1P14

分子軌道法プログラムでの利用に向けた ACP 通信ライブラリによる集団通信の実装

○本田宏明[1,3], 森江善之[1,3], 南里豪志[1,3], 稲富雄一[2,3], 高見利也[1,3]
(九大情基センター[1], 九大シス情[2], JST-CREST[3])

【はじめに】

京コンピュータを始めとした数万ノードからなるスーパーコンピュータが既に実現されている．現在研究開発が行われているエクサスケールクラスでは，$10^6 \sim 10^7$ ノードクラスの超高並列環境となる一方で，スパコン消費電力とコストの問題から 1 プロセスが使用可能なメモリ量は 10GB 程度に制限されるのではないかとの予想がされている．一方，エクサスケールに向け開発が進んでいる大規模並列アプリケーションの多くは，通信ミドルウェアとして Message Passing Interface (MPI) ライブラリ [1] を前提とした実装がなされており，分子軌道法計算アプリケーションも同様である．この現在標準ともいえる MPI ライブラリであるが，10^6 プロセス実行の場合では各プロセスの通信に関わる要求メモリ量は 30GB を越えると試算されており [2]，このままでは起動プロセスの増加とともにいずれプログラムの実行が困難になると危惧されている．そこで ACE プロジェクト [3] の研究グループでは，通信元と通信先のメモリ上のデータを片側のプロセスのみが関与し送受信を行う Remote Direct Memory Access (RDMA) の機構に基づき，低遅延・省メモリな記述が可能な Advanced Communication Primitives (ACP) ライブラリ [3] を新規に開発中である．Ethernet や Infiniband[4], Tofu[5], Tofu2[6] 等のネットワークデバイスに依存しない Basic Layer ライブラリ群の仕様はほぼ策定が終わり，実装についても既に公開が始まっている [3,7]．しかしながらアプリケーションの実装者からは低レベルな機能を持つ Basic Layer の直接利用は難しいといった問題がある．

そこで本研究では，ACP Basic Layer ライブラリの特徴である，遠隔間通信ならびに片側通信の特徴を活用した，利便性の高い集団通信ライブラリの仕様の提案ならびに実装を行なった．この際には，既存の MPI ライブラリの仕様から離れ，分子軌道法計算や他のアプリケーション向けに個別のインターフェースを提供可能な，基本となる集団通信関数を新規に開発することとした．

【ACP 通信ライブラリ】

ACP 通信ライブラリではライブラリによる暗黙のメモリ確保量は最小限に留められており，アプリケーション開発者が実際の通信に必要な分のみバッファの動的確保や破棄を制御可能としている．また，RDMA を利用した片側通信の機能を有し，MPI3 規格でもサポートされていない種々の通信機能をライブラリ利用者がボトムアップ的に実装可能であるといった利点を持つ．分散並列環境に対してグローバルアドレススペースモデルを採用しており，全プロセスにてユニークとなる Global Address (GA) 情報によりメモリアクセスのための情報を管理可能としており，プロセス間のデータコピーに対しては C 言語の memcpy 関数と同様な記述が可能である．また，Basic Layer では，データ送信元でも受信先でもない第三者の通信マスタープロセスによる遠隔間通信命令を発行可能といった特異的な特徴を持ち，Basic Layer を利用した新規通信ライブラリ開発を容易にしている．現状では ACP ライブラリの下層に位置している Basic Layer とそれを利用する Middle Layer の 2 層で構成されており，本研究で開発中の集団通信関数群は Middle Layer に属している．

