CMSI計算科学技術 特論B

第3回
アプリケーションの性能最適化2
(CPU単体性能最適化)

2014年4月24日

独立行政法人理化学研究所
計算科学研究機構 運用技術部門
ソフトウェア技術チーム チームヘッド

南 一生
minami_kaz@riken.jp

K computer

RIKEN ADVANCED INSTITUTE FOR COMPUTATIONAL SCIENCE

講義の概要

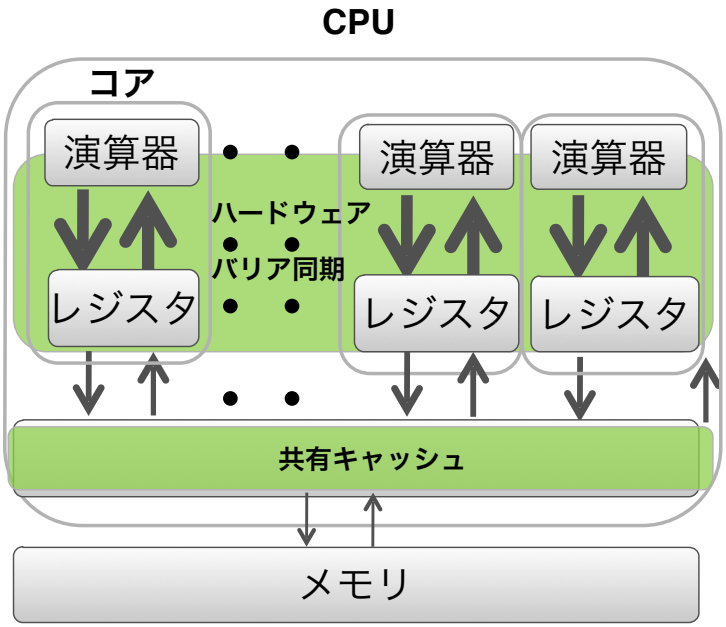
- スーパーコンピュータとアプリケーションの性能
- アプリケーションの性能最適化1 (高並列性能最適化)
- アプリケーションの性能最適化2 (CPU単体性能最適化)
- アプリケーションの性能最適化の実例1
- アプリケーションの性能最適化の実例2

内容

- スレッド並列化
- CPU単体性能を上げるための5つの要素
- 要求B/F値と5つの要素の関係
- 性能予測手法(要求B/F値が高い場合)
- 具体的テクニック

スレッド並列化

スレッド並列の概要

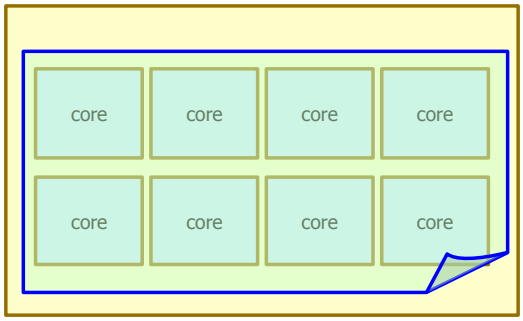


- 京の場合1CPUに8コア搭載.
- 8コアでL2キャッシュを共有.
- MPI等のプロセス並列に対しCPU内の8コアを使用するためのスレッド並列化が必要.
- スレッド並列化のためには自動並列化かOpenMPが使用可.
- 京の場合はハードウェアバリア同期機能を使用した高速なスレッド処理が可能.

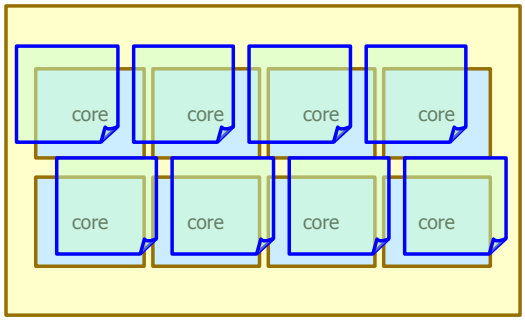
ハイブリッド並列とフラットMPI並列

- MPIプロセス並列とスレッド並列を組み合わせをハイブリッド並列という.
- 各コアにMPIプロセスをわりあてる並列化をフラットMPI並列という.
- 京では通信資源の効率的利用，消費メモリ量を押さえる観点でハイブリッド並列を推奨している.

1プロセス8スレッド
（ハイブリッドMPI）の場合



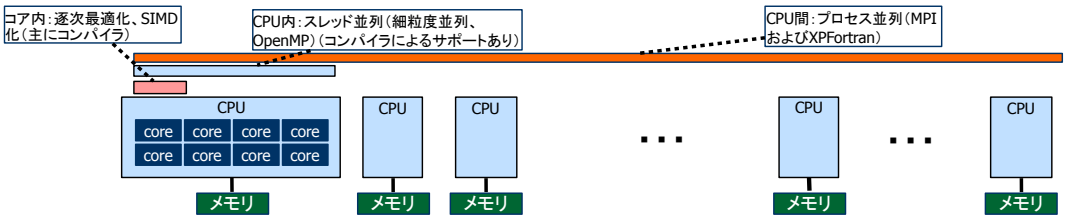
8プロセス（フラットMPI）の場合



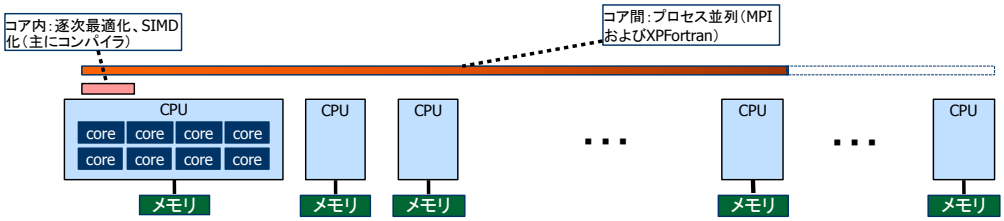
ハイブリッド並列とフラットMPI並列（京の場合）

- スレッド並列＋プロセス並列のハイブリッド型
  - コア内：コンパイラによる逐次最適化，SIMD化
  - CPU内：スレッド並列(自動並列化：細粒度並列化†，OpenMP)
  - CPU間：プロセス並列(MPI、XPFortran)

†細粒度並列化
高速バリア同期を活用し，内側ループをコア間で並列化
ベクトル向け（内側ループが長い）コードの高速化が可能



- プロセス並列型
  - コア内：コンパイラによる逐次最適化，SIMD化
  - コア間：プロセス並列(MPI、XPFortran)



