

Wisteria/BDEC-01 スーパーコンピュータシステムの運用

佐藤 孝明^{1,a)} 宮寄 洋¹ 前田 光教¹ 山本 和男¹ 福沢 秋津¹ 中張 遼太郎¹ 山田 新¹
中島 研吾² 塙 敏博²

The operation of Wisteria/BDEC-01 Supercomputer System

TAKAAKI SATOH^{1,a)} HIROSHI MIYAZAKI¹ MITSUNORI MAEDA¹ KAZUO YAMAMOTO¹ AKITSU FUKUZAWA¹
RYOTARO NAKAHARI¹ HAJIME YAMADA¹ KENGO NAKAJIMA² TOSHIHIRO HANAWA²

概要：本稿は、2021 年 5 月より運用を開始した Wisteria/BDEC-01 スーパーコンピュータシステムに関する導入の経緯とシステムの概要に加え、利用者が Wisteria/BDEC-01 を利用するための制度について説明する。また、試験運用から正式運用にかけて得られた利用状況のデータについて報告する。

1. はじめに

東京大学情報基盤センター^{[1], [2]}(以下、当センター)では、「計算・データ・学習」融合スーパーコンピュータシステム(Wisteria/BDEC-01)^{[3], [4]}を導入し、2021 年 5 月から試験運用、同年 8 月から正式運用を開始した。

Wisteria/BDEC-01 は、計算(Simulation)、データ(Data)、学習(Learning)の融合(「S+D+L」融合)を実現するためのプラットフォームとして Fujitsu FX1000 で構成されたシミュレーションノード群(Odyssey)と最新の GPU(NVIDIA A100 Tensor コア GPU)を搭載したデータ・学習ノード群(Aquarius)の 2 つのサブシステムで構成されている点の特徴となっている。当センター最初の GPU クラスタである Reedbush-H、Reedbush-L^[5]では、各ノードに NVIDIA P100 GPU がそれぞれ 2 基、4 基搭載されており、ノード単位での運用を実施していた。Aquarius では、各ノードに 8 基の GPU が搭載され、GPU 単位での運用が可能であり、利用者のより多様なニーズに応えることが可能となっている。

2. 導入の背景

2021 年 3 月時点では、当センターでは 4 式(Reedbush-H、Reedbush-L、Oakforest-PACS、Oakbridge-CX)のシステムを運用しており、総利用者数は約 2,600 名、そのうち 55 %



図 1 Wisteria/BDEC-01(Odyssey) 外観

は学外利用者という状況であった。各システムは、高い計算性能、利用しやすいプログラム開発環境、安定した運用が利用者に高く評価されてきた。

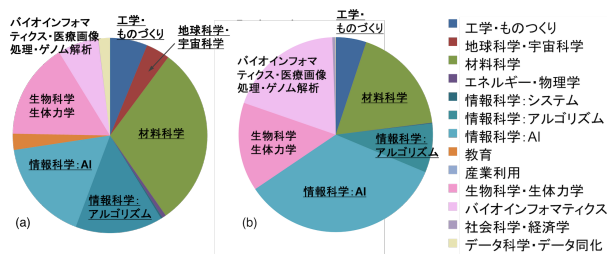


図 2 実行ジョブノード時間の分野別比率(2020 年度)、(a) Reedbush-H (ノード当たり 2GPU)、(b) Reedbush-L (ノード当たり 4GPU)

¹ 東京大学 情報システム部 情報基盤課
Information Systems Department, Information Technology Group,
The University of Tokyo.

² 東京大学 情報基盤センター
Information Technology Center, The University of Tokyo.

^{a)} takaaki@cc.u-tokyo.ac.jp

各システムの利用分野を分析すると、長年にわたって工学・ものづくり、地球科学・宇宙科学、材料科学が利用時間の多くを占めてきていたが、当センター初の GPU 搭載システムである Reedbush-H、Reedbush-L では人工知能、

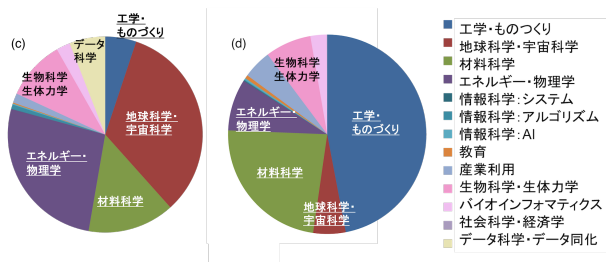


図3 実行ジョブノード時間の分野別比率 (2020 年度)、(c) Oakforest-PACS、(d) Oakbridge-CX

生物科学関連分野、医療画像処理を中心としたバイオインフォマティクスなど、より多様な分野で使用されていることがわかっている (図 2 (a, b))。また、本来計算科学・計算工学用途を念頭において導入された Oakforest-PACS、Oakbridge-CX においてもデータ科学、生物科学関連分野の利用が多くなってきている傾向がある (図 3 (c, d))。

