The University of Osaka
Institutional Knowledge Archive

Title	高性能計算・データ分析基盤システム（SQUID: Supercomputer for Quest To Unsolved Interdisciplinary Datascience）：わが国の学術・産業を支える研究基盤の実現に向けて
Author(s)	伊達, 進
Citation	サイバーメディアHPCジャーナル. 2021, 11, p. 3-32
Version Type	VoR
URL	https://doi.org/10.18910/87667
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

高性能計算・データ分析基盤システム
(SQUID: Supercomputer for Quest To Unsolved Interdisciplinary Datascience)
～わが国の学術・産業の発展を支える研究基盤の実現に向けて～

伊達　進
大阪大学 サイバーメディアセンター

1. はじめに

大阪大学サイバーメディアセンター（以下、CMC）は、2014 年 11 月に導入されたベクトル型スーパーコンピュータ SX-ACE [1]の後継機として、2021 年 5 月より高性能計算・データ分析基盤システムの運用を開始した。本センターでは、この高性能計算・データ分析基盤システムを通じて、我が国の学術・産業を支える研究者による未解決の学際的なデータサイエンス問題への探求を支援できるよう決意と願いを込めて、Supercomputer for Quest to Unsolved Interdisciplinary Datascience (SQUID)と命名した。

この SQUID は総理論演算性能 16.591 ペタフロップス(Peta Flops) 、つまり、浮動小数点演算を 1 秒間に 1 京 6591 兆回実行可能な演算性能を有する本センター初の 10 ペタフロップス超の性能を有するスーパーコンピュータである。この性能値は、一世代前の SX-ACE の総理論演算性能 423 テラフロップス(Tera Flops)と比較すると、約 40 倍向上した数字であり、同時に、本センターが国内屈指の演算性能を有するスーパーコンピュータの導入に成功したことを示している。

HPC ジャーナル読者の方々の多くはご存知かもしれないが、国立大学でのスーパーコンピュータシステムの導入は非常に長い時間と労力のかかる仕事となる。国立大学では、一定以上の金額や性能を超えるスーパーコンピュータを導入する場合、特定調達という枠組みで、導入したいスーパーコンピュータの仕様書の作成、導入すべきスーパーコンピュータが満たすべきベンチマークテストの設計を行う必要がある。その後、本センターで作成した仕様書やベンチマークを基に、スーパーコンピュータシステムを提案できる候補ベンダらが仕様書を満たす提案と見積もり金額を作成し、入札期日までに入札を行うこととなる。その後、複数の候補ベンダらの提案が仕様書を充足しているかについて審査し、その充足度合いがあらかじめ仕様書やベンチマーク作成時に作成されていた得点表に基づき、技術点として点数化される。点数化された提案は所掌事務課に戻され、入札金額に基づいて得点化された価格点とあわせて評価され、最終的な落札者が決まる。厳密にいえば、所掌事務課が想定する落札金額と折り合いがつかない場合、再入札が行われるなどの流れがあるようだが、基本的な流れは上記の通りである。

このような調達の進め方は候補となるベンダ間の公平性と価格競争を促すことに目的の一因があることから説明できる。このような国が定めたルールに基づいてスーパーコンピュータシステムを導入しなければならないため、例えば、3 ヶ月後に新製品として発売される、このプロセッサとあのアクセラレータを 1000 基導入して、などと、簡単にはできないのである。それゆえ、国立大学のスーパーコンピュータの調達では、導入するスーパーコンピュータシステムの導入時期をにらみながら、限られた予算で将来可能であろう性能値を推測しつつ、利用できるであろう製品・ソリューションを仕様書に落とし込んでいくこととなる。この仕様書の作成やベンチマークの設計が、導入されるスーパーコンピュータシステムを決定づけることとなり、調達を担当する教職員が最も労力と時間をかける部分となる。逆に言えば、ここに労力と時間をかけず手を抜いてしまえば、仕様書が非現実的な技術要求を含んでしまい、その結果不調となる場合もある。この場合、調達手続きを最初あるいは途中からやり直しとなり、新しいスーパーコンピュータの導入を待ち望んでいる利用者を失望させてしまう。一方、手を抜いて作成した仕様書でも、もちろん、その仕様をみたすスーパ

ーコンピュータがどこかのベンダから導入されうる。しかし、この場合、調達金額に見合わずコストパフォーマンスが悪いスーパーコンピュータであったり、時代遅れのテクノロジーを採用したスーパーコンピュータであったり、利用者にとって使い勝手の悪いスーパーコンピュータが導入されるといった結果が起こる。こうした場合、直接的な被害を被るのはセンターの利用者であり、ひいてはわが国の国民の血税を無駄遣いしてしまうことになる。それゆえ、数十億円規模のスーパーコンピュータシステムの調達を担当する教職員は、調達失敗の不安感に苛まされながら、強い使命感と責任感で調達に向かい合わなければならないのである。

今回の SQUID も例外ではなく、調達には著者が所属する応用情報システム研究部門教員、情報推進部情報基盤課技術系職員、情報推進部情報企画課職員、契約所掌課職員、本センターが設置した仕様策定委員会、技術審査職員委員会の教職員ら多くの関係者が一致団結して取り組んでいる。とりわけ、著者は、SQUID の構想立案、設計において中心的な役割を担ってきた。

手元のメモから過去を振り返ってみると、われわれは 2017 年後半頃より、将来導入すべきスーパーコンピュータ導入に向けた技術調査・検討を開始している。それ以降、入札候補ベンダだけでなく、スーパーコンピュータを構成するプロセッサ、アクセラレータ、ネットワーク、ソフトウェア等の構成技術を供給しうる候補ベンダとの打ち合わせ・情報交換を継続的に何度も行ってきた。同時に、導入するスーパーコンピュータがどうあるべきか？といったこともまた仕様策定員委員会を中心に議論・検討をつづけてきた。その結果、ほぼ 4 年の月日を経て、国内屈指の性能を有するスーパーコンピュータ SQUID を導入することに成功したのである。

以前の HPC ジャーナルで記載した OCTOPUS の特集記事 [2] にも記載したが、その時と同じくあえて同じ言い回しを使いたい。自画自賛のようでみっともないが、SQUID はいま CMC が導入できる最高のスーパーコンピュータシステムである。限られた予算内で当初想定していた性能目標を達成するとともに、次世代のスーパーコンピュータにつなげることができる新しいチャレンジも盛り込むことができた。事実、後述するが本調達の一つのチャレンジとして導入したクラウドバースティング機能は本調達以降に行われている他大学計算基盤センターの調達にも盛り込まれ、このことは本センターのスーパーコンピュータの設計が他大学のスーパーコンピュータを先駆していることを示しているに他ならない。また、もう一つの新しいチャレンジでもあるデータ集約基盤 ONION については、すでに学内外の研究者の高い関心と興味が寄せられ始めている。

くどいようであるが、再度記載する。スーパーコンピュータの調達は、多大なる労力と時間をかけて行う。担当する教員は強い使命感と責任感で調達に尽力している。一方、こうした調達は、“ただの業務”にすぎないという教員もいるし、一生懸命やっても評価されないという教員もいる。事実、大学教員の公募などでこうした調達経験などが考慮・評価されることはほとんどない。著者自身も評価されている実感はほとんどなく、調達にまわす時間や労力を研究に回して論文を一つでも書いた方が自身の業績やキャリアパスに有益であろうと心底思う。しかし、著者の所属するサイバーメディアセンターはわが国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スーパーコンピュータを提供していくことを目的としており、著者はその目的遂行のための中心的なポストにある。そうした使命感から、自身の心底にある思いと戦いながら、大阪大学ひいてはわが国の学術・産業の発展を支える研究基盤を構想し、2017 年に OCTOPUS、2021 年に SQUID を実現してきた。再度記載するが、これらのスーパーコンピュータは、限られた予算の中で大阪大学ひいてはわが国の学術・産業の発展を加速させることができる最高のスーパーコンピュータであり、そうあってほしいと願っている。

本稿では、CMC のスーパーコンピュータ事業の要であるスーパーコンピュータ調達の記録、また、当該調達の後任者への引き継ぎ・情報共有という視点も含めて、SQUID 導入の背景、SQUID 調達概要を記したのち、SQUID の構成・特徴について記したい。

徒然なるまま記すので、読みにくい部分があるかもしれないが、ご容赦いただければ幸いである。なお、SQUID の構成・特徴についてのみ関心がある方は、本稿の前半部を読み飛ばしていただければ幸いである。

2. SQUID 導入の背景・経緯

2.1 CMC のスーパーコンピューティングシステムを取り巻く背景

本センターでは、2017 年 12 月に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) [3] の運用を開始した。OCTOPUS 導入当時、OCTOPUS の前身であるスカラ型スーパーコンピュータ HCC （汎用コンピュータクラスタ: Hanyou Computer Cluster）[4] は、スーパーコンピュータとしての基本的な要件を満たすことができない状況であり、スカラ型スーパーコンピュータを期待する利用者の皆様方を失望させる結果となっていた。その結果、本センターのスカラ型スーパーコンピュータの利用者はほとんどいない状況になってしまっていた。それゆえ、総計 300 ノードを超える OCTOPUS 導入に際しては、このスーパーコンピュータは利用者に使ってもらえるだろうか？という不安もあった。しかし、OCTOPUS 導入当初こそ比較的余裕があったものの、導入後初の年度末となる 2018 年 1-3 月頃には、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のすべてのノード群で 80 % 以上の利用率を記録するような状態が継続する状況となっていた。その翌年度においては、年度始めの 4 - 5 月頃には利用率が一時的に低下する傾向がみられたが、それ以降は定常的にどのノード群も利用率が高い状態が継続し、年間を通じて常にその状態を維持していた。当初、われわれはこのような高負荷状態を OCTOPUS 調達の成功の証であると喜んでいたが、ジョブの待ち時間が定常的に長くなってしまい資源不足の状況に陥ってしまっていた。この傾向は年々激化傾向にあり、2019 年度、2020 年度においては、遅くとも 8 月には、年間で提供できる計算資源量が売り切れてしまうという状況が発生してしまう深刻な資源不足の状態であった。つまり、利用者の計算需要に対して供給量が不足してしまい、利用者である研究者の研究活動に支障を与えてしまう状況となっていた。

一方、今回導入した SQUID の前身である SX-ACE は、2014 年度に導入したベクトル型スーパーコンピュータであった。この SX-ACE は、OCTOPUS 導入時、多少老朽化が進みつつある状況ではあったが、気象・気候、流体シミュレーション等の高い演算性能とメモリバンド幅のバランスを要求するプログラムをお使いの利用者に対しては好評なシステムとなっていた。加えて、SX-ACE は比較的資源量に余裕があり、ジョブ投入からの待ち時間も現実的な時間内に収まっていた。それゆえ、SX-ACE をお使いの利用者の満足度は非常に高いものであった。しかし、その一方、SX-ACE に対して否定的な意見をもつ利用者もいた。例えば、研究開発コミュニティで開発・公開されるオープンソースソフトウェアやライブラリを利用して研究開発を行う利用者は、それらのソフトウェアやライブラリを SX-ACE に移植すらできないという状況が多々発生していた。また、たとえ移植できたとしても、インテル製プロセッサよりも実行時間が遅くなってしまうという状況も発生していた。これは SX-ACE が NEC 製の独自のオペレーティングシステムを採用していたこと、ダイナミックリンクをサポートしていなかったこと、NEC 製の独自ベクトルプロセッサであったことなどが起因している。こうしたことから、SX-ACE を利用できない利用者からは、次期スーパーコンピュータには、汎用的なプロセッサやアクセラレータを搭載した大規模なスーパーコンピュータを希望する声が高まる状況にあった。とりわけ、OCTOPUS が導入され高い利用率を継続するようになると、ますます OCTOPUS と同様な汎用型のスーパーコンピュータを SX-ACE の後継機として切望する利用者の声が高まる傾向にあった。

2.2 学術的な背景

SQUID に向けた検討・議論を開始した 2017 年後半期においては、数値解析・シミュレーションといったキーワードに代表される高性能計算(High Performance Computing)分野だけでなく、AI（人工知能：Artificial Intelligence）、ML（機械学習: Machine Learning）、DL （深層学習: Deep Learning）といったキーワードに代表される高性能データ分析(High Performance Data Analysis)分野の研究が急速に活発化する傾向にあった。その後、先進的な研究開発では、高性能計算と高性能データ分析を融合しようとする試みも報告されつつあった。例えば、高性能計算実行中に中間結果に対して AI あるいは DL といったデータ分析技術を適用することにより、それ以降の計算を取りやめるかどうかを判断するといった試み、あるいは、大規模なデータを取り扱う AI エンジンの開発に、複数の GPU ノードを併用する高性能計算技術を応用するといった試みはその一例としてあげることができる。

また、2.1 節で記載したが、利用者である研究者の計算ニーズはますます多様化傾向にあった。事実、次期スーパーコンピュータに対するアンケートを通じて、次期スーパーコンピュータが搭載すべきプロセッサやアクセラレータについて調査を行ったところ、ベクトル型プロセッサ、NVIDIA 製 GPU、インテル製プロセッサ、AMD 製プロセッサ、ポスト京プロセッサ、FPGA 等の回答がえられている。このような利用者の多様化する計算ニーズを次期スーパーコンピュータシステム SQUID でどのように収容するかが課題となっていた。

また、SQUID 調達では、急速に広域化・グローバル化する学術研究に対応することが課題となりつつあった。今日の学術研究では、同一の大学や研究機関というだけでなく、国内外の大学や研究機関で共同研究体制を形成して研究を推進することが日常的になりつつある。また、そうした研究は、大阪大学の OU ビジョン [5] に示されるように、産業と学術の共同・共創活動として行われることも期待され、さらに、そうした活動は新たな社会価値創造にむけて加速・発展していくことが求められる。そうしたことから、SQUID 調達においては、そのような広域化・グローバル化する学術研究を支える計算基盤としてどのようなスーパーコンピュータであるべきかを考えることが課題となっていた。

