Rにおける高性能コンピューティング

中野 純司　モデリング研究系　教授

【はじめに】

フリーの統計解析環境Rは現在最も標準的な統計解析ソフトウェアと言える。その完成度は高く、研究のみならず実務的なデータの解析にもよく利用されている。Rで大量データの解析や大量の計算を行うためには並列計算のパッケージを利用する必要があるが、その中でもパッケージsnow (Simple Network Of Workstations) が最もよく利用されている。snow はlapplyのような関数を並列環境に実装したものである。ただ、その実装には改善する余地がある。本稿では snow と同等ないくつかの関数のMPI(Message Passing Interface)を用いたより効率的な実装を与え、特にスーパーコンピュータ上のRのための簡易かつ高速な並列計算方法を提供する。
本研究は中間栄治氏（株式会社COM-ONE）との共同研究である。

【MPI】

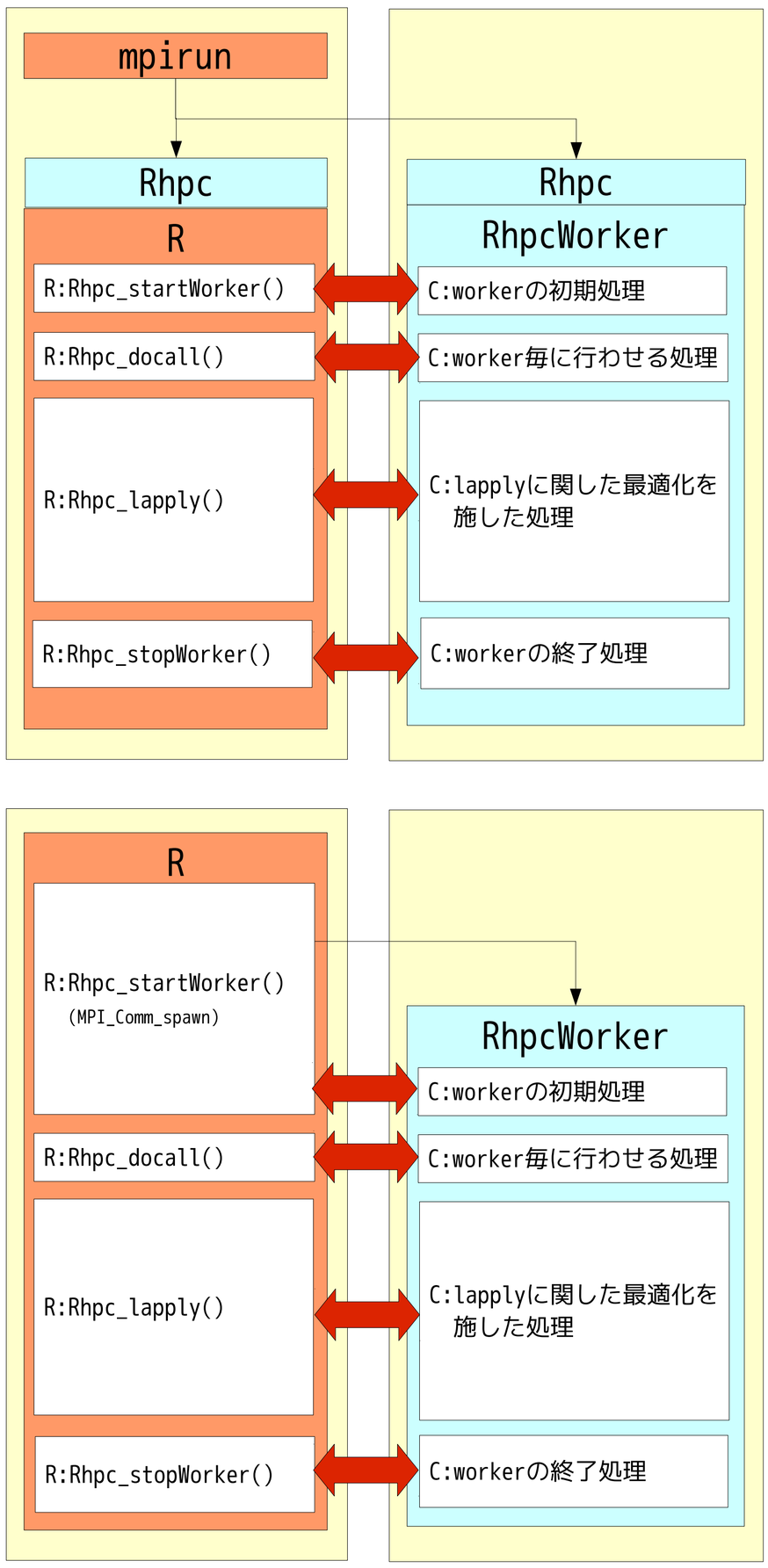
今日の共有メモリ型及び分散メモリ型のスーパーコンピュータでは並列計算を用いて高性能コンピューティングを行う。共有メモリ型においてはNUMAノード、分散メモリ型においては各ノードにいかに効率よく計算を分散させるかが演算効率を高める鍵である。これらの資源管理を行うためにスーパーコンピュータにはジョブスケジューラが搭載されているが それにより効率的かつ汎用的に扱える並列化のライブラリとしては、MPIが現状では最も妥当であり、数万並列の計算さえよく行われている。MPI は種々の計算機言語で利用でき、RからもRmpiというパッケージを用いれば利用可能である。ただし Rのプログラム及びデータ(例えばlist型)をMPIの考え方で扱うのは容易ではなく、Rmpiによりプログラムを組むと言うのは現実にはあまり行われていない。

