データ解析・シミュレーション融合スーパーコンピュータシステム **Reedbush** の紹介

塙　敏博　中島　研吾
東京大学情報基盤センター

1　はじめに

本稿では、本年7月1日から一部稼働を開始した「データ解析・シミュレーション融合スーパーコンピュータシステム **Reedbush** 」について紹介する。ReedbushはCPUのみの計算ノード群と、CPUに加えGPUを搭載した計算ノード群のハイブリッドなシステムである。

2　背景

本センターでは現在、Yayoi, Oakleaf-FX, Oakbridge-FXの3システムを運用している。また本センターは、筑波大学と共同で最先端共同HPC基盤施設(JCAHPC: Joint Center for Advanced High Performance Computing[1]) を立ち上げ、2013年3月よりPost T2Kシステムの調達を行ってきた[*1]。

本センターのスーパーコンピュータシステムは、図1に示す通り、工学、地球・宇宙科学、材料科学といった様々な分野の、2,000を超える利用者に活用されている。利用者の半数以上が学外からの利用者であり、また最近では生物学、生体力学、生化学などの分野の利用者が増加している。このような計算機利用需要の増加から、Oakleaf/Oakbridge-FXは大変混雑しているのが現状である。特に2015年のシステム利用率は80%以上に及び、計算資源の拡充が急務であった。しかしながら、Post T2Kシステムは当初の導入予定から遅れ2016年12月頃に稼働の予定であり、また最新アーキテクチャであることから、ユーザがプロダクトラン可能になるまでには、やや時間を要する可能性があった。

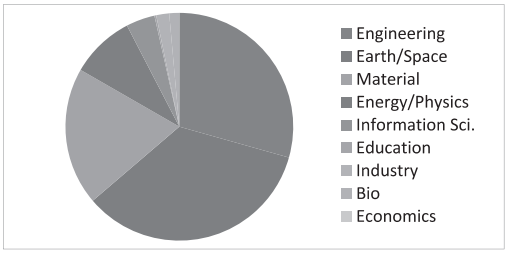

図 1　Oakleaf/Oakbridge-FX の 2015 年の分野ごと使用内訳 (CPU 時間)

一方、Oakleaf-FXは2018年3月に運用を終える予定であり、2018年の秋頃より新しいシステム(Post FX10)の運用開始を目指している。図1からもわかる通り、現在のシステムは主に計算科学や工学向けに利用されており、またFX10システムもそのような目的の利用を想定し設計されている。Post FX10システムにおいては、さらに多くの方々に利用していただくために、

[*1] 詳細については「メニーコア型大規模スーパーコンピュータシステム：Oakforest-PACS」の記事を参照のこと。

ビッグデータ解析や人工知能といった、近年盛り上がりを見せている新たな分野の要求をも満たすシステムの開発を目指している。

これらの現状を踏まえて、我々は新たにReedbushと呼ばれるシステムを導入することにした。このシステムには、以下の2つの大きな役割が期待されている。

1. Oakleaf/Oakbridge-FXの混雑緩和
2. Post FX10 システムに向けてのテストベッドシステム

我々が演算アクセラレータを搭載したシステムを導入するのは今回が初めてである。以前は、アクセラレータ用のプログラミング環境として、GPUにおいてはCUDAのような専用のプログラミング言語を用いて記述する必要があった。そのため、2,000人を超えるユーザに、そのような言語を習得してもらうのは困難であると考えてきた。しかし近年、OpenACCといった指示文ベースのアクセラレータ用並列プログラミング言語が標準的に使われるようになり、実用に耐えうる十分な性能が得られるようになってきた。さらに、データ科学や機械学習など、従来のユーザとは異なる分野からも、新たにGPU搭載スパコンへのニーズが高まっていることや、OpenACCやGPUクラスタなどに詳しい教員が増えたことから、演算アクセラレータとしてGPUを搭載したシステムの導入を決定した。

