OCTOPUS のクラウドバースティング拡張

伊達　進 [1)], 片岡洋介 [3)], 五十木秀一 [4)], 勝浦裕貴 [2)], 寺前勇希 [2)], 木越信一郎 [2)]

1) 大阪大学サイバーメディアセンター 2) 大阪大学 情報推進部
3) 日本電気株式会社　第一官公ソリューション事業部
4) 日本マイクロソフト株式会社 デジタルトランスフォーメーション事業本部

date@cmc.osaka-u.ac.jp

OCTOPUS with Cloud Bursting Functionality

Susumu Date[1)], Hiroaki Kataoka[3)], Shuichi Gojuki[4)],
Yuki Katsuura[2)], Yuki Teramae[2)], Shinichiro Kigoshi[2)]

1) Cybermedia Center, Osaka University 2) Dep. of Info. and Comm. Tech. Service, Osaka University
3) 1st Gov. and Pub. Solutions Division, NEC Corp. 4) Customer Success Unit, Microsoft Japan Co., Ltd.

概要

大阪大学サイバーメディアセンターでは、総理論演算性能 1.463 PFlops を有するスーパーコンピュータシステム OCTOPUS (Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer) を 2017 年 12 月に導入し、4 ヶ月間の試験運用の後、2018 年 4 月より本格運用している。本システムは導入当初より故障も少なく、安定したシステム運用を可能とする一方、利用者からの高い計算需要から高負荷状態が継続し、利用者の計算要求から計算完了までの待ち時間が増大化している。本研究では、このような背景から、OCTOPUS の高負荷状態の緩和を目的とし、クラウドベンダの提供するクラウド資源への負荷のオフロードによる問題解決を試行する。具体的には、オンプレミス環境の OCTOPUS と、マイクロソフト社の提供する Azure サービスを連動させたクラウドバースティング実証環境を構築し、OCTOPUS の高負荷状態を回避する手法の実現可能性を検証する。本稿では、クラウドバースティング実証環境を概説し、本稿執筆時までの検証成果を報告する。

1　はじめに

大阪大学サイバーメディアセンター (CMC: Cybermedia Center) では、2017 年 12 月にハイブリッド型クラスタシステム OCTOPUS (Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer) を導入した。CMC の利用者からのスカラ型スーパーコンピュータに対する多様な計算ニーズ・需要を収容可能であり、利用者に定常的かつ安定的に高い性能を提供することが期待される中で導入された OCTOPUS は、非常に高い利用率で利用される状況となっている。しかし、その一方で、利用者の計算要求から計算完了までの待ち時間が定常的に長時間になるという新たな問題が深刻になりつつあり、利用者からの問い合わせ・相談の声も大きくなりつつある。正式運用 2 年目となる 2019 年度においては、この待ち時間の問題は年度の早期段階から顕著になりつつあり、CMC の大規模計算機システム事業の利用者満足度を向上していく上で重大な問題となっている。

一方、近年では、利用者の計算ニーズ・要求に基づく、個別のソフトウェアスタックを配備する計算資源群を、オンデマンドに必要量だけ利用可能な IaaS (Infrastructure as a Service) 型クラウドサービスが成熟しつつある。Amazon AWS [2]、Microsoft Azure [3] 、Oracle Cloud [4] は、クラウドベンダの提供する IaaS 型クラウドサービスの一例である。今日、これらのクラウドサービスを活用することで、高性能計算、分散計算、ネットワーク等の深い知識を有していない利用者でも、直感的なグラフィカルインタフェース (GUI) を通じて、大規模な計算環境を容易に構築することが可能である。また、クラウドベンダの提供する IaaS 型クラウドサービスでは、クラウドベンダの保有する膨大な計算機資源量のため、その利用者はその資源量、負荷状況を気にすることなく利用できる。さらに、これらのクラウドベンダの提供する計算機資源は、クレジットカード決済により気軽に購入可能である。このような計算環境配備のオンデマンド性、簡易性、手軽性は、大学の計算機センターの提供する計算機資源の一部あるいは全部をクラウドベンダの提供する IaaS 型クラウドサービスで代用することへの関

心を高めている。

上述した OCTOPUS の高負荷状況、および、IaaS 型クラウドサービスへの関心を鑑み、本稿では、OCTOPUS と IaaS 型クラウドサービスを連動させることにより、オンプレミス環境である OCTOPUS の負荷を一時的にクラウド環境にオフロードできる計算環境を試行する。より技術的かつ具体的には、CMC の OCTOPUS とマイクロソフト社の提供する Azure サービスに、近年急速に期待と関心を高めているクラウドバースティング技術を応用することで、本研究では、OCTOPUS の負荷を Azure 上に構築した計算機資源にオフロードする。その際、本研究で構築する OCTOPUS-Azure クラウドバースティング環境での検証を通じて、将来の本格運用にむけた技術課題の抽出を行いつつ実現可能性を考察・評価する。

