Oakbridge-CX スーパーコンピュータシステムの運用

下條 清史, 宮嵜 洋, 田川 善教, 山本 和男,
佐島 浩之, 佐藤 孝明, 中張 遼太郎, 山田 新

東京大学大学 情報システム部 情報基盤課 スーパーコンピューティングチーム

shimojo@cc.u-tokyo.ac.jp

The operation of Oakbridge-CX Supercomputer System

Kiyofumi Shimojo, Hiroshi Miyazaki, Yoshiyuki Tagawa, Kazuo Yamamoto,
Hiroyuki Sajima, Takaaki Sato, Ryotaro Nakahari, Hajime Yamada

Supercomputing Team, Information Technology Group,
Information Systems Department, 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の概要と運用状況について報告する。

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する。Oakbrigde-CX はシステム全 1,368 ノードの内、128 ノードに SSD を搭載しており、ステージングを含む高速で効率的なデータ入出力を実現し、大容量データ処理をともなうデータ同化処理等に威力を発揮する。

Oakbridge-CX は、2018 年 3 月を以って運用を終了した Oakleaf/Oakbridge-FX（Fujitsu PRIMEHPC FX10）の後継システムとして位置づけられ、最新の技術に基づく高い計算・通信・入出力性能及び安定性を備え、Oakleaf/Oakbridge-FX システムの他，Reedbush-U[5]をはじめとする既存システムからのプログラム類の移行を容易に実施し、高い性能を維持する環境を提供する。

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)を搭載した富士通社製のスーパーコンピュータである。CPU は計算ノードが 1,368 台で内 128 台に SSD を搭載し、計算ノード単体の理論演算性能は、4.8384TFLOPS、主記憶容量は、192GiB(DDR4)であり、全体では 6.61PFLOPS、256.5TiB の性能を有している。フロントエンドにはサーバ 10 台(FUJITSU Server PRIMERGY CX2560 M5)をログインノードとしてサービスに提供している。ファイルシステムには、並列ファイルシステムを 12.4PB 備える。(図 2 および、表 1、表 2)

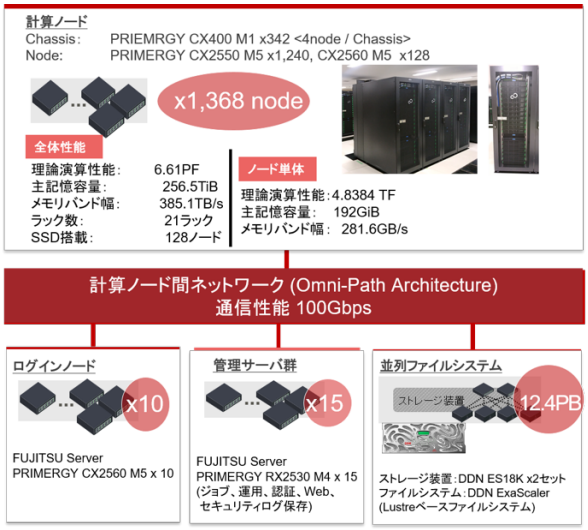

図 2. Oakbridge-CX 全体構成

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		FBB Fat Tree
並列ファイルシステム	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 (SSD 非搭載)	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
CPU	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)	
SSD	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 ソフトウェア

表 3 に示すとおり､多数の OSS を用意している。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. ソフトウェアー覧

OS	Red Hat Enterprise Linux 7、CentOS 7
言語 処理系	Intel® Pallalel 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® Pallalel 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、Vislt、POV-Ray
デバッガ プロファイラ	Arm DDT、Intel® Parallel Studio XE Cluster Editon(Vtune Amplifier XE、 Advisork、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[6] を提供予定である。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。

すでに Reedbush では HyperWorks を題材とした講習会が開催されており、Oakbridge-CX での開催も検討している。

3 運用形態

3.1 トークン制と利用コース

研究者個人単位で利用するための「パーソナルコース」、研究・グループ単位でまとまって利用するための「グループコース」によるサービスを行っている。

利用するコース、利用申込したノード数に応じて、計算ノードの利用可能時間である「トークン（ノード時間積)」を割り当て、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消費され、計算式は「経過時間×ノード数×消費係数」である。バッチジョブ実行において各コースで定められたノード数を超えると、超えた部分について消費係数が 2 倍となる。

