り当て交換法のアルゴリズムやパラメタの改善



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
- より制御された計算機性能ヘテロ環境での性能測定
- 高レイテンシなネットワークでの性能測定