PRIMEHPC FX10 における HPCG 最適化

富士通株式会社
TC ソリ事)計算科学ソリ統) 坂口 吉生
次世代 TC 本) アプリ開発統) 細井 聡

1 はじめに

本稿では、東京大学情報基盤センターで稼働している Oakleaf-FX (FUJITSU PRIMEHPC FX10 (以降 FX10))にて実施した High Performance Conjugate Gradients (HPCG)のチューニングと性能測定結果について報告する。

HPCG ベンチマークは、連立一次方程式 Ax=b を前処理付き CG 法(conjugate gradient method)で解く際の性能をベンチマークし、計算機パフォーマンスを評価するものである。TOP500 で利用する HPL(High Performance LINPACK)は、大規模密行列の連立一次方程式の直接解法であるが、実際のアプリケーションで求められる性能要件とは異なることが多い。一方 HPCG は、大規模疎行列の連立一次方程式の反復解法であり、実際のアプリケーションに多いという特徴がある。HPCG の詳細は、本号別記事[*1]で紹介されている。

International Supercomputing Conference (ISC'16) において HPCG ランキングが発表され、Oakleaf-FX (FX10 / 4,800nodes、76,800cores)で性能測定した HPCG 性能 0.0565 Pflops が、世界 27 位の性能[*2]を達成した。2014 年 11 月の SC'14 にて HPCG Version 2.2 で測定、登録した性能値(0.0448 Pflops)から、約 1.26 倍性能向上させた HPCG チューニング・測定について記述する。

2 HPCG ベンチマークの最適化

ISC'16 以降、HPCG ベンチマークの登録は HPCG3.0 を用いることが強く求められている。そのため、今回計測した HPCG ベンチマークは Version3.0 で実施している。HPCG Version 3.0 では、以下のような変更が Version 2 から行われている。

- Quick Path という実行モードで実行時間短縮を実現
- 問題の設定処理(行列生成部分)が HPCG 実行時間に含まれる
- HPCG version 2.4 と 3.0 の両方のスコアを表示する

HPCG ベンチマークのチューニングは、「2.1 Gauss-Seidel 前処理のブロックマルチカラーリング化」「2.2 GS と SPMV との融合」で、HPCG の Version 3.0 対応は、「2.3 HPCG3.0 対応」で記述する。

*1 中島研吾, HPCG について, スーパーコンピューティングニュース(東京大学情報基盤センター)18-5, 2016
*2 http://www.hpcg-benchmark.org/custom/index.html?lid=155&slid=288

2.1 Gauss-Seidel 前処理のブロックマルチカラーリング化

HPCG ベンチマークにおける Gauss-Seidel(以降 GS)前処理は、逐次処理であり、そのままではスレッド並列化(#pragma omp parallel for)できない。そこで、いわゆるブロックマルチカラーリング(BMC)[*3]を行って、スレッド並列化可能とする。

BMC では、行列中の連続する *k* 行を 1 つのブロックとし、ブロック間の依存関係を解析して同時実行可能なブロックを求め、同時実行可能なブロックに同じ色を割り当てる。そして、同じ色を割り当てられたブロックを複数スレッドで並列実行する。

たとえば、各ブロックが 0〜N-1 の N 個の色で塗り分けられた場合、これを M スレッドで以下「図 2-1 BMC によるスレッド並列化」のように並列実行する(異なる色の実行の間にはバリア同期が必要)。

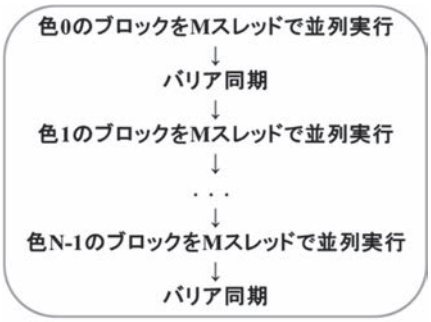

図 2-1 BMC によるスレッド並列化

また、BMC は、依存解析をブロック単位でなく 1 行単位で行う単純なマルチカラーリング(MC)に比べると、一般に反復回数の増加が少ないことが知られている。HPCG ベンチマークにおいては、反復回数が 1 回増えると 2%の性能低下として換算されてしまうので、反復回数の増加を極力抑えることは非常に重要である。

