Vol.2012-HPC-135 No.43
2012/8/3

大規模超並列スーパーコンピューターシステム Oakleaf-FX(Fujitsu PRIMEHPC FX10) の性能評価

大島 聡史[1,a)]　實本 英之[1]　鴨志田 良和[1]　片桐 孝洋[1]　田浦 健次朗[1,2]　中島 研吾[1]

概要：本稿では東京大学情報基盤センターにおいて 2012 年 4 月に稼働を開始したスーパーコンピューターシステム Oakleaf-FX (Fujitsu PRIMEHPC FX10) の性能について報告する．Oakleaf-FX は SPARC64IXfx プロセッサや FEFS ファイルシステムを採用した大規模な並列計算機システムであり，理論浮動小数点演算性能は 1.13PFLOPS である．また京速コンピュータ「京」とも高い互換性を備えており，計算科学・計算機科学の発展に大きく寄与できるものと考えられる．本稿では実機にていくつかのベンチマークを用いて性能評価を行った結果を報告する．

Performance Evaluation of Oakleaf-FX (Fujitsu PRIMEHPC FX10) Supercomputer System

Satoshi OHSHIMA[1,a)]　Hideyuki JITSUMOTO[1]　Yoshikazu KAMOSHIDA[1]
Takahiro KATAGIRI[1]　Kenjiro TAURA[1,2]　Kengo NAKAJIMA[1]

Abstract: We report the performance of Oakleaf-FX (Fujitsu PRIMEHPC FX10) supercomputer system which has begun in April 2012 at Kashiwa campus, Information Technology Center, The University of Tokyo. This system is a large-scale parallel computer with SPARC64IXfx CPU and FEFS file system. The peak performance is 1.13 PFLOPS. Moreover, this system is compatibility of the K computer and expected to contribute a lot to progress of computer/computational science. In this paper, we report some results of performance evaluation on this supercomputer system.

1. はじめに

東京大学情報基盤センター（以下，当センター）では 2011 年に既設のスーパーコンピュータシステム SR11000/J2 システム（以下 SR11000/J2）の使用期限を迎えた．後継のシステムとしては，従来の SR11000/J2 システム利用者の継続性を重視した「大規模 SMP 並列スーパーコンピューターシステム」と大規模超並列計算向けの「大規模超並列スーパーコンピューターシステム」の 2 システムを導入することとなった．前者については日立製作所社製の SR16000 モデル M1 システム（愛称[*1] Yayoi）に決定し，2011 年 10 月に運転を開始し，同年 11 月より正式サービスを開始している [1][2]．また後者については富士通社製の PRIMEHPC FX10 システム（愛称[*1] Oakleaf-FX）[3] に決定し，2012 年 4 月に試験運用開始，7 月に正式サービス開始となっている [4]．これら 2 システムはそれぞれ異なるアーキテクチャを採用した最新の計算機システムである．

本稿では Oakleaf-FX の性能について，実システムを用いたベンチマーク結果を用いて報告する．今回用いたベンチマークプログラムは以下の 6 種類である．

(1) STREAM ベンチマーク

(2) HPL ベンチマーク

(3) MPIFFT ベンチマーク

[1] 東京大学 情報基盤センター
Information Technology Center, The University of Tokyo
[2] 東京大学 大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
[a)] ohshima@cc.u-tokyo.ac.jp

[*1] 愛称は設置場所にちなんでおり，Yayoi は本郷キャンパス浅野地区（東京都文京区弥生，弥生=Yayoi），Oakleaf-FX は柏キャンパス（千葉県柏市柏の葉，柏の葉=oakleaf）に設置されている

© 2012 Information Processing Society of Japan

(4) GeoFEM ベンチマーク

(5) MDTEST ベンチマーク

(6) IOR ベンチマーク

本稿の構成は以下の通りである．2 章では Oakleaf-FX のハードウェア構成について述べる．3 章では Oakleaf-FX 上で様々なベンチマークプログラムを実行した結果を実行時に意味のあったコンパイラオプションなどの情報とともに報告する．4 章はまとめの章とする．

なお Oakleaf-FX は性能を向上させるためドライバやソフトウェアの更新，コンパイラオプションの最適化などを日々進めており，また本稿は試験運転期間中に執筆している．そのため本稿における内容と本サービス開始後の実システムでの測定値や設定は一部異なる可能性がある．

2. ハードウェア構成

2.1 全体構成

本章では Oakleaf-FX の主なハードウェア構成について述べる．

はじめに Oakleaf-FX の全体構成について述べる．図 1 に Oakleaf-FX の全体構成図を示す．また表 1 には Oakleaf-FX の性能諸元を旧システムである SR11000/J2 や現有システムである HA8000（T2K 東大版）および SR16000/M1 とあわせて示す．

Oakleaf-FX は 4800 の計算ノード，複数種類の大規模ストレージ，高速な内部ネットワーク（Tofu ネットワーク），そしてログインノードや各種管理用ノードから構成される計算機システムである．GPU などのアクセラレータを搭載しない均質な計算機システムであり，1.13PFLOPS の総理論演算性能を備えている．計算ノードの冷却には水冷と空冷が併用されており，Linpack 測定時最大消費電力は 1.40MW 未満（空調等を含めた総消費電力は 2.0MWh 未満）である．

