省メモリ技術と動的最適化技術によるスケーラブル通信ライブラリの開発

九州大学情報基盤研究開発センター
南里　豪志

2011年10月11日
CREST「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」領域会議

研究のねらい

- Exa FLOPS級計算機で、スケーラブルな並列アプリケーションの実行を可能とする

  スケーラブルな通信ライブラリ実装技術

  スケーラブルな並列アプリケーション作成のための通信ライブラリ応用技術

- 目標：

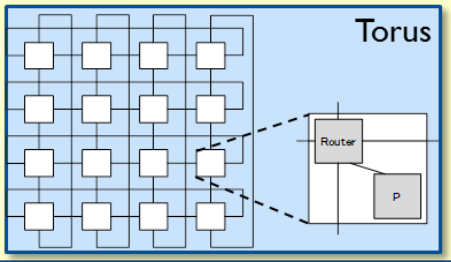
  通信バッファ領域を総メモリ容量の10%以内に抑えながら、**実アプリケーションで数千万～数億プロセスまでの性能向上**を維持する

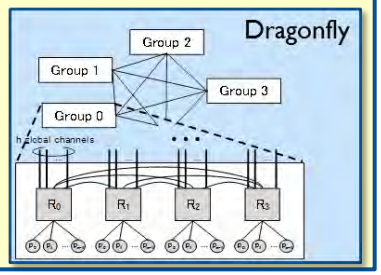
本プロジェクトで想定するExa FLOPS環境

システム全体

- ノード数： 数十万ノード
- インターコネクト：
  多次元トーラス もしくは
  直接網＋間接網（Dragon Fly等）

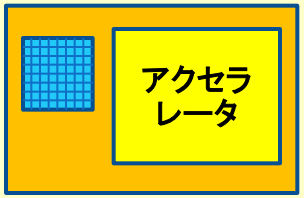
プロセスの配置が性能に大きく影響する



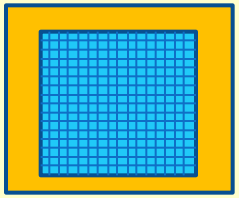


ノード内

- コア数：
  数十コア+アクセラレータ
  (heterogeneous)



もしくは
数百コア～千コア
(homogeneous)



- メモリ量： ～数百GB

コア当たりメモリ量の低下

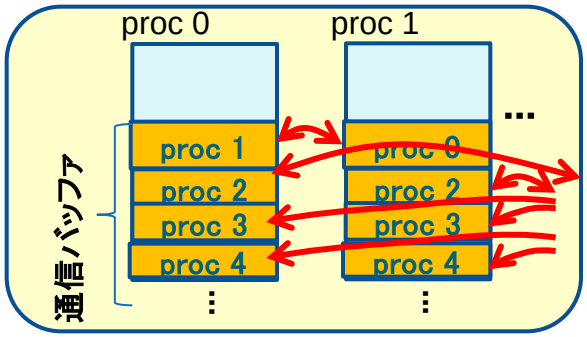
並列プログラミング環境

プロセス数：	数千万～数億プロセス
プロセス当たりメモリ量：	1~10GB/プロセス程度
プログラミングインタフェース：	高級並列言語（PGAS等） もしくは MPI + OpenMP

Exa FLOPSに向けた通信ライブラリの課題
1. 数億プロセスに耐えるメモリ管理機構

- プロセスあたりメモリ容量： 1～10GB程度

- 従来の通信ライブラリにおける使用メモリ量の問題
  - i) 各プロセスで全プロセスの情報を管理
    - ＝ 1プロセス 4Byte でも **400MB/プロセス**
  - ii) 通信相手プロセス毎にバッファを用意
    - ＝ 1プロセス 1KBでも、最悪 100GB/プロセス