実行効率の観点から、ハイブリッド型を推奨

CPU単体性能を上げるための5つの要素

CPU単体性能を上げるための5つの要素

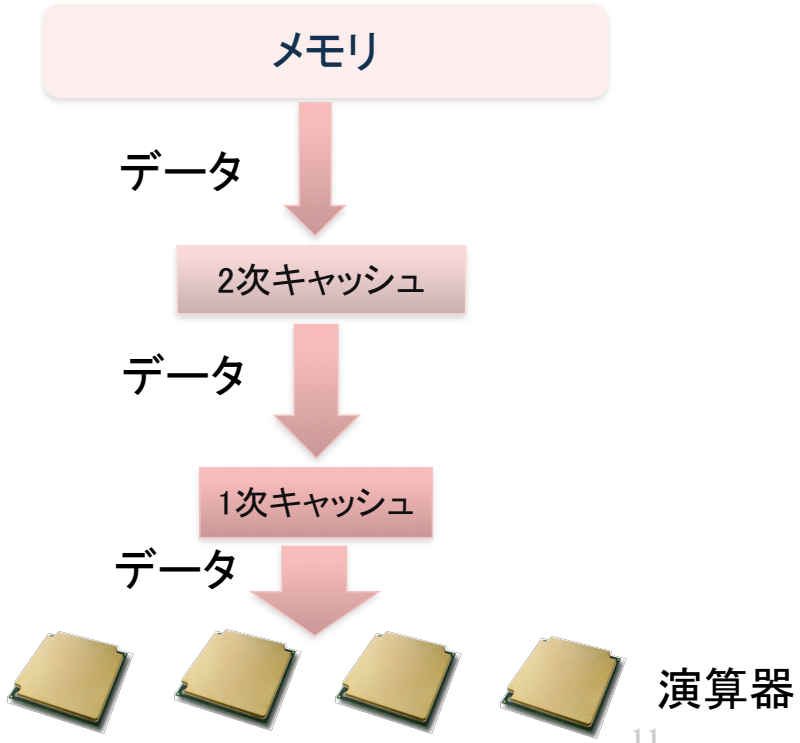
CPU内の複数コアでまずスレッド並列化が
できていると事は前提として

(1) ロード・ストアの効率化
(2) ラインアクセスの有効利用
(3) キャッシュの有効利用
(4) 効率の良い命令スケジューリング
(5) 演算器の有効利用

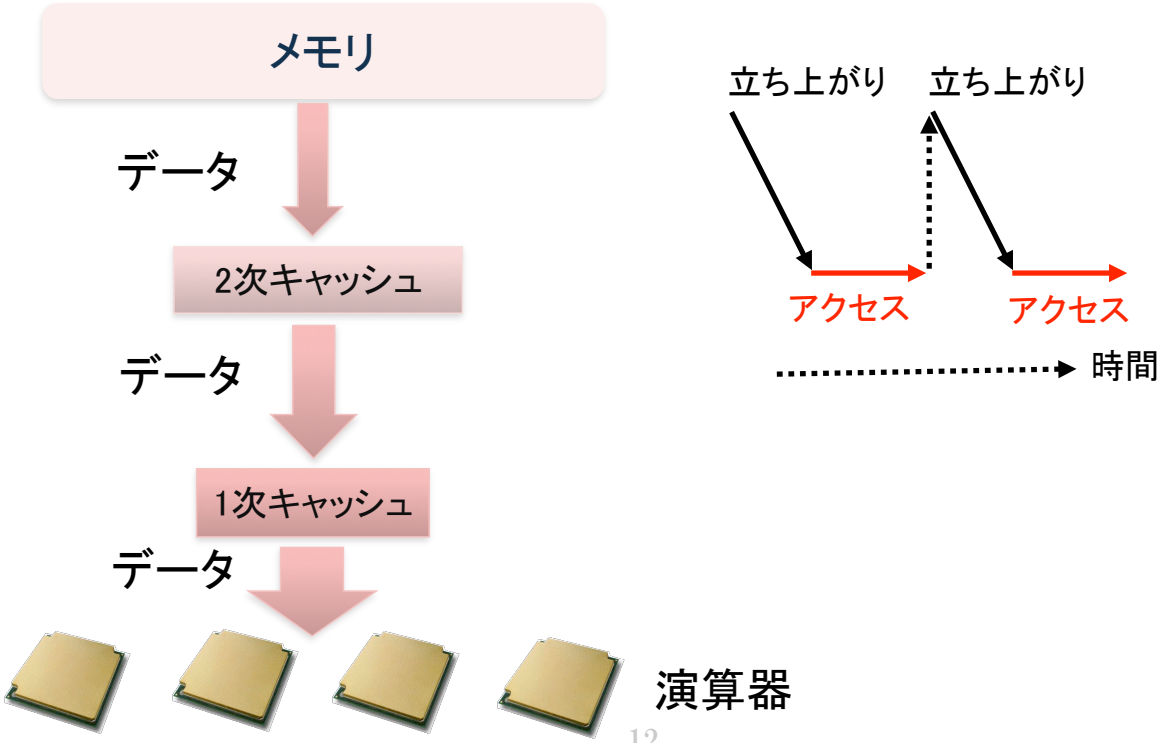
(1) ロード・ストアの効率化

- プリフェッチの有効利用
- 演算とロード・ストア比の改善

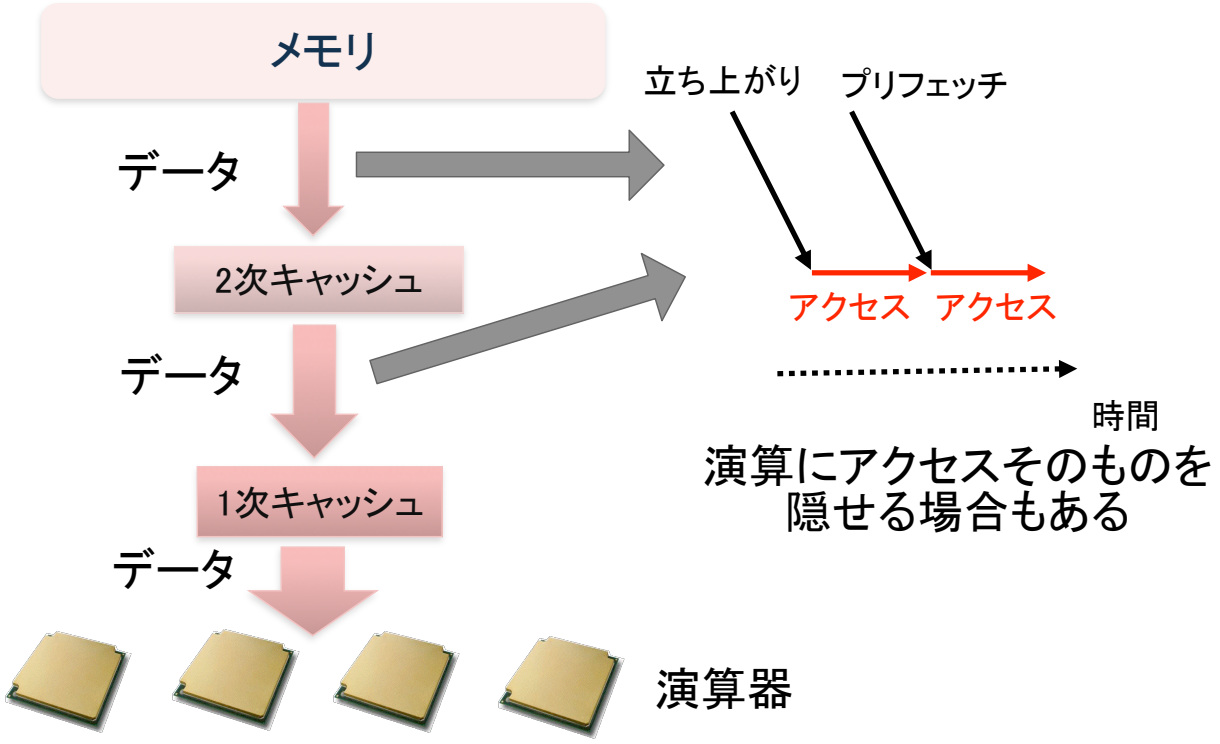
レイテンシ(アクセスの立ち上がり)



レイテンシ(アクセスの立ち上がり)



プリフェッチの有効利用

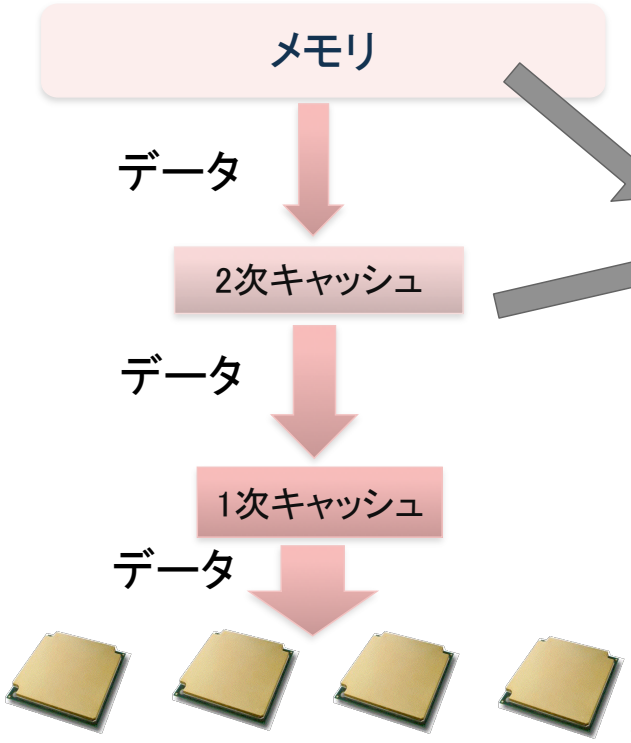


演算とロード・ストア比の改善

- 以下のコーディングを例に考える.
- 以下のコーディングの演算は和2個，積2個の計4個.
- ロードの数はx,a(i),a(i+1)の計3個.
- ストアの数はxの1個.
- したがってロード・ストア数は4個
- 演算とロード・ストアの比は4/4となる.
- なるべく演算の比率を高めロード・ストアの比率を低く抑えて演算とロードストアの比を改善を図る事が重要.

```
do  j=1,m
do  i=1,n
  x(i)=x(i)+a(i)*b+a(i+1)*d
end  do
```