HPCG ベンチマークは 4 段のマルチグリッドであるが、全てのグリッドにおいて、この BMC によるスレッド並列化を行っている。

2.2 GS と SPMV との融合

マルチグリッド(ComputeMG)の実行において、最も粗いグリッド以外においては、プリ・スムージングのループ中の最後の Symmetric Gauss-Seidel(SYMGS)実行に引き続き、疎行列ベクトル積(SPMV)が実行される。SYMGS においては、疎行列全体を順方向、逆方向の順にアクセスする。また、SPMV においても疎行列全体にアクセスする。

また、SYMGS および SPMV の最初には、隣接プロセス間で互いに必要なベクトルの要素を送受信し合う袖通信が存在する。

---

*3 Takeshi Iwashita, Hiroshi Nakashima, and Yasuhito Takahashi. Algebraic Block Multi-Color Ordering Method for Parallel Multi-Threaded Sparse Triangular Solver in ICCG Method. In International Symposium on Parallel and Distributed Processing (IPDPS), 2012.

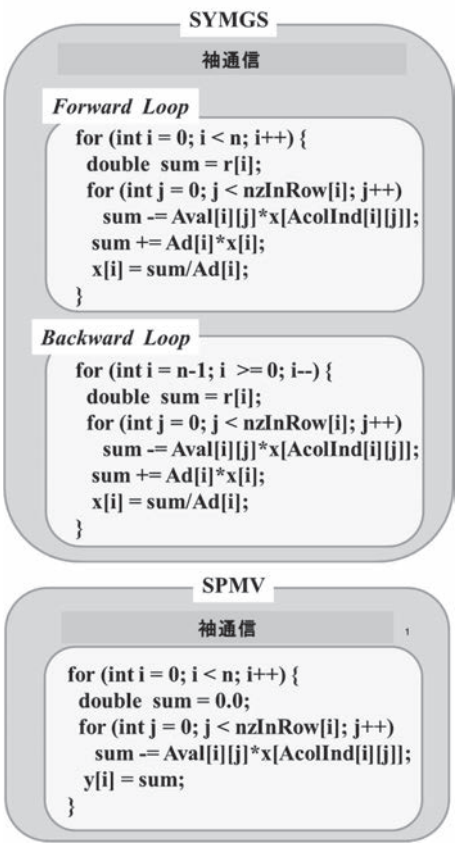

図 2-2 SYMGS と SPMV

各変数や値は以下である。

n	疎行列の行数
nzInRow[i]	疎行列の i 行目の非零要素数
Aval[i][j](0 <= j <= nzInRow[i])	疎行列の i 行目の非零要素の値
AcolInd[i][j](0 <= j <= nzInRow[i])	疎行列の i 行目の非零要素の列位置
Ad[i]	疎行列の i 行目の対角要素の値

疎行列 A は L2 キャッシュより充分大きいので、SYMGS の Backward ループ終了後 SPMV を実行する際には、Aval や AcolInd はキャッシュ上に残っておらず、再度メモリからロードする必要がある。また、SYMGS の Backwad ループにおいては、ベクトル *x* の各要素の値を更新し、更新後のベクトル *x* の要素を SPMV で参照する。SPMV は parallel for 可能なので、各行の処理は任意の順で行うことができる。

さらに、SPMV の *k* 目で参照する *x* の要素の値全てが隣接プロセスから送信してもらう必要がない場合は、それらの要素全てが SYMGS の Backward ループにおいて更新された後であれば、*k* 行目の計算を開始することができる。

そのため、SPMV の k 行目の計算をその時点で行い、SPMV の *k* 行目で参照する Aval[k][j] や AcolInd[k][j]がキャッシュに載っているうちに計算できれば高速化が期待できる。これらを GS と SPMV との融合と呼ぶことにする。

