

[ 大規模科学計算システム ]

## SX-Aurora TSUBASA でのプログラミング(並列化編)

### — 共有並列化と分散並列化 —

林 康晴 早坂 武  
日本電気株式会社

サイバーサイエンスセンターのスーパーコンピュータ AOBA-S の運用が開始されました。AOBA-S に採用されている SX-Aurora TSUBASA は、第三世代の Vector Engine(VE30A)カードを 8 枚搭載した Linux サーバー(Vector Host)が、InfiniBand の Fat-Tree ネットワークにより接続されたシステムです。各 VE カードは、16 個の CPU コアを搭載しています。SX-Aurora TSUBASA のハードウェア性能を十分に引き出すには、プログラムの並列化により、複数の CPU コアを有効に活用する必要があります。SX-Aurora TSUBASA は、並列処理環境としてプログラミング言語 Fortran、C、および C++、並びに 通信ライブラリ MPI を用意しています。本稿は、第 1 章で並列処理の基本事項をご説明した後、第 2 章では NEC Fortran コンパイラ(以後、単にコンパイラと記します)による共有並列化、第 3 章では MPI ライブラリ NEC MPI による分散並列化を中心にご紹介します。

## 1. 並列処理

並列処理とは、ひとつの仕事を複数の小さな仕事に分割し、それらの仕事を複数同時に実行することです。例えば、2 重ループを 4 つの仕事に分割し、4 個の CPU コア上で並列実行すると、図 1.1 のようになります。この場合、外側の do j のループの 100 回の繰返しを 4 等分し、各 CPU コア上で並列に実行しています。

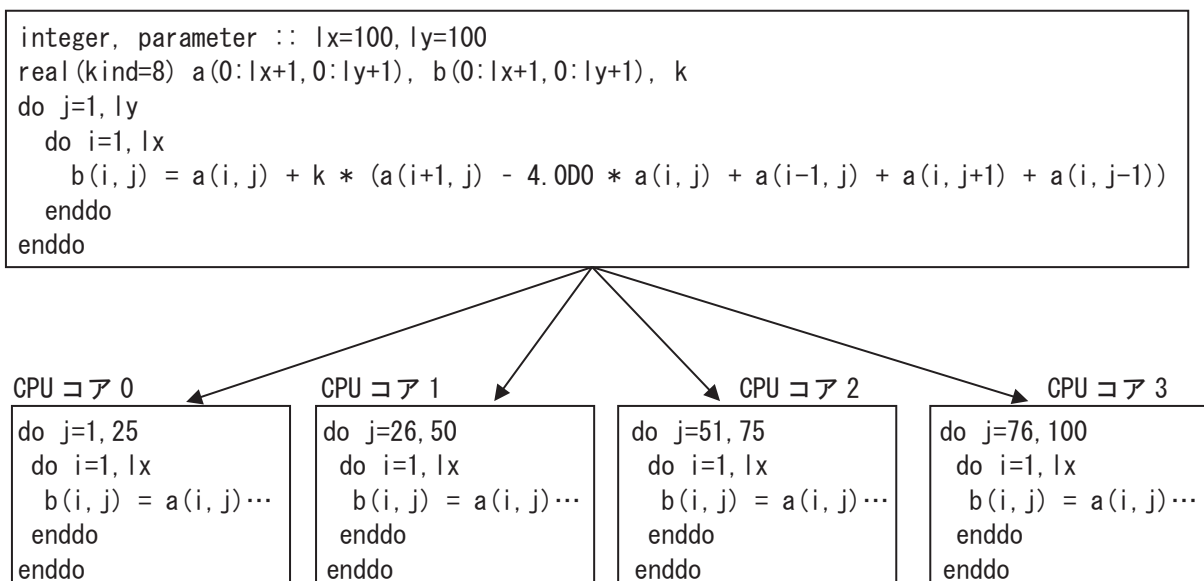


図 1.1 2 重ループの並列処理

### 1.1 並列化可能な条件

並列実行中、複数の仕事は、一般には実行タイミングの調整(同期)なしで、それぞれ独立に実行されます。そのため、異なる仕事内の計算の間には定まった実行順序はなく、一般には実行毎に異なる順序で実行されます。したがって、ループが並列実行可能であるためには、ループの繰返しをどのような順序で実行しても実行結果が変わらないことが必要です。例えば、図 1.2 の 2 つのループ(a)、(b)は、いずれもループ中で確定される各配列要素 a(i) および a(i,1), (i=2,3,...,n) が、各ループ(a)、(b)の全ての繰返

し中でそれぞれ 1 度しか出現しません。そのため、ループの繰返しをどのような順序で実行しても結果は同じになるので、並列実行可能です。一方、図 1.3 の(a)、(b)の場合、いずれもループの k 回目の繰返しで値が確定される配列要素 a(k)が、(a)の場合 k+1 回目の繰返しで、(b)の場合 k-1 回目の繰返しでそれぞれ引用されます。そのため、ループの繰返しの実行順序に依存して、確定前の値を引用するか、確定後の値を引用するかが変わってしまい、並列化することはできません。図 1.3 の(c)の場合も、ループの実行終了時に変数 ifound に残る値が、ループの繰返しの実行順序に依存して変わってしまうので、並列化することはできません。このように、ループを並列化するためには、ひとつの繰返しで確定したデータ要素は、他の繰返しでは確定も引用もしないようにループを記述する必要があります。図 1.3 の(d)のように、ループ外への飛越しを含むループも、ループの繰返しの実行順序に依存して飛越しのタイミングが変わってしまうので、並列化することはできません。

図 1.4 の 2 つのループ(a)、(b)は、それぞれスカラー変数 t、s をループの全ての繰返しで確定するので、図 1.3 の条件に該当し、一見並列化できないように思えます。しかし(a)の場合、変数 t は、ループの繰返し毎に新たに確定した値を、その繰返しの中だけで引用しています。そのため、ループ実行後に変数 t の値を引用しないのであれば、仕事毎に作業変数を確保して、並列実行中は、変数 t の代わりにその作業変数を参照すれば並列化可能です。また(b)のように、同一の変数に同一の演算を繰返し適用するパターンは集計計算と呼ばれます。集計計算の場合も、やはり仕事毎に作業変数を確保して、並列実行中はその作業変数に各仕事の局所的な計算結果を格納し、並列実行終了時に全ての仕事の計算結果を集計することにより並列化できます。

多重ループの場合、並列化できるかどうかはループ毎に判断します。図 1.5 の内側の do i のループは並列化可能ですが、外側の do j のループは、左辺の a(i,j)と右辺の a(i,j-1)との間に図 1.3(a)と同様の依存関係があるので並列化できません。

<pre>integer, parameter :: n=100 real(kind=8), dimension(n) :: a, b do i=2, n   a(i) = b(i) + b(i-1) enddo</pre>	<pre>integer, parameter :: n=100 real(kind=8), dimension(n, n) :: a do i=2, n   a(i, 1) = a(i, 2) + a(i-1, 2) enddo</pre>
(a)	(b)

図 1.2 並列化できるループ

<pre>integer, parameter :: n=100 real(kind=8), dimension(n) :: a do i=2, n   a(i) = a(i) + a(i-1) enddo</pre>	<pre>integer, parameter :: n=100 real(kind=8), dimension(n) :: a do i=1, n-1   a(i) = a(i) + a(i+1) enddo</pre>
(a) 依存	(b) 逆依存

<pre>integer, parameter :: n=100 real(kind=8), dimension(n) :: a integer ifound do i=1, n   if(a(i).ge.0) ifound = i enddo</pre>	<pre>integer, parameter :: n=100 real(kind=8), dimension(n) :: a do i=1, n   if(a(i).ge.0) goto 100 enddo 100 continue</pre>
(c) 出力依存	(d) 制御依存

図 1.3 並列化できないループ

<pre>integer, parameter :: n=100 real(kind=8) :: a(n), b(n), t do i=1,n-1   t = a(i) + a(i+1)   b(i) = t enddo</pre>	<pre>integer, parameter :: n=100 real(kind=8) :: a(n), s do i=1,n   s = s + a(i) enddo</pre>
--	--

(a) 繰返し内作業変数

(b) 集計計算

図 1.4 作業変数の使用により並列化できるループ

```
integer, parameter :: n=100
real(kind=8), dimension(n,n) :: a
do j=2,n
  do i=1,n
    a(i,j) = a(i,j-1) + a(i,j)
  enddo
enddo
```

図 1.5 内側ループだけ並列化できる多重ループ

## 1.2 並列化の効果

ひとつのプログラムを N 個の CPU コア上で並列実行した場合、最大 N 倍のスピードアップ(実行時間の短縮)が期待できます。このスピードアップを最大にするには、まず、どれだけ多くの部分を並列化できたか(並列化率)が重要です。これは、図 1.6 のように逐次実行時間の 10 %分の仕事が並列化できない場合、残りの仕事をどれだけ多くの小さな仕事に分割して並列化しても、スピードアップは 10 倍未満になってしまうことから分かります。

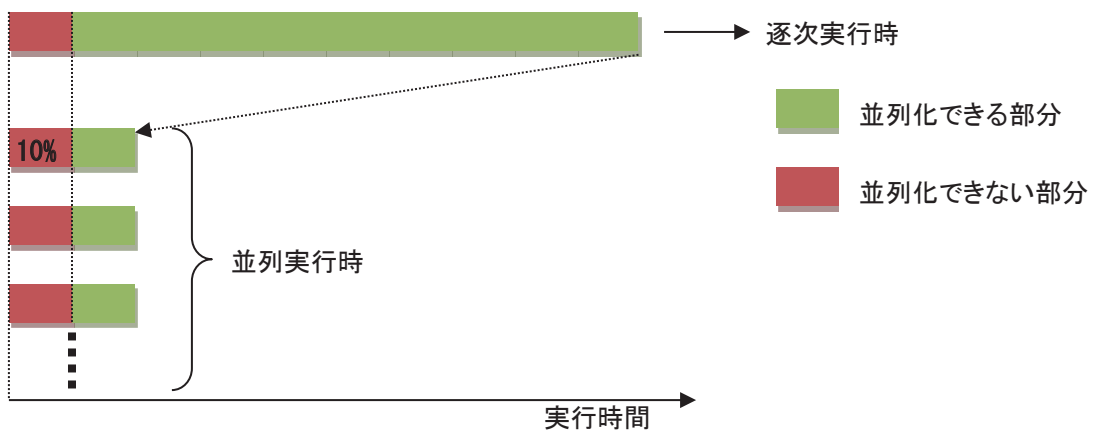


図 1.6 並列化率とスピードアップ(並列化率 90 %・スピードアップは 10 倍未満)

また、仕事を複数の小さな仕事に分割するための処理や、仕事間のデータのやり取り(通信)および同期といった逐次実行時には必要なかった処理(オーバーヘッド)をできるだけ小さくすることも重要です。これは、図 1.7 のように逐次実行時間の 10 %分のオーバーヘッドが並列化により発生すると、スピードアップはやはり 10 倍未満になってしまうことから分かります。特に、実行時間(粒度)の小さなループを並列化するような場合、オーバーヘッドの方が大きすぎ、並列化によりかえって遅くなる場合もあります。オーバヘ

