Vol.2022-HPC-185 No.13
2022/7/28

富岳におけるSPEC HPCの評価

児玉 祐悦[1,a)]　近藤 正章[1]　佐藤 三久[1]

概要：スーパーコンピュータ「富岳」において SPEC HPC ベンチマークの評価を行った。富岳は 2021 年 3 月から供用が開始された日本のフラグシップスーパーコンピュータで、A64FX をプロセッサとした 15 万ノード以上で構成され、倍精度理論最大性能は 488PFLOPS である。SPEC HPC 2021 は、SPEC の最新ベンチマークスイートで HPC 向けの複数のベンチマークからなっており、MPI および OpenMP、さらにアクセラレータを用いた評価が簡単に行える。データセットとして tiny, small, medium, large の 4 つが用意されており、それぞれについて、最小ノード数での評価、ノード数を増加させたときの性能スケーラビリティ、他のシステムとの比較などを行った。さらに、富岳の持つ 3 種類の電力制御機構を組み合わせた電力モードによる評価を行い、ブーストエコリテンション時に、ノーマルから性能を約 2 %向上させつつ、エネルギーを約 17 %削減できることを確認した。

1. はじめに

理化学研究所では、日本における次世代のフラグシップスーパーコンピュータとして富岳を開発し、2021 年 3 月より一般供用を開始した。本稿では、SPEC HPC ベンチマークを用いて富岳の評価を行ったので、その結果について報告する。また、その結果について他のプロセッサとの比較や電力性能などを考察する。

本稿では、まず、富岳およびそのプロセッサである A64FX の概要について述べた後、富岳における電力制御機構について説明する。次に、SPEC HPC ベンチマークについて概要を述べ、富岳の結果について報告する。その後、その結果についての種々の観点からの評価を行い、最後にまとめを述べる。

2. 富岳の概要

富岳は総ノード数 158,976 の大規模なスーパーコンピュータである。432 ラックから構成され、1 ラックに 384 ノードを搭載している。ただし、そのうち 36 ラックは半分の 192 ノードとなっている。各ノードは 6 次元メッシュ／トーラスで接続されており、ユーザからは最大 48 x 69 x 48 の 3 次元トーラスとして指定できる。各ノードは 2.0GHz で動作し、倍精度理論最大性能は 488PFLOPS である。

富岳の各ノードは 1 チップのプロセッサ Fujitsu A64FX[1] から構成される。図 1 にプロセッサの構成を示す。A64FX は Armv8-A の 64 ビットアーキテクチャであり、SVE(Scalable Vector Extension) を実装した最初のプロセッサである。A64FX は計算コアを 48 個持つメニコアプロセッサであり、内部では 4 つの CMG（Core Memory Group）に分かれている。各 CMG は 12 個の計算コア、1 個のアシスタントコア、8MiB の共有 L2 キャッシュ、メモリコントローラから構成されている。

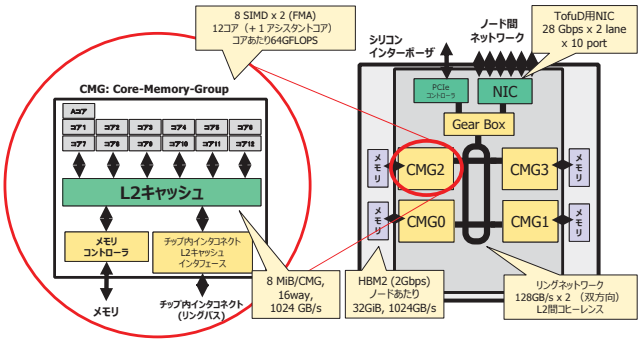

図 1　A64FX の構成

各コアは 512 ビットの SIMD 演算パイプラインを 2 本持っており、通常 2.0GHz で動作し、理論倍精度浮動小数点演算性能は 64GFLOPS である。L1 キャッシュは各コアにデータ・命令各 64KiB あり、ともに 4way で、キャッシュラインサイズは 256 バイトと SVE 向けに長めに設定されている。1 サイクルに最大 64 バイトのロード 2 回あるいはストア 1 回が可能で最大 L1 キャッシュバンド幅は 256GB/s である。L2 キャッシュは CMG 内に 16way で 8MiB あり、CMG 内のコアで共有される。L2 キャッシュの最大バンド幅は 1,024GB/s であるが、各コアの最大 L2 アクセスバンド幅は 128GB/s である。A64FX のマイクロ

[1] 理化学研究所 計算科学研究センター
[a)] yuetsu.kodama@riken.jp

アーキテクチャについては、[2] に詳細が公開されており、また各パラメータのコデザインによる決定については [3] に報告されている。

A64FX はパッケージ内に搭載された HBM2 を 1CMG あたり 1 個、1 チップでは 4 個搭載しており、容量 32GiB、スループット 1,024GB/s のメモリ性能を実現している。4 つの CMG はリングバスで接続されており、CMG 間のキャッシュコヒーレンスが維持される。ノード間ネットワークは Tofu Interconnect D (TofuD) という京と同様の 6 次元メッシュ／トーラス接続で、28Gbps x 2lane x 10ports の物理性能であり、ネットワークコントローラや PCI Express コントローラもチップ内に実装している。

A64FX はシングルチップに 48 コアを搭載したメニコアプロセッサであり、SIMD 幅を広げて性能を向上させつつ、1 コアあたりのチップ面積や消費電力は低く抑えるためアウトオブオーダリソースなどは制限されている。そのような制約から、演算レイテンシや L1 キャッシュレイテンシなどは他のプロセッサより長くなっており、性能を向上させるためには、これらの特徴を考慮したチューニングが必要になっている。一方で、メモリとして HBM2 を採用し、バンド幅は他のプロセッサの数倍を実現しておりメモリバンド幅がボトルネックとなるようなプログラムでは比較的容易に性能を向上させることができる。ただし、容量は 32GiB と少ないため、複数ノードによる並列化が不可欠である。

2.1 A64FX の電力制御機構

電力効率の向上のため、A64FX では、最先端の半導体プロセスとして TSMC 社の N7 を採用するとともに、電力効率を高める回路設計が行われている。その結果、2019 年の Green500 では GPU マシンを抑えて 1 位となった [4]。さらに、次にあげるようないくつかの電力制御機構を備えている。

電力効率を重視して、富岳ではノーマルモードとして周波数は 2.0GHz としている。一方で、少しでも性能を向上させたい、という要求に応えるために、2.2GHz のブーストモードを用意している。しかし、A64FX ではノーマルモード時の電源電圧は可能な限り低く抑えているため、2.2GHz で動作させるためには電源電圧を上げる必要があり、ブーストモードでは電力効率は低下してしまう。