本論文は以下の通り構成する。2 節では OCTOPUS の構成、現状の問題を記し、クラウドバースティング機能拡張にむけた要求要件をまとめる。3 節では、本研究で構築したクラウドバースティング環境について解説する。その後、4 節にて本稿執筆時点までに得られた知見をまとめる。5 節にて本論文をまとめる。

2　クラウドバースティングへの要求要件

本節では、OCTOPUS の構成と現状を紹介し、クラウドバースティングに向けた要求要件をまとめる。

2.1　OCTOPUS の構成

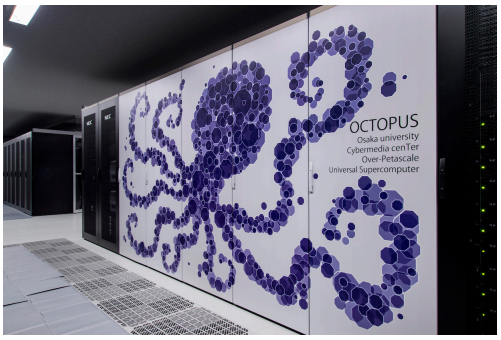

図 1　OCTOPUS の外観.

図 1 に OCTOPUS の外観を示す。OCTOPUS は、Intel 製プロセッサ (Skylake) を搭載した汎用 CPU ノード群、Intel 製プロセッサに加え NVIDIA 製 P100 を搭載した GPU ノード群、Intel 製プロセッサ (Knights Landing) を搭載した Xeon Phi ノード群、6TB の主記憶を搭載した大容量主記憶搭載ノード群、3PB の大容量ストレージ、利用者がプログラム開発を行うフロントエンドサーバ群から構成されるハイブリッド型クラスタシステムである [5]。これらのノード群、ストレージは、Infiniband EDR による相互結合網に接続されており、100Gbps の高帯域な通信が可能となっている。また、OCTOPUS では、上述の全てのノード群、ストレージは、Mellanox 製 InfiniBand EDR ディレクタースイッチ CS7500 1 台に収容されており、相互に 1-hop の低遅延での通信が可能である。さらに、汎用 CPU ノード群、GPU ノード群、Xeon Phi ノード群については、それらを構成する全てのプロセッサ、アクセラレータを直接液冷しており、高い性能を安定的に供給できるように構築されている。

OCTOPUS の運用では、OCTOPUS の計算ノード群をノード時間単位で利用者に負担金を課金する。すなわち、CMC では、ある計算ノード 1 台を 1 時間利用した場合の消費電力に相当 (冷却に伴う電力量を含む) する費用を基に、OCTOPUS の負担金制度が構築されている。このようなノード時間単位での利用負担金制度を行なうため、CMC では、計算ノードには同時に 2 つ以上の利用者ジョブ要求が割り当てられないように制御し、あるジョブ要求が他のジョブ要求と課金対象となる計算資源を共有する状況を回避している。

OCTOPUS では、このような計算機提供方法を実現するとともに、利用者が OCTOPUS を効率よく利用できるよう、NEC 製 NQSII、および、JobManipulator(NQSII/JM) [6][7] をジョブ管理サーバとして導入している。この NQSII/JM は利用者からのジョブ要求を受けつけ、各ノード群へのジョブ割り当てを行う機能を持つ。OCTOPUS は複数の異なるノード群から構成されるため、各ノード群に対応した実行用ジョブクラスに利用者のジョブ要求を振り分ける。この実行用ジョブクラスはさらに要求並列度数に基づいて複数段階に細分化されており、ジョブ要求の並列度数ごとのジョブクラスへと振り分ける。この際、上述したような状況が発生しないような制御も行いつつ、利用者の待ち時間を縮減するとともに、システム全体が定常的に高い利用率を維持できるようにしている。さらに、実際の運用では、OCTOPUS の管理者は、ジョブ投入状況、利用者の待ち時間等を監視し、ジョブクラスへマッピングされるノード群の設定を変更するなどの対応により、利用者の待ち時間の縮減、ジョブスループットの向上にむけて日々尽力している。

2.2　OCTOPUS の現状

しかし、高負荷状態を回避しようとする運用努力にも関わらず、導入当初から OCTOPUS の汎用ノード群、GPU ノード群は利用率が高い状態が続いている。

OCTOPUS の正式運用を開始して 2 年目となる 2019 年度においては、年度当初より汎用ノード群、GPU ノード群は 80% を超え 90% 程度の利用率で利用される状況が継続している。また、Xeon Phi ノード群についても、CPU ノード群の高負荷状態時にその負荷のオフロード先として使われる傾向もあり、定常的に高い利用率となっている。CMC の大規模計算機システム事業としては、OCTOPUS の高い利用率を歓迎していたが、利用者の待ち時間の増大が利用者の研究活動の遅延を招いてしまうことで、全国の大学の研究者が学術研究・教育に伴う計算及び情報処理を行う全国共同利用施設としての CMC の機能不全を起こしてしまうことを懸念している。