ッドを削減するには、できるだけ外側のループで並列化する、といった方法により、各仕事の粒度をできるだけ大きくすると効果的です。

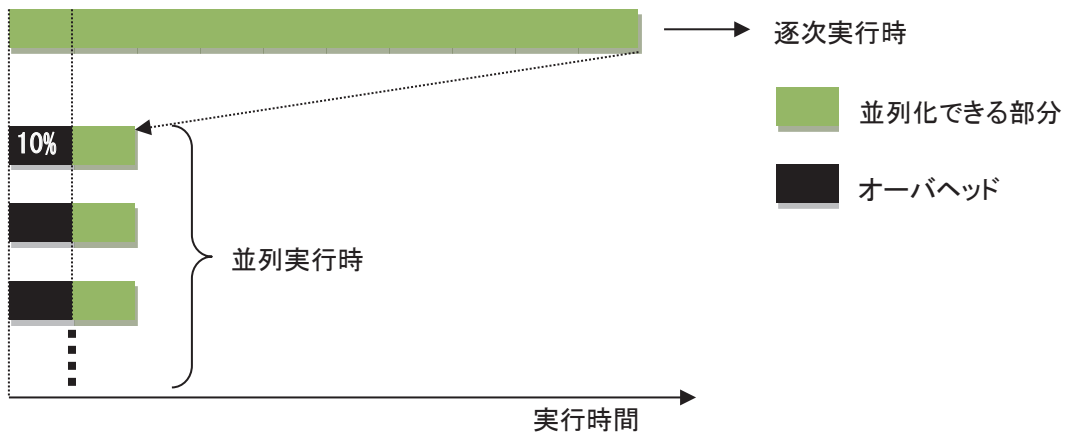


図 1.7 オーバヘッドとスピードアップ (オーバーヘッド 10 %・スピードアップは 10 倍未満)

さらに、各仕事の実行時間のバランス(負荷バランス)をできるだけ均等化することも重要です。これは、図 1.8 のように複数の仕事のうちひとつの仕事の実行に逐次実行時の 10 %分の時間がかかってしまったとすると、残りの仕事をどれだけ高速化しても、スピードアップは 10 倍以下になってしまうことからわかります。

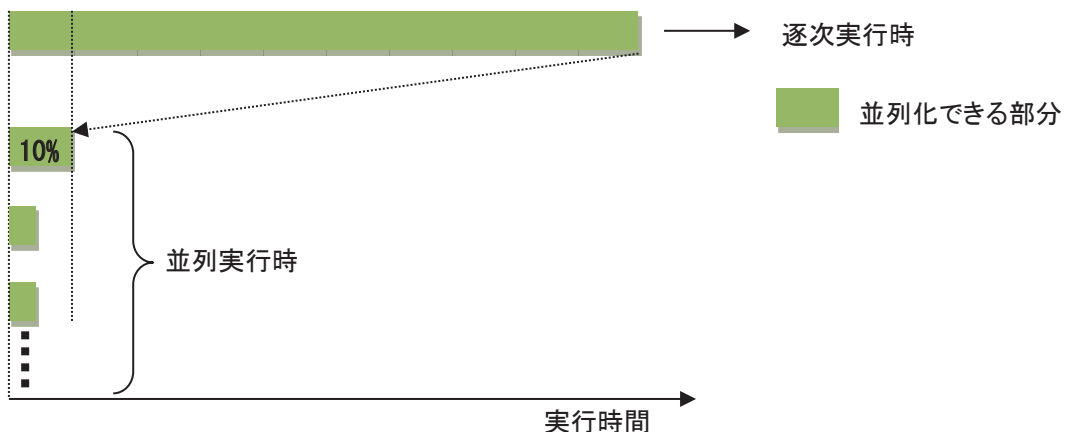


図 1.8 負荷バランスとスピードアップ (スピードアップは 10 倍以下)

このように、並列プログラムにおいて高いスピードアップを達成するには、高い並列化率・低いオーバーヘッド・均等な負荷バランスの 3 点が重要です。

### 1.3 経過時間と CPU 時間

並列処理は、ひとつの仕事を複数の小さな仕事に分割して、複数の CPU コア上で同時に実行することにより、仕事の実行時間(経過時間)を減らします。一方、分割した全ての仕事を実行するための CPU コアの処理時間の総計(CPU 時間)を減らすことはできません。実際には、ひとつの仕事複数の小さな仕事に分割するための処理や仕事間の同期などのオーバーヘッドも発生するので、CPU 時間は並列化によりむしろ増加することに注意してください。

## 1.4 経過時間とチューニング

並列化の主たる目的であるプログラム実行の経過時間短縮のためには、アルゴリズムの工夫、高速なライブラリの使用、およびベクトル化の促進などにより、逐次プログラムとしての性能を並列化の前に十分チューニングしておくことが重要です。並列化によるスピードアップは、使用する CPU コアの個数が上限となりますが、逐次プログラムのチューニングによる高速化は、場合により数十～数百倍に達することもあり、さらに CPU 時間も削減できるからです。

## 2. 共有並列化

一般に、コンピュータのプログラムは、プロセスにより実行されます。実行中のプロセスは、そのメモリ空間にプログラムのデータを格納しており、現在の実行状態をもっています。プロセスは、複数のスレッドから構成することもできます。ひとつのプロセス中の複数のスレッドは、それぞれ独自の実行状態をもち、別々の仕事を実行できますが、プロセスのメモリ空間は全てのスレッドが共有します。

共有並列化とは、ひとつのプロセス中の複数のスレッドによる並列処理のことです。AOBA-S の各 VE カード内では、主記憶装置を共有する 16 個の CPU コア上で実行されるスレッドに仕事を割り当て、共有並列化を行うことができます。例として、図 1.1 の 2 重ループの共有並列化を考えます。第 1 章と同様に、外側の `do j` のループを 4 つの仕事に分割して別々のスレッドに割り当てます。この場合、データ領域に関しては、図 2.1 のように、2 次元配列 `a, b` を `do j` のループに対応する 2 次元目で分割し、分割された各部分領域の処理を各スレッドが担当することになります。ここで、例えば `j=25` の繰返しを担当するスレッド 0 は、スレッド 1 の担当領域である配列要素 `a(i,26)`, ( $i=1,2,\dots,lx$ ) の値を引用する、といったように、配列の分割境界部分の処理では、自スレッドの担当領域外の配列要素を参照する必要があります。しかし、各スレッドは、メモリ上に配置された 2 次元配列 `a, b` の全ての領域を直接参照できるので、特に問題はありません。一方、DO 変数 `i, j` については注意が必要です。並列実行中、各スレッドは変数 `i, j` をそれぞれ独自のタイミングで参照します。そのため、配列 `a, b` と同様に、変数 `i, j` の領域を全てのスレッドが共有すると、あるスレッドが確定した変数 `i, j` の値を別のスレッドが引用してしまう可能性があり、実行が正しく行えません。従って、並列実行中に各スレッドが値を確定する変数は、図 2.2 のようにそれぞれのスレッドが固有の作業領域を割り付けて参照する必要があります。2 次元配列 `a, b` のような全てのスレッドがメモリ領域を共有するデータを SHARED データ、DO 変数 `i, j` のような各スレッドが専用のメモリ領域を割り付けるデータを PRIVATE データと呼びます。

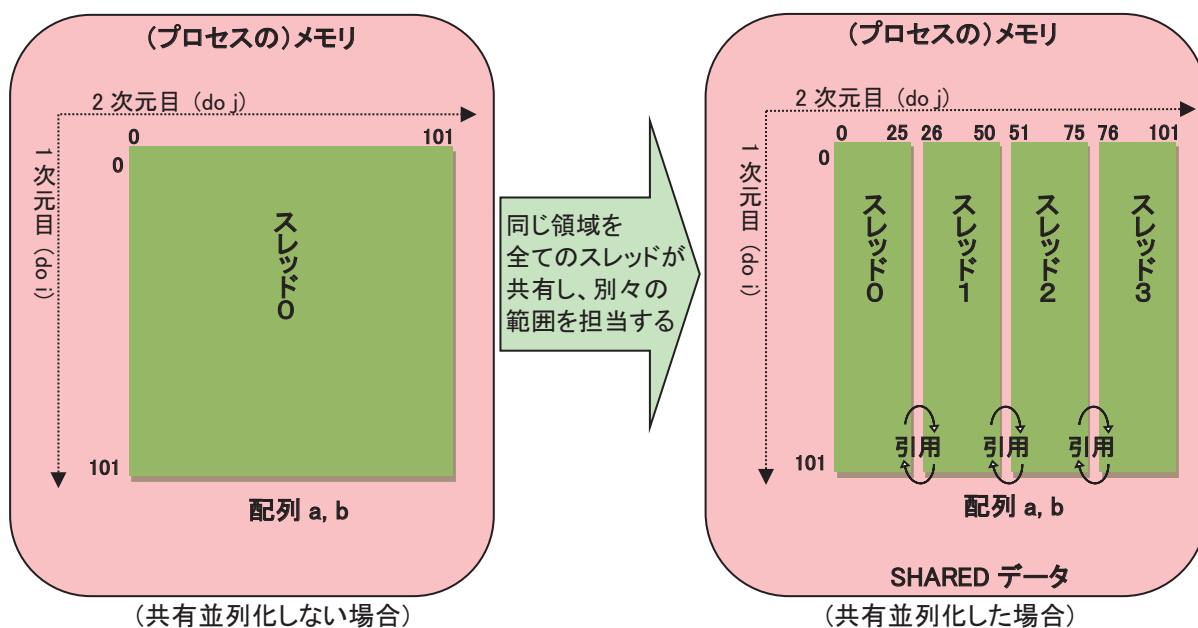


図 2.1 各スレッドが担当する配列 `a, b` の領域

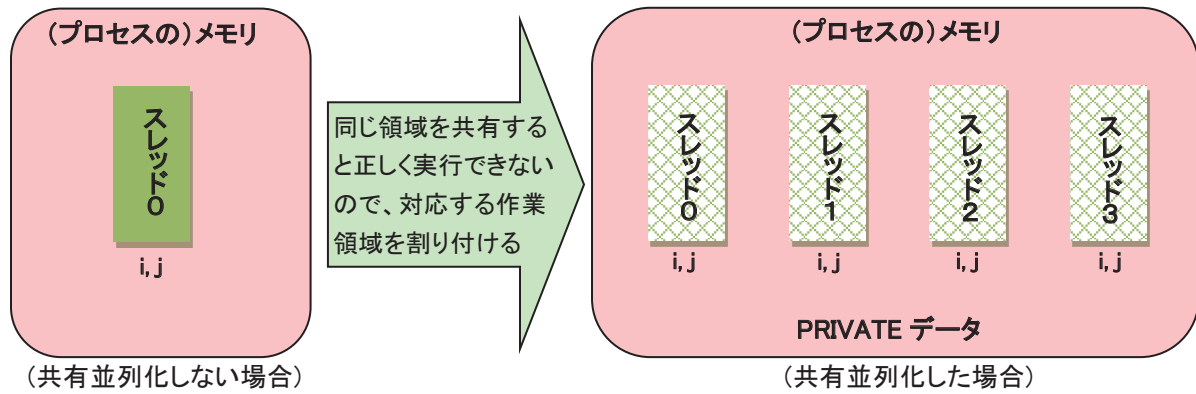


図 2.2 各スレッド専用の DO 変数 i, j 用の作業領域

## 2.1 自動並列化

並列化を行う場合、並列実行しても結果が不正にならないことを保証するために、通常はデータの依存関係の解析を行い、細心の注意を払ってプログラムの変形や指示文の挿入を行う必要があります。しかし、コンパイラの自動並列化機能を使用すると、コンパイラオプション `mparallel` を翻訳時に指定するだけで、それらの作業をコンパイラが自動的に行います。

