

## 並列処理数値シミュレーションのための スケーラビリティ検討方法

折居茂夫

日本原子力研究所 計算科学技術推進センター  
〒113 東京都文京区本駒込2-28-8 理化学研究所駒込分所  
E-mail: orii@koma.jaeri.go.jp

科学技術分野の数値シミュレーションを並列処理で行う場合、シミュレーションの問題の性質、規模、計算スキーム、アルゴリズム、プログラミング方法が、スケーラビリティに影響を与える。並列計算がシミュレーションプログラム中の並列性を利用するためである。更に、並列計算機のアーキテクチャ、性能仕様、言語等が並列計算機特有のオーバーヘッドを生み、並列性と密接に関係し、スケーラビリティの検討を困難にしている。本研究の目標は、並列処理時間のモデルを作り、数値シミュレーションコードのスケーラビリティ検討方法を確立することである。時間測定と共にこのモデルを使う時、並列性とオーバーヘッドを精度良く表すことができる。本論文では、この方法を示し、並列化コレスキープログラムを用いてVPP500, Paragonでのスケーラビリティ検討を行った。

## A Method of Analyzing Scalability for Parallel Processing Numerical Simulation

Shigeo Orii

Japan Atomic Energy Research Institute  
Center for Promotion of Computation Sci. & Eng.  
2-28-8, Honkomagome, Bunkyo-ku, Tokyo 113, Japan

In case of parallel processing for a numerical simulation in the field of science and technology, there are some factors contributing to scalability, such as a property and size of the problem, calculation scheme, algorithms, and programming style of the simulation code. The reason is that parallel calculations make use of parallelisms in the simulation code. In addition, the architecture, the performance and the language of parallel computers spawn specific overheads which relate closely to the parallelisms and result in the cause of difficulty in analyzing scalability. The target of this study is to establish the method of analyzing scalability of numerical simulation codes by making a model of parallel processing time. The model, combined with measurement of the time, accurately describes the parallelisms and the overhead. In this paper, the modeling methodology and two examples of scalability analyses are shown with a parallelized Cholesky program on VPP500 and Paragon.

## 1. はじめに

並列処理は、問題やシミュレーションコードの中にある並列性を利用して性能向上を図る。しかし並列処理の特有のオーバーヘッドとのバランスの上に全体の性能が決まる。

このように相殺する効果がある場合、アルゴリズムの処理時間をモデル化することによって現象の理解が容易になり、最適化のために役立つことが、ベクトル計算機時代から知られている[1]。並列計算機においても、処理時間の理想化したモデル化を行って並列計算機の特性を研究すること[2]、数値計算アルゴリズムとして意味のあるカーネルをモデル化して性能評価を行うこと[3]が行われてきた。

これらの研究にも関わらず、実際のシミュレーションコードのスケラビリティの検討では、未だ並列化作業を行って時間測定するのが現状である。この主な原因の一つは、シミュレーションコードのような大きなコードの処理時間のモデル化が困難なためである。

しかし、数値シミュレーションコードのスケラビリティを、系統的、定量的に検討するためには、この問題を避けて通ることはできない。

そこで本論文では、3章でシミュレーションコードの処理時間のモデルを作り、任意の並列計算機のスケラビリティを検討する方法について述べる。4章では、カラムコレスキ法を用いた検討例をVPP500, Paragonについて示す。

## 2. 研究のモチーフ

本研究のモチーフは、処理時間のモデル化によって、少ない測定で広い範囲の「問題の規模」と「プロセッサ数」に対し、スケラビリティを検討することである。

また複数の研究者が作成したモデルをアセンブルして利用したい。そこで、多くの計算機に適用でき、あるがままの問題とシミュレーションコードが記述できること、及び他の計算機に無い特定の測定機能を使わないでモデルパラメータを決定できることを原則としたモデル化方法を作成することにした。

本方法でモデル化パラメータを測定すれば、カーネルレベルの並列化方法と並列計算機のスケラビリティの検討ができる。このレベルをシミュレーションコードまで拡張したい。

その時、モデルとパラメータを手掛かりにして、ターゲット数値シミュレーションに対し、スケラビリティを持つ並列計算機を設計することも可能と考える。

## 3. スケラビリティ検討方法

### 3.1 処理時間のモデル化

数値シミュレーションコードに基づいた処理時間 $t$ のモデル化は、問題の規模 $n$ 、プロセッサ数 $N$ を用い次のように仮定して行う。

- 1プロセッサの処理時間 $t_{1pe}$ は四則演算の計算回数 $f(n)$ に比例する。これを式(1)に示す。
- 並列処理時間 $t_{npe}$ は $t_{1pe}$ に比例する。これを式(2)に示す。
- 通信処理時間 $t_c$ は通信回数 $g(n,N)$ に比例する。これを式(3)に示す。

