

「京」の後の時代を支えるスパコン

② 次期フラッグシップスーパーコン ピュータの概要—スーパーコン ピュータ「富岳」



石川 裕 | 理化学研究所計算科学研究センター

新庄直樹 | 富士通（株）

佐藤三久 | 理化学研究所計算科学研究センター

清水俊幸 | 富士通（株）

概要

文部科学省が行った2012年からの2年間プロジェクトである「将来のHPCI^{☆1}システムのあり方の調査研究」の成果を基に2014年からフラッグシップ2020プロジェクト（FLAGSHIP2020 Project）が開始された。本プロジェクトは、(1) スーパーコンピュータ（以下、スパコン）「京」の後継機の開発・整備、および(2) 開発した後継機を用いて重点的に取り組むべき社会的・科学的課題（重点課題）に向けたアプリケーションの開発から構成され、2021年頃に後継機の共用^{☆2}が開始する予定である。プロジェクト開始時、後継機の名称は決まっていなかったため、ポスト「京」という名前を使用していた。ポスト「京」の開発においては、世界最高水準の消費電力性能、計算能力、ユーザの利便・使い勝手の良さ、画期的な成果の創出というシステムの特色を兼ね備えた、総合力のあるスパコンを実現すべく、(1) システムの消費電力を30～40MWとし、(2) いくつかの実アプリケーションにおいて、「京」の100倍の性能を達成できること、をシステムの性能目標とした。なお、「京」の消費電力は2011年11月TOP500リストのベン

チマーク計測において約12MWだった。

ポスト「京」の開発は理化学研究所（以下、理研）に委託され、理研は開発企業を公募し、富士通（株）（以下、富士通）を開発パートナーに選び、また、文部科学省が選定した9つの重点課題実施機関とともにハードウェアとソフトウェアのコードデザイン（協調設計）を進めた。コードデザインにおいては、各重点課題実施機関からアプリケーションを選定してもらい、それらターゲットアプリケーションの実行効率を高めるべく、ハードウェア、ソフトウェアの設計を進めた。ターゲットアプリケーションのリストを表-1に示す。

2018年夏には富士通において試作機が完成しハードウェア試験が実施され、さらに、現在、システムソフトウェアの開発・試験が進んでいる。試作機によるターゲットアプリケーションの性能予測で

表-1 ターゲットアプリケーション

	名前	概要
①	GENESIS	タンパク質構造予測
②	Genomon	ゲノム解析
③	GAMERA	地震シミュレーション
④	NICAM+LETKF	ビッグデータとシミュレーションによる気象予測
⑤	NTChem	第一原理量子化学計算
⑥	ADVENTURE	大規模計算力学システム
⑦	RSDFT	第一原理計算電子状態計算
⑧	FrontFlow/blue	Large-Eddy Simulation（乱流シミュレーション）
⑨	LQCD	Lattice QCD（量子色力学）

☆1 High Performance Computing Infrastructure の略。

☆2 「特定先端大型研究施設の共用の促進に関する法律」に基づき設置・運用されます。

は、ターゲットアプリケーションのうち2つのアプリケーションが「京」の100倍以上の性能を達成する見込みである¹⁾。また消費電力に関しても目標値を達成する見込みである。ユーザの利便・使い勝手の良さについては、「京」と互換性のあるOpenMP^{☆3}によるノード内スレッド並列とMPI^{☆4}によるノード間並列プログラミングモデルを踏襲していること、また、標準的Linux Distributionの利用、オープンソース管理ツールの採用、仮想マシン、システムが提供する実行環境以外のユーザが望む実行環境を利用可能とするコンテナ技術などを導入することによりユーザの利便性を上げている。2019年5月、名称公募の末、スーパーコンピュータ「富岳」という名称が正式に決まった。「富岳」は富士山の異名で、富士山の高さが「富岳」の性能の高さを表し、また富士山の裾野の広がり「富岳」のユーザの広がりを意味する。本稿では、「富岳」のハードウェアおよびソフトウェアの概要を紹介する。

