スーパーコンピュータ OCTOPUS の混雑緩和に向けた取り組み

勝浦 裕貴 [1)], 寺前 勇希 [1)], 木越 信一郎 [1)], 伊達 進 [2)]

1) 大阪大学情報推進部情報基盤課
2) 大阪大学サイバーメディアセンター

katsuura-y@cmc.osaka-u.ac.jp

efforts to reduce congestion of OCTOPUS

Yuki Katsuura, Yuki Teramae[1)], Shinichiro Kigoshi[1)], Susumu Date[2)]

1) Department of Information and Communications Technology Services, Osaka Univ.
2) Cybermedia Center , Osaka Univ.

概要

大阪大学サイバーメディアセンターでは 2017 年 12 月よりスーパーコンピュータシステム OCTOPUS の運用を開始した。OCTOPUS は多くの研究者の方々からご好評いただき、運用開始から非常に高い利用率を誇っているが、その結果としてジョブ実行までの待ち時間が深刻な問題となっている。本稿ではこの待ち時間の削減に向けた取り組みについてまとめる。

1 はじめに

大阪大学サイバーメディアセンターでは 2017 年 12 月よりスーパーコンピュータシステム OCTOPUS の稼働を開始した[1]。OCTOPUS は合計 319 ノードの比較的小規模なシステムであるが、汎用 CPU ノード群、GPU ノード群、Xeon Phi ノード群、大容量主記憶搭載ノード群という 4 つのノード群で構成されており、ユーザは OCTOPUS への利用申請でこれらのノード群を横断的に利用することができる。各ノード群の概要を図 1 に示す。

汎用CPUノード 236ノード (471.24 TFLOPS)	プロセッサ：Intel Xeon Gold 6126 (Skylake / 2.6 GHz 12コア) 2基 主記憶容量：192GB
GPUノード 37ノード (858.28 TFLOPS)	プロセッサ：Intel Xeon Gold 6126 (Skylake / 2.6 GHz 12コア) 2基 GPU：NVIDIA Tesla P100 (NV-Link) 4基 主記憶容量：192GB
Xeon Phiノード 44ノード (117.14 TFLOPS)	プロセッサ：Intel Xeon Phi 7210 (Knights Landing / 1.3 GHz 64コア) 1基 主記憶容量：192GB
大容量主記憶搭載ノード 2ノード (16.38 TFLOPS)	プロセッサ：Intel Xeon Platinum 8153 (Skylake / 2.0 GHz 16コア) 8基 主記憶容量：6TB

図 1.OCTOPUS の概要

これらの様々なアーキテクチャを利用できるヘテロジニアスな構成や、低めの料金設定、広く普及している Intel 社製のプロセッサを導入していることなどから、サービス開始以来非常に人気の高いシステムとなっている。前述の通り比較的小規模なシステムなこともあって毎年のように資源の売り切れが発生しており、2020 年度は 5 月 18 日時点で資源が売り切れてしまい、新規の利用申請を停止する事態となった。人気が高いことは運用する身としては嬉しい限りなのだが、その反面、多数のユーザが多数のジョブを投入することによる待ち時間の長時間化が悩みの種となっている。

当センターでは年度末にユーザ向けのアンケートを実施しており、その中で OCTOPUS の不満点に関する設問を用意している。図 2 が 2018 年度のアンケートにおける OCTOPUS の不満点に関する回答の集計結果である。「不満点なし」が約 2 割であるのに対し、7 割以上が「ジョブ実行」を不満点として回答している。コメントでもジョブ実行までの待ち時間の改善に関する要望が多く、OCTOPUS の運用において大きな課題となっている事は間違いない。逆に、「計算資源の性能」や「開発環境」なども選択肢として用意していたが、これらを不満点として選択したユーザはおらず、新たな共用計算機としては十二分な性能を備えていたと言える。そのため、システムの性能向上よりも、ジョブの待ち時間を始めとした混雑の解消こ