また、一部の利用者からは、スーパーコンピュータシステムへのデータ移動の容易化を希望する声が寄せられていた。OCTOPUS 以前のスーパーコンピュータでは、予算的な制約もあり、スーパーコンピュータシステムへのデータ移動、あるいは、スーパーコンピュータシステムからのデータ移動は SCP や FTP といった CLI（コマンドラインインタフェース：Command Line Interface）による方法しか提供していなかった。それゆえ、上述した広域化・グローバル化する学術研究を鑑みつつ、SQUID 調達においても、利用者のデータ移動ニーズを充足できるデータ基盤をどのように提供するのかもまた重要な課題となっていた。

2.1 および 2.2 節に記した背景のもと、スーパーコンピュータシステム SX-ACE の契約更改をターゲットとして、新スーパーコンピュータ SQUID の導入に向けた検討・議論を 2017 年後半期より本格的に開始した。

3.「高性能計算・データ分析基盤システム」に向けた調達

3.1 調達概要

本センターでは約 4 年近くの歳月を費やし、SQUID の導入に向けた検討・議論を経て準備を重ねてきた。その調達活動の経緯について、以下に記す。

SQUID の前進となる SX-ACE は 2019 年 11 月に契約満了（実際は、スーパーコンピュータが利用できない空白時間が長くなってしまうと困るという利用者の声を最大限に配慮して、規模を縮小しながらの契約延長をしており 2021 年 2 月末日までサービスを提供した。）となることから、次期スーパーコンピュータに向けた調達準備について 2017 年度後半期より調査・検討を開始した。次期スーパーコンピュータに向けては、2 節に記載した高性能データ分析に対する期待と関心を背景に、2016 年度に本センターが主催した Cyber HPC Symposium [6] で取りあ

げたテーマでもある、高性能計算と高性能データ分析の融合を目指すという視点から、調達名称を「高性能計算・データ分析基盤システム」と設定した。仕様策定委員会設置後も、SX-ACE の契約満了予定の 2019 年 11 月頃利用可能なプロセッサ、アクセラレータ、メモリ、相互結合網等のスーパーコンピュータ構成要素技術についての調査を継続し、本センターで構想するスーパーコンピュータについてまとめた導入説明書を作成した。この導入説明書を基に、ベンダ企業からの意見を招請することを目的とした導入説明会を 2018 年 5 月 8 日に開催した。

その後、提案ベンダから提出された資料を精査するとともに、日々更新し続けるスーパーコンピューティングシステム技術動向を勘案し、「高性能計算・データ分析基盤システム」のターゲットを 2019 年 11 月から 13 ヶ月延長して 2020 年 12 月に再設定することを決定した。その後、仕切り直しとなる 2 回目の導入説明会を 2019 年 6 月 4 日に本センター吹田本館 2 階大会議室で開催し、再度ベンダ企業からの意見を招請した。

一方、本センターのスーパーコンピュータの利用者および利用候補者がどのような次期スーパーコンピュータを希望するかを把握するため、2019 年 8 月 30 日 ～ 9 月 30 日の 1 ヶ月間、利用者および本学の教職員に対して次期スーパーコンピュータ希望アンケートを実施している。アンケートでは、2020 年 12 月頃に利用可能な最新プロセッサ、アーキテクチャ、ソフトウェアだけでなく、本センターに対するユーザ支援についても幅広く意見を集約した。

その後、各社より提出された意見および上記次期スーパーコンピュータ希望アンケート結果を参考にしながら、再度 2020 年 12 月頃に利用可能なプロセッサ、アクセラレータ、メモリ、相互結合網等のスーパーコンピュータ構成要素技術についての調査を通じて、本センターにおいて仕様書案の作成を進めていた。しかし、仕様書案作成を鋭意進めている間の 2019 年 11 月に開催された、国際会議・展示会 SC で発表された最先端技術の動向を勘案し、次期スーパーコンピュータシステムの導入時期を 2020 年 12 月から 3 ヶ月延長し 2021 年 2 月と再設定した。スーパーコンピュータを構成するプロセッサや最新 GPU の選択の幅をひろげ、本センターにとって最良の高性能計算・データ分析基盤システムの導入のために、本センター設置の仕様策定委員会で判断した。

導入時期を 2021 年 2 月に設定した仕様書案を作成し、2019 年 12 月 3 日に仕様書案説明会を開催し、各ベンダ企業からの意見招請を行った。本センターでは、この仕様書案説明会で配布した仕様書案に対して提出された各ベンダ企業からの意見、および、上述した 2019 年度に実施した次期スーパーコンピュータ希望アンケートの結果を基に、最終仕様書を作成すべく準備を進めた。その結果、2020 年 3 月 18 日の仕様策定員会において、次期スーパーコンピュータシステムの導入時期を 2021 年 5 月と再設定することとなった。最終仕様書確定にむけ、仕様書案の修正・加筆を繰り返しながら、完成度を高めるべく尽力した。

2020 年度初めに最終仕様書が完成すると、7 月 6 日に入札官報公告をおこない、その 1 週間後の 7 月 13 日に最終仕様書を説明する入札説明会を実施した。入札説明会は、新型コロナウイルス感染症拡大への懸念、および、社内事情により出張のできない企業の方々への配慮から、本センター吹田本館 2F 大会議室およびオンラインのハイブリッドな環境で行った。

本調達では、2020 年 8 月 26 日に入札期限が設定されていたが、当日までに 2 社の応札があった。その直後より技術審査委員会による審査が行われ、10 月 14 日に開札となった。なお、当初予定では、10 月 7 日に開札予定であり、1 週間開札を延長した。これは、入札参加企業より仕様書に対する技術的な疑義がよせられたため、本センター側で当該疑義が何ら問題がないことを実証する期間を設けたことによる。その後、11 月 24 日には、落札官報公示が行われ、日本電気株式会社に落札されたことが報告された。

3.2 SQUID 調達方針

1 節に記載した通り、スーパーコンピュータの調達は、導入時期頃に利用可能な技術・ソリューション、またそれらの技術動向に関する綿密な調査を行

いつつ、本センターで導入すべきスーパーコンピュータを調達予算内で構想し、仕様書に落としこまなければならない。スーパーコンピュータの調達ではこの仕様書の完成度が、導入されるスーパーコンピュータの成否を決定づけることになる。

今回の高性能計算・データ分析基盤システムの調達においては、われわれは下記 5 つのチャレンジを設定して調達を行った。

(1) HPC・HPDA 融合
(2) クラウド連動・連携機能
(3) セキュアコンピューティング環境
(4) データ集約環境
(5) テイラーメイド型計算

以下、これらについて説明する。

(1) HPC・HPDA 融合

2 節で記載したように、数値シミュレーション、数値解析といったキーワードに示される高性能計算分野に加え、AI、ML、DL 等のキーワードに示される高性能データ分析分野からの計算ニーズが急速に高まっている。そういった背景から、本センターでは、SX-ACE の後継機には、HPC 分野だけでなく HPDA 分野からの多様な計算ニーズを収容することが求められた。さらに、今日では、HPC for AI、AI for HPC といった言葉に示されるように、HPC と HPDA の両分野は密接に関連しつつある。さらに、多種多様な IoT (Internet of Things) センサを利活用したビッグデータ解析といったことも盛んに行われている。

一般的に、HPC 分野の研究者がスーパーコンピュータに求める要求と HPDA 分野の研究者がスーパーコンピュータに求める要求は大きく異なる。例えば、HPC 分野の研究者が要求するソフトウェアとしては MPI (Message Passing Interface)、OpenMP などを利用する一方、HPDA 分野の研究者が要求するソフトウェアとしては Jupyter notebook、TensorFlow、PyTorch などがあげられ、大きく異なる。また、利用するプログラミング言語も HPC 分野の研究者の多くは、C、Fortran などを利用するのに対し、HPDA 分野の研究者は Python、Java などを主に利用する。さらにいえば、HPDA 分野の研究者は HPC 分野の研究者が当然のように利用できるスケジューラによるバッチ処理に対する強い抵抗感があり、インタラクティブ処理に対する強い要求がある。

このような考察や検討を経て、本調達では、HPC 分野および HPDA 分野の研究者がともにそれぞれの分野の計算を行うことができ、また HPC と HPDA を融合した新たな計算手法にチャレンジできる HPC・HPDA 融合計算基盤としてのスーパーコンピュータの実現を目指した。

(2) クラウド連動・連携機能

近年では、IaaS (Infrastructure As A Service)型クラウドが手軽に利用可能となってきた。例えば、マイクロソフト社の提供する Azure [7]、オラクル社の提供する Oracle Cloud Infrastructure(OCI) [8]、Amazon 社の提供する AWS [9]などはその一例である。これらの IaaS 型クラウドの特徴は、利用者が自分自身の計算環境を即座に構築・利用できる点にある。例えば、利用者はブラウザ上から利用したい Linux あるいは Windows の仮想計算機イメージを選択し、希望する規模で展開し、利用者自身の計算環境を構築することができる。

本センターでは、この IaaS 型クラウドの特徴を利用して、既設 OCTOPUS の高負荷状態を回避し、利用者の待ち時間の縮小を提言することを目的としたクラウドバースティング機能の研究開発を続けてきた [10, 11, 12]。その結果、このクラウドバースティング機能を利用することで、OCTOPUS の高負荷時に、OCTOPUS に投入された計算要求（ジョブ要求）を Azure および OCI にオフロードすることに成功している。

今回の調達では、本格的なスーパーコンピュータの運用に耐えうる、当該クラウドバースティング機能を中核としたクラウド連動・連携機能を次期スーパーコンピュータの一機能として実戦配備することを目指した。これにより、例えば、高負荷時のクラウドへのオフロードだけでなく、次期スーパーコンピュータ契約期間中に最新のプロセッサやアクセラレータが利用可能となった時、IaaS 型クラウドで提供されるそれら最先端技術を早期に実戦配備し、利

用者に提供することも可能となる。

(3) セキュアコンピューティング環境

2 節で記載したように、近年では AI、ML、DL に代表される HPDA 分野の研究開発が活発化傾向にある。その結果、様々な学術研究分野において、AI、ML、DL の応用を試みる動きがある。こうした動きは、医学、歯学、薬学といった高い秘匿性を有する研究分野においても例外ではなく、当該分野が有する大容量データに対して高性能データ分析を適用することへの高い期待がある。しかし、高秘匿性・高機密性が求められるデータを扱う学術研究分野においては、多くの利用者が共有して利用するスーパーコンピュータを利用することはデータセキュリティ上の問題があり困難となっている現状がある。このことは、わが国の学術、産業の発展・加速を妨げる一要因となりえ、早急に解決すべき課題となっている。

上述の背景と考察から、本調達では、高秘匿性・機密性を扱う分野の研究者が、安心してデータ処理を行うことができるセキュアコンピューティング環境の試験導入を目指すこととした。

(4) データ集約環境

本センターのスーパーコンピュータ環境では、利用者のローカル環境とスーパーコンピュータ間のデータ移動のために SCP および FTP を提供している。利用者のローカル環境からスーパーコンピュータ環境のログインノードあるいはフロントエンドノードに上述のプロトコルを用いて接続し、データを双方向に移動できる。SCP や FTP にはオープンソースとして利用できるクライアントソフトウェアも多くあり、これらのプロトコルに詳しくない利用者でも、クライアントソフトウェアの提供する GUI (Graphical User Interface)経由で直感的にデータ操作が可能である。

しかし、2017 年度に OCTOPUS を導入以降、新たにスーパーコンピュータの利用を開始した利用者も増加していること、また、近年の AI や ML といった高性能データ分析を活用とする研究活動の活発化もあり、より簡易的にデータをスーパーコンピュータに集約したいという要望もみられるようになった。また、今日では、新型コロナウイルス感染症拡大の影響はあるが、国内外の研究者と連携した共同研究の枠組みの中でスーパーコンピュータを利用しているケースも多く見られるようになってきた。学術機関の研究者だけでなく、企業の研究者らとも共同研究もまた活発化傾向にある。このような国際共同研究や産学共同研究では、共同研究に参画する研究者の所属する研究機関、大学、さらには、民間クラウドベンダのクラウド環境間でプログラムや入力データおよび計算結果などのデータを移動・共有することが日常茶飯事となりつつある。

本調達では、このような背景から、スーパーコンピュータをグローバル化・広域化する学術研究で要求される研究環境を構成する一計算基盤に位置付け、本センターのスーパーコンピュータに対して簡易にデータを集約することのできる機能性を実現することを目指した。さらに、利用者、クラウド、スーパーコンピュータ間でこれら研究データを自由に共有・移動させることができる仕組みの実現を目指した。

(5) テイラーメイド型計算

本センターのスーパーコンピュータ上では多種多様なプログラムが稼働している。本センターのスーパーコンピュータ上で一から開発されたプログラムもあれば、利用者環境で開発されたプログラム、オープンソースとして利用可能なプログラムなど多種多様なプログラムがスーパーコンピュータ上で動作することが期待されている。事実、2017 年度に OCTOPUS を導入以降、本センターにはオープンソースとして利用可能な各種プログラムの導入にむけた相談が数多く寄せられている。本センターでは、それらに個別にベストエフォートによる対応を行ってきてはいる [13] が、少ない人数でのスーパーコンピュータの運用を行っているため、今後さらに相談件数が増えると対応できなくなることも予測される。

オープンソースとして公開されているプログラム

や、利用者環境で開発されるプログラムの場合、コンパイラ、ライブラリ、ソフトウェア、アプリケーションのバージョンの違いによってスーパーコンピュータへの導入が難しい場合がある。一般権限で動作するソフトウェアやアプリケーションの場合、利用者のホームディレクトリ環境で行うなどの方法で回避できる場合はあるが、コンパイラなどのシステムソフトウェアやオペレーティングシステムなどの相違はインストールが困難あるいは不可能な場合もある。