ハードウェア

プロセッサ

富岳のプロセッサアーキテクチャは、本稿冒頭で述べた「将来のHPCIシステムのあり方の調査研究」の中の「レイテンシコアの高度化・高効率化による将来のHPCIシステムに関する調査研究」における成果²⁾が基になっている。この調査研究において、

表-2 A64FX 諸元

命令セット アーキテクチャ	Armv8.2-A SVE (512bit)
計算コア数と性能	48 + 2/4 アシスタントコア 倍精度: 2.7 TF 以上, 単精度: 5.4 TF 以上, 半精度: 10.8 TF 以上
キャッシュ・メモリ	L1D: 64 KiB, 4way, 230 GB/s (load), 115 GB/s (store) L2: 8 MiB, 16way, 115 GB/s (load), 57 GB/s (store)
主記憶	HBM2 容量 32 GiB, バンド幅 1024 GB/s
インターコネクト	Tofu Interconnect D (28 Gbps × 2 lane × 10 port)
I/O	PCI-Express Gen3 x16
半導体技術	7nm FinFET

☆3 コンパイラ指示文によりスレッド並列化する標準規格。

☆4 通信ライブラリ標準規格。

低電力型、高周波数型、高IPC (Instruction Per Clock) 型、行列積加速型を含む8つのタイプのアーキテクチャを検討した。SIMD (Single Instruction Multiple Data) 演算幅、演算回路数、キャッシュ容量、コア数、コア接続方式など基本要素について、ターゲットアプリケーション候補になっていたアプリケーション群に対してそれぞれのアーキテクチャにおける性能推定を実施し、アーキテクチャを絞り込んでいった。プロジェクトが開始するとターゲットアプリケーション群に対して同様の手法で基本アーキテクチャを決定し詳細化していった。

表-2に「富岳」のプロセッサ、富士通 A64FX³⁾の仕様の概要を示す。プロセッサチップは、7nm FinFETの半導体技術を使い、アプリケーションが使用する48コアのCPUとインターコネクトネットワークであるTofu Interconnect D (TofuD) インタフェース、I/OのためのPCI Expressインタフェースが搭載されたメニーコア・プロセッサである。図-1はチップのダイ写真である。12コアごとにグループ化され、これをCMG (Core Memory Group) と呼んでいる。同一CMGにあるコアはそれらのコア間で共有されるL2キャッシュとHBM2 (High Bandwidth Memory 2) と呼ばれるメモリ

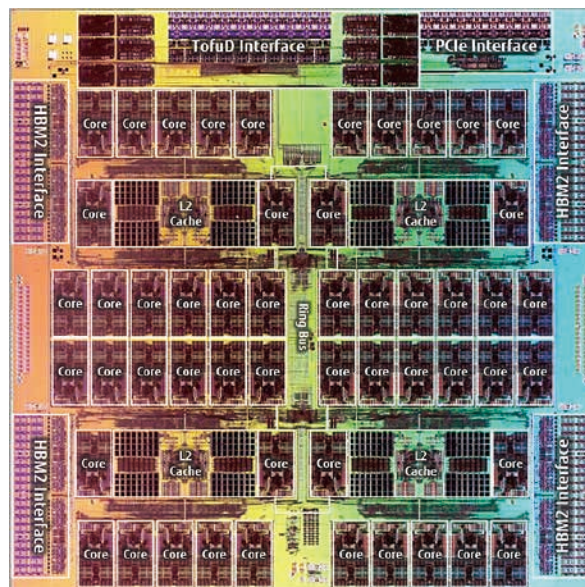


図-1 A64FX チップのダイ写真

を接続するメモリコントローラに接続されている。CMGは、チップ内のネットワークで接続されており、CMG間においてメモリは共有される。CMGには12コアに加えて1つのアシスタントコアとよばれるコアがオペレーティングシステムのデモンプロセスやI/O処理専用コアとして実装されている。命令セットは、Arm社の64bitアーキテクチャArmv8.2-Aに高性能計算のためのSIMD命令仕様であるSVE (Scalable Vector Extension)²⁾が採用されている。SIMD幅は512bitで、8つの倍精度浮動小数点(64bit)データを同時に演算でき、2.7 TFLOPS以上の理論性能を有する。