A64FX では、アプリケーション性能を出来るだけ維持しつつ、そのアプリケーションでは使用しない回路の電力を削減することにより、電力効率を上げるという基本コンセプトに基づき、いくつかのパワーノブを有している。その中に、2 本ある浮動小数点演算パイプラインを 1 本に制限するパワーノブ FLAonly がある。このパワーノブを適用するとともに、その際のピーク電力が低くなること見越した電力制御を行い電力削減を実現する機構をエコモードと呼ぶ。

ジョブが割り当てられないノードの電力を削減するために、富岳ではノードリテンションという状態を設定している。これは、1 つのアシスタントコア以外のコアをすべてコアリテンションという状態にする。コアリテンションでは、アイドル時の電力に適した制御を行うことにより、アイドル電力を低く抑えることができる。コアリテンションから通常のアイドル状態へは数ミリ秒で遷移可能である。

これらの電力制御は独立に設定が可能であり、計 8 つの組み合わせが可能である。例えば、ブーストエコはブーストモードとエコモードを組み合わせたモードである。ただし、ブーストモードはノード内の全コアで共通に設定されるのに対して、エコモードとコアリテンションはコアごとに設定が可能である。

富岳のノードには計算ノードと IO ノードがある。IO ノードは計算ノードと同じくユーザジョブを実行するとともに、各ジョブに対するファイル IO のサービスを行う。そのため、計算ノードではアシスタントコアが 2 個有効になっているのに対し、IO ノードでは 4 個有効になっている。さらに電力制御が IO 処理に影響を与えないように、IO ノードではブーストモード固定となっており、現状ではユーザが電力制御を行うことはできない。

さらに、ノードリテンションが全ノードで有効であると、大規模ジョブの実行時に富岳全体の電力変動が大きくなりすぎ、施設の冷却設備への影響が懸念される。そのため、現在は一部のラックへの適用にとどめて、どこまで適用拡大できるかを慎重に見極めている。また、コアリテンションについても、大規模ジョブに適用したときの電力変動について検討中であり、現在は 1 ラック（384 ノード）以下の実行を行うスケジューラのリソースグループに制限されている。

2.2 Power API

上記で述べた電力制御は、富士通のジョブスケジューラのスクリプトでジョブ単位で設定することが可能である。また、Power API を用いてプログラム内で制御することも可能である。この Power API は富士通が開発したライブラリとして富岳、およびその市販バージョンである FX1000 で提供されている。

富岳ではこの Power API を用いて、実測電力 (measured) と推定電力 (ideal) の 2 種類のノード電力を取得できる。実測電力は CMU(Core Memory Unit) とよぶプロセッサボード上の POL(Point of Load) デバイスで測定した電力であり、約 5 ミリ秒間隔で更新される。実測電力には、半導体プロセスのばらつきやノード構成の違いなどによりノード毎にばらつきがある。一方、推定電力はコア内の各ユニットの稼働率などから計算される電力であり、約 1 ミリ秒間隔で更新される。推定電力は、アプリケーションの

表 1　SPEC HPC 2021 ベンチマーク

Name	Application Area	Language	Dataset
lbm	Computational Fluid Dynamics	C	tiny,small,medium,large
soma	Physics / Polymeric Systems	C	tiny,small
tealeaf	Physics, High Energy Physics	C	tiny,small,medium,large
clvleaf	Physics, High Energy Physics	Fortran	tiny,small,medium,large
miniswp	Nuclear Engineering, Radiation Transport	C	tiny,small
pot3d	Solar Physics	Frotran	tiny,small,medium,large
sph_exa	Astropysics and Cosmology	C++14	tiny,small
hpgmgfv	Cosmology, Astrophysics, Combustion	C	tiny,small,medium,large
weather	Weather	Fortran	tiny,small,medium,large

電力評価・チューニングに用いることを想定している。推定電力では半導体プロセスによるばらつきはない。ノード構成によるばらつきをなくすために、実測電力に含まれているアシスタントコア・光モジュール・PCI-e の電力が、推定電力では除かれている。また、実測電力・推定電力ともに、DC 変換後のノードのみの電力であり、ラック内にある PSU による AC/DC 変換ロスやファイルシステム接続用のスイッチ、冷却電力などは含んでいない。推定電力は、ノード電力だけでなく、CMG 単位のコア電力や L2 キャッシュ電力、メモリ電力など電力内訳を取得することもできる。一部の推定電力は PMU(Performance Monitor Unit) を通じて Power API がサポートされていないクラスタ版の FX700 でも利用可能である [5]。

3. SPEC HPC 2021

SPEC（Standard Performance Evaluation Corporation）[6] は、計算機の性能とエネルギー効率を評価するための標準化されたベンチマークとツールを開発・維持するために設立された非営利団体であり、1988 年から活動している。ベンチマークには、SPEC CPU®、SPEC ACCEL®、SPEC MPI®、SPEC OMP®、SPEC Cloud® など、さまざまな種類がある。

SPEC HPC 2021[7] は、これまでの SPEC MPI、SPEC OMP および SPEC ACCEL の開発経験に基づいて、2021 に制定された HPC 向けの新しいベンチマークスイートである。フラット MPI から MPI と OpenMP のハイブリッド、さらにはアクセラレータを用いた実行まで、簡単な指定で実行することができる。他の SPEC ベンチマークと同様にベンチマーク指標としては、ratio と呼ぶリファレンスマシンと比較した相対性能を用いている。SPEC HPC の性能は、全ベンチマークの幾何平均をとる。また、全ベンチマークに共通のコンパイルオプションを適用した値を base, ベンチマーク毎に最適なオプションを適用した値を peak と区別する。本稿では base のみを評価した。SPEC HPC では tiny, small, medium, large の 4 つのデータセットがある。表 1 に示すとおり、tiny と small は 9 個のベンチマーク、medium と large は 6 個のベンチマークから構成されている。リファレンスマシンは Intel Xeon E5-2680v3 (Haswell, 12 cores, 2.5 GHz, Turbo boost off) を 2 基搭載し、メモリが 64GB (DDR4-2133 x 8port, 136 GB/s) を 1 ノードとした並列システムである。各データセットに対し 1 ノード, 10 ノード, 85 ノード, 340 ノードを使用してフラット MPI により実行した性能を 1 としている。

