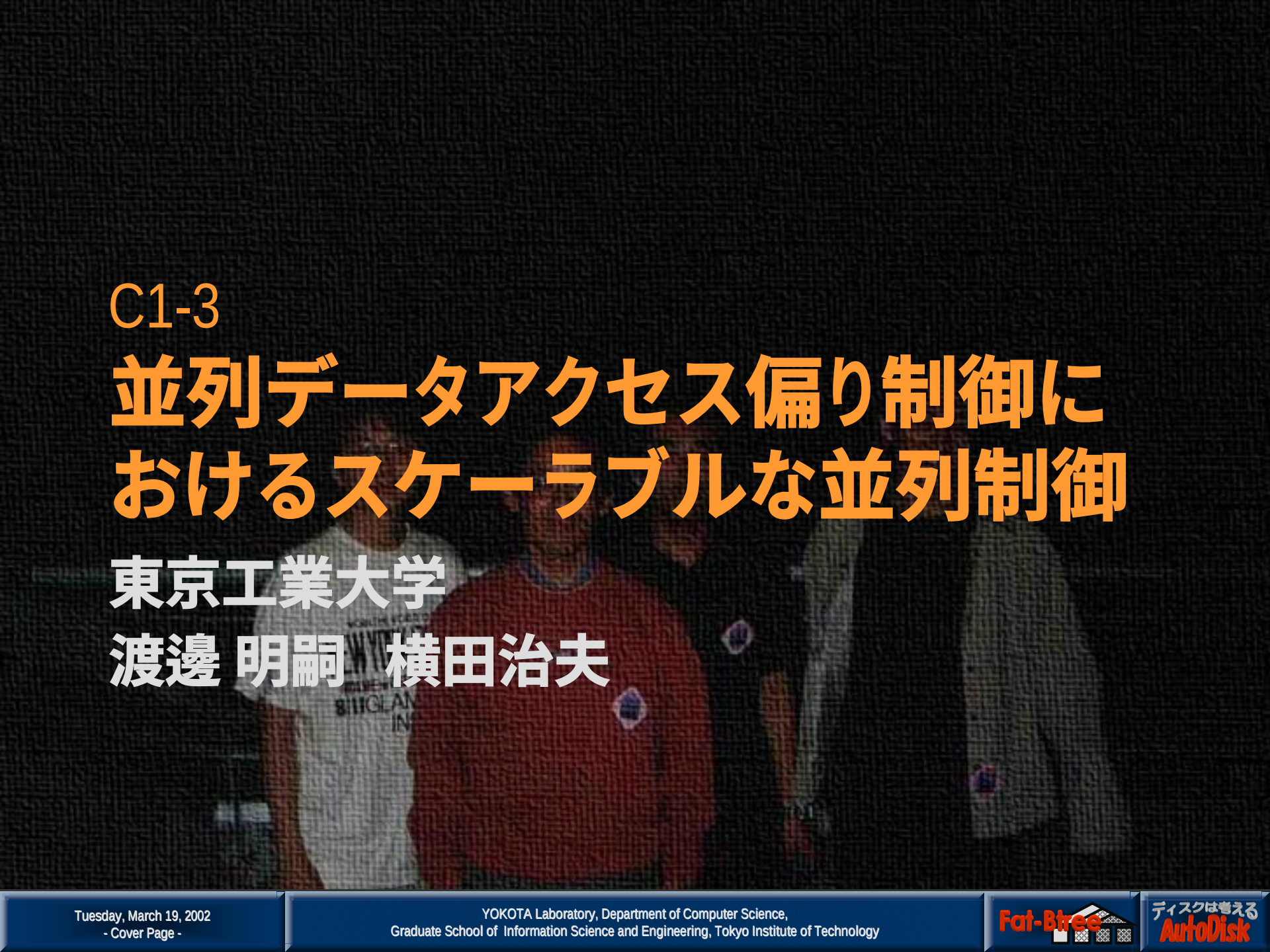


C1-3

並列データアクセス偏り制御におけるスケーラブルな並列制御

東京工業大学
渡邊 明嗣　横田治夫

概要

- 並列データアクセス偏り制御における並列制御手法の改良
  - *通信木を用いた並列度の向上* [DE 2001-110]
  - シミュレーションによる性能向上の検証
  - 適切な負荷評価手法の選択
  - 負荷評価の粒度が性能に与える影響の検証

