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実装と評価

大阪大学 大学院情報科学研究科
博士前期課程1年 谷口英二

発表内容

- 研究の背景
  - λコンピューティング環境
- 研究の目的
  - 共有メモリシステムの実装と評価
- 共有メモリシステムとメモリアクセス手法
  - AWG-STARシステム
- 共有メモリシステムの性能評価
- まとめと今後の課題

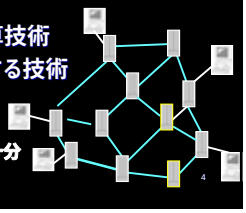
研究の背景

- グリッドコンピューティング環境
  - ノード間のデータの送受信は主にTCP/IP
    - 転送確認処理のオーバヘッド
    - 損失処理による転送レートの劣化
- λコンピューティング環境
  - 高速・高品質通信路をノード間に提供可能
  - フォトニックネットワーク上で分散計算を行う環境

グリッドコンピューティング環境

- グリッドコンピューティング環境では
  - ネットワークを介し、複数のノードを接続し計算資源・ストレージを共有
  - ノード間の通信は主にTCP/IP
- 必要となる技術
  - 広域・大規模計算の分散計算技術
  - 大容量データを高速に送受する技術

TCP/IPではパケット処理などのため不十分

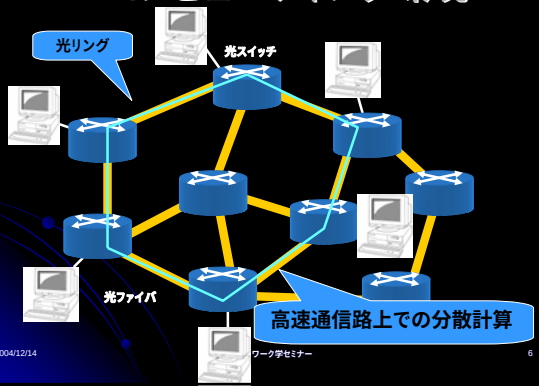


λコンピューティング環境

- λコンピューティング環境では
  - 各ノードやルータを光ファイバで接続
  - ノード間の通信は波長パスを利用
    - パケット通信とは異なりデータ損失が少ない
  - ハードウェアレベルでの高速なデータ共有
    - 明示的なデータ転送は行わない
- 高速・高品質な通信路をノード間に提供可能

→広域・大規模の分散計算に適用

λコンピューティング環境

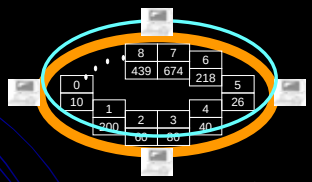


λコンピューティング環境におけるアーキテクチャ

- λコンピューティング環境におけるアーキテクチャ
  - 共有メモリ型アーキテクチャ
  - 高速チャネル型アーキテクチャ
- 共有メモリ型アーキテクチャ
  - 光リングを共有メモリとして利用する
- 高速チャネル型アーキテクチャ
  - 光リングを通信用の高速チャネルとして利用する

λコンピューティング環境におけるアーキテクチャ

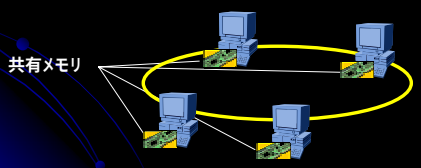
- 共有メモリ型アーキテクチャ [3]
  - 光リング自体を、接続している全ノードの共有メモリとして利用する



[3] 中本博久, 馬場健一, 村田正幸, "λコンピューティング環境における共有メモリアクセス手法の提案", 電子情報通信学会技術研究報告, 第104巻, 81号, pp.43-48, 2004

λコンピューティング環境におけるアーキテクチャ

- 高速チャネル型アーキテクチャ
  - データを共有する領域を各ノードに設ける
  - 光リングを専用の通信路として利用する

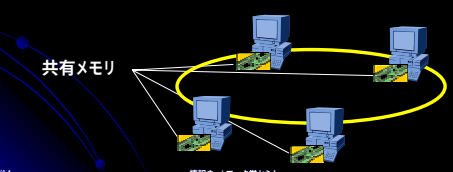


発表内容

- 研究の背景
  - λコンピューティング環境
- 研究の目的
  - 共有メモリシステムの実装と評価
- 共有メモリシステムとメモリアクセス手法
  - AWG-STARシステム
- 共有メモリシステムの性能評価
- まとめと今後の課題

