# 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

富士通株式会社
テクニカルコンピューティング・ソリューション事業本部
計算科学ソリューション統括部） 坂口 吉生

## 1. システムの概要

大規模超並列スーパーコンピュータシステム Oakbridge-CX*1 は、「FUJITSU Server PRIMERGY CX2550 M5 x1240 ノード」、および「FUJITSU Server PRIMERGY CX2560 M5 x128 ノード」を中核とした、総理論演算性能(倍精度浮動小数点演算性能) 6.6 PFLOPS のスーパーコンピュータシステムです。計算ノードは、Intel® Xeon® Platinum 8280(コードネーム:Cascade Lake)を搭載し、筐体当たり性能と消費電力当たり性能(高密度実装・省電力)に重点を置いた HPC 計算サーバです。各計算ノードは Intel® Omni-Path Architecture (OPA)により100Gbps 性能のフルバイセクションバンド幅で接続し、ノード間通信とファイル共有を行ないます。並列ファイルシステムのストレージ容量は 12.4PB で、データ転送帯域は 193.9GB/s です。システムの全体構成概要を、「図 1-1 Oakbridge-CX システム概要図」に記載します。

Oakbridge-CX は、ジョブスケジューラ(Technical Computing Suite)により、計算資源を最適化した効率的なジョブ運用でペタスケール規模において高いアプリケーション実効性能とスケーラビリティを実現します。Oakbridge-CX は、2019 年 6 月時点で TOP500 HPL *2で世界 45 位(日本 8 位)、Green500 *3で世界 28 位(日本 4 位)、HPCG *4で世界 48 位(日本 12 位)にランキングされています。

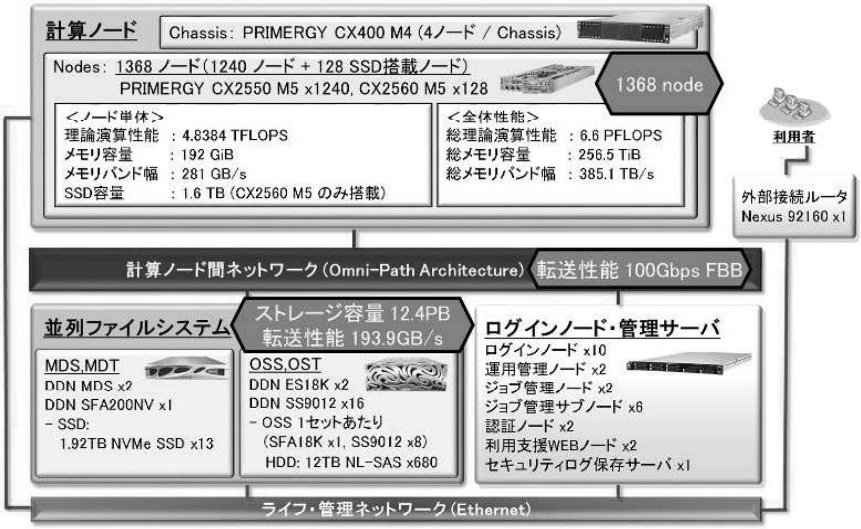



図 1-1 Oakbridge-CX システム概要図

*1 https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/
*2 https://www.top500.org/system/179606
*3 https://www.top500.org/green500/list/2019/06/
*4 https://www.top500.org/hpcg/lists/2019/06/

システム全体、ファイルシステム、計算ノードのシステム仕様を、「表 1-1 Oakbridge-CX システム仕様」に、ソフトウェア構成を、「表 1-2 Oakbridge-CX ソフトウェア一覧」に記載します。

表 1-1 Oakbridge-CX システム仕様

| 項目 | 仕様 | |
|---|---|---|
| 倍精度理論演算性能 | ノード単体：4.8384 TFLOPS (28 コア x 2.7GHz x 32fp x 2 CPU)<br>システム全体：6.6 PFLOPS (4.8384 TFLOPS x (1,240+128)計算ノード) | |
| 主記憶容量 | ノード単体：192 GiB (DDR4 16 GiB 2933MHz x 12 チャネル)<br>システム全体：256.5 TiB (192 GiB x (1,240+128)計算ノード/1024) | |
| メモリバンド幅 | ノード単体：281.568 GB/s (2933MHz x 8 バイト x 12 チャネル)<br>システム全体：385.1 TB/s (281.568 GB/s x (1,240+128)計算ノード/1000) | |
| インターコネクト | Intel® Omni-Path Architecture　　トポロジー：FBB Fat Tree | |
| ファイルシステム | 並列ファイルシステム | ストレージ装置「ES18K」<br>ディスクエンクロージャ装置「SS9012」 |
| | 実効利用可能容量 | 12.4 PB |
| | データ転送帯域 | 193.9 GB/s |
| | 物理構成 | MDS x 2<br>MDT(SFA200NV) x 1<br>OST コントローラ(ES18K) x 2<br>OST エンクロージャ(SS9012) x 16 |
| | インターコネクト | Intel® Omni-Path Architecture |
| 計算ノード | 計算ノード数/CPU 数 | 1,240 ノード / 2,480 CPU　　[CX2550M5] |
| | | 128 ノード / 256 CPU　　[CX2560M5] |
| | シャーシ台数 | PRIMERGY CX400M4 310 シャーシ [CX2550M5] |
| | | PRIMERGY CX400M4 32 シャーシ [CX2560M5] |
| | SSD 容量・性能<br>[CX2560M5] | NVMe SSD 1.6TB<br>読み出し性能：3.20GB/s<br>書き込み性能：1.32GB/s |
| | Omni-Path インタフェース | Omni-Path HFA カード x 1 |
| | Omni-Path 転送速度 | 片方向 12.5GB/s(全二重) |