スーパーコンピューティングは、従来の計算科学・計算工学シミュレーションに加えて、データ科学、機械学習等の知見を融合した新しい手法を適用することによって、サイバー空間とフィジカル空間の融合を通じた Society 5.0 実現に大きく貢献すると期待されているが、当センターでは 2015 年頃からこのような状況を念頭において、「計算・データ・学習 (Simulation+Data+Learning, S+D+L)」融合を実現するプラットフォームとして「計算・データ・学習」融合スーパーコンピュータシステム (通称「BDEC(Big Data & Extreme Computing) システム」) 構築を目指し、様々な研究開発を進めてきた。Reedbush や Oakbridge-CX は、いずれも「BDEC システム」設計のためのプロトタイプかつ実証システムとしても位置づけられている。

Wisteria/BDEC-01 はこのような背景のもと、「BDEC システム」の 1 号機として、設計、導入を実施し、運用を開始したシステムである。Odyssey と Aquarius を使用し、幅広いアプリケーションをカバーすることによって、最先端の科学技術計算を支える重要なインフラとなることだけでなく、東京大学各部局 (生産技術研究所、地震研究所、大気海洋研究所、物性研究所)、理化学研究所等との協力のもと、ものづくり、地球科学分野 (固体地球、大気・海洋)、物質科学などの分野における「S+D+L」融合により、Society 5.0 実現に向けた重要なプラットフォームとなることが期待されている。

3. ハードウェアの概要

3.1 ハードウェア全体構成

Wisteria/BDEC-01 は、2 つのサブシステム Odyssey と Aquarius で構成 (図 4、5、表 1、2) され、富士通株式会社により開発されたスーパーコンピュータである。総理論演算性能はそれぞれ 25.9 PFLOPS (Odyssey)、7.2 PFLOPS (Aquarius)、合計 33.1 PFLOPS である。Odyssey と Aquarius

を利用するためのログインノード及び各ファイルシステムは、管理サーバ (ジョブ管理等) も含め、同じ環境が利用できるよう構成されており、ログインノードの切り替えや、Odyssey と Aquarius の間でのデータ移行は発生しない。

表 1 全体構成

項目	Odyssey	Aquarius
総理論演算性能	25.9 PFLOPS	7.2 PFLOPS
総ノード数	7,680	45
総主記憶容量	240.0 TiB	36.5 TiB
インターコネクト	Tofu インターコネクト D	InfiniBand HDR (200Gbps) x 4
ネットワークトポロジー	6 次元メッシュ / トーラス	Full-bisection Fat Tree
共有 ファイル システム	システム名	FEFS (Fujitsu Exabyte File System)
	サーバ (OSS)	DDN SFA7990XE
	サーバ (OSS) 数	16
	ストレージ容量	25.8 PB
	ストレージ データ転送速度	504 GB/s
高速 ファイル システム	システム名	FEFS (Fujitsu Exabyte File System)
	サーバ (OSS)	DDN SFA400NVXE
	サーバ (OSS) 数	16
	ストレージ容量	1.0 PB
	ストレージ データ転送速度	1.0 TB/s

3.2 シミュレーションノード群 (Odyssey)

Odyssey の構成やスペックは、図 4、表 1、2 に示すとおりである。A64FX は、Armv8.2-A 命令セットアーキテクチャをスーパーコンピュータ向けに拡張した SVE(Scalable Vector Extension) を世界で初めて実装した CPU である。48 個の演算コアと 2 個または 4 個のアシスタントコアを有しており、Odyssey 全体でのコア数は 368,640 コアとなる。ノード間ネットワークについては、バイセクションバンド幅が 13.0 TB/s の Tofu インターコネクト D で各ノードが結合されている。

3.3 データ・学習ノード群 (Aquarius)

Aquarius の構成やスペックは、図 4、5、表 1、2 に示すとおりである。各ノードは、CPU が 2 基に対し、GPU は 8 基搭載されている。ノード間ネットワークは、データ転送速度が 200 Gbps の帯域を有する InfiniBand HDR、4 リンクを用いてフルバイセクションバンド幅で接続されている。

Aquarius の各ノードは、25 Gbps の Ethernet インタフェースを備えており、Aquarius は合計 800 Gbps のネットワーク転送速度で外部との通信が可能な構成となっている (図 5)。当センターの既存システムである Oakbridge-CX と同様に、Aquarius の一部のノードは SINET 等の外部ネットワークを介して、様々な外部リソースに直接アクセス可能であり、観測データをリアルタイムに取り込んで解析、シミュレーションに利用することも可能である。