一方、今日では、仮想計算機(Virtual Machine)技術や、コンテナ(Container)技術などが一般的に利用可能的となってきている。VMWare や KVM (Kernel Virtual Machine)などはすでにプロダクションレベルでのサーバ運用でも利用されているし、Docker はデファクトスタンダードなコンテナ技術として様々なクラウドサービスでも利用されている。HPC の分野でも、仮想計算機技術やコンテナ技術は近年でも利用されるケースが増えており、これらの技術を前提とした機能を具備するプロセッサも存在し、性能面でのオーバヘッドも小さくなってきている。しかし、これらの技術は、一般的に管理者権限を必要とするものが多く、これらをスーパーコンピュータ上で利用可能にした際、システムセキュリティ上の問題がある。

しかし、本調達では、利用者の計算ニーズ・要求の多様化は今後ますます進展していくと考え、利用者が利用者自らの計算環境をテイラーメイドできる環境が必要不可欠と考えた。それゆえ、本調達では、テイラーメイド型計算環境の利用者への提供は必須と位置付けている。

4. SQUID の概要

4.1 概要

3 節で記した調達の結果、本センターでは、総理論演算性能 16.591 PFlops の高性能計算・データ分析基盤システム SQUID の調達に成功した。SQUID（表 1）は、汎用 CPU ノード、GPU ノード、およびベクトルノード、20 PB のハードディスクと 1 PB の SSD から構成される大容量ストレージが 200 Gbps の相互結合網で接続されるハイブリッド型スーパーコンピュータである。後述するが、汎用 CPU ノードおよび GPU ノードに搭載されている。

CPU、GPU ノードに搭載されている GPU、ベクトルノードに搭載されているベクトルプロセッサはいずれも最新の製品が搭載されている。とりわけ、汎用 CPU ノードに搭載される CPU については、SQUID の 2021 年 5 月の稼働開始直前の 4 月 6 日（米国時間）に米国 Intel 社によって第 3 世代 Xeon Scalable Processor として発表[14]されたばかりの製品となっており、まさに最新製品である。

表 1: SQUID の構成概要

ノード構成	汎用 CPU ノード 1520 ノード(8.871 PFlops)	プロセッサ：Intel Xeon Platinum 83686 (Ice Lake / 2.4 GHz 38 コア) 2 基 主記憶容量：256 GB
	GPU ノード 42 ノード(6.797 PFlops)	プロセッサ：Intel Xeon Platinum 8368 (Ice Lake/ 2.4 GHz 38 コア) 2 基 GPU：NVIDIA HGX A100 8GPU ボード 主記憶容量：512 GB
	ベクトルノード 36 ノード(0.922 PFlops)	プロセッサ：AMD EPYC 7402P (ROME / 2.8 GHz 24 コア) 1 基 主記憶容量：128 GB ベクトルプロセッサ：NEX SX-Aurora TSUBASA Type20A 8 基
相互結合網	Mellanox InfiniBand HDR (200Gbps)	
ストレージ	DDN EXAScaler (Lustre/ HDD 20.0 PB, NVMe SSD 1.2 PB)	

また、GPU についても、本記事執筆時点においては、国内の学術計算基盤センターに大規模に導入された実績はいまだ少ない状況であり、最先端のテクノロジーとなっている。ベクトルプロセッサについてはNEC SX-Aurora TSUBASA Type20A[15]という最新モデルが導入されている。このように稼働時点で最新・最先端の技術を結集したスーパーコンピュータを導入できたことは、まさに仕様策定に長い時間と多大なる労力を費やした結果である。

さらに、今回導入する高性能計算・データ分析基盤システム SQUID には、3 節で記したように、これまでのスーパーコンピュータ調達にはなかった 5 つのチャレンジを盛り込んだ。これらは後述するが、SQUID ではすべてのチャレンジが実現されている。本記事執筆時点においては、新しいチャレンジを実現できる機能についてはまだまだ利用者に周知が行き届いていない部分もあるが、本センターでは今後の運用を通じて利用者の皆様方に積極的にご利用いただきたいと考えている。

さて、SQUID の性能について冒頭に少し触れたが、その性能値がどのような意味をもつのかを気にされる読者の方も多いと思うので、ここで再度性能面についてふれる。結論から言うと、この 16.591 PFlops という性能は、国内でも屈指のものである。もちろん国内の最大のスーパーコンピュータは理化学研究所計算科学研究センターが運用する富岳(理論演算性能: 537 PFlops）や、産業技術総合研究所の ABCI（理論演算性能: 54 PFLOPS)、東京大学と筑波大学が共同で運用する最先端共同 HPC 基盤施設の Oakforest-PACS（理論演算性能: 25 PFlops）とは比べ物にならないが、理論演算性能 15 PFLOPS 超のスーパーコンピュータを有する学術研究機関・大学は宇宙航空研究開発機構、名古屋大学、海洋研究開発機構等数えるほどしかないのが現状である。このことからも本センターの SQUID のポテンシャルをご理解いただけるであろうか。

さらに、2021 年 6 月に開催された国際会議・展示会 ISC2021（オンライン開催）で発表された、世界のスーパーコンピュータの性能をランキングする TOP500[16]においては、性能評価指標 LINPACK[17]による性能評価結果をエントリーした結果、SQUID の汎用 CPU ノード群だけで世界 67 位という記録を残している[18]。また、同様に HPCG [19]という性能指標によるランキングでも 54 位[20]、また電力効率のよい高性能計算の実現を評価する Green500 においても世界 57 位を記録する結果となった[21]。SQUID の全部ではなく一部の汎用 CPU ノード群のみを用いて世界のスーパーコンピュータの性能を評価するランキングで上位の成績を収めていることからも SQUID が高い性能を有することが示されている。

図 1: SQUID の外観 1

図 2: SQUID の外観 2

SQUID の外観を図 1、図 2 に示す。図 1 はラックに描画されたスーパーコンピュータ SQUID の“顔”となるラックペイントである。このラックに描画されたラックペイントは、2021 年 5 月に稼働を開始する前に、「スーパーコンピュータシステム SQUID ラックデザインコンテスト」[22]と称して、本センターで実施したコンテストでの最優秀賞の作品である。このコンテストは読者の皆様方は記憶にあるかもし

れないが、スーパーコンピュータシステムを一般の方に親しみを持ってもらうこと、スーパーコンピュータシステムの研究成果の一般の方へのアウトリーチや本センターのプレゼンス向上という視点から、OCTOPUS を導入した際に初めて実施した企画であり、導入するスーパーコンピュータシステムの顔となるべきラックデザインを一般公募するコンテストを今回の SQUID においても実施した。本コンテストでは、大阪大学サイバーメディアセンターが主催し、SQUID の導入業者となる日本電気株式会社が共催、SQUID を構成するプロセッサ、アクセラレータ、ストレージ、クラウド資源を供給する、インテル株式会社、エヌビディア合同会社、クラウディアン株式会社、株式会社データダイレクト・ネットワークス・ジャパン、日本オラクル株式会社、日本マイクロソフト株式会社が協賛し、事務局を担当する株式会社サイバースケッチが協力する体制にて、2021 年 2 月 15 日を締め切りとして実施した。図 1 に示したラックペイントデザインは全 59 作品のなかから選定されたデザインである。図 2 の左面には、本コンテストに本センターのロゴとともに、本コンテストに共催、協賛してくれた各企業のロゴをプリントしている。本記事を掲載時点では、新型コロナ感染症拡大が猛威をふるっている状況であるため難しいが、この状況が改善すれば見学会などを通じて読者の皆様にもご覧いただける機会を提供できればと切に思う。

以下の節では、スーパーコンピュータ SQUID を構成する要素について解説する。

4.2 計算ノード

4.2.1 汎用 CPU ノード

SQUID の CPU ノードは、Intel Xeon Platinum 8368 プロセッサを 2 基、256 GiB の主記憶容量を搭載した計算ノードである。この Intel Xeon Platinum 8368 プロセッサは、Intel 社の Ice Lake 世代のプロセッサであり、2.4 GHz 38 コアを有しており、単体での理論演算性能は 2.918 TFLOPS となる。それゆえ、CPU ノード単体での理論演算性能は 5.836 TFLOPS となる。さらに、この SQUID には 1,520 基の CPU ノードが搭載されているため、CPU ノード群全体の理論演算性能は 8.871 PFLOPS となる。この 8.871 PFLOPS という数字は、SQUID の総理論演算性能 16.591 PFLOPS の約 53.4 %となる。上述したが、この CPU ノード群だけで TOP500 では世界で 67 位の性能を有するスーパーコンピュータになっている。

この Ice Lake 世代の Intel 製プロセッサの最大の特徴は、PCI Express 4.0 (Gen4)に対応したことにある。これに伴い、CPU サーバでは、メモリチャネルも 8 チャネルまでサポートされており、メモリ帯域幅も増強されている。SIMD(Single Instruction Multiple Data)命令に関しては、OCTOPUS に導入されている Sky Lake 世代のプロセッサと同様に、AVX-512 命令セット[23]がサポートされている。そのため、512 ビットに拡張された SIMD は OCTOPUS と同様に利用することが可能である。

図 3: NEC LX103Bj-8

図 4: 汎用 CPU ノードを収容する Blade Enclosure

この CPU ノードは、NEC LX103Bj-8 という型番のサーバ（図 3）として供給される。表 2 に NEC LX103Bj-8 の構成を示す。図 4 は NEC LX103Bj-8 を搭載する 8U の Blade Enclosure である。LX103Bj-8 には、上述した第 3 世代 Intel Xeon Scalable Processor

である Intel Xeon Platinum 8368 (2.40Ghz, 24 コア)が 2 基搭載されている（図 3 中央部）。その 2 基のプロセッサを取り囲むように、16 GiB の DDR4-3200 DIMM が 16 本が搭載されている。この LX103Bj-8 を図 4 に示す Blade Enclosure の 1 シャーシに 19 台を収容させる。この Blade Enclosure を 4 シャーシ組み上げ、1 ラックに搭載している（図 5）。SQUID の汎用 CPU ノード群は上述の通り 1,520 ノードであるため、合計 20 本のラックが本センターの IT コア棟に設置されている。なお、汎用 CPU ノードに搭載されたプロセッサは、OCTOPUS と同様に、液冷方式で冷却を行い、安定的に高性能が供給できるよう設計・実装を行なっている。図 5 に見えるラック前面の黒いホースは、冷却水をサーバ上のプロセッサに供給し、また、プロセッサの熱を回収した冷却水をサーバから回収するためのものである。なお、プロセッサの熱を回収した冷却水は、サーバ上部 Asetek 製 CDU (Cooling Distribution Unit)を通じて、サーバラックの設置されている IT コア棟のサーバ室の階下にある冷凍機に送られ、冷却されたのちにサーバラック CDU に戻ってくる仕組みになっている。

表 2: NEC LX103Bj-8 構成

プロセッサ	Intel Xeon Platinum 8368 (Ice Lake) x 2
メモリ構成	256 GiB (16GiB DDR4-3200 DIMM x 16)
ストレージ	240 GB SSD
NW　I/F	InfiniBand HDR x 1, 25 G / 10GbE x 2, BMC
OS	Cent OS 8.2

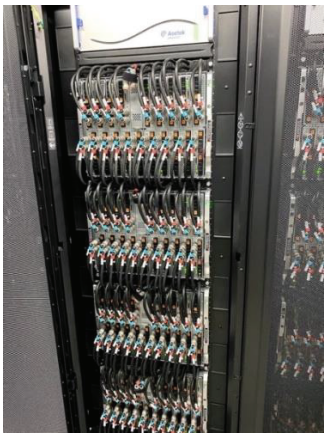

図 5: 汎用 CPU ノードラック

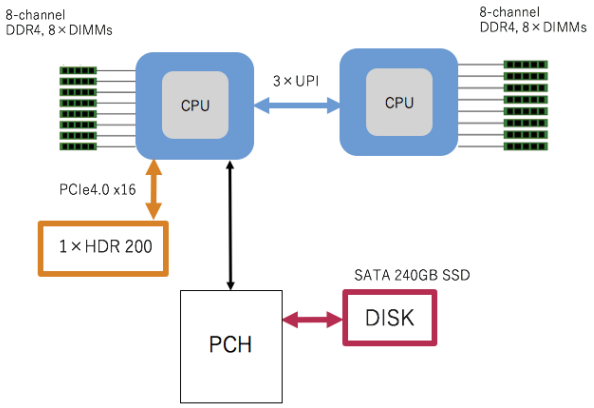

図 6: 汎用 CPU ノードブロック図

次に、内部構成について解説する。図 6 は LX103Bj-8 の内部構造を示したブロック図である。サーバ内では、2 基のプロセッサは UPI (Ultra Path Interconnect) 3 本によって接続されている（当該 CPU がサポートする最大の UPI 数である）。各リンクで 11.2 GT/s (Gigatransfer per second)×3 での信号速度で転送が可能となっている。また、上述したが、SQUID の汎用サーバに搭載されている Ice Lake 世代のプロセッサは PCI Express 4.0 (Gen4)に対応しており、メモリチャネルも PCE Express 3.0 (Gen3)時代の 6 チャネルから 8 チャネルに拡大している。LX103Bj-8 では、2 個ある CPU ソケットのそれぞれから 8 チャネルメモリコントローラを経て 8 個の 16 GiB DDR4-3200 DIMM に接続されている構造である。それゆえ、1CPU あたりのメモリとの転送速度は、3200 MHz×8 Byte×8 ch = 204.8 GB/s となる。OCTOPUS の汎用 CPU ノードの 127.968 GB/s と比較すると、SQUID の 1 CPU あたりのメモリ転送速度は 1.6 倍向上していることになる。また、汎用 CPU サーバには、NVIDIA 製 Mellanox InfiniBand HDR (200Gbps)が搭載されているが、PCI Express 4.0 (Gen4)×16 (256Gbps)で CPU と接続されているため、200 Gbps での高速通信がボトルネックなしに可能となっている。なお、図中にある SATA 240 GB の SSD はシステム領域として利用している。