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。この場合、払い出せる計算機資源に余裕があれば追加購入することができる。なお、トークンは利用期間内に消費できることを保証するものではなく、次年度への繰り越しや返金等はできない。

以上の方式は、既存システムである Reedbush や Oakforest-PACS[7]と基本的には同じである。

表 4. Oakbridge-CX 利用コース

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】 （申込口数 1 口当り、最大 3 口まで） トークン　8,640 ノード時間 （1 ノード/年） 並列実行ノード数　4 ノードまで消費係数 1.0 4 ノード超のとき消費係数 2.0 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 400,000 円】 （申込ノード 4 ノード当たり） トークン　34,560 ノード時間 （4 ノード/年） 並列実行ノード数　申込ノードまで消費係数 1.0 申込ノード超のとき消費係数 2.0 ディスク容量　16TB （申込ノード数 4 ノードあたり）
	ノード固定（要審査）	基本セット　【大学・公共機関等 600,000 円】 （申込ノード 4 ノード当たり） トークン　34,560 ノード時間 （4 ノード/年） 並列実行ノード数　申込ノードまで消費係数 1.0 申込ノード超のとき消費係数 2.0 ディスク容量　16TB （申込ノード数 4 ノードあたり）

3.1.1 ノード固定

グループコースのノード固定では、計算ノードを当該グループで専有して利用することができる。バッチジョブによる利用の他、商用プログラムや特殊なライブラリ等の利用、利用環境のカスタマイズ（インタラクティブ実行環境、ローカルディスク利用等）が可能である。専用ログインノードの設置も利用者負担により可能であり、高いセキュリティが求められる場合に有用である。申込に際しては審査がある。現在 1 グループが正式サービス開始後にノード固定を利用予定である。

3.1.2 トークン移行

本センターのスーパーコンピュータシステム相互にトークンの移行を可能とした。この仕組みにより新たな利用負担金が発生することなく他方のシステムを利用することができる。トークンの移行にあたっては表 5 の換算率により移行先のトークン量、表 6 の基準ノード数、換算係数より移行先のトークン消費係数切替点(ノード)が決まる。

表 5. トークン換算率

トークン移行			換算率
Oakbridge-CX	→	Reedbush	1.3
Oakbridge-CX	→	Oakforest-PACS	2.0
Reedbush	→	Oakbridge-CX	0.75
Oakforest-PACS	→	Oakbridge-CX	0.5

表 6. 消費係数切替点換算

トークン移行（基準ノード数）			換算係数
Oakbridge-CX (4 ノード)	→	Reedbush (4 ノード)	6
Oakbridge-CX (4 ノード)	→	Oakforest-PACS (8 ノード)	8
Reedbush (4 ノード)	→	Oakbridge-CX (4 ノード)	3
Oakforest-PACS (8 ノード)	→	Oakbridge-CX (4 ノード)	4

3.2 ジョブキューと制限値

Oakbrigde-CX では、表 7 に表記したようなジョブクラスを用意している。

ノード数が 256 ノードまで使用可能で、実行時間が最大 48 時間(x-large は 24 時間)の regular キュー(ノード数により small, medium, large, x-large の各キューに振り分けられる)と実行時間の短い debug キュー(16 ノードまで)、short キュー(8 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 8 ノードまで用意しており、このキューではトークンが消費されない。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ほか、可視化などの解析データの前処理や後処理のために、ログインノードと同じ構成のノードで対話的実行する prepost キューを提供する予定であり、利用支援ポータルからの予約利用形式での提供を予定している。

なお、SSD 搭載ノードは regular キューのみで使用可能である。

表 7. ジョブクラス制限値(正式サービス)

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

3.3 ファイルシステム

Oakbrigde-CX にはログイン・計算ノードの双方から利用できる並列ファイルシステム（/work）とログインノード専用のファイルシステム(/home)がある。/work の容量は申込コースにより基本量

が決まり、増量が可能である。/home は各ユーザ一律に 50GB を上限としており、バッチジョブからの利用はできない。

3.4 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、SSD 搭載ノードは regular キューのみで使用可能で、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。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、ジョブ終了時にデータがクリアされるため各自の領域への退避が必要である。

また、各計算ノードの SSD を個別に使用するだけでなく、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、ファイルステージング機能を有する。