ここに、問題の規模 $n$ とはメッシュ数等を表す量である。 $t_0$ と $E$ は、定数と比例定数である。特に $E$ は効率を表すこととし、1プロセッサの最大性能 $R_a(\text{flops})$ 、通信最大性能 $R_c(\text{Byte/秒})$ 、 $N$ で規格化する。

$$t_{1pe} = t_{s0} + f(n) \cdot R_a^{-1} \cdot E_s^{-1} \quad (1)$$

$$t_{npe} = t_{p0} + t_{1pe} \cdot N^{-1} \cdot E_p^{-1} + t_{povh} \quad (2)$$

$t_{povh}$ は並列オーバーヘッドで次式で表わされる。

$$t_{povh} = t_c + t_{cyn} + t_f + t_j$$

ここに、 $t_{cyn}$ は同期時間、 $t_f$ と $t_j$ はタスク生成と消滅時間である。

$$t_c = t_{c0} + g(n,N) \cdot R_c^{-1} \cdot E_c^{-1} \quad (3)$$

ここに、 $t_{c0}$ は通信立ち上がり時間、 $E_c$ は通信効率である。

$f(n)$ 、 $g(n,N)$ は、シミュレーションコードのコーディングより導出し、 $f(n)$ は、四則演算の数のみを数える。また、内部関数の呼び出しは、四則演算と同等に数えることとする。

コード全体の並列実行時間 $(T_{npe})^T$ は、並列処理

のカーネルの個数を $k_p$ 、逐次処理のそれを $k_s$ とする時、式(4)で表せる。

$$(T_{npe})^T = \sum_{k=1}^{k_p} (T_{npe})_k + \sum_{k=1}^{k_s} (T_{1pe})_k \quad (4)$$

また、コード全体の逐次処理時間 $(T_{1pe})^T$ は、式(5)となる。

$$(T_{1pe})^T = \sum_{k=1}^{k_p} (T_{1pe})_k + \sum_{k=1}^{k_s} (T_{1pe})_k \quad (5)$$

ここに、 $T_{1pe} = \sum_{i=1}^{i_{max}} (t_{1pe})_i$ 、 $T_{npe} = \sum_{i=1}^{i_{max}} (t_{npe})_i$ 、

で、 $i_{max}$  はループ等のイテレーション回数である。

### 3.2 スケーラビリティ決定要因とプログラムパラメータ

数値シミュレーションコードのスケラビリティ決定要因は、次のa.~c.で示される3つの大項目中に分類できる。式(1)~(3)を用いたモデル化によって、要因が細かく区別できるようになる。

#### a. シミュレーションコードの並列性に関するもの

- ・並列化率 (式(6)に示す) (P)

#### b. 並列オーバーヘッドに関するもの

- ・プロセッサの使用効率 ( $E_p$ )  
(ロードバランス、共有メモリアクセス、等)
- ・並列立ち上がり時間 ( $t_{p0}$ )
- ・通信時間 ( $t_c$ )
- ・同期時間 ( $t_{cyn}$ )
- ・タスク生成、消滅時間 ( $t_f, t_j$ )

#### c. 1プロセッサの処理に関するもの

- ・演算器の使用効率 ( $E_s$ )  
(複数演算器の稼働率等、メモリアクセス競合、キャッシュミスヒット等)
- ・演算立ち上がり時間 ( $t_{s0}$ )  
(メモリ・キャッシュアクセスの立ち上がり等)

### 3.3 測定によるプログラムパラメータの決定

測定は、シミュレーションコードに経過時間を測定する時計を挿入して行う。プログラムパラメータE、 $t_0$ 等は、問

題の規模 $n$ に対してプロセッサ数 $N$ を変えて経過時間測定を行い、式(1)、(2)、(3)を最小自乗法でフィットさせて求める。

### 3.4 スケーラビリティ検討

#### (1) $E$ 、 $t_0$ 、 $t_{povh}$ の検討

$E$ は、計算機のポテンシャルをどのぐらい使用しているかを示す目安となる。全体に対して、 $E$ が小さい箇所、 $t_0$ 、 $t_{povh}$ が大きい箇所に着目する。コーディング、計算アルゴリズムレベルの最適化が大きく寄与している場合がある。

