Oakleaf-FX フルノードによる HPCG 実行結果の概要（2014 年 10 月大規模 HPC チャレンジ）

中 島 研 吾

東京大学情報基盤センター

1. HPCG とは？

本稿は 2014 年 10 月 30〜31 日に実施された「大規模 HPC チャレンジ」における HPCG 実行結果の概要について述べるものである。

スーパーコンピュータシステムの性能指標としては Linpack という密行列を係数行列とする大規模連立一次方程式の解を求めるベンチマークがこれまで使用されて来ており，1993 年に始まった TOP500[1]では世界のスーパーコンピュータの性能のランキングの情報を得ることができる。TOP500 は年 2 回（11 月の SC-XY（アメリカで開催される International Conference on High Performance Computing, Networking, Storage and Analysis，XY：年号）と 6 月の ISC-XY[2]（ドイツで開催される International Supercomputing Conference）で更新される。

Linpack で高性能を出すためにはできるだけ大規模な問題を解く必要があるが，計算量が未知数の 3 乗に比例するため，PFLOPS 級のシステムでは「日（day）」のオーダーの計算時間が必要となる。EFLOPS 級ではこれが「週」から「月」へどんどん長くなっていくことが予想されている。システムの安定性の確認には適しているが，5 年程度の稼働時間の中で 1 ヶ月を性能計測に費やすことは無駄である，という意見が強くなって来た。1 ヶ月ということになると電気代だけでも億単位（円）の費用が必要になるのである。そこで，限られた資源を有効に使用するために，短時間で効率的にシステム性能を計測できるベンチマーク手法がここ数年求められてきた。また Linpack で実施している計算は実際のスパコンで稼働しているアプリケーションとも大きな隔たりがあり，より実アプリケーションに近いベンチマーク手法の確立も併せて議論されてきた。

HPCG（High Performance Conjugate Gradients）[3]はそのような背景から開発され，Linpack が密行列ソルバーなのに対して，HPCG は有限要素法から得られる疎行列を対象とした線形ソルバーである。三次元ポアソン方程式を差分格子のような規則的形状において有限要素法によって離散化して得られる疎行列を係数とする連立一次方程式を幾何学的マルチグリッド前処理（Geometric Multgrid Preconditioning）による共役勾配法（Conjugate Gradient Method）を使用して解いており，連立一次方程式を解いている部分の計算性能（FLOPS 値）によって性能を算出する（ただ現状のルールでは計算の準備に要する部分の計算時間も一部も求解部分の計算時間に加算して FLOPS 値を算出することになっているようである）。Multigrid の smoother としては Gauss-Seidel 法が使用されており，スレッド並列化にはマルチカラー法によるリオーダリング

[1] http://supercomputing.org/

[2] http://www.isc-hpc.com/

[3] http://hpcg-benchmark.org/

が適用されている。

大規模問題向けのマルチグリッド法を前処理手法として採用しているので，ポストペタスケール，エクサスケールのシステムでのアプリケーションを考慮すると，より実用的な手法であると考えられる。計測手法についてはまだ検討中の段階であるが，現在は最低 60 分実行すること，となっており，Linpack と比較するとだいぶ短い。

HPCG のアイディアが最初に示されたのは 2013 年 6 月の ISC'13（Leipzig, Germany），SC13（Denver）で Ver.1.0 が公開された。SC13 でのミーティング，ワークショップによる意見交換などを通じて，計測方法のルール，実施者がチューニングできる箇所や選択できるオプションなども固まりつつある。プログラムは C++で記述されており，上記 HP からダウンロード可能である。Intel MKL，Nvidia GPU 等に最適化されたバージョンも開発中とのことである。

ISC'14（Leipzig）で初めてランキング（15 システム）が公開された。SC14 では 25 システムに増加し，初めて Top 3 の表彰も行われた。Top5 の概要は表 1 の通りである

