情報科学実験／計算機システム実験／知能情報メディア実験

高性能並列プログラミング

実験テキスト

1. 実験の目的

本実験は、並列処理システムにおける並列プログラミングの代表的手法を学び、併せて並列処理システムの性能に関する理解を深めることを目的とする。実験では、並列プログラミング手法として、以下の 2 つを題材として取り上げる。

1) MPI (Message Passing Interface)
   MPI は分散メモリ (distributed memory) 型並列計算機において、メッセージパッシングによるプロセス間通信を行うための、業界標準 API (Application Programming Interface) である。現在、ほとんどの分散メモリ型並列計算機ではMPIを用いることが可能で、さらに PC (Personal Computer)やワークステーションのクラスタにおいても、標準的な並列プログラミング手法となっている。
2) OpenMP
   OpenMP は共有メモリ (shared memory) 型並列計算機において、並列プロセスが一つの論理アドレス空間上で共有メモリを介して通信しながら並列処理を行う際、プログラム上の並列処理可能部分を簡単な記述により明示し、効率的な並列処理を行うための業界標準 API である。共有メモリ型の並列計算機のほとんどで OpenMP を用いることが可能となっており、さらに、本来は分散メモリ型であるクラスタにおいても、ソフトウェアのサポートにより仮想的な共有メモリシステムを実現し、その上でOpenMPによる共有メモリ型並列処理を記述する例もある。

本実験では、数種のプログラムについて、MPI 及び OpenMP、あるいはそれらをミックスした並列プログラミングを行い、PC クラスタ上で実際にそれらのプログラムの実行し、台数効果等の性能測定を行う。さらに、得られた結果とプログラムの特性に関する考察を行い、並列処理プログラミングに対する理解を深める。

なお、MPI 及び OpenMP はいずれも、C (C++を含む) 及び Fortran の両言語に対する API が提供されているが、本実験では C (C++を含む) のみを用いる。

2. 並列計算機アーキテクチャとプログラミングパラダイム

並列計算機あるいは並列処理システムのアーキテクチャは、プロセッサ間の通信形態に着目すると、分散メモリ型と共有メモリ型に分類できる。

分散メモリ型並列計算機では、各プロセッサはそれぞれ固有のメモリを持つ。すなわち、1つのメ

モリシステムは1つのプロセッサによってのみアクセスされ、メモリの点だけを考えれば、通常の逐次計算機と同じである。並列処理においては、プロセッサ間での何らかのデータ通信が必要となるが、これはプログラム上で「通信を行う関数またはサブルーチンの明示的な呼び出し」によって行われる。例えば、プロセッサ A がプロセッサ B にデータを与える場合、プロセッサ A 上ではそのデータの送信（send）操作を行い、プロセッサ B 上ではそれに対応する受信（receive）操作を行う。この一連の操作を**メッセージパッシング**（message passing）と呼ぶ。また、複数のプロセッサ間で各自が持つデータの総和を求めたい場合、全員がそのデータを送信し、その後でデータを纏め上げる処理が行われ、これを受信する。全体の処理の歩調は、こうしたデータの送受信の関係により自然に取られる。

一方、共有メモリ型並列計算機では、1つのメモリシステムが複数のプロセッサによって共有され、各プロセッサは自由に任意の番地に対するデータの読み書きを行うことができる。プロセッサ A からプロセッサ B にデータを送りたい場合、特定のメモリ番地にプロセッサ A がデータを書き、その後でプロセッサ B が同じ番地からこれを読み出す。ここで重要なのは、「データが書かれた後に読む」という条件である。分散メモリにおけるメッセージパッシングでは、送信されていないデータを受信することはできないので（データがまだ到着していないため、受信プロセッサは待たされる）、プロセッサ間の待ち合わせは自然に行われる。これに対し、共有メモリでは任意の番地のデータを任意のタイミングで読み出せるため、そこにあるデータが意味のあるものであることを保証する必要がある。従って、共有メモリにおいては、何らかの形でプロセッサ同士が待ち合わせを行う必要がある。この操作を**同期**（synchronization）と呼ぶ。一般に、共有メモリプログラミングシステムでは、何らかの形でプロセッサ間の同期を取る手段が提供されている。

本実験では、分散メモリ並列プログラミングのための API として MPI を、共有メモリ並列プログラミングのための API として OpenMP を、それぞれ取り上げる。MPI 及び OpenMP の仕様・記述方法・プログラム実行方法については付録 A 及び付録 B を参照のこと。