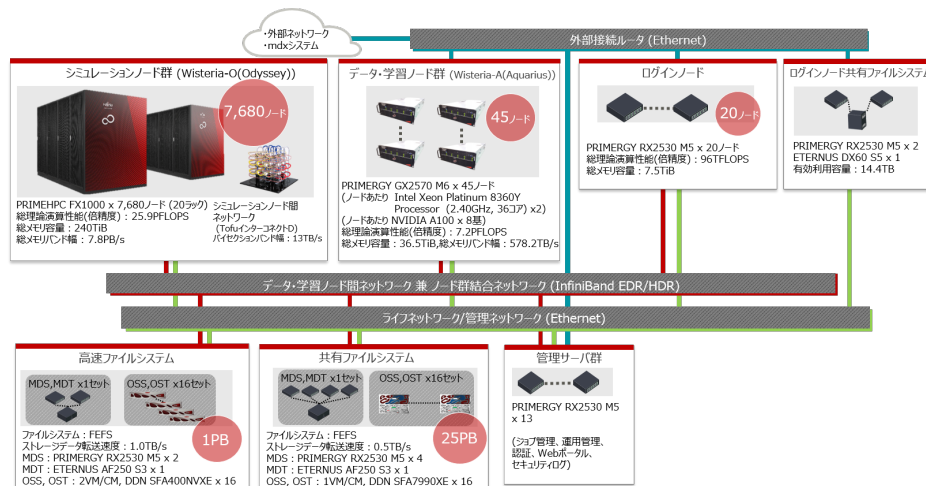


図 4 Wisteria/BDEC-01 全体構成図

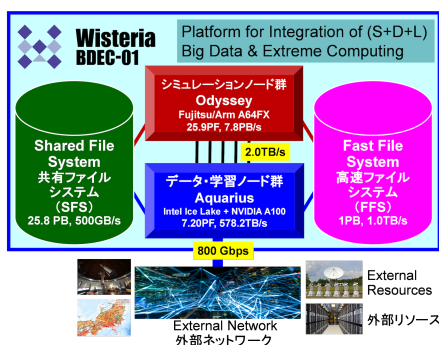


図 5 Wisteria/BDEC-01 の概要図

表 2 ノードの構成

項目		Odyssey	Aquarius
マシン名		FUJITSU Supercomputer PRIMEHPC FX1000	FUJITSU Server PRIMERGY GX2570 M6
CPU	プロセッサ名	A64FX	Intel®Xeon® Platinum 8360Y (Ice Lake)
	プロセッサ数 (コア数)	1 (48+アシスタント コア 2 or 4)	2 (36+36)
	周波数	2.2 GHz	2.4 GHz
	理論演算性能	3.3792 TFLOPS	5.53 TFLOPS
	メモリ容量	32 GB	512 GiB
	メモリ帯域幅	1,024 GB/s	409.6 GB/s
	搭載数	-	8
GPU	プロセッサ名	-	NVIDIA A100
	SM 数 (単体)	-	108
	メモリ容量 (単体)	-	40 GiB
	メモリ帯域幅 (単体)	-	1,555 GB/s
	理論演算性能 (単体)	-	19.5 TFLOPS
GPU 間接続	CPU-GPU 間接続	-	PCI Express Gen4 x 16 レーン (1 レーンあたり 片方向 32 GB/s)
	GPU 間接続	-	NVLink x 12 本 (1 本あたり 片方向 25GB/s)

3.4 ストレージシステム

ストレージ環境は、システムで利用するデータやプログラムを格納する用途としての「共有ファイルシステム」、加えて、高速にデータの読み書きが可能な SSD で構成された「高速ファイルシステム」の 2 つのストレージが用意されている。分散ファイルシステムとしては FEFS (Fujitsu Exabyte File System) が用いられており、図 5 に示すとおり、Odyssey、Aquarius のすべてのノードと、すべてのログインノード (プリポストノード) から、大規模なデータを高速に処理することが可能である。

4. ソフトウェアの概要

4.1 ソフトウェア構成

4.1.1 主なソフトウェア群

Wisteria/BDEC-01 では、Fortran、C/C++ コンパイラや、Python をはじめとするインタプリタ、MPI 通信ライブラリ等といったスーパーコンピュータで開発、研究を行うための基本的なソフトウェアに加え、計算科学、データ科学、機械学習、人工知能等様々な分野のライブラリ、ツール、アプリケーションをインストールしている。Odyssey と Aquarius で使用できる開発環境の違いを表 3 に示す。オープンソースのアプリケーションとしては、OpenFOAM^[6](数

値流体力学)、MateriApps アプリケーション群^[7](物質科学)、東京大学生産技術研究所で開発された革新的シミュレーションソフトウェア群^[8]、当センターで開発された自動チューニング機構を有するアプリケーション開発・実行環境 (ppOpen-HPC)^[9]と「S+D+L」融合のための革新的ソフトウェア基盤 (h3-Open-BDEC)^{[10], [11]}も利用可能となっている。

ソフトウェア開発環境は HPC ミドルウェアの FUJITSU Software Technical Computing Suite (TCS) で提供されており、TCS はこの他に、ファイルシステム環境などのシステ

ムの運用管理機能、ジョブ運用管理などの利用環境に関する機能を有している。システム使用者は、TCS を使用することで、CUI (pjsb コマンド等) や GUI (利用支援ポータル等) を介して Odyssey、Aquarius に対するジョブ実行やトークン量の管理などを行うことができる。