3　ハードウェア

Reedbushシステムは、CPUのみのノードからなる**Reedbush-U**と、演算アクセラレータとしてGPUを搭載したノードからなる**Reedbush-H**の2つのサブシステムから構成され、それぞれは独立のシステムとして運用される。図3に全体構成図、表2にシステム全体の仕様、表1に各計算ノードの仕様を示す。

Reedbushシステムは、浅野キャンパスにある情報基盤センター別館に設置される。消費電力は368.4 kW（冷却除く）で、全て空冷である。

図2　Reedbush システムの外観

3.1 汎用計算サブシステム：Reedbush-U

各計算ノードは、各ソケットに18コアの 最新世代 Intel Xeon E5プロセッサ (開発コード名: Broadwell-EP)を2ソケット搭載し、256 GBの DDR4メモリを搭載する。ノードあたり性能は 1.2 TFLOPS、メモリバンド幅は 153.6 GB/sec である。

Reedbush-Uサブシステム全体は 420台の計算ノードからなり、各ノードは100 GbpsのInfiniBand EDRによりフルバイセクションバンド幅を持つFat-treeトポロジで接続されている。ピーク演算性能は 508.03 TFLOPS, 総メモリ容量は 105 TByteである。

3.2 演算加速サブシステム：Reedbush-H

各計算ノードは、Reedbush-Uと同じCPU、メモリを搭載しており、加えて 2基の最新世代 NVIDIA Tesla P100 GPU (開発コード名: Pascal)を搭載する。このGPUは1基あたり、4.8〜5.3 TFLOPSと極めて高い性能を持ち、また16 GByteのHBM2 (High Bandwidth Memory)を搭載し、メモリバンド幅は 720 GByte/秒に達する[2]。

特徴的なのは、図4に示すように、

- 新しい高速インタコネクトであるNVLinkにより2基のGPU間が40 GByte/秒のバンド幅で接続されていること
- 各GPUに近接したInfiniBand FDRのHCA (Host Channel Adapter)が用意され、GPUメモリの内容を他のノードとの間で直接送受信できるように工夫されていること

である。

Reedbush-Hサブシステム全体は120台の計算ノードからなり、各ノードは56 Gbps のInfiniBand FDRを2リンク持ち、フルバイセクションバンド幅を持つFat-tree トポロジで接続されている。ピーク演算性能は 1287.4〜1418.2 TFLOPS、総メモリ容量は 30 TByteである。

3.3 ストレージ

Reedbushは、ストレージとして5.04 PByteの並列ファイルシステムと、209 TByteの高速ファイルキャッシュシステムを備える。並列ファイルシステムはLustreファイルシステムであり、実際にデータの格納するOSS (Object Storage Server)としてDataDirect Networks社のSFA14KEを3 セット備え、計算ノード群に対して合計145.2 GB/秒 のバンド幅を提供する。高速ファイルキャッシュシステムには、同じくDataDirect Networks社のIME14Kを6 セット用いる。これは、多くのSSDを搭載した複数のサーバを用いて、ファイル書き込みをまとめて高速化するバーストバッファや、ファイル読み込みのキャッシュなどといった機能を実現するもので、計算ノード群に対して合計436.2 GB/ 秒のバンド幅を実現する。

並列ファイルシステムについては、RAID6によるディスク冗長化、ファイルサーバやコントローラの二重化などにより、信頼性・可用性・耐故障性を高めていると同時に、無停電電源装置(UPS)を備え、万一の停電に備えてバッテリバックアップを行っている。

