

Oakbridge-CX スーパーコンピュータシステムの 運用状況と利用者支援の取り組み

山田 新¹⁾, 宮寄 洋¹⁾, 前田 光教¹⁾, 山本 和男¹⁾, 佐島 浩之¹⁾,
佐藤 孝明¹⁾, 福沢 秋津¹⁾, 中張 遼太郎¹⁾, 中島 研吾²⁾, 埴 敏博²⁾

1) 東京大学 情報システム部 情報基盤課

2) 東京大学 情報基盤センター

yamada@cc.u-tokyo.ac.jp

The operation of Oakbridge-CX Supercomputer System and user support efforts

Hajime Yamada¹⁾, Hiroshi Miyazaki¹⁾, Mitsunori Maeda¹⁾, Kazuo Yamamoto¹⁾,
Hiroyuki Sajima¹⁾, Takaaki Sato¹⁾, Akitsu Fukuzawa¹⁾, Ryotaro Nakahari¹⁾,
Kengo Nakajima²⁾, Toshihiro Hanawa²⁾

1) Information Technology Group, Information Systems Department, The University of Tokyo

2) Information Technology Center, The University of Tokyo

概要

2019 年 7 月より運用を開始した Oakbridge-CX スーパーコンピュータシステムについて
運用状況と利用者支援の取り組みについて報告する。

1 はじめに

東京大学情報基盤センター[1] (以下、本センター) では、大規模超並列スーパーコンピュータシステム (Oakbridge-CX) [2][3]を導入し、2019 年 7 月から試験運用を開始し[4]、2019 年 10 月から正式サービスを開始している。

本稿では、Oakbridge-CX についての概要と、運用開始から 2020 年 9 月下旬までの運用状況、同期間に行われたプリポスト予約システムの整備について紹介する。



図 1. Oakbridge-CX の外観

2 システム概要

2.1 ハードウェア

Oakbridge-CX は Intel® Xeon® Platinum 8280(開発コード名: CascadeLake)を搭載したスーパーコンピュータシステムである。計算ノードは 1,368 台で、そのうち 128 台に SSD を搭載している。計算ノード単体の理論演算性能は 4.8384TFLOPS、主記憶は DDR4 192GiB であり、全体では 6.61PFLOPS の性能、256.5TiB の主記憶容量を有している。サーバ 10 台をログインノードとしてサービスに提供している。並列ファイルシステムとして 12.4PB の Lustre ファイルシステムを備える(表 1、表 2)。

表 1. システム全体諸元 (計算ノード)

総理論演算性能		6.61 PFLOPS
総ノード数		1,368 (内 SSD 搭載 128)
総主記憶容量		256.5 TiB
ネットワーク トポロジー		フルバイセクションバンド幅 Fat Tree
並 列 シ ス テ ム フ ァ イ ル	システム名	DDN ExaScaler (Lustre ベースファイルシステム)
	サーバ(OSS)	DDN ES18K
	サーバ(OSS)数	2
	容量	12.4 PB
	転送速度	193.9 GB/sec

表 2. ノード諸元 (計算ノード)

項目		計算ノード	計算ノード (SSD 搭載)
マシン名		FUJITSU PRIMERGY CX2550 M5	FUJITSU PRIMERGY CX2560 M5
ノード数		1240	128
C P U	プロセッサ名	Intel® Xeon® Platinum 8280 (開発コード名 : CascadeLake)	
	プロセッサ数	2 (28+28 コア)	
	周波数	2.7 GHz	
	理論演算性能	4.8384 TFLOPS	
メ モ リ	容量	192 GiB (DDR4)	
	メモリバンド幅	281.5 GB/sec	
インターコネクト (通信性能)		Intel® Omni-Path ネットワーク (100 Gbps)	
S S D	容量	-	1.6 TB (NVMe)
	読み出し性能	-	3.20 GB/sec
	書き込み性能	-	1.32 GB/sec

2.2 SSD 搭載ノード

Oakbridge-CX は、全 1,368 ノードの内、128 ノードに SSD を搭載しており、1 ジョブで最大 112 ノードまで利用可能である。1 ノードの SSD 容量および読込性能、書込性能はそれぞれ 1.6TB、3.2GB/s、1.32GB/s であり、112 ノードでは 179.2TB、358.4GB/s、147.84GB/s である。

なお、SSD 搭載ノードは、ジョブ実行時の一時ファイルの置き場所としての利用を想定しており、ジョブ終了時にはデータがクリアされる。ジョブ終了後も利用するデータは、ファイルステージング機能を用いて各自の領域へ退避する必要がある。