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

C1-3: 並列データアクセス偏り制御におけるスケーラブルな並列制御

- 概要
- 本研究の背景
- TCSH — スケーラブルな並列制御
- シミュレーション実験と考察
- 結論

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

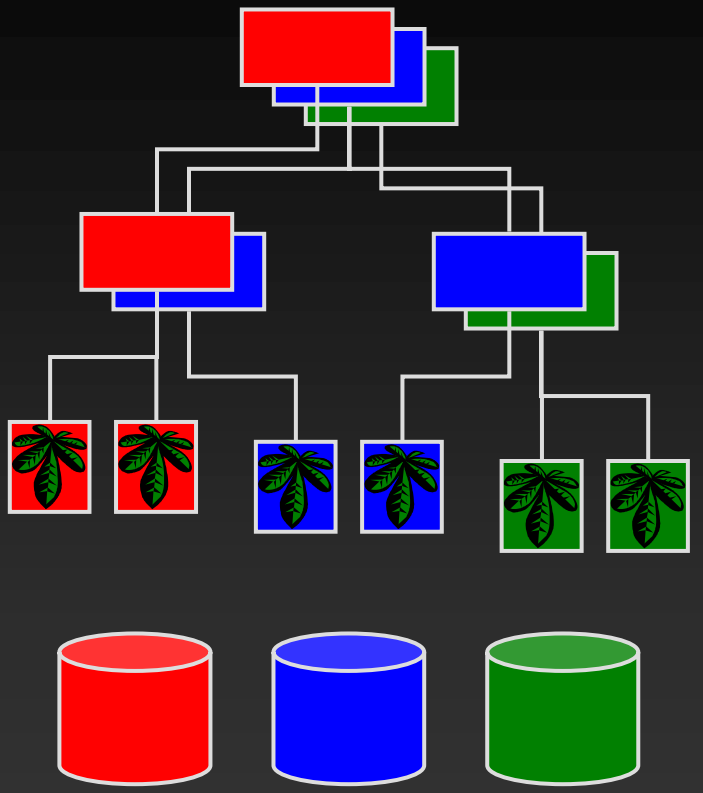
背景/
データの分散格納と偏り制御

- ☞ *データ量の増加と応答時間短縮の要求*
- 並列無共有システムを用いた
  データの分散格納
  - 高速化と安価な構成
  - ☞ *PE間の負荷偏りによる性能低下*
  - PE間の負荷を均等化する技術
    ～ **偏り制御の要求**
  - 動的なデータの再配置による負荷分散

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

背景/
偏り制御有りデータ分散格納に適した並列ディレクトリ構造

- Fat-Btree**並列ディレクトリ構造** [YOKOTA:1999]
  - **探索･更新が効率的**
  - **インデクス構造変化のコストが小さい** [YOKOTA:1999]
    - **データ移動コストが小さい**
    - ☞ **動的再配置のコストを小さくできる**
  - **葉ノードは全てのPEから到達可能**



YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

背景/ Fat-Btree偏り制御の並列制御手法

- Detect_Skew_&_Invoke_Migration [YOKOTA:1999], Full-window**並列制御** [Feelifl:1999Nov.]
  - **トークンベースの並列制御 (定数個のトークンがPEを巡回する並列制御)**
  - ✕ **PE数の増加とともに実行時間が *O(PE数)* で増加**
    - ☞ **大規模なシステムにおいては負荷が偏った状況の解消に長い時間が必要**