そが OCTOPUS にとって最も対処すべき課題であると考えられる。

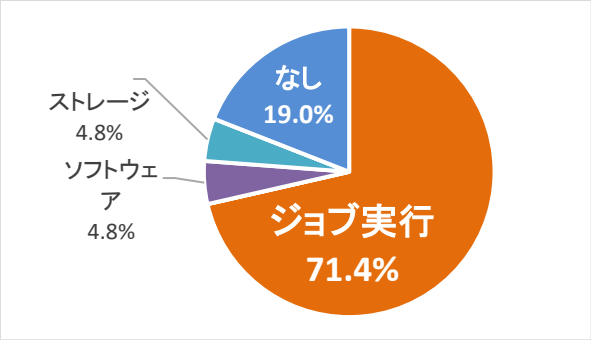

図 2.OCTOPUS の不満点（2018 年度アンケートの集計結果からの抜粋）

2　OCTOPUS の状況と課題

2018 年度の OCTOPUS において、待ち時間が 24 時間を超えたジョブの件数と、全ジョブの平均待ち時間を月別に集計したグラフを図 3 に示す。

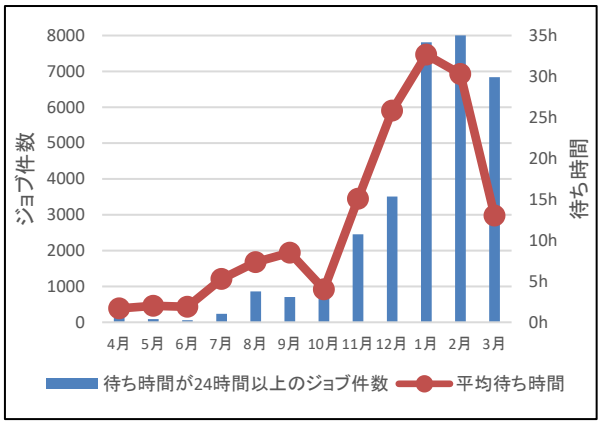

図 3.OCTOPUS の平均待ち時間と待ち時間が 24 時間以上のジョブ件数（2018 年度）

正式サービスの開始が 2018 年 4 月からということもあり、年度前半は比較的落ち着いているものの、年度後半になると待ち時間が急増しており、特に 12 月～2 月には平均で 24 時間以上の待ち時間が発生してしまっている。

当センターの計算機は従量課金制ではなく、利用する資源（OCTOPUS ポイント[2]）を事前に購入し、そのポイントを消費して利用する、所謂プリペイド式の利用負担金制度をとっている。当たり前だが、OCTOPUS が 1 年間で動作できる時間は限りがある（319 ノード×365 日）ため、1 年間で提供可能な資源量上限を予め設定し、上限に達した時点で利用申請の受付を停止するという運用を行っている。これが前述の資源の売り切れである。しかし、年度の前半は購入した資源を温存し、年度後半に集中的に利用するというユーザが一定数存在する。そのため、年度前半にはノードに空きができて提供可能時間を無駄にしてしまう一方で、年度後半は混雑して長時間の待ちが発生するという現象が起こる。これによる長時間の待ちが不満になっているのはもちろん、購入したポイントは年度末で失効するため、混雑で思うようにジョブ実行が進まず余らせたポイントが無駄になってしまうという点も不満を加速させていると思われる。

また、利用時期の偏りだけでなく、利用するノード群の偏りも大きな問題となっている。2018 年度のノード群ごとの月別利用率を図 4 に示す。

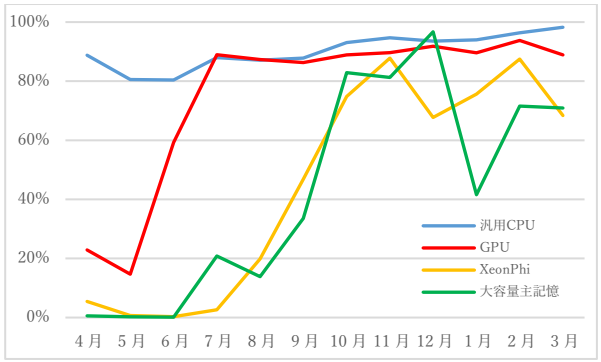

図 4. OCTOPUS の 2018 年度利用率

汎用 CPU ノード群は 1 年を通して高い利用率を保っているが、GPU ノード群、Xeon Phi ノード群、大容量主記憶搭載ノード群については年度前半の利用率が低くなっている。このように汎用 CPU ノード群だけに利用が集中すると他のノード群に空きができてしまい、提供可能時間が無駄になってしまう一方、汎用 CPU ノード群では長い待ち時間が発生してしまう。こうなるとユーザのポイント消費が思うように進まず、ポイントを余らせないために年度後半は更に躍起になってジョブを投入するようになり、年度前半にポイントを温存していたユーザのジョブとあわせて年度後半の混雑が更にひどいものになる。