Oakleaf-FX(Fujitsu PRIMEHPC FX10) は理化学研究所 計算科学研究機構に設置されている京速コンピュータ「京」の商用版後継機システムの 1 つである．そのため Oakleaf-FX と「京」はハードウェア・ソフトウェアともに共通点が多く共通の最適化手法が有効となる可能性が高い．

2.2 計算ノード

図 2 に計算ノードの構成を示す．

Oakleaf-FX は計算ノードの CPU として富士通が開発した SPARC64IXfx を搭載している．SPARC64IXfx は 16 コアによって構成されているマルチコアプロセッサであり，SPARC64 系 (SPARC64V9 + HPC-ACE) アーキテクチャの CPU である．Oakleaf-FX に搭載されている SPARC64IXfx の動作周波数は 1.848GHz である．SMT 機能は搭載されておらず，1CPU あたり 236.5GFLOPS(1.848GHz×8IPC×16 コア) の理論倍精度浮動小数点演算性能を備えている．キャッシュについては，L1 データキャッシュと L1 命令キャッシュを各コアごとに 32KB ずつ，L2 キャッシュを 1CPU あたり 12MB 搭載しており，L3 キャッシュは搭載していない．特徴的な仕様や機能としては，命令レベル並列性を活用するためのレジスタ数拡張，再利用性のあるデータを選択的に残すためのセクターキャッシュ機能，高速な並列計算をサポートする VISIMPACT・ハードウェアバリア機能などがあげられる．

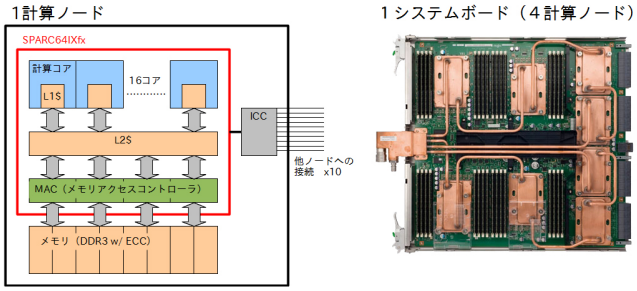

図 2 SPARC64IXfx CPU と計算ノードの構成

計算ノード 1 ノードには SPARC64IXfx 1 基（1 ソケット）に加えて，メインメモリ（ECC 付き DDR3 メモリ）32GB とノード間通信を担当するインターコネクトコントローラ（Inter Connect Controller, ICC）が搭載されている．また計算ノード 4 組によってシステムボードが構成されており，システムボード単位でラックに搭載されている．計算ノード群全体では 4,800 の計算ノード（=76,800 の計算コア）が搭載されており，総理論演算性能は 1.13PFLOPS，総主記憶容量は 150TByte である．

2.3 計算ノード間ネットワーク

Oakleaf-FX の計算ノード間ネットワークには 6 次元メッシュ/トーラスインターコネクト (Tofu[*1] インターコネクト) が用いられている．ネットワーク構成の概要を図 3 に示す．

Oakleaf-FX の各計算ノードに搭載された各 ICC は最大 10 個の ICC と相互接続しており，4 方向送信と 4 方向受信を同時に行うことが可能である．10 個の接続はそれぞれ X 軸 (X+,X-)，Y 軸 (Y+,Y-)，Z 軸 (Z+,Z-)，A 軸，B 軸 (B+,B-)，C 軸に対応しており，X 軸と Y 軸は筐体間，Z 軸と B 軸はシステムボード間，A 軸と C 軸はシステムボード上を接続している．さらに X,Y,Z 座標が同一の計算ノード 12 ノードごとに Tofu 単位を形成している．

以上の構成により Oakleaf-FX の計算ノードネットワークは高速性と高信頼性（冗長性）を兼ね備えている．またシステムを利用するユーザから見た計算ノード間ネット

