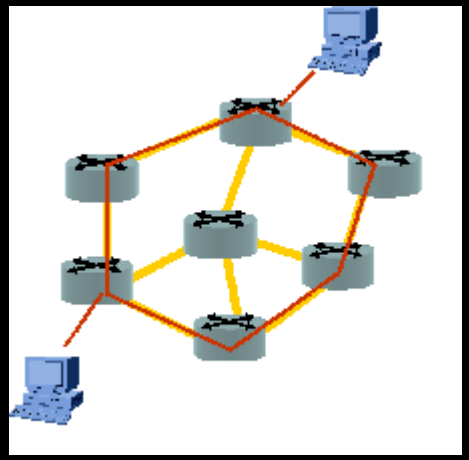
λコンピューティング環境構築のための共有メモリシステムの評価

大阪大学 基礎工学部情報科学科
ソフトウェア科学コース4年　宮原研究室
谷口英二

e-tanigu@ics.es.osaka-u.ac.jp

研究の背景

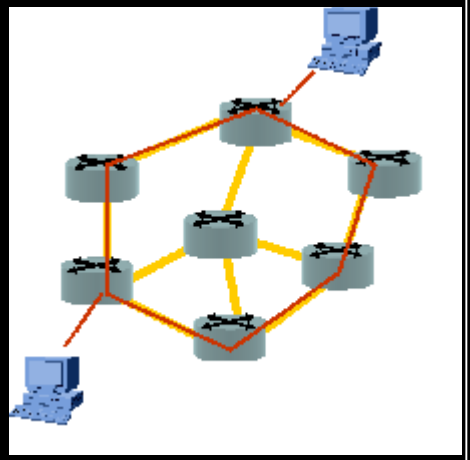
- ネットワークを用いた分散計算環境としてグリッド技術がある
  - TCP/IPによるパケット交換を用いている
    - 転送確認処理のオーバヘッド
    - 損失処理の転送レートの劣化



- λコンピューティング環境
  - 計算機と接続しているネットワークを仮想的な光リングネットワークとして利用
  - 光リングを利用して高速・高品質通信の実現の可能性

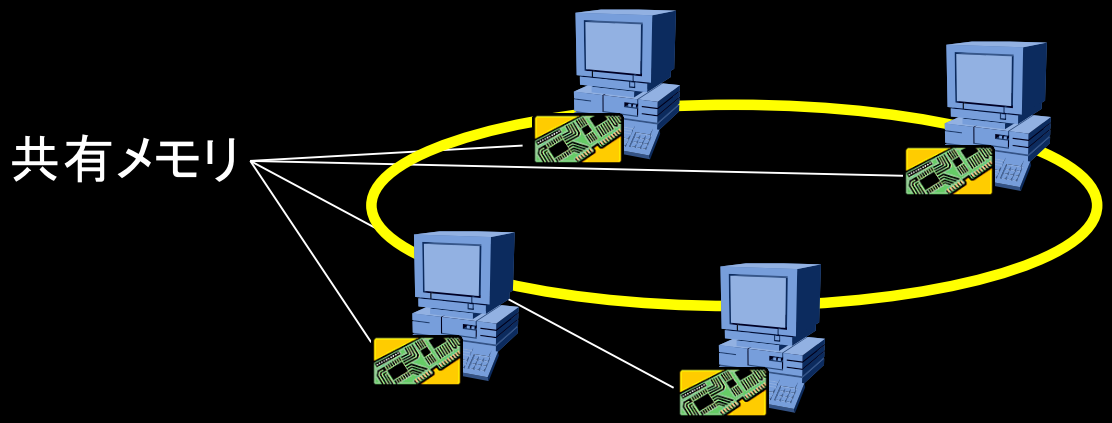
λコンピューティング環境における分散計算システム

- 共有メモリ型システム
  - 光リングを各ノードの共有メモリとして用いる
- 高速チャネル型システム
  - 光リングを高速伝送路として用いる
  - 各ノードにデータを共有する領域を設ける
- 光リングに適合する共有メモリ方式が必要となる