表 1　HPCG の上位 5 システム（SC14，2014 年 11 月）

	System	TOP500 Ranking	HPCG (TFLOPS)	Ratio Linpack	Ratio Peak
1	Tianhe-2	1	632	1.8%	1.1%
2	京	4	461	4.4%	4.1%
3	Titan	2	322	1.8%	1.2%
4	Mira	5	167	1.9%	1.7%
5	Piz Daint	6	105	1.7%	1.3%

各システムの詳細は TOP500 の HP，後掲の表 5 を参照されたいが，Titan と Piz Daint は Intel Xeon と NVIDIA Kepler の複合システム，Mira は IBM BlueGene/Q である。最後の 2 列は HPCG ベンチマークの性能の Linpack 性能，ピーク性能に対する比である。HPCG は疎行列を対象とするため対ピーク性能比は密行列を対象とした Linpack と比較して低い。チューニングにも大きく依存するが，上の表で示したように「京」と「Titan」では TOP500 と HPCG では順位が入れ替わっていることがわかる。

2. HPCG 実行結果

今回は 2014 年 10 月 30 日に実施された大規模 HPC チャレンジにおいて，Oakleaf-FX（Fujitsu PRIMEHPC FX10）のフルノード（4,800 ノード）を使用して HPCG を実行した。実行プログラムは理化学研究所計算科学研究機構において京コンピュータ向けに最適化されたもの（表 1 に示したものと同じ）を使用した[4]。

今回は，以下の 2 点をパラメータとして様々なケースを実施したが，そのうち最適な結果を得られたものの中からいくつかの計算例を紹介する：

・並列プログラミングモデル（Flat MPI，Hybrid（HB）2×8，HB 4×4，HB 8×2，HB 16×1）
・1-MPI プロセスあたり問題サイズ

[4] http://www.hpcg-benchmark.org/downloads/sc14/HPCG_on_the_K_computer.pdf

Oakleaf-FX は各計算ノード上に 16 コアを有している。Flat MPI は，MPI のみを使用して並列化したものであり，計算ノード上の各コアを独立の MPI プロセスとしたものである。それに対して HB M×N は OpenMP+MPI のハイブリッド並列プログラミングモデルを適用しており，M=MPI プロセス当りスレッド数，N=ノード内 MPI プロセス数，である。Oakleaf-FX では 1 コア当 1 スレッドが上限であるので，M と N の積が 16 となるように各ケースを設定している。例えば HB 8×2 は 8 スレッドの MPI プロセスがノード内に 2 つある，ということになる。

HB M×N において，M の値が大きくなるほど，同じ計算機リソースを使用しても MPI プロセスは少なくなるため，通信のオーバーヘッドは減少することが予想されるが，1 プロセス当のデータ量は概して大きくなる。

HPCG では各 MPI プロセスあたりの問題サイズを nx，ny，nz というパラメータで制御することができ，これは X，Y，Z 方向の節点数に相当する。したがって，各 MPI プロセスあたり問題サイズは nx×ny×nz である。今回は各ケースにおいて nx=ny=nz として計算を実施した。また，全体モデルの配置も各方向におけるプロセス数（npx，npy，npz）によって定めることができる。

表 2 は MPI プロセスあたりの問題サイズを変化させて，4,800 ノード（76,800 コア）を使用して HB 4×4 によって計算した例である。一般的に 1 プロセスあたりの問題規模が大きい方が通信のオーバーヘッドの影響が相対的に少なくなるため，ノード数が大きい場合は有利であるが，逆にメモリへの負担が増えてプロセスあたりの性能が低下する場合があるため，最適値の選定にあたっては注意が必要である。表 2 の場合は nx=ny=nz=144 の場合に最適値が得られている。表 3 は各並列プログラミングモデルにおいて，最適な問題サイズを使用した場合の 4,800 ノード（76,800 コア）における測定結果である。問題サイズは異なっているが TFLOPS 値では HB 4×4 が最も良い性能である。