コンパイラは、まずプログラムを解析して並列実行可能なループや文の集まりを抽出し、次にそれらを複数の仕事に分割してスレッドに割り当て、経過時間を短縮します。また、SHARED データ・PRIVATE データの判定も自動的に行います。コンパイラの自動並列化対象は、図 2.3 のとおりです。

対象となる構文	DO ループ、配列式
対象ループ中に書ける文	代入文、IF 構文、GOTO 文、CONTINUE 文、CALL 文、CASE 構文
対象となる演算	四則演算、べき乗演算、論理演算、関係演算、型変換、組込み関数

図 2.3 コンパイラの自動並列化対象

## 2.2 コンパイラの最適化

ここでは、共有並列化の効果を高めるためにコンパイラが行う最適化をいくつかご紹介します。

### 2.2.1 ループ変形

ループ融合・ループ一重化などのループ変形による最適化が可能な場合、最適化後に共有並列化を行い、共有並列実行時の仕事の粒度を最大化します。図 2.4 にループ融合の例を示します。

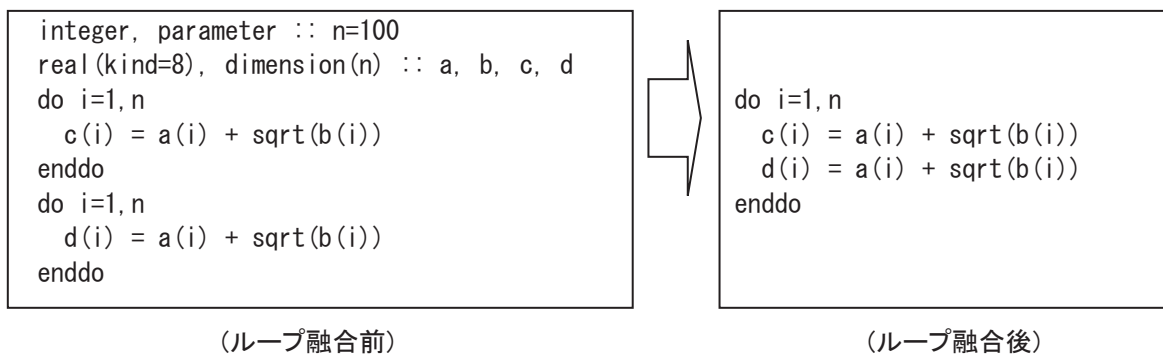


図 2.4 ループ融合の例

## 2.2.2 条件並列化

並列化できるかどうか、または並列化により高速化できるかどうかを翻訳時に判断できない場合、実行時に依存関係やループ長を調べて、共有並列化するかどうかを選択できるように条件並列化を行います。図 2.5 では、条件並列化後の IF 構文の論理式  $nx*ny > n$  によって、共有並列化により高速化するのに十分なループ長があるかどうかを判定しています。この際、ループ中の各演算の演算コストに基づいて、共有並列化の効果が期待できる値 (n) を自動的に算出し、条件並列化を行います。また、論理式  $id-ic==0 .or. abs(id-ic)>=nx$  によって、ループ中で確定される配列要素  $y(ic+i)$  と  $y(id+i)$  の領域に重なりがなく、並列化可能であるかどうかを判定しています。

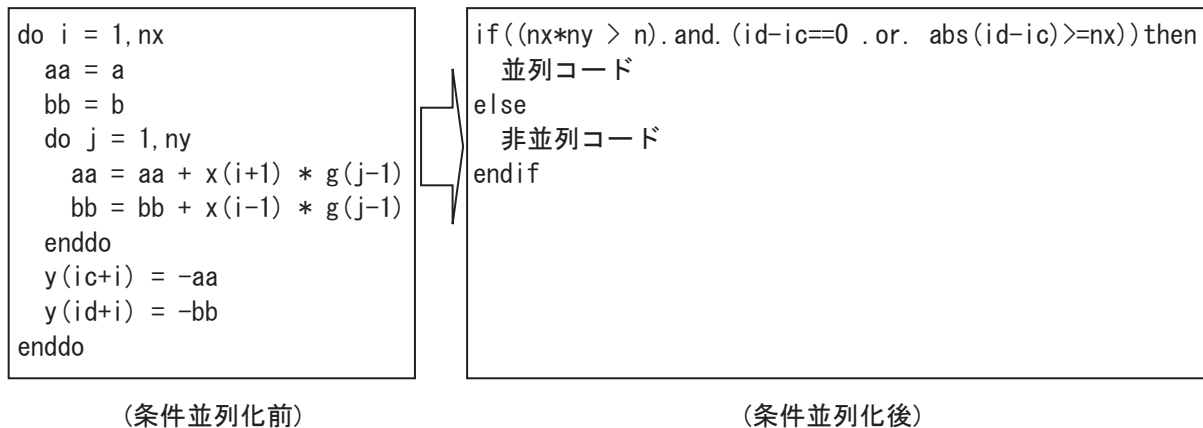


図 2.5 条件並列化の例

## 2.3 自動並列化促進のためのプログラミング

コンパイラオプション `-mparallel` の指定だけでは自動並列化できない場合でも、コンパイラ指示行の挿入やプログラムの修正により自動並列化できる場合もあります。ここでは、主な並列化用指示行とプログラムの修正例をご紹介します。

### 2.3.1 並列化指示行

コンパイラ指示行の書式は、図 2.6 のとおりです。

Fortran の場合:

```
!NEC$ コンパイラ指示オプション
```

C・C++ の場合:

```
#pragma _NEC コンパイラ指示オプション
```

図 2.6 コンパイラ指示行の書式

自動並列化のために用意されている主なコンパイラ指示オプションは、次のとおりです。

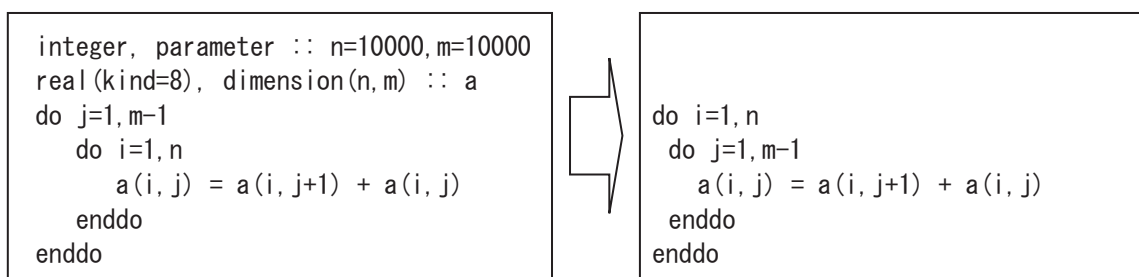
- `concurrent [schedule(種別[, 繰返し回数])] / noconcurrent`  
直後のループの自動並列化を許可する / しないを指定します。コンパイラ指示オプション `concurrent` には、省略可能な `schedule` 指示句を指定して、ループの繰返しをどのようにスレッドに割り当てるかを指定できます。種別として指定できるのは、`static`、`dynamic`、または `runtime` です。 `static`

は、ループの繰返しを、スレッド番号の昇順にラウンドロビン方式で割り当てます。dynamic は、ループの繰返しを、その時点で仕事が割り当たっていないスレッドに割り当てます。runtime は、実行時に環境変数 OMP\_SCHEDULE で指定された種別に従います。static および dynamic には、一度に割り当てるループの繰返し回数を指定できます。コンパイラ指示オプション noconcurrent は、粒度が小さく共有並列化するとかえって性能が低下するループに指定します。

- inner / noinner  
内側ループまたは一重ループを自動並列化の対象とする / しないを指定します。多重ループの内側ループは、既定値では自動並列化の対象にならないので、共有並列化したい場合、コンパイラ指示オプション inner を指定します。一重ループは、共有並列化の効果が翻訳時に不明の場合、自動並列化の対象になりませんが、コンパイラ指示オプション inner を指定することにより自動並列化の対象とすることができます。
- select\_concurrent  
多重ループの内、直後のループを優先して自動並列化します。
- nosync  
ループ中で参照される配列に依存がないことを指定します。依存関係が翻訳時には分からないので自動並列化できないときでも、依存がないことを利用者が分かっている場合、コンパイラ指示オプション nosync によりそれをコンパイラに教えてやることによって、自動並列化できる場合があります。
- cncall  
Fortran の組込み関数以外の手続引用を含むループは、既定値では自動並列化の対象になりませんが、並列化しても問題ないことを利用者が分かっている場合、コンパイラ指示オプション cncall を指定すると自動並列化の対象とすることができます。自動並列化されたループ中で引用される手続の仮引数は、結合する実引数が SHARED データの場合、原則として SHARED データとなるので、そのような仮引数を手続中で確定した結果、図 1.3 のような依存関係が発生すると結果が不正となることに注意してください。

### 2.3.2 並列化のためのソース変形

図 2.7(a)の 2 重ループは、外側の do j のループには依存があるので並列化できませんが、コンパイラ指示オプション inner を内側の do i のループに指定すると、内側ループで自動並列化されます。ただし、内側ループを並列化すると、並列化のオーバーヘッドが外側ループの繰返し回数分発生してしまいます。このような場合、図 2.7(b)のように内側ループと外側ループを利用者が交換すると、外側ループで自動並列化され並列化のオーバーヘッドを削減できます。



(a) ループ交換前

(b) ループ交換後

図 2.7 ループ交換の例



図 2.8(a)のループネストは、仮引数  $w$  に関するデータ依存により、外側の  $do\ j$  のループを自動並列化できません。このような場合、図 2.8(b)のように仮配列引数  $w$  を次元拡張し、 $do\ j$  のループの繰り返し毎に異なる要素を参照するように修正すれば、自動並列化可能となります。この際、仮引数  $w$  に対応する実引数にも同様の修正が必要となることに注意してください。また、手続  $sub$  の引用元が、仮引数  $w$  に対応する実引数を参照しておらず、配列  $w$  が単なる作業変数として利用されている場合は、図 2.8(c)のように仮引数  $w$  を手続  $sub$  の局所変数に変更すると、ループネスト中ではコンパイラにより PRIVATE データとして割り付けられるので、自動並列化できます。