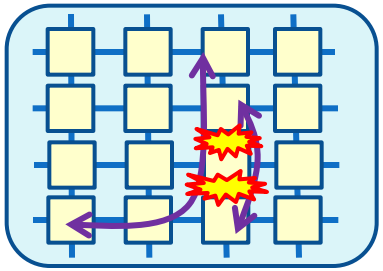
各プロセス用に
通信バッファを用意

Exa FLOPSに向けた通信ライブラリの課題
2. 数億プロセスにおける性能チューニング

- 従来の人手による手法や実行前の静的な手法の限界

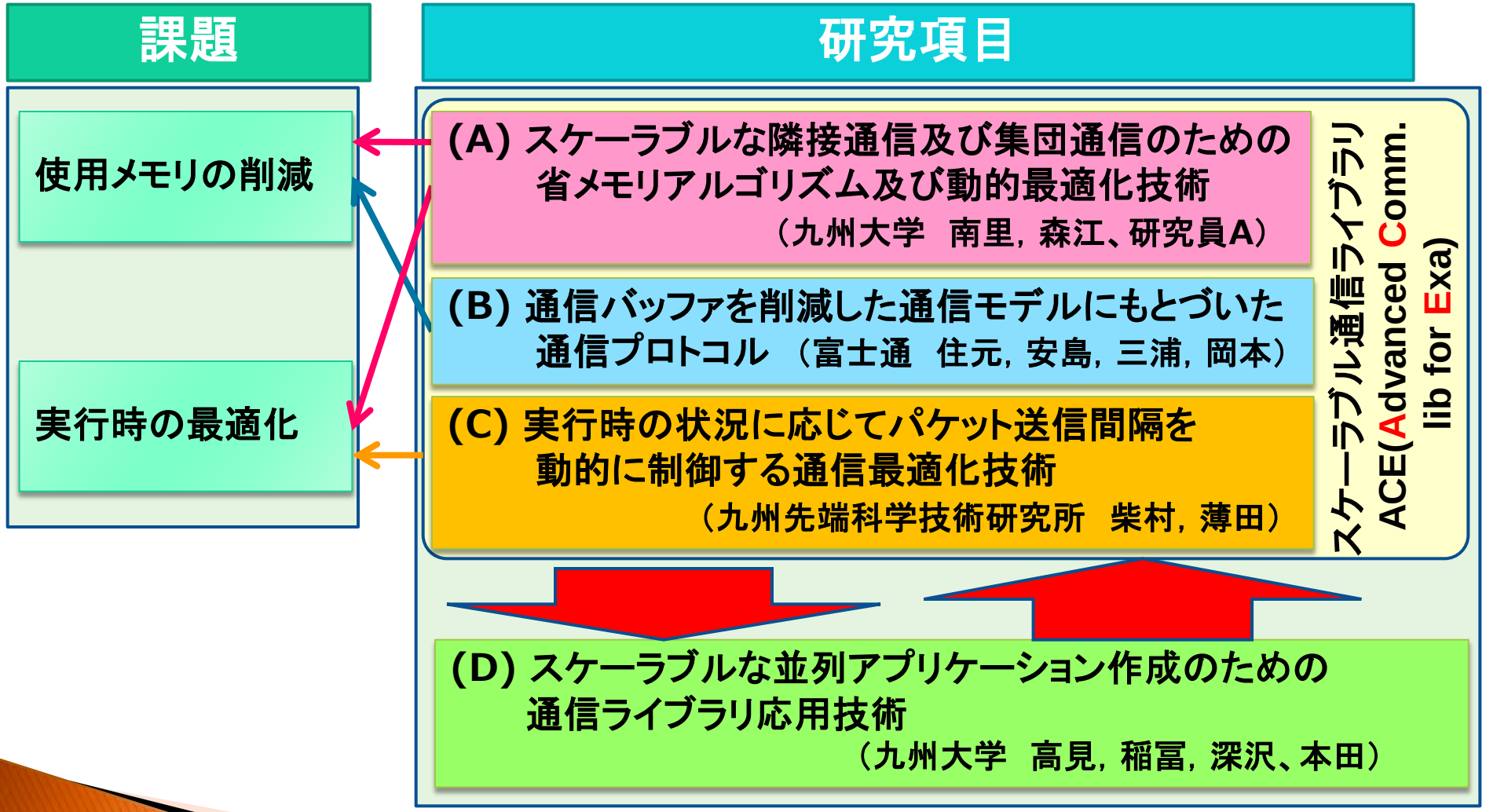
i) 性能最適化の所要時間:
プロセス数増に伴って最適化の探索空間が爆発的に拡大