表 1-2 Oakbridge-CX ソフトウェア一覧

| 項目 | 計算ノード | ログインノード |
|---|---|---|
| OS | CentOS 7 | Red Hat Enterprise Linux 7 |
| 言語処理系 | Intel® Parallel Studio XE Cluster Edition , GCC コンパイラ , OpenJDK | |
| ライブラリ | Intel® Parallel Studio XE Cluster Edition (MPI , MKL , IPP DAAL) | |
| | HDF5, FFTW, METIS, MT_METIS, ParMETIS, NetCDF, Parallel netCDF, PETSc, SuperLU, SuperLU_MT, SuperLU_DIST, Scotch, PT-Scotch, Xabclib, GSL, ppOpen-HPC, ppOpen-AT, MassiveThreads,Trilinos, CMake, Python, Anaconda | |
| デバッガ・ツール | Arm DDT | |
| | Intel® Parallel Studio XE Cluster Edition<br>(VTune Amplifier XE , Advisor , Inspector , Trace Analyzer & Collector) | |
| アプリケーション | mpijava, OpenFOAM, ABINIT-MP, PHASE, FrontFlow/blue, FrontISTR, REVOCAP, OpenMX, xTAPP, AkaiKKR, MODYLAS, ALPS, feram, GROMACS, BLAST, R Packages, Bioconductor, BioPerl, BioRuby, BWA, GATK, SAMtools, Quantum ESPRESSO, Xcrypt, ParaView, VisIt, POV-Ray | OpenFOAM, Globus Toolkit, Gfarm |
| ファイルシステム | 並列ファイルシステム ：DDN Lustre<br>分散ファイルシステム ：BeeGFS On Demand | |
| ジョブ管理 | 富士通 HPC ミドルウェア ：Technical Computing Suite | |

# 2. ハードウェアの特長

## 2.1 計算サーバの特長

### 2.1.1 FUJITSU Server PRIMERGY CX400 M4/CX2550 M5, CX2560 M5

FUJITSU Server PRIMERGY CX400 M4 は 2U サイズのシャーシで、FUJITSU Server PRIMERGY CX2550 M5 および、CX2560 M5 を 4 台搭載できます。最新の Intel® Xeon CPU を搭載し、2U サイズに 4 台搭載という高密度設計で、柔軟なストレージ・ネットワーク I/O を兼ね備えたマルチノードサーバです。また、水冷方式によりサーバ内部の CPU クーラーの内蔵ポンプで CPU とメモリを冷却し、冷却ファンの回転数を下げることで消費電力を削減しています。また、ログインノードは計算ノードと同一 CPU を搭載した CX2560M5 です。

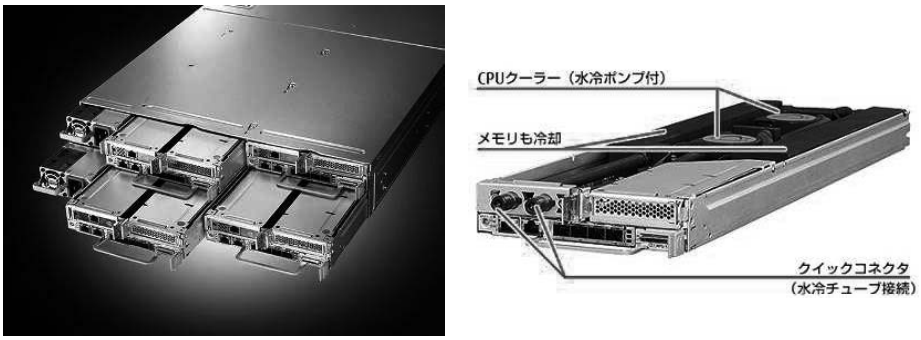



図 2-1 CX400 M4 および CX2550 M5/CX2560 M5 の外観イメージ

### 2.1.2 Intel® Xeon® Platinum 8280

Intel® Xeon® Platinum 8208 プロセッサは、ハイパフォーマンス・コンピューティング、高度な人工知能および分析に最適に設計されています。計算ノードの CPU は、64 ビット拡張された IA32 アーキテクチャに対応しており、多数の均一なコアがハードウェア共有メモリ(NUMA)によって接続されたメニーコアアーキテクチャです。各コアは並列処理性能を高める AVX-512(SIMD 演算)に対応しています。