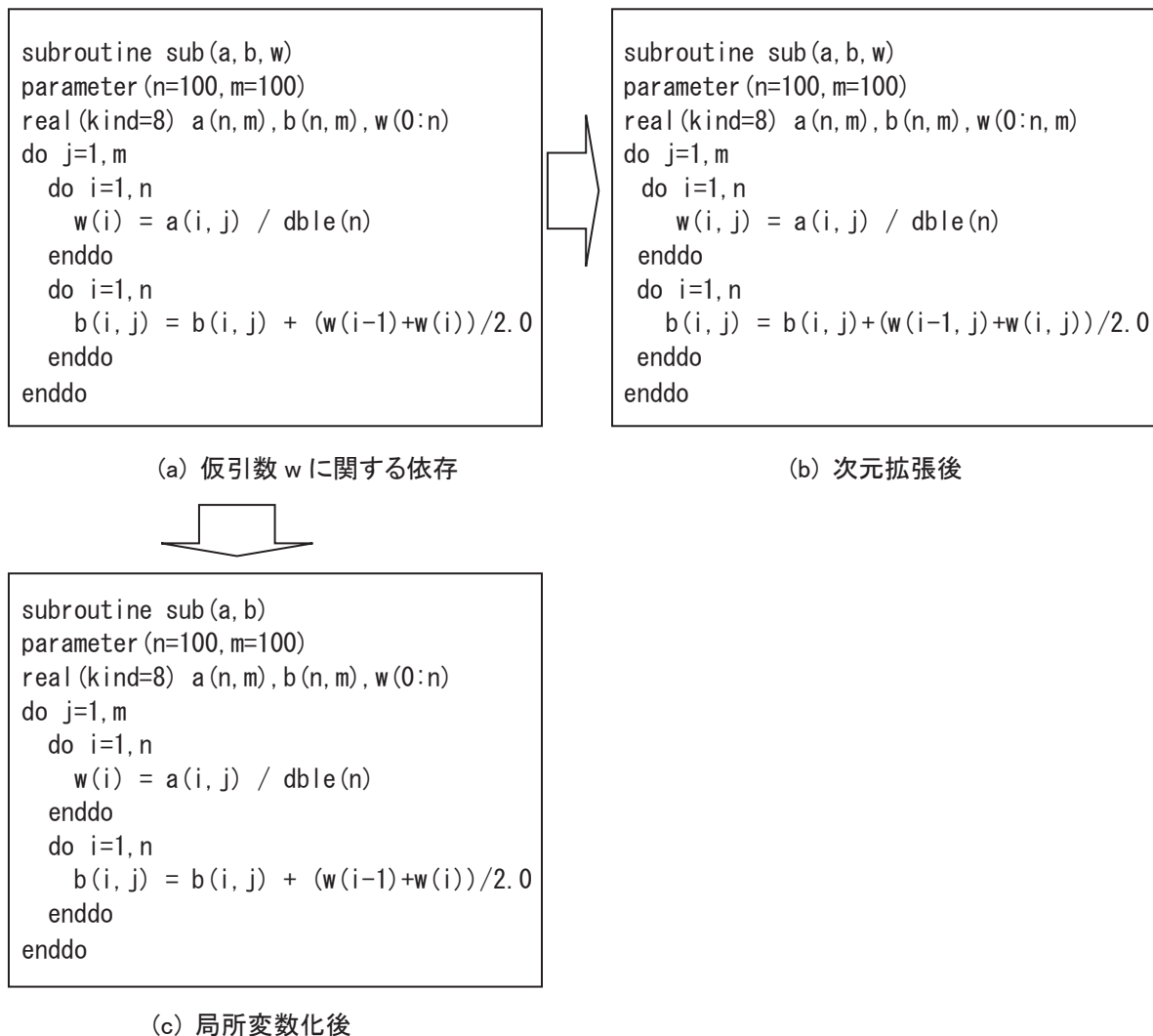


図 2.8 仮引数の修正により自動並列化可能となる例

## 2.5 OpenMP による共有並列化

コンパイラは、OpenMP による手動並列化も用意しています。OpenMP は、the OpenMP Architecture Review Board (ARB)<sup>[1]</sup> によって策定された共有並列化のためのオープンな API であり、共有並列化のデファクトスタンダードとして広く使用されています。OpenMP では、主として、ソースプログラムに OpenMP ディレクティブを挿入することによって共有並列化を行います。利用者は、OpenMP ディレクティブに必要な応じて指示句を追加し、並列化方法を詳細に制御できます。OpenMP ディレクティブの書式は、図 2.9 のとおりです。

Fortran の場合：

```
!$OMP OpenMP ディレクティブ
```

C・C++の場合：

```
#pragma omp OpenMP ディレクティブ
```

図 2.9 OpenMP ディレクティブの書式

SX-Aurora TSUBASA で OpenMP を使用する場合、コンパイラオプション `-fopenmp` を翻訳時に指定してください。プログラム実行の際は、1 つのプロセスが利用するスレッド数を環境変数 `VE_OMP_NUM_THREADS` で指定してください。

### 2.5.1 OpenMP 使用時の注意点

自動並列化の場合、コンパイラがプログラムを解析し並列ループ間の不必要な同期を自動的に抑制します。一方 OpenMP による共有並列化時には、`nowait` 指示句を指定しない限り並列ループの直後に毎回同期が生成されます。同期が必要ない場合には、`nowait` 指示句を指定すると共有並列化のオーバーヘッドを削減できます。

また自動並列化の場合、図 1.4 のような作業変数や集計変数は、コンパイラによって自動的に認識され適切に処理されます。一方 OpenMP による共有並列化時には、DO 変数以外の作業変数および集計変数をそれぞれ `private` 指示句および `reduction` 指示句中に指定する必要があります。

OpenMP の `parallel do` 指示文をループに指定して共有並列化した例を図 2.10 に示します。繰返し内作業変数 `t` を `private` 指示句中に、集計演算子 `+` および集計変数 `s` を `reduction` 指示句中に指定することに注意してください。なお、DO 変数 `i` は、この場合既定値で `PRIVATE` データとなるので、`private` 指示句を指定する必要はありません。

```
integer, parameter :: n=100
real(kind=8) :: a(n), t, s
!$OMP parallel do private(t) reduction(+: s)
do i=1,n-1
  t = a(i) + a(i+1)
  s = s + t
enddo
```

図 2.10 OpenMP による並列化例

## 2.6 性能解析

### 2.6.1 プログラム実行性能情報 (PROGINF)

逐次 (ベクトル) プログラムの場合と同様に、共有並列プログラムの場合も、プログラム実行性能情報 (PROGINF) が使用できます。PROGINF は、環境変数 `VE_PROGINF` の値を `YES` または `DETAIL` に設定してプログラムを実行することによって採取できます。図 2.11 に、環境変数 `VE_PROGINF` の値を `DETAIL` に設定して共有並列実行した場合の出力例を示します。ここで図中の `※` は、共有並列実行時に固有の情報です。Max Active Threads の値により、同時に実行したスレッドの個数の最大値を確認できます。また、図中 `◇` は、`VE30A` から追加された情報であり、レベル 3 キャッシュの利用率情報を出力します。

Real Time (sec)	:	18.914921	
User Time (sec)	:	297.227732	
Vector Time (sec)	:	204.714487	
Inst. Count	:	225908581946	
V. Inst. Count	:	60747556381	
V. Element Count	:	15380715794317	
V. Load Element Count	:	4512736316356	
FLOP Count	:	6217859387690	
MOPS	:	58868.836694	
MOPS (Real)	:	925048.247358	
MFLOPS	:	20919.512947	
MFLOPS (Real)	:	328723.308860	
A. V. Length	:	253.190691	
V. Op. Ratio (%)	:	99.056085	
L1 Cache Miss (sec)	:	14.285157	
CPU Port Conf. (sec)	:	0.000000	
V. Arith. Exec. (sec)	:	132.062536	
V. Load Exec. (sec)	:	64.239533	
LD L3 Hit Element Ratio (%)	:	39.529025	◇
VLD LLC Hit Element Ratio (%)	:	42.832553	
FMA Element Count	:	1951573996260	
Power Throttling (sec)	:	0.000000	
Thermal Throttling (sec)	:	0.000000	
Max Active Threads	:	16	※
Available CPU Cores	:	16	※
Average CPU Cores Used	:	15.713929	※
Memory Size Used (MB)	:	2778.000000	
Non Swappable Memory Size Used (MB)	:	102.000000	

図 2.11 共有並列実行時の PROGINF

## 2.6.2 性能解析機能(FTRACE 機能)

逐次(ベクトル)プログラムの場合と同様に、共有並列プログラムの場合も、FTRACE 機能によって手続単位又は利用者が指定した区間単位の詳細な性能情報が取得できます。

基本的な使用方法は逐次プログラムの場合と同様ですが、コンパイラは、共有並列化する各ループネストを切り出して、それらを独立した手続に変形します。そのため共有並列化された各ループネストは、FTRACE の表示上は独立したひとつの手続として表示されることに注意してください。切りだされた各ループネストは、ソースプログラム中の出現順に、元の手続名に加えて\_ $\$1$ 、\_ $\$2$ 、 $\dots$ とサフィックスが付いた名前の手続として表示されます。図 2.12 に、共有並列実行時の FTRACE の表示例を示します。\_ $\$1$  のついた手続 JACOBI $\$1$  は、手続 JACOBI の最初に出現する共有並列化されたループネストの性能情報です。-thread0、-thread1、 $\dots$ は、各スレッドの性能情報です。JACOBI のような $\$$ のついていない手続名は、手続 JACOBI 中の共有並列化されなかった部分の性能情報です。図中の「LD L3 HIT E.%」は VE30A から追加された情報であり、レベル 3 キャッシュの利用率情報を出力します。

FREQUENCY	EXCLUSIVE TIME[sec]( % )	AVER. TIME [msec]	MOPS	MFLOPS	V. OP RATIO	AVER. V. LEN	VECTOR TIME	L1CACHE MISS	CPU PORT CONF	LD L3 HIT	VLD LLC E. %	PROC. NAME
32	246.090(100.0)	7690.305	71014.7	25266.6	99.18	253.2	204.711	14.283	0.000	39.53	42.83	JACOBI\$1
2	15.382( 6.3)	7691.243	72129.3	25664.6	99.18	253.2	13.167	0.856	0.000	39.34	41.86	-thread0
2	15.380( 6.2)	7689.957	72142.2	25668.9	99.18	253.2	13.030	0.887	0.000	39.52	43.15	-thread1
2	15.380( 6.2)	7689.916	72141.6	25669.0	99.18	253.2	13.111	0.890	0.000	39.50	43.01	-thread2
2	15.381( 6.3)	7690.571	72136.3	25666.8	99.18	253.2	13.104	0.905	0.000	39.52	42.94	-thread3
:												
2	0.000( 0.0)	0.010	383.1	0.0	0.00	0.0	0.000	0.000	0.000	46.91	0.00	JACOBI

図 2.12 共有並列実行時の FTRACE

主に浮動小数点演算を実行する共有並列プログラムの性能を分析する場合、FTRACE 中の MFLOPS の値に着目してください。図 2.11 の場合、並列ループネスト JACOBI\$1 は、4 個のスレッドそれぞれが約 26000MFLOPS の性能で実行したことが分かります。MFLOPS 値がスレッド間でほぼ均等となっているので、うまく共有並列化されていると判断できます。逆に図 2.13 のように、スレッド 0 およびスレッド 1 だけ MFLOPS 値が大きく、残りのスレッドが小さい場合は、負荷バランスが悪い、と判断できます。このようなプログラムは、コンパイラ指示オプション concurrent の schedule 指示句などを指定して負荷バランスを均等化すると、経過時間を短縮できる可能性があります。