ii) 性能見積もりの複雑化:
実行時に決定する要素で性能が大きく変動
例) プロセス配置による通信衝突や通信距離変化



通信衝突

本プロジェクトの研究項目

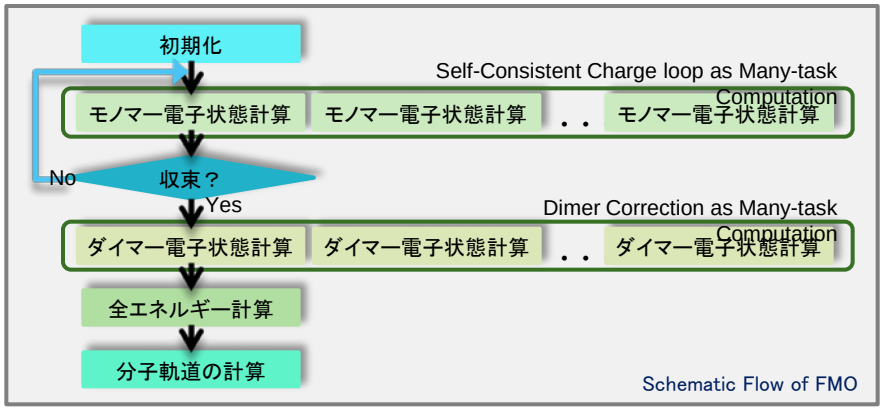


研究のポイント

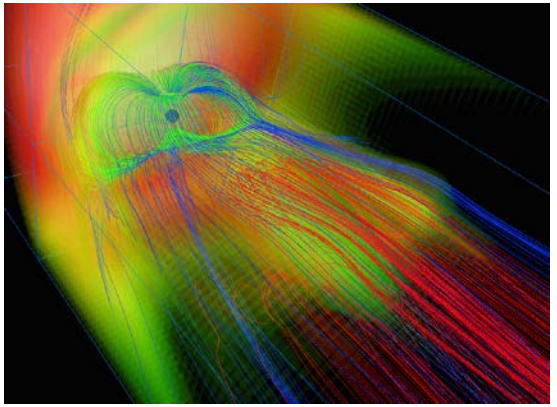
- 通信ライブラリの低遅延化、高スループット化、および省メモリ化
  - アルゴリズム、通信路制御、プロトコルの開発
- ハードウェア、通信ライブラリ、アプリケーションのco-design

Exa FLOPSに向けたアプリケーションからの要請

- 通信を削減、隠蔽、効率化する技術
- 例) 量子化学計算プログラム OpenFMO
  - 集団通信(Allreduce等)を隠蔽したい
  - 負荷均等化(master-worker)のための排他制御用通信を効率よく行いたい