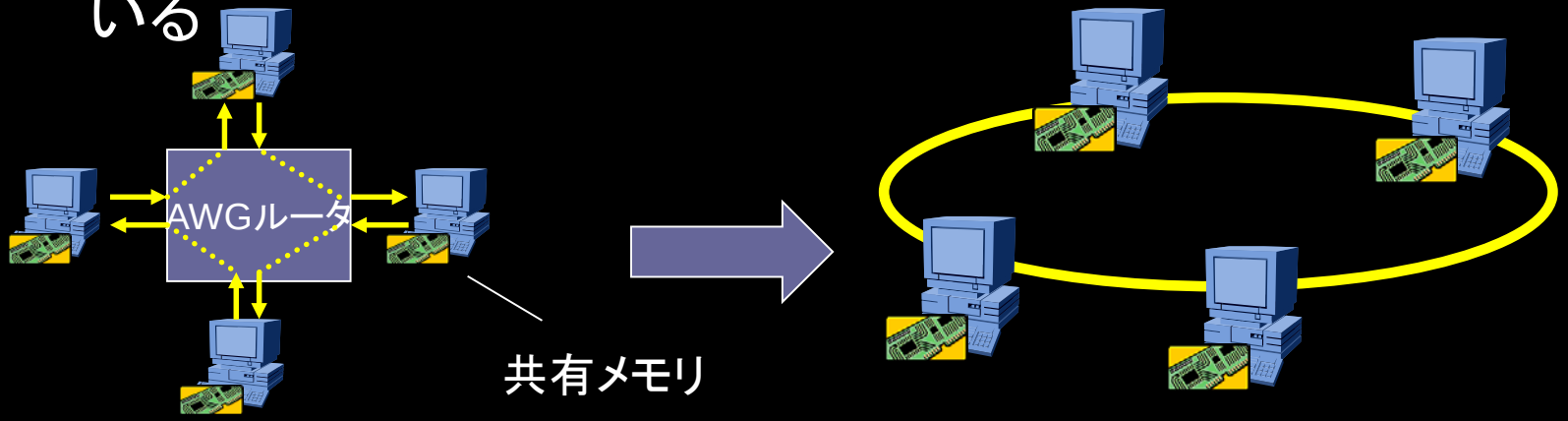
研究の目的

- 高速チャネル型システムを対象とした共有メモリ方式の評価
  - 具体的なシステムとしてAWGーSTARを使用する

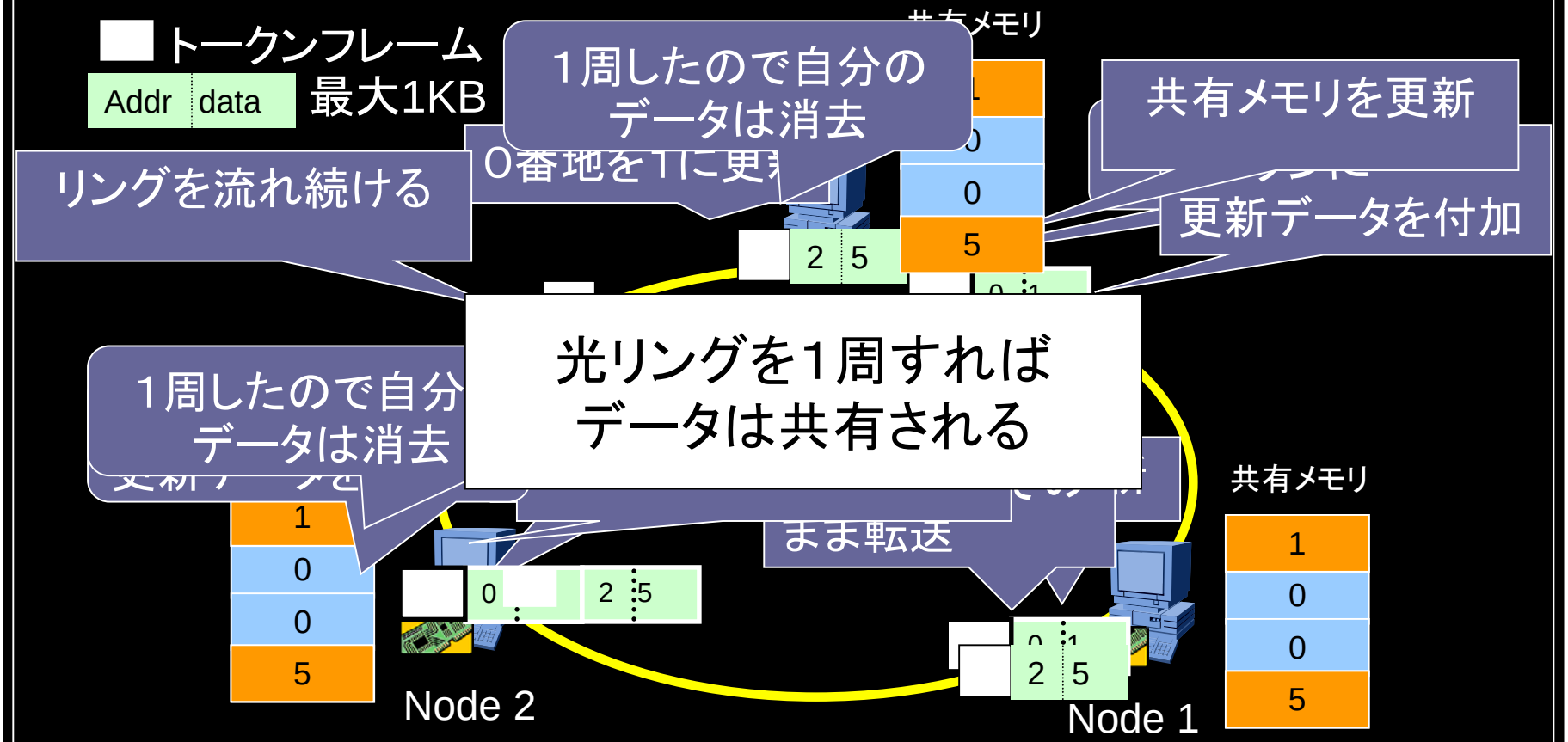


AWGーSTAR　システム構成

- 各ノードを光ファイバを用いて波長ルータ (AWG) に接続し光リングを構成
- 各ノードの共有メモリを光リングを通じて共有