FREQUENCY	EXCLUSIVE TIME[sec]( % )	AVER. TIME [msec]	MOPS	MFLOPS	V. OP RATIO	AVER. V. LEN	VECTOR TIME	L1CACHE MISS	CPU PORT CONF	LD L3 HIT	VLD LLC E. %	PROC. NAME
32	676.916(100.0)	21153.626	11577.0	4082.8	98.30	253.2	81.652	4.896	0.000	41.05	43.12	JACOBI\$1
2	42.309( 6.3)	21154.469	90631.2	33179.7	99.24	253.2	38.506	2.139	0.000	42.71	44.64	-thread0
2	42.307( 6.2)	21153.667	87815.0	32144.0	99.24	253.2	37.900	2.142	0.000	42.37	43.45	-thread1
2	42.307( 6.2)	21153.285	490.9	0.0	73.98	254.0	0.400	0.036	0.000	0.01	0.01	-thread2
2	42.307( 6.2)	21153.746	490.9	0.0	73.98	254.0	0.421	0.044	0.000	0.04	0.04	-thread3

図 2.13 負荷バランスの悪い共有並列実行時の FTRACE

### 3. 分散並列化

分散並列化とは、それぞれ独自のメモリ空間をもつ複数のプロセスによる並列処理のことです。SX-Aurora TSUBASA では、ひとつまたは複数の VE カードの CPU コア上で実行されるプロセスに仕事を割り当て、分散並列化を行うことができます。

例として、図 1.1 の 2 重ループの分散並列化を考えます。第 1 章と同様に、外側の do j のループを 4 つの仕事に分割して別々のプロセスに割り当てます。この場合、データ領域に関しては、図 3.1 のように、2 次元配列 a、b を do j のループに対応する 2 次元目で分割し、分割された各部分領域の処理を各プロセスが担当することになります(分割配置・1 次元分割)。ここで、例えば j=25 の繰返しを担当するプロセス 0 は、プロセス 1 の担当領域である配列要素 a(i,26), (i=1,2,...,lx)の値を引用する、といったように、配列の分割境界部分の処理では、自プロセスの担当領域外の配列要素を参照する必要があります。ところが各プロセスは、他のプロセスに割り付けられた配列の領域を直接参照できないので、通信を行って必要な配列要素の値を取得しなければなりません。このように分散並列化の場合、共有並列化と異なり、他のプロセスの担当範囲のデータを参照するために、必要に応じて通信が必要です。一方 DO 変数 i、j については、図 3.2 のように各プロセスがそれぞれ専用の領域を割り付ける(重複配置)と、並列実行中、各

プロセスは変数  $i, j$  をそれぞれ独立に参照できるので通信の必要はありません。すなわち、分散並列実行時にどのような通信が必要となるかは、各データ要素をどのようにプロセスに割り付けるか(データマッピング)およびひとつの仕事をどのように複数の小さな仕事へと分割してプロセスに割り当てるか(計算マッピング)に依存します。このように分散並列化においては、データマッピング、計算マッピング、および通信の3点を全て考慮してプログラムする必要があります。

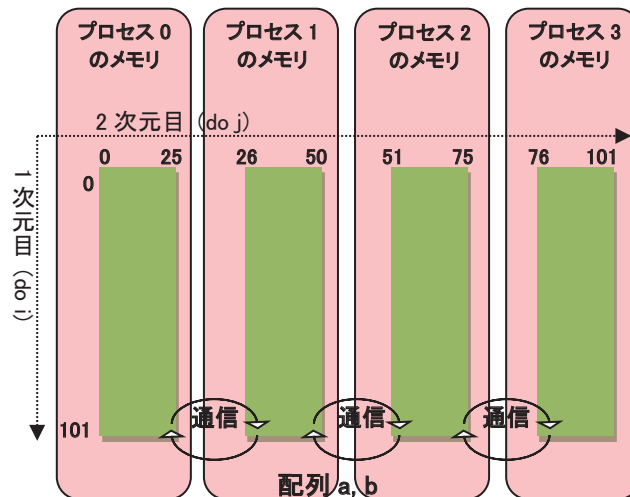


図 3.1 配列 a, b の分割配置(1次元分割)

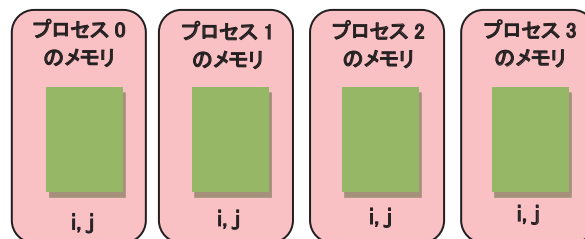


図 3.2 DO 変数  $i, j$  の重複配置

### 3.1 データマッピングと性能

再び図 1.1 の 2 重ループの分散並列化を考えます。今度は図 3.3 のように、外側・内側のループをそれぞれ 2 つの仕事に分割してみます。この場合、データ領域に関しては、図 3.4 のように 2 次元配列 a, b を 1 次元目・2 次元目ともに分割し、分割された各部分領域の処理を各プロセスが担当することになります(分割配置・2次元分割)。ここでも、例えば  $j=50$  の繰返しを担当するプロセス 0 は、プロセス 2 の担当領域である配列要素  $a(i, 51)$ , ( $i=1, 2, \dots, 50$ ) の値を引用する、といったように、配列の分割境界部分の処理では、自プロセスの担当領域外の配列要素を参照するために通信が必要です。

```
integer, parameter :: lx=100, ly=100
real(kind=8) a(0:lx+1,0:ly+1), b(0:lx+1,0:ly+1), k
do j=1, ly
  do i=1, lx
    b(i, j) = a(i, j) + k * (a(i+1, j) - 4.0D0 * a(i, j) + a(i-1, j) + a(i, j+1) + a(i, j-1))
  enddo
enddo
```

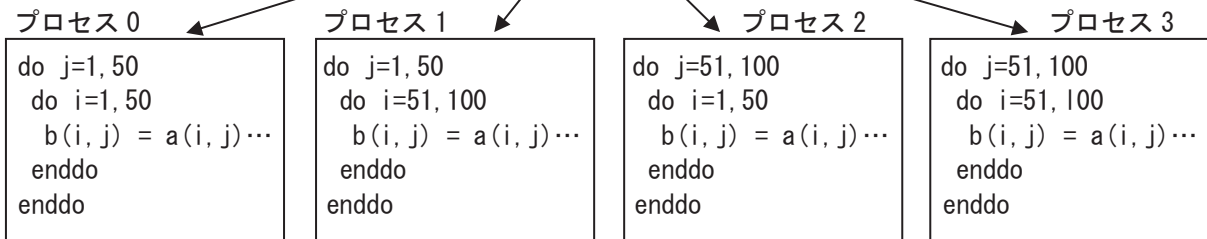


図 3.3 2つのループをともに並列化する例

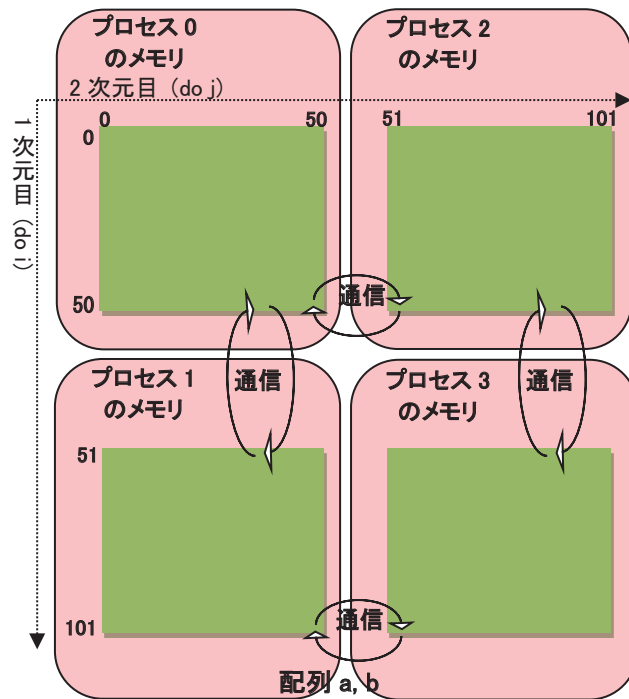


図 3.4 配列 a, b の分割配置(2次元分割)

これを前節の外側ループだけによる分散並列化(1次元分割)と比較してみます。各プロセスの通信に関しては、1次元分割の場合、送受信回数はそれぞれ2回(2次元目の上下方向)、送受信量はそれぞれ1次元目の寸法である  $lx \times 2$  となります。一方2次元分割の場合、送受信回数はそれぞれ4回(1・2次元目それぞれの上下方向)、送受信量はそれぞれ  $(lx/1 \text{次元目分割数} + ly/2 \text{次元目分割数}) \times 2$  となります。従って、1次元分割と比べると、2次元分割の方が通信回数が多いですが、通信量はプロセスの個数が増えるほど少なくなります。1回の通信時間は、典型的にはセットアップ時間+通信量/通信バンド幅と表せます。そのため、通信量が少なく(配列の寸法が小さく)セットアップ時間の影響が大きい場合には1次元分割の方が通信コストは少なく、逆に通信量が多く(配列の寸法が大きく)通信量の影響が大きい場合には2次元分割の方が通信コストは少なくなると考えられます。一方、各プロセスの演算性能を比較

すると、ベクトル長は内側の `do i` のループに対応する配列の 1 次元目の寸法に依存します。そのため、1 次元目をプロセス間に分割することによってベクトル長が短くなりすぎると、ベクトル演算の効率が低下します。従って、配列の寸法がそれほど大きくない場合、2 次元目だけを分割したほうが演算時間は短い可能性があります。このように、これら 2 つの分散並列化方法のどちらが経過時間を短縮できるかは、プロセスの個数や配列の寸法に依存して変わります。分散並列化時には、通信コストや並列化後の演算性能を考慮して、データマッピングを決定することが重要です。

## 3.2 Message Passing Interface (MPI)

### 3.2.1 MPI 概要

MPI は、Message Passing Interface Forum<sup>[2]</sup> によって策定された通信ライブラリのオープンな規格であり、分散並列プログラミングのデファクトスタンダードとして広く使用されています。MPI を使用した分散並列化では、図 3.1 や図 3.4 中の通信部分を、MPI 手続の引用によってプログラムすることになります。

分散並列化時に高いスピードアップを達成するには、通信コストの削減が特に重要です。プロセス内のメモリアクセスコストと比較して、プロセス間の通信コストは圧倒的に大きいからです。

SX-Aurora TSUBASA は、高速な通信環境を提供するため、MPI ライブラリ NEC MPI を用意しています。NEC MPI は、プロセスの VE カードへの配置方法や通信サイズに応じて、VE カード上の主記憶装置や InfiniBand のハードウェア機能を利用した最適な通信プロトコルを選択し、SX-Aurora TSUBASA の性能を最大限引き出しています。

### 3.2.2 MPI プログラミングの基本

MPI プログラミングでは、まず MPI 規格で規定されている各種定数などが定義されたヘッダファイル `mpi.h` をインクルード (C・C++ の場合)、あるいは、システムモジュール `mpi` または `mpi_f08` を引用します (Fortran の場合)。なお、Fortran の場合は、ヘッダファイル `mpif.h` をインクルードすることもできますが、`mpif.h` の利用は推奨されません。次に、手続 `mpi_init` または `mpi_init_thread` を引用して MPI の初期化を行い、手続 `mpi_finalize` を引用して MPI の使用を終了します。利用者は、MPI の初期化から使用終了までの間で、一対一通信、集団通信、または片側通信により通信を行うことができます。