表 2-1 Intel® Xeon Platinum 8280 の主な仕様

| 項目 | 説明 |
|---|---|
| 物理コア数 | 28 |
| 論理コア数（スレッド数）*5 | 56(1 コアあたり 2 スレッド) |
| プロセッサ・ベース動作周波数 | 2.70 GHz |
| ターボ・ブースト利用時の最大周波数 *6 | 4.00 GHz |
| CPU キャッシュ | 38.5 KB |
| UPI（Intel® Ultra Path Interconnect）リンク数 | 3 |
| 命令セット拡張 | Intel AVX-512 |
| 理論ピーク性能 | 2.4192TFLOPS |

*5 ログインノードおよび計算ノードは実効効率を考慮し Hyper Threading は無効化しています。
*6 アプリケーション特性、温度条件等により大きく異なります。

## 2.2 並列ファイルシステムの特長

### 2.2.1　DDN ES18K, DDN SS9012

並列ファイルシステムは、ストレージ装置「ES18K」、ディスクエンクロージャ装置「SS9012」で構成します。並列ファイルシステムは、計算ノード、ログインノードとファイルを共有します。DDN ES18K は、DDN SFA18KE に Lustre 並列ファイルシステムを内蔵した DDN 並列ファイルシステムソリューションです。DDN ES18K は、1 台あたり 49GB/s のスループット性能を持ち、2 つの MDS および複数の OSS で構成され、必要に応じてスループットを拡張することが可能なスケールアウト型ストレージシステムです。

MDS は「DDN MDS サーバ」を 2 台、MDT は「DDN SFA200NV」を 1 台で構成します。OSS/OST は、コントローラは「DDN ES18K」を 2 台、エンクロージャは「DDN SS9012」を 16 台で構成します。DDN ES18K は RAID 構成に Declustered RAID(DCR)を採用したストレージ装置です。DCR は複数のドライブをプール化し、プール内にデータ、パリティ、スペア領域を分散配置する新しいデータ保護方式です。並列ファイルシステムは 12.4PB(RAID6(8D+2P)相当のパリティ分散、NL SAS ディスク)の有効利用容量を有し、12TB のハードディスク 1,360 本で構成します。

計算ノード群から並列ファイルシステムへのストレージデータ転送速度は、ホストポート(Omni-Path)の転送速度である、193.9GB/秒です。

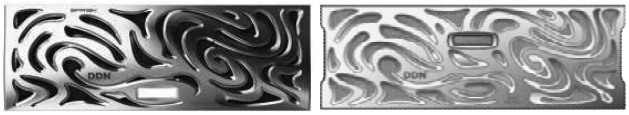


図 2-2 ES18KE、SS9012 外観

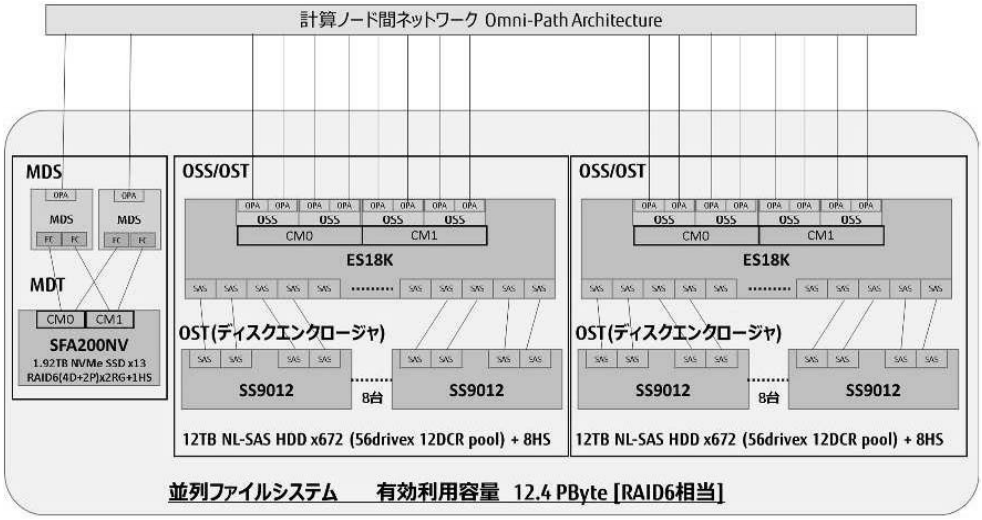



図 2-3 並列ファイルシステム構成図

## 2.3 インターコネクトの特長

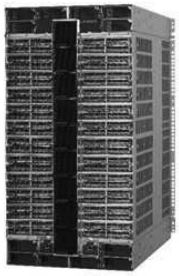
### 2.3.1 Intel® Omni-Path Architecture

計算ノード間通信・ファイル共有インターコネクトは、Intel® Omni-Path Architecture(Intel® OPA)です。インターコネクトネットワークを構成する Spine スイッチは、「Intel® Omni-Path Director Class Switches 100 Series」2 台で、Leaf スイッチは、「Intel® Omni-Path Fabric Edge Switches 100 Series」 64 台です。