また、各計算ノードの SSD を個別に使用するだけでなく、ジョブ割り当ての都度 BeeGFS on

Demand (BeeOND)により 200TB 以上の単一の並列共有ファイルシステム(/dist)を構成して利用することも可能である。

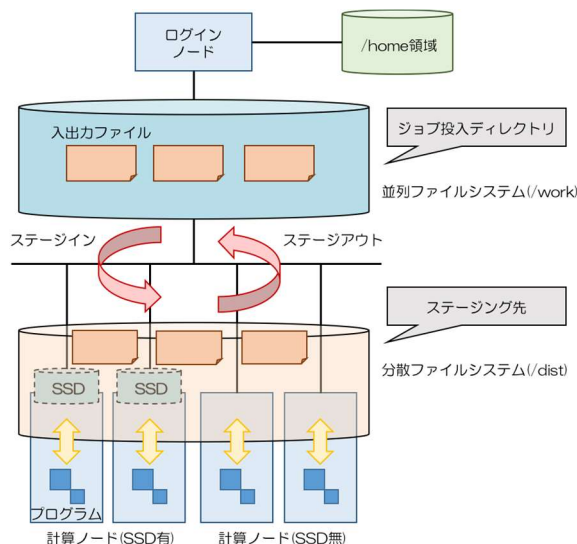


図 2. ステージング概要

2.3 外部接続ノード

SSD 搭載ノードのうち 16 ノードは、外部接続ノードとして用意されており、現在 5 ノードをノード固定利用として資源提供中である。外部接続ノードは、学術情報ネットワーク(SINET)を経由して、外部計算資源(サーバ、ストレージ、センサーネットワーク等)に直接接続が可能であるため、大量のデータをリアルタイムに処理する場合に威力を発揮する。実際の活用例として、地震観測装置やデータアーカイブと連携した利用が行われている。

外部接続ノードの運用には、システムの構成やセキュリティ等、通常の計算ノードとは異なった配慮が必要であり、運用にあたって得られた知見は本センターへ導入を予定している「計算・データ・学習」融合スーパーコンピュータシステムや mdx (調達システム名称:データ活用社会創成プラットフォーム基盤システム)へ還元していく予定である。

2.2 ソフトウェア

表 3 に示すとおり、多数のオープンソフトウェアを事前に用意している。ライブラリ、アプリケーションは東京大学で開発しているソフトウェアを導入している。

表 3. ソフトウェア一覧

OS	Red Hat Enterprise Linux 7, CentOS 7
言語処理系	Intel® Parallel Studio XE Cluster Edition, GCC コンパイラ, OpenJDK
メッセージ通信ライブラリ	Intel® MPI, Open MPI, Intel® Omni-Path Fabric Software
ライブラリ	Intel® Parallel Studio XE Cluster Edition (MPI, MKL (BLAS, CBLAS), IPP DAAL), HDF5, FFTW, METIS, MT-METIS, ParMETIS, NetCDF, Parallel netCDF, PETSc, SuperLU, SuperLU MT, SuperLU DIST, Scotch, PT-Scotch, Xabclib, GNU Scientific Library, ppOpen-HPC, ppOpen-AT, MassiveThreads, Trillinos, CMake, Python, LAPACK, ScaLAPACK, Anaconda
アプリケーション	mpijava, OpenFOAM, ABINIT-MP, PHASE, FrontFlow/blue, FrontISTR, REVOCAP-Coupler, REVOCAP-Refiner, OpenMX, xTAPP, AkaiKKR, MODYLAS, ALPS, feram, GROMACS, BLAST, R packages, Bioconductor, BioPerl, BioRuby, BWA, GATK, SAMtools, Quantum ESPRESSO, Xcrypt, Paraview, VisIt, POV-Ray
デバッグプロファイラ	Arm DDT, Intel® Parallel Studio XE Cluster Edition (Vtune Amplifier XE, Advisor, Inspector, Trace Analyzer & Collector)

2.3 商用ソフトウェア

商用の科学技術計算用アプリケーションとして、総合 CAE プラットフォーム Altair HyperWorks を提供中である。本センター所有のライセンスで、国内アカデミックユーザ(大学、短大、大学校、高専等に所属)はライセンス料無料で利用可能であり、HyperMesh、HyperView などクライアント PC で動作するソフトも無料で利用できる。その他のユーザ(企業や研究機関に所属)は本センター所有のライセンスでは、利用できないが、ライセンスを個別に購入することで利用可能となる。

3 利用制度

3.1 トークンと消費係数

一般利用制度で Oakbridge-CX の利用を申し込むと、申込時のセット数に応じて、計算ノードの利用可能時間である「トークン（1セット当たり 8,640 ノード時間）」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（一部のコースを除き）利用できるノード数制限などはなく、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。トークンはバッチジョブの実行ごとに「経過時間×ノード数×消費係数」で計算される量が消費され、トークンを使い切るとバッチジョブの投入ができなくなる。

