第2章 分散メモリ型の並列計算システムでアプリを作る際に使われるMPIを利用して

48コア，RAM96Gバイトのクラスタ作り

宮田 賢一

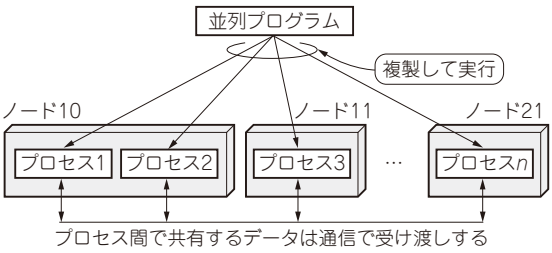

図1 MPIの実行モデル

大規模演算を複数のコンピュータに分散するプロトコル MPI

Open MPIによる分散メモリ型のクラスタ環境を実際にラズベリー・パイで構築し，幾つかの並列計算プログラムを動かしてみます．これによりスーパ・コンピュータの世界の一端を体験できるでしょう．

● 移植性とスケーラビリティに優れる

複数のプロセス間でデータの送受信を行うための通信プロトコルにメッセージ・パッシング・インターフェース(Message Passing Interface; MPI)があります．Open MPIはこれを実装したソフトウェアの1つです．

MPIは分散メモリ型の並列計算システムでアプリケーションを記述するのに広く使われています．MPIで書かれたプログラムは移植性が高いのが特徴です．

つまりラズベリー・パイやPCからスーパ・コンピュータに至るまで，同じプログラムを異なるアーキテクチャで実行できます．

また高いスケーラビリティも得られます．ここでいうスケーラビリティとは，同じプログラムを幾つまで複数実行して並列化できるかという拡張性のことです．MPIを使うと1台から10万台クラス[注1]まで拡張できます．

● 同じプログラムを必要な数だけ複製し独立したプロセスとして実行する

MPIの実行モデルを**図1**を使って説明します．MPIを使って作成した並列処理では，同じプログラムを必要な数だけ複製し，各ノード上で独立したプロセスとして実行します．

ノードとは，ラズベリー・パイやPCのような，計算を実行するコンピュータのことです．各プロセスは複数のコンピュータ上に分散するのが基本ですが，同じコンピュータ上で複数のプロセスとして実行しても構いません．

ここでいう「プロセスとして実行する」ということは，異なるプロセスの間ではメモリのアドレス空間を共有しないということを意味します．同じプログラムを実行する複数のプロセスが，たとえ1つのノード上で実行されていたとしても同様です．つまりプロセス間でデータの受け渡しが必要な場合は，プロセス間でのデータ通信が必要ということになります．

MPIの規格では通信路として何を使うかは規定していませんが，小規模な構成ではイーサネット上のTCP/IPネットワークが使われます．

数万ノードを超えるような大規模なスーパ・コンピュータでは，InfiniBandと呼ばれる広帯域・低遅延な通信路などが採用されています．

一方ノード内のプロセス間の場合は，UNIXドメイン・ソケットを使ったプロセス間通信(OSを介したメモリ間コピー)によって効率良く処理することもできます．

ラズパイ群で並列分散処理できる環境を構築する

● 構築するシステムの特徴

ラズベリー・パイを使ってMPIの実行環境を構築していきます．実際に構築したラズベリー・パイ・クラスタの外観が**写真1**です．

作成するシステムは次の特徴を持ちます．

注1：日本のスーパ・コンピュータである京では8万ノードでMPIを動作させていました(1)．ただしこれだけの大規模な並列計算機の場合，ノード当たりのメモリ使用量の増大やノード間の通信方式・帯域確保などの構成上の課題が発生してきます．