HPCG ベンチマークは 4 段のマルチグリッドであるが、最も粗いグリッド以外の 3 つのグリッドにおいて、GS と SPMV との融合を行っている。

2.3 HPCG3.0 対応

HPCG Version 2.4 からの変更点は、行列生成部分も性能測定対象に含まれるようになったことである。行列生成部に floating 演算は含まれないので、Version 2.4 に比べ全体の FLOPS 値は下がることになる。as is の行列生成部の#pragma omp parallel for 構文中には、リダクション演算とスレッドセーフでない map 型データへのアクセスに対して、#pragma omp critical 構文が用いられている。

前者に対しては、OpenMP の reduction 節を用い、後者に対してはスレッドセーフなデータ型に変更することにより、critical 構文を削除し、高速化を図っている。

3 HPCG 測定結果

Oakleaf-FX 4,800 ノード(76,800 コア)を使用して、1 プロセスあたりのグリッドサイズ(x/y/z 方向のグリッドサイズ)および、プロセス数を変えて、スラッシング影響を考慮しながら性能測定した。測定結果を「表 3-1 グリッドサイズとプロセス・スレッド数」に記載する。最も性能が高かったのは、グリッドサイズ 152×168×152、かつ、19200 プロセス×4 スレッドの場合で、この時の 56.517TFLOPS を採用し、HPCG 申請を実施した。

表 3-1 グリッドサイズとプロセス・スレッド数

グリッドサイズ	プロセス数 x スレッド数	HPCG 性能(TFLOPS)	申請
152 x 168 x 152	19200p x 4t	56.51	○
136 x 136 x 136	19200p x 4t	55.51	
136 x 248 x 232	9600p x 8t	54.11	

FX10 システムでは、メモリアクセス高速化のために、ラージページ対応しているが、指定可能なサイズは 8KB、4MB、32MB、256MB である。指定可能な種別として、スタックセグメント、データセグメント、ヒープセグメント、スレッドスタックセグメントがあるが、すべてのセグメントで、最高性能であるラージページサイズ 32MB を適用した。

また、インターコネクト通信(Tofu [*4]：6 次元メッシュトーラス) のネットワーク形状は、最大形状である 20 x 15 x 16 で、ノード割り当ては rank-map-bynode (1 プロセスを割り当てると、別のノードへ移動してプロセスを割り当て、すべてのノードに一通り割り当てると、最初に割り当てたノードに戻る)を適用した。

*4 Tofu (Torus fusion) は、富士通の高速インターコネクトの呼称。

HPCG version 3.0 では、3.0 性能以外に、移行を考慮して 2.4 のスコアも、YAML 出力結果ファイルの最終行付近に、以下のように表示される。

```
⋮ (省略)
_____ Final Summary _____:
HPCG result is VALID with a GFLOP/s rating of: 56516.8
HPCG 2.4 Rating (for historical value) is: 57674.5
```

以下に、SC'14(2014 年測定時)の HPCG 性能値と、ISC'16(2016 年今回測定時)の HPCG 性能値を「表 3-2 HPCG 性能」に記載する。

表 3-2 HPCG 性能

Version	PFLOPS	理論ピーク(1.135PFLOS)比	HPCG/HPL
2.2	0.044847	3.9%	4.3%
3.0	0.056516	5.0%	5.4%
3.0 (2.4 換算)	0.057674	5.1%	5.5%

4 まとめ

ISC'16 において HPCG ランキングが発表され、Oakleaf-FX (FX10 / 4,800nodes、76,800cores)で測定した HPCG 性能 0.0565 Pflops が、世界 27 位の性能を達成した。

本稿では、SC'14(HPCG Version 2.2)で測定、登録した値から、約 1.26 倍性能向上させた HPCG ベンチマークのチューニング・測定について報告した。

HPCG の実行により、HPC 分野で利用される実際のアプリケーション実行に必要なハードウェア・ソフトウェア性能が判る。今回測定結果の高い実効性能(ピーク性能比で 5%)は、東京大学情報基盤センターFX10 (Oakleaf-FX/Oakbridge-FX)において、アプリケーション実行の性能向上に貢献できると考える。

謝辞

本稿で紹介した計算は、2016 年 6 月 2～3 日に実施された「大規模 HPC チャレンジ」において実行された。この場をお借りして深く感謝いたします。