主記憶として採用されているHBM2は、シリコン基板上にメモリチップをSi貫通電極(TSV: Through-Silicon Via)により接続した積層メモリである。1つの積層メモリあたり256GB/sのメモリバンド幅を持つ。CPUには4つの積層メモリが接続されており、全体では32GiBの容量、1024GB/sのメモリバンド幅を有する。外部接続されるDDRメモリはない。TofuDは、「京」のネットワークであるTofuインターコネクトと同様の6次元メッシュ／トーラストポロジを採用している。

表-3に「富岳」と「京」の性能比較を示す。「京」は2011年に完成したシステムであり、全系性能は約11PFLOPSある。「富岳」は400以上のラックから構成される予定であるが、そのうちの10ラックでほぼ「京」全体の性能を有することになる。

SVE

SVE^{☆5}はArm社が富士通の協力を得て策定したHPCアプリケーションのためのArmv8で拡張されたSIMD命令セットである。以下の特徴を持つ。

☆5 Arm limited, "ARM Architecture Reference Manual Supplement - the Scalable Vector Extension (SVE), for Armv8-a." 2018, <https://developer.arm.com/docs/ddi0584/latest/arm-architecture-reference-manual-supplement-the-scalable-vector-extension-sve-for-armv8-a>

(1) ベクトル長非依存プログラミング (Vector-length agnostic programming) が可能。VLと呼ばれるレジスタにSIMD命令のベクトル長が格納されており、各命令はこれを参照して動作する。これにより1つのバイナリを異なるベクトル長のハードウェアで動作させることができる。

(2) SVEのほとんどのSIMD演算命令は、ベクトルの各要素を演算するかしないかを制御する機構があり、これを駆使することによってコードをコンパクトにすることができる。たとえば、SIMD命令を使ったループの最初と最後の処理を簡単にすることができる。

(3) アドレスをベクトル要素として持つオペランドとして、複数のメモリアドレスの要素を読み書きする、Gatherロード命令、Scatterストア命令がある。

(4) ベクトル要素のデータ型は、FP64(倍精度浮動小数点)、FP32(単精度浮動小数点)に加えてFP16(半精度)やINT8(8ビット整数)のデータ型がある。FP16やINT8を使って、機械学習等のAIアプリケーションを加速することが可能である。

基本性能

倍精度浮動小数点行列計算DGEMMの性能およびSTREAMベンチマークプログラムの中のTriad($a[j] = b[j] + \text{scalar} * c[j]$ の計算)の性能はそれぞれ以下のとおりである。

- DGEMM : 2.5+ TFLOPS
- STREAM Triad : 840 GB/s

また、図-2に示す通り、1ノードにおけるHimeno

表-3 「富岳」と「京」の比較

		「富岳」	「京」
CPUアーキテクチャ		A64FX	SPARC64 VIIIfx
ノードあたり	コア数	48	8
	理論倍精度浮動小数点演算性能	2.7 TFLOPS以上	0.13 TFLOPS
	主記憶容量	32 GiB	16 GiB
	理論メモリバンド幅	1024 GB/s	64 GB/s
	理論ネットワーク性能	40.8 GB/s x (双方向)	20 GB/s x (双方向)
ラックあたり	ノード数	384	102
	理論倍精度浮動小数点演算性能	約1PFLOPS	約0.013PFLOPS
半導体技術		7 nm FinFET	45 nm

ベンチマーク（ポアソン方程式をヤコビの反復法で解くプログラム）、WRF ベンチマーク（気象予測系プログラム）の性能が測定されており高い性能を発揮している。

ハードウェア構成

図-3 に示す通り「富岳」本体は計算ノードと計

算および I/O 処理を行う計算兼 I/O ノードから構成され、15 万ノード以上（理論倍精度浮動小数点演算性能 400PFLOPS 以上）が設置される予定である。予想ピーク演算・メモリ性能を表-4 に示す。計算兼 I/O ノードにはユーザ用ストレージが搭載されている計算兼ストレージ I/O ノードがあり 1.6 TB の SSD が搭載される。計算ノードのアシスタ