(2) ラインアクセスの有効利用



キャッシュライン
128byte毎のアクセス

全部の要素を
計算に利用

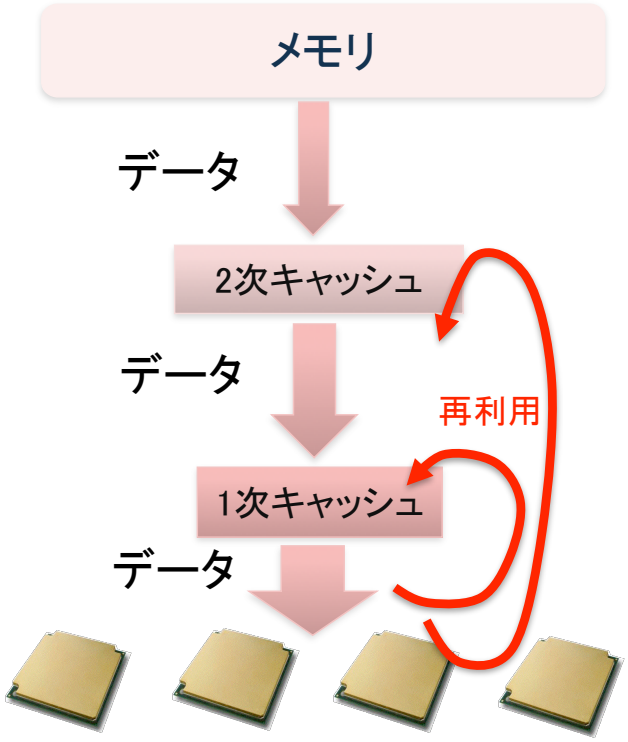
効率が良い

キャッシュライン
128byte毎のアクセス

1個の要素のみ
を計算に利用

効率が悪い

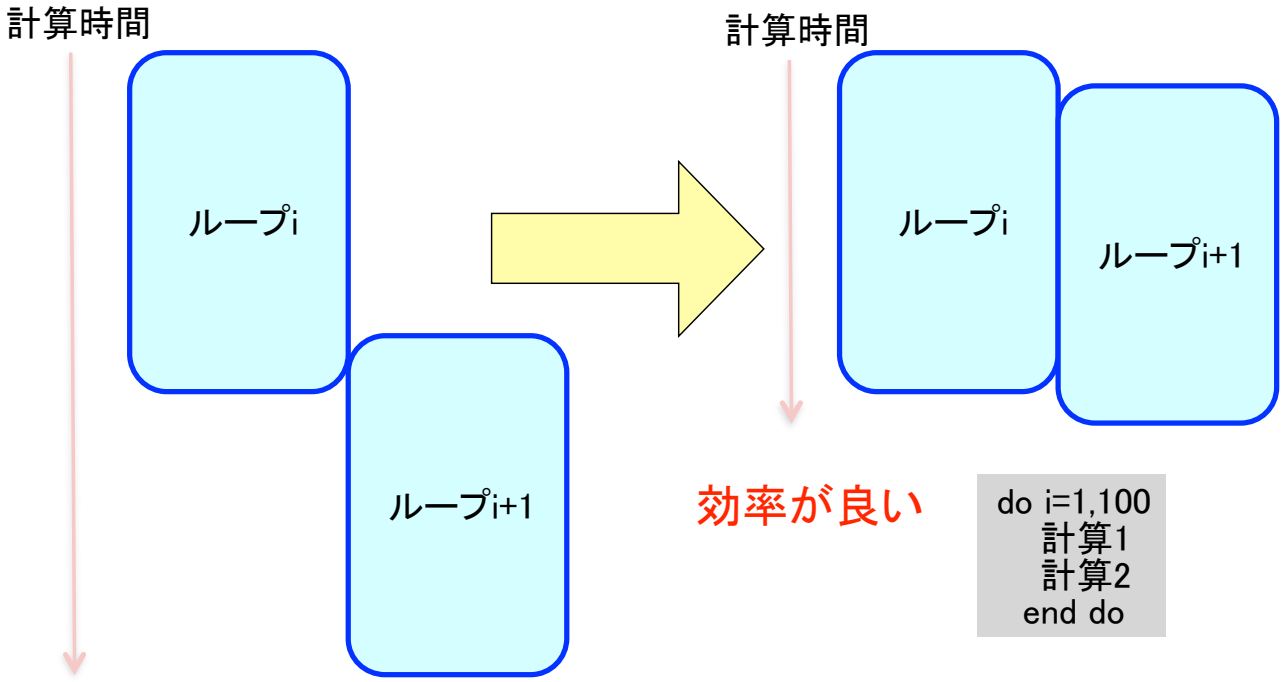
(3) キャッシュの有効利用



・データ供給能力の高いキャッシュを設ける
・キャッシュに置いたデータを何回も再利用し演算を行なう
・こうする事で演算器の能力を十分使い切る

- **ブロッキング**

(4) 効率の良い命令スケジューリング



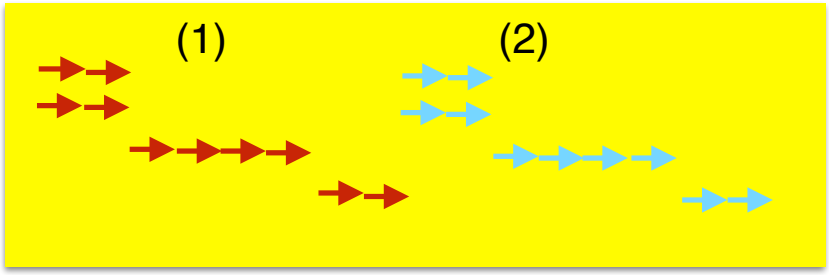
並列処理と依存性の回避

< ソフトウェアパイプラインニング> **コンパイラ**

<前提>
・ロード2つorストアと演算とは同時実行可能
・ロードとストアは2クロックで実行可能
・演算は4クロックで実行可能
・ロードと演算とストアはパイプライン化されている

例えば以下の処理をを考える。

```
do i=1,100
  a(i)のロード
  b(i)のロード
  a(i)とb(i)の演算
  i番目の結果のストア
end do
```



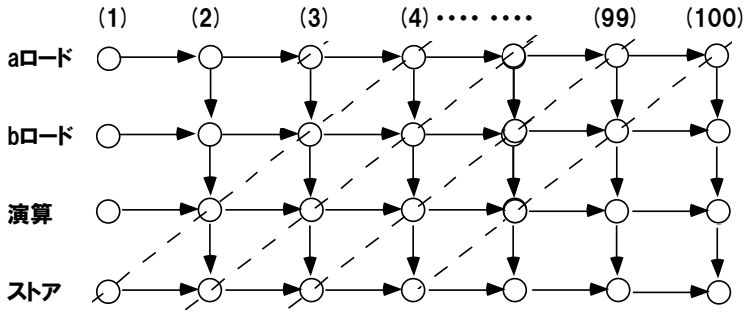
<実行時間>
・8クロック×100要素=800クロックかかる

並列処理と依存性の回避

< ソフトウェアパイプラインニング> **コンパイラ**

左の処理を以下のように構成し直す事を
ソフトウェアパイプラインニングという

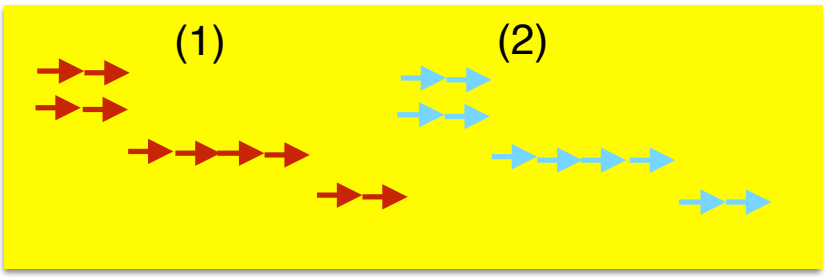
a(1)a(2)a(3)のロード
b(1)b(2)のロード
(1)の演算
do i=3,100
　a(i+1)のロード
　b(i)のロード
　(i-1)の演算
　i-2番目の結果のストア
end do
b(100)のロード
(99)(100)の演算
(98)(99)(100)のストア