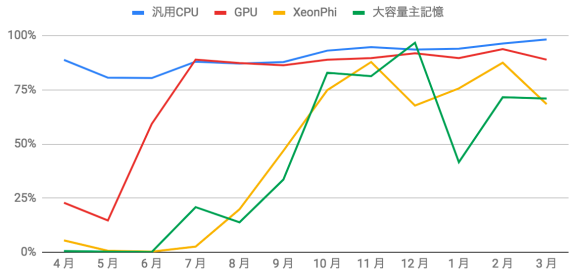

図 2　2018 年度の OCTOPUS の利用率 (月平均).

そのような懸念から、2018 年度の OCTOPUS 利用率が年度当初から年度後半にむけて急激に高まった傾向に着眼し、2019 年度においては、年度開始当初の 3 ヶ月 (4-6 月) における計算機利用負担金を、それ以外の期間 (7-3 月) における計算機利用負担金よりも軽減する季節係数の制度 [8] を導入し、利用者の利用を年度前半に誘導することを試行している。本稿執筆時点において、ある程度の計算負荷を年度早期に誘導できる効果が得られていると考えられる。しかし、2018 年度の状況 (図 2) を考慮すれば、それ以降の期間においても依然高い利用率および長い待ち時間が継続していく状況が予想され、さらなる対策の実施が急務であると考えている。

2.3　クラウドバースティング機能拡張への要求要件

本研究では、OCTOPUS のクラウドバースティング機能拡張にむけ、下記 5 点の要求要件を設定した。

1. オンデマンド性：クラウド資源の利用は必要に応じて最小限であるように構成すること。
2. 透過性：利用者は、クラウドとオンプレミス環境で異なるジョブ投入方法とならないこと。
3. 選択性：利用者は、クラウド資源の利用可否を選択できること。
4. 同一性：クラウド環境とオンプレミス環境での計算結果に相違がないこと。
5. ハイスループット性：システム全体としてのスループット向上がなされること。

以下、これらの要求要件について説明する。

2.3.1　オンデマンド性

2.2 節で記したように、OCTOPUS は高負荷状況が継続し、利用者の長い待ち時間が深刻化しつつある。その対策の一つとして、OCTOPUS へのクラウドバースティング機能拡張に期待を寄せているが、クラウドバースティング機能の実現可能性・有用性が検証できれば、将来的には、CMC の管理者の判断により適宜クラウド資源の利用可否を制御したいと考えている。つまり、OCTOPUS の高負荷状況が継続し、長い待ち時間が発生しているような状態が検出された際には、ある一定量のクラウド資源を CMC が準備し、ジョブ管理サーバのキューで待ち状態になっているジョブ要求の一部をクラウド資源に誘導する。また、待ち時間が緩和された際、CMC の判断でジョブ要求のクラウドへの誘導を停止したり、あらかじめ準備したクラウド資源量が消費された際は、ジョブ管理サーバによってクラウド資源へのジョブ要求の誘導を停止する。このように、できる限りオンプレミス環境で計算実行が行われることを前提とし、管理者の判断によって必要な資源量を必要な時にだけ提供できる必要がある。

2.3.2　透過性

CMC の判断によってクラウド資源の利用を制御し、待ち状態にあるジョブ要求をクラウドに誘導する場合を想定すれば、クラウドとオンプレミス環境でジョブ投入方法に相違がない必要がある。例えば、クラウド環境とオンプレミス環境での利用者 ID やパスワードの相違は、その相違を解決するための ID 連携などの技術が新たに必要となる。また、クラウドとオンプレミス環境で利用するジョブ管理システムの相違はジョブ要求スクリプトの記述の相違を生む。このため、利用者にはジョブ実行の際にオンプレミス環境とクラウド環境での相違を意識させないことが望ましい。

2.3.3　選択性

オンプレミス環境での計算を前提としたサービス提供を行う視点から、利用者の中には、データセキュリティなどへの懸念等の理由から、高負荷状態においてもクラウド資源への誘導を望まない場合がある。こうした場合への対応のため、利用者のクラウド資源利用

可否の選択を確認する何らかの手法が必要となる。

2.3.4　同一性

オンプレミス環境とクラウド環境での計算結果は同一である必要がある。利用者がクラウド利用について承認し、その結果、システムの高負荷状況が改善するとともに利用者の待ち時間が縮小したとしても、2 つの環境での計算結果に相違があることは許されない。

