Oakforest-PACS スーパーコンピュータシステムの運用

宮嵜 洋[†], 山本 和男[†], 小林 弘幸[‡], 田川 善教[†], 佐島 浩之[†], 坂井 朱美[†], 安部 達巳[†]

[†]東京大学情報システム部情報基盤課

[‡]筑波大学計算科学研究センター

miyazaki@cc.u-tokyo.ac.jp

概要： 2016 年 12 月に運用を開始した Oakforest-PACS スーパーコンピュータシステム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]と筑波大学計算科学研究センター[2]は最先端共同 HPC 基盤施設[3]（JCAHPC : Joint Center for Advanced High Performance Computing）を立ち上げ、共同でスーパーコンピュータを調達し、メニーコア型スーパーコンピュータシステム（Oakforest-PACS システム）[4]を設置、2016 年 12 月から運用を開始した。Oakforest-PACS はメニーコアプロセッサ技術を用いた米インテル社の高性能プロセッサ Intel Xeon Phi （Knights Landing）を搭載した 2017 年 6 月時点での国内最高性能のスーパーコンピュータである。2 つの大学により共同でスーパーコンピュータシステムを調達・運用するという初の試みにより、「京」コンピュータを超える理論演算性能を有する計算資源の導入、提供を実現することができたので報告する。

図 1．Oakforest-PACS の外観

2 導入の経緯

世界最高水準の計算科学の推進のため、過去に共同仕様を策定した T2K オープンスーパーコンピュータアライアンス（筑波大、東大の両センターと京都大学学術情報メディアセンターからなる協定）による連携をさらに推し進め、2013 年 3 月、筑波大学と東京大学で「最先端共同 HPC 基盤施設の設置及び運営に関する協定」を締結し、本協定のもとに最先端共同 HPC 基盤施設を設置した。本施設の理念に則り、2 センターが相互に連携・協力してスーパーコンピュータシステムの開発、運用することを決め、設置場所は筑波大と東大（本郷キャンパス）のほぼ中間地点にある柏キャンパスとした。

2013 年 7 月、調達を開始する時点では、一体運用と調達手続きについて様々な実現方法を模索していたため、資料提供招請は両大学からそれぞれ官報公示を行ったが、その後は共同で実施し、落札後の貸借契約は筑波大、東大と納入業者の 3 者間（実際はリース会社も含め 4 者間）とすることで合意した。ただし、調達を円滑に進めるため、事務手続きについては東大で行うこととした。2014 年 6 月、調達手続きに関する委任状を交わし、両大学共同で仕様書原案の策定に入る。2015 年 1 月の仕様書原案説明会を経て、8 月には計算資源の持ち分や賃貸借料、光熱水料等の負担割合について覚書を交わした。割合は東大と筑波大で 2:1 としている。2016 年 1 月までに仕様書が完成、入札公告がなされた。結果は富士通が落札し、11 月 30 日に納品が完了、12 月 1 日から運用を開始した。借入期間は 5 年半（2016 年 10 月からの部分的納入も含め、2022 年 3 月まで）である。

3 システム概要

3.1 ハードウェア

Oakforest-PACS システムはメニーコアプロセッサ Intel Xeon Phi 7250 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 8,208 台で、計算ノード単体の理論演算性能は 3.0464TFLOPS、主記憶容量は、96GB(DDR4) ＋ 16GB(MCDRAM)であり、全体では 25.004PFLOPS、

897TB の性能を有している。フロントエンドには Xeon E5-2690v4 のサーバ 6 台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 26PB、高速ファイルキャッシュシステム 940TB を備える。後者は DDN 社製 IME14K で、転送性能 1TB/sec のバーストバッファとして機能する。(図 2 および表 1、表 2)