表3 ソフトウェア構成

項目	Odyssey	Aquarius
OS	Red Hat Enterprise Linux 8	
コンパイラ	GNU コンパイラ	
	富士通社製 コンパイラ (Fortran77/90/95/2003/2008、C、C++)	Intel コンパイラ (Fortran77/90/95/2003/2008、C、C++) NVIDIA HPC SDK (Fortran77/90/95/2003/2008、C、C++、OpenACC 2.7) NVIDIA CUDA SDK (CUDA C、CUDA C++)
メッセージ通信ライブラリ	富士通社製 MPI	Intel MPI、Open MPI
ライブラリ	SuperLU、SuperLU MT、SuperLU DIST、METIS、MT-METIS、ParMETIS、Scotch、PT-Scotch、PETSc、Trillinos、FFTW、GNU Scientific Library、NetCDF、Parallel netCDF、HDF5、Parallel HDF5、CMake、Miniconda、Xabclib、ppOpen-HPC、MassiveThreads、Boost C++、mpiJava	
	富士通社製ライブラリ (BLAS、CBLAS、LAPACK、ScaLAPACK)	Intel 社製ライブラリ (MKL) (BLAS、CBLAS、LAPACK、ScaLAPACK)、 cuBLAS、cuSPARSE、cuFFT、 MAGMA、cuDNN、NCCL
アプリ ケーション	OpenFOAM、ABINIT-MP、PHASE、FrontFlow/blue、FrontISTR、REVOCAP-Coupler、REVOCAP-Refiner、OpenMX、MODYLAS、GROMACS、BLAST、R packages、bioconductor、BioPerl、BioRuby、BWA、GATK、SAMtools、Quantum ESPRESSO、Xcrypt、ROOT、Geant4、LAMMPS、CP2K、NWChem、DeepVariant、Paraview、VisIt、POV-Ray、TensorFlow、Chainer、PyTorch、Keras、Horovod、MXNet	
	-	Theano
フリー ソフトウェア	autoconf、automake、bash、bzip2、cvs、emacs、findutils、gawk、gdb、make、grep、gnuplot、gzip、less、m4、python、perl、ruby、screen、sed、subversion、tar、tcsh、tcl、vim、zsh、git など	
	-	Globus Toolkit、Gfarm、FUSE
コンテナ 仮想化	singularity	

4.1.2 利用支援

利用支援としては、利用支援ポータルにおいて、既存システムと同様に、以下の機能を提供している。

- SSH 公開鍵登録
- 計算リソースに関する表示
- グループ内計算リソースの管理

- プリポスト予約
- マニュアルの閲覧

また、この他に、開発環境として利用できる Web アプリケーションである JupyterLab 環境、Web ブラウザを使用したりリモートアクセスを可能にする Apache Guacamole 環境を提供している。

4.2 本システムにおける特徴

4.2.1 GPU 単位での利用

Wisteria/BDEC-01 では、TCS で計算資源の管理を行っているが、Aquarius のノード群に搭載されている 8 基の GPU については、それぞれの GPU 毎での資源管理が可能になっている。したがって、TCS を使用することで、Reedbush とは異なり、例えば、1 つのノード内で 1 GPU を使用する、それぞれ異なるジョブを 8 つ同時に実行するといった管理ができる。これにより、Aquarius は、45 ノードをノード単位で運用するだけではなく、360 GPU を GPU 単位で運用することも可能である。Aquarius の利用制度においては、ノードを効率よく運用するため、この点を踏まえた制度及びジョブキューの設定が追加されている。

なお、NVIDIA A100 においては、NVIDIA MIG (NVIDIA Multi-Instance GPU)^[12] の機能が実装されており、GPU 内を 7 つのインスタンスに分割し、それぞれのインスタンスに対し個別にジョブを割り当てることができる。しかしながら、Aquarius ではスーパーコンピュータとして運用する場合の計算効率等を考慮して、この機能を使用していない。

4.2.2 h3-Open-SYS/WaitIO

h3-Open-SYS は、Wisteria/BDEC-01 のようなヘテロジニアスなシステムにおいて、シミュレーションとデータ処理実行を統合するソフトウェア群であり、当センターで開発が進められている「h3-Open-BDEC(「S+D+L」融合のための革新的ソフトウェア基盤)」^{[10],[11]} の一部である。

h3-Open-SYS/WaitIO^[13] は、その中核機能の一つで、2 つ以上の並列プログラム間で通信を行うためのライブラリである。各並列プログラムは典型的には MPI プログラムであるが、MPI 以外の形の並列プログラムからも利用できるよう汎用的に設計されている。

h3-Open-SYS/WaitIO は多様な並列計算機環境を想定している。特に、通信を行う並列プログラムが同一の並列計算機で動作する場合はもちろん、Odyssey と Aquarius のように、各々の並列プログラムが異なる並列計算機で動作する場合にも対応する。そのために通信経路としてファイルと TCP/IP を想定する。

現状では、Odyssey と Aquarius は個別に運用されているが、h3-Open-SYS/WaitIO の利用により、両者を連携した「S+D+L」融合が実現可能となる。h3-Open-SYS/WaitIO は現在検証中であり 2021 年 10 月以降、順次利用者に公開していく予定である。

5. 利用の仕組み

5.1 トークン制と利用形態

利用申込みを行った際に、使用できる資源として、ノードまたは GPU の利用可能時間に相当する「トークン」とファイルシステムを使用するための「ディスク容量」が利用者のアカウントに設定される。これらは、申込み時のセット数に応じて決まり、トークン量は 1 セット当たり 8,640 トークン (24 時間 x 360 日相当)」、ディスク容量は、共有ファイルシステム用として、1 セット当たり 2 TB が割り当てられる。この割り当てられたトークン内であれば、一部の利用形態を除き、申込みセット数に伴う利用可能ノード数制限はなく、最大利用可能ノード数までバッチジョブの実行が可能である。トークン量はバッチジョブの実行ごとに、Odyssey の場合は「ノード数×経過時間×消費係数」、Aquarius の場合は「GPU 数×経過時間×消費係数 (消費係数は 3)」で計算された量が消費され、トークン量を使い切るとバッチジョブの投入ができなくなる。