Intel® OPA ネットワークは、Intel® OPA スイッチ「Intel® Omni-Path Edge Switch 100 Series」と「Intel® Omni-Path Director Class Switches 100 Series」で構成しており、スイッチと各ノード群の間は、Intel® Omni-Path Architecture で接続します。本インターコネクトネットワークは、計算ノード、並列ファイルシステム、ログインノード、管理サーバ間のデータ転送に利用されます。

データ転送速度はエンコード後の理論ピーク値で、1 ノードあたり 100Gb/s(12.5GB/s)です。全計算ノード間をフルバイセクションバンド幅(FBB)構成で結合しています。ログインノードと並列ファイルシステムは、計算ノード間ネットワークにより全計算ノードと等距離に接続し、それぞれのトラフィックが偏らないように均等に接続しています。また、ログインノードと並列ファイルシステム間のトラフィックが計算ノード間の通信への影響を最小限にした構成となっています。

Intel® OPA の装置外観を「図 2-2 ES18KE、SS9012 外観」、装置仕様を「表 2-2 Omni-Path Director Class Switches 100 Series と Edge Switches 100 Series の仕様」に記載します。



← 左図が Director Switch ↑上図が Edge Switch

図 2-4 Omni-Path Director Class Switches 100 Series と Edge Switches 100 Series の外観

表 2-2 Omni-Path Director Class Switches 100 Series と Edge Switches 100 Series の仕様

| 項目 | Edge Switch 100 (1 台あたり) | Director Switch 100 (1 台あたり) |
|---|---|---|
| サイズ(高さ) | 1U | 20U |
| 通信速度 | 100Gbps | 100Gbps |
| ポート形状 | QSFP28 | QSFP28 |
| ポート数 | 48 ポート | 816 ポート(48 port Leaf x 17slot)<br>1152 ポート(最大) |
| スイッチング容量 | 9.6Tbps | 19.2Tbps |

# 3. ソフトウェアの特長

## 3.1 コンパイラ・開発環境

### 3.1.1 Intel® Parallel Studio XE Cluster Edition の概要

プログラム開発環境は、Intel® Parallel Studio XE Cluster Edition です。Intel® Parallel Studio XE Cluster Edition は、HPC 開発者が直面する課題に対応し、HPC アプリケーションのパフォーマンスとスケーラビリティを大幅に向上させるための完全なツール群を提供します。実績あるクラスターツールと、高度なスレッド化/メモリ正当性検出ツールおよびパフォーマンス・プロファイリング・ツールを組み合わせることで、現在および将来の HPC クラスターシステム向けにスケーリングするアプリケーションの開発を実現します。

表 3-1 Intel® Parallel Studio XE 2019 Cluster Edition のコンポーネント一覧

| 主な用途 | ツール名 |
|---|---|
| Fortran コンパイラ | Intel® Parallel Studio XE Composer Edition for Fortran |
| C/C++コンパイラ | Intel® Parallel Studio XE Composer Edition for C++ |
| MPI(メッセージ・パッシング・ライブラリ) | Intel® MPI ライブラリ |
| 科学技術計算向けライブラリ | Intel® MKL |
| マルチメディア、データ処理、<br>通信アプリケーション向けライブラリ | Intel® IPP |
| マルチスレッド化テンプレート・ライブラリ | Intel® TBB |
| 解析、最適化用ツール | Intel® Trace Analyzer & Collector |
| パフォーマンス・プロファイラ | Intel® VTune Amplifier XE |
| メモリ/スレッドエラー検出ツール | Intel® Inspector XE |
| マルチスレッド化のアドバイスツール | Intel® Advisor XE |

### 3.1.2 コンパイラ

#### 3.1.2.1 Intel® Parallel Studio XE Composer Edition for Fortran

Intel® Parallel Studio XE Composer Edition for Fortran は、Co-Array Fortran 等の最新の技術に対応し、より高速なアプリケーションを開発するための Fortran コンパイラです。インテルプロセッサ・ベースのプラットフォームおよび他の互換プラットフォームにも対応する最適化をご提供します。また、機能拡張した GNU デバッガ(インテル GDB)も同梱されます。

Intel® Parallel Studio XE Composer Edition for Fortran は、以下の言語規格に対応しています。

- ISO/IEC 1539-1:2010 (Fortran 2008 規格の一部をサポート)
- ISO/IEC 1539-1:2004
- ISO/IEC 1539:1997
- ISO/IEC 1539:1991
- ISO 1539-1980
- OpenMP API V4.0 (OpenMP 4.0 の User-Defined Reduction (UDR) を除く)

#### 3.1.2.2 Intel® Parallel Studio XE Composer Edition for C++

Intel® Parallel Studio XE Composer Edition for C++は、ハイパフォーマンスを実現するための最新テクノロジーが集約された C/C++コンパイラです。