研究の目的

- 高速チャネル型アーキテクチャを対象とした共有メモリ方式の実装と評価
  - 具体的なシステムとしてNTTフォトニクス研究所が開発したAWG-STARシステム利用



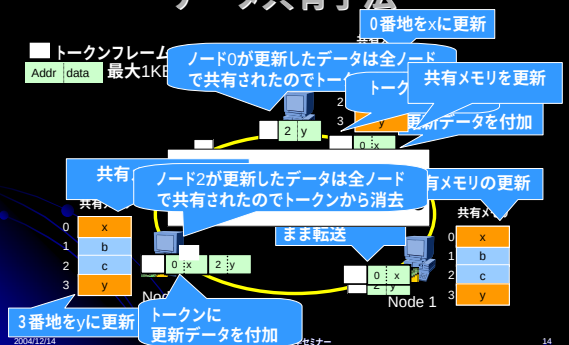
発表内容

- 研究の背景
  - λコンピューティング環境
- 研究の目的
  - 共有メモリシステムの実装と評価
- 共有メモリシステムとメモリアクセス手法
  - AWG-STARシステム
- 共有メモリシステムの性能評価
- まとめと今後の課題

AWG-STARシステム

- 各ノードは波長ルータ(AWG)に接続し光リングネットワークを構成
- 各ノードは共有メモリボードに共有メモリを搭載
  - 光リング上の全ノードの共有メモリが同じアドレスに同じデータを保持

AWG-STARシステム データ共有手法



AWG-STARシステムの処理遅延

- 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
  - 共有メモリへのデータの反映
- 共有メモリボードへのPCIバスの遅延時間
  - ローカルメモリへのアクセスよりも時間を要する
- 光リングへのアクセス遅延時間
  - 制御トークンの待ち時間

AWG-STARを用いたλコンピューティング環境の構築

- 分散計算に必要となる機能はない
  - AWG-STARはサイズの大きいデータをリアルタイムで共有するためのシステム
- 分散計算に必要な機能の設計および実装
  - 同期機構
  - 共有変数のための領域確保機能
  - アプリケーションをAWG-STARに適用するためのコード修正

発表内容

- 研究の背景
  - λコンピューティング環境
- 研究の目的
  - 共有メモリシステムの実装と評価
- 共有メモリシステムとメモリアクセス手法
  - AWG-STARシステム
- 共有メモリシステムの性能評価
- まとめと今後の課題

実験環境：システム

- 共有メモリボードの仕様
  - 伝送速度：2 Gbps
  - 1回あたりの最大データ転送サイズ：1 KB
  - 共有メモリへの書き込み：最大60MB/s
  - 共有メモリからの読み出し：最大67MB/s