2.3.5　ハイスループット性

構築するクラウドバースティング機能によって、ジョブの待ち時間の軽減とオンプレミス環境の負荷軽減がなされる必要がある。

3　クラウドバースティング実証環境

本節では、本研究で構築した OCTOPUS-Azure クラウドバースティング実証環境についてまとめる。

3.1　クラウドバースティング実証環境の概要

Microsoft Azure は、HPC 分野で利用される、Intel あるいは AMD 製の高性能 CPU、NVIDIA 製 GPU (Kepler, Maxwell, Pascal, Volta)、低遅延高帯域幅ネットワーク InfiniBand 等から構成される様々な仮想マシンを世界中のリージョンで提供している。この Azure を利用することで、利用者の目的・用途にあわせたスーパーコンピュータをオンデマンドに構築可能である。また、Azure では、最新の CPU、GPU、高性能なノード間通信等の HPC 関連技術がリリースされた際には、対応する仮想マシンの迅速な提供が予定されている。加えて、大阪大学では、学習の効率化、学習環境の質の向上を目的としてマイクロソフト社との包括契約 (EES) を締結済みであり、Azure サービスの追加修正は比較的容易である。このような理由から、本研究では、OCTOPUS のクラウドバースティング機能拡張にむけ、Microsoft Azure を採用した。

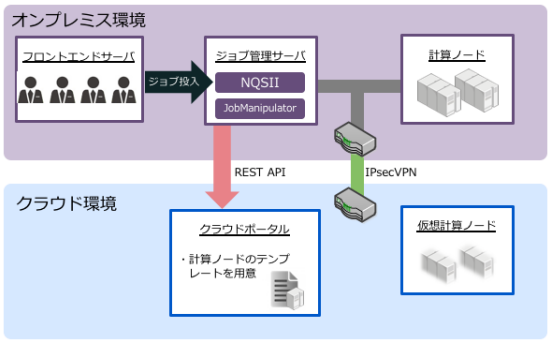

図 3　OCTOPUS-Azure クラウドバースティング環境の概要.

図 3 に構築した OCTOPUS-Azure クラウドバースティング環境の概要を示す。オンプレミス環境 OCTOPUS から Azure を利用するクラウドバースティング機能を実現するために、本研究では、主として (1) クラウドブリッジネットワーク、(2) ジョブ管理サーバ、(3) クラウドへの OCTOPUS 計算ノードイメージの配備の 3 点の拡張を実施した。これらの拡張を通じて、利用者がジョブ管理サーバにあらかじめ設置された実行ジョブクラスにジョブ要求を投入した際、ジョブ管理サーバは Azure の提供するクラウドポータルにあらかじめ配備しておいた計算ノードのテンプレートイメージに基づき、Azure 上に仮想計算機として実装された OCTOPUS 計算ノードを起動し、要求されたジョブを実行することが可能となる。

以下、主要 3 点の拡張部分について記す。

3.2　クラウドブリッジネットワーク

2.3 節で示したように、OCTOPUS-Azure クラウドバースティング環境では、利用者はオンプレミス環境とクラウド環境の区別なく利用できなければならない。そのためには、OCTOPUS の計算ノード群と同様に、Azure 側で構成される仮想計算ノード群を、後述するジョブ管理サーバから監視管理させる必要がある。また、Azure 側に配備される仮想計算ノードとオンプレミス側に配備される計算ノードに差異が生じないように、構成されるソフトウェアスタックを同一にするだけでなく、OCTOPUS の大容量ストレージも同一の操作性で利用できるようにする必要もある。このような状況を考慮すれば、Azure クラウド環境上で配備される仮想計算ノードはオンプレミス側と同等の論理ネットワーク内で稼働できることが望ましい。

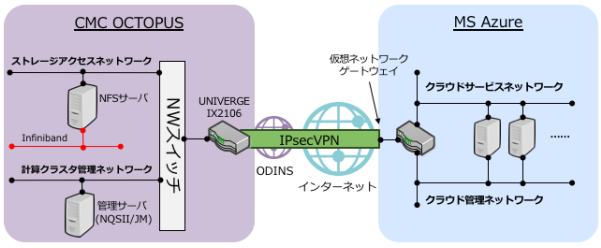

図 4　ブリッジネットワークの構成.

図 4 に本研究で実現した OCTOPUS-Azure 間のクラウドブリッジネットワークの概要を示す。本研究では、OCTOPUS-Azure 間のネットワークの実現に際して、IPsec による VPN (Virtual Private Network) を構築する。これにより、OCTOPUS 内のプライベートな論理ネットワークを Azure 上の仮想ネットワークへ延伸する構成とした。このために、オンプレミス

側では NEC 製 UNIVERGE IX2106 を導入し、大阪大学の学内ネットワーク ODINS、インターネット経由で、Azure 側の仮想ネットワークゲートウェイとの間に VPN 接続を行った。クラウド側では計算ノードの管理を行うためのクラウド管理ネットワーク、計算ノード間のプロセス間通信や、OCTOPUS 側のストレージへのトラフィック等に利用されるサービスネットワークから構成した。