- 例) 電磁流体プログラム
  - 領域間の隣接通信を効率よく行いたい



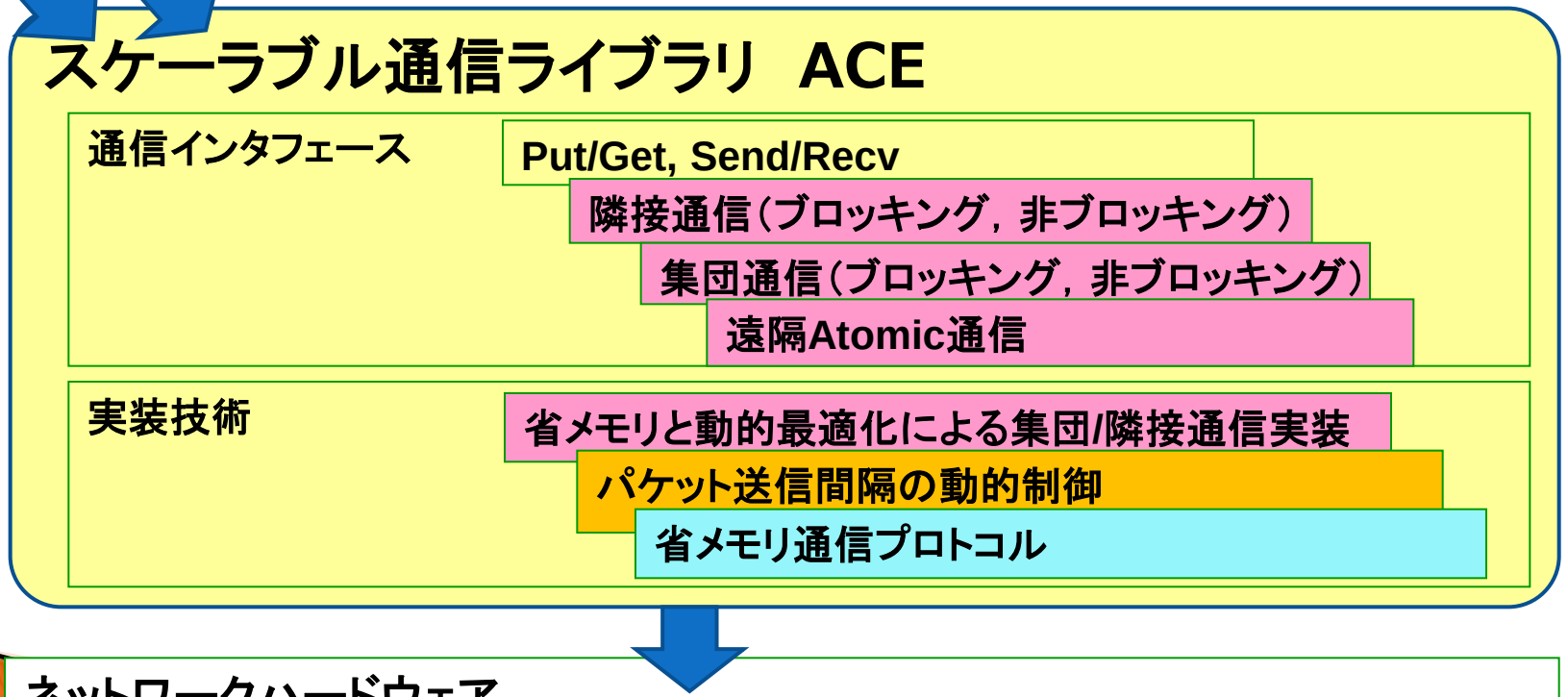
電磁流体シミュレーション結果

ACE

Advanced Communication library for Exa

既存のアプリケーション：
- ACEの省メモリ技術と動的最適化技術によりスケーラビリティ向上

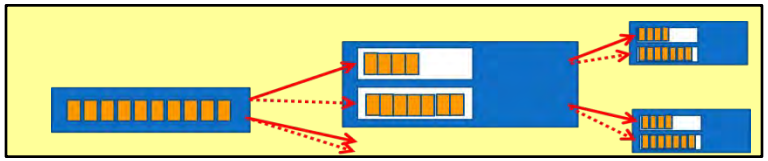
ACEのインタフェースを活用するアプリケーション
- さらに非ブロッキング集団通信，隣接通信等の活用で通信を削減・隠蔽・効率化