4.2.2 GPU ノード

SQUID の GPU ノードは、Intel Xeon Platinum 8368 プロセッサを 2 基、512 GiB の主記憶、NVIDIA HGX A100 8 GPU ボード（開発コード: Delta）を搭載した計算ノードである。GPU ノードに搭載されているプロセッサは、CPU ノードに搭載されているプロセッサと同等であり、単体での理論演算性能は 2.918 TFLOPS となる。なお、この構成は、OCTOPUS で発生しているように汎用 CPU ノード群の混雑状況が激化し利用者の待ち時間が大きくなった際に一次的に GPU ノードを汎用 CPU ノードとして開放するといったこと、また、利用者のプログラムの CPU ノードから GPU ノードへの移植を容易にするといったことを想定して設計されている。

GPU ノードの中核である NVIDIA HGX A100 8 GPU ボード（開発コード名：Delta）は、名前の示す通り、GPU アクセラレータが 8 基搭載されたボードである。1 個の A100 あたりは、8 GB の HBM2 (High Bandwidth Memory2)を 5 個、すなわち、HBM2 を 40 GB、108 個の有効化されたストリーミングマルチプロセッサ(SM)が搭載されている。その結果、HBM2 へのメモリバンド幅は 1.555 TB/s、理論演算性能は、倍精度で 19.5 TFlops となっている。

この数字を基に OCTOPUS に搭載されている NVIDIA Tesla P100 と SQUID で導入された A100 を単体で比較すると、メモリ帯域幅では 732 GB/s から 1.555 TB/s へ約 2.1 倍、メモリ容量では 16 GB から 40 GB へ 2.5 倍、理論演算性能（倍精度）では 5.3 TFlops から 19.5 TFlops へ約 3.7 倍の性能向上がみられることがわかる。さらに、ノード単位でみれば、OCTOPUS に搭載された Tesla P100 4 基の理論演算性能が 21.2 TFlops であるのに対し、SQUID に搭載された A100 8 基の理論演算性能は、OCTOPUS の GPU ノード 1 基あたりの GPU 演算性能の約 7.35 倍の 156 TFlops となっている。さらに、整理すれば、OCTOPUS の GPU ノード単体に搭載された CPU と GPU の総理論演算性能は 23.196 TFlops であるが、SQUID の GPU ノード単体の CPU と GPU の総理論演算性能は約 6.98 倍の 161.836 TFlops に性能向上していることになる。

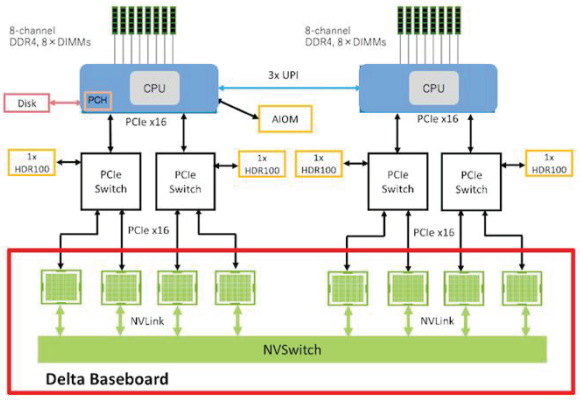

図 7: GPU ノード ブロック図

次に、GPU ノードの内部構成について解説する。図 7 に GPU ノードのブロック図を示す。CPU およびメモリの構成は、メモリサイズが違う点をのぞき、4.2 節で記載した汎用 CPU ノードと同一であるので、ここでの再度の説明は省略する。図中下部の黄緑色で囲まれた部分が GPU ノードの中核となる前述の Delta ボードである。この Delta ボードの最大の特徴は、8 基の A100 は NVIDIA 製の第 3 世代 NVLINK、および、これら NVLINK を収容し相互に接続する第 2 世代 NVSwitch から構成されている。NVLINK や NVLINK の仕様詳細については、NVIDIA 社のウェブサイト[24]やホワイトペーパーをご覧いただければと思うが、この第 3 世代の NVLINK では、仕様上、1 リンクあたり双方向で 50 GB/s のデータ転送が可能であり、第 2 世代の NVSwitch では最大 600 GB/s の通信が可能である。GPU ノードに搭載されている NVIDIA HGX A100 8 GPU ボード （開発コード名：Delta）の内部構成[25]を図 8 に示すが、各 A100 より 12 本の NVLINK が NVSwitch に接続されていることからも上記に示した性能値をお分りいただけると思う。

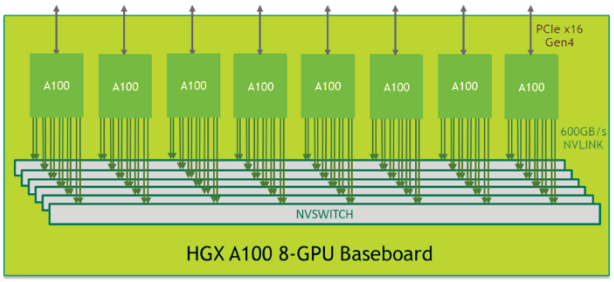

図 8: NVIDIA HGX A100 8GPU Baseboard [25]

また、GPU ノードでは、汎用 CPU ノードと異なる点として、相互結合網への接続が 200 Gbps から 400 Gbps となっている点があげられる。相互結合網への接続のために、GPU ノードからは 4 本の InfiniBand HDR 100 が用いられている。GPU との InfiniBand HDR 100 は、PCIe Gen4 x 16 で PCI スイッチ経由で実現されており、100 Gbps の帯域を確保している。GPU ノード内では、各 CPU ソケットあたり 2 本の HDR 100 が接続されており、相互接続網への接続は CPU ソケットに対して対照的な構成となっている。

図 9: NEC LX106Rj-4G

表 3: NEC LX106Rj-4G 構成

プロセッサ	Intel Xeon Platinum 8368 (IceLake) x 2
メモリ構成	512 GiB (32 GiB DDR4-3200 DIMM x 16)
ストレージ	240 GB SSD
NW　I/F	InfiniBand HDR100 x 4, 25 G / 10 GbE x 2, BMC
GPU	NVIDIA A100 (SXM4) x 8 (Delta)
OS	CentOS 8.2

図 9 に GPU ノードの外観を示す。この GPU ノードは、NEC LX106Rj-4G という型番のサーバ（図 9）として供給される。表 3 に LX106Rj-4G の構成を示す。この GPU ノードは 4U サイズであり、図 7 に示すブロック図の構成で 2 基のインテル製 Ice Lake プロセッサ、および、NVIDIA HGX A100 8 GPU ボード（開発コード: Delta）が搭載される構成である。SQUID では、この LX106Rj-4G の 6 ノードが 1 ラックに搭載され、合計 7 ラックの GPU ノードが IT コア棟に設置されている（図 10: 奥から 2 番目から 7 ラック）。もちろん、この GPU ノードについても、搭載されているプロセッサ、および、GPU はすべて汎用 CPU ノードと同様に液冷方式により冷却し、高い性能を安定的に供給できるよう設計・実装している　（図 11）。

図 10: GPU ノードラック

図 11: GPU ノードラック拡大図

4.2.3 ベクトルノード

SQUID のベクトルノードは、AMD 製 EPYC 7402P（2.8 GHz, 28 コア）を 1 基、128 GB の主記憶、NEC 製 SX-Aurora TSUBASA Type20A を 8 基搭載した計算ノードである。この AMD 製 EPYC 7402P はコードネーム ROME として開発された、PCI Express 4.0 (Gen4) 対応の第 2 世代 EPYC であり、理論演算性能は 2.150 TFlops となる。

図 12: NEC SX-Aurora TSUBASA Type20A

ベクトルノードの中核であるベクトルプロセッサ NEC SX-Aurora TSUBASA Type20A（図 12）は、世界最速クラスのコア（倍精度での演算性能 307 GFlops）を 10 コア搭載しており、3.07 TFLops の理論演算性能を有する。また、このプロセッサには、8 GB の HBM2 を 6 個、すなわち 48 GB の HBM2 が搭載されており、1.53 TB/s のメモリ帯域幅が利用可能となっている。ベクトルノードはこのベクトルプロセッサを 8 基搭載しており、ベクトルノード全体の理論演算性能は 26.71 TFlops となる。

SQUID の前身となるスーパーコンピュータ SX-ACE では、256 GFlops の理論演算性能を有するベクトルプロセッサを搭載していた。図 12 に示す SX-Aurora TSUBASA Type20A は、前身の SX-ACE の約 12 倍の性能を有しており、さらに、SQUID と SX-ACE の単体ノードを比較すると、276 GFLOPS から 26.71 TFlops への約 96.78 倍の性能向上が得られていることになる。

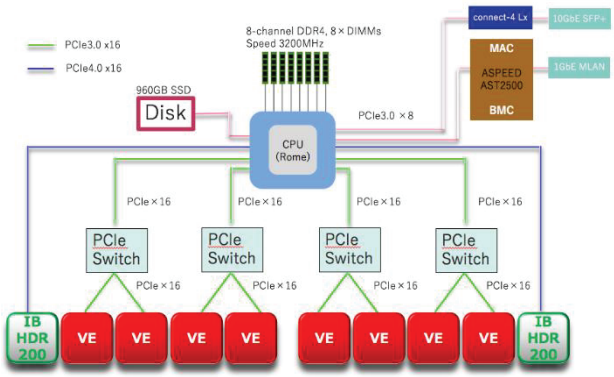

図 13: ベクトルノード ブロック図

次に、ベクトルノードの内部構成について解説する。図 13 にベクトルノード内のブロック図を示す。冒頭に記載したが、ベクトルノードに搭載されている AMD EPYC 7402P は PCI Express4.0 (Gen4)世代のプロセッサであり、8 チャネルのメモリコントローラ経由で 8 枚の 16 GiB DDR4-3200 DIMM が接続されている。このため、CPU とメモリ間の転送速度は、3200 MHz × 8 Byte × 8 ch = 204.8 GB/s となり、CPU ノードおよび GPU ノードと 1CPU あたりのメモリ転送速度は同等である。一方、CPU とベクトルプロセッサ(VE: Vector Engine)への接続は、PCIe Switch を経由する。 具体的には、2 基のベクトルプロセッサが 1 個の PCIe Switch にそれぞれ PCI Gen3 x 16 (128 Gbps)で接続され、PCI Gen3 x 16 (128 Gbps)で CPU に接続される。

また、ベクトルノード単体は、GPU ノード同様に 400 Gbps で相互結合網に接続される構成である。この構成のため、IB HDR 200 の HCA (Host Channel Adapter) を 2 基搭載し、CPU と各 HCA は PCI Gen4 x 16 (256 Gbps)で接続されている。

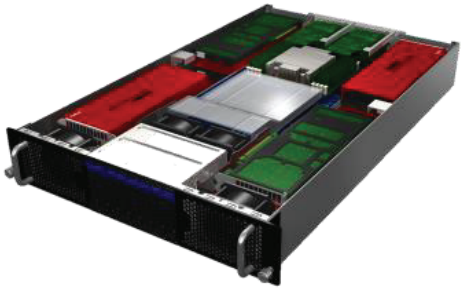

図 14: NEC SX-Aurora TSUBASA B401-8

表 4: NEC SX-Aurora TSUBASA B401-8 の構成

プロセッサ	AMD EPYC 7402 (2.8Ghz/24c) x 1
メモリ構成	128 GiB (16GiB DDR4-3200 DIMM x 8)
ストレージ	960 GB SSD
NW I/F	InfiniBand HDR x 2
ベクトル プロセッサ	VE Type 20A x 8
OS	Cent OS 8.2

図 14 にベクトルノードの外観を示す。このベクトルノードは、NEC SX-Aurora TSUBASA B401-8 の

型番が付与されたサーバとして供給される。表 4 に NEC SX-Aurora TSUBASA B401-8 の構成を示す。このベクトルノードは 2U サイズであり、図 13 に示すブロック図の構成で 1 基の AMD 製 Rome プロセッサ、および、8 基の NEC SX-Aurora TSUBASA が搭載される構成である。図中赤く見える箇所が NEC SX-Aurora TSUBASA である。SQUID では、この SX-Aurora TSUBASA B401-8 の 18 ノードが 1 ラックに搭載され、合計 2 ラックのベクトルノードが IT コア棟に設置されている(図 15: 手前 2 ラック; 図 16)。このベクトルノードについても、搭載されているプロセッサおよびベクトルプロセッサはすべて汎用 CPU ノードや GPU ノードと同様に液冷方式により冷却し、高い性能を安定的に供給できるよう設計・実装されている。