[*1] Torus fusion

© 2012 Information Processing Society of Japan

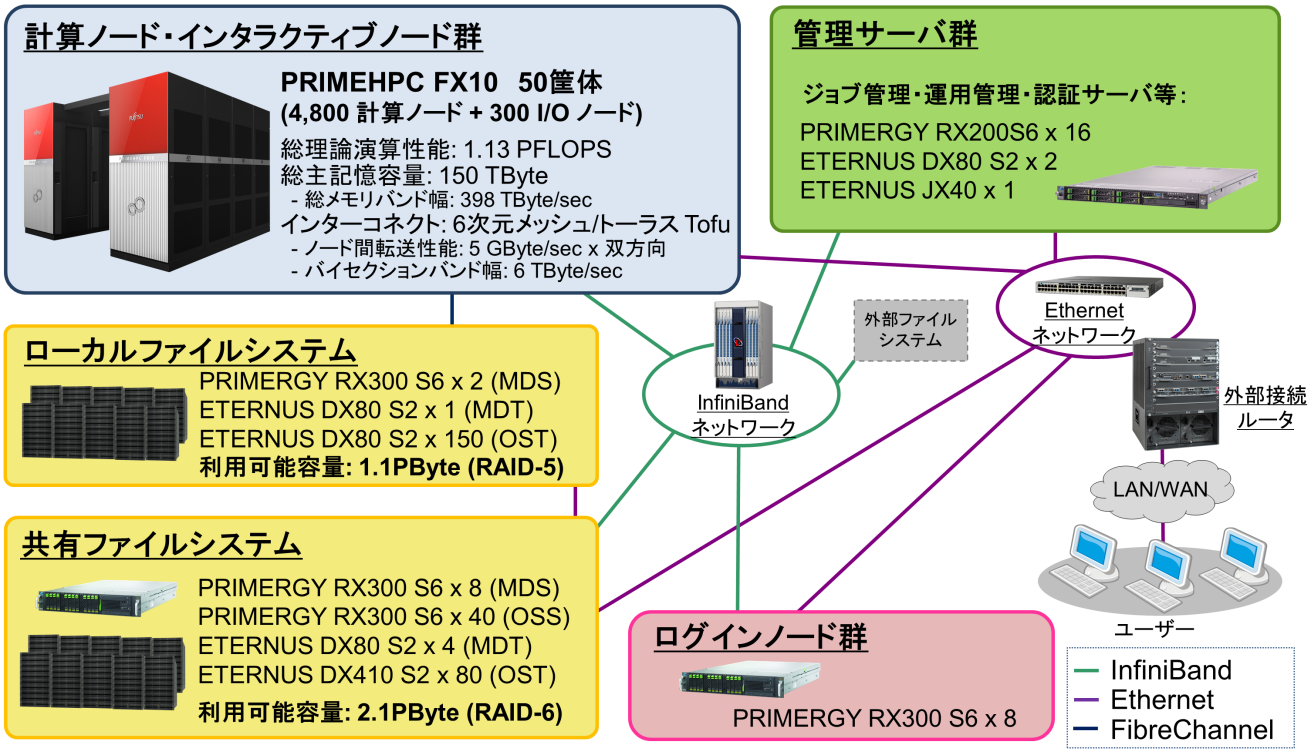

図 1　Oakleaf-FX の全体構成

表 1　Oakleaf-FX の性能諸元

	PRIMEHPC FX10 (Oakleaf-FX)	SR16000/M1 (Yayoi)	SR11000/J2 (旧システム)	HA8000 (T2K 東大版)
CPU	SPARC64IXfx 1.848 GHz	Power7 3.83 GHz	Power5+ 2.3 GHz	Opteron8356 2.3 GHz
総計算ノード数	4800	56	128	952
コア数/計算ノード	16	32	16	16
理論演算性能/コア	14.784 GFLOPS	30.64 GFLOPS	9.2 GFLOPS	9.2 GFLOPS
理論演算性能/計算ノード	236.5 GFLOPS	980.48 GFLOPS	147.2 GFLOPS	147.2 GFLOPS
理論演算性能/全計算ノード	1.13 PFLOPS	54906.88 GFLOPS	18841.6 GFLOPS	140.1344 TFLOPS
主記憶容量/計算ノード (使用可能容量)	32 GByte (28 GByte)	200 GByte (170 GByte)	128 GByte (112 GByte)	32 GByte (28 GByte)
主記憶容量/全計算ノード	150 TByte	11200 GByte	16384 GByte	31.25 TByte
B/F 値	0.36	0.52	1.39	0.29
SMT 機能	非対応	最大 4 スレッド/コア 運用時最大 2 スレッド/コア	非対応	非対応
計算ノード間 ネットワーク構成	6 次元メッシュ/トーラス (Tofu ネットワーク)	階層型完全結合	3 段クロスバー	フルバイセクション バンド幅 FatTree
計算ノード間転送性能	20 GByte 双方向 4 方向同時通信可能	96 GByte/sec 双方向	12 GByte/sec 双方向	A 群 5 GByte/sec 双方向 B 群 2.5 GByte/sec 双方向
ストレージ容量	1.1 PByte + 2.1 PByte (+ 3.6 PByte)	556 TByte	94.2 TByte	1 PByte
CPU/主記憶間物理転送 性能/計算ノード	85 GByte/sec	512 GByte/sec	204.6 GByte/sec	42 GByte/sec

ワークは最大で 3 次元のトーラス空間となり，複雑な 6 次元の形状を強く意識せずとも常に高いネットワーク性能を得ることができる．

2.4　ストレージ

Oakleaf-FX は 2 系統のストレージシステム，ローカルファイルシステムと共有ファイルシステムを備えている．ローカルファイルシステムはステージング用に用意されたシステムである．PRIMERGY RX300 S6 と ETERNUS DX80 S2 から構成されており，1.1PByte の容量と 131GByte/sec の性能を備えている．一方の共有ファイルシステムは全計算ノードに加えてログインノードからも