#### (2) 並列化率の検討

次式(6)の並列化率 $P$ から、そのシミュレーションコードが持つ並列性を調べることができる。倍率 $1/(1-P)$ が、目安となる。

$$P = \frac{\sum_{k=1}^{k_p} (T_{1pe})_k}{(T_{1pe})^T} \quad (6)$$

#### (3) 並列処理時間検討

プロセッサ数 $N$ に対する並列処理時間 $T_{npe}$ を検討する。リアリティ、ターゲット性能に必要なプロセッサ数を検討する。

また、 $t_{povh}$ を検討することにより、その問題とシミュレーションコードに対する、並列化アルゴリズム、並列計算機のアーキテクチャ、通信性能仕様と演算性能のバランス、等が検討できる。

リアリティが無い場合、項目(2)、(1)の順に遡ることによって原因を特定または、絞ることができる。

## 4. スケーラビリティ検討例

### 4.1 検討対象

図-1に示すカム・コルスキー法のLU分解部を並列化して、そのスケラビリティをVPP500、Paragonについて検討した。プログラムはFORTRANで書き、全ての浮動小数点の計算は単精度で行った。

```
parameter(n=5000)
```

```
real*4 rL(n,n),s(n)
```

```
do 10 j=1,n
```

```

do 11 i=1,n
  s(i)=0.
11 continue
do 20 k=1,j-1
do 20 i=j,n
  s(i)=s(i)+rL(j,k)*rL(i,k)
20 continue
  tj=L(j,j)-s(j)
  rL(j,j)=tj/sqrt(tj)
do 30 i=j+1,n
  ti=L(i,j)-s(j)
  rL(i,j)=ti/sqrt(tj)
30 continue
10 continue

```

図-1 カラム・コレスキー法のLU分解部

並列化は、cyclic手続き分割をdo 20の外側のループに対して行う。配列sとrLは重複マージン配列とし、データ分割は行わないアルゴリズムとした。各プロセッサが計算結果sをdo 30で使用するため、20 continue直後で全プロセッサで総和計算を行う。

do 20のマルチ化は、do 10を考慮して行い、式(7)、(8)となる。

$$\begin{aligned}
 (T_{1pe})_{do10,20} &= \sum_{j=1}^n \sum_{k=1}^{j-1} \sum_{i=j}^n (t_{1pe})_{do20} \\
 &= (T_{1pe})_{do20} \cdot n \quad (7)
 \end{aligned}$$

$$\begin{aligned}
 (T_{1pe})_{do20} &= 1/2 (n-1) \cdot (t_{s0})_{do20} \\
 &+ 1/3(n^2-1) \cdot R_a^{-1} \cdot (E_s)_{do20}^{-1} \\
 (T_{npe})_{do10,20} &= \{(t_{p0})_{do10} \\
 &+ (T_{1pe})_{do20} N^{-1} \cdot (E_p)_{do20}^{-1} \\
 &+ \{povh\} \cdot n \quad (8)
 \end{aligned}$$

また、do 10を考慮してdo 30をマルチ化すると式(9)となる。

$$(T_{1pe})_{do10,30} = n \cdot (t_{s0})_{do30}$$

$$+ (n^2 + n) \cdot R_a^{-1} \cdot (E_s)_{do30}^{-1} \quad (9)$$

#### 4.2 VPP500における検討

VPP500の並列化は、図-1に次の追加を行う。

```

parameter(N$=16)
!xocl processor pe(N$)
!xocl spread do/(pe,index=1:n,part=cyclic)
do 20 k=1,j-1
. . . . .
20 continue
!xocl end spread sum(s)

```

図-2 VPP500の並列化プログラム

総和計算sum(s)はトランスポート転送を用いて行われ[4]、ネットワークにクロススイッチを使用しているため、式(10)のようにマルチ化する。

$$\begin{aligned}
 (T_c)_{do10} &= 2 \{ (t_{c0})_{do20} + (N-1)/N \\
 &\cdot [ (4 \text{byte}) \cdot n / N ] \\
 &\cdot R_c^{-1} \cdot (E_c^{-1})_{do20} \} \cdot N \cdot n \quad (10)
 \end{aligned}$$

ここで、右辺の係数2は、総和計算にトランスポート転送が2回必要なことを示す。また、大括弧の右にあるnは、総和計算が、do 10のイテレーションでn回行われることを示す。その隣のNは、n/N個の4byteデータをN個のプロセッサに転送することを表す。また、(N-1)/Nは自プロセッサへは転送を行わないことを表す。また、VPP500の演算速度が1.6Gflopsであるのに対し、転送速度が100Mデータ/sec(単精度)のため、演算時間は無視できると仮定した。

