12-NA02

片桐孝洋（東京大学情報基盤センター）

高精度行列-行列積アルゴリズムにおける並列化手法の開発

1. 概要

行列-行列積に代表される基本線形計算を集約したライブラリBLAS(Basic Linear Algebra Subprograms)は、多くの線形計算で必須の処理である。

従来の数値計算ライブラリは、演算速度は考慮しているが演算精度の考慮が不十分である。解の精度保証が重要な課題である。BLASを用いた汎用的な数値計算ライブラリ（例：LAPACK）において精度保証をする研究は多くない。

BLASを精度保証する研究が、早稲田大学の大石教授のグループにより進められている。本研究は大石グループで開発された高精度行列-行列演算（尾崎の方法[1]）を基本とし、その並列化を中心する以下の研究開発を行う。

i. 「疎行列-密行列積」、もしくは、「疎行列-疎行列」の実装方式の開発
ii. スレッド並列化手法の開発
iii. 分散メモリ型計算機による並列化手法の開発

2. 共同研究体制

- 片桐孝洋（東大・情報基盤センター）
- 尾崎克久（芝浦工業大・システム理工学部）
- 荻田武史（東京女子大・現代教養学部）
- 大石進一（早大・理工学術院・応用数理学科）

3. 尾崎の方法の実装に対する自動チューニング（AT）記述言語ABCLibScriptの適用

- 尾崎の方法のメインカーネルの＜実コード＞に、AT記述言語ABCLibScriptを適用
- インストール時AT
- GOTO BLASの静的ライブラリの関数コールについて、以下の実装を自動選択する機能

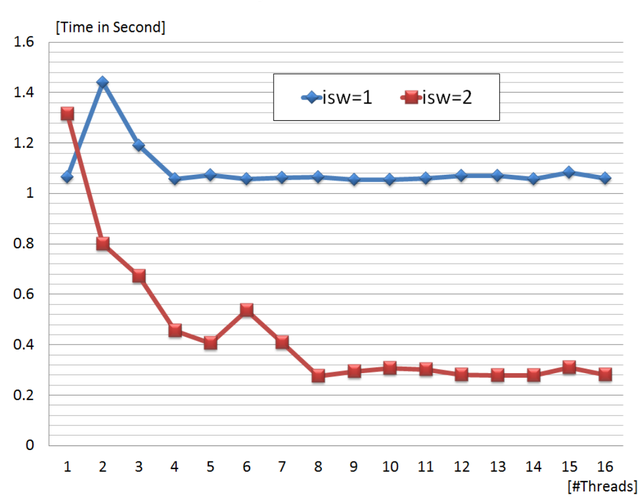
（1）スレッド並列化した関数 dgemm_(…)をコールする実装
（2）逐次BLAS関数 dgemm_seq (…)を問題サイズの並列性を利用してコールする実装

```
#pragma ABCLib install select region start
#pragma ABCLib name SelectOzaki
#pragma ABCLib select sub region start
for (i=0; i<Ak; i++) {
   for (j=0; j<Bk; j++) {
     dgemm_(chn, chn, &m, &p, &n, &one,
       A+i*m*n, &lda, B+j*n*p, &ldb, &zero,
       C+m*p*(j+Bk*i), &ldc);
   } }
#pragma ABCLib select sub region end
#pragma ABCLib select sub region start
#pragma omp parallel for private( i, j)
  for (k=0; k<Ak*Bk; k++) {
    i = k / Bk; j = k % Bk;
    dgemm_seq(chn, chn, &m, &p, &n, &one,
      A+i*m*n, &lda, B+j*n*p, &ldb, &zero,
      C+m*p*(k), &ldc);  }
#pragma ABCLib select sub region end
#pragma ABCLib install select region end
```