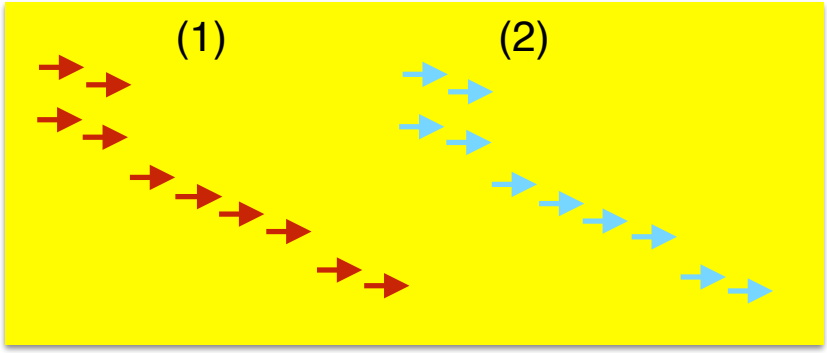
並列処理と依存性の回避

< ソフトウェアパイプラインニング> **コンパイラ**

do i=1,100
　a(i)のロード
　b(i)のロード
　a(i)とb(i)の演算
　i番目の結果のストア
end do



do i=1,100
　a(i)のロード
　b(i)のロード
　a(i)とb(i)の演算
　i番目の結果のストア
end do

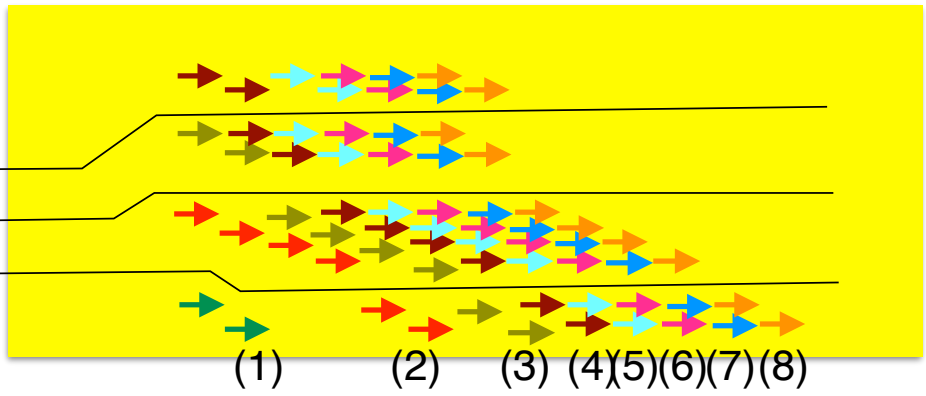


並列処理と依存性の回避

< ソフトウェアパイプラインニング>

左の処理を以下のように構成し直す事を
ソフトウェアパイプラインニングという

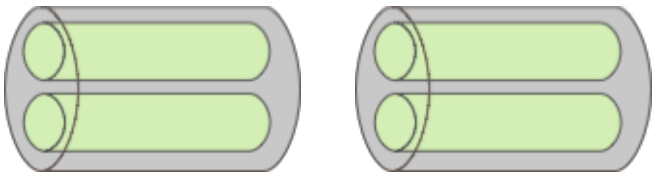
```
a(1)a(2)a(3)のロード
b(1)b(2)のロード
(1)の演算
do i=3,100
    a(i+1)のロード
    b(i)のロード
    (i-1)の演算
    i-2番目の結果のストア
end do
b(100)のロード
(99)(100)の演算
(98)(99)(100)のストア
```



<実行時間>
・前後の処理及びメインループの立ち上がり部分を除くと
・1クロック×100要素=100クロックで処理できる

(5) 演算器の有効利用

2 SIMD Multi&Add演算器×２本



乗算と加算を4個同時に計算可能

(1 + 1) × 4 = 8

この条件に
近い程高効率

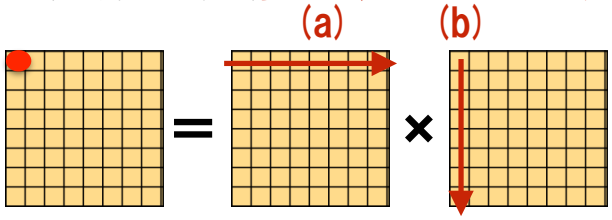
1コアのピーク性能：8演算×2GHz =16G演算/秒

要求B/F値と5つの要素の関係

要求B/F値と5つの要素の関係

アプリケーションの要求B/F値の大小によって性能チューニングにおいて注目すべき項目が異なる

行列行列積の計算（要求B/F値が小さい）

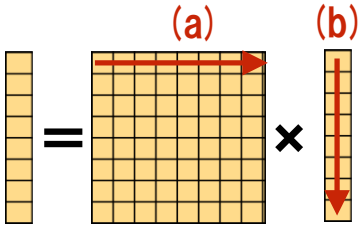


$2N^3$個の演算　N^2個のデータ　N^2個のデータ

B/F値
=移動量(Byte)/演算量(Flop)
$=2N^2/2N^3$
$=1/N$

原理的にはNが大きい程小さな値

行列ベクトル積の計算（要求B/F値が大きい）



$2N^2$個の演算　N^2個のデータ　N個のデータ

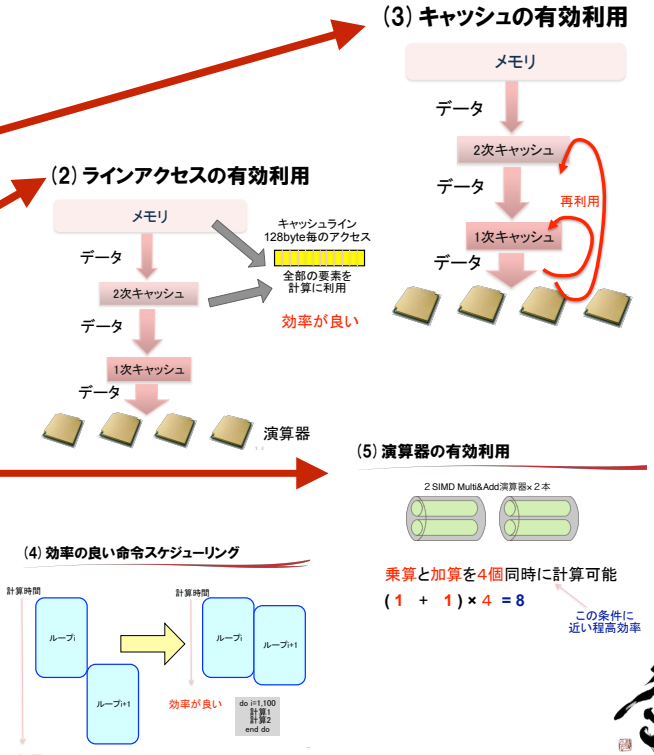
B/F値
=移動量(Byte)/演算量(Flop)
$=(N^2+N)/2N^2$
$\fallingdotseq 1/2$

原理的には1/Nより大きな値

要求B/F値と5つの要素の関係

要求するB/Fが小さいアプリケーションについて

- 原理的にキャッシュの有効利用が可能
- まずデータをオンキャッシュにするコーディング：(3)が重要
- つぎに2次キャッシュのライン上のデータを有効に利用するコーディング：(2)が重要
- それが実現できた上で(4)(5)が重要