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

C1-3: 並列データアクセス偏り制御におけるスケーラブルな並列制御

- 概要
- 本研究の背景
- TCSH — スケーラブルな並列制御 [WATANABE, DE 2001-110]
- シミュレーション実験と考察
- 結論

TCSH — スケーラブルな並列制御/ 値域分割の性質を利用した通信量削減

- 値域分割
  - データは、隣接するPEにのみ移動できる
  - 隣接PEへ送る負荷が決定できれば十分
- 全PEの詳細な情報は不要

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

TCSH — スケーラブルな並列制御/ 通信量の削減

- 隣接するPEへ送るべき負荷の決定に必要な動的情報は以下の3つだけ
  - PE間の負荷平均: $L_{avg.}$
  - 右側にあるPE全ての負荷の合計: L_{Right}
  - 左側にあるPE全ての負荷の合計: L_{Left}
  - *静的情報: 右側にあるPEの数: $PE\#_{Right}$*
  - *静的情報: 左側にあるPEの数: $PE\#_{Left}$*

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

TCSH — スケーラブルな並列制御/ 移動すべき負荷

- 左 (右) 側のPEに移動すべき負荷 $\Delta L_{\{Left/Right\}}$ は下式で求められる
  - $\Delta L_{\{Left/Right\}} = L_{avg.} \times PE\#_{\{Left/Right\}} - L_{\{Left/Right\}}$
  - 負なら左 (右) 側のPEから負荷を受け取る
- 上式は、左 (右) 側のPE群の負荷合計が、偏りが無い場合の負荷合計に比べてどれだけ足りないかを表す