3. PC クラスタ

近年、PC はその性能をますます増し、その対価格性能比はもはやワークステーション(WS)とは比べ物にならないほど良くなっている。そこで、この PC を多数並べ、何らかの形で通信させることにより、並列処理または分散処理形態を取り、高性能計算に用いようという研究が盛んになっている。このように、多数の PC を集め、一塊にしてシステム化したものを PC クラスタ（PC cluster）と呼んでいる。PC クラスタにおけるハードウェアとソフトウェアの構成は、主に以下のような形が主流となっている。

1) ハードウェア

PC 自体は ATX 等の標準的な規格の汎用品を使うことが可能である。メモリやハードディスク等についても、一般的なものである。注意すべきは PC 間を結合し、クラスタを構築するためのネットワークである。現在、100base-TX Ethernet (Fast Ethernet)の価格は、PC 本体に接

続するカードと、ケーブルをまとめるスイッチの双方が非常に安価になっており、低価格の割に性能の高い PC クラスタを構築するための一つの標準となっている。

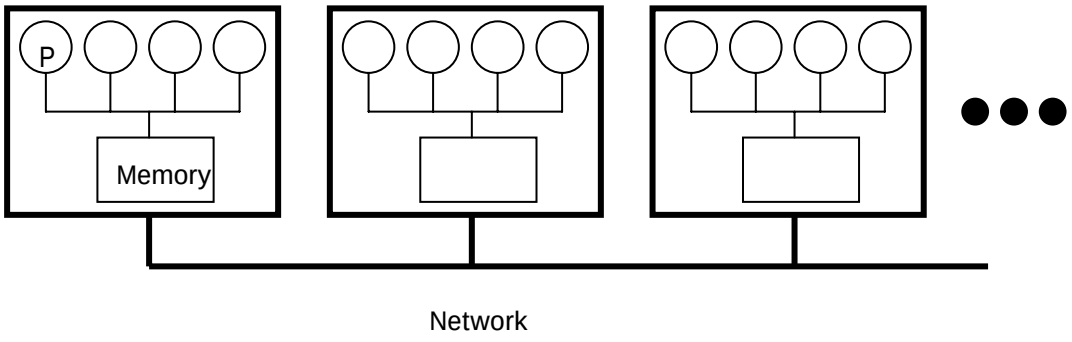
この他に、高性能通信を実現するために Myrinet や 1000base-T Ethernet 等を用いる場合もある。

2) ソフトウェア

最も重要なのは、オペレーティングシステムの選択である。これには Windows 等ではなく、Linux を用いるのが標準となっている。Windows は PC 単体で各種アプリケーションを使うには適当であるが、ネットワーク構成を強く意識し、効率的なシステムを作ろうとする場合は不適当である(特にその不透明性が)。Linux はオープンソース UNIX であり、各種改良がユーザの手で加えられており、ネットワークの仕組みや振る舞いも明確なため PC クラスタでは非常に多く用いられる。

また、PC クラスタにおける PC(**ノード**と呼ぶ)間通信をソフトウェア的にどう見せるかについては、一般的に MPI によるメッセージパッシングが用いられる。また、SMP 型の PC をクラスタに用いる場合、その SMP 上での並列プログラミングも重要になる。これには Pthread を用いる場合と、OpenMP を用いる場合がほとんどである。これらのソフトウェアは無償で提供されているものも多く、特に MPI に関しては、MPICH と呼ばれるライブラリが標準的に使われる。Pthread は Linux に標準のものが添付される。また、OpenMP についてはフリーのものがいくつか存在する。

本実験では、SMP-PC クラスタ上で並列プログラミングを行う。MPI としては MPICH を用い、SMP 上ではフリーの OpenMP コンパイラである Omni OpenMP を用いる。



SMP-PC クラスタのイメージ図

3. 並列処理と性能評価

1) スケーラビリティ(台数効果)

PC クラスタであれ何であれ、並列処理を行う一つの大きな目的は、アプリケーションの高速化である。従って、単に物事が並列に動くだけでは(パズル的興味は別にして)だめで、複数のプロセッサを同時に使っただけの性能向上が求められる。この際、最もよく用いられる指標は**プロセッサ数**

に対するスケーラビリティ（scalability)である。これは、**台数効果**とも呼ばれ、要するに、N 台のプロセッサを用いた場合、1 台のみを用いた場合に比べ、どれくらい速くなるかということを表す。