4. 評価

富岳上で、SPEC HPC 2021 をインストールして、評価を行った。ベンチマークのビルドには Fujitsu compiler tcsds-1.2.34 を用いた。具体的なコンパイルオプションは以下の通りである。

```
COPTIMIZE = -Nclang -Ofast -mcpu=a64fx+sve
-ffj-eval-concurrent -fsave-optimization-record
-Nlst=t -Koptmsg=2
CXXOPTIMIZE = -Nclang -Ofast -mcpu=a64fx+sve
-ffj-eval-concurrent -fsave-optimization-record
-Nlst=t -Koptmsg=2
FOPTIMIZE = -Kfast -Kopenmp -Nlst=t -Koptmsg=2
```

また、ポータビリティの指定として以下を追加している。

```
532.sph_exa_t,632.sph_exa_s=default=default:
PORTABILITY += -std=c++14
513.soma_t,613.soma_s=default=default:
PORTABILITY += -DSPEC_NO_VAR_ARRAY_REDUCE
```

富岳では基本的にノーマルモード (2.0GHz) で、1 ノードあたり 4 ランク、つまりランクあたり 1CMG(12thread) で実行している。なお、昨年度の SPEC CPU/OMP の評価 [8] では、一部コンパイラオプションの検証指定の不備により invalid run となっていると報告していたが、今回の実行では flags.xml を正しく修正することにより valid run として実行できている。

表 2 に SPEC HPC 2021 tiny の結果を示す。表には各ベンチマークの名前、Rank 数 x スレッド数、富岳の結果（実行時間と ratio）を示している。Tiny は 60GB のメモリを必要とする。富岳は 1 ノードあたり 32GiB のメモリであるため、1 ノードでは実行できない。2 ノードで実行を試したが、メモリが足りないために実行できなかった。シ

表 2 SPEC HPC 2021 tiny 結果

	Ranks x Thds	Time (s)	Ratio
505.lbm_t	12x12	800	2.81
513.soma_t	12x12	1114	3.32
518.tealeaf_t	12x12	412	4.01
519.clvleaf_t	12x12	141	11.7
521.miniswp_t	12x12	594	2.69
528.pot3d_t	12x12	121	17.5
532.sph_exa_t	12x12	1530	1.27
534.hpgmgfv_t	12x12	464	2.53
535.weather_t	12x12	147	21.9
tiny	12x12	5323	4.84

表 3 SPEC HPC 2021 small 結果

	Ranks x Thds	Time (s)	Ratio
605.lbm_s	80x12	882	1.76
613.soma_s	80x12	632	2.53
618.tealeaf_s	80x12	751	2.73
619.clvleaf_s	80x12	210	7.86
621.miniswp_s	80x12	385	2.86
628.pot3d_s	80x12	151	11.1
632.sph_exa_s	80x12	1413	1.63
634.hpgmgfv_s	80x12	568	1.72
635.weather_s	80x12	171	15.2
small	80x12	5163	3.70

ステムや MPI のバッファなどでユーザが使えるメモリが若干足りなかったものと思われる。本結果は 3 ノードによる実行で、1 ランクあたり 1CMG（12threads）であり、12 ランクでの実行である。Tiny は 9 個のベンチマークからなる。ベンチマークごとの ratio 値は 1.27 から 21.9 と大きく異なっており、幾何平均は 4.84 となっている。基本的にはメモリバンド幅がボトルネックとなるベンチマークでは性能が高く、SIMD 化があまり適用できていないベンチマークでは性能が低くなっていると思われるが、プロファイルなどによる評価は考察で述べる。リファレンスマシンは SPEC HPC の概要で述べたとおり、Intel Xeon Haswell 2 基からなる 1 ノードで、24 ランクのフラット MPI で実行した結果である。富岳では 3 ノードで 4.84 倍の性能であり、まずまずの性能だと言える。スレッド数を変化させた場合の評価や、最新のプロセッサとの比較は考察で述べる。

表 3 に SPEC HPC 2021 small の結果を示す。表の項目は tiny と同じである。Small は 480GB のメモリを必要とするため、富岳では 20 ノードによる実行とした。1 ランクあたり 1CMG（12threads）であり、80 ランクでの実行である。Small は tiny と同じ 9 個のベンチマークからなる。リファレンスマシンは 10 ノードで 240 ランクのフラット MPI で実行した結果である。これは、富岳 20 ノードと同じプロセッサ数である。ベンチマークごとの ratio 値は 1.63 から 15.2 と大きく異なっており、幾何平均は 3.70 となっている。ベンチマーク毎の傾向は tiny と同様である。

表 4 SPEC HPC 2021 medium 結果

	Ranks x Thds	Time (s)	Ratio
705.lbm_m	576x12	1224	1.00
718.tealeaf_m	576x12	550	2.45
719.clvleaf_m	576x12	237	7.81
728.pot3d_m	576x12	213	8.68
734.hpgmgfv_m	576x12	648	1.54
735.weather_m	576x12	186	12.9
medium	576x12	3058	3.86

表 5 SPEC HPC 2021 large 結果

	Ranks x Thds	Time (s)	Ratio
805.lbm_l	2112x12	972	2.80
818.tealeaf_l	2112x12	311	4.66
819.clvleaf_l	2112x12	158	13.3
828.pot3d_l	2112x12	378	12.0
834.hpgmgfv_l	2112x12	751	4.46
835.weather_l	2112x12	174	19.7
large	2112x12	2744	7.53

表 4 に SPEC HPC 2021 medium の結果を示す。表の項目は tiny と同じである。Medium は 4TB のメモリが必要とするため、富岳では 144 ノードによる実行とした。1 ランクあたり 1CMG（12threads）であり、576 ランクでの実行である。Medium は 6 個のベンチマークからなる。リファレンスマシンは 85 ノードで 2040 ランクのフラット MPI で実行した結果である。ベンチマークごとの ratio 値は 1.0 から 12.9 と大きく異なっており、幾何平均は 3.86 となっている。ベンチマーク毎の傾向は tiny, small と同様である。富岳の性能が低かったスケーラビリティが低い 3 つのベンチマークが除かれているが、全体性能としては大きな違いはない。

表 5 に SPEC HPC 2021 large の結果を示す。表の項目は tiny と同じである。Large は 14.5TB のメモリを必要とするため、富岳では 528 ノードによる実行とした。1 ランクあたり 1CMG（12threads）であり、2,112 ランクでの実行である。Large は medium と同じ 6 個のベンチマークからなる。リファレンスマシンは 340 ノードで 8,160 ランクのフラット MPI で実行した結果である。ベンチマークごとの ratio 値は 2.8 から 19.7 と大きく異なっており、幾何平均は 7.53 となっている。ベンチマーク毎の傾向は medium と同様である。