  - 共有メモリを分散計算に必要なデータ共有手段として用いる

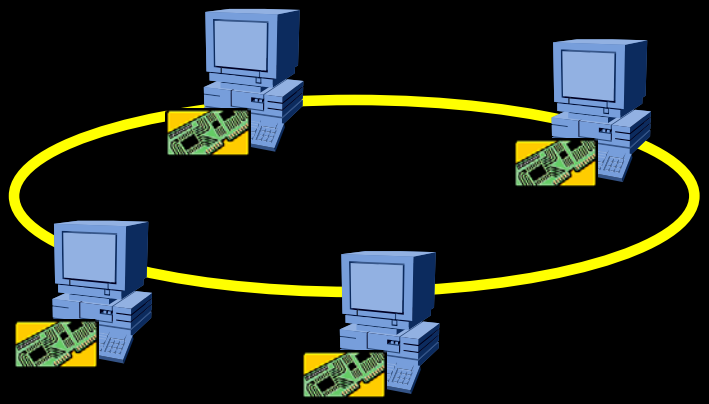


AWG-STAR　データ共有手法



AWG－STAR　処理遅延

- AWG－STARにおける処理遅延
  - 光ファイバによる伝播遅延: 5 ns/m
  - 各ノードにおける処理遅延: 500 ns
    送信フレームの削除と追加、共有メモリへの反映



