advanced network architecture research group

λコンピューティング環境構築のための Globus Toolkit を用いた MPI ライブラリの実装と評価

大阪大学 基礎工学部 情報科学科
村田研究室　井本 舞

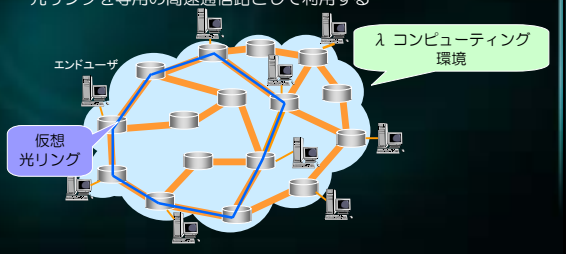
研究の背景

- グリッドコンピューティング
  - ネットワークを介して複数の計算機を接続し、計算資源、ストレージを共有
    - 広域で大規模な計算
    - 大容量データの高速転送
  - 通信オーバーヘッドが問題
    - TCP/IP が通信に使われる
      - パケット処理によるオーバーヘッド
      - パケットロスによる再送遅延

高速かつ、高信頼な通信パイプを
エンドユーザに提供する技術が必要

λコンピューティング環境の提案

- 計算機、ルータを光ファイバで接続
- 波長パスを通信の最小粒度とする
- 仮想光リングを構成
  - 光リングを専用の高速通信路として利用する



研究の目的

- λコンピューティング環境に Globus Toolkit を導入
- λコンピューティング環境における MPI ライブラリの実装と評価
  - NTTフォトニクス研究所が開発したAWG-STARを利用

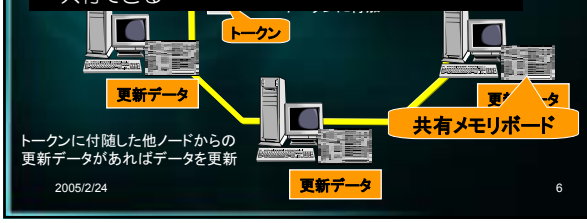
Globus Toolkitによるグリッド環境構築

- Globus Toolkit はグリッド環境を構築するためのミドルウェア
  - 各計算機の実装に依存しないインターフェースを提供する
  - 通信、認証、ジョブ管理などを行う

λコンピューティング環境にGlobus Toolkit を導入することにより、ユーザはλコンピューティング環境を意識することなく高速な分散計算環境を使うことができる

AWG-STARにおけるデータ共有手法

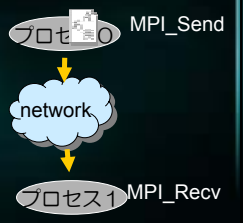
- 共有メモリから読み込み
  - 自ノードのメモリボードへアクセスする時間がかかる
- 共有メモリへの書き込み
  - 自ノードへのメモリボードアクセスする時間とトークンが光リングを一周する時間がかかる
- 自ノードのメモリボードへの読み書きでメモリを共有できる



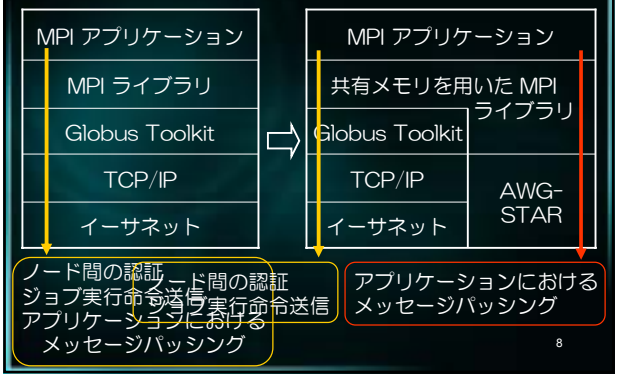
MPI (Message Passing Interface)

- 並列計算ではメッセージ交換を行いながら計算を進める
- MPI はプロセス間でメッセージを交換するための仕様

```
If(my_rank == 0){
  sprintf(message,"%s", Hello);
  MPI_Send ( 送信先ランク：1
             メッセージ：message);
}
 else if (my_rank == 1){
  MPI_Recv ( 送信元ランク：0
             メッセージ：message);
}
```