最新のインテルアーキテクチャ・プロセッサに対応したアプリケーションを簡単に開発できます。OpenMP を使用したベクトル化とスレッド化により、AVX-512 命令を含む最新の SIMD(single-instruction-multiple-data)対応のハードウェアを最大限に活用できます。

また、機能拡張した GNU デバッガ(インテル GDB)も同梱しています。Intel® Parallel Studio XE Composer Edition for C++は、以下の言語規格に対応しています。

- ISO/IEC 9899:1999
- ISO/IEC 9899:1990
- ISO/IEC 14882:2003
- OpenMP API V4.0

#### 3.1.2.3 MPI ライブラリ

Intel® MPI ライブラリは、ハイパフォーマンスな MPI-3.1 規格を複数のファブリックに実装することにより、インテルアーキテクチャー・ベースのクラスタで、アプリケーションのパフォーマンス向上を支援します。

Intel® MPI ライブラリは、必要な場合のみ動的に接続を確立し、メモリー・フットプリントを削減します。また利用できるトランスポートの中から最も高速なものを自動で選択します。必要なメモリ空間のみ割り当てる 2 段階からなる通信バッファーの拡大機能を含むいくつかの手法により、メモリ使用量を抑えます。

レイテンシーの低下、メッセージレートの向上、RMA/片方向通信の向上等、パファーマンス最低化と、スケーラビリティを向上しています。

#### 3.1.2.4 科学技術計算ライブラリ Intel® MKL

Intel® MKL(Math Kernel Library)は、工学、科学、金融系アプリケーションのパフォーマンスを最大限に引き出すために、高度に最適化され、広範囲にスレッド化された演算ルーチンのライブラリです。

- 密線形代数 – BLAS/CBLAS、LAPACK、信頼領域ソルバ
- スパース線形代数 - スパース BLAS、スパース形式変換、PARDISO 直接法スパースソルバ、反復法スパースソルバ、プリコンディショナ
- 高速フーリエ変換(FFT)
- 最適化された LINPACK ベンチマーク
- ディープニューラルネットワーク
- ベクトル統計とデータ適合
- 統計関数 - ベクトル乱数生成器、サマリー統計ライブラリ
- クラスタサポート - ScaLAPACK、クラスタ FFT

### 3.1.3 開発支援

#### 3.1.3.1 Intel® VTune Amplifier XE

Intel® VTune Amplifier XE は、インテル VTune パフォーマンス・アナライザの成功を基に開発された、インテルの最新のパフォーマンス・プロファイラです。多数の高度な機能として、以下の機能が利用可能です。

- 高速で正確なパフォーマンス・プロファイル
  - hotspot(統計コールツリー)
  - ハードウェア・イベント・ベースのサンプリング
- スレッド・プロファイリング
  - スレッドの相互作用の時間軸表示
  - ワークロードをバランス良く分散
- 簡単なセットアップ
  - 事前定義済みのパフォーマンス・プロファイル
- 互換性
  - インテルプロセッサおよび互換プロセッサ
  - インテル、GNU コンパイラ
  - C/C++、Fortran、アセンブリ
- 不具合を素早く発見
  - 無関係なデータをフィルタリング
  - ソース/アセンブリ・レベルでの結果表示
  - イベントの多重化

#### 3.1.3.2 Intel® Trace Analyzer & Collector

Intel® Trace Analyzer & Collector は、ハイパフォーマンスなアプリケーションの解析、最適化、配備を行うことができるツールです。MPI 通信によるパフォーマンスのボトルネックを素早く発見することで、MPI クラスタのパフォーマンスを理解および最適化するための重要な情報を提供します。

### 3.1.4 Arm DDT

Arm DDT（旧 Allinea DDT）は、分散並列環境用のデバッガです。Intel Xeon 用に開発したプログラムの最適化やデバッグが可能です。以下の機能があります。

- MPI を使ったプログラムなど 、複数のサーバで起動するネイティブ Linux アプリケーションのデバッグ
- プロセスおよびスレッドに対し、プログラムのステップ実行、変数の変更やエラー、あるいはユーザーが追加したブレークポイントでのプログラム停止
- メモリーデバッグ機能で、メモリーリーク、ダングリング・ポインター、境界を超えた配列アクセスなどに起因する見つけ難い問題や断続的な問題を検出
- 強力で拡張可能な C++ STL デバッグで、STL コンテナを直観的に表示
- マルチプロセスの変数および配列を表示し、自動的に変更を検出