ネットワークハードウェア
（RDMA，遠隔Atomic操作，パケット送信間隔調整等）

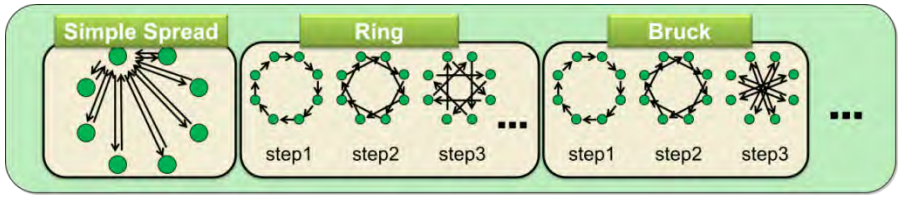
動的最適化技術と省メモリ技術による通信インタフェースの実装

- ランク配置に応じた隣接通信
- 省メモリ集団通信アルゴリズム



共有バッファを用いた木構造による省メモリアルゴリズム

- スケーラブルな動的最適化技術
  - 集団通信、隣接通信のアルゴリズム選択、経路選択、セグメントサイズ調整

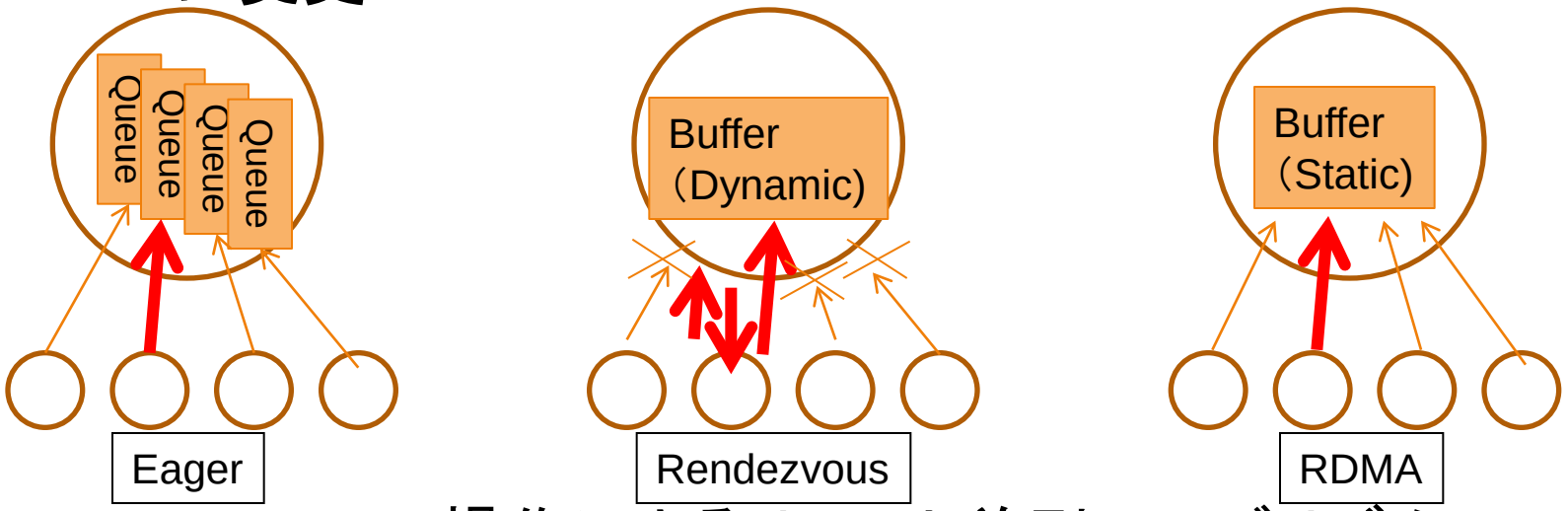


適切な集団通信アルゴリズムの選択

  - 非ブロッキング通信の動的最適化技術
  - 動的最適化機構の低オーバヘッド化、省メモリ化
    - 静的な情報の活用、分散型の管理機構

省メモリ通信プロトコル

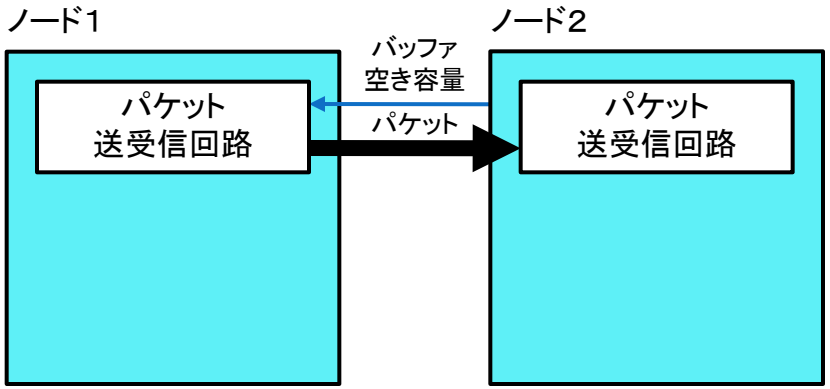
- 基本通信プロトコル： メモリ使用量の多い Eager、遅延の大きい Rendezvous から、省メモリ・低遅延のRDMAに変更