表 4 一般/公募制度利用負担金表

区分	利用負担金額等	
一般申込	Odyssey/Aquarius 基本セット (年額) 【大学・公共機関等 60,000 円】 (複数セット申込可, 利用期間は 1 ヶ月単位で設定可)	
	トークン量	8,640 トークン (24 時間× 360 日相当, 1 セット当たり)
	トークン	Odyssey: 1.00 (1 ノード当たり) ※
	消費係数	Aquarius: 3.00 (1 GPU 当たり)
	ディスク容量	共有ファイルシステム グループにつき 2 TB (1 セット当たり)
	利用登録番号数	制限なし
公募制度による申込 (要審査)	Odyssey 基本セット (年額) 【大学・公共機関等 60,000 円, 企業 72,000 円】 (複数セット申込可, 利用期間は 1 ヶ月単位で設定可)	
	トークン量	8,640 トークン (24 時間× 360 日相当, 1 セット当たり)
	トークン	1.00 (1 ノード当たり) ※
	消費係数	
	ディスク容量	共有ファイルシステム グループにつき 2 TB (1 セット当たり)
	利用登録番号数	制限なし
	Aquarius 基本セット (年額) 【大学・公共機関等 180,000 円, 企業 216,000 円】 (複数セット申込可, 利用期間は 1 ヶ月単位で設定可)	
	トークン量	25,920 トークン (1 GPU, 24 時間× 360 日相当, 1 セット当たり)
	トークン	3.00 (1 GPU 当たり)
	消費係数	
	ディスク容量	共有ファイルシステム グループにつき 6 TB (1 セット当たり)
	利用登録番号数	制限なし

※ 消費係数 1.50 のノード群 (優先利用向け) を全体の 15%程度設ける。

表 4 に利用区分に対する負担金や資源量が記載されている。一般申込においては、「S+D+L」融合に向けた機会が増えるように Odyssey と Aquarius の両方を利用できる制度設計になっている。当センターでは、利用負担金や消費係数はシステムの消費電力 (電気料金) に基づいて決めている。一方で、企業利用などの公募制度や HPCI への資源提供など様々な公募制度の内容に対応できるようにする事情から Reedbush と同様に Odyssey と Aquarius とを別のメニューに分けた負担金表も設定している。

5.2 ノード固定 と GPU 専有利用

当センターでは、グループによる固有の要望 (独自のアプリケーションのインストール、ジョブ実行環境の設定、外部資源への接続など) に対応できるように、申込みグループ内で占有してノードを利用できるようにするノード固定サービスを行っている。Wisteria/BDEC-01 でも、表 5 で示のようにノード固定サービスが設定されているが、Odyssey においては、Tofu インターコネクト D によって結合され、最適化されたノード形状を維持するため、サービスを Aquarius だけに限定している。Odyssey においては、混雑対策を兼ねた消費係数 1.5 の優先利用向けノード群を全体の 15 % 程度設けている。

表 5 GPU 専有/ノード固定利用負担金表

区分	利用負担金額等			
GPU 専有申込	Aquarius 1GPU セット (年額) 【大学・公共機関等 270,000 円, 企業 324,000 円】 (最大 7 GPU まで, 申込単位は下表参照, 利用期間は 1 ヶ月単位で設定可)			
	トークン量	25,920 トークン (1 GPU, 24 時間× 360 日相当, 1 セット当たり)		
	トークン	Odyssey: 1.00 (1 ノード当たり) ※		
	消費係数	Aquarius: 3.00 (1 GPU 当たり)		
	ディスク容量	共有ファイルシステム グループにつき 6 TB (1 セット当たり)		
	利用登録番号数	制限なし		
	GPU 数	トークン量	大学・公共機関	企業
	1	25,920 トークン	270,000 円	324,000 円
	2	51,840 トークン	540,000 円	648,000 円
	4	103,680 トークン	1,080,000 円	1,296,000 円
ノード固定 (要審査)	Aquarius 基本セット (年額) 【大学・公共機関等 2,160,000 円, 企業 2,592,000 円】 (1 ノード 8 GPU 1 年分, 1 セットの申込み可, 利用期間は 1 ヶ月単位で設定可)			
	トークン量	207,360 トークン (8 GPU, 24 時間× 360 日相当)		
	トークン	Odyssey: 1.00 (1 ノード当たり) ※		
	消費係数	Aquarius: 3.00 (1 GPU 当たり)		
	ディスク容量	共有ファイルシステム グループにつき 48 TB		
	利用登録番号数	制限なし		

※ 消費係数 1.50 のノード群 (優先利用向け) を全体の 15%程度設ける。

ノード固定サービスは、Aquarius のノード数が少ないこともあるが、消費電力 (電気料金) の関係上 高価なサービスとなっている。これに対し、より安価に利用できる専有サービスとして、TCS が GPU 単位でジョブ実行を管理できる機能を活用し、GPU 専有利用というサービスを設けている。これは、ノードの占有ではなく、申込みセット数分の GPU を、Aquarius 全体の混雑状況に関係なく専有して使用できるというサービスである。ノード固定サービス及び GPU 専有利用においては、割り当てられる Aquarius の資源に加えて、トークン量を消費係数 1 で消費することで、Odyssey も利用することができる。