表 4. Oakbridge-CX 一般利用制度概要

項目		利用負担金、他	
パーソナルコース	一般申込	基本セット	【大学・公共機関等 100,000 円】
		トークン	(最大3セットまで) 8,640 ノード時間 (1 ノード/年) 消費係数 1.0
		ディスク容量	4TB
グループコース	一般申込	基本セット	【大学・公共機関等 100,000 円】
		トークン	(複数セット申込可) 8,640 ノード時間 (1 ノード/年) 消費係数 1.0
		ディスク容量	4TB
グループコース	ノード固定(要審査)	基本セット	【大学・公共機関等 150,000 円】
		トークン	(複数セット申込可) 8,640 ノード時間 (1 ノード/年) 消費係数 1.0
		ディスク容量	4TB

3.2 ジョブキューと制限値

Oakbridge-CX では、表 5 に示すようなジョブクラスを用意している。

regular キューは、最大 256 ノード、最長 48 時間(x-large は 24 時間)までのジョブが実行可能であり、ノード数により small, medium, large, x-large の各キューに自動的に振り分けられる。さらに、実行時間が 8 時間までの short キュー(最大 8 ノードまで)、30 分までの debug キュー(16 ノードまで)、を提供している。

インタラクティブ実行用のキューは最大 8 ノードまで利用可能であり、トークンは消費されない。このほか講習会や講義利用については専用のキューを用意して対応している。

なお、SSD 搭載ノードは regular キューのみで使用可能である。

表 5. ジョブクラス制限値

キュー名	ノード数	制限(経過)時間
debug	1-16	30min
short	1-8	8h
regular		
(small)	1-16	48h
(medium)	17-64	48h
(large)	65-128	48h
(x-large)	129-256	24h
interactive)		
(interactive_1)	1	2h
(interactive_8)	2-8	10min

4. プリポストキュー予約システム

4.1 プリポストキュー

プリポスト (prepost) キューは、可視化などの解析データの前処理や後処理のために、インタラクティブなジョブを実行することができるキューで、ノード構成は計算ノードと同様である。しかし、このプリポストキューは他のキューと利用方法が異なり、ジョブを直接投入することはできず後述のプリポストキュー予約システムから予約して利用する必要がある。

4.2 プリポストキュー予約システム

プリポストキュー利用のために必要なプリポストキュー予約システムを 2019 年 9 月に整備し、2019 年 10 月より正式提供している。プリポストキュー予約システムを使うことで、表 6 に示す条件の範囲内で、希望の時間帯にプリポストキューの資源を予約して利用することができる。このよ

うな予約利用は、Oakbridge-CX 利用支援ポータル内の「プリポスト予約」から、必要事項を入力した上で予約申込を行うことで利用が可能である。

表 6. プリポストキュー予約条件

利用時間	1h から 6h (1h 単位)
利用ノード数	1 か 8 ノード(1ノード単位)
予約期間	予約日時の 15 分後から 60 日以内
予約数上限	実行中を含め 5 件

プリポストキュー予約システムは、図 3 の通り 2019 年度中は利用が低調であったが、2020 年度に入ってから徐々に利用が増えつつある。しかし、利用増加の余地は十分あるため、認知度向上に努めていきたい。

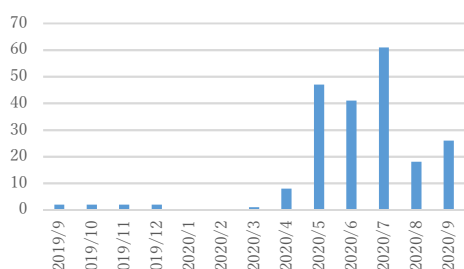


図 3. プリポストキュー予約システム
月別予約件数

5 利用状況

5.1 ノード利用率

Oakbridge-CX の利用率は、正式サービス直後は、比較的低い状況が続いていたが、その後は増加傾向にあり、直近の数ヶ月では 7 割程度の高い水準の利用率を維持している。

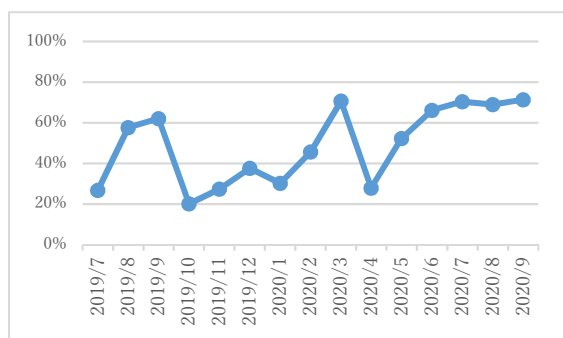


図 4. ノード利用率

5.2 バッチジョブ

Oakbridge-CX のキュー別ノード時間積割合を

図 5 に示す。small キューの占める割合が大きい
が、large, x-large キューが割当てられるような、
並列度の高いジョブ利用も相当数存在している。