© 2012 Information Processing Society of Japan

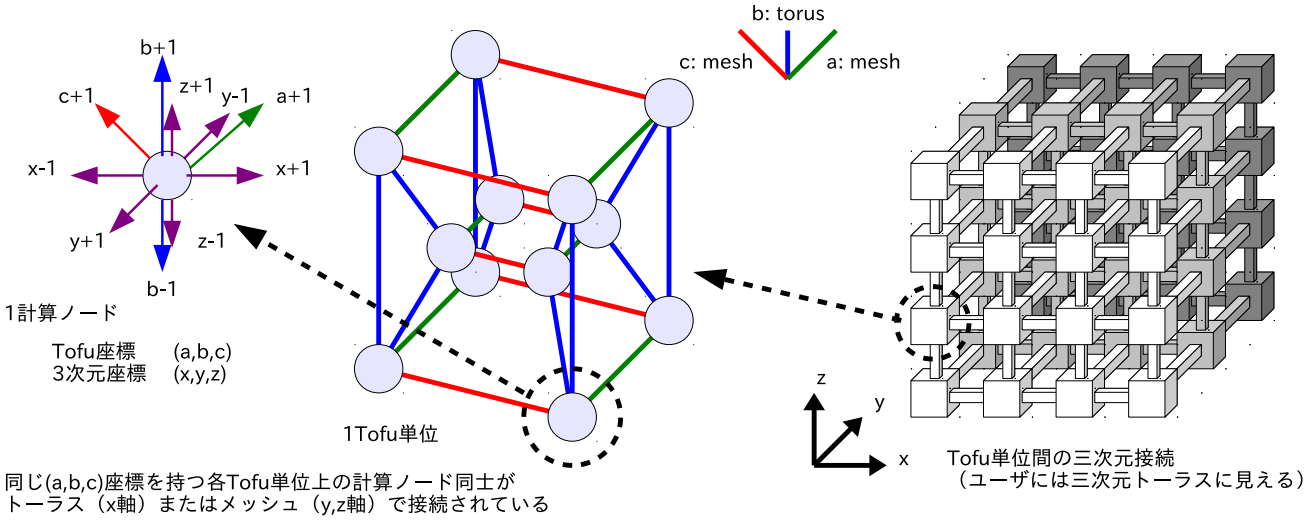

図 3　Oakleaf-FX のネットワーク構成（Tofu インターコネクトの概要）

利用可能なシステムである．PRIMERGY RX300 S6 と ETERNUS DX80 S2 に加えて ETERNUS DX410 S2 から構成されており，2.1PByte の容量と 136GByte/sec の性能を備えている．ホームディレクトリは共有ファイルシステムに置かれる．

これら 2 種類のストレージシステムはいずれもファイルシステムとして FEFS(Fujitsu Exabyte File System) を用いている．FEFS は富士通が Lustre をベースに開発したファイルシステムであり，大規模システム向けに最大ファイルサイズや最大ファイル数などの拡張がなされている．

また Oakleaf-FX の外部に**外部ファイルシステム**も用意されている．外部ファイルシステムは 3.6PByte の容量を持つ大容量のストレージシステムであり，一時的な作業ファイルを置くいわゆるワーク領域として使用することが想定されている．外部ファイルシステムのファイルシステムには Lustre が用いられている．

3. ベンチマークによる性能評価

3.1 STREAM ベンチマーク

STREAM ベンチマーク [6] を用いて計算ノードのメモリ性能を測定した．STREAM ベンチマークは配列に対して “特定の処理” を繰り返し実行した際の実行時間からメモリ性能（MB/s）を算出する．“特定の処理” としては，

- 配列のコピーを行う Copy（c[j] = a[j]）
- 配列とスカラーとの乗算を行う Scale（b[j] = scalar*c[j]）
- 二つの配列を加算する Add（c[j] = a[j]+b[j]）
- スカラーとの乗算と配列加算を組み合わせた Triad（a[j] = b[j]+scalar*c[j]）

が用意されている．計算内容によってメモリのロード回数とストア回数に違いがあるため，ハードウェアアーキテクチャにより値の傾向が異なる．

表 2　STREAM 測定結果（単位は MB/sec，括弧内は理論性能比）

	Oakleaf-FX PRIMEHPC FX10	Yayoi SR16000/M1
Copy	59987.3012 (68.9%)	224825.3361 (42.9%)
Scale	59768.9227 (68.7%)	226349.5329 (43.2%)
Add	64640.5627 (74.3%)	256364.6680 (48.9%)
Triad	64712.2441 (74.3%)	255192.6583 (48.7%)

1 計算ノード上で OpenMP を用いて 16 スレッド実行した際の性能を測定した．プログラムの作成には富士通 Fortran コンパイラを使用した．主なコンパイルオプションとしては-Kopenmp -Kfast -KXFILL -Kprefetch_sequential=soft -Kprefetch_double_line_L2 -Kprefetch_line_L2=64 -Koptmsg -Qt を指定した．（実行時間測定部のみ C 言語による記述を用いたため富士通 C コンパイラを使用した．）問題サイズ (N) は 80,000,512，計算繰り返し回数 (NTIMES) は初期値 (10) とした．

実行時環境変数 OMP_NUM_THREADS および PARALLEL に 16 を設定してプログラムを実行した．測定結果を Yayoi(SR16000/M1，32 スレッド実行，参考文献 [2] から引用) と比較して表 2 に示す．性能値自体を比較すると Oakleaf-FX のメモリ性能は Yayoi の 25%程度であるが，理論性能比については Yayoi が 50%未満であるのに対して Oakleaf-FX は 68%以上の性能が得られている．

3.2 HPL ベンチマーク

HPCC ベンチマーク (HPCC 1.4.0)[7] に含まれる HPL

© 2012 Information Processing Society of Japan

の性能を測定した．このベンチマークは LU 分解による連立一次方程式の求解を行うものであり，特に行列-行列積計算（BLAS3 DGEMM）の性能がベンチマークスコアに大きな影響を与えるベンチマークである．

プログラムの作成には富士通 C コンパイラ，および富士通 Fortran コンパイラを利用した．BLAS ライブラリは富士通社が独自開発した BLAS ライブラリを使用した．主なコンパイルオプションとして-O3 -Kopenmp,parallel,fast -Nsrc,sta -Koptmsg（C 言語，翻訳時），-Kopenmp,parallel,ocl,fast -Koptmsg -Qt（Fortran 言語，統合時）を指定した．