5.3 ジョブキューと制限値

表 6 が Odyssey 用のジョブクラスと制限値、表 7 が Aquarius 用のジョブクラスと制限値である。最も利用可能ノード数が大きい regular キューは、ジョブスクリプトで指定するノード数により、自動で small、medium、large、x-large に振り分けられるキューで、最大 48 時間ジョブを実行できる (large キューまたは x-large キューは 24 時間)。この他に短時間ジョブ用のキューとして short キュー、デバッグ用途の debug キュー、インタラクティブ実行用の interactive キューを用意している。debug キュー及び interactive キュー利用時にはトークンは消費されない。また、Odyssey における優先利用向けの消費係数 1.5 のノード群を使用するためのキューは、priority-o キューである。

表 6 Odyssey ジョブクラス制限値

キュー名	ノード数	制限 (経過) 時間	メモリー 容量 (GiB)
debug-o	1 ~ 144	30 分	28
short-o	1 ~ 72	8 時間	28
regular-o			
(small-o)	1 ~ 144	48 時間	28
(medium-o)	145 ~ 576	"	"
(large-o)	577 ~ 1152	"	"
(x-large-o)	1153 ~ 2304	24 時間	"
priority-o	1 ~ 288	48 時間	28
interactive-o			
(interactive-o_n1)	1	30 分	28
(interactive-o_n12)	2 ~ 12	10 分	28

Aquarius のジョブクラスは、ジョブに割り当てるリソースがノード単位であるジョブクラス (Odyssey も同様) と、GPU 単位であるジョブクラスの大きく 2 種類のジョブクラスが設定されている。前者は他ジョブと実行時の割り当てノードを共有しないが、後者は割り当てノードを共有する (後者は share キューと呼ぶ)。share キューの場合は、1 GPU 当たりの割り当て CPU コア数は 9 コアであり、最大使用メモリー量は 56 GiB になっている。1 ノード内で実行される全ジョブのメモリー量の合計が 448 GiB 以内となるよう設定されている。

表 7 Aquarius ジョブクラス制限値

キュー名 (ノード単位)	ノード数	制限 (経過) 時間	メモリー 容量 (GiB)
debug-a	1	30 分	448
short-a	1 ~ 2	2 時間	448
regular-a			
(small-a)	1 ~ 2	48 時間	448
(medium-a)	3 ~ 4	"	"
(large-a)	5 ~ 8	24 時間	"
interactive-a	1	10 分	56
キュー名 (GPU 単位)	GPU 数	制限 (経過) 時間	メモリー 容量 (GiB)
share-debug	1, 2, 4	30 分	56
share-short	1, 2, 4	2 時間	56
share			
(share-1)	1	48 時間	56
(share-2)	2	"	"
(share-4)	4	24 時間	"
share-interactive	1	10 分	56

表 8 に示すジョブクラスと制限値は、当センターの Oakbridge-CX などでも機能を有するプリポストキューに関するものである。プリポストキューは、解析データから可視化を行うための処理など、ログインノードで行う処理の中で特に負荷の高い処理を実行するために設けられている。プリポストノードは、ログインノードと同一の環境であり、インタラクティブなジョブ実行を行うことができる。プリポストキューの使用にあたっては、トークンの消費は発生しない。

表 8 プリポストジョブクラス制限値

キュー名	ノード数	制限 (経過) 時間	メモリー 容量 (GiB)
(予約無し)			
prepost	1	6 時間	340
(予約有り)			
prepost1_n1 ~	1	1 ~ 6 時間	340
prepost4_n1	1 ~ 4	1 ~ 6 時間	340
prepost1_n4	1 ~ 8	1 ~ 6 時間	340
prepost1_n8			

5.4 ファイルシステムの利用

Wisteria/BDEC-01 では、表 9 で示されるファイルシステムを提供している。一時領域では、1 週間更新がないファイルが自動削除される仕組みである。

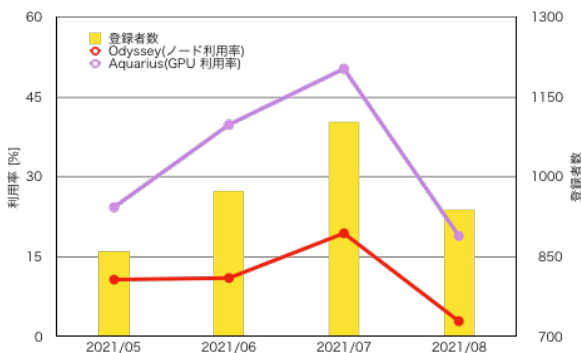
表9 ファイルシステム一覧

ファイルシステム	割り当て容量	主な利用用途
共有ファイルシステム	2 TB/1 セット (グループ当たり 、一般申込の場合)	作業領域
高速ファイルシステム (一時領域)	共有ファイルシステム 割り当て容量の 25 分の 1 (グループ当たり)	作業領域 (高速な入出力用)
高速ファイルシステム (恒久領域)	申込制	作業領域 (高速な入出力用)
ログインノード 共有ファイルシステム	50 GB (ユーザ当たり)	ホームディレクトリ

6. 利用状況

6.1 ノード利用率

図6は、試験運用期間(2021年5月14日から7月29日まで)と正式運用期間(8月2日から)にかけてのOdyssey、Aquariusそれぞれのシステム利用率と登録者数の推移である^[14]。

図6 Wisteria/BDEC-01 システム利用率と登録者数^[14]