例えば、1 プロセッサの場合の処理時間を T1 とし、p プロセッサを用いた場合の処理時間を T(p)とする。この時、速度向上率（speed-up ratio）は

$$S(p)=T1/T(p) \qquad (式\ 1)$$

として表すことができる。並列処理がうまく行けば、T(p)は小さくなるため、S(p)が大きいほど速度が向上したことになる。

2) 並列処理効率

並列処理速度向上率 S(p)は大きいほど理想的である。その目安として、S(p)が p に対してどの程度大きいかに着目する。並列処理による速度向上の一つの理想は「p プロセッサを用いた場合、速度も p 倍速くなる」ということである。即ち、

$$S(p)=p \qquad (式\ 2)$$

が理想だということになる。一般には、様々な要因によってこの状況は成立しづらく、$S(p)<p$ となるのが普通である。そこで、プロセッサ数を投入したことに対する見返りがどれくらいあったかを**並列処理効率（efficiency)**として、以下のように定義できる。

$$E(p)=S(p)/p \qquad (式\ 3)$$

もし $S(p)=p$ ならば $E(p)=1$ (=100%) であり、$S(p)<p$ であれば、1 未満となる。

3) 並列処理効率低下の要因

並列処理効率を落とす要因は様々なものがあるが、大きく分けると

・並列処理可能な部分が全処理量に対し十分でない（並列度の不足）
・並列処理をするための余分な処理時間がある（並列処理オーバヘッド）

の 2 種類に分けられる。前者は、例えば仕事を p 台のプロセッサに十分分割するほど大きくない場合である。後者は、例えば仕事を並列に分割する際に通信等の余計な処理が発生することに基づく。

並列処理不可能な部分が多いと並列処理効率が上がらないということを示した、アムダールの法則（Amdahl's Law)というのがある。これは、ある処理の逐次実行時間 T1 が、並列化できない部分 Ts と、並列化可能部分 Tp から成る場合、p プロセッサで理想的に並列化できたとしても、その j 並列処理時間 T(p)は

$$T(p)=Ts + Tp/p \qquad (式\ 4)$$

であるということを言っている。この状況で p を増やしていっても、S(p)は一定値に向かって収束してしまう。

[課題] アムダールの法則に従い、式 4 を元に、プロセッサ数 p が無限大になった極限において、

S(p)及び E(p)がどうなるかを考察せよ。

一方、並列処理オーバヘッドは、分散メモリ方式におけるメッセージパッシング、共有メモリ方式における同期待ちなどの形で現れてくる。本来、逐次処理であれば不要であったこれらの処理は、単純にオーバヘッドとして処理時間に追加される。例えば、N 個のデータを p プロセッサに分割して処理し、最後に結果を取りまとめるような処理においては、最初に各プロセッサにデータを N/p 個ずつ分配する作業や、最後に結果を集める作業で通信が必要となる。このような通信にかかる余計なコストを**通信オーバヘッド**と呼ぶ。

4) メッセージパッシングにおける通信オーバヘッド

メッセージパッシングプログラムにおいては、通信オーバヘッドは比較的単純に予測可能である。モデルを単純化するために、プログラムの実行時間は、正味の計算に要する部分 Tcalc と、通信を行う部分 Tcomm に分かれると考える。総実行時間 T は Tcalc+Tcomm として表される。

通信時間は、一般的に、通信するデータの量に依存する。ここで注意しなければいけないのは、この関係は**比例関係ではない**、ということである。通常、通信時間は**通信データ量に依存しない固定コスト**と、**データ量に比例するコスト**の和で成り立つ。すなわち、データの量を N で示すと、通信時間 Tcomm は

$$Tcomm = a + N/b \qquad (式\ 5)$$

で表される。a は固定コスト、b は単位時間当たりに転送できる通信量単位である。

[課題] 実験 PC クラスタシステムにおける、ノード間 MPI 通信の性能を調べ、式 5 の a 及び b に相当するパラメータを求めよ。同様に、ノード内のプロセッサ間で MPI 通信を行った場合についても調べよ。

一般的に、メッセージパッシングによる並列プログラムでは、各プロセスは何らかの内部処理(演算)を行った後、通信を行う。その後さらに内部処理を行い、次に通信、…というように、内部処理と通信を交互に繰り返す。当然、内部処理に要する時間が長く、通信データ量が少ないほど、通信オーバヘッドは軽くなる。そこで、この両者の比を概念的に**並列処理粒度 (granurality)**と呼ぶ。内部演算が長く、通信時間が短い場合を**粒度が粗い (coarse grain)**と呼び、内部演算が短く、通信時間が長い場合を**粒度が細かい (fine grain)**と呼ぶ。

5) 共有メモリプログラムにおけるオーバヘッド

共有メモリパラダイムの場合、通信のオーバヘッドはメッセージパッシングの場合と違って目に見える形では現れにくい。しかし、共有データの参照のために排他制御を行ったり、同期待ちを行う処理は明らかにオーバヘッドである。