表 2　HPCG 計算結果（HB 4×4，Oakleaf-FX（4,800 ノード，76,800 コア)），問題サイズの影響

nx, ny, nz	MPI プロセス数	合計問題サイズ	HPCG 性能（TFLOPS）
136	19,200	48,296,755,200	41.8
144	19,200	57,330,892,800	44.8
152	19,200	67,426,713,600	43.2

表 3　HPCG 計算結果（Oakleaf-FX（4,800 ノード，76,800 コア）における最適値）

	最適 nx, ny, nz	MPI プロセス数	合計問題サイズ	HPCG 性能（TFLOPS）
Flat MPI	104	76,800	86,389,555,200	42.5
HB 2×8	136	38,400	96,593,510,400	43.2
HB 4×4	144	19,200	57,330,892,800	44.8
HB 8×2	144	9,600	28,665,446,400	43.2
HB 16×1	144	4,800	14,332,723,200	39.5

また，表 4 は HB 4×4 において nx=ny=nz=144 としてノード数を 64 ノードから 4,800 ノードまで変化させた場合の計算結果と，64 ノードを規準とした並列化効率である。4,800 ノードの場合でも 97%を超えるような高い並列化効率が得られているが，この理由は，HPCG は問題を最後まで解いているわけではなく，あくまでも計算効率を求めているためである。

表 4　HPCG 計算結果（HB 4×4，nx=ny=nz=144），ノード数による性能の変化

ノード数	MPI プロセス数	HPCG 性能（TFLOPS）	並列化性能（%），64 ノードの 100%
64	256	0.614	100.0
512	2,048	4.91	99.9
2,048	8,192	19.2	97.9
4,800	19,200	44.8	97.4

3. SC14 における結果

表 5 は表 1 に示した，SC14 における HPCG ランキングの結果（抜粋）である。25 システムのうち，Oakleaf-FX は 13 位となっている。ここで順位そのものについて論ずることはあまり意味が無いが，TOP500 で 7 位，30 位と Oakleaf-FX よりも上位のシステムよりも高い値が得られていることがわかる。

謝 辞

本計算の実行に当たって，プログラムを貸与いただいた熊畑清博士，南一生氏（理化学研究所計算科学研究機構）に深甚なる謝意を表するものである。

表 5　SC14 における HPCG ランキング（抜粋）

	Site	Computer	Cores	HPL Rmax	TOP500 Rank	HPCG
1	NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.863	1	0.623
2	RIKEN Advanced Inst for Comp Sci	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	705,024	10.510	4	0.461
3	DOE/SC/Oak Ridge Nat Lab	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x	560,640	17.590	2	0.322
4	DOE/SC/Argonne National Laboratory	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom	786,432	8.587	5	0.167
5	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x	115,984	6.271	6	0.105
6	Leibniz Rechenzentrum	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR	147,456	2.897	12	0.083
7	DOE/SC/LBNL/NERSC	Edison - Cray XC30, Intel Xeon E5-2695v2 12C 2.4GHz, Aries interconnect	133,824	1.655	18	0.079
8	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x	76,032	2.785	13	0.073
9	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR	65,320	1.283	27	0.061
10	CEA/TGCC-GENCI	Curie thin nodes - Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR	77,184	1.359	26	0.051
13	Information Technology Center, The University of Tokyo	Oakleaf-FX - PRIMEHPC FX10, SPARC64 IXfx 16C 1.848GHz, Tofu interconnect	76,800	1.043	36	0.045
14	Texas Advanced Computing Center/Univ. of Texas	Stampede - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P	462,462	5.168	7	0.044
15	International Fusion Energy Research Centre (IFERC), EU(F4E) - Japan Broader Approach collaboration	Helios Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR	70,560	1.237	30	0.043
18	Cyberscience Center Tohoku University	NEC SX-ACE 4C+IXS	2,048	0.123	N/A	0.013
24	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	2,720	0.150	435	0.004
25	SURFsara	Cartesius - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4_ FDR, Nvidia K40m	3,036	0.154	419	0.003