利用負担金のかからない試験運用期間においては、利用率及び登録者数共に増加傾向であり、利用負担金の発生する正式運用期間に入ると大きく落ち込む様子となった。

また、図6からAquariusの方がOdysseyよりも利用率が高いように見えるが、Aquariusのregular-aキュー等の統計情報において、ジョブが使用するノード内での使用GPU数を8GPUと換算している箇所があり、Aquariusの方が高い利用であるように見えている点に注意を要する。ジョブにおいて実際に使用したGPU数でのデータを示せるよう検討を行っている。

6.2 バッチジョブ

図7, 8, はOdyssey、Aquariusそれぞれにおけるジョブ処理件数を示したグラフである。各キューにおける件数を積み上げたグラフとなっている。

ジョブ処理件数で見た場合、OdysseyよりAquariusの方が全体的に件数が多く、Odysseyではsmall-oキュー(1から144ノード)、Aquariusではshort-aキュー(8～16GPU)

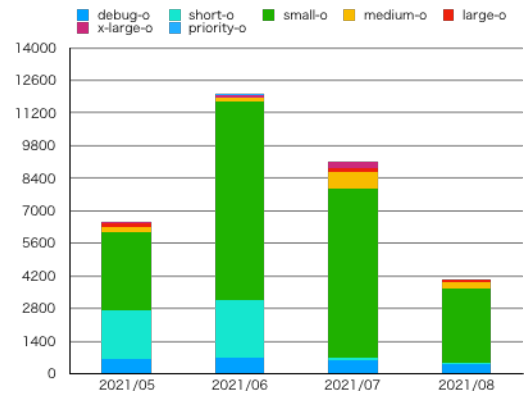


図7 ジョブ処理件数 (Odyssey)

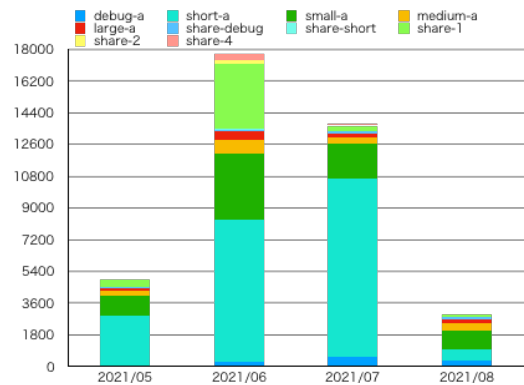


図8 ジョブ処理件数 (Aquarius)

が使われる傾向となっていた。shareキューについては、6月にshare-1キューの件数が伸びているが、その他のshareキューの利用は目立ってはいない。

図9, 10は、Odyssey、Aquariusそれぞれにおけるノード時間積、GPU時間積で見た場合の、キューごとの割合を示したグラフである。

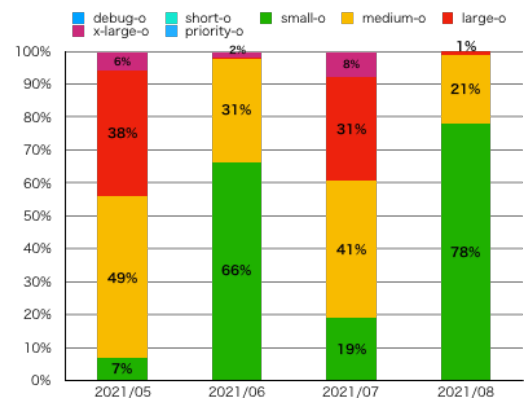


図9 ジョブ実行ノード時間積 (Odyssey)

Odyssey、Aquarius共に試験運用期間では大きな規模のジョブ(medium-o, large-o: 145～1152ノード、medium-a, large-a: 24～64GPU)が実行され、正式運用に入ると規模の大きいジョブは控えられた傾向が見えている。shareキューの利用はほとんど目立たない結果となった。

OdysseyにおけるFX1000では、順調に試用を行ったこ

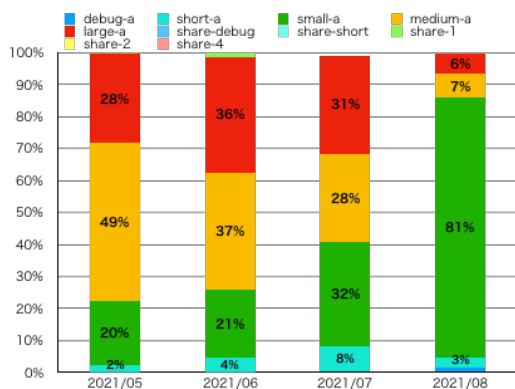


図 10 ジョブ実行 GPU 時間積 (Aquarius)

とが推測される。一方で、Aquarius では、利用者が、試験運用期間を通して試行錯誤を頻繁に行っていたことが期待される。また、利用が目立っていない share キューについては、その有用性自体について、利用者に伝わっていないことが推測される。share キューは、トークン消費を抑える上で有用であることを利用者に明確に案内していく予定である。