また、OpenMP の場合、並列化を行う際に並列プロセス(スレッド)を起動したり停止したりする必要がある。さらに、それらが並列動作中に、排他制御を必要とする場合、その制御に時間がかかる。その他、private 属性を持つ変数を大量に使うような場合(付録 A 参照)も並列プロセスの起動に

時間がかかることになる。

6) システムの共有

一般に、UNIX を用いたシステムでは TSS (time sharing system)による処理が行われており、単一ノードに複数のユーザプロセスが走り得る。この状況は、同時に複数のユーザに対するサービスを行うことができるが、並列処理システムとして見た場合、CPU やメモリ、ディスク、ネットワーク等のリソースを互いに奪い合うことになり、一般に性能向上の妨げとなる。少なくとも、これまでに述べたような性能評価モデル(performance model)を論じる上で、他のユーザの仕事が含まれると性能予測や性能評価が困難になる。そこで、何らかの形でシステムを排他制御し、あるユーザプログラムがシステムを占有している状態を作り出すことが望ましい。

Linux ベースの PC クラスタでこの状況を作ることは難しいが、SCore と呼ばれるミドルウェアを使い、Linux 上で並列プロセスを制御することにより、これを実現することができる。しかし、システムを完全に排他的に用いると、他のユーザがプログラミングすら行うことができなくなってしまう。従って、TSS を許す状況と許さない状況を、適宜使い分ける必要がある。

本実験においては、実験中の作業を**プログラムの作成・デバッグを行っているフェーズ**と、**並列処理性能評価を行っているフェーズ**の 2 つに分け、前者に関しては TSS 環境で、後者に関しては SCore 環境でプログラムを実行することにする。実験日程中、適宜**性能評価を行う時間帯**を設け、その時間だけは全員が SCore 環境で実験を行うよう、強制することにする。

4. 実験内容

1) OpenMP による並列化実験

以下の各種処理を OpenMP によって並列化し、1 つのノード上の 2 つのプロセッサで性能評価せよ。まず、付録Aの「OpenMPチュートリアル」にある例(hello world, laplace 等)を実験してみよ。その後で、以下の例題を自分で記述し、実行してみよ。

a) 配列データの総和を求めよ。(整数、浮動小数)
b) 配列データに適当な重みの関数を作用させ総和を求める。例えば、1/m sin(1/m)を多数の m(=1,2,...,N)に対して計算し、総和を求めてみよ。
c) S 行 T 列の行列 A と、T 行 U 列の行列 B の積を求め、S 行 U 列の行列を求めよ。

2) MPI による通信実験

付録 B「MPI による並列プログラミング」に従い、以下の各通信実験を MPI によって行い、通信性能を評価せよ。

a) ノード間の point-to-point 通信性能を各種データ量について測定せよ。
b) ノード内の SMP 上のプロセス間での point-to-point 通信性能を測定せよ。
c) 各種データサイズに対し、collective 通信(Allgather 等)の性能を測定せよ。
d) point-to-point 通信を 2 者間のみで行う場合と、そういったペアの通信を多数同時

に行った場合について性能を比較せよ。

3) OpenMP 及び MPI を独立に用いた応用プログラミング

N 点の質点間の重力相互作用に基づき、ニュートンの運動方程式の積分を行うプログラムを OpenMP 及び MPI でそれぞれ別個に作成せよ。

各質点間には互いに以下のような加速度が存在する。

$$aij(d)=(ri(d)-rj(d))*mi*mj/\{r(ij)*r(ij)\} \quad (式 6)$$

ここで、mi 及び mj は質点 i 及び質点 j の各質量、r(ij)は 2 つの質点間の距離である。また、d は方向を表し、3 次元問題では X,Y,Z がこれに相当する。ri(d)は、d 次元方向における粒子 i の座標である。これらにより、質点 i と質点 j の関係における d 方向の加速度 aij(d)が求まる。なお、式をよく見ると、粒子ペア i と j に関し、aij(d)と aji(d)は互いに符号が逆で絶対値が等しいことがわかる(このことは、プログラム上で計算量を削減するのに役立つ)。質点 i の次のタイムステップにおける d 方向の位置 ri(d)next は以下のようにして求まる。