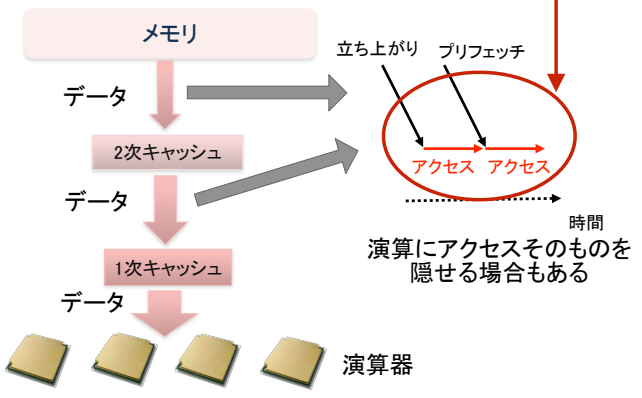


要求B/F値と5つの要素の関係

要求するB/Fが大きいアプリケーションについて

・メモリバンド幅を使い切る事が大事

レイテンシーが隠れた状態にする事．この状態でメモリバンド幅のピーク(京の場合の理論値は64GB/sec)が出せる．レイテンシーが見えているとメモリバンド幅のピーク値は出せない．

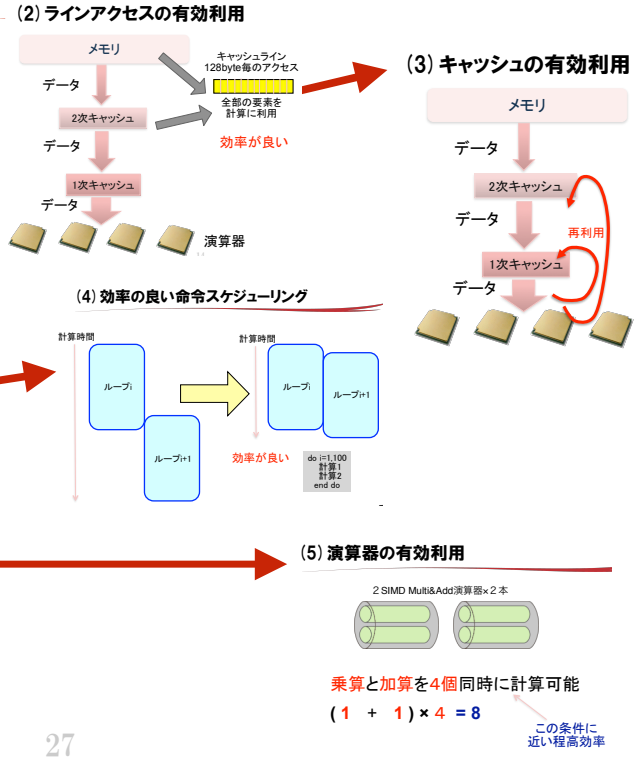


演算にアクセスそのものを隠せる場合もある

要求B/F値と5つの要素の関係

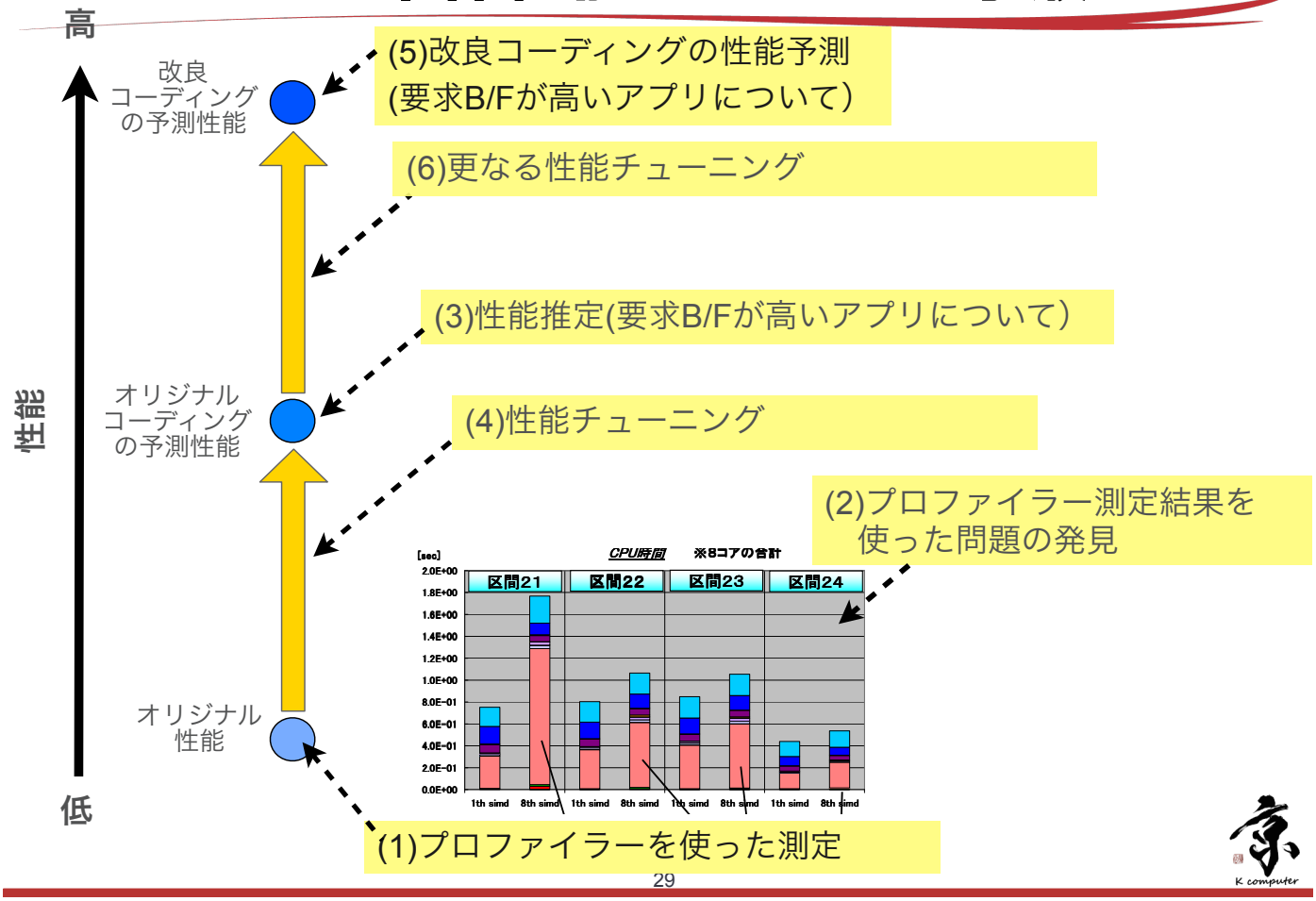
要求するB/Fが大きいアプリケーションについて

- 一番重要なのは(1)(2)
- 次にできるだけオンキャッシュする(3)が重要
- これら(1)(2)(3)が満たされ計算に必要なデータが演算器に供給された状態で、それらのデータを十分使える程度に(4)のスケジューリングができて、さらに(5)の演算器が有効に活用できる状態である事が必要

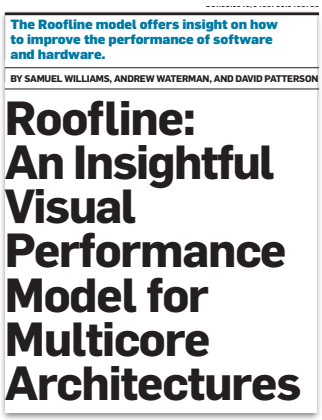


性能予測手法
(要求B/F値が大きい場合)