全体の処理時間は、次式(11)、(12)となる。

$$\begin{aligned}
 (T_{1pe})^T &= (T_{1pe})_{do10,20} + (T_{1pe})_{do10,30} \quad (11)
 \end{aligned}$$

$$\begin{aligned}
 (T_{npe})^T &= (T_{npe})_{do10,20} \\
 &+ (T_{1pe})_{do10,30} + (T_c)_{do10} \quad (12)
 \end{aligned}$$

経過時間測定のため、サビスタプログラムの時計

gettodをdo 20、do 30の前後に挿入した。またsumの時間を測定するため、sumを行わない同一の処理をするdoループを作成し、それらのdo 20の時間差をsumの処理時間とした。測定は、ループnを500から5000まで変え、またNを1から16まで変えて行った。コンパイラは、VPP FORTRAN77 EX/VPP V12L20を用い、コンパイルオプションは、-Wxを用いた。またsumの演算を最適化するためのワーク領域を、実行時オプション-Wl,-Pg2000で確保した。

測定した結果に最小自乗法を用い、式(7)~(10)の定数を求めた結果を表-1に示す。表中の $E_s$ は、この計算がVPP500の1プロセッサのポテンシャルを46~60%を引き出していることがわかる。尚、計算には $R_a=1.6\text{Gflops}$ ,  $R_c=400\text{MB/秒}$ を用いた。

表-1 VPP500のモデルパラメータ値

条件	n=5000	条件	N=1
$(E_p)_{do20}$	0.999	$(E_s)_{do20}$	0.605
$(t_{p0})_{do20}$	40 $\mu\text{s}$	$(t_{s0})_{do20}$	0 $\mu\text{s}$
$(E_c)_{do20}$	0.493	$(E_s)_{do30}$	0.46
$(t_{c0})_{do20}$	1.10 $\mu\text{s}$	$(t_{s0})_{do30}$	1.7 $\mu\text{s}$

式(6)から得られる並列化率はほぼ100%である。do 30の計算時間が無視できるためである。

図-3に倍率を示す。○は $(T_{1pe})_{do10,20}/(T_{npe})_{do10,20}$ 、□は $(T_{1pe})_{do10,20} + (T_{1pe})_{do10,30} / [(T_{npe})_{do10,20} + (T_{1pe})_{do10,30} + (T_c)_{do10}]$ である。この並列化アルゴリズムの性能が、総和計算の性能に

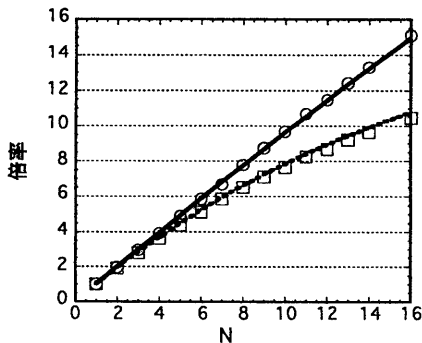


図-3 VPP500のスケラビリティ

依存していることが判る。尚、図中の記号が実測値、線がモデルからの計算値である。

#### 4.3 Paragonにおける検討

Paragonの並列化プログラムを図-4に示す。図-1に次の変更、追加を行う。iamは自プロセッサ番号、nodesはプロセッサ数である。 $l_r, l_c$ は各々、列と行のプロセッサ数の $\log_2$ をとったものである。tgdsunで、全プロセッサ間の総和計算を行う。尚、tmpは総和計算のためのワーク配列である。

```

dimension tmp(n)
iam=mynodes()
nodes=numnodes()
istat=nx_app_rect(l_r,l_c)
.....
do 20 k=iam+1, j-1, nodes
do 20 i=j,n
.....
20 continue
call tgdsun(s,n,tmp)

```

図-4 Paragonの並列化プログラム

総和計算を行うサブルーチンtgdsunの作成に当たっては、メッシュネットワークにおける衝突を避けるため、ネットワークの列方向と行方向の順にツリー状に転送を行う。データは、ツリーの頂点の1プロセッサに集めながら総和計算され、計算結果は来た道を逆戻りする形で配布される。これを、式(13)のようにモデル化する。

$$(T_c)_{do10} = \{ 2 [(t_{c0})_{tgdsun} + n(l_r+l_c) \cdot (4_{\text{byte}}) \cdot R_c^{-1} \cdot (E_c^{-1})_{tgdsun}] + (t_{s0})_{tgdsun} + n(l_r+l_c) \cdot R_a^{-1} \cdot (E_s^{-1})_{tgdsun} \} n \quad (13)$$