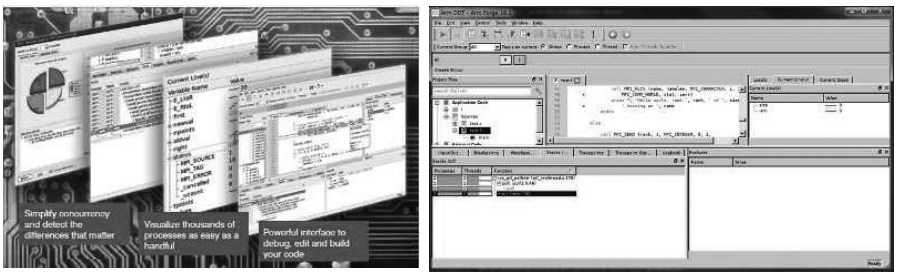



図 3-1 Arm DDT の利用画面

## 3.2 HPC ミドルウェア「Technical Computing Suite」

### 3.2.1 ジョブ運用管理機能

HPC 分野における富士通のバッチジョブ管理システムは、ベクトル計算機時代から NQS をベースに開発し、大規模センター向けジョブスケジューラとして様々な開発ならびに運用実績を積み重ねてきました。ジョブ運用管理機能により、ユーザーアプリケーションの実行で計算機資源を効率的に利用し、システムの最大性能を引き出します。ジョブ運用管理機能は以下の機能で構成されます。

表 3-2 ジョブ運用管理機能の特長

| 名称 | 説明 |
|---|---|
| ジョブマネージャ機能 | ジョブの実行制御と管理を行います。 |
| ジョブスケジューリング(PJS)機能 | 投入された複数のジョブを効率よく実行できるように制御します。 |
| ジョブ資源管理機能 | ジョブが必要とする資源の管理を行います。 |

ユーザーが投入したジョブは、ジョブマネージャ(PJM)機能で受け付けられ、ジョブ ACL をもとに各種制限値に基づいて、利用者のジョブ投入・実行が制限され、ジョブスケジューラ(PJS)機能に実行すべきジョブのスケジューリング(ジョブ選択、資源選択)を依頼します。ジョブ資源管理(PRM)機能は、システムが持つ CPU やメモリなどの計算機資源の使用量を管理します。

ジョブスケジューリング機能(PJS)は実行待ちジョブに対して、複数のジョブが資源を効率的に利用して実行するようジョブ選択優先度順にソートし、以下のようにスケジューリングを行います。

- 現在の空き資源で実行可能なジョブに資源を割り当て、実行を開始します。
- 現在の空き資源で実行できないジョブに未来の空き資源を割り当てます。
- 各ジョブの経過時間制限値を元に資源が解放する時刻を予測し割り当てます。

PJS は以下「表 3-3 ジョブスケジューリング機能」に示すジョブスケジューリング機能を備えています。

表 3-3 ジョブスケジューリング機能

| スケジューリング機能 | 説明 |
|---|---|
| フェアシェア | ユーザー、グループの資源利用実績に基づくスケジューリング機能 |
| 優先度制御 | ユーザー、グループ、ジョブが持つ優先度に基づくスケジューリング機能 |
| FCFS | 投入順序に実行するスケジューリング機能 |
| バックフィル | 優先度の高い実行待ちジョブの実行を妨げない範囲で優先度の低い実行待ちジョブを先に実行するスケジューリング機能 |
| デッドライン | 縮退運転や保守等で運用を停止する資源がある場合、指定された期間、指定された資源を避けてジョブを実行するスケジューリング機能 |

Oakbridge-CX システムでは、ジョブ管理システムによりバッチジョブおよびインタラクティブジョブの実行が制御されます。ユーザーはジョブ実行に必要なリソースグループ名、ノード数、経過時間などを指定し、ジョブ管理システムにジョブ実行を依頼します。依頼されたジョブはジョブ管理システムにより資源の空き状況やジョブ間の優先度などを元に自動で実行開始されます。利用可能なジョブ形式は以下の 2 種類です。

表 3-4 ジョブの形式

| ジョブ形式 | 用途 |
|---|---|
| バッチジョブ | バッチ形式でジョブを実行する。<br>ノードダウンなどの異常発生時、オプションによりジョブ自動再実行が可能。 |
| インタラクティブジョブ | ユーザー端末からのデータ入力により、会話形式でジョブを実行する。<br>主にデバッグ等での利用を想定。異常発生時、ジョブは再実行されない。 |

バッチジョブは、以下の 3 種類の種別に分類されます。

表 3-5 バッチジョブの種別

| バッチジョブ種別 | 用途 |
|---|---|
| 通常ジョブ | バッチジョブの基本単位のジョブ |
| ステップジョブ | 投入した複数のジョブを 1 つのまとまりとして扱い、その中で実行順序、依存関係をもつジョブ |
| バルクジョブ | 1 度のジョブ投入で、複数の通常ジョブを生成し、実行するジョブ |

ジョブ操作のためのコマンドは複数準備されており、「表 3-6 ジョブ操作コマンド一覧」に記載します。

表 3-6 ジョブ操作コマンド一覧

| 機能 | コマンド名 |
|---|---|
| ジョブ投入 | pjsub |
| ジョブ参照 | pjstat |
| ジョブ削除 | pjdel |
| ジョブ保留 | pjhold |
| ジョブ保留の解除 | pjrls |
| ジョブ履歴の表示 | pjlogf |
| ジョブ負荷の表示 | pjsar |
| ジョブスクリプト表示 | pjcat |