ほとんどの MPI 手続では、コミュニケータと呼ばれる引数で、通信を行うプロセスの集合を指定します。プログラム開始時の全てのプロセスに対応するコミュニケータは、MPI の定数 `MPI_COMM_WORLD` です。各プロセスは、ランクと呼ばれるコミュニケータ中で一意な整数値をもちます。ランクの値は、0 以上、プロセス数-1 以下です。手続 `mpi_comm_size` および `mpi_comm_rank` により、それぞれプロセスの個数および自プロセスのランクを取得でき、これらを利用して各プロセスが行う処理をプログラムできます。

MPI プログラムの典型的な例として、図 1.1 の 2 重ループを外側ループで 4 つの仕事に分割し、分散並列化したプログラムを図 3.5 に示します。MPI プログラミングでは、プログラム全体のデータ・処理ではなく、各プロセス用に分割した後のデータ・処理をプログラムします。そのため、定数 `ly` の値を 25 に変更して (文番号 100)、配列の宣言 (文番号 200) とループ長 (文番号 300) を分割後の範囲に縮小しています。この際、通信時に受信したデータを配置する領域が必要なので、図 3.6 のように各プロセス上の配列 `a` の領域を大きめに割り付けていることに注意してください。このような隣接プロセスとの通信のための領域は袖領域と呼ばれます。並列ループ (文番号 300) 実行前に、分割境界の処理に必要な配列 `a` の要素を、一対一通信手続 `mpi_sendrecv` により隣接プロセス間で送受信 (文番号 400) し、並列ループ実行中は、隣接プロセス上のデータの代わりに自プロセス上の袖領域を参照しています。両端のプロセスには片側の隣接プロセスが存在しないので、通信相手を設定する IF 構文 (文番号 500) において、通信相手を示す変数 `men`、`mep` に、該当するプロセスが存在しないことを示す MPI の定数 `MPI_PROC_NULL` を設定しています。送受信先として定数 `MPI_PROC_NULL` を指定すると、その通信は実行されません。

```

use mpi
integer :: nproc, me, men, mep
100 integer, parameter :: lx=100, ly=25 ! ly は分割後の寸法
200 real(kind=8) a(0:lx+1,0:ly+1),b(0:lx+1,1:ly), k
! 初期化
call mpi_init(ierr)
! プロセスの個数を変数 nproc に取得する
call mpi_comm_size(MPI_COMM_WORLD, nproc, ierr)
! 自プロセスのランクを変数 me に取得する。
call mpi_comm_rank(MPI_COMM_WORLD, me, ierr)
:
! 通信相手の設定 : 両端のプロセスは特別処理を行う。
500 if (me .eq. 0) then
    men = MPI_PROC_NULL ! 下端のプロセスの下隣は存在しない
else
    men = me - 1 ! 下隣のプロセスのランク
endif
if (me .eq. nproc-1) then
    mep = MPI_PROC_NULL ! 上端のプロセスの上隣は存在しない
else
    mep = me + 1 ! 上隣のプロセスのランク
endif
! 隣接プロセス間の通信 : 上方向 → 下方向
400 call mpi_sendrecv(a(1, ly), lx, MPI_DOUBLE_PRECISION, mep, 0, a(1, 0), & ! 上方向
& lx, MPI_DOUBLE_PRECISION, men, 0, MPI_COMM_WORLD, MPI_STATUS_IGNORE, ierr)
call mpi_sendrecv(a(1, 1), lx, MPI_DOUBLE_PRECISION, men, 0, a(1, ly+1), & ! 下方向
& lx, MPI_DOUBLE_PRECISION, mep, 0, MPI_COMM_WORLD, MPI_STATUS_IGNORE, ierr)
! 並列化後のループ
300 do j=1, ly
    do i=1, lx
        b(i, j) = a(i, j) + k * (a(i+1, j)-4.0D0*a(i, j)+a(i-1, j)+a(i, j+1)+a(i, j-1))
    enddo
enddo
:
! 後始末処理
call mpi_finalize(ierr)
end
    
```

図 3.5 MPI プログラム例

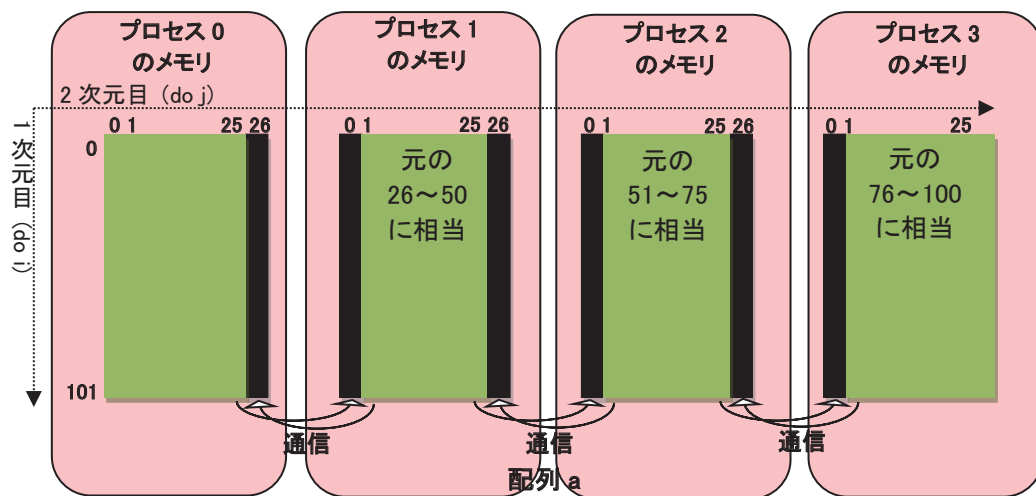


図 3.6 1次元分割時の配列 a の割付け (黒色は袖領域)



### 3.3 NEC MPI の活用方法と注意点

#### 3.3.1 MPI プログラムへの標準入力

MPI プログラムに標準入力から入力データを与える場合、図 3.7 のように、実行ファイルに対して直接入力データを与えると、正常に実行できない場合があります。

```
mpirun -np 4 /execdir/a.out < /datadir/inputfile
```

図 3.7 MPI プログラムへの不正な標準入力の例

このような場合、図 3.8 のように、コンパイラの変数 `VE_FORT5` に入力ファイル名を設定して事前接続した上で、MPI プログラムを実行してください。

```
export VE_FORT5=/datadir/inputfile
mpirun -np 4 /execdir/a.out
```

図 3.8 MPI プログラムへの標準入力

#### 3.3.2 プロセス毎に別々のファイルを対象にした入出力

環境変数 `MPIRANK` は、コミュニケータとして引数 `MPI_COMM_WORLD` を指定して、手続 `mpi_comm_rank` を呼び出すことにより得られるランクと同一の値をもっています。これを利用して、図 3.9 のようなシェルスクリプトを作成し(シェルスクリプトのファイル名を、例えば `/execdir/run.sh` とします)、図 3.10 のように、`mpirun` コマンドでそのシェルスクリプトを実行すると、各プロセスのランクと入出力ファイル名とを一対一に対応させ、プロセス毎に別々のファイルを対象にした入出力が可能です。

```
#!/bin/sh
export VE_FORT5=infile.$MPIRANK
export VE_FORT6=outfile.$MPIRANK

exec /execdir/a.out
```

図 3.9 環境変数 `MPIRANK` を利用した入出力のためのシェルスクリプト例

```
mpirun -np 4 /execdir/run.sh
```

図 3.10 シェルスクリプトを介した MPI プログラムの実行例

プログラム中で接続したファイルに対して入出力する場合は、手続 `mpi_comm_rank` により取得したランクを利用してファイル名を決めると、やはりプロセス毎に別々のファイルを対象にした入出力が可能です。

#### 3.3.3 プロセス毎に別々のファイルへの標準出力・標準エラー出力

複数の MPI プロセスが同時に標準出力または標準エラー出力を行った場合、それらは入り交じって表示される場合があります。各プロセスの標準出力および標準エラー出力を別々のファイルに容易に出力できるように、NEC MPI はシェルスクリプト `/opt/nec/ve/bin/mpisep.sh` を用意しています。このスクリプトを使用して、図 3.11 のように MPI プログラムを実行すると、環境変数 `NMPI_SEPSELECT` の実行時の値に応じて、次のように各プロセスの標準出力および標準エラー出力が別々のファイルに出力されます。なお下記の `u` は通常 0 となり、`r` は環境変数 `MPIRANK` の値となります。

- 環境変数 `NMPI_SEPSELECT` の値が 1 の場合  
各プロセスの標準出力が、ファイル `stdout.u:r` に出力されます。

- 環境変数 `NMPI_SEPSELECT` の値が 2 の場合 (既定値)  
各プロセスの標準エラー出力が、ファイル `stderr.u:r` に出力されます。
- 環境変数 `NMPI_SEPSELECT` の値が 3 の場合  
各プロセスの標準出力および標準エラー出力が、それぞれファイル `stdout.u:r` および `stderr.u:r` に出力されます。
- 環境変数 `NMPI_SEPSELECT` の値が 4 の場合  
各プロセスの標準出力および標準エラー出力が、ともにファイル `std.u:r` に出力されます。

```
mpirun -np 4 /opt/nec/ve/bin/mpisep.sh /execdir/a.out
```

図 3.11 シェルスクリプト `/opt/nec/ve/bin/mpisep.sh` を使用した実行例

### 3.3.5 MPI 集団手続デバッグ支援機能

NEC MPI は、実行時でなければ見つけにくい MPI プログラム特有のバグを検出する機能を用意しています。コンパイラオプション `-mpiverify` を指定して翻訳し、環境変数 `NMPI_VERIFY` の値を 3 または 4 に設定して実行すると、集団手続に対する呼出し順序の誤りや、引数のプロセス間の不整合などの情報が出力されます。

### 3.3.6 Fortran からの MPI 使用時の注意点

Fortran プログラム中で引用する非ブロッキングな MPI 手続の実引数として、部分配列、配列式、配列ポインタ、または形状引継ぎ配列を指定すると、引数がメモリ上連続であることを保証するために、コンパイラが一時配列を割り付けて実引数をコピーする場合があります。このとき、実際の引数としては、その一時配列が MPI 手続に渡されます。しかし、その一時配列は、非ブロッキングな MPI 手続の完了以前に解放されてしまうので、プログラムが正常に実行されない可能性があります。非ブロッキングな MPI 手続の実引数には、メモリ上連続な配列を指定してください。形状引継ぎ配列については、コンパイラオプション `-fassume-contiguous` を指定すると、コンパイラによる一時配列の生成を抑制できます。

## 3.4 ハイブリッド並列化

複数のプロセスだけによるプログラムの分散並列化をフラット並列化と呼びます。SX-Aurora TSUBASA では、分散並列化と第 2 章でご説明した共有並列化を併用することもできます。これをハイブリッド並列化と呼びます。ハイブリッド並列化は、MPI プログラムを自動並列化または OpenMP により共有並列化するだけで使用可能です。

図 3.5 では、図 1.1 の 2 重ループを外側ループで 4 つの仕事に分割し、4 プロセスに割り当てフラット並列化しました。これを、それぞれ 2 スレッドからなる 2 つのプロセスに割り当てハイブリッド並列化するには、図 3.5 中の定数 `ly` の値を 50 に変更した上で、自動並列化のためのコンパイラオプション `-mparallel` を指定して翻訳し、共有並列化します。さらに、環境変数 `VE_OMP_NUM_THREADS` の値を実行時に 2 に設定すると、各プロセスは 2 スレッドで実行されます。フラット並列化の場合、配列 `a` を各プロセス上に図 3.6 のように割り付けました。一方、ハイブリッド並列化の場合、配列 `a` を各プロセス上に図 3.12 のように割り付けることとなります。ひとつのプロセス中の複数のスレッドは、そのプロセスの全てのデータ領域を直接参照できるので、スレッド間で通信を行う必要はなく、従ってスレッド間には袖領域も必要ないことに注意してください。