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

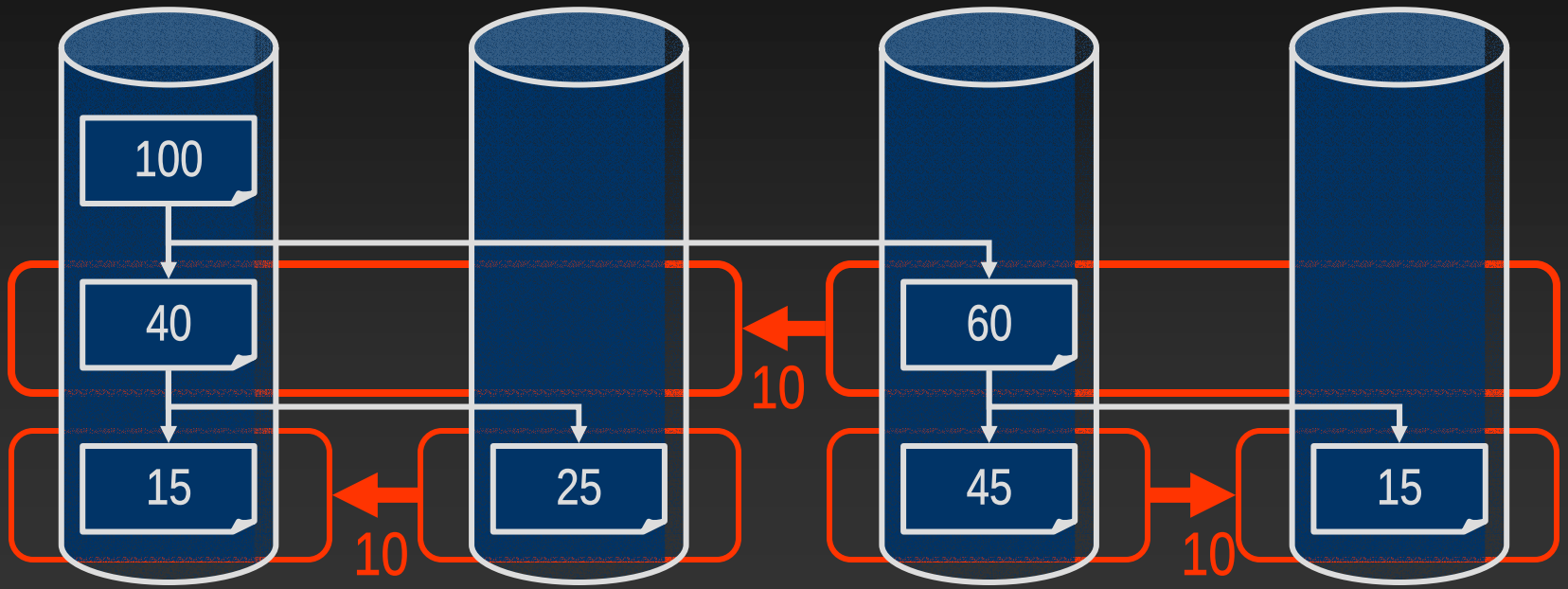
TCSH — スケーラブルな並列制御/ 通信木の利用

- ☞ *通信する情報の量を削減*
- 通信木の利用が可能
  - システムを再帰的に分割し、木構造を持ったネットワークを構築する
  - 偏り除去に必要な、負荷分布の取得、移動計画の伝達を *O(log PE数)* で行える
    - 同時処理を行うPE数を大幅に増やせる

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

TCSH — スケーラブルな並列制御/ 並列制御の仕組み

- 通信木を構築し、木の根に近いノードから順に、負荷の移動量を決定する



C1-3: 並列データアクセス偏り制御におけるスケーラブルな並列制御

- 概要
- 本研究の背景
- TCSH — スケーラブルな並列制御
- シミュレーション実験と考察
- 結論

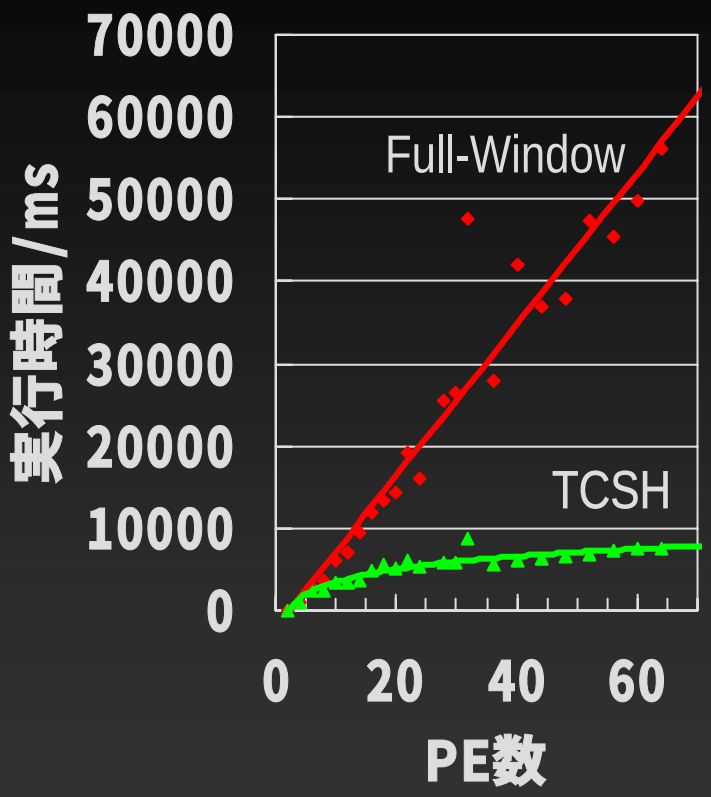
YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

シミュレーション実験と考察/ シミュレーション手法

- PE数による性能変化の測定
  - 最大負荷率が120%以下になるまで偏り除去を繰り返し、実行時間の合計を測定
  - ただし、偏り除去の間に行われたクエリ処理を除く
- 実験の設定と用語 (予稿集参照)
  - PE当りの葉ノード数=64K
  - キャッシュ: 128セット8連想度LRU
  - ディスクアクセス時間10ms、キャッシュ1 μs
  - speed parameter α = 0.25