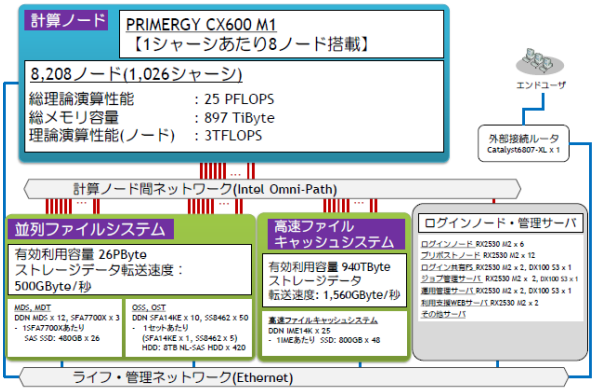

図２．Oakforest-PACS 全体構成

表１．システム全体諸元（計算ノード）

総理論演算性能		25.004 PFLOPS
総ノード数		8208
総主記憶容量		897 TB
ネットワークトポロジー		Full-bisection Fat Tree
並列ファイルシステム	システム名	Lustre ファイルシステム
	サーバ(OSS)	DDN SFA14KE
	サーバ(OSS)数	10
	容量	26 PB
	データ転送性能	500 GB/sec
高速ファイルキャッシュシステム	サーバ	DDN IME14K
	サーバ数	25
	容量	940 TB
	データ転送性能	1560 GB/sec

表２．ノード諸元（計算ノード）

マシン名		PRIMRGY CX1640M1
CPU	プロセッサ名	Intel Xeon Phi 7250 (Knights Landing)
	コア数	68 コア
	周波数	1.4 GHz
	理論演算性能	3.0464 TFLOPS
メモリ	DDR4	96 GB 115.2GB/sec(ピーク)
	MCDRAM	16 GB 490GB/sec(実効)
インターコネクト		Intel Omni-Path ネットワーク (100 Gbps)

3.2 ソフトウェア

表 3 に示すとおり、主に 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、筑波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
コンパイラ	GNU コンパイラ Intel コンパイラ (Fortran77/90/95/2003 / 2008、C、C++)
メッセージ通信ライブラリ	Intel MPI、Intel Omni-Path Fabric Software
ライブラリ	Intel 社製ライブラリ(MKL)、BLAS、LAPACK、ScaLAPACK、FFTW、GNU Scientific Library、NetCDF、Parallel netCDF、Xabclib、ppOpen-HPC、ppOpen-AT、MassiveThreads、SuperLU、SuperLU MT、SuperLU DIST、METIS、MT-METIS、ParMETIS、Scotch、PT-Scotch、PETSc、Boost
アプリケーション	mpijava、omnicompiler、OpenFOAM、ABINIT-MP、PHASE、FrontFlow/blue、FrontISTR、REVOCAP、OpenMX、xTAPP、AkaiKKR、MODYLAS、ALPS、feram、GROMACS、BLAST、R、bioconductor、BioPerl、BioRuby
ﾃﾞﾊﾞｯｶﾞ ﾌﾟﾛﾌｧｲﾗ	Total View、Intel VTune、Trace Analyzer & Collector Allinea DDT

4 運用形態

4.1 利用区分

Oakforest-PACS の利用申込は原則として東京大学と筑波大学でそれぞれ募集を行っている。計算ノードに対するリソース制限の方法によって、トークンというノード時間積を付与し、利用とともにトークンを消費する「バジェット制限型」と、1 か月単位でノードが割り当てられ、同時使用ノード数の上限まで使用することができる「ノード制限型」という 2 種類の利用区分がある。そのほか HPCI 資源提供と大規模 HPC チャレンジについては、両センター共同で実施しており、最先端共同 HPC 基盤施設（JCAHPC）として申し込みを受け付けている。(表 4)

表４．利用区分と申込コース

利用区分	コース名	申込先
バジェット制限型	パーソナルコース	東京大学
	グループコース	
	HPCI 資源提供	JCAHPC
	大規模HPCチャレンジ	
	学際共同利用	筑波大学
ノード制限型	大規模一般利用	

4.2 バジェット制限とノード制限

東京大学の「パーソナルコース」、「グループコース」（表 5）と筑波大学の「学際共同利用」（公募）ではバジェット制限型を採用している。利用するコースや利用申込したノード数（申込ノード数）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）」を割り当てるが、この割り当てられたトークン内であれば申込ノード数によらず、各コースにおける最大利用可能ノード数（グループでの利用なら 2,048 ノード）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申込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しない。