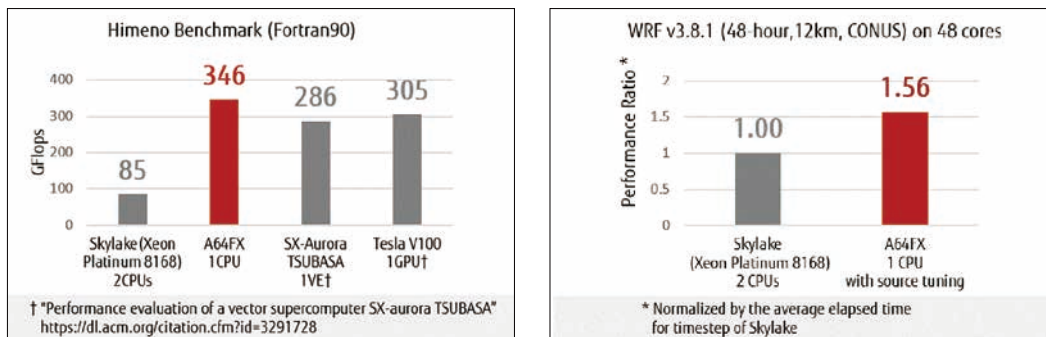


図-2 Himeno ベンチマーク, WRF ベンチマークの結果

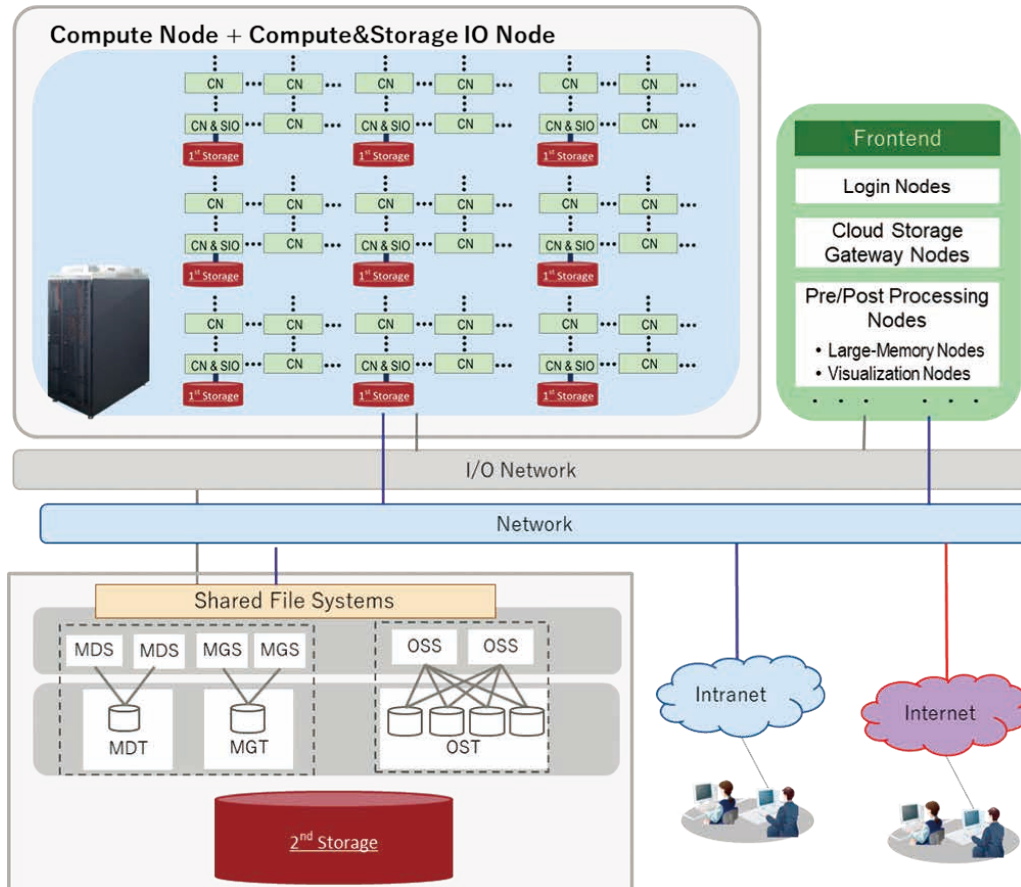


図-3 「富岳」ハードウェア構成概要

ントコアは2基、計算兼ストレージ I/O ノードのアシスタントコアは4基である。これらコアは他の48コアと同じ仕様である。16台の計算ノードごとにそのうちの1台が計算兼ストレージ I/O ノードとして構成される。