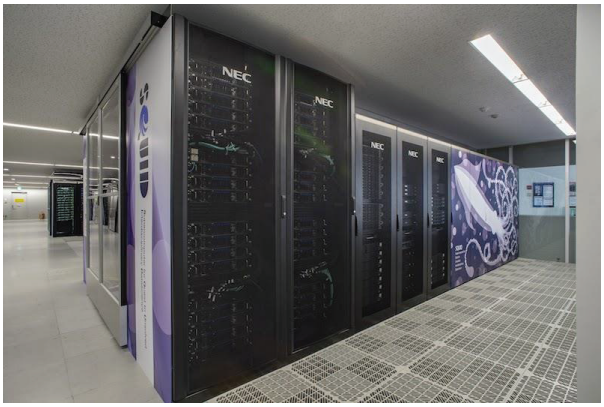

図 15: ベクトルノードラック（表）

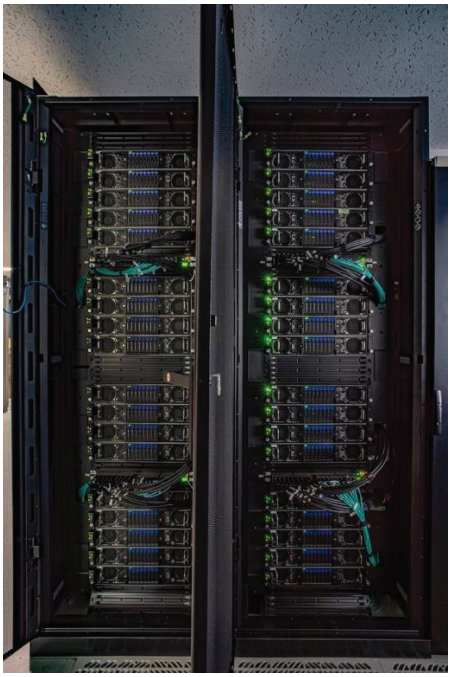

図 16: ベクトルノードラック（裏）

4.3 データ集約基盤 ONION

4.3.1 概要

3.2 節調達方針において、データ集約環境が一つのチャレンジであることについて記載した。これまでのスーパーコンピュータ調達のデータ基盤に対しては、いかに本センターのスーパーコンピュータ上で動作するプログラムが高速に処理されるか、という点を重視し、高性能計算のストレージへの入出力性能 (read、write)、ストレージ容量、および、耐障害性といったストレージそのものに対する性能評価指標に基づいてきた。本調達では、このストレージそのものに対する性能評価指標に加え、広域化・グローバル化する学術研究の中でデータを利活用できるかという視点を取り入れている。3.2 節で記したように、今日の学術研究においては、地理的に分散した研究者らがチームとなって国内、国外、産学の枠を超えて連携・協働することが日常的になりつつある。さらに、大阪大学では、OU ビジョン 2021[26]のもと、専門分野を超え、広く世界と協働する新たな知の創出を目指す「Open Research」、社会のニーズに基づく基盤研究の課題を発掘し、新たな社会的価値の創出につながる「Open Innovation」が示され、より一層国内、国外、産学の枠を超えて連携・協働することが求められている。

このような背景から、本センターでは、本学のスーパーコンピュータが具備すべきデータ基盤のあり方を検討し、本学のビジョンに示される Open Research および Open Innovation を加速・進展させることのできるデータ集約基盤を構想してきた。本調達では、そのデータ集約基盤を ONION (Osaka university Next-generation Infrastructure for Open research and Open innovatioN) と名付け、ONION の PoC (Proof of Concept) 実装として試験的に導入することとした。本調達で ONION を試験的な導入に留める理由としては、(1)本格的に大規模に ONION を整備することは、主目的である高性能計算・データ分析計算基盤で必要となる計算資源のサイズを大幅に縮減してしまい、ますます拡大傾向にある計算ニーズ・要求を満たすことができなくなる恐れがあること、(2)本センターで構想した ONION が有用では

なく、利用されない場合もありえることがあげられる。本センターでは、ONION を試験導入し、利用者に実際に利用していただきながら、改善、機能拡張・増強を経て成長させる方針を選択した。

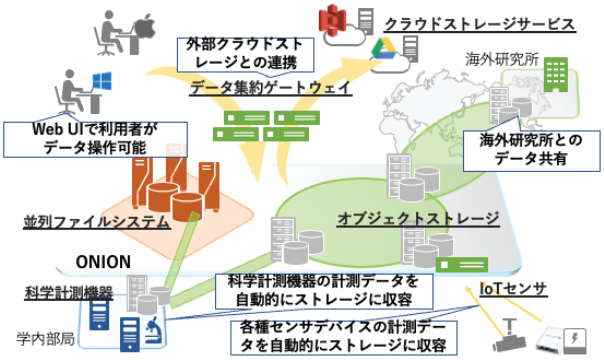

図 17: ONION の概要

図 17 に本センターで構想する ONION の概要を示す。広域化・グローバル化する学術研究、産学共創の支援には、国内外、産学の枠を超えて、計算に必要となるデータ、プログラム、計算結果など様々な研究データが広域研究環境上で共有されなければならない。また、本調達で導入する SQUID の計算資源もストレージ資源も広域研究環境での一拠点として構成されなければならない。そういった視点から、本調達では、ONION の PoC 実装を実現している。例えば、本センターでは、スーパーコンピュータでの計算結果を即座に外部のスーパーコンピュータを利用していない共同研究者に共有・公開したいという要望がある。また、学内に設置された科学計測機器で生成されるデータを速やかにスーパーコンピュータのストレージに移動し、データ解析をしたいという要望や、計算結果やプログラムを民間クラウドストレージに保存したいなどの要望がある。また、単にスーパーコンピュータのストレージをウェブインタフェースからアクセスしたいという要望もある。本調達で実現した ONION は、上述した要望をすべて実現するように設計・実装されている。

この ONION の実現に際して、SQUID は異なる 2 種のストレージを導入し、クラウドストレージでのデファクトスタンダードプロトコルである S3 (Amazon Simple Storage Service) プロトコルを中核として統合している（図 18）。今日では、S3 に対応した IoT センサも存在し、逆に、S3 に対応していれば民間クラウドストレージへのデータ保存も容易である。ONION では、この S3 を中核とすることで、学内の各種科学計測機器やセンサからのデータ集約、スーパーコンピュータからクラウドへのデータ移動を容易にしている。その上で、オンラインストレージソフトウェア NextCloud を用いて、2 種の異なるストレージの違いを隠蔽することで、利用者にシングルディスクイメージ上での直感的なデータ操作を行うことを可能にしている。

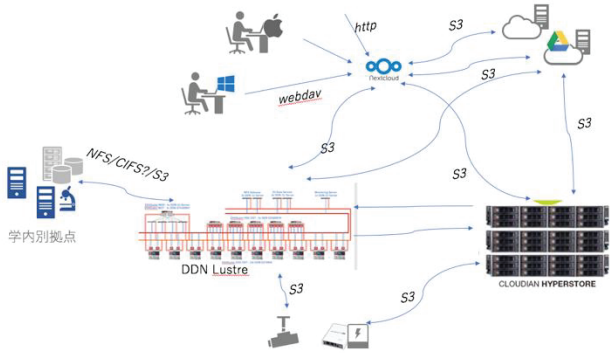

図 18: ONION でのプロトコル連携

以下では、ONION を構成する並列ファイルシステム DDN(DataDirect Networks)製 EXAScaler(Lustre)、オブジェクトストレージ Cloudian 製 HyperStore、および NextCloud について概説する。

4.3.2 並列ファイルシステム　EXAScaler

SQUID を構成する計算ノードは合計 1,598 ノードにも及ぶ。それゆえ、SQUID のストレージシステムには、SQUID 上で同時実行されるジョブの行うファイル操作の高速性が要求される。SQUID では、このようなファイル I/O の高速性を十分に満たすことのできる並列ファイルシステム DDN 製 EXAScaler (Lustre)[27]が採用されている。

図 19 に ONION を構成する並列ファイルシステム DDN EXAScaler の概要を示す。表 5 に ONION 構成並列ファイルシステム DDN EXAScaler の性能を示す。ONION 構成並列ファイルシステムは、約 80 億個の inode 数をサポートし、合計 21 PB を並列ファ

イルシステム Lustre ベースのファイルシステムである DDN 製 EXAScaler を基軸として提供する。この DDN 製 EXAScaler は OCTOPUS でも使用されている。SQUID では、この 21 PB の大容量ストレージをわずか 3 ラックに高密度に収容している（図 20）。

この大容量ストレージは、ExaScaler ファイルサービス（Lustre ファイルサービス）実現のためのサーバ群(Object Storage Server (OSS)、Object Storage Target (OST)、Meta Data Server(MDS)、Meta Data Target (MDT)、ストレージ管理サーバ（モニタリングサーバ）、NFS サーバ、 S3DS サーバから構成される。この構成は、それぞれのサーバ数は異なるが、構成要素としてはほとんど OCTOPUS と同様の構成となっている。詳細は OCTOPUS の特集記事[2]をご参照いただければ幸いである。

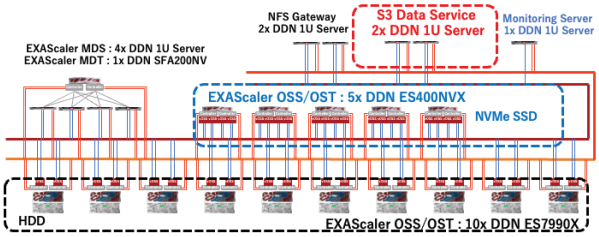

図 19: DDN EXAScaler の概要

表 5: ONION 構成並列ファイルシステムの性能

実効容量(HDD)	20.00 PB
実効容量(NVMe)	1.20 PB
最大合計 inode 数	約 80 億個
最大想定実効スループット(HDD)	160 GB/s 以上
最大想定実効スループット(NVMe SSD)	write:160 GB/s 以上 read:180 GB/s 以上

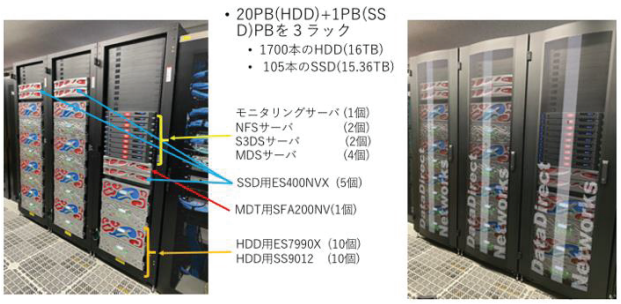

図 20: DDN EXAScaler 搭載ラック

しかし、OCTOPUS と SQUID では 2 点の相違がある。1 点目は、OCTOPUS ではハードディスクドライブ(HDD)のみの並列ファイルシステムの構成であったのに対し、SQUID では HDD および SSD のハイブリッド構成による並列ファイルシステムとなっている点である。2 点目は、4.3.1 で記載した S3 プロトコルを中核としたデータ集約基盤 ONION 実現のために、本調達では新しく S3DS サーバが導入されている点である。Lustre ファイルシステムを構成する OSS、OST などの各要素がどのような役割を担い、どのように動作するかについては OCTOPUS の特集記事の際に詳解したので、そちらを参照していただくとして、本特集記事では上述の 2 点の相違について説明する。

まず 1 点目の HDD および SSD のハイブリッド構成を採用した理由であるが、これはもちろんストレージへの read、write 性能の向上を目的としている。今日においては、SSD の価格も低下傾向にあり、SSD のみを用いた並列ファイルシステムを構成することも考えられる。しかし、本調達のスーパーコンピュータに見合う規模で並列ファイルシステムの導入を考える場合、未だ現実的ではなく、同一予算では導入できるストレージ容量は小さくせざるを得ない。本調達では、上記のストレージプロダクトの価格動向を考慮しつつも、大規模・大容量データを扱う利用者のストレージ I/O 性能に対する要望を満たせるよう、ハイブリッド構成とした。

さらに、SSD と HDD のハイブリッドでファイルシステムを構成する際、SQUID では、SSD、HDD それぞれに対してファイルシステムを構成し、それぞれの領域を異なるディレクトリにマッピングした。

これにより、HDD のデータを SSD に明示的に利用者に移動させるという手間を利用者にとらせるのではなく、利用者はディレクトリを使い分けるだけで SSD の高いストレージ I/O 性能を享受できるようにしている。読者の中で大規模・大容量データを扱っている方がおられれば、是非 SSD 領域の利用を検討いただけば幸いである。

次に、2 点目の相違点である S3DS サーバについて解説する。この S3DS（S3 データサービス）サーバは、Lustre ストレージ上で S3 API を提供する。これにより、外部の S3 対応クラウドストレージと Lustre ファイルシステム間でのデータのやり取りを可能にする。この S3DS サーバを用いることで、利用者は S3 に対応したクライアントから S3DS サーバ経由で並列ファイルシステム EXAScaler 上にデータをファイルとして格納できる。さらにいえば、この仕組みを利用することで、学内に設置された科学計測機器あるいはセンサ等からデータを自動的に集約する仕組みを構築可能となる。

4.3.3 オブジェクトストレージ HyperStore

図 21: Cloudian 製 HyperStore アプライアンス HSA-1610

ONION を構成するオブジェクトストレージとして、Cloudian 製 HyperStore アプライアンス HSA-1610（図 21）が 7 ノード導入されている。この HSA-1610 にはデータ格納用に 10 TB の SAS HDD が 12 本の合計 120 TB、メタデータ格納用に 960 GB の SSD が 2 本の約 2 TB が搭載されている。SQUID では、この HSA-1610 7 ノードを一つのストレージシステムとして構築している（図 22）。

図 22: ラックに搭載された ONION 構成オブジェクトストレージ

この Cloudian 製 HyperStore は、前述した S3DS 同様に、S3 API をサポートしている。そのため、外部のクラウドや、S3 対応ストレージとの相互のデータ交換を容易に行うことができる。さらに、この HyperStore は、高信頼性、高拡張性、高可用性を特徴としている。この HyperStore では、最小 3 ノード以上からストレージシステムを構成でき、必要に応じてサーバ 1 ノード単位で拡張を行うことができる柔軟な拡張性を備えている。本調達では、上述したが、ONION は試験的な導入という側面もあるため、最低限を導入し、今後利用状況をみながら拡張を行うことを想定している。また、この HyperStore は、データが自動的に保護される機能を有しており、高信頼性を提供している。その方式として、レプリカ（複製）方式（図 23）と、イレージャー・コーディング(EC)方式（図 24）が利用可能である[28]。前者は HyperStore にデータが格納される際に、自動的に書き込まれるデータのコピー（複製）をノード間に分散する方法であり、後者は格納されるデータに加えて、データを復元するための符号を付加して、それらを複数のサーバ間に分散する方法である。このようなデータ保護機能をもたせることで、オブジェクトストレージを構成する一部のサーバの障害時においてもデータを失うことのない仕組みになっている。本センターでは、後者のイレージャー・コーディング方式を採用している。なお、一部のノードが障害により利用できない場合においても、システム