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

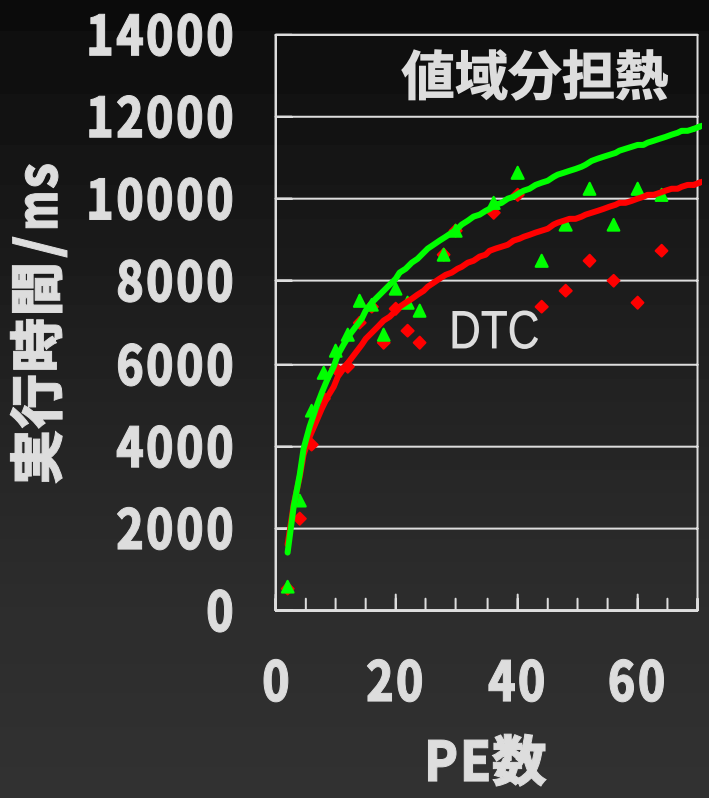
シミュレーション実験と考察/ PE数と実行時間



- グラフは初期偏り=zipf(0.5)の結果の一部 (DTCとの組み合わせ)
- シミュレーション結果は理論値にほぼ一致
  - Full-Window ~ O(PE#)
  - TCSH ~ O(log PE#)
- TCSHは大規模な場合の実行時間を短縮する

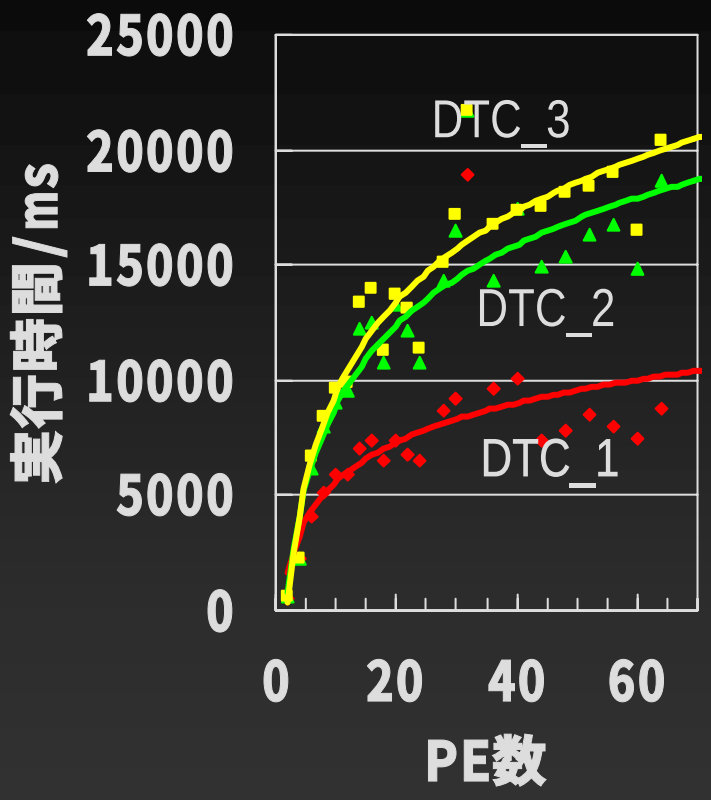
YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