ストレージシステムは2階層構成とし、第1階層を計算兼ストレージ I/O ノードに搭載されているストレージ、第2階層として約150PB容量を有する共有並列ファイルシステムを提供する。

本体システムに接続されるネットワークは共有ファイルサーバに接続するネットワーク、フロントエンドに接続するネットワークなどから構成される。フロントエンドネットワークから日本全国の大学、研究機関等の学術情報基盤ネットワークであるSINETに接続される。

将来の拡張性についての検討

一部の計算ノードのPCI Express経由で他のデバイスに接続しバッチジョブ運用できるよう検討を進めている。たとえば、将来、アプリケーション実行に特化した高性能FPGAクラスターを接続し、「富岳」の計算ノードとともに協調計算できるような仕組みをバッチジョブシステムに組み込んでいる。

ソフトウェア

ソフトウェアスタックの概要を図-4に示す。「富岳」は「京」同様LinuxをOSとして採用している。「京」のプロセッサアーキテクチャであるSPARC向けのLinux Distributionは存在しなかった。このため「京」ユーザがLinux系ツールやライブラリを利用したいときは、ソースコードを入手してコンパイルしなければならなかった。「富岳」ではRed Hat Enterprise Linux 8によりツールなどのソフトウェアが整備さ

れ「京」のような不便さは解消される。

また、ますます、さまざまなオープンソースの活用が見込まれるが、同一ソフトウェアでもユーザによって異なるバージョンを利用したい、あるいは、新しいシステムソフトウェアを利用したい、などさまざまな要求が出てくるだろう。これらのニーズに応えるべく「富岳」においては、オープンソース管理ツール、仮想マシン、コンテナなどを導入する。

プログラミング言語・ライブラリ

富士通製コンパイラとして以下のプログラミング言語環境が提供される。

- Fortran 2008 と Fortran 2018 の一部仕様
- C11仕様ならびにGNUおよびClang^{☆6}拡張
- C++14 と C++17 の一部仕様ならびにGNUおよびClang拡張
- OpenMP4.5ならびにOpenMP 5.0の一部仕様

☆6 オープンソースコンパイラの1つ。

表-4 総演算・メモリ性能

	「富岳」	「京」
総倍精度浮動小数点演算 (64bit)	400 PFLOPS 以上 (34 倍以上)	11.3 PFLOPS
総単精度浮動小数点演算 (32bit)	800 PFLOPS 以上 (70 倍以上)	11.3 PFLOPS
総半精度浮動小数点演算 (16bit)	1,600 PFLOPS 以上 (141 倍以上)	—
総メモリバンド幅	150 PB/sec 以上 (29 倍以上)	5,184TB/sec