表５．Oakforest-PACS 利用コース（東大の例）

コース		利用負担金額、他
パーソナルコース	コース 1	【大学・公共機関等 100,000 円】 トークン：17,280 ノード時間 （2 ノード/年） 並列実行ノード数：16 ノードまで 消費係数：8 ノードまで 1.00, 8 ノード超で 2.00 ディスク容量：1 TB
	コース 2	【大学・公共機関等 200,000 円】 トークン：34,560 ノード時間 （4 ノード/年） 並列実行ノード数：64 ノードまで 消費係数：16 ノードまで 1.00, 16 ノード超で 2.00 ディスク容量：1 TB
グループコース		【大学・公共機関等 400,000 円, 企業 480,000 円】（8 ノード当たり） トークン：69,120 ノード時間 （8 ノード/年） 並列実行ノード数：2,000 ノードまで 消費係数：申込ノード数まで 1.00, 申込ノード数超で 2.00 ディスク容量：グループにつき 2 TB（8 ノード当たり）
トークン追加		パーソナルコース 1【大学・公共機関等 8,300 円】 1,440 ノード時間 パーソナルコース 2【大学・公共機関等 16,600 円】 2,880 ノード時間 グループコース【大学・公共機関等 33,300 円，企業 40,000 円】 5,760 ノード時間

筑波大学の「大規模一般利用」は、ノード制限型である。ノード制限型は契約に応じてノード数に制限があり、月単位に同時実行ノード数の上限が設定される。

4.3 ジョブキュー

4.3.1 メモリモード

Oakforest-PACS では、MCDRAM をどのように使うかで 2 種類のジョブキューを用意している。

MCDRAM をキャッシュ領域として使用する cache モード（DDR メモリ領域 96GB + MCDRAM を DDR のキャッシュとして使う領域 16GB）と flat モード（DDR メモリ領域 96GB + MCDRAM メモリ領域 16GB）があるが、BIOS によってモードを設定するため、ジョブ実行時の動的な変更ができない。利用者がどちらのモードを選択するか、初期の運用で動向を見極めていたが、運用開始から 4 ヶ月のノード時間の平均はほぼ半数（flat 49%、cache 45%、その他 5%）であった。2017 年度の 4～8 月の平均では、flat が 70%、cache が 30%と flat モードが優勢となっているが、一方に倒すほど顕著な結果とは言えず、当面は両方を提供し様子を見ることとしている。（図 3）

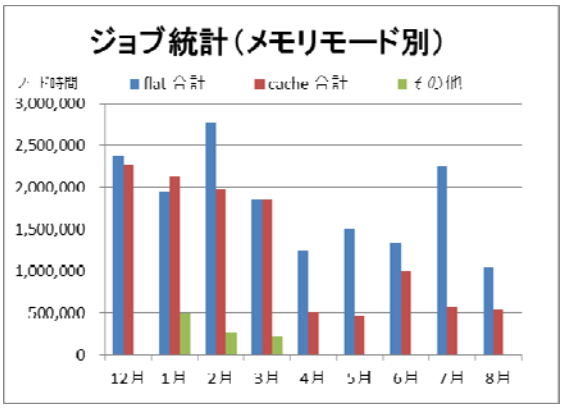

図３．Oakforest-PACS メモリモード別ジョブ統計

4.3.2 ジョブキューと制限値

ノード数が 2,048 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(128 ノードまで)を用意している。メモリモードごとにそれぞれのキューが用意されているのも前項の理由による。(表 6)

インタラクティブ実行用のキューは 16 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使用する tutorial、lecture キューや、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 キューを用意している。

表６．ジョブクラス制限値

キュー名	ノード数	制限(経過)時間
debug-flat debug-cache	1～128 1～128	30 分 30 分
regular-flat (small-flat) (medium-flat) (large-flat) (x-large-flat) regular-cache (small-cache) (medium-cache) (large-cache) (x-large-cache)	 1～ 128 129～ 512 513～1024 1025～2048 1～ 128 129～ 512 513～1024 1025～2048	 48 時間 48 時間 48 時間 24 時間 48 時間 48 時間 48 時間 24 時間
interactive-flat (interactive_n1-flat) (interactive_n16-flat) interactive-cache (interactive_n1-cache) (interactive_n16-cache)	 1 2～16 1 2～16	 2 時間 10 分 2 時間 10 分
tutorial lecture	16	15 分
challenge-flat challenge-cache	1～8208	24 時間
prepost	1	6 時間

4.4 ファイルシステム

Oakforest-PACS 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(申込 8 ノードにつき 2TB)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。高速ファイルキャッシュシステムはステージング機能により、ジョブ投入後、入力ファイルを並列ファイルシステム /work から高速ファイルキャッシュ/cache へ転送し、ジョブ終了時にジョブの結果を/cache から /work へ転送することができるため、高速な入出力を必要とするジョブの実行に効果がある。