1 年間の提供資源量の上限は全てのノード群が 100%利用される前提で計算しているため、ノードに空きができ、提供可能時間を無駄にしてしまうことは、年度末の混雑に直結する。当センターに限った話ではないと思うが、11 月～2 月頃は学生

の論文シーズンなこともあって利用が集中するため、利用時期の偏りはある程度仕方のない部分ではある。しかし、この利用時期の偏り、そして利用ノード群の偏りを緩和し、提供可能時間の無駄を減らすことが、年度後半の待ち時間の緩和にもつながると考えられる。

なお、2018 年度後半には Xeon Phi ノード群、大容量主記憶搭載ノード群の利用率が向上しているが、これはセンター側で何か対策を実施したという訳ではなく、繁忙期に入り混雑を極めた汎用 CPU ノード群の待ち時間にしびれを切らし、空いている Xeon Phi ノード群、大容量主記憶搭載ノード群にユーザが流れたことが原因と考えられる。この時期に Xeon Phi ノード群へ大量にジョブを投入していたユーザに話を聞く機会があったが、Xeon Phi に最適化されたプログラムではないため実行時間自体は伸びてしまうものの、待ち時間を加味すると汎用 CPU ノード群よりも遥かに早くジョブが終わるとのことであった。また、GPU ノード群は 6 月から急激に利用率が上昇しているが、この原因については後述する。

3　混雑緩和に向けた取り組みと成果

第 2 章で述べた OCTOPUS の課題について、2018 年度～2019 年度に行ったいくつかの取り組みを行った。本章ではこれらの詳細について報告する。

3.1 GPU ノードを汎用 CPU ノードとして利用可能に

図 1 を見ると、GPU ノード群は汎用 CPU ノード群と基本的な構成が同一となっている事が分かる。というのも、OCTOPUS の調達活動を行っていた段階から汎用 CPU ノード群が最も利用されるのではないかと予想していたため、そういった場合に GPU ノードを汎用 CPU ノードとしても利用できるよう仕様を定めたためである。

予想通り、正式サービス以前の無料開放期間（2017 年 12 月～2018 年 3 月）からすでにこの傾向が表れており、正式サービスを開始してからも汎用 CPU ノード群のみ待ち時間が発生するという状況が続いたため、2018 年 6 月から GPU ノードを汎用 CPU ノードとしても利用できるよう設定変更を行った。その結果、図 4 のように GPU ノード群の利用率が 6 月から急激に上昇することとなった。この説明からも分かると思うが、GPU ノード群の利用率の大半は汎用 CPU ノードとして利用されたものである。図 4 の利用率はあくまで”ノードが利用されているかどうか”を表しており、GPU ノードで実行されているジョブが GPU を利用するかどうかに関わらず、ジョブが実行されていれば利用率として計上している。

図 4 からも分かる通り、4～5 月は GPU ノードの利用率が非常に低かったため、当初は全ての GPU ノードに対し、汎用 CPU ノードに投入されたジョブ（以下、「汎用 CPU ジョブ」という）が実行されるよう設定変更を行った。これにより汎用 CPU ジョブの待ち時間は減少したものの、GPU ノードに投入されたジョブ（以下、「GPU ジョブ」という）の待ち時間が大幅に増加することになった。この時点では GPU ジョブの数が少なかったため影響は大きくなかったのだが、運用を続けているうちに GPU ジョブが増加するだけでなく、それを遥かに上回る速度で汎用 CPU ジョブも増加していったため、「GPU ノードなのに汎用 CPU ジョブばかり実行されていて肝心の GPU ジョブが流れない」という問題が表面化するようになった。この対策として GPU ジョブ専用ノードの確保や GPU ジョブのアサイン可能時間の延長等を実施し、GPU ジョブが実行されやすくなるよう調整した。GPU ジョブが増えたとはいえ、汎用 CPU ジョブの数が圧倒的に多いことは変わりないため、上述した対策のノード数や時間などを適宜調整し、ジョブのバランスを取りながら運用を行っている。ノードの基本的な構成を同一にしなければならないという制約はあるが、ノード群間での負荷分散や利用率向上において非常に有効であるため、複数ノード群で構成されるシステムの構築を行う際は、複数のノード群の役割をこなせるノード設計にすることが望ましいのではないかと思う。