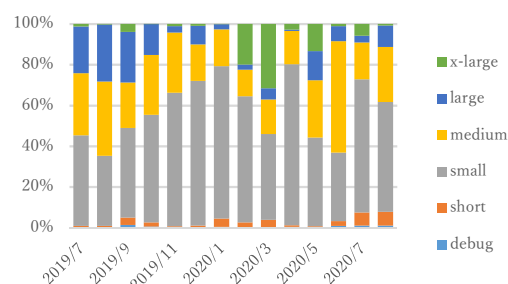


図 5. キュー別ノード時間積割合

5.3 利用分野

Oakbridge-CX で利用されている研究の分野は図 6、図 7 に示す通りである。従来から本センターのスーパーコンピュータシステムの利用が多い物理・工学系の分野が大きな割合を占めているが、データ科学・データ同化の分野でも多く利用されている。

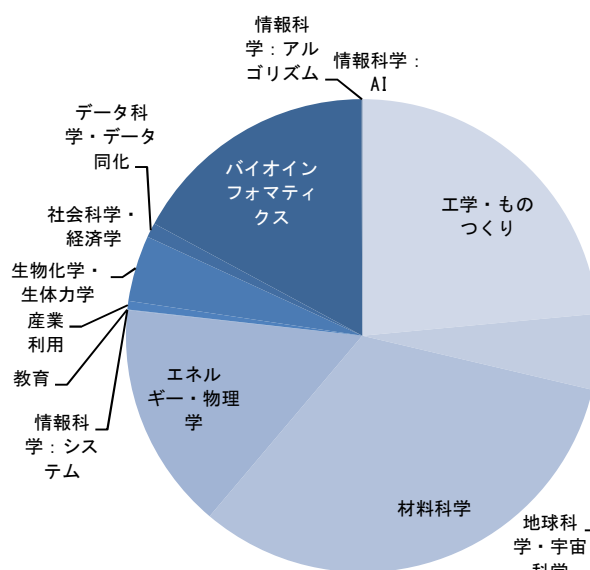


図 6. 研究分野別ノード時間積割合
(2019 年 10 月～2020 年 3 月)

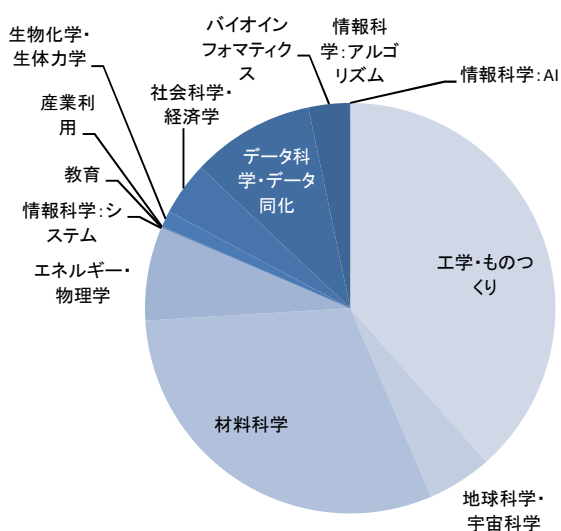


図 7. 研究分野別ノード時間積割合
(2020 年 4 月～2020 年 9 月)

6 まとめと今後の課題

本稿では、2019 年 7 月より運用を開始した Oakbridge-CX スーパーコンピュータシステムについて運用状況と利用者支援の取り組みを報告した。外部接続ノードや SSD 搭載ノードを備えた Oakbridge-CX は、データ同化など本センターの伝統的な利用分野とは異なる分野での利用が広がっている。

今後の課題としては、より安定した運用とさらなる利用拡大があり、具体的には以下のことが挙げられる。

- ・ 利用向上に向けた取り組み（講習会、手引書の充実）
- ・ 新型コロナウイルス感染による活動制限下での大規模 HPC チャレンジの運用
- ・ コンパイラやアプリケーションの更新拡充
- ・ SSD 搭載ノード活用法の周知
- ・ プリポストキュー予約機能の認知度向上

これらの課題に取り組みながら、より使いやすいシステムを目指して運用に努めていく予定である。

参考文献

- [1] 東京大学情報基盤センター
<http://www.cc.u-tokyo.ac.jp/>
- [2] Oakbridge-CX スーパーコンピュータシステム,
<http://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/>
- [3] 富士通株式会社 坂口 吉生、「大規模超並列

スーパーコンピュータシステム Oakbridge-CX の特長」、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、Vol.21, No.4 pp.23-35 2019.

- [4] 東京大学情報システム部 「Oakbridge-CX スーパーコンピュータシステム運用開始のお知らせ」、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、Vol.21, No.4 p.5 2019.