$$ri(d)next=ri(d)+ \Delta t \Sigma aij \quad (式 8)$$

ここで、Σは質点 i に対する他の質点 j からの力の総和である。Δt は時間の刻み幅に相当する定数である。

各質点の初期位置は 3 次元空間上でランダムに与えることとし、各質点の質量 m(i)は 1 から 2 の間でランダムに与えるものとする。

4) 重力プログラムのハイブリッド化

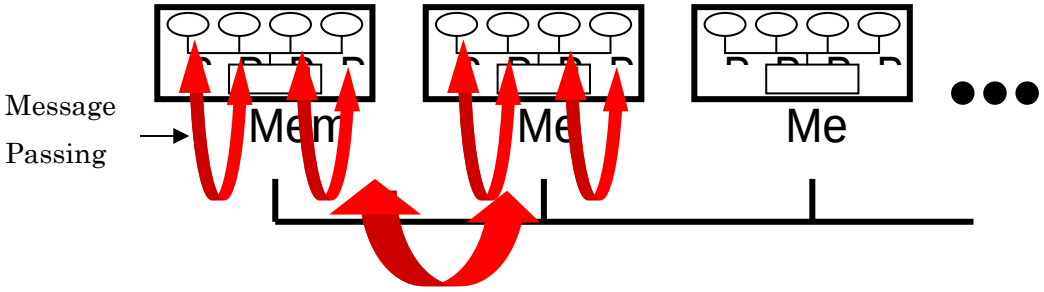
MPI と OpenMP の両方を用いた形で、重力プログラムをハイブリッド化せよ。

ここで、MPI と OpenMP のハイブリッドプログラムとは、以下のようなものを指す。実験で用いる SMP-PC クラスタでは、OpenMP のプログラムは同一ノード上の SMP 結合されたプロセッサ間でしか有効でない。これに対し、MPI プログラムは machinefile をどのように設定するかによって、並列プロセスをどのプロセッサに割り当てるかが決まり、各ノード上に1つずつしかプロセスを置かない場合(この場合、全プロセス数はノード数に等しくなる)や、各ノード上に2つずつプロセスを置く場合(この場合、全プロセス数はノード数×2となる)等が考えられる。MPICH における machinefile は、記述されたホストに対しラウンドロビンでプロセスを割り当てるため、1つのノード名が複数回現れれば、対応する複数のプロセスが同一ノードに割り当てられることになる。この時、UNIX は複数のプロセスを SMP 上のそれぞれ別のプロセッサで実行しようとするため、結果としてノード内の複数のプロセッサの間でも MPI による通信が行われる。

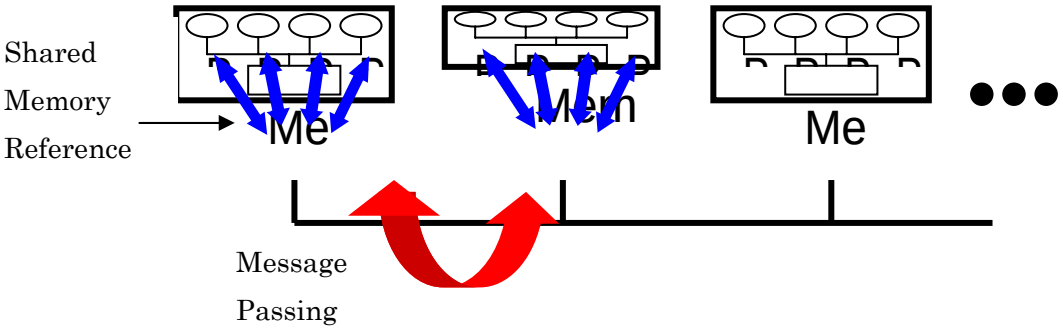
これに対し、ハイブリッドプログラミングでは、まず各ノードに最大1つずつの MPI 並列プロセスを立ち上げる。そして、各プロセス内で適宜 OpenMP 記述を行うことにより、ノード内でも部分的な並列処理を行う。こうすると、ノード内では OpenMP による共有メモリ並列処理が、ノード間では MPI による分散メモリ並列処理が行われることになる。ノード内でのプロセッサ間通信は、MPI 関数を用いてメッセージパッシングを行うより、共有メモリを直接読み書きした方が速いと考えられる。すなわ

ち、ハイブリッドプログラミングによって、ノード内通信が最適化される可能性があり、MPI だけを用いた場合より性能が上がる可能性がある。

しかし、実際にはハイブリッドプログラミングは複雑であり、また種々の要因によって、必ずしもうMPI のみを使う場合よりも高速化されるとは限らない。そこで、ハイブリッドプログラミングを実際に行い、MPI のみを用いた場合と性能比較をせよ。比較は様々な角度から行い、自分なりの結論をまとめること。



MPI のみを用いたプロセッサ間通信



MPI+OpenMP ハイブリッドプログラミングによるプロセッサ間通信