4.5 大規模 HPC チャレンジ

2017 年 10 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、Oakforest-PACS がもつ最大計算ノード数である 8,208 ノードを、最大 24 時間、1 研究グループで計算資源の占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、全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。初年度は 6 回実施する予定で 9 月末時点で 3 課題が採択されており、残りの 3 回を募集中である。

5 利用状況

5.1 利用ノード時間

Oakforest-PACS は 2016 年 12 月に運用を開始したが、2017 年 3 月までは試験運用期間とした。Oakforest-PACS のキュー別ノード時間積ジョブ統計を図 4 に示す。

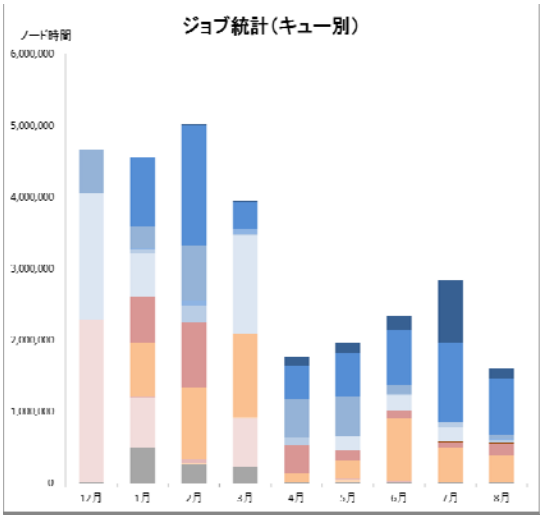

図４．Oakforest-PACS キュー別ジョブ統計

12 ～ 3 月までは試験運用期間として希望者に無償で提供していたため、80%を超える利用率であったが、2017 年度に入り、利用負担金が課されてから、ノード時間積は 4 月に一旦下がった。その後は次第に回復しつつあり 7 月で 47%である。8 月は夏季の節電により約 8 日間のシステム停止を行ったため、利用されたノード時間も少なくなっている。なお、グラフはキュー別に色分けしているが、寒色系が flat モード、暖色系が cache モードのキューである。

5.2 利用申込

9 月末時点での登録ユーザ数は 111 グループ 1,197 ユーザ（筑波大 45 グループ、327 ユーザ、東大 66 グループ、870 ユーザ)、お試しアカウント付き講習会での利用が 3 グループあった。

利用者へのトークンの払い出しについては 9 月末時点で総量の 64% 程度である。総量までの払い出しを上限としている。

2017 年度からは 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）に資源を供出している。

また、通常の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）についても募集を行っている。

6 今後の課題

Oakforest-PACS の運用を開始して約 1 年になるが、空調の温度や風量、冷却水の温度などシステムの発熱との調整を行いつつ、全体としての稼働は安定している。ただし、利用者の目線では使い始めにはまだ戸惑いもあるようだ。メモリモードについても、適切な使用方法を周知していく必要がある。利用者がメニーコアプロセッサの性能を発揮するためにはお試しアカウント付き講習会（「KNL 実践」など）に参加することが近道だが、まずはこのような機会があるという情報を発信してくことが重要である。また、高速ファイルキャッシュについては高い性能の反面、現状では構成する装置 1 台の障害でファイルシステム全体の復旧が必要となる不具合がある。このため、利用者には問題点を認識してもらった上で利用してもらっているが、この問題は早急に解決しなければならない。

7 おわりに

Oakforest-PACS は Top500 では 2016 年 11 月に世界 6 位、国内 1 位を達成して順調なスタートを切ることができ、2017 年 6 月にも国内では最高性能であった。2 大学が共同で導入することで、より大規模な計算資源を提供できるようになり、全体性能の向上だけでなく、利用者のジョブの流れやすさ、様々な環境やサービスを投入できる運用の自由度などメリットは大きい。両大学の教員のみならず、技術職員が協力して日々の運用に当たるという体制も新鮮である。これらの利点を活かして今後も利用者にとって使い勝手の良いスーパーコンピュータを提供することに尽力していきたい。

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] 筑波大学計算科学研究センター
https://www.ccs.tsukuba.ac.jp/
[3] 最先端共同 HPC 基盤施設（JCAHPC）
http://jcahpc.jp/
[4] Oakforest-PACS スーパーコンピュータシステム http://www.cc.u-tokyo.ac.jp/system/ofp/