3.2.1 1 ノード性能

実行環境としては，計算ノード 1 ノード，1 ノードあたり 16CPU コアを使用し，各計算ノードにおける MPI プロセス数を 1，プロセスあたりスレッド数を 16（ハイブリッド MPI を想定，ただし 1 ノード実行では MPI 実行なし）として実行した．主な問題設定（hpccinf.txt に指定する値）としては以下の値を用いた．

- Ns = 56000, NBs = 448, Ps = 1, Qs = 1

実測性能および理論演算性能に対する性能割合は以下の通りとなった（小数点第 3 位以下切り捨て）:

- 1 ノード 0.21 TFLOPS, 90.59%

Yayoi における 1 ノード性能（0.83 TFLOPS, 84.65%）[2] と比較すると，絶対性能は低いが理論性能比は高いことが確認できた．

3.2.2 全系性能

実行環境としては，計算ノード 4800 ノード，1 ノードあたり 16CPU コアを使用し，各計算ノードにおける MPI プロセス数を 1，プロセスあたりスレッド数を 16（ハイブリッド MPI）として実行した．合計のスレッド数 × プロセス数は 4800 となる．主な問題設定（hpccinf.txt に指定する値）としては以下の値を用いた．

- Ns = 4058880, NBs = 448, Ps = 30, Qs = 160

実測性能および理論演算性能に対する性能割合は以下の通りとなった（小数点第 3 位以下切り捨て）:

- 4800 ノード 1.04 PFLOPS, 91.89%

なお，2012 年 6 月発表の TOP500 List において上記の性能は 18 位にランキングされた [14]．この時の電力は 1176.80kW であった．

3.3 MPIFFT ベンチマーク

HPCC ベンチマーク (HPCC 1.4.0) に含まれる FFT の性能を測定した．このベンチマークは一次元の高速フーリエ変換を行うものであり，全対全通信（MPI_Alltoall）の性能がベンチマークスコアに大きな影響を与えるベンチマークである．

プログラムの作成には富士通 C コンパイラ，および富士通 Fortran コンパイラを利用した．数値計算ライブラリとして，FFTW, SSLII，SSLII スレッド並列機能および BLAS/LAPACK スレッド並列版を利用した．主なコンパイルオプションとして-Kfast -Kopenmp -Nsrc,sta -Koptmsg（C 言語，翻訳時），-Kfast -Kopenmp -mlcmain=main -SSL2BLAMP（Fortran 言語，統合時）を指定した．

コンパイル時の HPCC 固有のマクロオプションは以下を使用した : -DHPCC_FFT_235 -DHPCC_MEMALLCTR -DRA_SANDIA_NOPT (mpifft.o, wrapmpifftw.o, pzfft1d.o のみ-DUSING_FFTW を追加)

実行環境としては計算ノード 8 ノード（インタラクティブノード）を使用し，計算ノード 1 ノードあたり MPI プロセス数を 1，プロセスあたりスレッド数を 16（ハイブリッド MPI）として実行した．合計のスレッド数 × プロセス数は 128 となる．主な問題設定（hpccinf.txt の設定値）としては以下の値を用いた．

- Ns = 160000, NBs = 80, Ps = 1, Qs = 8

なお，この時の Vector Size は以下となる．

- Vector size: 3,200,000,000

実測性能としては

- 30.213 GFLOPS, 1.59%

の性能を得た．

Yayoi における 8 ノード性能（151.121 GFLOPS, 1.92%）[2] と比較すると絶対性能・理論性能比ともに低いが，B/F 値の差などを踏まえると Oakleaf-FX の性能は良好であると判断できる．

3.4 GeoFEM ベンチマーク

3.4.1 概要

GeoFEM プロジェクト [8] で開発された並列有限要素法アプリケーションを元に整備した性能評価のためのベンチマークプログラム GeoFEM-Cube[9] による評価を実施した．オリジナルの GeoFEM ベンチマーク [10] は，

(1) 三次元弾性静解析問題（Cube 型モデル，PGA モデル）

(2) 三次元接触問題

(3) 二重球殻間領域三次元ポアソン方程式

に関する並列前処理付き反復法ソルバーの実行時性能（GFLOPS 値）を様々な条件下で計測するものである．

© 2012 Information Processing Society of Japan

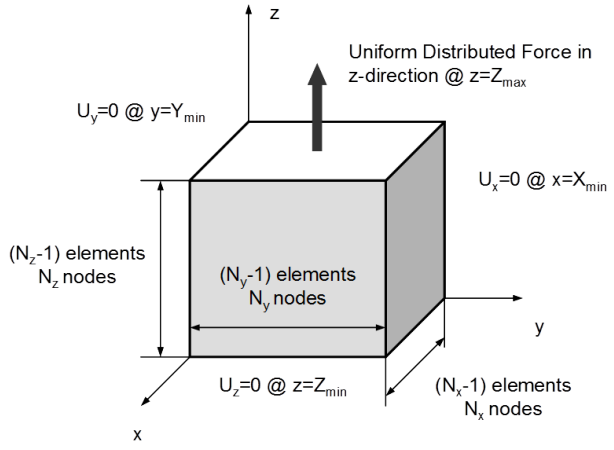

図 4 Cube 型ベンチマークの境界条件

プログラムは全て OpenMP ディレクティヴを含む FORTRAN90 および MPI で記述されている．各ベンチマークプログラムでは，GeoFEM で採用されている局所分散データ構造 [8] を使用しており，マルチカラー法等に基づくリオーダリング手法によりベクトルプロセッサ，SMP，マルチコアプロセッサにおいて高い性能が発揮できるように最適化されている．また，MPI，OpenMP，Hybrid（OpenMP ＋ MPI）の全ての環境で稼動する．