ここに、 $N=2^{l_r+2^{l_c}}$ である。

ここで、右辺の第一、二項は転送時間を表し、係数2はデータ転送が往復行われることを表す。

また第三、四項の総和計算は、往路のみ行われる。また、大括弧の右にあるnは、総和計算が、do 10のイテレーションでn回行われることを示す。

測定のため、サビ関数clockをdo 20、do 30、tgdsunの前後に挿入した。時間測定は、 $n$ を500から2500まで変え、またNを1から64まで変えた。コンパイラは、PGFTN(Version Rel 4.5)を用い、コンパイルオプションは、-nx -mvecdを用いた。-mvecdはベクトル計算を行うオプションであり、tgdsunには、必須であった。

測定した結果に、式(7)~(9)及び(11)をフィッティングさせた結果を表-2に示す。表中の $E_s$ は、この計算でParagonの1プロセッサの効率が0.4~34%とばらついていることがわかる。尚、計算には $R_a=0.1\text{Gflops}$  (単精度)、 $R_c=200\text{MB/秒}$ を用いた。

表-2 Paragonのモデルパラメータ値

条件	n=2500	条件	N=1
$(E_p)_{do20}$	0.983	$(E_s)_{do20}$	0.338
$(t_{p0})_{do20}$	$0.40 \mu s$	$(t_{s0})_{do20}$	$0.94 \mu s$
$(E_c)_{tgdsun}$	0.113	$(E_s)_{do30}$	0.00380
$(t_{c0})_{tgdsun}$	$0 \mu s$	$(t_{s0})_{do30}$	$.0017 \mu s$
$(E_s)_{tgdsun}$	0.0684		
$(t_{s0})_{tgdsun}$	$0 \mu s$		

式(6)の並列化率Pは90%で、倍率 $1/(1-P)$ は10倍である。

図-5に倍率を示す。○は $(T_{1pe})_{do10,20} / (T_{npe})_{do10,20}$ 、+は $[(T_{1pe})_{do10,20} + (T_{1pe})_{do10,30}] / [(T_{npe})_{do10,20} + (T_{1pe})_{do10,30}]$ 、□は $[(T_{1pe})_{do10,20} + (T_{1pe})_{do10,30}] / [(T_{npe})_{do10,20} + (T_{1pe})_{do10,30} + (T_c)_{do10}]$ である。 $(E_s)_{do30}$ が極端に低いことが並列性に影響を与え最大性能を10倍にしている様子、更にtgdsunにおける総和計算がスケラビリティに影響し、採用した並列アルゴリズムとの相性を悪くしていることがわかる。尚、図中の記号が実測値、線がモデルからの計算値である。

### 5. 議論

モデルはロードバランス等の効果を取り入れる必要がある。また、1プロセッサで実行できないコードの、

$E_s, t_{s0}$ の定義の方法の検討が必要である。これらは、各アーキテクチャの計算機上で、シミュレーションコードを並列化することによって研究する予定である。

意味があり、精度の高いモデルを作成するためには、アプリケーション側の視点を持った人の協力が必須である。

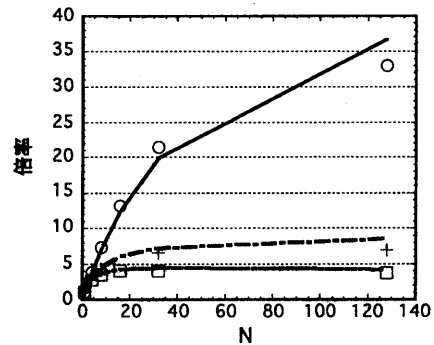


図-5 Paragonのスケラビリティ

### 6. 謝辞

Paragonの総和計算プログラムは、日本原子力研究所、計算科学技術推進センターの大田敏郎氏作成のものを利用した。感謝したい。

### 【参考文献】

- [1]Oriei S., Fully Vectorizable Particle-Mesh Model Codes on the FACOM VP-100/200, Proc. ANS/ENS/JAES Int. Topical Meeting, Knoxville, Tennessee, April 9-10, Vol. 1, pp.22 (1985)
- [2]Hockney R., Curington I,  $f_{1/2}$ :A parameter to characterize memory and communication bottlenecks, Parallel Computing Vol. 10, pp.277-286 (1989)
- [3]Formella A., Muller S., Paul W., Bingert A., Isolating the Reasons for the Performance of Parallel Machines on Numerical Programs, Automatic Parallelization, Wieweg Publishing (1994)
- [4]岡田、坂本、浅井：VPP500システムのソフトウェア、電子情報通信学会論文誌 D-I Vol. J78-D-1, No.2, pp.149-161 (1995)