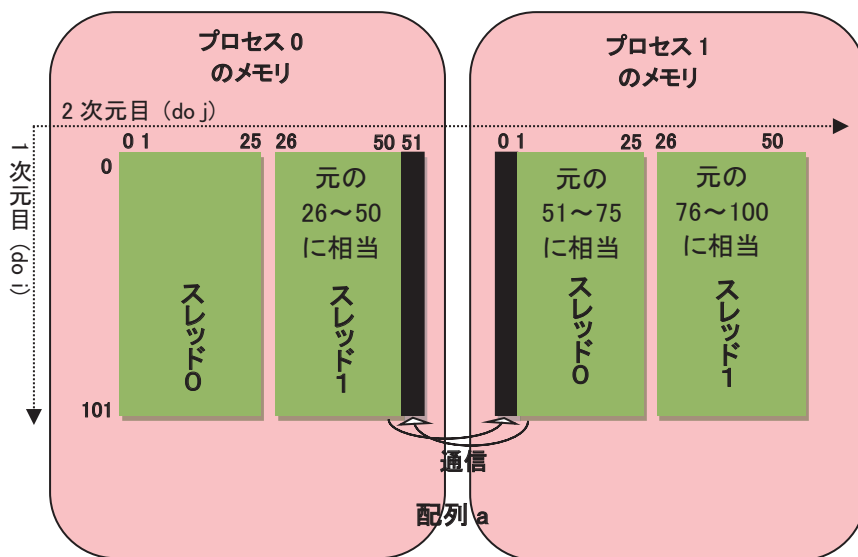


図 3.12 ハイブリッド並列化時の配列 a の割付け(黒色は袖領域)

ひとつの仕事を複数の小さな仕事に分割して、複数のプロセスに割り当てる場合(フラット並列化)と、各プロセス内の複数のスレッドに割り当てる場合(ハイブリッド並列化)とを比較すると、ハイブリッド並列化によりプロセスの総数が削減されます。そのため、プロセス数に依存してコストが増大する通信を含むようなプログラムは、ハイブリッド並列化により経過時間を短縮できる場合があります。その反面、分散並列化と共有並列化の2重のオーバーヘッドが発生するので、プログラムによってはハイブリッド並列化により経過時間が増加する場合があります。一方、コンパイラ・MPI の処理系内部のバッファや、全てのプロセスに重複配置するデータの個数は、プロセス数の減少とともに減少するので、ハイブリッド並列化の方がフラット並列化よりも使用メモリの総量を少なくできる傾向があります。

### 3.4.1 ハイブリッド並列化時の注意点

共有並列化の場合、既定値では各プロセスが VE カード内の全ての CPU コアを使用して共有並列実行を行います。そのため、ハイブリッド並列化時に、ひとつの VE カード上で2つ以上のプロセスを実行すると、CPU コアの個数がスレッドの個数より少なくなり、大幅に性能が低下する可能性があります。ハイブリッド並列化時には、各 VE カード上のプロセスの個数を1にするか、あるいは共有並列化のスレッドの個数を指定する環境変数 `VE_OMP_NUM_THREADS` を指定して、各 VE カード上のプロセスの個数とスレッドの個数の積が各 VE カードのコア数(VE30A の場合は 16) 以下となるようにしてください。

NEC MPI のスレッドサポートレベルは、`MPI_THREAD_SERIALIZED` です。従って、全てのスレッドが MPI 手順を呼び出せますが、OpenMP で共有並列化する場合、同一プロセス中の複数のスレッドが同時に MPI 手順を呼び出さないように利用者がプログラムする必要があります。

## 3.5 性能解析

### 3.5.1 MPI プログラム実行性能情報

プログラム実行時に環境変数 `NMPL_PROGINF` の値を、`YES`、`ALL`、`DETAIL`、または `ALL_DETAIL` に設定することによって MPI プログラムの実行性能情報を採取できます。図 3.13 に、環境変数 `NMPL_PROGINF` の値を `DETAIL` に設定して実行した場合の出力例を示します。User Time が最小のプロセス(Min)、最大のプロセス(Max)、および平均(Average)の情報が表示されており、プログラム全体の性能を簡便に把握できます。図中の「LD L3 Hit Element Ratio (%)」は VE30A から追加された情報であり、レベル 3 キャッシュの利用率情報を出力します。

MPI Program Information:			
=====			
Note: It is measured from MPI_Init till MPI_Finalize.			
[U,R] specifies the Universe and the Process Rank in the Universe.			
Times are given in seconds.			
Global Data of 4 Vector processes	:	Min [U,R]	Max [U,R] Average
=====			
Real Time (sec)	:	64.877 [0,0]	64.878 [0,1] 64.878
User Time (sec)	:	60.075 [0,0]	64.875 [0,1] 63.675
Vector Time (sec)	:	54.302 [0,0]	57.173 [0,3] 56.427
Inst. Count	:	59322128013 [0,0]	64282152733 [0,2] 63012467399
V. Inst. Count	:	17494773621 [0,0]	17632751395 [0,2] 17593712254
V. Element Count	:	4423763325401 [0,0]	4424371657756 [0,3] 4424173912218
V. Load Element Count	:	1302728711714 [0,0]	1302966137322 [0,3] 1302891734718
FLOP Count	:	1775592960917 [0,1]	1775592961762 [0,0] 1775592961128
MOPS	:	77505.531 [0,1]	83610.524 [0,0] 79032.784
MOPS (Real)	:	77421.537 [0,0]	77504.364 [0,2] 77483.037
MFLOPS	:	27369.443 [0,1]	29556.218 [0,0] 27916.186
MFLOPS (Real)	:	27368.310 [0,1]	27368.418 [0,0] 27368.353
A. V. Length	:	250.915 [0,2]	252.862 [0,0] 251.466
V. Op. Ratio (%)	:	99.072 [0,2]	99.167 [0,0] 99.097
L1 Cache Miss (sec)	:	3.268 [0,1]	3.344 [0,2] 3.292
CPU Port Conf. (sec)	:	0.000 [0,0]	0.000 [0,0] 0.000
V. Arith. Exec. (sec)	:	37.520 [0,0]	37.629 [0,1] 37.595
V. Load Exec. (sec)	:	14.320 [0,0]	16.283 [0,3] 15.708
LD L3 Hit Element Ratio (%)	:	41.257 [0,0]	41.714 [0,3] 41.503
VLD LLC Hit Element Ratio (%)	:	43.127 [0,0]	43.912 [0,3] 43.419
FMA Element Count	:	557320568364 [0,0]	557320568364 [0,0] 557320568364
Power Throttling (sec)	:	0.000 [0,0]	0.000 [0,0] 0.000
Thermal Throttling (sec)	:	0.000 [0,0]	0.000 [0,0] 0.000
Memory Size Used (MB)	:	607.666 [0,0]	609.000 [0,1] 608.000
Non Swappable Memory Size Used (MB)	:	95.667 [0,0]	161.000 [0,1] 112.000
Overall Data of 4 Vector processes			
=====			
Real Time (sec)	:	64.878	
User Time (sec)	:	254.700	
Vector Time (sec)	:	225.707	
GOPS	:	78.947	
GOPS (Real)	:	309.932	
GFLOPS	:	27.885	
GFLOPS (Real)	:	109.473	
Memory Size Used (GB)	:	2.375	
Non Swappable Memory Size Used (GB)	:	0.438	
VE Card Data of 1 VEs			
=====			
Memory Size Used (MB) Min	:	2431.998 [node=vh0, ve=0]	
Memory Size Used (MB) Max	:	2431.998 [node=vh0, ve=0]	
Memory Size Used (MB) Avg	:	2431.998	
Non Swappable Memory Size Used (MB) Min	:	448.000 [node=vh0, ve=0]	
Non Swappable Memory Size Used (MB) Max	:	448.000 [node=vh0, ve=0]	
Non Swappable Memory Size Used (MB) Avg	:	448.000	

図 3.13 MPI プログラムの実行時性能情報

### 3.5.2 性能解析機能 (FTRACE 機能・NEC Ftrace Viewer)

コンパイラオプション `-ftrace` を指定して翻訳した MPI プログラムを実行すると、解析情報ファイル `ftrace.out.u.r` (`u` は通常 `0`, `r` は環境変数 `MPIRANK` の値) が各プロセスから出力されます。MPI プログラムの場合、コンパイラによる性能解析情報に加えて、通信時間などの情報が得られます。

SX-Aurora TSUBASA が用意する NEC Ftrace Viewer は、性能解析情報をグラフィカルに表示でき、並列プログラムの性能を分析する際にも有効です。例として、8 プロセスにより並列実行したある MPI プログラムの負荷バランスを解析してみます。まず、解析情報ファイルを読み込んだときに表示される Process Breakdown Chart を見ると、図 3.14 のように手続 `ITERATE` の実行時間 (EXCLUSIVE TIME) が大きいことが分かります。手続 `ITERATE` の負荷バランスを把握するため、左上の「Chart」メニューから Function Statistics Chart を選択します。すると、図 3.15 中の折れ線グラフが示すように、手続 `ITERATE` の EXCLUSIVE TIME の標準偏差が大きく、負荷バランスが悪いことが分かります。状況をより詳細に調査するため、「Chart」メニューから Process Metrics Chart を選択します。次に、「Metrics Selection」メニューから、MPI COMM. TIME (MPI 通信時間) および MPI IDLE TIME (MPI 待ち時間) を選択します。すると、図 3.16 のように、MPI 通信時間および MPI 待ち時間を示す 2 つの折れ線グラフがほぼ重なって見えます。これは、MPI 通信時間のほとんどが MPI 待ち時間であることを意味しています。また、棒グラフで示される手続 `ITERATE` の EXCLUSIVE TIME と、折れ線グラフの MPI 待ち時間とを比較すると、EXCLUSIVE TIME に占める MPI 待ち時間が、両端のプロセスだけ短いことが分かります。従って、両端のプロセスの MPI 待ち時間以外の実行時間、つまり演算時間が他のプロセスよりも大きく、相対的に演算時間の小さな他のプロセスにおいて、MPI による通信時に待ちが発生していると推定できます。以上のことから、両端のプロセスの演算を他のプロセスに分配すれば、負荷バランスを改善できる可能性があることが分かります。このように、NEC Ftrace Viewer のグラフ機能を使用すると、負荷バランスなど複数の仕事間の関係を容易に把握できます。