著者らは参考文献 [10] において，3 種類の GeoFEM ベンチマークのうち図 4 に示すような一様な物性を有する単純形状（Cube 型）を対象とした三次元弾性静解析問題について cc-NUMA アーキテクチャを有する HA8000 に対して様々な最適化を試みた．この成果を性能評価用のベンチマークプログラムとして整備したものが GeoFEM-Cube である．

GeoFEM-Cube では，係数行列が対称正定な疎行列となることから，SGS(Symmetric Gauss-Seidel)[10] を前処理手法とし共役勾配法 (Conjugate Gradient，CG) 法によって連立一次方程式を解いている（以下 SGS/CG 法と呼ぶ）．三次元弾性問題では 1 節点あたり 3 つの自由度があるため，これらを 1 つのブロックとして取り扱っている．

連立一次方程式の係数マトリクスの格納法としてオリジナルの GeoFEM ベンチマークでは

(a) CRS(Compressed Row Storage)

(b) DJDS(Descending order Jagged Diagonal Storage)

の 2 種類の方法が準備されているが，GeoFEM-Cube ではスカラープロセッサ向けの CRS 法を使用している．

SGS 前処理では，係数行列 A そのものが前処理行列として利用されるため ILU 分解は実施しないが，前処理における前進後退代入はグローバルなデータ依存性を有するプロセスのため，並列性を抽出するためのリオーダリングが必要である [10]．GeoFEM ベンチマークでは，マルチカラー法（Multicoloring，MC）法，Reverse Cuthill-McKee（RCM）法，更に RCM 法にサイクリックに再番号付けする Cyclic マルチカラー法（cyclic multicoloring，CM）を適用する手法（CM-RCM）の 3 種類が利用可能となっている．

並列プログラミングモデルとしては各コアを独立に扱う Flat MPI と Hybrid 並列プログラミングモデルの両者を扱うことができる．Hybrid については「**Hybrid a×b（HB a×b）**」（a：MPI プロセス当りの OpenMP スレッド数，b：ノード内 MPI プロセス数）という形で，ノード構成に応じてスレッド数，MPI プロセス数を自由に決められるようになっている．

表 3 は，コア当りの問題サイズを 40^3 節点＝3×64,000＝192,000 自由度とした場合の 1 ノードの性能を様々な計算機（Hitachi SR11000/J2（Hitachi SR11K/J2），Hitachi SR16000/M1（Hitachi SR16K/M1），Hitachi HA8000 クラスタシステム（T2K 東大），FX10，「京」）で比較したものである．並列プログラミングモデルとしてはいずれも Flat MPI を使用している．Oakleaf-FX の対ピーク性能比は 6.77%となり，「京」（8.59%）と比較して若干低い．これは「京」（SPARC64 VIIIfx）と Oakleaf-FX（SPARC64 IXfx）では，ノード当りのコア数は 8 から 16 と増えているが，クロック数が 8%弱減少し，コア当りメモリバンド幅も約 25%低下しているためである．Oakleaf-FX と SR16K/M1（Power7）は，Byte/Flop 値はほぼ同じであるが，コア当りキャッシュサイズの大きい SR16K/M1 の方が対ピーク性能比は高い．

3.5 MDTEST ベンチマーク

ローカルファイルシステム，共有ファイルシステム上で MDTEST ベンチマークを実行し，性能評価を行った．MDTEST ベンチマークは Lawrence Livermore National Laboratory (LLNL) の Livermore Computing Center が公開している I/O ベンチマーク [13] であり，メタデータアクセス性能を計測するものである．

MDTEST では多数のプロセスが一斉に共有ファイルシステムにアクセスし，一定の処理を行う時間から共有ファイルシステムのメタデータアクセス性能を測定する．今回の性能評価では以下の条件で計測を行った．

- ファイル・ディレクトリの作成・削除の速度を測定
- 1 プロセスでの実験では上記の操作を 10,000 回ずつ実行，複数プロセスでの実験ではプロセスごとに上記の操作を 5,000 回ずつ実行
- プロセスごとに個別の作業ディレクトリを作成して処理を実行
- 10 回の測定を行い，平均値からアクセス速度を計算

図 5 は 1 プロセスで MDTEST を実行した結果である．

© 2012 Information Processing Society of Japan

表 3 各計算機環境における GeoFEM-Cube 性能評価結果，1 ノード，Flat MPI，コア当り問題サイズ：40^3 節点＝ 3×64,000 ＝ 192,000 自由度，Hitachi SR11000/J2（Hitachi SR11K/J2），Hitachi SR16000/M1（Hitachi SR16K/M1），Hitachi HA8000 クラスタシステム（T2K 東大），Fujitsu PRIMEHPC FX10（Oakleaf-FX），「京」