### 3.2.2 システム運用管理機能

システム運用管理機能(PSM)は、システムの運用管理に必要な機能を網羅的に備えており、運用管理の統合化(複数のコンポーネントの一元管理)、自動化・省力化(自動監視・運用制御などの自動運転機能)を実現します。システムの運用管理に必要な機能として、システムの電源制御、フェイルオーバー等の運転制御、管理者向けのシステム監視機能等を備えています。パワーキャッピング機能は、ノード毎に最大消費電力を設定しておくことで、実測した各ジョブの消費電力に基づき、設定した消費電力以下で動作するように CPU の周波数を制御します。主要な機能は以下の通りです。

表 3-7 システム運用管理機能の特長

| 機能 | 説明 |
|---|---|
| 制御：<br>電源制御 | システムの起動、停止を制御します。<br>システム全体の起動、停止、特定範囲や特定ノードの起動、停止 |
| 監視：<br>状態監視 | ハードウェアやソフトウェア(サービス)の状態を常時監視します。<br>ノードの稼動状態(電源状態、OS 状態)の確認、ジョブ状態の確認 |
| 運用支援：<br>一括操作、ログ管理<br>調査資料採取等 | 管理者が運用を行う際の支援ツールとして、ノード群に対し、コマンド実行を一括で行う機能や、ログ管理機能等があります。<br>ノード群に対するコマンド一括実行、ログ収集、ログ抽出 |
| パワーキャッピング機能 | 設定した最大消費電力以下で動作するように CPU 周波数の制御やジョブスケジューリングをします。<br>電力マネージメント機能、省電力ジョブスケジューリング機能 |

## 3.3 利用支援ポータル

Oakbridge-CX 利用支援ポータルは、利用者が Web ブラウザにて、ユーザアカウントの SSH 公開鍵登録、メール転送設定、パスワード変更、ドキュメント閲覧、グループ管理者機能等の機能を提供します。
ポータルの機能について、「表 3-8 利用支援ポータルの機能一覧」に、画面イメージを「図 3-2 利用者ポータルの画面イメージ(SSH 公開鍵登録、グループ管理者機能)」に記載します。

表 3-8 利用支援ポータルの機能一覧

| 機能名 | 機能概要 |
|---|---|
| ログイン認証 | ポータルログイン時の LDAP によるユーザ認証。 |
| お知らせ表示 | ログインメッセージ、障害情報、バージョンアップ情報表示。 |
| SSH 公開鍵表示・削除 | 登録済みの SSH 公開鍵を表示・削除。 |
| SSH 公開鍵登録 | SSH 公開鍵をアップロードし、ホーム領域に鍵を設定。 |
| メール転送設定表示・削除 | 登録済みのメール転送設定を表示・削除。 |
| メール転送設定登録 | 転送先のメールアドレスを追加登録。 |
| パスワード変更 | LDAP パスワードの変更。 |
| ドキュメント閲覧 | 利用者用マニュアルや製品マニュアルなどの一覧表示・参照。 |
| トークン表示 | 利用可能なトークン量・使用トークン量を表示。グループコースは、プロジェクト全体の利用可能なトークン量・使用トークン量を表示。 |
| ディスク使用量表示 | 利用可能なディスク量・ディスク使用量を表示。グループコースは、プロジェクト全体の利用可能なディスク量・ディスク使用量を表示。 |
| トークン設定 *7 | プロジェクトメンバの使用したトークン使用量表示およびプロジェクトメンバのトークン使用制限値および最大使用ノード数を変更。 |
| ディスク使用量設定 *8 | プロジェクトメンバのグループ領域ディスク使用量表示およびプロジェクトメンバのディスク使用量制限値およびファイル数制限値を変更 |

図 3-2 利用者ポータルの画面イメージ(SSH 公開鍵登録、グループ管理者機能)

*7 トークン設定は、グループ管理者のみが利用できる機能です。
*8 ディスク使用量制限値設定は、グループ管理者のみが利用できる機能です。

## 3.4 SSD 搭載計算ノードを利用したステージング機能

### 3.4.1 分散ファイルシステム BeeGFS On Demand

計算ノードのローカルストレージ(NVMe SSD)用分散ファイルシステムとして、BeeGFS On Demand (BeeOND)を利用しています。

- ジョブ実行前処理で BeeOND を起動し、ジョブに割り当てられた計算ノードの SSD を使用して、オンデマンドで一つの分散ファイルシステムを構築します。
- BeeOND で構築した分散ファイルシステムは、ジョブに割り当てられた計算ノード間で共有可能(同一ネームスペースを提供)です。

### 3.4.2 ステージング機能

Oakbridge-CX システムのステージング機能は、並列ファイルシステムと計算ノードのローカルディスク(SSD)から構成される分散ファイルシステム(BeeOND)の間で、ジョブの入力ファイル、出力ファイル等を転送する機能です。多数のジョブが実行される状況で並列ファイルシステムを利用すると、各ジョブによる入出力の競合が起こり、ジョブの入出力性能が低下する可能性があります。このような場合、ジョブが利用するファイルシステムを分散ファイルシステムにすることで、入出力の競合を低減できます。