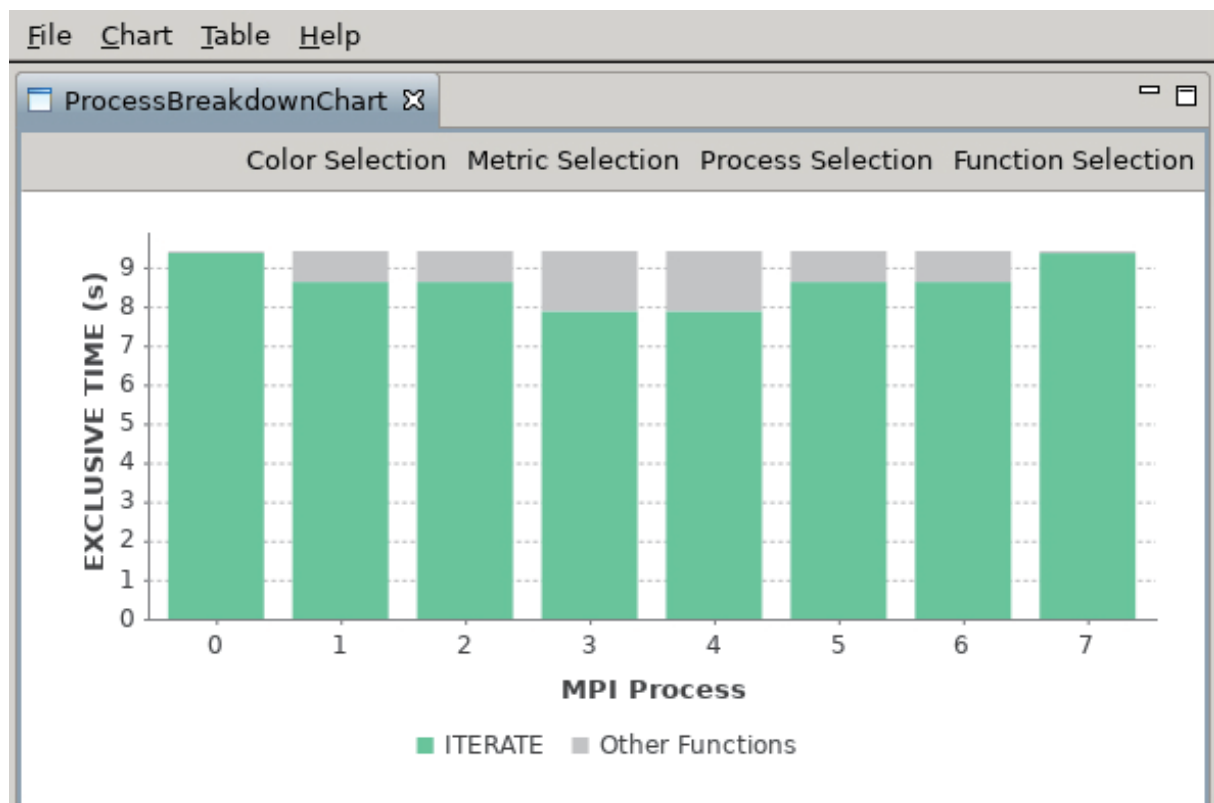


図 3.14 Process Breakdown Chart

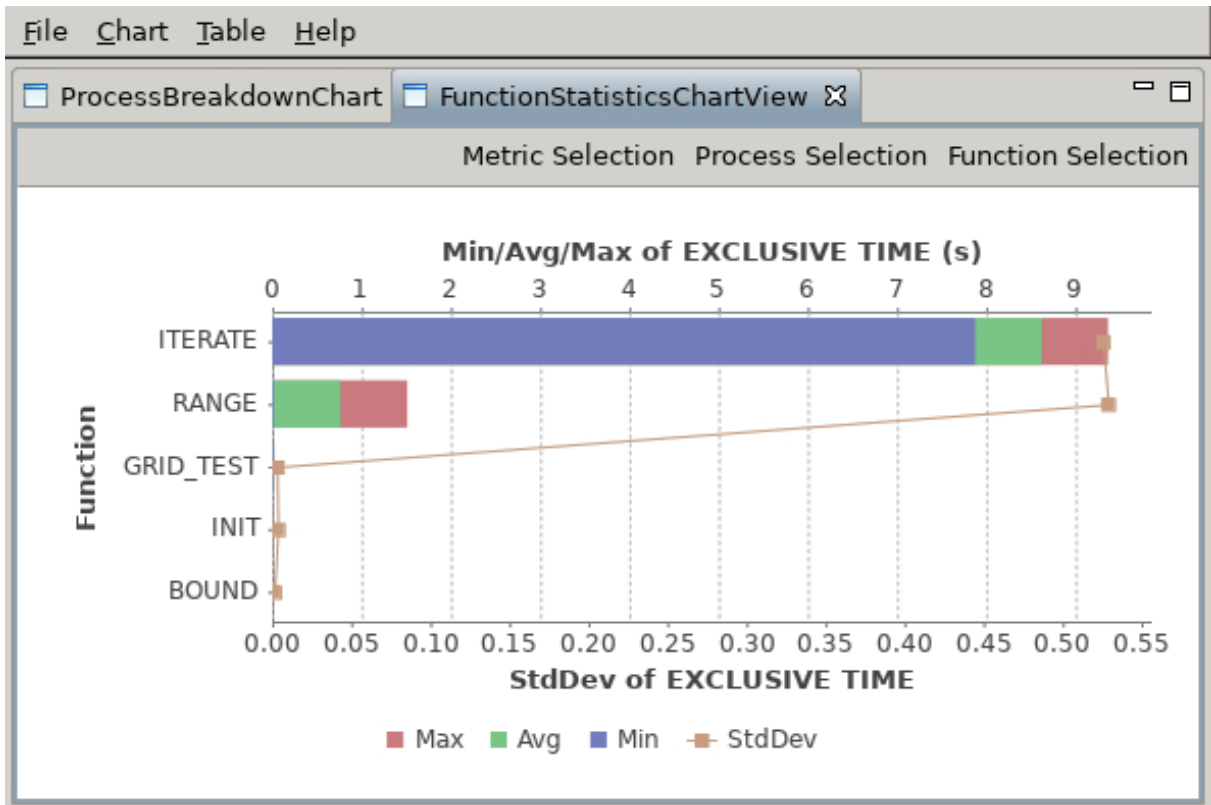


Figure 3.15 Function Statistics Chart

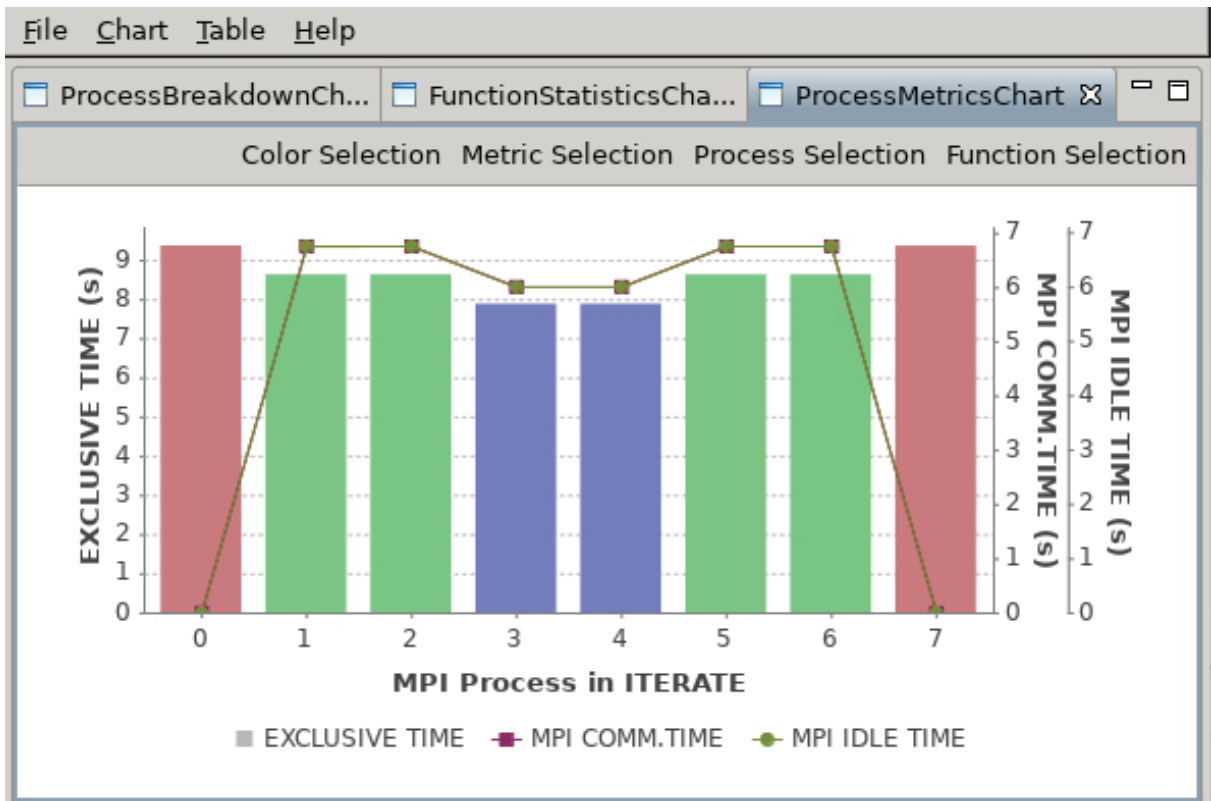


Figure 3.16 Process Metrics Chart

## 4. あとがき

本稿では、コンパイラの共有並列化機能 および NEC MPI を使用した分散並列化の基本的事項を中心にご紹介しました。SX-Aurora TSUBASA が用意する並列処理環境の詳細な使用方法については、「Fortran コンパイラ ユーザーズガイド」<sup>[3]</sup>、「C/C++コンパイラ ユーザーズガイド」<sup>[4]</sup>、および「NEC MPI ユーザーズガイド」<sup>[5]</sup>をご覧ください。また、プログラム実行性能情報(PROGINF)および性能解析機能(FTRACE)の詳細な使用方法については、「PROGINF/FTRACE ユーザーズガイド」<sup>[6]</sup>および「NEC Ftrace Viewer ユーザーズガイド」<sup>[7]</sup>をご覧ください。SX-Aurora TSUBASA をご使用になる上で、本稿がお役に立てば幸いです。

## 参考文献

- [1] The OpenMP Architecture Review Board <http://openmp.org/>
- [2] Message Passing Interface Forum <http://www.mpi-forum.org/>
- [3] Fortran コンパイラ ユーザーズガイド、日本電気株式会社  
[https://sxaoratsubasa.sakura.ne.jp/Documentation\(Japanese\)](https://sxaoratsubasa.sakura.ne.jp/Documentation(Japanese))
- [4] C/C++コンパイラ ユーザーズガイド、日本電気株式会社  
[https://sxaoratsubasa.sakura.ne.jp/Documentation\(Japanese\)](https://sxaoratsubasa.sakura.ne.jp/Documentation(Japanese))
- [5] NEC MPI ユーザーズガイド、日本電気株式会社  
[https://sxaoratsubasa.sakura.ne.jp/Documentation\(Japanese\)](https://sxaoratsubasa.sakura.ne.jp/Documentation(Japanese))
- [6] PROGINF/FTRACE ユーザーズガイド、日本電気株式会社  
[https://sxaoratsubasa.sakura.ne.jp/Documentation\(Japanese\)](https://sxaoratsubasa.sakura.ne.jp/Documentation(Japanese))
- [7] NEC Ftrace Viewer ユーザーズガイド、日本電気株式会社  
[https://sxaoratsubasa.sakura.ne.jp/Documentation\(Japanese\)](https://sxaoratsubasa.sakura.ne.jp/Documentation(Japanese))