	Hitachi SR11K/J2	Hitachi SR16K/M1	T2K 東大	Fujitsu FX10 Oakleaf-FX	「京」
Processor	IBM Power5+ 2.3 GHz	IBM Power7 3.83 GHz	AMD Opteron8356 2.3 GHz	SPARC64 IXfx 1.848 GHz	SPARC64 VIIIfx 2.0 GHz
Core #/Node	16	32	16	16	8
Peak Performance (GFLOPS)	147.2	980.5	147.2	236.5	128.0
STREAM Triad (GB/s)	101.0	264.2	20.0	64.7	43.3
Byte/Flop	0.686	0.269	0.136	0.274	0.338
GeoFEM-Cube (GFLOPS)	19.0	72.7	4.69	16.0	11.0
% to Peak	12.9	7.41	3.18	6.77	8.59
Last Level Cache/core (MB)	18.0	4.00	2.00	0.75	0.75

横軸にはアクセス速度 (Operations per second) を，ファイル作成等の各操作の実行回数を 1 秒あたりの値に正規化して示している．ローカルファイルシステムと共有ファイルシステムの間で大きな差はないが，ローカルファイルシステムのほうが若干高い性能となっている．メタデータサーバの構成はどちらのファイルシステムもほぼ同じであるが，ローカルファイルシステムは I/O ノードと実データを置くディスクが直接接続されているため，処理のレイテンシが小さくなっている．このことが性能差の理由であると考えられる．

図 6 は 32 ノードを使用し，ノードあたり 1 プロセスを起動して MDTEST を実行した結果である．横軸にはアクセス速度を，各操作の実行回数を全プロセスで合計した数を 1 秒あたりの値に正規化して示している．こちらもローカルファイルシステムと共有ファイルシステムの間で大きな差はなく，どの操作でも 2 万回/秒以上の値が計測された．

Oakleaf-FX と Yayoi を比較してみると，Yayoi における 1 ノード・8 プロセス使用時のディレクトリ作成・削除，ファイル作成・削除の性能はそれぞれ 5,892 回/秒，8,302 回/秒，7,044 回/秒，5,796 回/秒であった．Oakleaf-FX の共有ファイルシステム，ローカルファイルシステムの性能は Yayoi の半分から 1/4 程度であることがわかる．これは Yayoi で使用されている GPFS が Lustre や FEFS とは異なりメタデータが分散管理されていることが原因であると考えられる．

3.6 IOR ベンチマーク

ローカルファイルシステムと共有ファイルシステム上で IOR ベンチマークを実行し，性能評価を行った．IOR ベンチマークは MDTEST と同様に LLNL の Livermore Computing Center が公開している I/O ベンチマークであり，ブロック入出力のスループットを計測するものである．

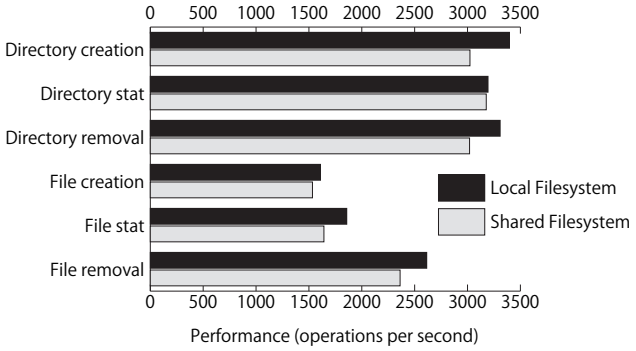

図 5 MDTEST (1 ノード) の実行結果

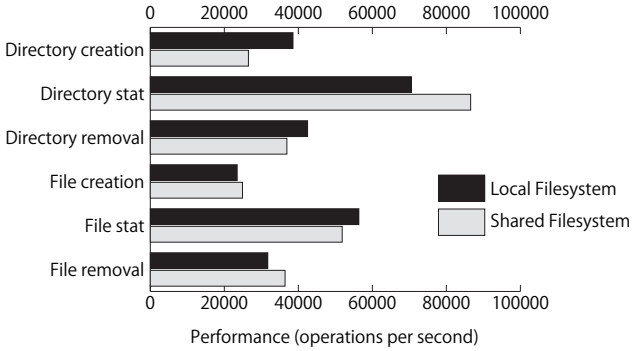

図 6 MDTEST (32 ノード) の実行結果

表 4 IOR の実行結果

	ローカル ファイルシステム	共有 ファイルシステム
1 ノード (MB/sec) (ior-multi)	4,023.40	3,964.92
複数ノード (MB/sec) (ior-multi)	139,008.00	134,734.62
複数ノード (MB/sec) (ior-single)	N/A	80,724.43

IOR では多数のプロセスが一斉に共有ファイルシステム上のファイルを読み書きし，データ転送性能を測定する．

© 2012 Information Processing Society of Japan

使用するファイルのプロセスへの割り当てについては，プロセスごとに別のファイルを割り当てるか，単一ファイル内でプロセスごとに別々の領域に割り当てるかを選択することが可能である．以下では前者を ior-multi，後者を ior-single と呼ぶことにする．今回の性能評価では，両者について以下の条件で計測を行った．