5. 考察

5.1 プロファイリング

各ベンチマークの実行の詳細をプロファイラにより確認した。富岳では、浮動小数点演算数やメモリバンド幅などの一部の性能カウンタについては、ジョブの統計情報として取得することができるが、ジョブ単位の情報であり、SPEC HPC の評価で用いた reportable な実行では複数の

表 6　SPEC HPC 2021 tiny 基本プロファイル情報

benchmark	SPEC ratio	Exec. time (s)	FP peak rate (%)	Memory peak rate (%)	SIMD rate (%)	SVE rate (%)
519.clvleaf_t	11.7	133	6.38	43.77	32.06	91.15
528.pot3d_t	17.5	117	2.30	73.13	43.18	99.93
535.weather_t	21.9	146	6.04	36.67	48.98	100.00
505.lbm_t	2.81	788	4.00	2.05	20.53	60.31
513.soma_t	3.32	111	1.42	1.75	9.36	49.28
518.tealeaf_t	4.01	409	0.94	15.11	0.86	8.72
521.miniswp_t	2.69	533	1.66	1.92	0.57	0.23
532.sph_exa_t	1.27	1540	1.17	0.89	4.64	0.16
534.hpgmgfv_t	2.53	473	1.27	11.02	0.36	0.79

ベンチマークを一つのジョブとして実行するため、個別のベンチマークの性能値を取得することはできない。そのため、個別にベンチマークを実行する必要がある。

また、SPEC ベンチマークの実行では runhpc というツールからベンチマークのビルドや実行環境の設定、実行などを指定したベンチマーク群に対して行う仕組みとなっている。富士通の基本プロファイラは、プログラム全体のプロファイル取得をソフトウェアの修正を行わずに行うことが可能であるが、直接ベンチマークを実行するわけではないため、そのままでは実行できない。そのため、runhpc で実行したログからベンチマークの実行コマンドを抽出して、そのコマンドを再度プロファイラで実行することによりプロファイル情報を取得した。

表 6 に基本プロファイル情報を示す。表には、各ベンチマーク名、SPEC 性能値 (SPEC ratio)、実行時間 (Exec. time)、浮動小数点演算性能の実測値の理論値に対する比率 (FP peak rate)、メモリスループットの実測値の理論値に対する比率 (Memory peak rate)、SIMD 命令数の実行された全命令数に対する比率 (SIMD rate)、実行された SVE 演算数の浮動小数点演算の総数に対する比率 (SVE rate) を示している。実行は tiny を 3 ノードで実行したうちのランク 0 の 12 スレッド分の情報である。メモリスループットは CMG あたりの合計、それ以外は各コアの値の平均である。上の段に SPEC のスコアが 10 を超えている 3 つのベンチマークを、下の段にそれ以下の 6 つのベンチマークを並び替えて示している。

上の段のベンチマークではメモリバンド幅ピーク比が 35 %を超えており、富岳のメモリ性能の向上が性能向上に寄与していることがわかる。また、これらのベンチマークでは SIMD 命令の割合が 30 %を超えており、浮動小数点演算の SVE 化率は 90 %を超えており、SIMD 化の寄与も高いことがわかる。この 3 つのプログラムは Fotran で記述されており、利用した富士通コンパイラでは指定した最適化オプションでソフトウェアパイプラインが適用されることも一因であると思われる。それでも FPU 利用率は 10 %未満であり、演算やキャッシュレイテンシの隠蔽による性能向上に、まだ改良の余地があることがわかる。

一方で、下の段のベンチマークではメモリバンド幅ピーク比が約 15 %以下で、SIMD 化率は約 20 %以下、浮動小数点演算の SVE 化率も 60 %以下となっている。これらが性能が芳しくない理由であると思われる。また、これらは C で記述されており、指定した Clang オプションではソフトウェアパイプラインが適用されないことも、性能が抑えられている一因と考えられる。Trad モードによるソフトウェアパイプラインの適用については、次のコンパイラオプションの節で考察する。

5.2　コンパイラオプション

本評価で用いている富士通コンパイラの C および C++ には、Trad モードと Clang モードという 2 つのモードがある。Trad モードは、これまでの富士通コンパイラとの互換性が高く、ソフトウェアパイプラインなどの独自の最適化オプションがサポートされている。Clang モードは、オープンソースの LLVM コンパイラのフロントエンドを用いて、他のコンパイラとの互換性が高く、オープンソースのプログラムのコンパイルに向いている。このため、先の評価では Clang モードを用いていた。

SPEC HPC 2021 を Trad モードでビルドしてみたところ、ビルドは特に問題なく完了した。実行結果を表 7 に示す。ただし、Fortarn に関しては違いはないので、C および C++で記述されている 6 つのベンチマークのみ示している。実行は 4 ノード 16 ランクで行い、Clang モードと Trad モードを比較している。Trad モードでは-Kfast -Kresp=all をオプションとして指定している。Trad モードではポインタの参照先が衝突するかどうかの判定が他のコンパイラと比較して厳しく、他のコンパイラでは SIMD 化されるループなどでも SIMD 化がされないことがある。それを避けるためのオプションが-Kresp=all である。

表 7 に示すとおり、4 つのベンチマークでは性能が向上し、特に 518.tealeaf では 4 倍以上の性能向上が確認された。一方で、2 つのベンチマークでは性能が低下し、特に 521.miniswp では性能が 20 %まで低下している。この原因

表 7　富士通コンパイラでの Clang mode と Trad mode の比較

	Clang	Trad	Trad/Clang
505.lbm_t	3.78	5.60	1.48
513.soma_t	4.37	3.57	0.82
518.tealeaf_t	5.10	22.0	4.31
521.miniswp_t	3.53	0.71	0.20
532.sph_exa_t	1.65	2.14	1.30
534.hpgmgfv_t	2.55	5.95	2.33
tiny	6.16	7.01	1.14

を調べたところ、本ベンチマークでは OpenMP 4.0 の機能である task directive の depend 節が使われていた。Trad モードではこの機能はサポート外であるが、デフォルトでリンクされる LLVM 互換の omplib と組み合わせると、コンパイラは正常に行われるが、性能が大きく低下してしまうことがわかった。しかし、サポート外の機能のため、改善の見込みはない。

Fortran で記述されている 3 つのベンチマークも含めた幾何平均では、Trad モードの方がやや良いという結果であるが、その差はそれほど大きくはないため、評価としては Clang モードのままとしている。SPEC ベンチマークには、base というすべてのベンチマークで同じオプションを指定する指標の他、 peak とよぶベンチマークごとに最適なオプションを選択できる指標がある。peak の評価として、ベンチマークごとに Clang モードと Trad モードを適切に選択する最適化は、他の最適化オプションのチューニングを含めて今後の課題である。また、最新の Clang モードではソフトウェアパイプラインのオプションも追加されており、こちらの評価も今後の課題である。