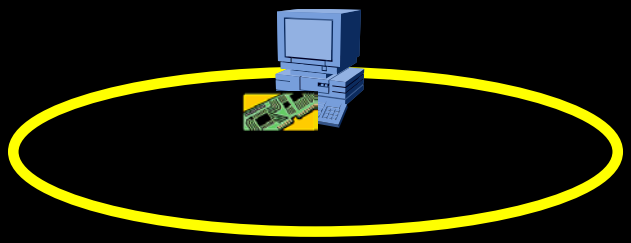
リング長を40 m、ノード数を4とすると

1周に要する時間は

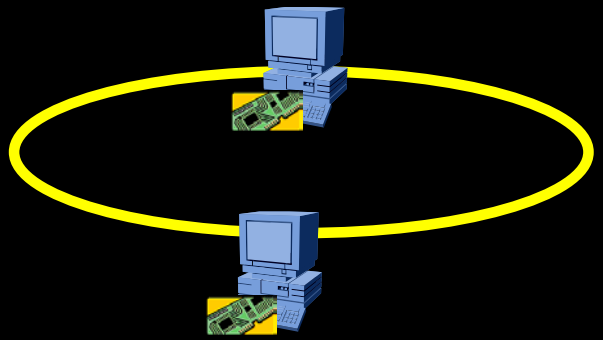
500×4＋5×40 =2200 ns

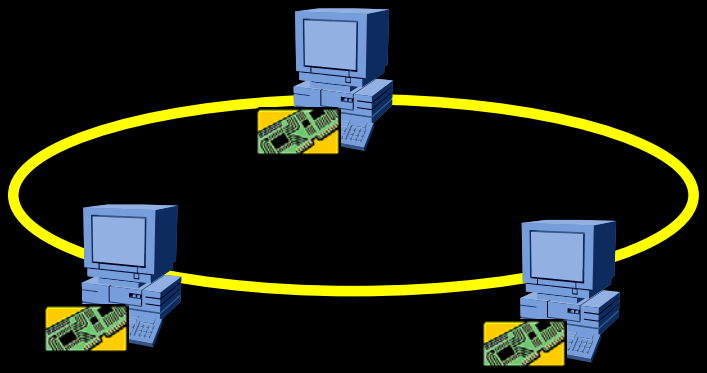
実験環境（システム構成）

ノード数1 リング長 10m 遅延 550ns



ノード数2 リング長 20m　遅延 1100 ns



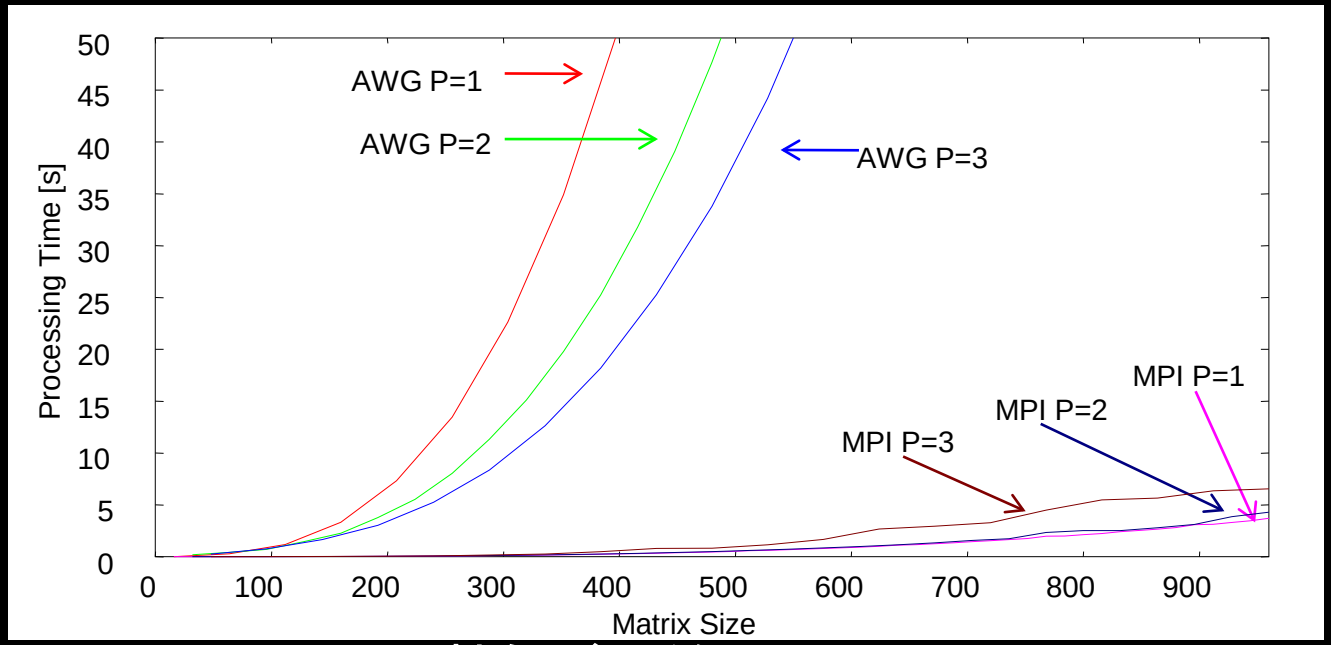


ノード数3 リング長 30m 遅延 1650ns

実験環境（アプリケーション）

- SPLASH2（分散計算用ベンチマーク集）
  - LU分解
    - 共有メモリへのアクセスが多い
- MPI (Message Passing Interface)によるTCP/IPとの比較