シミュレーション実験と考察/ TCSHと負荷評価の組み合わせ



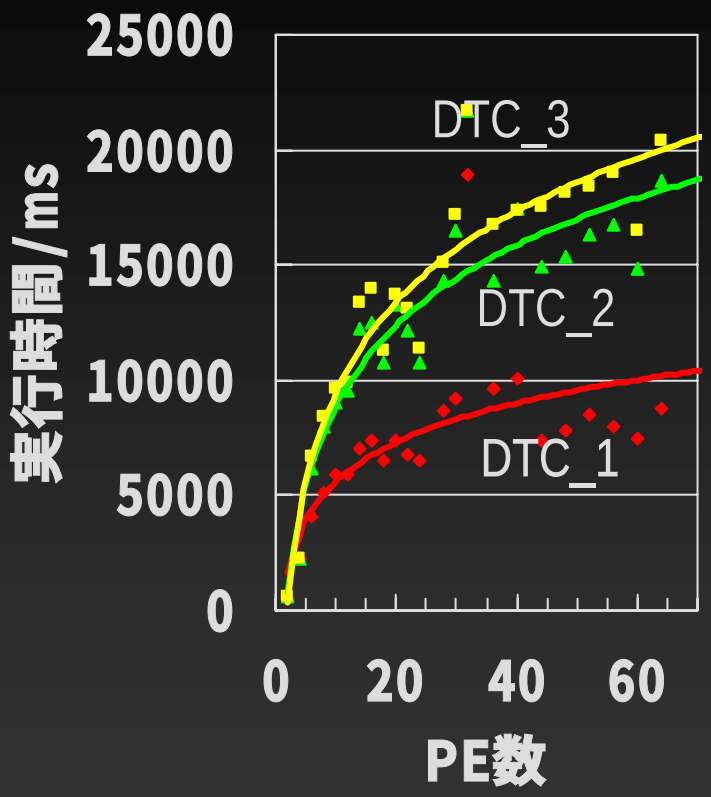
- DTC : アクセスコストを重視した負荷評価 [DE2001-82]
- 値域分担熱 : 値域へのアクセス頻度で評価する負荷評価 [Feelifl, 1999]
- グラフは初期偏り=zipf(0.9)の結果
- TCSHにはDTCが適する

シミュレーション実験と考察/ 粒度と実行時間



- DTCの粒度
  - DTC_1: 最も粗い
  - DTC_3: 細かい
- グラフは初期偏り=zipf(0.9) の結果の一部 (DTCとの組み合わせ)

シミュレーション実験と考察/ 粒度と実行時間



- 粗い粒度で好結果
  - 細粒度：実行時間 大
    - ○ ステップ数 小 [DE2001-82]
    - × 負荷計測コスト 大
  - 粗粒度：実行時間 小
    - × ステップ数 大 [DE2001-82]
    - ○ 負荷計測コスト 小
  - 負荷計測コストが大きな影響

C1-3: 並列データアクセス偏り制御におけるスケーラブルな並列制御

- 概要
- 本研究の背景
- TCSH — スケーラブルな並列制御
- シミュレーション実験と考察
- 結論

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

結論

- 本研究では、偏り制御の並列制御方式をシミュレーションを用いて検証･改善した
  - TCSH並列制御方式 [DE 2001-110] が理論的予測の O(log PE#) を達成していることを確認
    ⇒ TCSHのスケーラビリティーを確認
  - シミュレーション結果から、TCSHに適した負荷評価として DTC_1 [DE2001-82] を選択

YOKOTA Laboratory, Department of Computer Science,
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

今後の研究

- より詳細なシミュレーション実験による検証
  - 条件を変えた計測
- 偏り制御による平常動作への影響の軽減
  - 一度に移動するデータ量を制限する案
  - 高精度負荷評価の低コスト実現策の研究
- 負荷評価の実装手法の検討