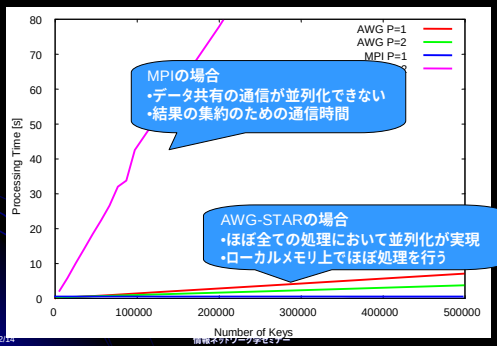
ノード数3 リング長 60m 1800 ns/周

- CPU Intel Xeon 3.06GHz
- OS Redhat9 (kernel 2.4.20)
- コンパイラgcc 3.2

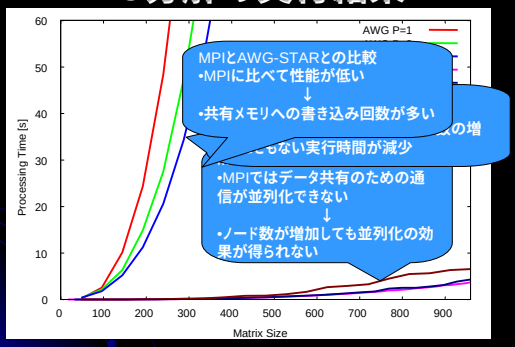
実験環境：アプリケーション

- SPLASH2(並列分散計算用ベンチマーク集)
  - 以下のプログラムを使用
    - 基数ソート
      - 共有メモリへのアクセスが少ない
    - LU分解
    - 高速フーリエ変換(FFT)
      - 共有メモリへのアクセスが多い
  - 処理の対象となるデータを分割し、各ノードで処理を行うことで並列化
- MPI(mpich1.2.5)を用いた場合と比較

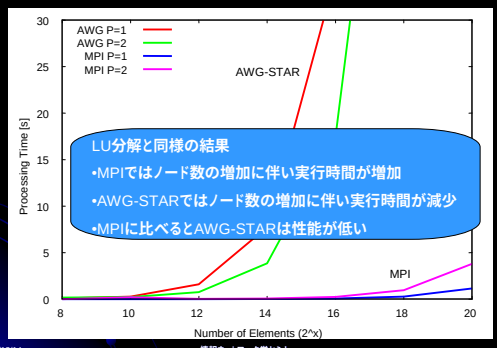
基数ソートの実行結果



LU分解の実行結果



FFTの実行結果



考察

- AWG-STARにおいてLU分解･FFTで性能低下
  - これらは共有メモリへのアクセス回数が多く、これが影響していると考えられる
- 共有メモリへのアクセスによる影響
  - 共有メモリへのアクセス遅延
  - 書き込みの際のトークンの待ち時間
  - 書き込みに伴うデータの光リングの周回時間
- 現状ではシステムの条件のため共有メモリへのアクセスが影響を与えている

改善方法

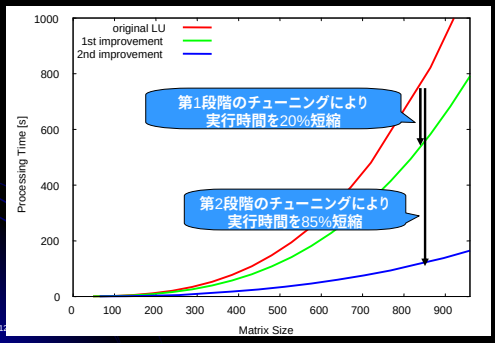
- 共有メモリへの処理遅延が性能に影響

- 改善方法
  - ハードウェアとドライバの改善
    - 共有メモリへのアクセス速度の向上
  - アプリケーションソフトウェアの改善
    - 共有メモリへのアクセス回数を削減

アプリケーションソフトウェアの改善

- LU分解のチューニング
  - 共有メモリへのアクセス回数を削減する
  - ローカルメモリをキャッシュとして使う
- 第1段階：複数のデータをまとめて書く
  - チューニング前：行列の1要素単位 (8B)
  - チューニング後：行列の1ブロック単位 (2KB)
- 第2段階：第1段階に加えて、再利用するデータはローカルメモリに保持
  - 書き込み回数の9割以上が他ノードで利用されず、自ノードでのみ再利用するデータの書き込み

共有メモリへのアクセス回数を削減した場合の実行結果 (ノード数3)



ソフトウェアの改善の結果

- LU分解のチューニングにより実行時間の短縮が行えた
  - 共有メモリへのアクセス回数の削減
  - ローカルメモリをキャッシュとして利用
- FFTでも同様のチューニングにより実行時間を最大で80%短縮

まとめと今後の課題

- λコンピューティング環境における共有メモリ方式
  - 具体的なシステムとしてAWG-STARシステム
- 共有メモリ方式の評価
  - 現状では共有メモリへのアクセスが性能に影響
  - ソフトウェアレベルでの改善で性能向上
- 今後の課題
  - ハードウェアの改善の検討
  - 実践的アプリケーションによる評価
  - 効率のよいメモリアクセス手法の提案