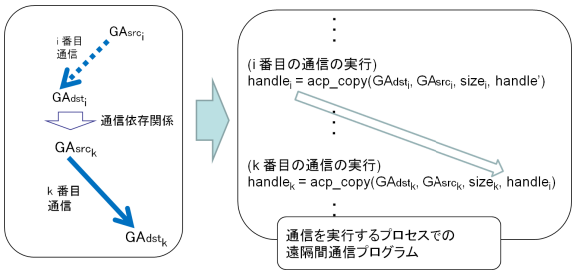

図 1: ACP の遠隔間通信を利用した依存関係を含む複数の通信の記述

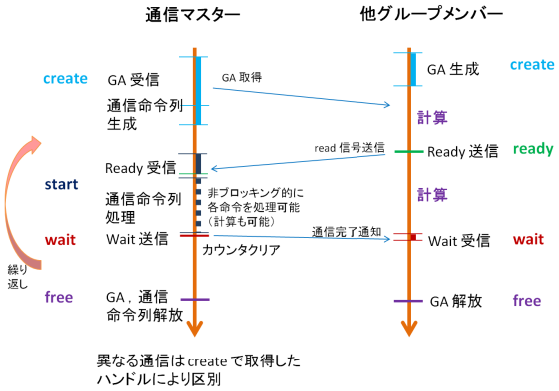

図 2: 基本集団通信における通信手続き

【基本集団通信の実装】

本研究で実装している基本集団通信は，ACP ライブラリの特徴の一つである遠隔間通信の機能を利用し実装されており，Persistent 型かつ片側通信に基づく非ブロッキングのインターフェースを実現している．集団通信は一般的に一対一通信の組合せとして記述可能であり，親ランクのデータを自ランクを経由して子ランクに送信する場合など，それぞれの通信には依存関係がある．これに対し，Basic Layer でサポートされている acp_copy 関数の遠隔間通信機能を利用することにより，通信マスタープロセスのみにおいて依存関係のある通信素過程の組合せとして記述可能である (図 1 参照)．この依存関係は通信命令列なるデータとして記述可能であり，通信命令列を変更することで種々のタイプの通信が実現される．

また図 2 に示す様に，各通信を create, ready, start, free からなる手続きに分解したインターフェース (Persistent 型) としており，各手続きを片側通信かつ非ブロッキング通信とすることで，通信マスターが発行する通信とワーカープロセスにおける計算のオーバーラップを可能とした．通信を実際に行なう start 関数と，主たるオーバーヘッドとしての create 関数を分離することで，Hartree-Fock 計算等における繰返しの複数回の集団通信に対し，create 関数の実行を繰返し前の 1 回のみにすることが可能である．

現実装における要求メモリ量は，10^6 プロセスの一斉集団通信に通信マスタープロセス 1 つといった最も多くなる条件においても，全ての GA 情報ならびに通信命令列の保存に必要な 40MB 程度となった．これに Basic Layer 分を併せても 200MB と 10GB の制限以下である．現実の利用の際には MPI のコミュニケータに対応するグループ内通信を利用することにより要求メモリ量が減少し，Basic Layer についてもさらなる省メモリ化を現在実施していることから，将来のエクサ環境においても十分に利用可能なメモリ量に抑えることが可能である．

当日は，本通信ライブラリを利用した量子化学計算アプリケーション実装についても報告する予定である．

【参考文献】

[1] Message Passing Interface Forum (online) available from ⟨http://www.mpi-forum.org/⟩.
[2] 住元真司 他, 情報処理学会研究報告, Vol.2014-HPC-143, No.8, 2014., 安島雄一郎 他, 同研究報告, No.9.
[3] ACE project (online) available from ⟨http://ace-project.kyushu-u.ac.jp/index.html⟩.
[4] Infiniband Trade Association (online) available from ⟨http://www.infinibandta.org/⟩.
[5] Y.Ajima *et al.*, *FUJITSU Sci.Tech.J.*, Vol.48, pp.280-285, 2012.
[6] Y.Ajima *et al.*, *Supercomputing Lecture Notes in Computer Science*, Vol.8488, pp.498-507, 2014.
[7] 森江善之 他, 情報処理学会研究報告, Vol.2015-HPC-148, No.33, 2015., 野瀬貴史 他, 同研究報告, No.32.