5.3　スケーラビリティ

評価では最小のノード数での実行を行ったが、各サイズにおいてスレッド数やノード数の違いに対する性能のスケーラビリティを評価した。

まず最初に、 tiny データサイズを用いて、スレッド数に対する性能への影響を評価した。富岳で、ランク数やスレッド数を変化させて評価した結果が表 8 である。ここではランク数 x スレッド数として、3x48, 12x12, 24x6, 24x1 の 4 通りを比較している。12x12 の結果は表 2 と同じである。すべて 3 ノードでの実行であり、最初の 3 つは 144 個の全コアを利用しているのに対し、最後の設定は 24 コアとリファレンスマシンと同じコア数での実行としている。リファレンスマシンは SPEC HPC の概要で述べたとおり、Intel Xeon Haswell 2 基からなるノードで、1 ノード 24 ランクでフラット MPI で実行している。富岳では 1CMG あたり 2 ランクを割り当て、各ランクが 1 スレッドで実行している。

表 8 に示すとおり、12x12 と比較して、3x48 では性能が 2 割ほど低下しており、スレッド数 48 では性能が低下す

表 8　SPEC HPC 2021 tiny ランク数 x スレッド数の比較

Ranks x Thds	3 x 48	12 x 12	24 x 6	24 x 1
Ratio	3.89	4.84	4.85	1.07

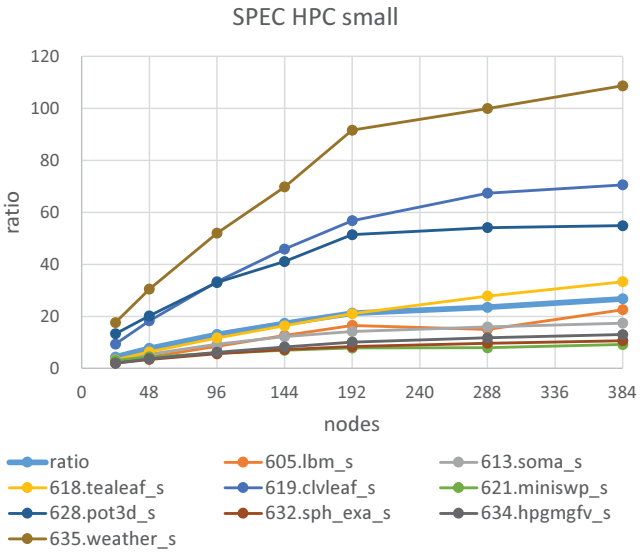

図 2　SPEC HPC small のノード数に対するスケーラビリティ

ることがわかった。これは A64FX では 48 コアが 4 つの CMG に分割されており、他の CMG 上のメモリアクセスは L2 キャッシュコヒーレンスによるオーバヘッドが入ることも影響していると思われる。一方、24x6 では性能はほぼ同じであり、スレッド数 12 は十分にスレッド性能が出ていると判断し、1 ランク 12 スレッドを基本とすることした。また、24x1 では性能が 22 %に低下し、ほぼリファレンスマシンと同じ性能となっている。これよりコアの性能はほぼ Haswell と同じであると見積もられる。

次にノード数を変化させたときのスケーラビリティを評価した。

図 2 は small データサイズでノード数を 24 ノードから 384 ノードに増やしたときの各ベンチマークの性能と全体の性能（ratio）を示している。SPEC HPC では最低 2 回の実行が必要で、その lower median を各ベンチマークの性能として採用することになっている。本評価では、問題サイズ一定でノード数を増加させているため、実行時間が徐々に短くなっていき、実行時間のばらつきが見えてくる。そのため、本評価では各ベンチマークの実行を 5 回に増やしている。また、選択したノード数は 12 ノードの倍数としている。これは富岳のネットワークである Tofu では 12 ノードが単位となっているためである。富岳ではノード間接続が 3D トーラスという直接網であり、ノード割当方法によりネットワーク通信性能が異なってくることが考えられる。富岳では 384 ノードまでの実行を行うスケジューラの small リソースグループでは、基本は 1 ノード単位でノード割当が行われるが、本評価ではノード指定を 2x3x2:torus のように指定することにより、Tofu 単位でノード割当が行

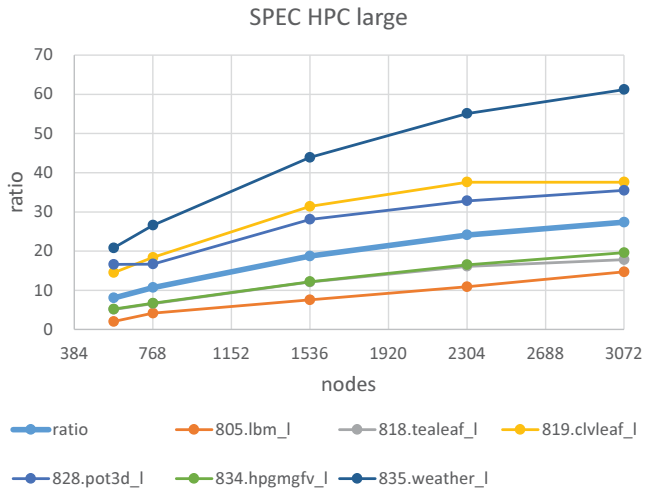

図 3 SPEC HPC large のノード数に対するスケーラビリティ

われるようにした。しかし、どのような形状が最適かは不明であり、その評価は peak 指標の最適化とともに今後の課題である。

SPEC HPC の性能は 192 ノードで 21.2 であり、24 ノードの性能 4.36 に対して、ノード数 8 倍で 4.9 倍の性能であり、スケーラビリティは 61 %となっている。しかし、その後はかなり性能向上は鈍化している。

各ベンチマークは大きく 3 つのグループに分けられる。最初のグループは、性能の良い 635.weather, 619.clvleaf, 628.pot3d の 3 つのベンチマークで、ノード数を増やしたときの性能の伸びも良く、ノード数 192 で 50 を超える性能を示している。ただし、いずれも 192 ノードくらいで性能向上の伸びは落ち始めている。2 番目のグループは、618.tealeafs, 605.lbms, 613.soma の 3 つのベンチマークで、24 ノードのときの性能はそれほど高くないが、ノード数を増やしたときには着実に性能は向上している。3 番目のグループは、621.miniswp, 632.sph_exa, 634.hpgmgfv の 3 つのベンチマークで 4 ノードでの性能も低く、ノード数に対する性能向上も小さい。