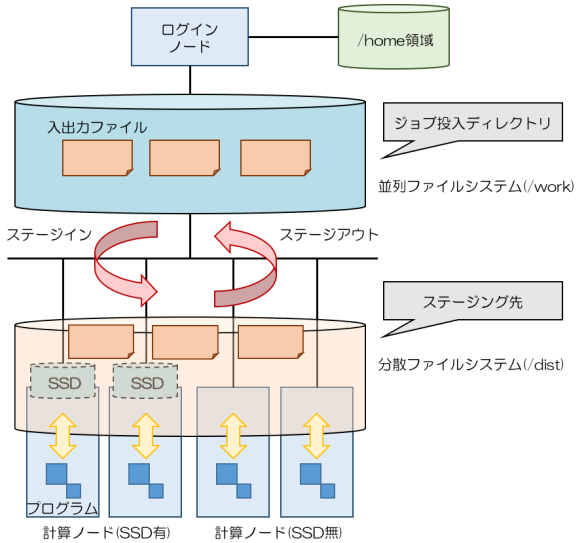

図 3. ステージング概要

SSD 搭載ノードのうち 16 ノードは、外部接続ノードとして使用することとしており、一例としては、VLAN を使用して SINET 経由で特定の外部ネットワークと直結し、遠隔でリアルタイムなデータ処理を実現する計画がある。データ利活用型の利用形態が求められており、個別のニーズにも対応していく予定である。

3.5 大規模 HPC チャレンジ

2019 年 12 月より、「大規模 HPC チャレンジ」を実施する。「大規模 HPC チャレンジ」は、Oakbridge-CX が持つ最大規模の計算ノード数である 1,280 ノード(内 SSD 搭載 112 ノード)を、最大 24 時間、1 研究グループで計算資源の専有利用ができる公募型プロジェクトである。月に 1 回、午前 9 時から翌日の 9 時までの 24 時間、1,280 ノードを使用できる環境に切り替える。年数回公募を行い、課題審査で採択されると利用することができる。正式サービス開始後に公募を開始する予定である。

3.6 その他

令和元年度 10 月からは HPCI（革新的ハイパフォーマンス・コンピューティング・インフラ）および JHPCN（学際大規模情報基盤共同利用・共同研究拠点）に資源を拠出予定である。令和元年度は、HPCI で採択された 3 課題に Oakbridge-CX の資源を 133,680 ノード時間提供し、JHPCN で採択された 5 課題に Oakbridge-CX の資源を 101,837 ノード時間提供する予定で、令和 2 年度は通年で資源を拠出する予定である。

一般利用に加えて教育利用や企業利用、若手・女性利用、トライアルユース（有償・無償・無料体験）についても正式サービス開始後に順次募集を開始する予定である。

4 利用状況

4.1 試験運用期間

Oakbridge-CX は、2019 年 7～9 月の予定で試験運用を行なっており、試験運用期間中はジョブの実行時間制限を正式サービスの 1/4 程度の時間とし、ユーザに開放している。システムの動作に問題がないか、リソースの制限値など設定が適切か、ユーザの従来機からの移行など検証を行っている段階であるが、9 月途中時点では、日別最大 74.57%の利用率を記録している。10 月 1 日からは正式サービスを開始する予定である。

4.2 利用申込

2019 年 9 月中旬時点での試験運用登録ユーザ数はグループコースが 43 グループ 277 ユーザ、パーソナルコースが 31 ユーザである（本センタースタッフを除く）。その内、2019 年 9 月中旬時点で、正式サービス開始後も継続利用予定のユーザ数は、2019 年 9 月中旬の時点で、グループコースが 24 グループ 145 ユーザ、パーソナルコースが 13 ユーザとなっており、正式サービス開始に向けて増加傾向にある。正式サービス開始後は、企業利用、HPCI および JHPCN への資源提供を開始し利用者数の増加が見込まれる。講義やセンターが

開催している講習会での教育利用も提供予定である。

利用者へのトークンの払い出しについては年間総量の 120% までの払い出しを上限としている。

5 今後の課題

今後の課題として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）
- 12 月から実施予定の大規模 HPC チャレンジの安定運用
- 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
- prepost キュー予約機能の提供

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.

[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] Altair HyperWorks http://altairhyperworks.jp/

[7] Oakforest-PACS スーパーコンピュータシステム, https://www.cc.u-tokyo.ac.jp/supercomputer/ofp/service/