- POSIX I/O を使用
- ファイルの書き込みの性能を測定
- 1 回のシステムコールあたりの書き込みサイズは 1MiB

実行結果を表 4 に示す．1 ノードでの ior-multi の実験は，ノード内で 16 プロセスを起動して実施した．プロセスごとに 256GiB，合計で 4TiB の書き込みを行った．ローカルファイルシステム，共有ファイルシステムともに約 4GB/秒の性能となった．接続されているディスクの書き込み速度または，Tofu インターコネクトの 5GB/秒の転送速度のどちらかがボトルネックになっていると考えられるが，詳細については未調査である．

複数ノードでの ior-multi の実験では，ノードあたり 1 プロセスを起動し，合計書き込みサイズがほぼ 32TiB となるように設定して IOR を実行した．結果は，ローカルファイルシステムは 139GB/秒，共有ファイルシステムは 134GB/秒の性能となった．ノード数は複数のケースを試し，最高の性能だったものを表に示している．ローカルファイルについては，1,200 ノードを使用し，ノードあたりの書き込みサイズを 26.86 GiB (合計 32.2 TiB) とした場合を示している．共有ファイルシステムについては，1,872 ノードを使用し，ノードあたりの書き込みサイズを 17.09 GiB (合計 32.00 TiB) とした場合を示している．複数ノードでの ior-single の実験は，共有ファイルシステムにおいてのみ実施した．1,920 ノードで，ノードあたりの書き込みサイズを 17.09 GiB (合計 32.81 TiB) として実行した結果，80.7 GB/秒の性能であった．単一ファイルに対しての書き込みを行う場合，FEFS や Lustre では複数の OST にファイルをストライピングすることでデータ転送性能を向上させることができる．通常の Lustre 1.8 ではストライプ数の最大値は 160 であるが，FEFS では大規模環境向けの拡張を行っているため，ストライプ数は最大で 20,000 まで設定することができる．今回の実験ではファイルのストライプ数は OST の数と同じ 480，ストライプサイズは 4095MiB として計測を行った．

IOR についても Yayoi と比較してみると，Yayoi の最大性能が約 10GB/秒であるのに対して Oakleaf-FX は転送帯域・ディスク本数共に増加させたため 13 倍程度の性能が得られている．Yayoi では ior-multi と ior-single の間で大きな性能差は確認できなかったが，Oakleaf-FX では 4 割程度の性能低下が確認された．10GB/秒程度の帯域でのテストでも同程度の性能差が確認されており，もともとブロックごとに I/O サーバが分散されている GPFS と，そうではない Lustre の特性の差が表れていると考えられる．

4. おわりに

本稿では Oakleaf-FX(Fujitsu PRIMEHPC FX10) の性能について実システムにおけるベンチマーク測定結果を用いて評価した．本稿にて紹介したように Oakleaf-FX は様々な特徴があり，最大の性能を得るためには様々な最適化が必要である．今後も Oakleaf-FX を用いたプログラムの最適化について研究を推進し，またシステム自体の改善についても進めていく予定である．さらに，最適化やシステム改善の結果を基にしたライブラリやフレームワークの開発などにも取り組む予定である．

謝辞 システムの導入・実験にあたっては富士通株式会社および東京大学情報基盤センターの皆様にご協力いただきました．

参考文献

[1] SR16000 システム（SMP）(Yayoi)，東京大学情報基盤センター http://www.cc.u-tokyo.ac.jp/system/smp/.
[2] 大島聡史, 實本英之, 鴨志田良和, 片桐孝洋, 田浦健次朗, 中島研吾: 大規模 SMP 並列スーパーコンピューター (HITACHI SR16000 モデル M1) の性能評価, 情報処理学会研究報告 (HPC-133) (2012).
[3] スーパーコンピュータ PRIMEHPC FX10，富士通 http://jp.fujitsu.com/solutions/hpc/products/primehpc/.
[4] FX10 スーパーコンピュータシステム（Oakleaf-FX)，東京大学情報基盤センター http://www.cc.u-tokyo.ac.jp/system/fx10/.
[5] HA8000 クラスタシステム（T2K 東大)，東京大学情報基盤センター http://www.cc.u-tokyo.ac.jp/system/ha8000/.
[6] STREAM BENCHMARK http://www.cs.virginia.edu/stream/.
[7] HPC Challenge Benchmark http://icl.cs.utk.edu/hpcc/.
[8] GeoFEM http://geofem.tokyo.rist.or.jp/.
[9] UT-HPC benchmark http://www.cspp.cc.u-tokyo.ac.jp/ut-hpc-benchmark/.
[10] 中島研吾，片桐孝洋：マルチコアプロセッサにおけるリオーダリング付き非構造格子向け前処理付反復法の性能，情報処理学会研究報告（HPC-120-6）(2009)．
[11] Mattson, T.G., Sanders, B.A., Massingill, B.L.: Patterns for Parallel Programming, Software Patterns Series (SPS), Addison-Wesley (2005).
[12] Nakajima, K.: New Strategy for Coarse Grid Solvers in Parallel Multigrid Methods using OpenMP/MPI Hybrid Programming Models, ACM Proceedings of PPoPP/PMAM 2012, New Orleans, LA, USA (2012).
[13] Scalable I/O Benchmark Downloads, Lawrence Livermore National Laboratory https://computing.llnl.gov/?set=code&page=sio_downloads.
[14] TOP500 List - June 2012 (1-100) — TOP500 Supercomputing Sites http://www.top500.org/list/2012/06/100.

© 2012 Information Processing Society of Japan