図 3 は large データサイズでノード数を 576 ノードから 3,072 ノードに増やしたときの各ベンチマークの性能と全体の性能（ratio）を示している。SPEC HPC の性能は 3,072 ノードで 27.3 であり、768 ノードの性能 10.7 に対して、ノード数 4 倍で 2.5 倍の性能であり、スケーラビリティは 63 %となっている。

5.4 他のシステムとの比較

SPEC HPC のホームページには多くの結果が登録されている。ここでは富岳の結果をこれらと比較して評価を行う。

tiny データサイズで最も性能が高い結果であるテキサス先端計算センター（TACC) の Frontera と富岳を比較したのが表 9 である。比較対象の詳細は [9] を参照。プロセッサは Cascade Lake 世代の Xeon で 28 コアのプロセッサを各ノードが 2 個搭載しており、32 ノードで 64 rank x 28 threads での実行結果である。ランク数とスレッド数から Hyper Thread はオフであると思われる。周波数は 2.7GHz で Turbo Boost は最大 4.0GHz であるが、実行時にどれくらいの周波数で実行されていたのかは不明である。富岳では接続網である Tofu が 12 ノードを単位としているため、比較的近いノード数ということで、36 ノードと比較している。SPEC HPC 全体では、Xeon32 ノードで 78.3 に対して、富岳 36 ノードで 38.4 と半分程度の性能となっている。ちなみに、プロセッサ数が近い富岳 72 ノードで比較すると、富岳は 57.1 で 73 %という結果である。ノード数を 2 倍にしても性能は 1.5 倍くらいにしかなっていないため、Xeon に比べると性能がやや低いという結果である。一方、各ベンチマークの性能を見ると、3 つのベンチマークでは富岳のほうが性能が高い。これはメモリバンド幅の効果であると考えられる。

表 9 SPEC HPC 2021 Tiny 比較

	Frontera	Fugaku	(%)
Processor	Intel Xeon Platinum 8280 x 2	Fujitsu A64FX	
nodes	32	36	
Ranks x Thds	64 x 28	144 x 12	96.4
Frequency (GHz)	2.7 (Turbo up to 4.0)	2.0	74.1 (50.0)
Memory /node (GB/s)	280 (PC4-2933 x 12	1,024 (HBM2 x 4)	365.7
505.lbm_t	104	31.2	30.0
513.soma_t	99.6	36.0	36.1
518.tealeaf_t	117	38.5	32.9
519.clvleaf_t	59.6	98.6	165.4
521.miniswp_t	51.0	14.6	28.6
528.pot3d_t	96.2	105	110.3
532.sph_exa_t	61.9	11.5	18.6
534.hpgmgfv_t	37.9	17.4	45.9
535.weather_t	135	138	102.2
tiny base	78.3	38.4	49.0

富岳を他の計算機と比較する際に、同じノード数で比較することが妥当であるかは議論の余地があるところである。同じ消費電力での性能比較を行いたいが、SPEC に登録されているデータには電力情報がないため、別途自分たちで測定することが必要であり、今後の課題である。

SPEC HPC のホームページに登録されている small サイズでは大規模なノードでの結果はほとんどなく、MPI と OpenMP のハイブリッドタイプでは Lenobo の ThinkSystem SR665（AMD EPYC 7763 x2）を 6 ノード使用して 96 rank x 8 threads の結果が載っている。詳細は [10] を参照。これに対して、富岳で 24 ノード使用して 96 rank x 12 threads で実行して、比較した結果が表 10 である。rank 数

表 10 SPEC HPC 2021 Small 比較

	ThinkSystem SR665	Fugaku	(%)
Processor	AMD EPYC 7763 x 2	Fujitsu A64FX	
nodes	6	24	
Ranks x Thds	96 x 8	96 x 12	150.0
Frequency (GHz)	2.45 (Boost up to 3.5)	2.0	81.6 (57.1)
Memory /node (GB/s)	409 (PC4-3200 x 16	1,024 (HBM2 x 4)	250.4
605.lbm_s	4.37	2.20	50.3
613.soma_s	5.65	2.99	52.9
618.tealeaf_s	2.06	3.26	158.3
619.clvleaf_s	1.74	9.34	536.8
621.miniswp_s	4.44	3.08	69.4
628.pot3d_s	1.64	13.3	811.0
632.sph_exa_s	7.68	1.95	25.4
634.hpgmgfv_s	1.86	2.02	108.6
635.weather_s	9.43	17.6	186.6
small base	3.54	4.36	123.2

表 11 SPEC HPC 2021 Large 比較

	Frontera	Fugaku	(%)
Processor	Intel Xeon Platinum 8280 x 2	Fujitsu A64FX	
nodes	512	768	
Ranks x Thds	1,024 x 27	3,072 x 12	133.3
805.lbm_l	9.89	4.16	42.1
818.tealeaf_l	6.56	6.62	99.3
819.clvleaf_l	6.85	18.4	268.6
828.pot3d_l	6.47	16.7	258.1
834.hpgmgfv_l	10.9	6.70	61.5
835.weather_l	11.8	26.6	225.4
large base	8.47	10.7	126.4

は同じで、スレッド数が富岳のほうが 50 %多く、性能は 23 %富岳が高い。9 個のベンチマークのうち、EPYC の性能が高いのが 4 個で、最大 4 倍高速である。一方、富岳の性能が高いのが 5 個で最大 8 倍となっている。EPYC は高速な DDR メモリを 16 ポート備えるが、それでも富岳の方が倍以上高いメモリバンド幅となっていることが、主な違いであると思われる。

SPEC HPC のホームページに large サイズの結果として登録されているテキサス先端計算センター（TACC) の Frontera と富岳を比較したのが表 11 である。比較対象の詳細は [11] を参照。Tiny の比較を行ったノードと同じであるが、こちらは 1 rank 27 スレッドであり、512 ノードの実行で 8.47 である。富岳では 768 ノードを用いた 3,072 rank x 12 threads の実行で 10.7 と 26 %高い。スレッドあたりの性能としては 95 %の性能となる。6 個のベンチマークのうち、Xeon の性能が高いのが 2 個で、最大 2.4 倍である。ほぼ性能が同じベンチマークが 1 個、富岳の性能が高いのが 3 個で最大 2.7 倍となっている。Tiny と比較して富岳のスレッドあたりの性能比が向上したのは、tiny から large で減ったベンチマークが富岳では性能が低かったベンチマークであることによると思われる。ただし、Frontera では 2,048 ノードとノード数を 4 倍使用したときに 3.7 倍の性能を出しているのに対して、富岳では 2.6 倍にとどまり、差が開いている。これはベンチマークがネットワークトポロジーに対して最適化されていないため、InfiniBand を用いた間接網である Frontera の方が性能スケーラビリティが高いためと思われる。