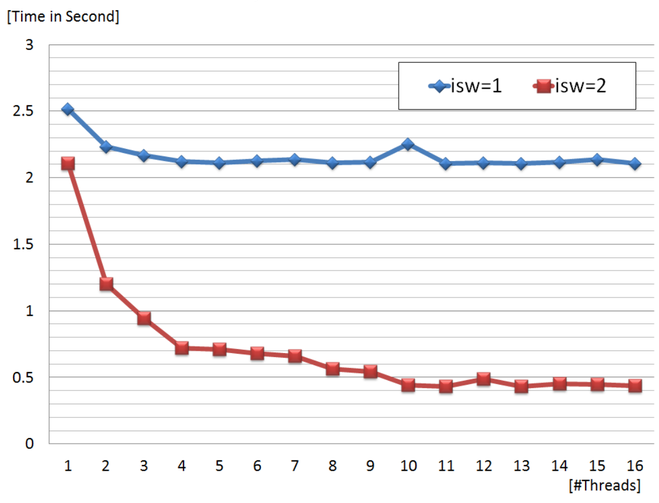
4. 性能評価（T2Kオープンスパコン1ノード（16コア））

- 日立Cコンパイラ、OpenMP並列化、最適化オプション：-Os -omp
- 後藤 BLAS ver.1.26（スレッド並列版、および逐次版）
- **ABCLibScriptによる自動生成コードをコンパイルして実行**
- 試験行列：*A*, *B*の要素を **pow(10,rand()%Φ)** で生成
- 自動チューニングの効果（実行時間[秒]）

***n* = 1000, *Φ* = 1**

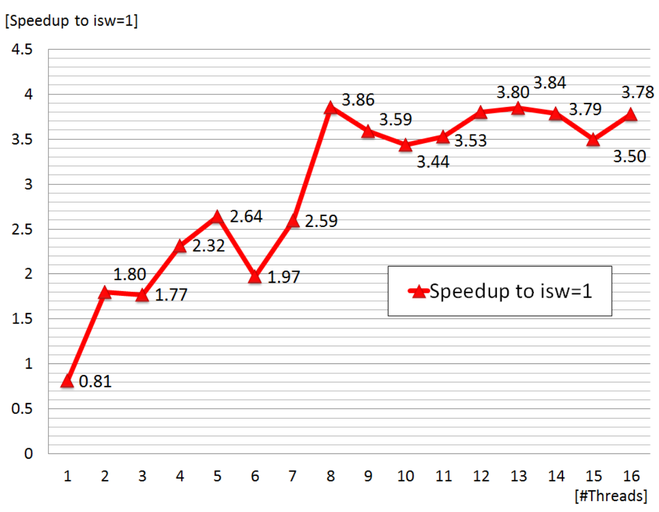


***n* = 1000, *Φ* = 10**

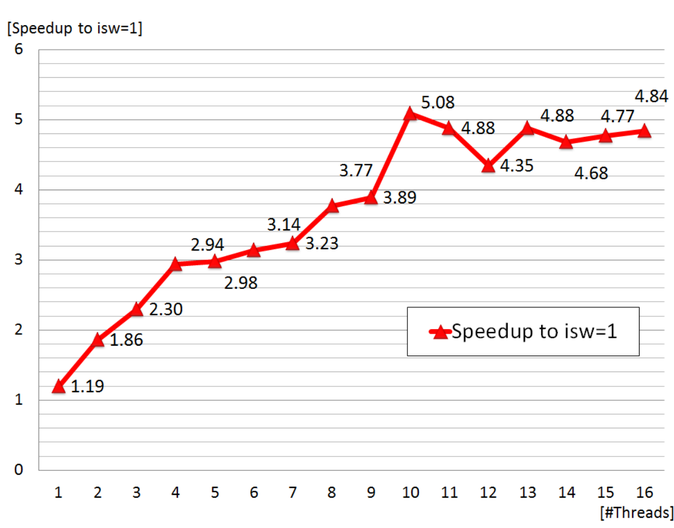


- 自動チューニングの効果（dgemmでスレッド並列化(isw=1)に対する速度向上）。isw=1はdgemmスレッド並列化。isw=2は逐次dgemmで問題レベル並列化。

***n* = 1000, *Φ* = 1**



***n* = 1000, *Φ* = 10**



- 考察
  - 逐次dgemmスレッド並列化（isw=1）が有効なのは、*Φ* = 1でスレッド数＝1の時のみ
  - それ以外は、逐次dgemmで問題レベル並列化(isw=2)が高速
  - 従来実装である逐次dgemmスレッド並列化（isw=1）に対するATの効果（速度向上）は、最大で4.8倍程度
  - スレッド数が増えるほど一般にisw=2が有効

5. 今後の課題

- ccNUMA構成特有の最適化（ファーストタッチ等）と、そのAT機能への取り込み
- 動的に切り替える「疎行列-疎行列積」用の特化ルーチンの開発
  - 実行時の演算量削減を行い高速化
  - 疎行列に切り替える条件と、そのAT機能への取り込み
- MPIによる分散メモリ並列化方式の開発

参考文献

[1] K. Ozaki, T. Ogita, S. Oishi, S. M. Rump: Error-Free Transformation of Matrix Multiplication by Using Fast Routines of Matrix Multiplication and its Applications, Numerical Algorithms, Vol.59:1, pp. 95-118, 2012.