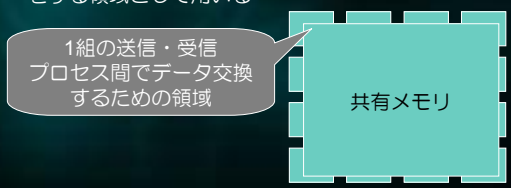


λコンピューティング環境における並列計算



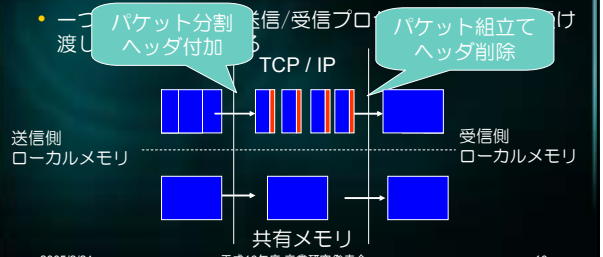
共有メモリを用いた MPI ライブラリの実装

- 共有メモリ上で動的にメモリをアロケートすることができない
  - プロセスの個数をnとすると、共有メモリを$n \times n$に分割
  - 一つの領域を一組の送信/受信プロセス間でのデータ交換をする領域として用いる



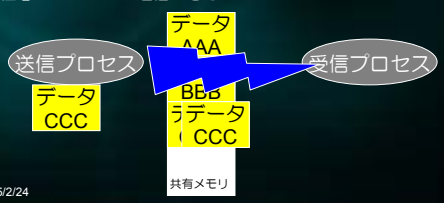
共有メモリを用いた MPI ライブラリの実装

- 共有メモリ上で動的にメモリをアロケートすることができない
  - プロセスの個数をnとすると、共有メモリを$n \times n$に分割
  - 一つ 送信/受信プロ け 渡し



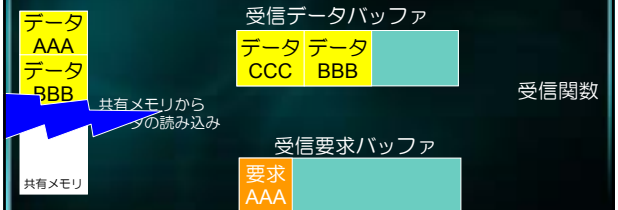
メッセージ送信時

- 送信関数が呼ばれると、送信データを共有メモリに書き込む
- 送信データを書き込んだ後、受信プロセスにシグナルを送る
  - シグナルはAWG-STARが提供する機能
  - 任意のプロセスに送信できる



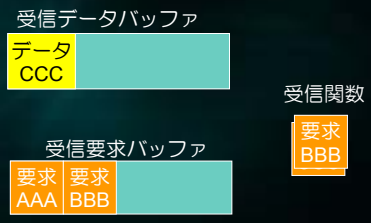
メッセージ受信時

- 受信関数が呼ばれるタイミングと、データを受信するタイミングが異なることを考慮
  - 受信データバッファと受信要求バッファをローカルメモリ上に設ける



メッセージ受信時

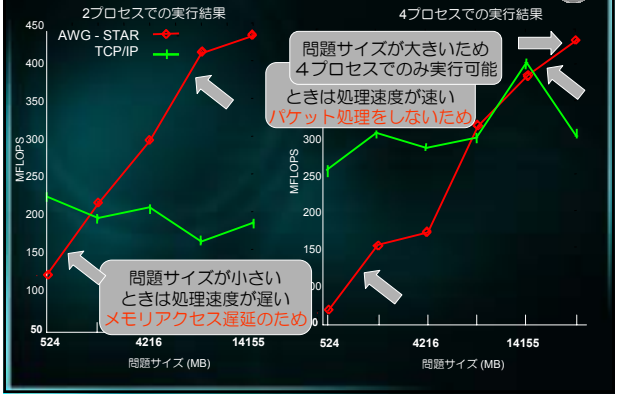
- 受信関数が呼ばれるタイミングと、データを受信するタイミングが異なることを考慮
  - 受信データバッファと受信要求バッファをローカルメモリ上に設ける



評価モデル

- 実験環境
  - ノード計算機最大４台で実行
    - １ノード計算機上で１プロセス
- 評価アプリケーション：姫野ベンチマーク
  - メモリの性能が処理性能に現れる
    - 共有メモリの性能評価に適用
  - データ交換に MPI を利用
  - 様々な問題サイズを設定できる
    - 問題サイズと交換するデータサイズが比例
    - 問題サイズとデータ交換の回数が反比例
- TCP/IPを用いた MPI ライブラリと比較を行う

姫野ベンチマークの実行結果



まとめと今後の課題

- まとめ
  - λコンピューティング環境における MPI ライブラリの設計、開発、実装、評価を行った
  - 共有メモリへのアクセス遅延が大きい
    - NTTフォトニクス研究所に改善要望
  - 交換データサイズが大きい場合は有効
- 今後の課題
  - 共有メモリ上のメモリを動的にアロケートするインターフェースの開発