5.5 電力モード

Large データセットを 1,728 ノードで実行する際に、異なる電力モードを設定して評価を行った。実行に用いたスケジューラのリソースグループでは、コアリテンションは禁止されているため、ノーマル、ブースト、エコ、ブーストエコの 4 つのみを比較した。コアリテンションの評価は、以下で別途行う。表 12 に 4 つの電力モードの結果と、ノーマルモードに対する相対値を示す。表には、各ベンチマークを 5 回実行した実行時間の合計、推定 (ideal) と実測 (measured) の 2 種類の消費エネルギー、それから計算される 2 種類の電力をベンチマークスイート全体の値として示すとともに、各ベンチマークの性能（5 回の実行の中央値）とベンチマーク全体の性能を示す。

ブーストモードでは、周波数が 10 %向上しており、SPEC HPC 全体の性能も 6.2 %向上している。ただし、電力は実測電力で 10.7 %増大し、エネルギーも 3.0 %増大している。ブーストモードでは性能は向上するが、エネルギー効率は低下することが確認できた。実測電力ではノード間のばらつきがあるため、他のノードの実行と比較することはできないが、この評価は 1 つのジョブとして実行して同一のノードを利用しており、また、ノード数も 1,728 ノードと十分に大きいため、相対的な比較には意味があると考えている。

エコモードでは、全体で 5.5 %ほどの性能低下となり、819.clvleaf の 13.3 %が最大の性能低下となっている。しかし、実測電力としては 21.8 %削減されており、実測エネルギーとしても 17.9 %削減されている。特に 828.pot3d では性能低下が 2.1 %と小さいため、このようなプログラムでは積極的にエコモードを使ってもらいたい。

ブーストエコモードでは、全体の性能が 1.7 %向上し、実測電力は 14.7 %、実測エネルギーでも 17.1 %削減されている。819.clvleaf では 7.3 %性能が低下してしまっているが、その他のベンチマークではノーマルより性能向上しており、ブーストとエコを組み合わせることにより、より多くのプログラムでエコモードによる電力削減の効果を得

表 12　電力モードの評価

	Absolute Value				relative to normal			
	normal	boost	eco	boosteco	normal	boost	eco	boosteco
Time (5runs)	5155.5	4795.3	5411.6	5008.1	100.0%	93.0%	105.0%	97.1%
Energy (ideal) kWh	306.7	306.6	268.1	265.9	100.0%	100.0%	87.4%	86.7%
Energy (measured) kWh	304.2	313.2	249.6	252.1	100.0%	103.0%	82.1%	82.9%
Power (ideal) kW	214.1	230.2	178.3	191.1	100.0%	107.5%	83.3%	89.3%
Power (measured) kW	212.4	235.1	166.1	181.2	100.0%	110.7%	78.2%	85.3%
805.lbm_l	8.29	9.08	7.74	8.47	100.0%	109.5%	93.4%	102.2%
818.tealeaf_l	13.1	14.0	12.7	13.8	100.0%	106.9%	96.9%	105.3%
819.clvleaf_l	31.6	32.9	27.4	29.3	100.0%	104.1%	86.7%	92.7%
828.pot3d_l	28.8	29.8	28.2	29.3	100.0%	103.5%	97.9%	101.7%
834.hpgmgfv_l	12.5	13.5	12.0	13.1	100.0%	108.0%	96.0%	104.8%
835.weather_l	45.3	47.6	43.6	47.1	100.0%	105.1%	96.2%	104.0%
spec hpc ratio	19.6	20.8	18.5	19.9	100.0%	106.2%	94.5%	101.7%

ることができることを確認した。

一方、ブースト時の推定エネルギー (ideal) はノーマル時と同じという結果になっている。そこで、各ノードのエネルギーのヒストグラムを確認したのが、図 4 である。ただし、この結果は表 12 とは別の実行結果であり、各ベンチマークは 1 回のみ実行している。図では推定エネルギー (e.E) を棒グラフで、実測エネルギー (m.E) を折れ線で示している。実測エネルギーは広くばらつきがあるのに対し、推定エネルギーはばらつきが抑えられていることがわかる。

例えばエコの実測エネルギー (eco m.E) の場合、26Wh と 31Wh あたりにピークがあり、その他 47Wh や 52Wh にも小さなピークが見える。1 つの山の分布は半導体プロセスのばらつきによるものであり、各ピークはノードの特性によるものである。エネルギーが小さいピークの 2 つは計算ノードであり、エネルギーの大きいピークの 2 つは IO ノードである。計算ノードの違いはアクティブ光ケーブル (AOC) のエネルギーを含むか含まないかである。

ノーマルの推定エネルギー (normal e.E) の場合、39Wh あたりに集中しているが、他に 43Wh あたりにもある。これはブースト固定の IO ノードの推定エネルギーである。ジョブの推定エネルギーはこれらの合計となるため、計算ノードだけよりも高めになってしまう。一方、ブーストではほぼすべてのノードが 40Wh あたりに集まっている。計算ノードだけを比較すると、ブーストの推定エネルギーはノーマルよりも大きいことがわかる。ただし、推定エネルギーは実測エネルギーの一番エネルギーが低いピークに近い値となることが望ましいが、それよりはやや高い値に見積もられている。また、ノーマルとブーストの差も、推定エネルギーでは実測エネルギーよりは小さめに見積もられている。しかし、推定エネルギーの精度は、相対評価には十分であると考えている。

次に、small データセットを 24 ノードで実行して、コアリテンションの評価を行った。実行では、ランク数を 96 に固定して、スレッド数を 12 と 8 に変化させて評価を行った。各結果を 12 スレッドをコアリテンションありで実行した結果との相対値として示したのが、表 13 である。電力モードはノーマルである。

12thread 実行のコアリテンションなしでは、性能が 0.3 % 向上しているが、電力は 2.9 %増大している。逆に言うと、コアリテンションを設定することにより、ほぼ性能影響無しで 3 %ほど電力を削減できると言える。8thread 実行では、12thread 実行と比べてスレッド数が減るため、性能が約 25 %低下している。一方、電力を見ると、コアリテンションありでは 8thread の電力が 12thread よりも 21.3 % 削減されているのに対し、コアリテンションなしでは 3.6 % に過ぎない。これは、使用されていない 4thread 分のコアの電力がコアリテンションにより削減されている効果であり、使用されていないコアがあるときのコアリテンションの効果を示している。SPEC HPC 全体では性能が低下しているため、エネルギーで見ると 10 %増大してしまっているが、628.pot3d では、性能低下が 1.5 %程度なので、エネルギー的にも削減効果が期待できる。このように、メモリバンド幅がボトルネックとなっているようなベンチマークでは、コア数を少なくしても性能が維持でき、コアリテンションを適用すると、さらなる電力削減を実現できることが確認できた。