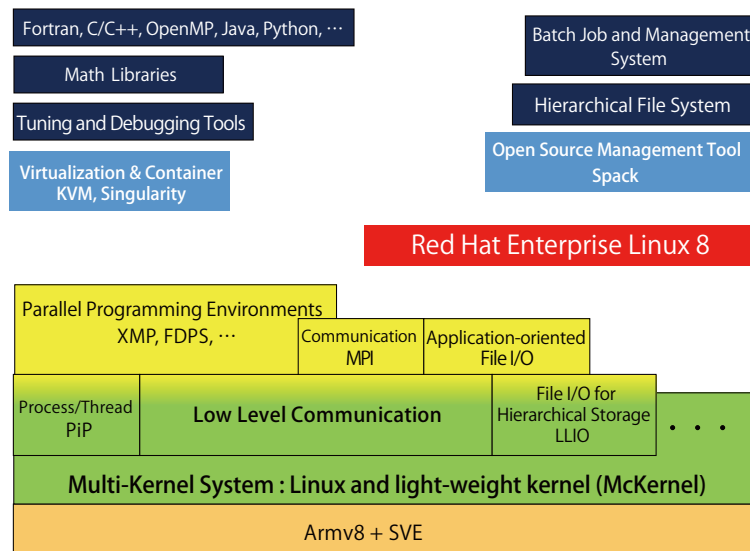


図-4 ソフトウェアスタックの概要

富士通製コンパイラに加えてGCC, LLVM, Arm製コンパイラも利用できる。また, Java, Ruby, Python (NumPy や SciPy 等のモジュールを含む) は富士通製コンパイラによる最適化された環境が提供される。

さらに, 理研が開発している XcalableMP 言語^{☆7}, FDPS ライブラリ^{☆8} が提供される予定である。XcalableMP は, Fortran, C および C++ の拡張として定義される PGAS 言語であり, 指示文に基づく平易なグローバルビュー並列化と, Fortran 2008 から導入された Coarray 機能に基づく, より精密なローカルビュー並列化の両方を提供する点を特徴とする。

FDPS は粒子法の大規模並列化を容易にするフレームワークであり, 粒子を表すクラスおよび粒子間相互作用を計算する関数を与えると, それらを使って

- 分散メモリ並列化のための領域分割
- 粒子のノード間移動
- 粒子間相互作用の計算

を効率的に行う MPI および OpenMP 並列化された関数群を生成する。アプリケーション開発者は C++, Fortran, C 言語からこれら関数を呼び出すことで, 開発者が並列化プログラミングしなくても高性能なプログラムを容易に開発できる。

数値計算ライブラリ

数値計算ライブラリとして, 富士通により BLAS (行列とベクトルに関する基本線型代数ライブラリ), LAPACK (線型計算のための数値解析ライブラリ), ScaLAPACK (並列版線型計算のための数値解析ライブラリ), 富士通製数値計算ライブラリである SSLII が提供される。また, 理研は 150K ノード規模の高並列環境下でのスケーラビリティや小規模問題に対応したライブラリについても開発を進め

ており, 固有値ソルバ EigenExa, 分散並列行列積ルーチン 2.5D-PDGEMM, バッチ BLAS カーネル BatchedBLAS が理研より提供される予定である^{☆9}。

通信ライブラリ

MPI 通信ライブラリの実装として, Open MPI, MPICH と呼ばれる 2 つの流れがある。富士通製通信ライブラリは Open MPI をベースにしている。理研は MPICH をベースにした通信ライブラリを開発してきており MPICH も「富岳」で利用できるようになる予定である。

プログラム開発支援環境

統合開発環境 (IDE) として, オープンソースである Eclipse を提供する。「京」で提供してきたデバッガ, 解析ツールはそれらの改良版が富士通により提供される予定である。また, HPC の分野で広く使われている性能解析ツール Scalasca が利用できる。

システムソフトウェア

オープンソース管理ツール

多くの並列アプリケーションはシステムごとにバイナリを作る必要がある。これはたとえばインターコネクトハードウェアには Tofu, InfiniBand, Omni-Path, Slingshot などさまざまな製品があるが, あるインターコネクト向けに作られたバイナリはほかのインターコネクトでは動作しない。このように, アプリケーションによってはソースコードからコンパイルする必要がある。これを支援するツールとしてパッケージ管理ツール Spack を導入する予定である。Spack は米国ローレンス・リバモア国立研究所 (LLNL) で開発されているツールである。

バッチジョブシステム, LLIO

「京」ではバッチジョブシステムによりユーザの

^{☆7} <https://xcalablemp.org/>

^{☆8} <http://fdps.jmlab.jp/>

^{☆9} <https://www.r-ccs.riken.jp/labs/lpnctr/projects/>

ジョブを実行する（なお、interactive 属性を付けてバッチジョブシステムに要求すると、ユーザが計算ノードを会話的に使うこともできる）。ファイルステージング手法が採用されており、ユーザはバッチジョブスクリプトに共有ファイルサーバに置かれているどのファイルを「京」のローカルストレージにコピーするのか、ジョブ実行後ローカルストレージのどのファイルを共有ファイルサーバにコピーするのかを指示する必要があった。「富岳」ではファイルステージングを採用せず、LLIO (Lightweight Layered IO Accelerator)⁴⁾ と呼ばれるソフトウェアにより第1階層ストレージをキャッシュあるいは一次ファイル置き場として利用し、第2階層ストレージとは非同期 I/O 処理することにより性能および利便性の向上を図る。