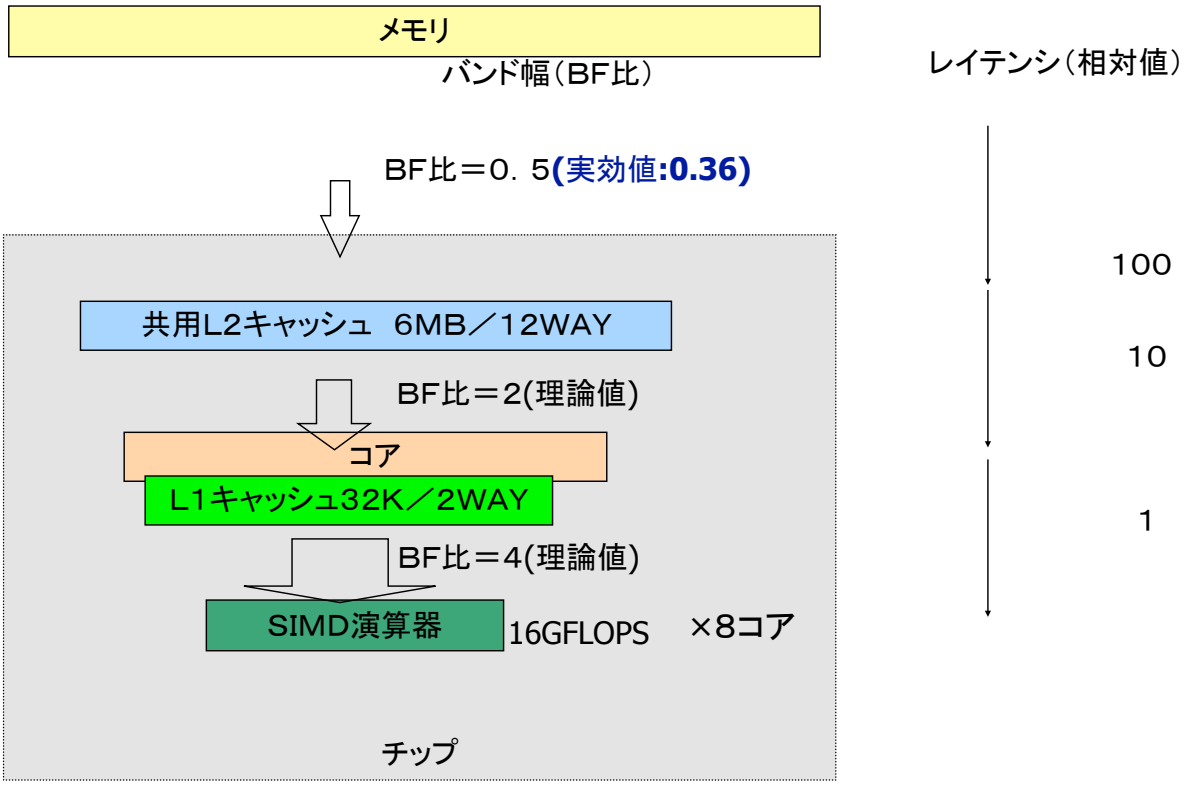
CPU単体性能チューニング手順



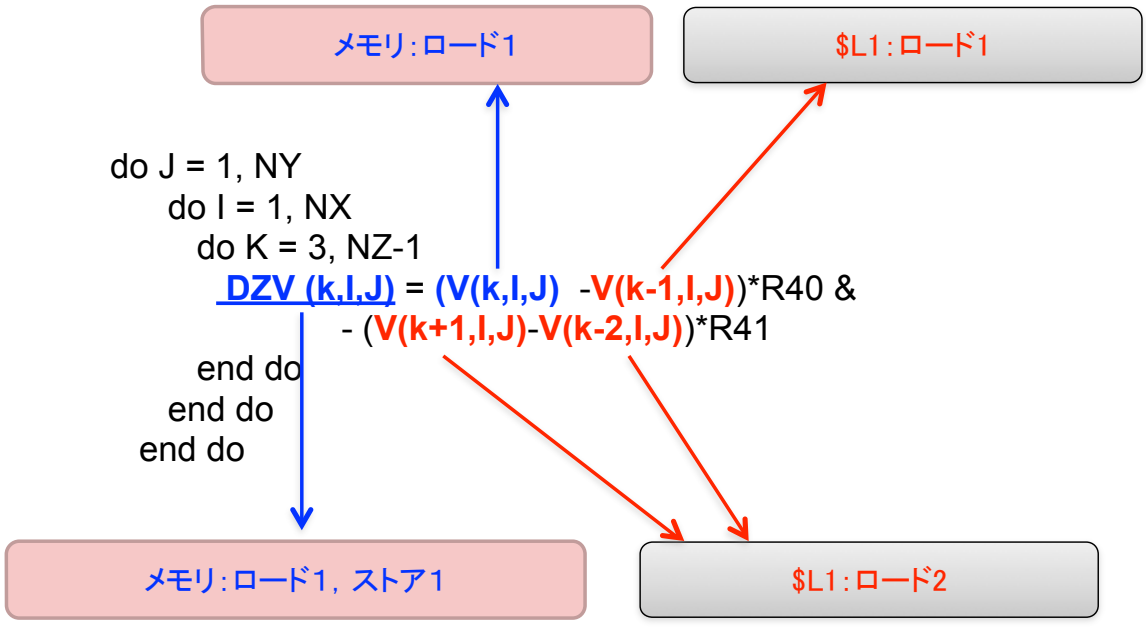
ルーフラインモデル



ベースとなる性能値

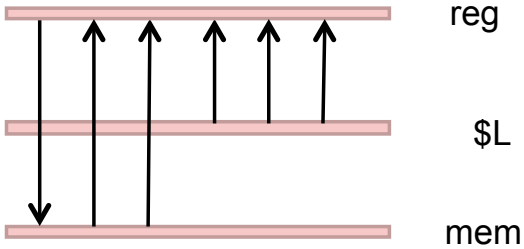


メモリとキャッシュアクセス（1）



メモリとキャッシュアクセス (2)

```
do J = 1, NY
    do I = 1, NX
      do K = 3, NZ-1
        DZV (k,I,J) = (V(k,I,J)  -V(k-1,I,J))*R40 &
                    - (V(k+1,I,J)-V(k-2,I,J))*R41
      end do
    end do
  end do
```



	Store	Load	バンド幅比 ($L1)	データ移動時間の比(L1)	バンド幅比 ($L2)	データ移動時間の比(L2)
$L	1	5	11.1 (8*64G/s)	0.5= 6/11.1	5.6 (256G/s)	1.1=6/5.6
M	1	2	1(46G/s)	3=3/1	1 (46G/s)	3=3/1

データ移動時間の比を見るとメモリで律速される
→ メモリアクセス変数のみで考慮すれば良い。

性能見積り

```
do J = 1, NY
    do I = 1, NX
      do K = 3, NZ-1
        DZV (k,I,J) = (V(k,I,J)  -V(k-1,I,J))*R40 &
                    - (V(k+1,I,J)-V(k-2,I,J))*R41
      end do
    end do
  end do
```

- 最内軸(K軸)が差分
- 1ストリームでその他の3配列は$L1に載っており再利用できる。

要求B/F	12/5 = 2.4
性能予測	0.36/2.4 = 0.15
実測値	0.153

要求Byteの算出:

1store,2loadと考える

4x3 = 12byte

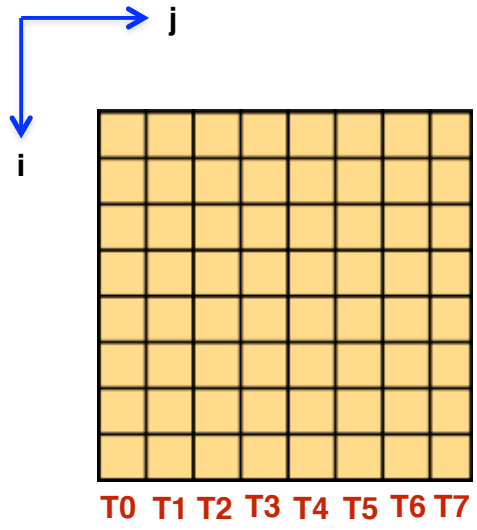
要求flop:

add : 3 mult : 2 = 5

具体的テクニック

（以降のハードウェア・コンパイラについての
話題は京を前提とした話です）

スレッド並列化



Ti：スレッドiの計算担当

jループをブロック分割

```
do j=1,n
  do i=1,n
    x(i,j)=a(i,j)*b(i,j)+c(i,j)
```

CG法前処理のスレッド並列化

- 以下は前処理に不完全コレスキー分解を用いたCG法のアルゴリズムである.
- 前処理には色々な方法を使うことが出来る.
- 例えば前回説明したガウス・ザイデル法等である.
- CG法の本体である行列ベクトル積・内積・ベクトルの和等の処理は簡単に前頁のブロック分割されたスレッド並列化を烏滸なことが出来る.
- しかしガウス・ザイデル前処理は前回講義のようにリカレンスがある.