6. まとめ

スーパーコンピュータ「富岳」において SPEC HPC ベンチマークの評価を行った。SPEC HPC にはデータセットとして tiny, small, medium, large の 4 つが用意されており、それぞれについて、最小ノード数での評価、ノード数を増加させたときの性能スケーラビリティ、他のシステムとの比較などを行った。さらに、富岳の持つ電力制御機構を組み合わせた電力モードによる評価を行い、ブースト

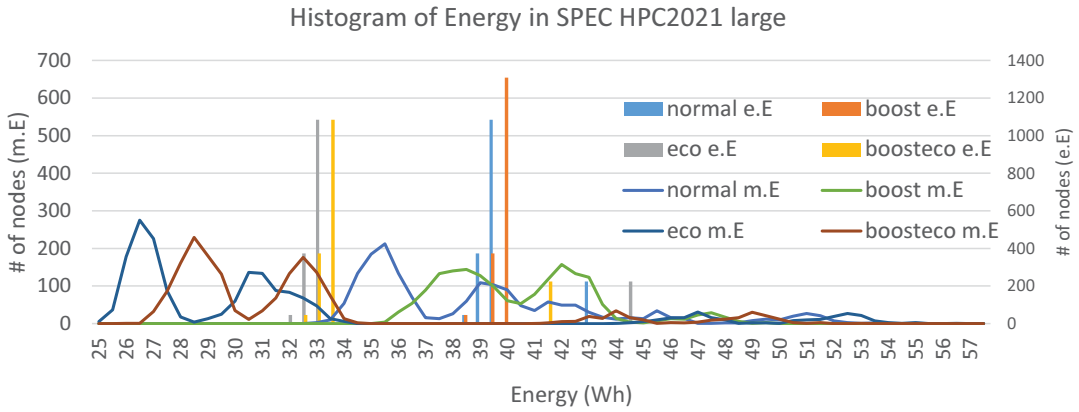

図 4 SPEC HPC large のノード消費エネルギーのヒストグラム

表 13 コアリテンションの評価

core retention	on	off	on	off
thread/rank	12	12	8	8
time (s)	100.0	99.7	140.9	91.8
Energy(ideal) (MWh)	100.0	102.5	110.9	139.5
Power(ideal) (W)	100.0	102.9	78.7	99.2
605.lbm_s	100.0	100.0	67.3	109.2
613.soma_s	100.0	100.7	68.7	68.7
618.tealeaf_s	100.0	100.3	68.7	68.7
619.clvleaf_s	100.0	101.0	78.8	79.4
621.miniswp_s	100.0	100.3	71.4	71.4
628.pot3d_s	100.0	102.3	98.5	98.5
632.sph_exa_s	100.0	99.5	70.9	70.9
634.hpgmgfv_s	100.0	100.0	71.8	72.3
635.weather_s	100.0	100.6	75.6	76.1
small base	100.0	100.5	74.1	74.3

エコ時に、ノーマルからの性能を約 2 %向上させつつ、エネルギーを約 17 %削減できることを確認した。

一方、課題としては、ベンチマークごとに性能差が大きく、要求メモリバンド幅が高いベンチマークでは性能が高いが、SIMD 化率が低いベンチマークでは性能が低いことが確認された。SIMD 化率を向上させることが性能向上に一番有効であると思われ、コンパイラの最適化の改良に期待したい。本評価では全ベンチマークで共通の最適化オプションを用いる base 指標を求めたが、この他にベンチマークごとに最適化オプションを選択できる peak 指標もある。今後は、個別のベンチマークに有効な最適化オプションの調査を行っていきたい。また、ノード数が多くなったときの富岳のスケーラビリティは、他のシステムに比べるとやや悪い。これは、富岳で採用している Tofu ネットワークは直接網であることが影響していると思われる。直接網は、アプリケーション側でネットワーク上のプロセス位置を最適に配置することにより通信のオーバヘッドを軽減することが可能であるが、汎用のベンチマークではそのような最適化を期待することが難しい。完全自動で最適化することは難しいと思われるが、ユーザの最適化の手間を軽減するツールが期待される。

ベンチマークの評価は、評価それ自体だけが目的ではなく、評価に関連して確認された最適化手法などを広く他のアプリケーションにも適用可能な方法を検討することも重要な目的である。今後は、そのような最適化手法、特にエネルギー削減を実現する手法について検討を行っていきたい。

参考文献

[1] Y. Yoshida, *Fujitsu High Performance CPU for the Post-K Computer*, 2018 IEEE Hot Chips 30 Symposium, 2.13, Aug. 2018.
[2] https://github.com/fujitsu/A64FX/, *A64FX Microarchitecture Manual*
[3] M. Sato, Y. Kodama, M. Tsuji and T. Odajima, "Co-Design and System for the Supercomputer "Fugaku"" in IEEE Micro, vol. 42, no. 02, pp. 26-34, 2022. doi: 10.1109/MM.2021.3136882
[4] http://top500.org/green500/lists/2019/11, *Green500 Nov. 2019*
[5] E. Arima, Y. Kodama, T. Odajima, M. Tsuji and M. Sato, "Power/Performance/Area Evaluations for Next-Generation HPC Processors using the A64FX Chip," 2021 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS), 2021, pp. 1-6.
[6] https://www.spec.org/, *SPEC: Standard Performance Evaluation Corporation*
[7] B. Holger et al., "First Experiences in Performance Benchmarking with the New SPEChpc 2021 Suites", https://doi.org/10.48550/arxiv.2203.06751, Mar. 2022.
[8] 児玉, 近藤, 佐藤, "A64FX における SPEC CPU および SPEC OMP の評価", 研究報告ハイパフォーマンスコンピューティング（HPC）, 2021-HPC-180(13), 1-9, July. 2021.
[9] https://www.spec.org/hpc2021/results/res2021q4/hpc2021-20210918-00058.txt, *SPEC HPC 2021 Tiny Result Dell PowerEdge C6420*
[10] https://www.spec.org/hpc2021/results/res2021q4/hpc2021-20210908-00028.txt, *SPEC HPC 2021 Small Result Lenovo ThinkSystem SR665*
[11] https://www.spec.org/hpc2021/results/res2021q4/hpc2021-20210919-00065.txt, *SPEC HPC 2021 Large Result Dell PowerEdge C6420*