6.3 利用者状況

図 11, 12 は、実行ジョブのノード時間または GPU 時間について、利用者の分野別に割合を示したグラフである。

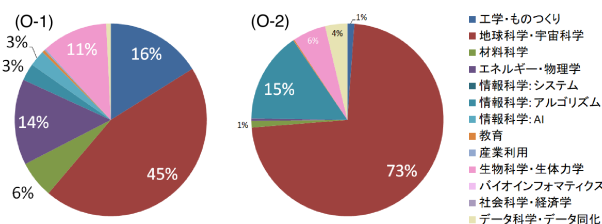


図 11 実行ジョブノード時間の分野別比率 (Odyssey)、(O-1) 2021 年 5 月～7 月 (試験運用期間)、(O-2) 2021 年 8 月 (正式運用)

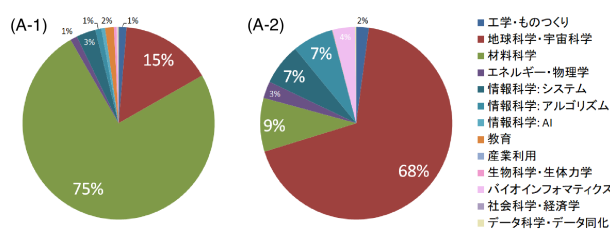


図 12 実行ジョブ GPU 時間の分野別比率 (Aquarius)、(A-1) 2021 年 5 月～7 月 (試験運用期間)、(A-2) 2021 年 8 月 (正式運用)

Odyssey では、試験及び正式運用 (8 月) を通して、45 % 以上が地球科学・宇宙科学分野の利用であるが、Aquarius では、試験運用期間では 75 % が材料科学分野であった。正式運用データは 8 月の一ヶ月分のみのため、今後とも利用分野別比率については分析を進める。一方で、データ科学・生物科学関連分野では、Aquarius の利用率は Reedbush-H/L と比べてかなり低い割合である。当該分野の Reedbush から Aquarius への移行が進んでいない可能性がある。

7. 今後の展開

Wisteria/BDEC-01 は、「S+D+L」融合を実現するためのプラットフォームの 1 号機として正式運用を迎えることができた。今後は、先に述べたように、より深いレベルでの Odyssey と Aquarius の融合を目指して研究開発を行うとともに、実 GPU 時間による統計分析や share キューの普及に関する課題を含め、GPU の利用向上や、アプリケーションの拡充、より利用しやすい環境の整備などの課題に取り組みながら、運用に努めていく予定である。

謝辞 本論文を作成するにあたって、様々なご教示をいただいた情報基盤課、情報基盤センターの関係者の皆様に深く感謝致します。また、システム構築及びその後の運用に携わっていただいている富士通株式会社の方々、加えて、試験運用期間において、Wisteria/BDEC-01 を試用し、様々な内容の指摘や質問を行ってくださった多くの利用者様に深く感謝致します。この場を借りてお礼を申し上げます。

参考文献

- [1] 東京大学情報基盤センター: <https://www.itc.u-tokyo.ac.jp/> .
- [2] 東京大学情報基盤センタースーパーコンピューティング部門: <https://www.cc.u-tokyo.ac.jp/> .
- [3] Wisteria/BDEC-01 スーパーコンピュータシステム: <https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/> .
- [4] 中島研吾, 堀 敏博, 下川辺隆史, 伊田明弘, 芝 隼人, 三木洋平, 星野哲也, 有間英志, 河合直聡, 坂本龍一, 岩下武史, 八代尚, 長尾大道, 松葉浩也, 荻田武史, 片桐孝洋, 古村孝志, 鶴岡 弘, 市村 強, 藤田航平, 近藤正章: 「計算・データ・学習」融合スーパーコンピュータシステム「Wisteria/BDEC-01」の概要, 情報処理学会研究報告 (2021-HPC-179-01) 2021 .
- [5] Reedbush スーパーコンピュータシステム: <https://www.cc.u-tokyo.ac.jp/supercomputer/reedbush/service/> .
- [6] OpenFOAM: <https://www.openfoam.com/> .
- [7] MateriApps: <https://ma.issp.u-tokyo.ac.jp/> .
- [8] 革新的ハイパフォーマンズ・コンピューティング・インフラ (HPCI): <http://www.hpci-office.jp/> .
- [9] ppOpen-HPC: <https://github.com/Post-Peta-Crest/ppOpenHPC> .
- [10] h3-Open-BDEC: <https://h3-open-bdec.cc.u-tokyo.ac.jp/> .
- [11] Iwashita, T, Nakajima, K., Shimokawabe, T., Nagao, H., Ogita, T., Katagiri, T., Yashiro, H., Matsuba, H.: *h3-Open-BDEC: Innovative Software Platform for Scientific Computing in the Exascale Era by Integrations of (Simulation + Data + Learning)*, Project Poster, ISC-HPC 2020 .
- [12] NVIDIA MIG(NVIDIA Multi-Instance GPU): <https://www.nvidia.com/ja-jp/technologies/multi-instance-gpu/> .
- [13] 住元真司, 荒川 隆, 坂口吉生, 松葉浩也, 八代 尚, 堀 敏博, 中島研吾: WaitIO-Socket:異種システム上の複数 MPI プログラムを結合する通信ライブラリの試作: 情報処理学会研究報告 (2021-HPC-181-07) 2021 .
- [14] 東京大学情報基盤センター: 東京大学情報基盤センタースーパーコンピューティングニュース: 6・7 月のジョブ統計, Vol.23 No.5 pp.15-16 (2021) .