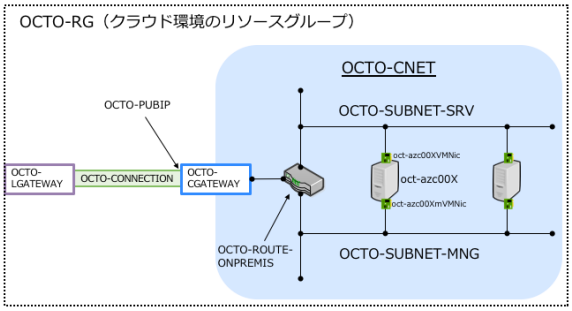

図 5　Azure 内ネットワークの構成.

図 5 に Azure 内に構築されたネットワークを示す。OCTO-LGATEWAY、OCTO-CGATEWAY、OCTO-CONNECTION は上述した OCTOPUS-Azure 間で IPSEC による VPN 接続を行うために、Azure 上で構築したリソースである。Azure 内には、OCTO-CNET という仮想ネットワークを構築し、クラウドサービスネットワーク OCTO-SUBNET-SRV、クラウド管理ネットワーク OCTO-SUBNET-MNG のサブネットを分割している。Azure 上に配備される計算ノード (oct-azc00X) には、それぞれのサブネットワークに対する仮想 NIC を設定する。これらの仮想 NIC を通じて OCTOPUS 側ネットワーク向けに送受信される計算ノードのトラフィックは、OCTOPUS 側の対応するネットワークへルーティングされるようルーティングテーブル OCTO-ROUTE-ONPREMIS を設定している。

3.3　ジョブ管理サーバの拡張

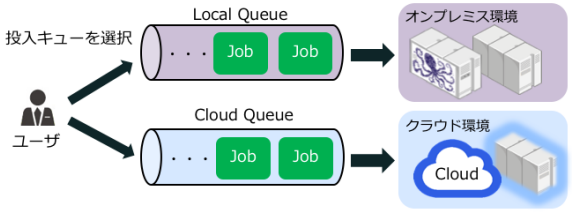

図 6　ジョブ管理サーバのキュー構成.

2.3 節で記したように、クラウド利用を許容できる利用者にとっては、クラウド環境とオンプレミス環境でジョブ投入方法に差異が生じないほうが望ましい。また、そのような利用者にとっては、クラウド資源とオンプレミス環境のいずれであってもジョブが迅速に完了することが望ましい。そのような観点からは、ジョブ投入が行われるジョブ管理サーバが、クラウド環境とオンプレミス環境の利用状況等を基に、ジョブの割り当てノードをを動的に制御できる必要がある。

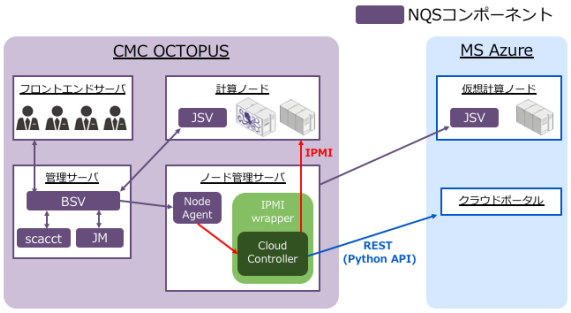

図 7　OCTOPUS-Azure 間でのジョブ管理サーバの構成.

しかし、本研究段階では、システム全体のスループットを向上させながら、利用者の待ち時間縮減を実現するために、投入されたジョブ要求の中からどのジョブを優先的にクラウド資源へと誘導するかを決定するアルゴリズムや運用ポリシは未実装である。そのため、本研究では、オンプレミス側 OCTOPUS のキューとは独立に、新たにクラウド資源投入用のキューを設定した (図 6)。実際運用の視点では、CMC の管理者判断により当該キューへのジョブ投入可否を設定し、クラウドバースティング機能を制御することを想定している。

図 7 に OCTOPUS-Azure 環境に配備したジョブ管理サーバの構成を示す。NQSII/JM は、ジョブの受付・実行を管理するバッチサーバ (BSV)、スケジューリングを行うスケジューラ (JM)、管理対象ノードを監視するジョブサーバ (JSV)、統計情報を管理するアカウンティング (SCACCT)、計算ノードの障害検知・電源制御を行うノード管理エージェント (NodeAgent) から構成される。

2.3 節で示した、クラウド資源のオンデマンド性を実現するためには、クラウド資源投入用のキューにジョブ投入がなされた場合に Azure 上の計算ノードを起動させ、ある一定期間利用がない場合に Azure 上の計算ノードを停止させる機能を OCTOPUS-Azure 環境上に実装しなければならない。本研究では、投入されたジョブの要求量に応じて、クラウド資源の起

動・停止を行う機能をジョブ管理サーバに実装した。NQSII/JM には、消費電力を最小限に抑えながらジョブの要求量に応じてオンプレミスの計算ノードの起動制御を行うことが可能な省電力運用機能が備わっている。本研究では、本機能をクラウド上の計算ノードに対応できるよう拡張するために、ノード管理エージェントから起動される IPMI Wrapper およびクラウドコントローラを開発した。

3.4 OCTOPUS 計算ノードイメージの配備

OCTOPUS の汎用 CPU ノード群は、Intel 製 Xeon Gold 6126(Skylake/ 2.6GHz 12cores) を 2 基、主記憶を 192GB 搭載した計算ノード 236 台が InfiniBand EDR の相互結合網に接続された構成である。本研究では、OCTOPUS で最も高い利用率を観測している汎用 CPU ノード群の負荷のオフロードを試行する。OCTOPUS を利用する利用者の全てのジョブ要求を Azure への誘導対象とするためには、利用者が OCTOPUS へのジョブ投入時にジョブ管理サーバに対して要求を指定するメモリ容量、コア数以上の高性能な計算資源を配備しておく必要がある。また、OCTOPUS-Azure 環境には、オンプレミスおよびクラウド間のデータ通信を考慮して、OCTOPUS とクラウド環境を接続するネットワークには広帯域性と低遅延性が要求される。このような背景から、本研究では、基盤となるネットワーク性能を重視し、西日本リージョンの Azure サービスを選択した。さらに、仮想計算機の利用に際しては、搭載主記憶容量は OCTOPUS の半分となる 96GB であるが、Intel 製 Xeon Platinum 8168(Skylake / 2.7GHz 24cores) を基盤とし、48 仮想コアを備えた仮想計算機 F48s_v2 を基に、OCTOPUS 計算ノードイメージ 32 セット (oct-azc001〜oct-azc032) を構築した。同時に、利用者がオンプレミス側、クラウド側の区別なく、利用者 ID/パスワード、同一のホーム領域で、ジョブ実行ができるよう、クラウド上に配備された計算ノードイメージはオンプレミス側に配備された同一のソフトウェア構成・設定を維持しつつ、Azure 環境向け汎用化といったクラウド特有の設定を行なっている。

4 検証と考察

本節では、2.3 節で示した 5 点の要求要件視点から、定性的かつ定量的な検証と考察を行う。

4.1 オンデマンド性

OCTOPUS の計算負荷をクラウド環境にオフロードするためには、3.3 節で示したジョブ管理サーバに設置しているクラウド資源投入用キューへのジョブ投入可否を設定するだけで可能となる。この投入可否の設定は、NQSII/JM で準備されている標準コマンド *qmgr* を用い、図 8 に示すように管理者権限で直感的かつ簡便な操作を通じて、OCTOPUS の負荷状態に合わせて管理者判断で迅速に行うことができる。図例では、Azure の仮想計算ノード群に対して INT-AZT および O-AZT の 2 つのキューの enable 設定をしている。そのため、クラウド資源を必要な時だけ利用可能とすることにより、クラウド資源に費やすコストを小さくすることができる。

さらに、NQSII/JM の省電力機能を適用し、クラウド上で稼働する計算ノードに 5 分以上ジョブ投入がない場合は、本研究で構築したジョブ管理サーバによって当該計算ノードは自動的に停止状態となる。また、ジョブ投入がなされた場合は、ジョブ管理サーバが当該計算ノードを起動する。これにより、クラウド上の計算ノードの稼働状態を最小に抑えることができる。

```
octopus $  qmgr -Pm
 Mgr: enable interactive_queue=INT-AZT
 Mgr: start interactive_queue=INT-AZT
 Mgr: enable interactive_queue=O-AZT
 Mgr: start interactive_queue=O-AZT
 Mgr: exit
```

図 8　クラウド資源投入用キューの制御.

4.2 透過性

2.1 節で記したように、OCTOPUS へのジョブ要求は、図 9 に示すようなジョブスクリプトを記載し、NQSII/JM に準備された標準コマンド *qsub* にジョブスクリプトを指定する。OCTOPUS-Azure 環境では、利用者は前述した O-AZT あるいは INT-AZT を (キュー名) に指定するだけで、Azure 上の計算ノードが利用可能となっている。

```
#!/bin/bash
#PBS -q (キュー名)
#PBS -l elapstim_req=1:00:00
cd $PBS_O_WORKDIR
./a.out
```

図 9　クラウド資源投入用キューへのジョブ投入.

本論文の研究段階では、利用者はキューの相違を意識しなければならず、利用者はクラウド環境とオンプレミス環境の相違を意識せざるを得ず、今回構築した

OCTOPUS-Azure 環境に高い透過性があるとは言い難い。この点については、後述の選択性とあわせた検討が今後必要になると考えている。

4.3 選択性

クラウド環境にデータを置きたくない、データセキュリティが気になる、等の理由から、クラウドの利用を許可・承認できない利用者も数多く存在する。本稿執筆時点での OCTOPUS-Azure 環境では、上述の通り、利用者はオンプレミス環境およびクラウド環境へのジョブ投入をジョブスクリプト内で指定でき、クラウド利用とオンプレミス利用の選択性があるといえる。しかし、本研究で目的とする OCTOPUS 高負荷状態の改善を考慮した場合、現状のように利用者がクラウド資源へのジョブ投入を行うのではなく、ジョブ管理システムが OCTOPUS に投入されたジョブ要求から、高負荷状態を改善できると考えられるジョブ要求をクラウド環境に誘導できることが望ましい。

本研究では、このような視点から、今後、図 9 に示したジョブスクリプトを拡張し、例えば、クラウド利用の可否を利用者に指定してもらうための #PBS -cloud yes といった拡張を検討していきたいと考えている。利用者のクラウド利用の透過性と選択性のトレードオフ問題は本研究で構築したクラウドバースティング機能における将来課題の一つと考えている。

4.4 同一性

本研究では、OCTOPUS および Azure における計算実行に差異が発生しないことを確認するために、MPI (Message Passing Interface) の一実装である MPICH 3.3.1 [10] に付属の円周率の計算を行うサンプルプログラム *cpi.c*、および、本センターの利用者による Fortran コードを IntelMPI でコンパイルし、OCTOPUS および Azure 環境上で実行した。CPI はノード間並列、利用者コードはノード内並列を行なうが、いずれも結果が同一であることを確認した。

4.5 ハイスループット性

ハイスループット性評価のために、OCTOPUS-Azure 間クラウドブリッジネットワークの性能、Azure 上の仮想計算ノード間データ通信性能、および仮想計算ノード上の計算性能について計測した。

4.5.1 OCTOPUS-Azure 間クラウドブリッジネットワークの性能

OCTOPUS-Azure クラウドブリッジネットワークの性能を評価するため、OCTOPUS のフロントエンドノードから OCTOPUS の汎用 CPU 計算ノードおよび Azure 側の仮想計算サーバに対して、*ping* コマンドにより RTT(round-trip time) を計測した。それぞれのサーバに対する計測は各パケットサイズに対して 100 回行なった。また、この計測を朝、昼、夕方に 5 セット回実施した。

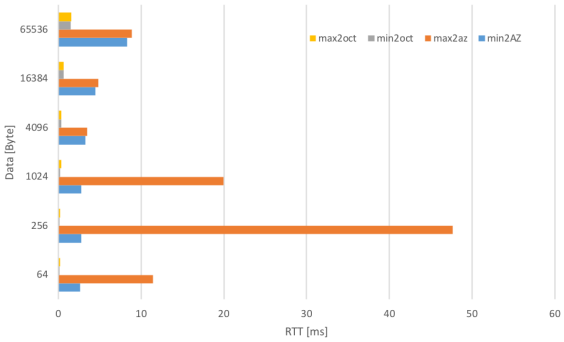

図 10 OCTOPUS フロントエンドからの RTT.

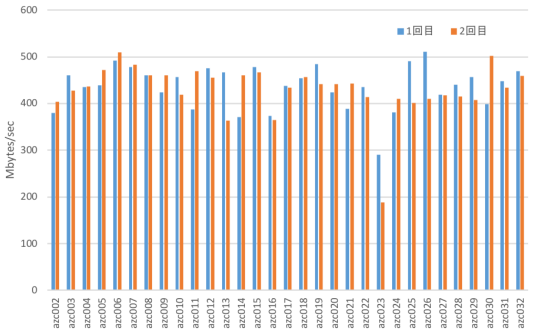

図 11 MPI PingPong 計測結果 (データサイズ = 4MB).

図 10 に OCTOPUS 計算ノードへの RTT 計測値の最大値と最小値、および Azure 仮想計算ノードへの RTT 計測値の最大値と最小値を示す。この結果から、OCTOPUS に対しては、全体の計測を通じて RTT は小さく安定していた一方、Azure に対しては、最大値と最小値で大きく差が開いたことがわかる。

4.5.2 Azure 内仮想計算ノード間データ通信性能

Azure へのクラウドバースティングのため配備した 32 台の仮想計算ノード間での通信性能を検証するため、oct-azc001 から他の仮想計算ノード全てに対して MPI PingPong によって計測した。図 11 にデータサイズ 4MB 時の計測結果を示す。この結果から、ノードの組み合わせによってスループット性能にばらつきがあることがわかる。今回の計測では、平均 432.6871MB/sec の性能が確認できたが、最大性能は 510.35MB/sec、最小性能は 188.6MB/sec であった。

本計測で観測された性能のばらつきは利用者の実行ジョブ時間に影響を及ぼしうる。また、この特性は時間的要因によりも変動しうることも観測された。本計

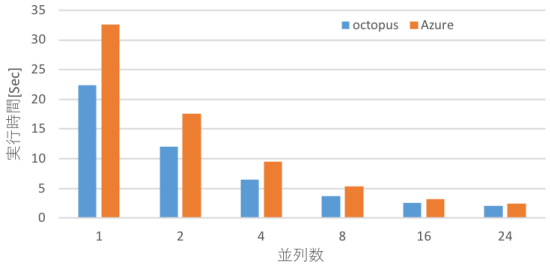

図 12　GROMACS を用いた性能比較.