LU分解による実行結果



AWG－STARの性能がよくない

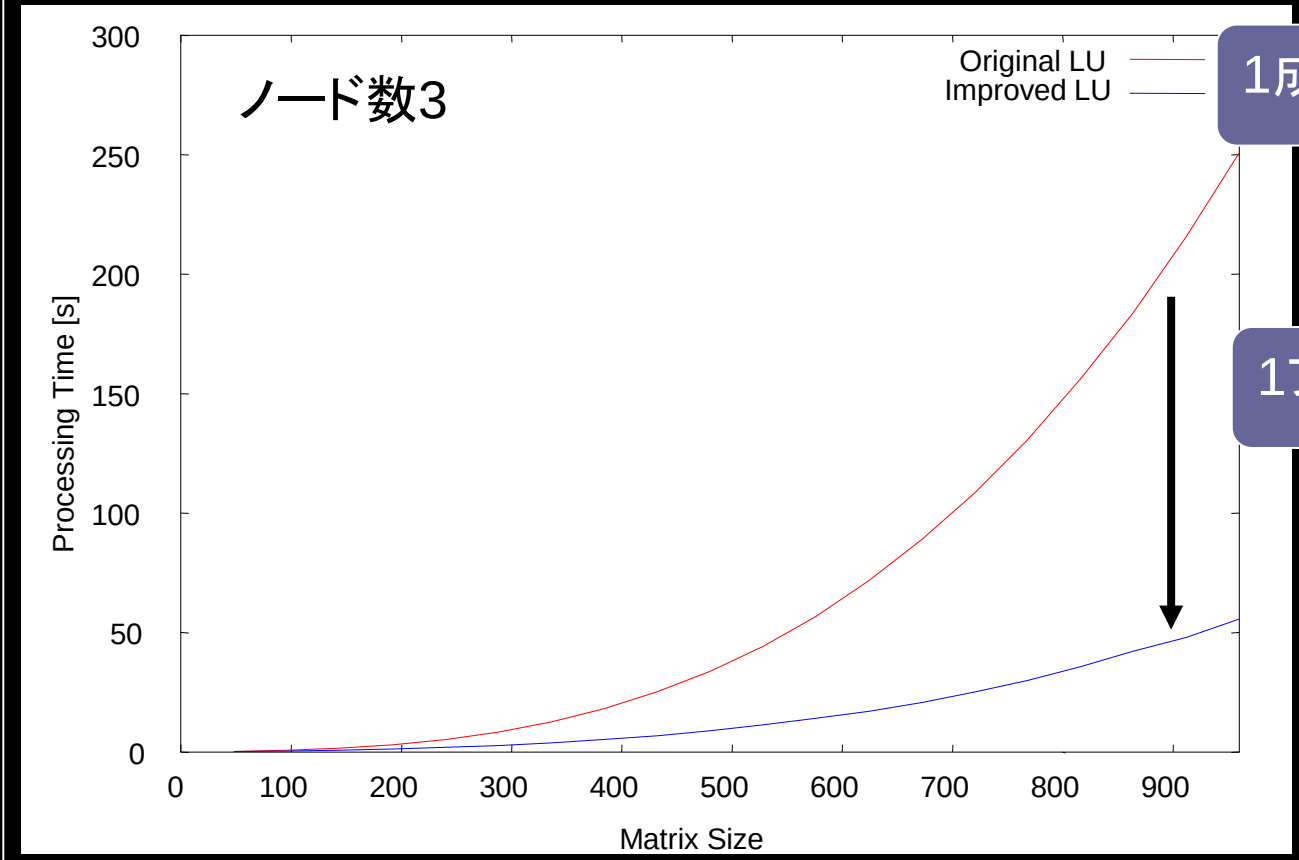
- MPIによるデータの共有のための通信に関しては並列化ができない
- さらにノード数が増えるため通信量が増加する

- 共有の必要のないデータの共有に関しての書き込み
- 全ノードでの通信が発生
- 書き込み回数が増加
- この書き込みによる遅延が影響

共有メモリアクセス方式の改善

- 共有メモリへの書き込み回数が性能に影響
  - 書き込み回数に応じて周回数が増えるため実行時間が増大
  - AWG－STARではトークンを利用するためにトークンの待ち時間が必要
- 改善方法
  - ローカルメモリを活用し共有の必要のないデータの共有メモリへの書き込みは行わない
  - データをまとめて書き込むことで書き込み回数を削減する
  - 書き込み後、即座にデータの周回を開始するようにする
    - AWG－STAR ではハードウェアにより制約される

共有メモリへの書き込み回数を削減した場合



1成分毎の書き込み

1ブロック毎の書き込み

実行時間を
約20%に短縮

まとめと今後の課題

- 共有メモリ方式の性能評価
  - 共有メモリへの書き込み回数
  - 光リングの周回時間
- 性能改善方法を検討
- 今後の課題
  - 効率のよい共有メモリアクセス手法の考案
  - 他のアプリケーションによる比較