仮想マシン、コンテナ

仮想マシンとして Armv8 に対応した KVM を導入する。これにより異なる Linux バージョンの利用も可能となる。またコンテナとして Singularity を導入すべく現在動作検証を進めている。Singularity は、米国ローレンス・バークレー国立研究所で開発されているコンテナ・ソフトウェアである。

ランタイム、OS カーネル

Linux カーネルに加えて理研は、軽量カーネル IHK/McKernel^{☆10} を開発している。Linux と共存して動作し、ユーザはバッチジョブ実行時に IHK/McKernel を使うかどうか選択できる。また、理研が研究開発してきている新しいプロセス・スレッド実行モデル PiP^{☆10} も「富岳」で提供予定である。

今後の開発スケジュール

「富岳」は、「京」を撤去し同じフロアに設置される。2019 年 8 月に「京」の運用を停止し、2019 年末か

ら「富岳」の設置を開始する。2021 年頃の共用開始を目指して設置・調整が続くが、2020 年の 4 月頃から、文部科学省の成果創出加速プログラム等において限定的に利用できるよう計画している。

内閣府の総合科学技術・イノベーション会議の専門部会では、「富岳」について、Society 5.0 における研究開発基盤として活かされるようにビッグデータ、AI 等のアプリケーションについても開発を促進するように指摘されており、今後、これら分野への利用促進の取り組みが進むであろう。

参考文献

- 1) 理化学研究所「富岳」関連技術情報, <https://postk-Web.r-ccs.riken.jp>
- 2) 石川 裕 他: レイテンシコアの高度化・高効率化による将来の HPCI システムに関する調査研究, 第 11 回戦略的高性能計算システム開発に関するワークショップ, 2014 年 3 月, <http://www.open-supercomputer.org/workshop/sdhpc/sdhpc-11/>
- 3) 富士通「富岳」関連参考資料, <http://www.fujitsu.com/jp/solutions/business-technology/tc/catalog/>
- 4) 岡本拓也, 酒井憲一郎, 住元真司, 石川 裕: 計算ノード Burst Buffer 実現のためのデータ処理方式の検討と評価, 情報処理学会, 2018-HPC-165 巻, 7 号 (2017).
(2019 年 8 月 7 日受付)

石川 裕 (正会員) yutaka.ishikawa@riken.jp

1987 年慶応義塾大学大学院工学研究科電気工学専攻博士課程修了。同年通産省電子技術総合研究所入所。2002～2014 年東京大学大学院情報理工学系研究科コンピュータ科学専攻。現在、理化学研究所計算科学研究センター、フラッグシップ 2020 プロジェクトリーダ。工学博士。

佐藤三久 (正会員) msato@riken.jp

2018 年度より、理化学研究所計算科学研究センター副センター長。2014 年より、フラッグシップ 2020 プロジェクト副プロジェクトリーダ、プログラミング環境研究チームリーダ。筑波大学連携大学院教授、筑波大学名誉教授、理学博士。

新庄直樹 shinjo@jp.fujitsu.com

1987 年東京大学大学院理学系研究科修士課程修了、富士通（株）入社。以来、スーパーコンピュータの開発に従事。2007 年次世代テクニカルコンピューティング開発本部システム開発統括部長代理、2014 年同本部長。

清水俊幸 t.shimizu@jp.fujitsu.com

1988 年東工大大学院理工学研究科修士課程修了、同年（株）富士通研入社、1996 年米 Wisconsin 大 Madison 校客員研究員。2010 年より、九州大学情報基盤研究開発センター客員教授、現在、富士通（株）次世代テクニカルコンピューティング開発本部に勤務、1993 年電子情報通信学会論文賞。

☆10 <http://www.sys.r-ccs.riken.jp/>