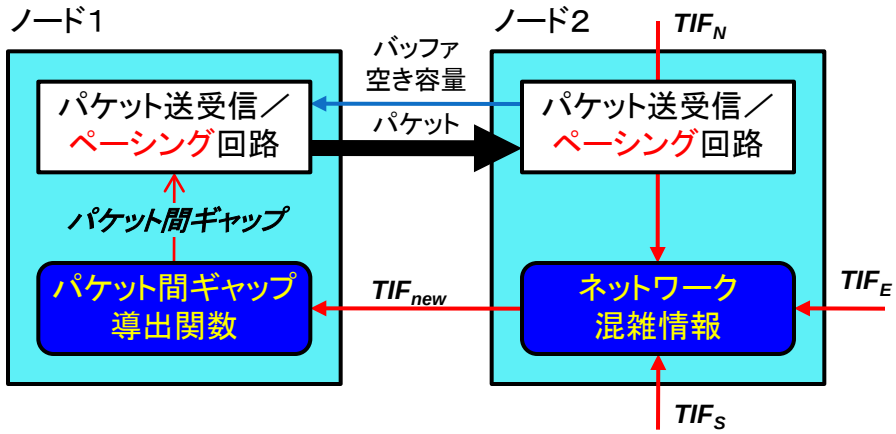


- RDMA, Atomic操作によるリモート並列アルゴリズム・ライブラリを整備
  - 片方向通信で、リモートノードの資源を動的利用
    - Mutex, Lock-Free List, Lock-Free Queue

動的パケット送信間隔制御技術

- ネットワークの混雑情報を収集し，パケット送信間隔を調整　⇒通信スループット最適化
  - TIF(Traffic Information Flit：経路上の混雑情報通知用フリット)による情報収集技術
  - 最適なパケット送信間隔の導出関数
    - 特に集団通信，隣接通信等のパターン通信





スケーラブルな
アプリケーション作成技術

- 量子化学計算プログラム OpenFMO
  - 非ブロッキング集団通信による通信時間隠蔽
  - 遠隔Atomic通信による効率的な排他制御
- 電磁流体プログラム
  - ランク配置に応じた動的最適化を行う非ブロッキング隣接通信による通信の隠蔽と効率化
  - 複数の領域をまとめることによる通信回数削減
- その他、幅広いアプリケーションに応用

- 通信ライブラリへのフィードバック

全体スケジュール


2011
2012
2013
2014
2015
2016
(A)省メモリと動的最適化による通信ライブラリ
ランク配置に応じた隣接通信パターン
非ブロッキング通信の自律通信推進機構
集団通信動的最適化技術
隣接通信、集団通信のプロトタイプ実装
省メモリ型アルゴリズム開発
非ブロッキング通信動的最適化技術
通信ライブラリACE 実装
ACEの性能検証，改良，公開
(B)省メモリ通信レイヤ
メッセージパッシング通信プロトコルの低遅延化、省メモリ化
データ構造遠隔操作ライブラリ
PGAS処理系向けライブラリの低遅延化、省メモリ化
分散データ構造ライブラリ
機能拡張、改良
(C)パケットペーシングによる動的最適化
隣接通信，集団通信におけるパケットペーシングの評価とモデル化
TIFを用いたネットワーク混雑情報収集
最適なパケット送信間隔導出関数
通信ライブラリへの実装
性能検証，改良
並列計算環境の性能予測技術
非ブロッキング通信の性能予測
Exa環境向けシミュレータ
Exa環境のアプリケーション性能予測技術
(D)並列アプリケーションへの応用
OpenFMOの通信隠蔽，動的負荷均等化
電磁流体プログラムの通信隠蔽
並列アプリ調査
並列アプリ通信効率化
省メモリ性，動的最適化性能の検証とフィードバック
高スケーラブルアプリケーション作成技術

本研究の波及効果

- 高級並列言語処理系との連携
  - 省メモリ、動的最適化技術： 下位通信レイヤのスケーラブルな実装
  - アプリ作成技術： コンパイラの自動最適化機能への活用

- MPI規格への寄与
  - アプリケーションによる評価にもとづいた、実用的なインタフェースの提案