入力ファイルを分散ファイルシステムへ転送する動作をステージイン、出力結果などを並列ファイルシステムへ転送する動作をステージアウトと呼び、これらを総じてステージングと呼びます。

また、ステージング機能は、ステージイン・ステージアウトを行わずテンポラリなスクラッチ利用も可能です。なお、複数ファイルまたは大容量ファイルの場合は、ステージングに時間がかかる場合があり、ステージングの効果が見込まれない場合は、並列ファイルシステムの利用を推奨します。

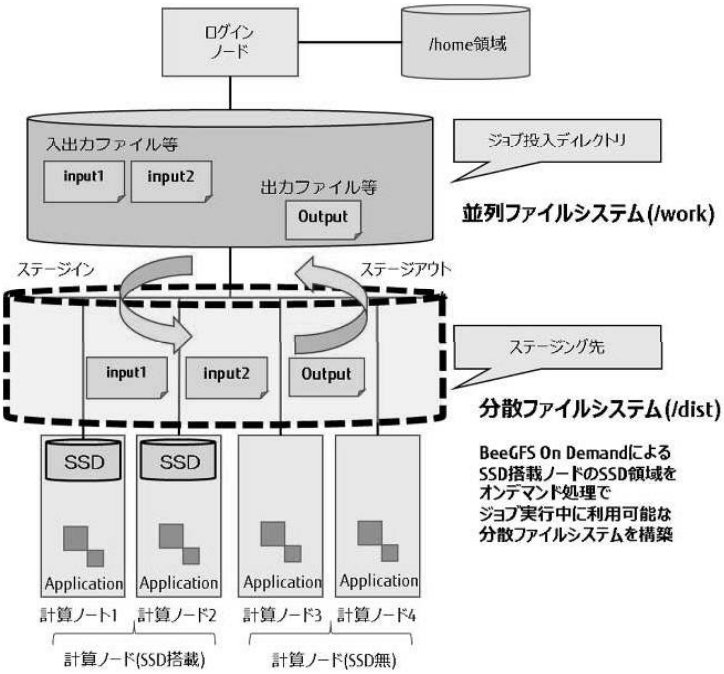



図 3-3 ステージング機能

## 4. 利用イメージ

Oakbridge-CX システムでは、全ての計算ノードとログインノード間で、並列ファイルシステム(/work)の領域が共有されます。共有領域の/home は、ログインノードからのみのアクセス可能です。

ログインノードに対しては、SSH(公開鍵認証方式)によりログインし、コマンドの対話的実行が可能であり、プログラムの作成・編集、コンパイル/リンク、ジョブ投入・表示などを行います。

計算ノード向けのプログラムを実行する場合、バッチジョブもしくはインタラクティブジョブとしてジョブ管理システムに処理を依頼します。インタラクティブジョブは、プログラムのデバッグや会話的なアプリケーション、可視化ソフトウェアの実行に使用します。

Oakbridge-CX のシステム利用について、「図 4-1 システムの利用イメージ」に記載します。

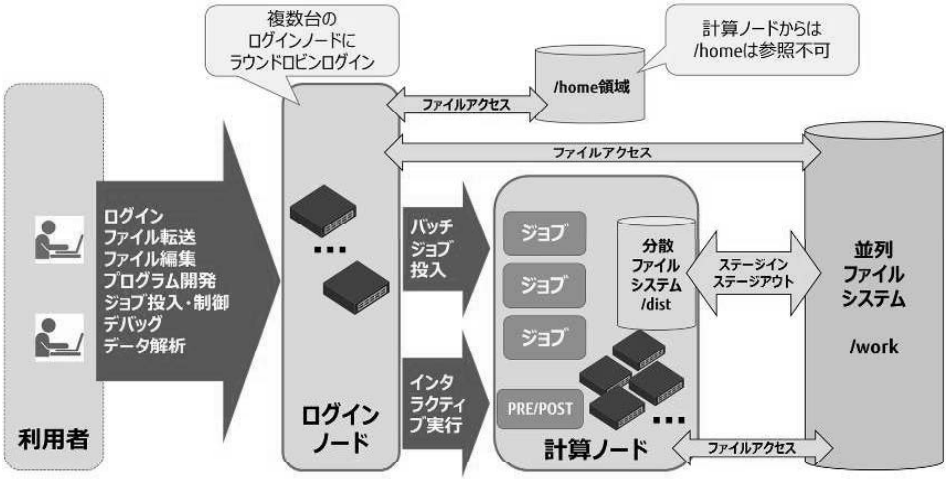



図 4-1 システムの利用イメージ

## 5. 参考 URL

Oakbridge-CX の利用コースやサービスの詳細は、下記 URL を参照してください。

https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/

Oakbridge-CX の利用手引書は、利用者ポータルより入手することが可能です(ただし利用者のみ)。

https://obcx-www.cc.u-tokyo.ac.jp/