は停止することがなく、可用性も高いものとなっている。

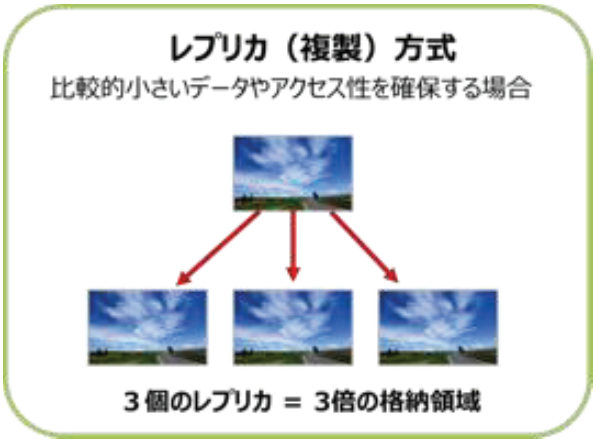

図 23: HyperStore のデータ保護機能（レプリカ方式）

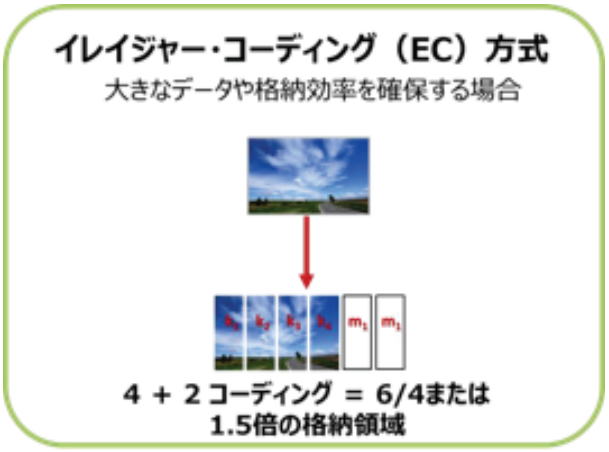

図 24: HyperStore のデータ保護機能（イレージャー・コーディング方式）

ところで、読者のなかには、4.3.2 での S3DS サーバと、本節で記載した HyperStore をどのように使い分けるのか？という疑問をお持ちの方がおられるかもしれない。この質問は、著者が講演会、シンポジウム等で話をする際にはよくきかれる質問である。確かに、S3DS も HyperStore も S3 API を提供しており、S3API 経由で外部ストレージとデータのやりとりが可能である。しかし、データの格納は、前者はファイルとして行われ、後者はオブジェクトとして行われる。したがって、技術的な詳細については、オブジェクストレージとファイルの相違に帰着していくところが大きいため、ここでの説明は控えるが、例えば、スーパーコンピュータで処理や解析するようなデータに関しては、スーパーコンピュータの計算処理で追記処理なども発生しファイルで格納・管理する S3DS が適している。一方で、データを保管、公開、共有したりする場合には、オブジェクトとして管理する HyperStore のほうが適している。このような視点で使い分けをご検討いただければ幸いである。

現段階では、この ONION 構成ストレージは IT コア棟に設置されている。そのため、IT コア棟のメンテナンス停電などの際には停止せざるを得ない。が、今後利用者が順調に増え、さらなる拡張が必要となった際に、例えば、豊中キャンパス、あるいは箕面キャンパスに一部のノードを拡張していくことで、無停止で運転することも可能になっていく。4.3.1 節に記載したが、試験的に導入したサービスであり、利用者の利用状況に応じて拡張を考えている。本学だけでなく我が国の学術研究の現場では、学術研究の再現性を担保し研究不正を防止する観点から研究データの適切な管理が求められている。本センターでは、そういった観点からもご利用いただけるよう機能拡張・増強も計画をしている。

4.3.4 NextCloud

図 25: NextCloud によるウェブインタフェース

NextCloud[29]はオンラインストレージソフトウェアである。読者の皆様の多くも研究室などでお使いかと思うが、本調達では、このオンラインストレージを利用者とのインタフェースとして採用している。

図 25 は著者が実際に NextCloud で自身のストレージにアクセスした際のスナップショットである。利用者はブラウザから自身の並列ファイルシステム内に置かれたデータを閲覧できる。また、外部ストレージ連携機能によって、外部のクラウドサービスや上述の HyperStore オブジェクトストレージを連携

させておくことによって、それらのストレージの違いを意識することなく、利用者はデータにアクセスすることが可能である。

図 26: スマートフォンを用いた NextCloud の利用例

また、NextCloud はブラウザ上で動作するため、お手元のスマートフォンからもアクセス可能である（図 26）。それゆえ、例えば、ジョブ終了後にスケジューラからジョブ完了通知をメール受信するように設定しておくことによって、ジョブ完了後に即座にブラウザを通じて ONION のインタフェースである NextCloud にアクセスすることで、その計算結果を確認することも可能となる。

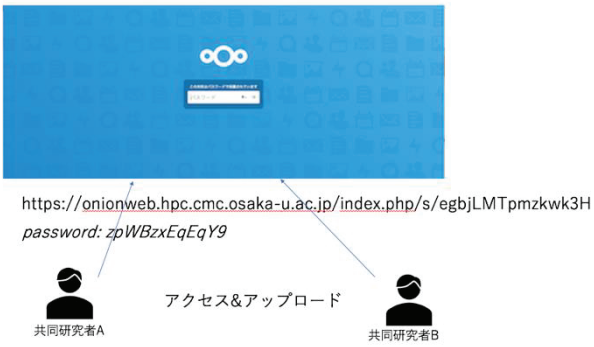

図 27: スーパーコンピュータを利用していない研究者へのデータ共有例

さらに、特定のファイルを、スーパーコンピュータのアカウントを利用していない外部の研究者に対してデータを共有することも可能である。図 25 の例では、公開・共有すべき特定のデータを選択し、外部の方にアクセスしてもらうデータの所在を示す URL を発行した上でパスワード設定を行なっているが、外部の研究者の方にこの URL とパスワードを発行することで、簡単に外部の研究者に公開・共有させることができる（図 27）。もちろん、この設定もスマートフォンから可能であるので、外出先でも研究者に計算結果後即座に情報共有が可能となる。この仕組みを利用することで、得られた計算結果を即座に外部の研究者と共有可能となったり、逆に、外部の研究者に入力データを特定のディレクトリにアップロードしてもらうことも可能となる。

このように NextCloud は直感的でわかりやすいウェブインタフェースを提供するとともに、背後にあるストレージの地理的分散状況や違いを隠蔽し、利用者のデータ操作を容易にしている。この NextCloud も試験的に導入したものであるが、是非積極的にお使いいただければ幸いである。

4.4 クラウドバースティング機能

SQUID では、オンデマンドに民間のクラウドサービスが提供するクラウド資源をオンデマンドに利用可能とするクラウドバースティング機能を配備している。この機能は、OCTOPUS の定常的な高負荷状態に鑑み、本センター、NEC、マイクロソフト、オラクルとの連携・協働により、OCTOPUS 上でプロトタイプ実装されているが、本調達で正式に運用レベルでサポートできるよう配備した。

SQUID では、このクラウドバースティング機能を通じて、マイクロソフト社の提供する Azure、および、オラクル社の提供する OCI を利用可能としている。本センターでのクラウドバースティング機能の運用は、(1)SQUID の高負荷時のクラウドへの負荷オフロードによる待ち時間の縮減、(2)SQUID 運用期間中に発表された新製品の評価を目的とした早期実戦配備を目的としている。

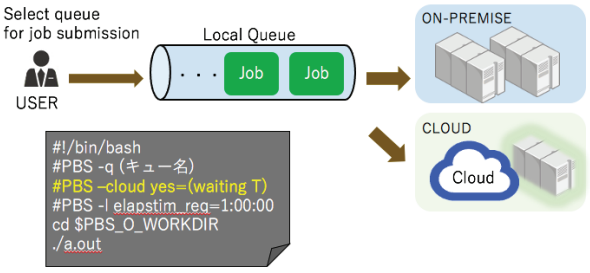

図 28: クラウドバースティング機能

図 28 に SQUID に導入されたクラウドバースティング機能の概念図を示す。通常では、大学の計算機センターのスーパーコンピュータにはスケジューラが配備されており、利用者のジョブ要求をキューに格納する。計算資源の利用状況を見ながら、ジョブ要求に対してオンプレミス、すなわち本センター配置のスーパーコンピュータの計算資源を割り当てていく。しかし、本センターでは SQUID が高負荷状態になった際には、クラウド環境上に本センターと同一のソフトウェアスタックをもつ計算ノードを動的に配備し、利用者にクラウド計算ノードを提供することを想定している。その際、本センターに配備されたクラウドバースティング機能では、利用者から投入されたジョブ要求の中からクラウド利用を承認されているジョブを順次クラウド資源に投下していく。これにより、システム全体の高負荷状態を緩和し、利用者のジョブ完了までの待ち時間を低減させる。また、SQUID 運用中に新たな製品が発表されクラウドで利用できるようになった際には、上記のようなクラウドバースティングキューだけでなく、クラウド専用キューを配備し、当該新製品を利用可能なクラウドノードとして利用者に早期展開する企画やイベントを想定している。

クラウド資源は、オンプレミス資源と比較して、同程度の性能をもつクラウド資源はどうしても価格が高い。それゆえ、クラウド資源の動的配備は、相当のコストが要求され、本センターの財布事情によるところもある。そのため、いつも利用できるというわけではないが、機会があれば利用を検討いただければ幸いである。

4.5 セキュアコンピューティング

サイバーメディアセンターは、全国の大学の研究者を対象として、その学術研究・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スーパーコンピュータを提供している。そのため、運用されるスーパーコンピュータは、多数の研究者によって共有利用される。このような共有型のスーパーコンピュータで、医学、薬学、バイオなど高い秘匿性や機密性が要求されるデータを扱う科学分野のシミュレーションや解析計算を行うのは非常に難しい現状がある。そのような高秘匿性・機密性が要求されるような分野の科学研究では、扱うデータに対する高いセキュリティ要求があるのが一般的である。一方、3.2 節に記載したが、今日の高性能データ分析分野への期待と関心の急速な高まりは、高いデータセキュリティが要求される科学研究分野におけるスーパーコンピュータの利用要求を高めている。

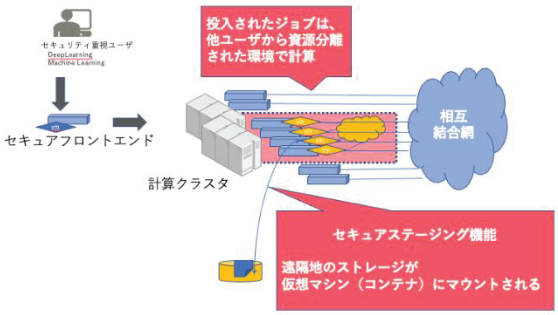

図 29: セキュアコンピューティング概念

本センターでは、このようなセキュアコンピューティングの実現にむけ、長らく研究開発を行ってきた。このセキュアコンピューティングは、そのような研究開発の経験と成果をもとに、SQUID に試験的に導入した機能である。図 29 に SQUID で導入されているセキュアコンピューティング環境の概念図を示す。セキュアコンピューティング機能は、SQUID に設けられたセキュアフロントエンド上から利用可能としている。なお、本特集記事執筆時点においては、このセキュアコンピューティング機能の利用は、本センターへの個別相談が必要となる。

以下、セキュアコンピューティング機能がどのように動作するかについて説明する。セキュアコンピューティング機能を利用する利用者グループには、セキュアフロントエンド上でコンテナ仮想化技術によって当該利用者グループだけがログインできる占有型フロントエンドノードが準備される。利用者はセキュアフロントエンドにログインし、スケジューラにジョブ要求を行う。この際、利用者のジョブ要求はコンテナジョブ（コンテナ仮想化技術により利用者個別の環境を計算ノードに配備し、その環境上でジョブを実行する）で行われることを前提とする。

このジョブが実際に計算ノード群に割り当てられると、当該ノード群上で実行されるコンテナは相互結合網から論理的に切り離され、それらコンテナ群だけを接続したネットワークが形成（セキュアパーティショニング機能：コンテナ間にオーバレイネットワークが形成）され、当該利用者のジョブのみが実行されるセキュアコンピューティング環境が動的に形成される。ジョブ実行が終了すると、当該利用者のためのセキュアコンピューティング環境は解消され、利用されたコンテナやデータもデータ流出防止の観点から計算ノードから削除される。

さらに、上記のセキュアコンピューティング機能に加えて、データを本センターに持ち出したくないという利用者向けにさらに高度なセキュリティ機能であるセキュアステージング機能を設けてある。この機能は、簡単にいうと、ジョブが割り当てられた計算ノードに対してのみ、そのジョブ実行中に遠隔地のストレージにアクセス可能とさせる。例えば、遠隔地のストレージに計算に利用するコンテナおよびデータを設置しておくことで、ジョブ実行時にジョブが実行される計算ノードにストレージがマウントされ、計算が行われる。計算結果はそのストレージに書き込まれ、計算が終わると計算ノードから切り離される。

このセキュアコンピューティング機能およびセキュアステージング機能は、SQUID でも最もチャレンジングな技術である。本センターでは大阪大学歯学部附属病院と S2DH (Social Smart Dental Hospital) [30] という共同研究を推進していることもあり、本特集記事執筆時点においては、歯学部附属病院の研究グループが、まず、最初の利用者になる予定であり、その準備を進めている。今後機会があれば、またこのセキュアコンピューティング機能およびセキュアステージング機能について報告したいと思う。

4.6 テイラーメイド型コンピューティング

3.2 節に記したが、今日の利用者の計算要求・ニーズは多様かつ複雑になっており、利用者によって計算方法も多様に異なる。それゆえに、利用するソフトウェアも多様に異なり、同一のソフトウェアであったとしてもそのバージョンが異なるという場合が多々ある。そうした場合、共有型のスーパーコンピュータでは、ある利用者は使いやすく利用できるが、他の利用者にとっては別途ソフトウェアのインストールが必要だったり、あるいは、ソフトウェアの組み合わせ的な問題や不具合から利用できない場合がある。特に、AI、ML といった分野においては、そのソフトウェア開発状況が活発であり、ソフトウェアやツールのバージョンが頻繁にアップデートされる。それゆえに、利用者の多様な計算ニーズを収容したいという管理者側の視点からもまた困った問題である。