【Rにおける並列環境】

Rで並列計算を行うためにはパッケージsnowを使う方法が主流となっている。snowは比較的高水準なlapply系の関数を提供し、Rの考え方に沿っているので使いやすい。ただ、MPIやSOCK(ソケット)などの低水準な機能を抽象化して統一的に利用するために、多くの場合最適な処理が行われているとは言えない。
また、その実装にはいくつかの問題点がある。まず、SOCKを用いる場合、利用するソケットの番号を利用者が管理しなければいけない。システムを個人が使う場合は問題ないが、 複数の利用者が共有している環境では管理が困難である。Rmpiを用いる場合はMPIにおけるオブジェクトの長さが int (従って最大$2^{31}-1$)に規定されているため、それ以上のデータの受け渡しができない。

【Rhpcの概要】

われわれは snow の欠点を克服し、Rでよりよい高性能コンピューティングを実現するために、パッケージ Rhpc を作成した。Rhpc はsnowと同等のいくつかの機能をMPIを用いてより効率的に実装したものである。
複数ユーザが共同で利用するコンピュータシステムでは ジョブスケジューラによってプロセスが割り振られるために MPIの起動はmpirunやmpiexecによって行われる。われわれのパッケージではmpirun等のコマンドから呼び出す場合はRhpcと言うシェルスクリプトを経由してこれらを実行させ MPIの環境変数(実装により異なる)によりmasterではRを、workerではRhpcWorkerを起動する。
またパーソナルコンピュータ上やパーソナルクラスタ上ではMPI2のMPI_Comm_spawnによる動的プロセス生成も可能であり、この場合はmasterがRhpcWorkerを直接起動する。
われわれのパッケージではオーバーヘッドを削減した２つの関数を提供する。



- Rhpc_docall(cl,FUN,...)
  workerに対して指示を一斉に出す場合(例えば乱数の指示)、MPI_BCASTを使った同報通信を利用しているのでオーバーヘッドの軽減が期待される(軽減の程度はMPIの実装に依存する)。
- Rhpc_lapply(cl,X,FUN,...)
  各 worker 毎に必要な X 引数の要素のみを MPI_SEND で渡し関数(FUN)及び引数群(...)など、不変の値は初回のみにMPI_BCASTにて同報通信によって与える。これにより通信量を削減する。

なお大きなデータを送信する場合、自動的に複数に分割してMPIを用いる。その他、最適化BLASの並列数の制御関数やRにおける最適なHPC環境のための様々な機能を提供する。

【おわりに】

Rは優秀な統計解析ソフトウェアではあるが、使いやすさを重視したインタープリタ言語で実装されているため、計算速度は速いとは言えない。ただ、統計計算には並列計算に向いているもの（各並列計算の間で情報のやりとりがほとんどないもの）が多く、また、lapplyのような関数は自然に並列化できる。従ってsnowやそれを改良したわれわれのパッケージは今後ますます利用されることが多くなると考えられる。