測実験後の調査の結果、この性能特性は、Azure 上での仮想計算ノードに仮想スイッチがデフォルトで組み込まれることに一因があることが判明している。今後、仮想計算ノードの構築時に仮想スイッチを利用しない Accelerated Network [9] の設定に切り替え検証を進めていく予定である。

4.5.3　Azure 仮想計算ノードの性能

OCTOPUS 汎用 CPU ノードと Azure 上の F48s_v2 による仮想計算ノードで、CMC においても利用者が多い GROMACS [11] を用いて性能を比較した。なお、本稿執筆時において、OCTOPUS では CPU 利用向けに GROMACS 2016.1 を導入している [12]。本研究では、この CPU 版 GROMACS を用い、OCTOPUS 汎用 CPU ノードと Azure の仮想計算ノードでの実行時間を計測・比較した。

図 12 に計測結果を示す。横軸は GROMACS の実行に用いたノード内コア並列数である。今回構築した Azure 上の仮想計算ノードでは、OCTOPUS 汎用 CPU ノードよりも高性能なプロセッサを基盤としているが、GROMACS の性能については OCTOPUS 以上の性能が観測できなかった。本稿執筆時点においては、GROMACS の性能低下要因を未解析であり、引き続き調査・検証を続けて行くことを予定している。さらに、実際に OCTOPUS の負荷をオフロードし、OCTOPUS-Azure 環境として高いスループットが得られ、結果として利用者の待ち時間が縮減できるかどうかについては、実際ジョブセットを投入して検証を行う必要があり、本研究の今後の課題である。

5　結論

本稿では、OCTOPUS と IaaS 型クラウドサービスを連動させることにより、OCTOPUS の負荷を一時的に Azure にオフロードできるクラウドバースティング環境の実現可能性を試行した。具体的には、構築した OCTOPUS-Azure 環境について説明し、本稿執筆時点までに得られた、オンデマンド性、透過性、選択性、同一性、ハイスループット性の観点からの検証結果および考察を報告した。本稿での検証を通じて、OCTOPUS から Azure へのクラウドバースティングの実現可能性を高く評価する一方、実際運用にむけてはクラウド上に構築された仮想計算機の不安定な性能への対策、クラウドバースティング機能の選択性・透過性での機能追加、利用者管理システム等の運用支援システムとの親和性などの技術課題の必要性を本研究では再確認した。さらに、実際に OCTOPUS の負荷状態を軽減し、利用者の待ち時間を縮減できるかの検証は不可欠である。CMC では、今後も引き続き実際運用に向けたクラウドバースティング機能の検証・評価を進めていきたいと考えている。

参考文献

[1] OCTOPUS, http://www.hpc.cmc.osaka-u.ac.jp/en/octopus/.

[2] Amazon AWS, https://aws.amazon.com/.

[3] Microsoft Azure, https://azure.microsoft.com/.

[4] Oracle Cloud, https://cloud.oracle.com/.

[5] 伊達 進, ”ペタフロップス級ハイブリッド型スーパーコンピュータ OCTOPUS：Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer ～サイバーメディアセンターのスーパーコンピューティング事業の再生と躍進にむけて～”, HPC ジャーナル, サイバーメディアセンター, no. 8, pp. 3-20, Sep. 2018.

[6] ”NQSII 利用の手引 [管理編] 第 6 版”, 日本電気株式会社, 2018 年 6 月.

[7] ”NQSII 利用の手引 [JobManipulator 編] 第 3 版”, 日本電気株式会社, 2017 年 3 月.

[8] 季節係数, http://www.hpc.cmc.osaka-u.ac.jp/system/manual/octopus-use/octopus_point/#seasonality.

[9] ”高速ネットワークを使った Linux 仮想マシンの作成”, https://docs.microsoft.com/ja-jp/azure/virtual-network/create-vm-accelerated-networking-cli.

[10] MPICH, https://www.mpich.org/.

[11] GROMACS, http://www.gromacs.org/.

[12] GROMACS 利用方法 (OCTOPUS), http://www.hpc.cmc.osaka-u.ac.jp/system/manual/octopus-use/gromacs/.