ステップ1: $\alpha^k = (r_i^k \bullet (LL^T)^{-1} r_i^k) / (Ap_i^k \bullet p_i^k)$

ステップ2: $x_i^{k+1} = x_i^k + \alpha^k p_i^k$

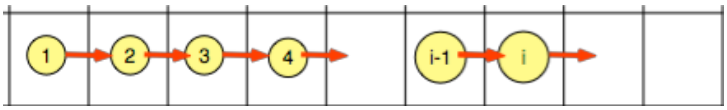
ステップ3: $r_i^{k+1} = r_i^k - \alpha^k Ap_i^k$

ステップ4: $\beta^k = (r_i^{k+1} \bullet (LL^T)^{-1} r_i^{k+1}) / (r_i^k \bullet (LL^T)^{-1} r_i^k)$

ステップ5: $p_i^{k+1} = (LL^T)^{-1} r_i^{k+1} + \beta^k p_i^k$

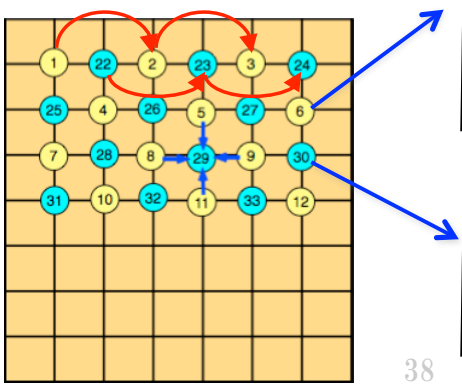
CG法前処理のスレッド並列化

- リカレンスがあると, 前回講義のように依存関係があり並列処理ができない.
- つまりこのままではスレッド並列化もできない.



- そこで例えばRED-BLACKスィープに書換えリカレンスを除去しそれぞれのループをブロック分割によるスレッド並列化を行う.

RED-BLACKスイープ



黄色ループをブロック分割

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

青色ループをブロック分割

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

ロード・ストアの効率化

演算とロード・ストア比の改善 < 内側ループアンローリング>

- 以下の様な2 つのコーディングを比較する。

```
do j=1,m
do i=1,n
  x(i)=x(i)+a(i)*b+a(i+1)*d
end do

do j=1,m do i=1,n,2
  x(i)=x(i)+a(i)*b+a(i+1)*d
  x(i+1)=x(i+1)+a(i+1)*b+a(i+2)*d
 end do
```

- 最初のコーディングの演算量は4 ,ロード/ストア回数は4 である.2 つ目のコ ーディン グの演算量は8 ,ロード/ストア回数は7 である.
- 最初のコーディングの演算とロード/ ストアの比は4/4,2つ目のコーディングの演算とロード/ストアの比は8/7となり良くなる.

ロード・ストアの効率化

演算とロード・ストア比の改善

<外側ループストリップ・マイニング>

・i j 型のコーディングを以下のようなものとする。

```
do  i=1,n
  do  j=1,m
    y(i)=y(i)+a(i,j)*x(j)
```

・この場合a(i,j)、x(j)の２個をロードして２個の演算を実施する。y(i)はレジスタ上に保持しておけばよい。

・したがって演算とロード/ストアの比は1/1である。

・j i 型のコーディングを以下のようなものとする。

```
do  j=1,n
  do  i=1,m
    y(i)=y(i)+a(i,j)*x(j)
```

・この場合a(i,j)、y(j)の２個をロードし、さらにy(j)をストアし２個の演算を実施する。

・したがって演算とロード/ストアの比は2/3である。

ロード・ストアの効率化

演算とロード・ストア比の改善

・以下のようなコーディングを外側ループストリップ・マイニングという。

```
do is=1,m,10
 do j=1,n
  do i=is,min(is+9,m)
    y(i)=y(i)+a(i,j)*x(j)
```

・このコーディングの最内ループをアンローリングする。

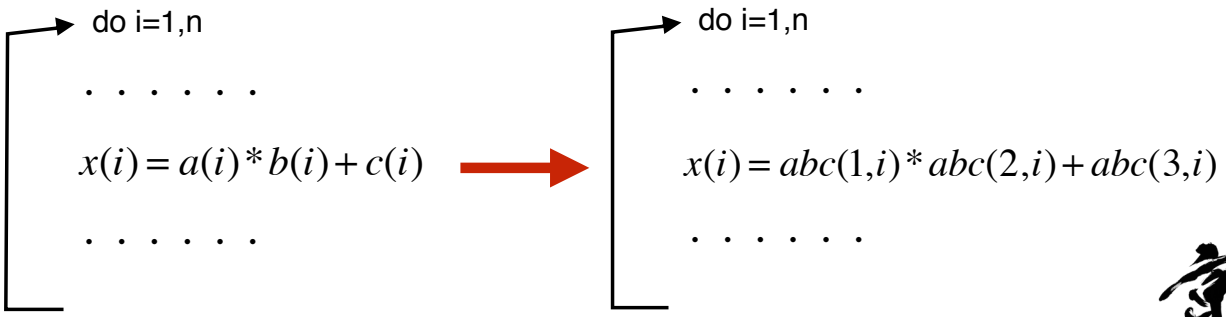
```
do is=1,m,10
 do j=1,n
    y(is)=y(is)+a(is,j)*x(j)
    y(is+1)=y(is+1)+a(is+1,j)*x(j)
      ・・・・          ・・・・
    y(is+9)=y(is+9)+a(is+9,j)*x(j)
```

・この場合a(is,j) ・・a(is+9,j)の10個とx(j)の1個をロードするのみですむ。y(js) ・・ y(is+9)はレジスタ上に保持しておけばよい。この間20個の演算を実行する。
・したがって演算とロード/ストアの比は20/11である。

ロード・ストアの効率化

プリフェッチの有効利用

- プリフェッチにはハードウェアプリフェッチとソフトウェアプリフェッチがある.
- ハードウェアプリフェッチは連続なキャッシュラインのアクセスとなる配列について自動的に機能する.
- ソフトウェアプリフェッチはコンパイラが必要であれば自動的に命令を挿入する.
- ユーザーがコンパイラオプションやディレクティブで挿入する事もできる.
- ハードウェアプリフェッチはループ内の配列アクセス数が一定の数を超えると効かなくなる.
- その場合はアクセスする配列を統合する配列マージで良い効果が得られる場合がある.



ロード・ストアの効率化

プリフェッチの有効利用

- ユーザーがコンパイラオプションやディレクティブで挿入する事もできる.
- 以下ディレクティブによるソフトウェアプリフェッチ生成の例.

改善後ソース（最適化制御行チューニング）

```
51              !ocl prefetch
                <<< Loop-information Start >>>
                <<< [PARALLELIZATION]
                <<<   Standard iteration count: 616
                <<< [OPTIMIZATION]
                <<<   SIMD
                <<<   SOFTWARE PIPELINING
                <<<   PREFETCH      : 32
                <<<     c: 16, b: 16
                <<< Loop-information
52    1  pp  2v          do i = 1 , n
53    1   p  2v            a(i) = b(d(i)) + scalar * c(e(i))
54    1   p  2v          enddo
```

インダイレクトアクセス（配列 b, c）に対するプリフェッチが生成された

RIKEN AICSチューニングチュートリアルより

ラインアクセスの有効利用

- リストアクセスでx,y,z,u,v,wの座標をアクセスするような場合配列マージによりキャッシュラインのアクセス効率を高められる場合がある.

改善前ソース

```
parameter(n=1000000)
real*8  a(n), b(n), c(n)
integer d(n+10)
  :
do iter = 1, 100
 do i = 1 , n
  a(d(i)) = b(d(i)) + scalar * c(d(i))
 enddo
enddo
  :
```

配列マージ

改善後（コンパイラ最適化イメージ）

```
parameter(n=1000000)
real*8  abc(3, n)
integer d(n+10)
  :
do iter = 1, 100
 do i = 1 , n
  ii = d(i)
  abc(1,ii) = abc(2,ii) + scalar * abc(3,ii)
 enddo
enddo
  :
```

（L1Dキャッシュ）

a(d(i))
・・・
a(d(i+1))
・・・
b(d(i))
・・・
b(d(i+1))
・・・
c(d(i))
・・・
c(d(i+1))
・・・

異なるキャッシュラインのアクセス

（L1Dキャッシュ）

abc(1, d(i))
abc(2, d(i))
abc(3, d(i))
・・・
abc(1, d(i+1))
abc(2, d(i+1))
abc(3, d(i+1))
・・・

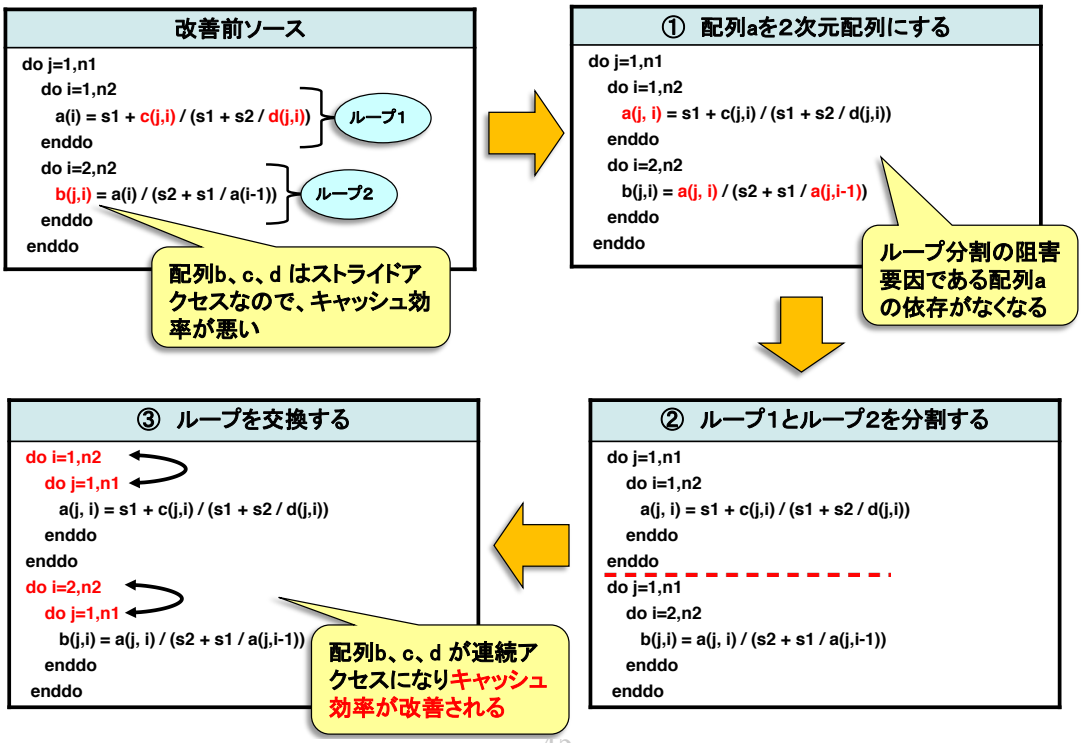
同じキャッシュラインのアクセス

→ キャッシュへのアクセス順番

RIKEN AICSチューニングチュートリアルより

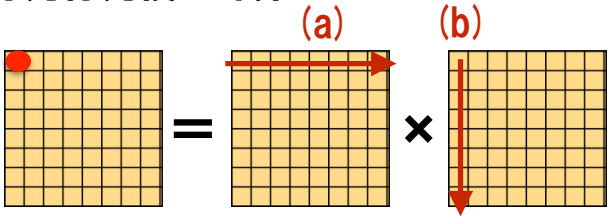
ラインアクセスの有効利用

- ストライドアクセス配列を連続アクセス化することによりラインアクセスの有効利用を図る.



キャッシュの有効利用

行列行列積の計算

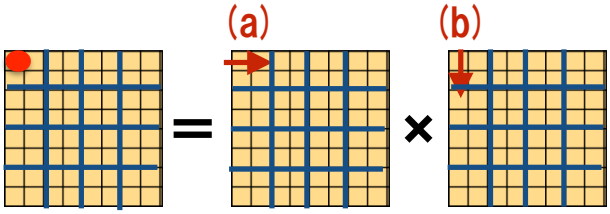


ブロッキング

B/F値
=移動量(Byte)/演算量(Flop)
$=2N^2/2N^3$
$=1/N$

原理的にはNが大きい程小さな値

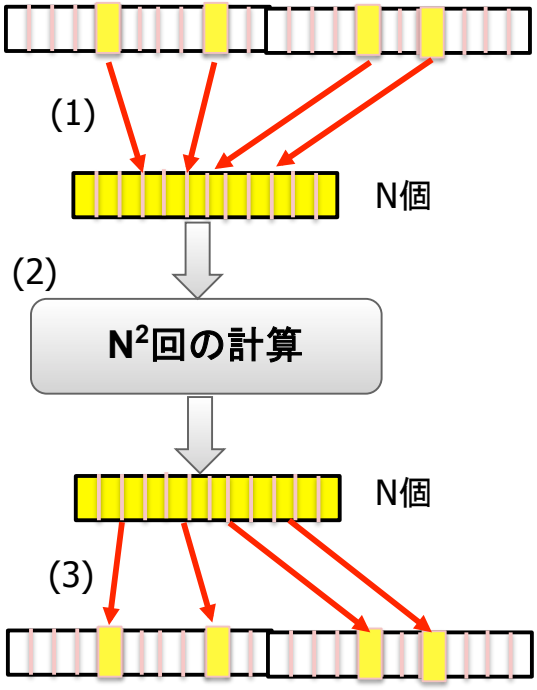
- 現実のアプリケーションではNがある程度の大きさになるとメモリ配置的には (a) はキャッシュに乗っても (b) は乗らない事となる



- そこで行列を小行列にブロック分割し (a) も (b) もキャッシュに乗るようにしてキャッシュ上のデータだけで計算するようにする事で性能向上を実現する.

キャッシュの有効利用

ブロッキング

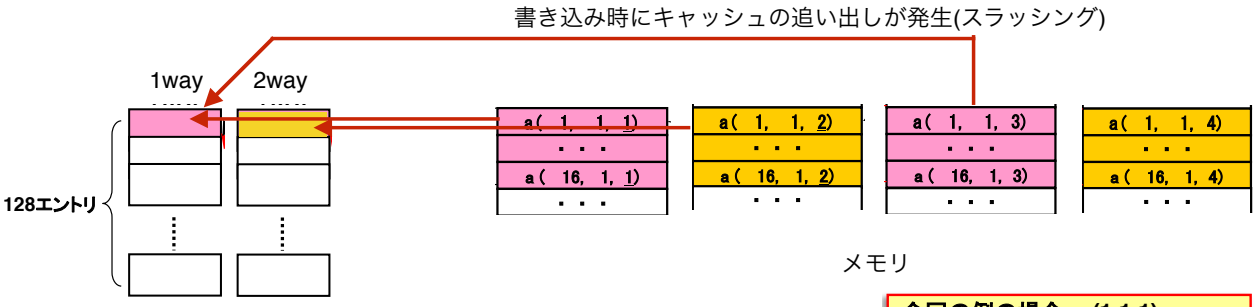


- 不連続データの並び替えによるブロッキング
- (1) N個の不連続データをブロッキングしながら連続領域にコピー
- (2) N個のデータを使用してN²回の計算を実施
- (3) N個の計算結果を不連続領域にコピー

- 一般的に (1) (3) のコピーはNのオーダーの処理であるためN²オーダーの計算時間に比べ処理時間は小さい.

キャッシュの有効利用

スラッシング



L1Dキャッシュ

128バイト×128エントリ×2way=32Kバイト

今回の例の場合、a(1,1,1)、a(1,1,2)、a(1,1,3)、a(1,1,4)はそれぞれ32×16KBずつ離れている(16KB境界にある)ため4つが同じインデックスに割り当てられる。そのため1つ目、2つ目のデータが3つ目、4つ目のデータに上書きされる。

ソース例

```
subroutine sub(a, n, m) ※n=256, m=256
real*8  a(n, m, 4)
  do j = 1 , m
    do i = 1 , n
      a(i, j, 4) = a(i, j, 1) + a(i, j, 2) + a(i, j, 3)
    enddo
  enddo
End
```