高速ファイルキャッシュシステムについては、erasure codingという技術で、IMEを構成する

サーバ間でパリティデータを保持することで、高いバンド幅と高い信頼性を両立させている。

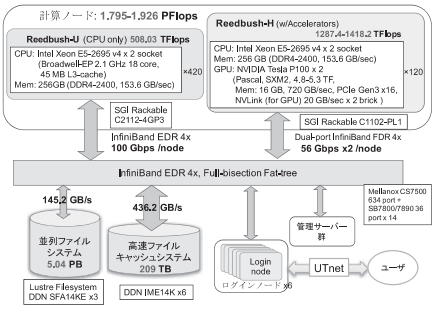

図 3　Reedbush システムの概要

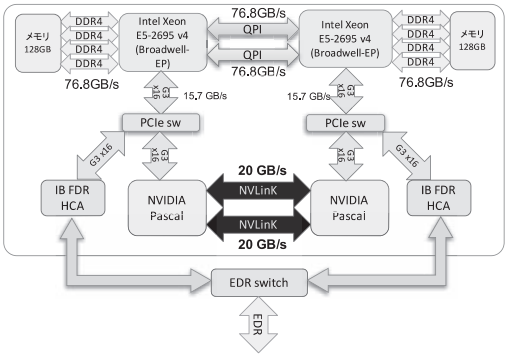

図 4　Reedbush-H ノードの構成

表 1　計算ノードの仕様

		Reedbush-U	Reedbush-H
製品名		SGI Rackable C2112-4GP3	SGI Rackable C1100 シリーズ (開発中)
CPU	プロセッサ名	Intel Xeon E5-2695v4 (Broadwell-EP) × 2 ソケット	
	周波数・コア数	2.1 GHz, 18 コア × 2 ソケット	
	ピーク性能	1209.6 GFLOPS	
メモリ	種別・構成	DDR4-2400, 4 チャネル × 2 ソケット	
	容量	256 GB	
	バンド幅	153.6 GB/秒	
GPU	プロセッサ名	搭載なし	NVIDIA Tesla P100 (Pascal)
	演算ユニット (単体)		56 SM (Symmetric Multiprocessor) × 64 CUDA コア (単精度), 32 CUDA コア (倍精度)
	ピーク性能（単体）		4.8〜5.3 TFLOPS
	メモリ種別（単体）		HBM2
	メモリ容量（単体）		16 GByte
	メモリバンド幅（単体）		720 GByte/秒
	搭載数		2 基
	GPU 間接続		NVlink × 2 brick (40 GB/秒)
	CPU-GPU 間接続		PCI Express Gen3 x16 レーン (16 GB/秒)
計算ノード間ネットワーク		InfiniBand EDR 4x (100 Gbps)	InfiniBand FDR 4x ×2 リンク (56 Gbps ×2)

4　ソフトウェア

以下に主なインストール済みソフトウェアについて紹介する。表3に主なソフトウェアの一覧を示す。

OSには、Red Hat Enterprise Linux 7を採用している。

表 2　Reedbush システム全体仕様

Reedbush-U	総理論演算性能	508.03 TFLOPS
	総ノード数	420
	総主記憶容量	105 TByte
Reedbush-H	総理論演算性能	1297.15〜1417.15 TFLOPS (うち GPU: 1152.0〜1272.0 TFLOPS)
	総ノード数	120
	総主記憶容量	30 TByte
計算ノード間ネットワーク		InfiniBand EDR 4x フルバイセクションバンド幅 Fat-tree
並列ファイルシステム	種類 サーバ (OSS)	Lustre ファイルシステム DDN SFA14KE
	サーバ (OSS) 数	3 セット (6 ノード、12 サーバ)
	ストレージ容量	5.04 PByte
	バンド幅	145.2 GB/秒
高速ファイルキャッシュシステム	サーバ	DDN IME14K
	サーバ数	6 セット (12 ノード)
	容量	209 TByte
	バンド幅	436.2 GB/秒

4.1　コンパイラ

まず、本システムの計算ノードのCPUは x86_64アーキテクチャのBroadwell-EPであり、ログインノードも同じである（コア数等は異なる）。従って、ソースコードをコンパイルする際、Oakleaf-FX/Oakbridge-FXなどのように、クロスコンパイラを用いる必要はない。