3.2 季節係数の導入

OCTOPUS の共有利用制度では、前述の通り OCTOPUS ポイントと呼ばれる資源をユーザが購入し、それを消費しながら OCTOPUS を利用するプリペイド式の利用負担金制度となっている。ジョブを実行した際のポイントの消費量は下記の計算式によって算出される。

消費量 = 使用ノード時間 × 消費係数 × 季節係数

「消費係数」は各ノード群の消費電力を元に定められた値である。例えば、GPU ノードは汎用 CPU ノードに対して約 4 倍の電力を消費するため、

消費係数も約 4 倍に設定されている。それによって、同じ時間、同じノード数の利用でも、汎用 CPU ノードの約 4 倍のポイントを消費するという仕組みである。これは、ノード群毎の個別の利用申請を必要とせず、“OCTOPUS ポイント”という一元的な資源で管理して 4 つのノード群を横断的に利用できるようにすることを目的とし導入した係数である。なお、汎用 CPU ジョブが GPU ノード群で実行された場合は、汎用 CPU ノード群の係数が適用される。

一方、「季節係数」は利用時期と利用ノード群の偏りの緩和を目指して導入された係数である。前年度の各ノード群の利用率を元に 0 より大きい 1 以下の値が設定され、特定期間、特定のノード群のポイント消費量が割引されるという仕組みである。正式サービス開始初年度でなる 2018 年度は通年 1 で運用したが、2019 年度の季節係数は 2018 年度の利用率を元に学内の委員会で議論し、表 1 の値で運用することとした。

ノード群	4～6 月	7～9 月	10～12 月	1～3 月
汎用 CPU	1			
GPU	1			
Xeon Phi	0.5	0.7	1	1
大容量主記憶	0.8	0.8	1	1

表 1. 2019 年度の季節係数

汎用 CPU ノード群及び（汎用 CPU ジョブと共用となった後の）GPU ノード群については常に高い利用率を保っていたため、引き続き通年 1 での運用とした。一方、2018 年度前半の利用率が低かった Xeon Phi ノード群は 4～6 月を 0.5、7～9 月を 0.7 とした。大容量主記憶搭載ノード群も同じく低かったが、もともと 2 ノードしかないことを考慮し、下げ幅は控えめにして 0.8 とした。この場合、例えば 4～6 月に Xeon Phi ノード群で実行したジョブは、そのポイント消費量が 5 割引される。上表の内容で 2019 年度は運用することを 2019 年 2 月にユーザへ周知し、 1 年間運用を行った結果、ノード群ごとの月別利用率は図 5 のようになった。

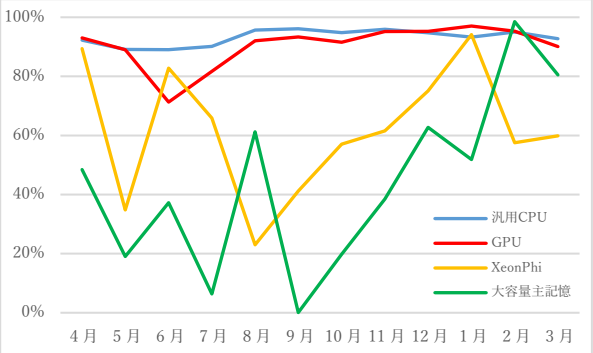

図 5. OCTOPUS の 2019 年度利用率

利用率を見る限りでは一定の成果はあったように見受けられる。特に Xeon Phi ノード群は季節係数の低い 4~6 月の利用率が昨年に比べて大きく上昇しており、利用時期の偏りにはある程度の効果はあったと言える。一方で、4~6 月に Xeon Phi を利用していたユーザの殆どが前年度にも同ノード群を利用していたユーザであり、季節係数の低さを目的として普段利用しているノード群以外にジョブ投入を試みるユーザはごく一部であるということも分かった。つまり、利用ノード群の偏りに対してはほとんど効果が得られなかったということになる。

また、季節係数を低くするとポイント消費量が抑えられるため、その分だけユーザが利用できるノード時間は増えることになる。季節係数の導入によって利用時期の偏りがある程度解消され、閑散期の利用率アップにはつながったものの、利用できるノード時間が増えている以上、それが繁忙期の混雑の解消にも効果があったとは言い切ることは難しいかもしれない。とはいえ、繁忙期の待ち時間は 2018 年度と比べると多少は短くなっていたため、一定の効果はあったと言える。

この季節係数に関して、「年度末の混雑を避けるため、年度末の季節係数を 1 より大きくすることに賛成か、反対か」という問いを 2019 年度のアンケートに設けた。季節係数が 1 より大きくなるということはポイント消費量が通常よりも割高になるということであり、ポイント消費量という点ではユーザにとってデメリットでしかない。当然筆者も反対多数を予想していたのだが、その結果は賛成 56%、反対 41%、どちらとも言えないが 3% と、賛成多数の結果となった。賛成派の意見としては、年度初めからの計画的な利用を促せる、年度末の混雑緩和に繋がるというものが多数であった。一方、反対派の意見には下記のようなものが

見られた。

- 年度末以外の利用は増えるかもしれないが、年度末に混雑するのは変わらないと思う。
- 料金の問題でなく、年度末近くに論文提出や学会が多いのが問題であり、季節係数増加による抑制効果は小さいのではないか。
- 年度末は（資源が売り切れているため）ポイントの追加購入ができなくなるのに、その年度末にポイントが不足する事態に陥りかねない。また、そのような事態に注意しながら利用したくない。
- 研究計画にも影響するので、上限値は 1 で固定してもらいたい。負担金を増やすか、全体の提供時間を減らすかして対応するべき。

本設問はあくまでユーザの意識調査を目的として設けたものであり、現時点では季節係数を 1 より大きく設定する予定はないが、アンケートの結果としては非常に興味深いものであった。2020 年 9 月現在調達を行っている高性能計算・データ分析基盤システム（以下、「次期スパコン」という）は OCTOPUS と同じく特徴の異なる複数のノード群から構成されるシステムを予定しており、課金制度についても OCTOPUS と同様に一元的なポイント制となることを想定している。次期スパコンの課金制度をより使いやすいものとするためにも、アンケートの意見も参考にしつつ OCTOPUS を運用し、季節係数等のポイント制度の運用ノウハウを蓄積しておきたい。いずれにしても、ユーザの研究計画に悪影響を及ぼすことは避けなければならないため、季節係数については引き続き慎重に運用を進めていくこととする。

3.3 クラウドバースティング機能の実装

“クラウドバースティング機能”とは OCTOPUS が混雑した際にジョブの一部をパブリッククラウド上の計算ノード（以下、「クラウドノード」という）に割り当て、OCTOPUS の計算負荷を肩代わりさせる機能である。この機能は本センターの伊達進准教授主導のもと実装した機能であり、2019 年度の AXIES 年次大会に投稿した論文「OCTOPUS のクラウドバースティング拡張」[3]にて転送の仕組みや環境構築について解説しているため、詳細についてはここでは触れない。

本機能の実装当初は Microsoft Azure[4]との連携のみであったが、2020 年 9 月より新たに Oracle Cloud Infrastructure[5]（以下、「OCI」という）との連携も可能となった。Azure 上のクラウドノードは OCTOPUS の汎用 CPU ノードにできるだけ近い構成のインスタンスを選択しているが、仮想化環境と物理環境という差異が原因となり、プログラムによっては計算結果に違いが出る、エラーで実行できないなどの問題が発生していた。新たに連携可能となった OCI 上の計算ノードはベアメタルサーバである為、この仮想と物理の差異に起因する問題が解決されるのではないかと期待されている。

本機能は未だ実証実験の段階ではあるが、純粋に”混雑時に一時的に計算ノードが増える”ことになるため、混雑の緩和に効果的であることは言うまでもない。一般ユーザ向けに機能開放するに当たっては、前述の技術的な問題の解決はもちろん、クラウドノードの利用料金の扱いやジョブの転送ポリシーと利用方法の検討、それらに伴う利用規定の改正やユーザへの周知等、やるべきことは多く残されているが、混雑緩和だけでなく様々な目的に応用できる機能であり、引き続きクラウドバースティング機能の検証・検討を進めることとする。差し当たり、OCI について当センターのユーザを対象に実証実験を行う予定である。

3.4 その他の対策

ここまでいくつか混雑緩和に向けた取り組みを紹介したが、直接的に混雑の緩和にはつながらないまでも、これ以外にも様々な混雑対策を行っている。

例えば、DBG キュー（デバッグキュー）及び DBG キュー専用ノードの新設である。OCTOPUS ではフロントエンドノード（プログラムのコンパイル、ジョブ投入等を行う作業用のノード）でのプログラム実行を禁止しているため、ジョブスクリプトを用いてジョブ投入し、計算ノードで実行する必要がある。しかしながら、汎用 CPU ノード群は全ノードのスケジュールが埋まっていることがよくあり、簡単なプログラムの実行ですら数日の待ち時間が発生することも珍しくない状況であった。これではプログラムの動作確認やデバッグもままならないため、1 ノード実行かつ最長 10 分までのジョブのみ投入可能な DBG キューを新設した。また、キューを新設してもノードに空きがなくては意味がないため、DBG キュー専用のノー

ドもあわせて用意している。流石に DBG キュー専用ノードは頻繁に空きができている（=提供可能時間を無駄にしている）が、利便性を考えると必要なノードであるのは間違いない。

他にも、ジョブスクリプトに記載する経過時間リクエストと実際のジョブの実行時間に大きな乖離があるユーザへの個別連絡なども行っている。OCTOPUS のスケジューラは基本的には先入れ先実行方式でノードが確保され、ジョブが実行される。ノードの確保はジョブスクリプトに記載されている経過時間や並列数のリクエストを元に行われるのだが、大規模なジョブが投入されると空きノード数などの関係で隙間時間ができることがある（図 4, 5 において慢性的に混雑している汎用 CPU ノード群の利用率が 100%に満たないのはこのためである）。ここで新たにジョブが投入された場合、通常は大規模ジョブの後にスケジューリングされるが、経過時間リクエストが隙間時間内におさまる短時間ジョブであった場合はこの隙間時間にスケジューリングされ、実行される。長時間に設定した場合よりも早く実行が始まるだけでなく、運用側にとってもこういった隙間時間を有効活用できると（微々たるものだが）混雑緩和にもつながるため、実際の実行時間が短い傾向であるにも関わらずジョブスクリプトの経過時間リクエストが長時間になっているユーザには積極的に連絡を行っている。

また、これは混雑対策を主目的にしたものではないが、ユーザサポートの一環としてプログラムチューニング支援サービスを行っている。当センターのユーザを対象にプログラムを募集し、OCTOPUS のベンダーである日本電気（株）のエンジニアがチューニングを施すというもので、作業内容はユーザの希望によって様々だが、多くの場合は並列化や OCTOPUS への最適化などである。応募するのは当センターのヘビーユーザの方が多く、こういった方のプログラムを高速化することで他のユーザも待ち時間が減るという恩恵も得られ、間接的に待ち時間対策になっているとも言える。本サービスは 2016 年度から毎年数件行っており、今後も継続して実施する予定である。その他、キューごとのアサイン数、実行数上限の定期的な調整などの細々とした対策も適宜行っており、混雑している中でもできる限り使いやすい環境になるよう取り組んでいる。

4 結果

4.1 取り組みの成果

以上、2019 年度までに行ったいくつかの混雑対策を紹介した。取り組みの効果の指標として、2018 年度と 2019 年度の平均待ち時間を比較したグラフを図 6 に示す。

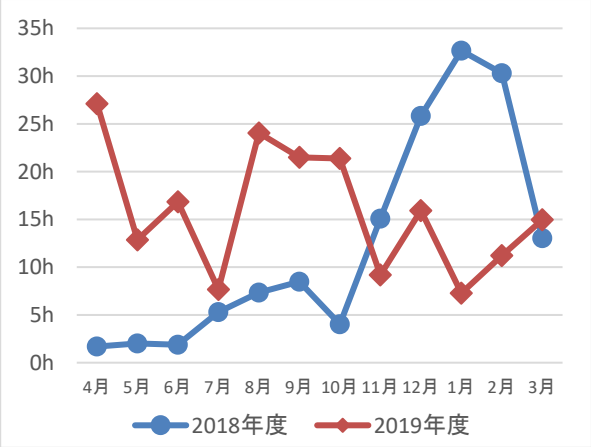

図 6. 2018 年度と 2019 年度の平均待ち時間

結果としては待ち時間の解消に対してあまり良い成果を挙げられなかったと言わざるをえない。また、2018 年度は OCTOPUS の正式サービス開始初年度であり、特に年度前半は利用者の絶対数が少ないため単純比較はできないが、2 章で問題点として挙げた利用時期の偏りについては概ね解消され、繁忙期の混雑の軽減と閑散期の利用率向上には成功したと言える。しかしその結果として、ほぼ 1 年を通して平均待ち時間が 10 時間を超えるという事態となった。最大の要因はやはり汎用 CPU ノード群への利用の偏りであるが、なかなか有効な解決策を打てないでいる。

4.2 今後の対策

今後着手しようとしている取り組みとしては、コア単位でのジョブ実行が挙げられる。当センターは長らく 1 ノード 1 ジョブ運用を行っており、これは、あるジョブが実行されているノードは当該ジョブに占有され、他のジョブと同時利用されることがないというものである。複数ジョブでの同時利用による性能低下の心配がない、利用時間の管理が簡単などのメリットがあるが、ジョブの利用資源量に関わらずノードが占有されるという大きなデメリットがある。例えば、OCTOPUS の汎用 CPU ノードは 1 ノードあたり 24 コア有してい

るが、並列化されていない 1 コア実行のジョブであっても 1 ノード（= 24 コア）占有するため、その間は 23 コア分の実行時間が無駄になっていると言っても過言ではない。ここで、ノード単位ではなくコア単位で資源を確保するよう設定し、1 ノードで複数のジョブが同時実行されるようにすることでこの問題は回避できる。先の例で言えば無駄になっていた 23 コア分の実行時間を活用できるようになるということであり、混雑緩和への効果を期待できる。

複数ジョブで 1 ノードを共有することで性能が低下し、ジョブあたりの実行時間が長くなる可能性はあるが、待ち時間の短縮に比べれば無視できるレベルのものであると思われる。実行時間が長くなるとポイントの消費量が増えるというデメリットもあるが、ノード単位での利用か、コア単位での利用かをユーザ側で選択できるようにすれば問題ない。先のアンケートでもポイント消費量より待ち時間の削減を優先するユーザが多かったため、選択式にしても一定のユーザはコア単位での利用を選択すると考えられる。また、コア単位での提供に対応することで GPU の分割提供も可能となる。GPU ノードには 4 基の GPU が搭載されているが、現状は 1 基のみ利用するジョブが多いように見受けられるため、GPU ジョブの混雑緩和にも効果的であると思われる。

しかしながら、システムの設定変更や利用制度の新設、規定の改正等の作業が伴うため、実装した場合に起こりうる問題やデメリットなどを十分に吟味した上で進める必要がある。また、あくまで 1 ノードで複数ジョブが同時に実行されるようにすることで提供可能なノード時間を擬似的に増やすという対症療法的な対策であり、利用ノード群の偏りに対する根本的な解決にはならない。

利用ノード群の偏りによる混雑を解決するだけであれば、OCTOPUS ポイントでの一元的な管理を廃止し、ノード群ごとの個別の利用申請にするという案がある。これによりそれぞれ個別に提供資源量上限を設けられるため、汎用 CPU ノード群の申請量が上限に達した時点で申請受付を打ち切れば、利用時期の集中による一時的な混雑はあるにせよ、慢性的な混雑は解消すると予想される。しかし、これでは OCTOPUS の一つの売りである複数のアーキテクチャを横断的に利用できる自由度の高さを犠牲にすることになる。今の OCTOPUS にとっては混雑の解消が最重要課題であるため一考の余地はあるが、本案もシステム改修など大掛かりな作業を伴うため、いずれにしてもメリット/デメリットをしっかりと考えたうえで判断する必要がある。

また、今回はジョブの平均待ち時間を成果の簡単な指標として示したが、これだけで混雑状況を正確に把握することは難しい。大規模並列の長時間ジョブはノードの確保が難しく待ちが長時間化してしまう一方、1 ノードの短時間ジョブは待ち時間が短くなりやすいほか、Xeon Phi などの空いているノード群に投入されたジョブが多いほど平均待ち時間も短くなるなど、同じ 1 件のジョブとして扱うには各ジョブの実行条件が違いすぎる。それでもなお平均して 10 時間以上の待ちが発生しているのは大問題だが、状況を詳しく把握し、有効な対策を考えるにはジョブの実行条件を加味した統計的な分析が必要であると考えられる。コア単位での提供に対応したところで、1 コア～数コア利用のジョブが少なければ期待した効果は得られないし、GPU の分割提供も同様である。対策効果の予測のためにも、より詳細な分析が必要である。もちろん統計データの収集は行っているが、各種資料の作成等に利用しているに過ぎず、データを有効活用できていないというのが現状である。

5 まとめ

本稿では OCTOPUS の混雑緩和に向けて行った様々な取り組みを紹介した。結果としては大きな成果を挙げられず、2020 年度も変わらず混雑した状況が続いている。OCTOPUS の運用においては、ヘテロジニアスなシステム特有の各ノード群の需要や特色の違いに起因する扱いの難しさを痛感すると同時に、結局のところ、高性能な加速器や特殊なアーキテクチャの計算機よりも純粋に性能の高い、使い慣れたスカラ型プロセッサを搭載したシステムが (少なくとも当センターのユーザでは) 頭一つ抜けて需要が高いということを知らしめられる結果となった。

また、今回様々な取り組みや改善案を述べてきたが、個人的には 2021 年度に運用開始予定の次期スパコンが OCTOPUS の混雑をすっかり解消してしまうのではないかと思っている。次期スパコンも OCTOPUS と同じくヘテロジニアスなシステムとなる予定だが、OCTOPUS の反省を活かし、スカラ型プロセッサを搭載したノード群は非常に大規

模なものになるよう仕様を定めている。搭載されるプロセッサも OCTOPUS より新しい世代のものになるよう調達を進めており、このシステムが稼働を始めたあかつきには多くのユーザが OCTOPUS から流れることが予想される。現在は利用率の高さに頭を悩ませているが、場合によっては利用率の低さに悩まされることになるかもしれない。

前述の通り、次期スパコンは OCTOPUS の運用経験を活かして調達を行っているシステムであるので、OCTOPUS の運用を通して次期スパコンにも通用するノウハウを蓄積しつつ、また、次期スパコンとの同時運用が始まっても多くのユーザに利用される人気なシステムであり続けられるよう、引き続き OCTOPUS の課題解決と改善に向けた様々な機能追加/改修、制度改正等の取り組みを進めることとする。

参考文献

[1] 伊達進,木越信一郎, 全国共同利用大規模並列計算システム調達の背景, 大学推進 ICT 推進協議会 2017 年度年次大会, 2017 年 12 月.

[2] OCTOPUS ポイントについて, http://www.hpc.cmc.osaka-u.ac.jp/system/manual/octopus-use/octopus_point/

[3] 伊達進, 片岡洋介, 五十木秀一, 勝浦裕貴, 寺前勇希, 木越信一郎, OCTOPUS のクラウドバースティング拡張, 大学推進 ICT 推進協議会 2019 年度年次大会, 2019 年 12 月.

[4] Microsoft Azure, https://azure.microsoft.com/ja-jp/

[5] Oracle Cloud Infrastructure, https://www.oracle.com/jp/cloud/infrastructure/iaas/overview/