この問題を解決するために、SQUID では、コンテナ仮想化技術 Singularity[31]を採用し、利用者が各自の計算環境を“テイラーメイド”できるよう設計している。読者の皆様方のなかには、なぜ Docker じゃないのか？コンテナ仮想化のデファクトといえばコンテナだろう？と思われる方もおられるかもしれない。確かにコンテナ仮想化といえば今日では Docker であり、Docker Hub などで気軽に利用できるコンテナイメージが数多く公開されていることも本センターでは把握している。しかし、実際運用の視点からは、Docker を利用者の皆様方にお使いいただくのは難しい。

Docker の利用については、OCTOPUS において利用できるよう環境を整備し、運用レベルで利用可否について検証した[32]。Docker の利用については技術的には可能であることは確認できているが、Docker をどのように利用者の皆様にご利用いただくのか（どのような環境を提供することで使いやすくなるのか？）ということだけでなく、セキュリティ上の問題をどのようにクリアするのか？という問題がある。具体的には、Docker の利用は管理権限を利用者に与えなければならない。このことは研究室レベルでの運用では問題とならないかもしれないが、センターレベルでの運用となると、利用者全員に対して管理権限を付与してしまうことは現実的ではない。一方で、Singularity は Docker のイメージを利用可能であることに加え、管理権限を利用者に付与することは必要ではない。そのため、HPC 分野では

Docker よりも Singularity を利用するケースが多々見られる。

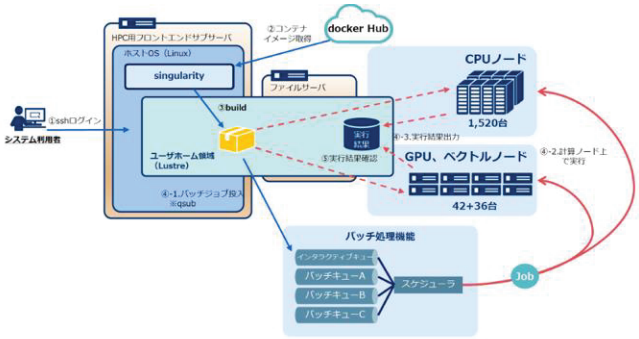

図 30: SQUID での Singularity 利用例

図 30 に SQUID での Singularity 利用例を示す。利用例は、SQUID の 3 種あるうちの 1 つの HPC 用フロントエンドノードからの利用例を示したものである。まず利用者は、ssh によってフロントエンドノードにログインする（図中①）。このログイン方法は、通常の Singularity を利用しない方法と同様である。その後、例えば、Docker HUB で公開されているコンテナイメージを取得し（②）、singularity でコンテナをビルドする（③）。その後、コンテナを利用しないジョブと同様に、qsub コマンドを利用して、Singularity コンテナジョブを投入し（④）、計算がおわると実行結果を確認できる（⑤）。Singularity の利用方法詳細については、本センターの大規模計算機システム事業 Web ページ[33]をご覧いただければ幸いである。

4.7 HPC・HPDA 融合

SQUID では、従来型の高性能計算向け利用者のための HPC 用フロントエンドノードに加え、高性能データ分析向け利用者のための HPDA 用フロントエンドノード、および、上述した高度なデータセキュリティを必要とするセキュアフロントエンドノードの異なる 3 種のフロントエンドを用意している。HPC 用フロントエンド、および、HPDA 用フロントエンドにそれぞれ 4 基、セキュアフロントエンドノードは 1 基を配備している。

HPC 用フロントエンドノードには、NEC LX112Rj-2 という型番のサーバ（図 31）を導入している。HPC 用フロントエンドノードの構成は表 6 に示す通りである。その名前の通り、OCTOPUS 同様に HPC 向けの開発環境を準備しており、エディタを利用してプログラムを開発し、各種コンパイラを用いてコンパイル、スケジューラへのジョブ投入・ジョブ確認といった一連の処理が実行可能である（図 32）。SQUID では、OCTOPUS で導入されているスケジューラ NQS II よりも新しいバージョンである NEC 製スケジューラである NQS V が導入されているが、基本的な利用方法については OCTOPUS とほぼ同様である。もちろん NQS V の機能拡張にあわせて利用方法も変更となっている部分は存在する。この辺りについては、Web ページ [34] に記載してあるので参照されたい。

図 31: HPC 用フロントエンドノード NEC LX 112Rj-2

表 6: HPC 用フロントエンドノード NEC LX 112Rj-2 の構成

プロセッサ	Intel Xeon Platinum8368 (2.4Ghz, 38 コア) x 2
メモリ構成	256 GiB (16GiB DDR4-3200 DIMM x 16)
ローカルストレージ	960GB SSD x 2
インタフェース	InfiniBand HDR x 1, 10GBase-SR x 2, 1000Base-T x 1, BMC
GPU	NVIDIA Quadro RTX6000 x 1
OS	Red Hat Enterprise Linux 8.2

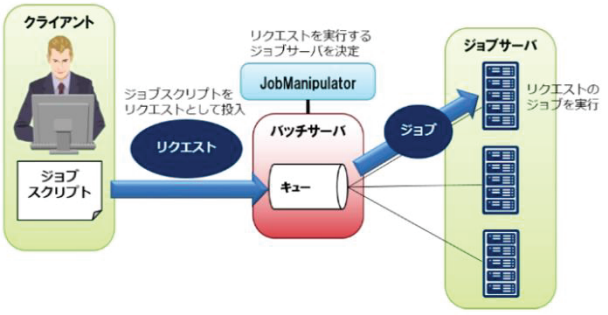

図 32: ジョブスケジューラを通じたジョブ投入

HPDA 用フロントエンドノードには、HPC 用フロントエンドノードと同様に、NEC LX 112Rj-2 という型番のサーバを導入している。表 7 に HPDA 用フロントエンドノードの性能を示す。表に示されるように、プロセッサ、ネットワークインタフェースは同様の構成であるが、メモリ構成は HPC 用フロントエンドと比べて 2 倍のメモリ容量となっている。またローカルストレージは SSD ではなく HDD となっている。さらに、HPC 用フロントエンドノードで搭載されていた GPU を HPDA 用フロントエンドノードでは搭載してない。これは、HPDA 用フロントエンドノードでは、NICE DCV Server による遠隔可視化を想定しないことによる。

SQUID では、この HPDA 用フロントエンドノードを通じて、利用者ごとに Jupyter notebook コンテナを起動し、利用できるよう設計している。この Jupyter notebook コンテナ上で Python を用いて試行錯誤的なデータ分析ができる環境を準備している（図 33）。これは AI、ML を扱う高性能データ分析分野の研究者は、従来の高性能計算分野の研究者のようにバッチ処理ではなく、試行錯誤的にインタラクティブ処理を希望していることに基づいている。この Jupyter コンテナを利用し、インタラクティブな処理をできるように設計している。ただし、現状では、計算ノードを用いた大規模なデータ分析を行う際には、高性能データ分析処理を行う場合にはスケジューラを通じたバッチ処理を行なっていただく必要がある。これは現時点では、限られた計算資源に対する計算要求を、HPC 利用者、HPDA 利用者で平等に扱いたいと考えたことによる。OCTOPUS では、待ち時間が数日以上に及ぶこともあり、その際には現状利用者の皆様には、スケジューラのスケジュール方針に基づいて平等に待っていただかざるを得ない。それゆえ、HPDA 利用者がインタラクティブ処理を要求することを理由に、HPC 利用者よりも優先的にジョブを割り当てることはむずかしい。この点については、今後の運用を通じた利用負担金や利用率の変化を見ながら、検討をしていきたいと考えている。

表 7: HPDA 用フロントエンドノード NEC LX 112Rj-2 の性能

プロセッサ	Intel Xeon Platinum8368 (2.4Ghz, 38 コア) x 2
メモリ構成	512 GiB (32 GiB DDR4-3200 DIMM x 16)
ローカルストレージ	1TB SAS HDD x 2
インタフェース	InfiniBand HDR x 1, 10GBase-SR x 2, 1000Base-T x 1, BMC
OS	Red Hat Enterprise Linux 8.2

図 33: HPDA フロントエンドの利用例

最後にセキュアフロントエンドノードについて記しておく。このセキュアフロントエンドノードには、HPC および HPDA 用フロントエンドノード同様、NEC LX 112Rj-2 という型番が付与されている。このサーバの外観は図 31 と同じである。表 8 に構成を示す。この構成は HPC フロントエンドノードとほぼ同様であるが、搭載されているストレージが HDD である。4.5 節で記したように、このセキュアフロントエンドノードから投入したジョブはセキュアな計算環境でジョブ実行が可能となる。

表 8: セキュアフロントエンドノード NEC LX 112Rj-2 の構成

プロセッサ	Intel Xeon Platinum8368 (2.4Ghz, 38 コア) x 2
メモリ構成	256 GiB (16 GiB DDR4-3200 DIMM x 16)
ローカルストレージ	1TB SAS HDD x 2
インタフェース	InfiniBand HDR x 1, 10GBase-SR x 2, 1000Base-T x 1, BMC
OS	Red Hat Enterprise Linux 8.2

5. 利用のために

本節では、これまでに説明した SQUID を実際に利用者の皆様がご利用いただくための参考情報を記す。ここまで読んでいただけた読者の皆様方には、ぜひ SQUID のご利用をご検討いただければ幸いである。なお、ここに記載する内容は、本センターの大規模計算機システム（スーパーコンピューティング）事業の Web ページにも記載しているので、詳細については当該 Web ページを参照いただければ幸いである。

5.1 利用負担金

利用負担金についての考え方は、OCTOPUS の特集記事[2]にまとめた。が、しかし、利用負担金についてのご意見やご相談は、利用者の皆様から頻繁にいただく。大規模計算が必要不可欠となっている研究者の方にとっては利用負担金の件は最大の関心事であろうと思う。ここでは、この利用負担金についての考え方を利用者の皆様方にご理解いただけることは本センターにとっても極めて重要であるので、以前の OCTOPUS の特集記事の際に記載した内容をベースに再度記載する。ぜひ本センターでの利用負担金の考え方にご理解いただき、利用者の皆様方のご支援をいただければ幸いである。

多くの利用者の方の関心が高いのは、計算機利用に伴う負担金であろう。事実、現利用者の方からも負担金についての多くの質問・相談・要望が寄せられている。とりわけ、多く寄せられる要望としては、利用負担金がもう少し安くならないか？あるいは、無料にならないのか？というものである。

スーパーコンピュータ上で大規模なシミュレーションやデータ解析を行う研究者にとっては、どうしても利用負担金が大きくなってしまう。このことは容易に想像できる。スーパーコンピューティング事業を担当する者としても、研究者としても、上記の利用負担金に対する要望は痛いほどよくわかる。しかし、言い訳ではないが、サイバーメディアセンターの利用負担金は、計算機利用に相当する消費電力量に基づいており、できるかぎり低額な設定としている。

スーパーコンピュータの保守・運用には、スーパーコンピュータシステムをただ設置し「はい、どうぞ」と置いておくだけではだめで、ハードウェア保守、ソフトウェア整備、テクニカルサポート等の費用も発生する。サイバーメディアセンターのスーパーコンピュータの調達においては、スーパーコンピュータを構成するハードウェア機器だけでなく、ハードウェア保守、整備・サポートに関する人件費等の運用・保守にかかる費用の多くを含めて調達を行なっている。そのため、サイバーメディアセンターのスーパーコンピュータの利用負担金は、利用者の計算機利用に伴う消費電力量相当のみを負担いただく構造となっている。言い換えれば、スーパーコンピュータを運転するのに必要最低限となる費用だけを利用者に負担いただいている。この利用負担金をゼロにするためには、あらかじめ調達に消費電力を見込んで含めておく、あるいは、どこからかそのための予算をとってくるしかないのが現状である。

ここで、前者でいいのではないか？と思われる読者の方もおられるかもしれない。その場合、消費電力分相当の費用をスーパーコンピュータシステム調達より取り置いておかなければならなくなり、調達できるスーパーコンピュータの規模は小さくならざるを得ない。場合によっては、スーパーコンピュータではなく、ただのコンピュータになってしまい、利用者の計算要望やニーズを充足することはさらに難しくなってしまう。今日科学研究に伴うデータ量

はますます大きくなり、計算要求・ニーズもまたますます大きくなりつつある。結果として、より大規模なシミュレーション、データ分析が可能なスーパーコンピュータの必要性･重要性が高まりつつある。このことを鑑みれば、計算機を利用した利用者自身に消費電力相当分を利用負担金として負担いただく受益者負担が最善ではないだろうか。

電気、水道等の生活ライフラインのように、科学研究のインフラであるスーパーコンピュータの運転にもどうしても費用が発生する。この費用については誰かが払わなければならない構造である。この構造をご理解いただき、SQUID をご利用いただければ幸いである。

5.2 SQUID 利用負担金

SQUID の利用負担金について説明する。表 9 に設計・策定した負担金制度の概要を示す。SQUID の利用負担金では、OCTOPUS 同様に、SQUID ポイントを導入している。SQUID もまた OCTOPUS と同様に、プロセッサ、アクセラレータの構成が異なる複数のノード群より構成されるハイブリッド型スーパーコンピュータである。そのため、それぞれのノード群ごとに消費電力が異なるため、原則に従えば異なる利用負担金単価となるのが適切であるが、利用者の皆様方にノード毎に契約いただく方式をとることはかなり不便である。そこで、SQUID では、SQUID ポイントを採用することで、利用者は事前にノード種別を確定しておく必要がなく、臨機応変に複数のノードを使い分けることができるよう利用負担金を設定した。

設計した利用負担金制度は、ノードを一利用者がある一定期間を予約して占有利用する占有､および、複数の利用者が時間分割で利用する共有の種別がある。占有利用は、現時点では、サイバーメディアセンターの計算拠点としての一角を担う革新的ハイパフォーマンス・コンピューティング・インフラストラクチャ(HPCI)[35]や、サイバーメディアセンターが参画する学際大規模情報基盤共同利用・共同研究拠点(JHPCN)[36]で採択された課題の利用を想定しており、多くの場合は共有利用となる。汎用 CPU ノードについては、利用状況に応じて占有利用での提供を検討して行く。

表 9: SQUID 利用負担金

(A) 占有	
基本負担額	占有ノード数
575,000 円/年	汎用 CPU ノード群 1 ノード
3,516,000 円/年	GPU ノード群 1 ノード
2,168,000 円/年	ベクトルノード群 1 ノード

(B) 共有		
	基本負担額	SQUID ポイント
コース	10 万円	1,000 ポイント
	50 万円	5,250 ポイント
	100 万円	11,000 ポイント
	300 万円	34,500 ポイント
	500 万円	60,000 ポイント

(C) ストレージ容量追加	
基本負担額	提供単位
2,000 円/年	HDD 1TB
5,000 円/年	SSD 1TB

表 10 にノード種別の消費係数を示す。表は 1 ノード時間あたりの消費係数を示している。SQUID では、OCTOPUS が非常に好評で非常に長い待ち時間となっていること、および、利用者からの利用負担金を高くしてでも優先度をあげてほしいという声があることを考慮して、通常優先度、高優先度、シェアの利用種別を設けている。通常優先度の場合は、OCTOPUS と同様の資源提供方法で、計算資源を時分割で利用者に利用いただく方法である。これを基準とし、シェアの利用種別では、自身のジョブが実行されている計算ノード上に他の利用者のジョブが割り当てられ実行される可能性がある。一方、高優先度の種別では、通常優先度、シェアと比較して、

高い優先度で計算資源の割り当てが行われる方法である。これに基づいて、高優先度、通常優先度、シェアの順に消費係数が多くなっている。

表 10: SQUID ノード種別消費係数

消費係数			
ノード群	高優先度	通常優先度	シェア
汎用 CPU ノード群	0.1873	0.1499	0.1124
GPU ノード群	1.1467	0.9174	0.6881
ベクトルノード群	0.707	0.5656	0.4242

例えば、汎用 CPU ノードの場合、1 ノードを 1 時間通常優先度利用した場合、0.1499 SQUID ポイントが消費される。言いかえれば、1000 ポイント（10 万円税別）で、CPU ノードの 1 ノードを 6,671 時間利用できる。一方、高優先度利用では、CPU ノードの 1 ノードは 10 万円で 5,339 時間利用できることとなる。本利用負担金の設計は、近い将来の SQUID の混雑時に、利用者による利用負担金を高くしてでも優先度をあげてほしいという声に対応したものである。

なお、本センターでは季節係数[37]を導入している。季節係数は、平成 31 年度以降の運用時に前年度の利用状況に基づいて、季節ごとに消費係数を変更することで利用者の計算計画を制御しようとすることを目的として導入した。この季節係数の仕組みは、消費係数から 1 未満の数字を乗算することにより、いわゆる利用負担金の割引を行う仕組みである。一般的に、年度末には、ジョブが多く投入され混雑する傾向にある一方、年度初めには比較的ジョブが走りやすい状況がある。これを緩和するために、例えば、年度初めころには消費係数を低くし、利用者の利用を年度初めに移動させることをねらう。OCTOPUS 同様に、SQUID の利用率等を加味しながら運用して行く予定としている。

5.3 公募利用制度

本センターでは、本センターの大規模計算機システムを活用する研究開発の育成・高度化支援の観点から、本センターが参画する「ネットワーク型」学際大規模情報基盤共同利用・共同研究拠点(JHPCN) [36]や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(HPCI) [35]の目的を踏まえつつ、今後の発展が見込まれる萌芽的な研究課題や本センターのスーパーコンピュータを最大限活用することで成果が見込まれる研究課題を公募する制度を 2016 年度より開始している。本制度には、それぞれ JHPCN および HPCI での課題として採択されることを目的とし、(1)若手・女性研究者支援萌芽枠、(2)大規模 HPC 支援枠を設定している。さらに、2019 年度からは、大規模な計算能力を必要とする人工知能分野の研究支援の観点から、(3)人工知能研究特設支援枠を設けている。本センターが設置する高性能計算機委員会において厳密な審査の上採択された研究課題については、本センターが研究課題推進のための利用負担金の全額を負担する。また、本センターの施設（会議室）もまた利用可能となる。ただし、その一方で、本センターが開催する公募型利用制度成果報告会での発表、成果報告、発表論文への謝辞記載等の義務も存在するので、その点はご留意いただきたい。

本制度は、例年、11 月半ば頃から 12 月半ば頃、3 月から翌年度の 4 月半ば頃までの期間の 2 期にわけて、翌年度の公募型利用制度への課題提案・応募を受け付けている。第 1 期の応募期間は、科学研究費の応募を終えた後に設定しており、研究者にとって応募しやすい（科研を提出した後に引き続き申請書を記載しないといけないが。）よう考慮している。また、2 期の応募期間は、HPCI および JHPCN への応募が不採択になった場合においても応募できる期間にしている。

本公募型利用制度の制度設計を行い、運営を行う筆者の立場からは、上記の公募利用制度整備趣旨をご理解頂ける方は、本センタースーパーコンピューティング事業ウェブページ内の公募型利用制度募集要項に従い、是非 SQUID を利用する課題の提案・応

募を積極的に検討いただければと思う。

5.4 HPCI & JHPCN

また、上述してきたように、サイバーメディアセンターは、学際大規模情報基盤共同利用・共同研究拠点(JHPCN) [36]や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(HPCI)[35]に参画している。HPCI では、国立研究開発法人理化学研究所のスーパーコンピュータを中核とし、国内の大学や研究機関の計算機システムやストレージを高速ネットワークで結ぶことにより、全国の HPC リソースを全国の幅広い HPC ユーザ層が効率よく利用できる科学技術計算環境を実現することを目的とする。また、JHPCN においては、北海道大学、東北大学、東京大学、東京工業大学、名古屋大学、京都大学、大阪大学、九州大学にそれぞれ附置するスーパーコンピュータを持つ 8 つの施設を構成拠点とし、超大規模計算機と超大容量のストレージおよび超大容量ネットワークなどの情報基盤を用いて、学際的な共同利用・共同研究を実施することにより、我が国の学術・研究基盤のさらなる高度化と恒常的な発展を目的としている。これらの枠組みを利用して、サイバーメディアセンターのスーパーコンピュータを利用することが可能である。

HPCI および JHPCN においても、前節 5.3 で記した公募型利用制度と同様に、課題申請書を記載の上、課題提案・応募を行う必要がある。応募された課題は、それぞれで設置された課題審査委員会で採否が決定される。採択された課題は、利用負担金を研究者が負担することなく利用することが可能となる。例年、JHPCN では 12 月半ば頃より 1 月初旬、HPCI では 3 月頃から 5 月初旬にかけて課題募集が行われている。本センターのスーパーコンピューティング事業からもウェブページやメーリングリストを通じて、これらの課題募集についてはアナウンスを行っている。HPCI および JHPCN に対して、SQUID を含む本センターのスーパーコンピュータを利用する研究課題を是非提案・応募して頂ければ幸いである。

6. おわりに

本特集記事では、2021 年 5 月に導入され、3 ヶ月間の試験運用を経て、2021 年 8 月から正式に運用を開始した SQUID（調達名称 : 高性能計算・データ分析基盤システム）について、その調達の背景、調達の方針、SQUID の概要と特徴、および、本センターのスーパーコンピュータを利用するための方法についてまとめた。本特集記事を依頼された頃は SQUID の導入がきまり、その具体的な設計を開始したか、始める直前のタイミングだったと思う。今回の調達では、本特集記事で記載したように、学術研究の広域化・グローバル化、高性能データ分析に代表される計算ニーズの多様化などを背景として、HPC・HPDA 融合、クラウド連動・連携機能、セキュアコンピューティング環境、データ集約環境、テイラーメイド型計算の 5 つのチャレンジを盛り込んだ。それゆえ、導入後の具体的なシステム設計については、これまで本センターで経験したことのない労力と時間が必要であり、本特集記事を時間までに完成できるかどうか不安であった。しかし、今なんとか本特集記事の終わりに近づき安堵しつつある。ただ、それでもやはり、今回の特集記事については徒然なるままに文章を書かせていただくのが精一杯で、十分に推敲する時間を割く余裕がなかった。それゆえに、読みづらい部分、説明が不十分である点もあろうかと思う。その点についてはご容赦いただければ幸いである。

最後に今一度記す。SQUID は、本センターが現時点で提供できる最高のスーパーコンピュータであると自負している。当然、できるだけ多くの研究者にご利用いただき、研究活動にお役立ていただきたいと考えている。SQUID に少しでも興味を持ってくださる方がおられたら、是非ご利用いただければ幸いである。SQUID がわが国の学術・産業の発展を支える研究基盤へと成長・発展してくことを願っている。皆様方のご支援がサイバーメディアセンター スーパーコンピューティング事業の発展につながっていく。是非とも、率直なご意見、ご要望等のフィードバックを引き続きお寄せ頂けるようお願いしたい。

参考文献

[1] SX-ACE, http://www.hpc.cmc.osaka-u.ac.jp/sx-ace-sys/.

[2] 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, Sep. 2018. [DOI:10.18910/70826]

[3] OCTOPUS, http://www.hpc.cmc.osaka-u.ac.jp/octopus/, 2021.

[4] 汎用コンクラスタ (HCC), http://www.hpc.cmc.osaka-u.ac.jp/hcc-sys/.

[5] OU ビジョン 2021 -社会変革に貢献する世界屈指のイノベーティブな大学へ-, https://www.osaka-u.ac.jp/ja/oumode/OUvision2021/y66s8j.

[6] Cyber HPC Symposium 2017, http://www.hpc.cmc.osaka-u.ac.jp/lec_ws/cyberhpcsympo-3rd/.

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

[8]Oracle Cloud Infrastructure, https://www.oracle.com/jp/cloud/.

[9] AWS, https://aws.amazon.com/.

[10] 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 ICT 推進協議会 2019 年度年次大会, 福岡, Dec.2019.

[11] Susumu Date, Hiroaki Kataoka, Shuichi Gojuki, Yiki Katsuura, Yuki Teramae, and Shinichiro Kigoshi, “First Experience and Practice of Cloud Bursting Extension to OCTOPUS”, 10th International Conference on Cloud Computing and Services Science, CLOSER2020, pp.448-455, May 2020. [DOI: 10.5220/0009573904480455]

[12] “増大するスパコン需要にこたえる新技術 パブリッククラウドを利用したクラウドバースティングを実装”, https://resou.osaka-u.ac.jp/ja/research/2019/20191115_2, Research at Osaka University, 2019.

[13]Software and Library (OCTOPUS), http://www.hpc.cmc.osaka-u.ac.jp/system/manual/octopus-use/software/.

[14]Welcome to How Wonderful Gets Done 2021, https://www.intel.com/content/www/us/en/events/how-wonderful-gets-done.html.

[15]SX-Aurora TSUBASA ラインナップ, https://jpn.nec.com/hpc/sxauroratsubasa/specification/index.html.

[16] TOP500, https://www.top500.org/.

[17]The LINPACK BENCHMARK, https://www.top500.org/project/linpack/.

[18]TOP500 LIST – JUNE 2021, https://www.top500.org/lists/top500/list/2021/06/.

[19] HPCG Benchmark, https://www.hpcg-benchmark.org/.

[20]HPCG – JUNE 2021, https://www.top500.org/lists/hpcg/list/2021/06/.

[21]GREEN500 LIST – JUNE 2021, https://www.top500.org/lists/green500/list/2021/06/.

[22] スーパーコンピュータシステム SQUID ラックデザイン コンテスト, http://www.hpc.cmc.osaka-u.ac.jp/squid-rack-design/.

[23] インテル アドバンスト・ベクトル・エクステンション 512 (Intel AVX-512), https://www.intel.co.jp/content/www/jp/ja/architecture-and-technology/avx-512-overview.html.

[24] NVLINK と NVSwitch, https://www.nvidia.com/ja-jp/data-center/nvlink/.

[25] NVIDIA Developer BLOG HGX A100 8-GPU connected with NVSwitch for fastest time-to-solution, https://developer.nvidia.com/blog/introducing-hgx-a100-most-powerful-accelerated-server-platform-for-ai-hpc/.

[26] OU ビジョン 2021 -社会変革に貢献する世界屈指のイノベーティブな大学へ-, https://www.osaka-u.ac.jp/ja/oumode/OUvision2021.

[27] DDN EXAScaler, https://ddn.co.jp/products/storage_software/exascaler.html.

[28] “Cloudian HyperStore Technical Whitepaper”, Cloudian Inc., 2015. (CLO-WP-13002-EN-02)

[29] NextCloud, https://nextcloud.com/.

[30] Social Smart Dental Hospital (S2DH), https://s2dh.org.

[31] Singularity, https://sylabs.io/singularity/.

[32] 渡場康弘, 李 天鎬, 伊達 進, “OCTOPUS 上における既存サービスと共存した Docker による資源提供環境の構

築, 大学 ICT 推進協議会 2019 年度年次大会, pp.239-241, 福岡, Dec.2019.
[33] コンテナ利用方法(SQUID), http://www.hpc.cmc.osaka-u.ac.jp/system/manual/squid-use/singularity/.
[34] フロントエンド利用方法, http://www.hpc.cmc.osaka-u.ac.jp/system/manual/squid-use/frontend/.
[35] 革新的ハイパフォーマンス・コンピューティング・インフラストラクチャ (High Performance Computing Infrastructure), *http://www.hpci-office.jp/*.
[36] 学際大規模情報基盤共同利用・共同研究拠点 (Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructure), *https://jhpcn-kyoten.itc.u-tokyo.ac.jp/ja/*.
[37] 勝浦裕貴, 寺前勇希, 木越信一郎, 伊達進, “スーパーコンピュータ OCTOPUS の混雑緩和に向けた取り組み”, 大学 ICT 推進協議会 2020 年度年次大会, Dec.2020