コンパイラとして、OSに標準で付属しているGNUコンパイラ環境(Fortran, C, C++)、およびIntelコンパイラ(Fortran, C, C++)がインストールされている。

Reedbush-Hについては、以上に加えて、NVIDIA GPU向けのプログラミング言語であるCUDA Cコンパイラ、PGI CUDA Fortran、さらに、PGI OpenACCコンパイラ (Fortran, C, C++) が用意される。

4.2　メッセージ通信ライブラリ

Intelコンパイラと親和性の高いIntel MPI、導入ベンダであるSGIから提供される SGI MPT、InfiniBandベンダである Mellanoxから提供される Mellanox HPC-X に加えて、オープンソースの Open MPI、MVAPICH2も使うことができる。

Reedbush-Hについては、GPUメモリを直接通信可能にするGPUDirect for RDMA (GDR) 機能を用いることができる、MVAPICH2-GDRおよびOpen MPIが用意される。

4.3　ライブラリ

IntelによるMath Kernel Library (MKL)により、BLAS, LAPACK, ScaLAPACKなどがサポートされる。その他、オープンソースの数値計算ライブラリである、SuperLUやFFTW、グラフ

計算ライブラリのMetisなどがインストールされる。

Reedbush-Hについては、GPU向けのライブラリとして、NVIDIAによるcuBLAS, cuSPARSE, cuFFTなどの数値計算ライブラリに加えて、MAGMAなどのオープンソースの数値計算ライブラリが用意される。また、cuDNN、Theanoなどの機械学習ライブラリ、OpenCVなどの画像処理ライブラリなども利用可能である。

4.4 アプリケーション

オープンソースの流体解析ソフトウェアOpenFOAMを始め、計算科学、工学、バイオインフォマティクス、機械学習などのライブラリが利用可能である。

表3 主なソフトウェア一覧

OS		Red Hat Enterprise Linux 7
コンパイラ		GNU コンパイラ、Intel コンパイラ (Fortran77/90/95/2003/2008、C、C++)
	(-H)	PGI コンパイラ (Fortran77/90/95/2003/2008、C、C++、OpenACC 2.0、CUDA Fortran) NVCC コンパイラ (CUDA C)
MPI ライブラリ		Intel MPI, SGI MPT, Mellanox HPC-X, Open MPI, MVAPICH2
	(-H)	GPUDirect for RDMA support: MVAPICH2-GDR, Open MPI
ライブラリ		Intel 社製ライブラリ (MKL): BLAS、LAPACK、ScaLAPACK その他ライブラリ: SuperLU、SuperLU MT、SuperLU DIST、METIS、MT-METIS、ParMETIS、Scotch、PT-Scotch、PETSc、FFTW、GNU Scientific Library、 NetCDF、 PnetCDF など
	(-H)	cuBLAS、cuSPARSE、cuFFT、MAGMA、OpenCV、ITK、Theano、Anaconda、 ROOT、 TensorFlow など
アプリケーション		OpenFOAM、ABINT-MP PHASE、FrontFlow、FrontISTR、REVOCAP、ppOpen-HPC など
デバッガ、プロファイラ		Total View, Intel VTune, Trace Analyzer & Collector

5 スケジュール

本システムは、2015年8月に調達を開始した。2016年3月に開札が行われ、納入業者がSGIに決定した。

2016年7月1日にReedbush-Uが稼働開始し、試験運用期間を経て、2016年9月1日から、本運用を開始する予定である。また、Reedbush-Hは2017年3月1日に稼働開始の予定で、その後1ヶ月の試験運用期間を経て、2017年4月3日より全系による運用を開始する予定である。

参考文献

[1] 最先端共同HPC基盤施設. http://jcahpc.jp.

[2] NVIDIA. Whitepaper: NVIDIA Tesla P100. https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf.