

ABCIクラウドストレージサービスの構築と評価

谷村 勇輔^{1,a)} 滝澤 真一朗¹ 小川 宏高¹ 浜西 貴宏¹

概要: AI 橋渡しクラウド (ABCI) は世界最大規模の人工知能 (AI) 処理向けの計算インフラストラクチャである。ABCI は AI の研究開発を加速するオープン・プラットフォームであることを目指し、AI ソフトウェア資産の蓄積や相互利用、AI の開発や応用を促す大規模データの集積を可能にするため、コンテナ技術のサポート、数十ペタバイトの高速ストレージ等を提供している。そして、データの共有と活用をさらに進めるため、ABCI のデータハーバーとしての役割を担い、Amazon S3 互換インタフェースとエンタープライズクラスの通信・データ暗号化を備えた ABCI クラウドストレージサービスを構築した。評価試験を通して、本サービスは 32 台のクライアントノードによる同時アクセスにおいて、100% の Write 負荷で 3.1GiB/sec、100% の Read 負荷で 4.5GiB/sec の合計性能を提供できることを確認した。また、ABCI の外からのアクセス性能やストレージ側での暗号化有効時のオーバーヘッドを明らかにした。

1. はじめに

AI 橋渡しクラウド (AI Bridging Cloud Infrastructure, ABCI) は、国立研究開発法人 産業技術総合研究所 (以下、産総研) が構築し、運用している世界最大規模の人工知能 (AI) 処理向け計算インフラストラクチャである [1-3]。ABCI は世界トップレベルの計算能力とデータ処理能力を有し、産学官共同の AI 研究開発を加速するオープン・プラットフォームであることを目指している。その実現のため、様々な AI ソフトウェア資産を蓄積して相互に利用するためにコンテナ技術をサポートし、大量のデータを必要とする AI の開発や応用の実証研究を行うために大規模データの集積を可能にする数十ペタバイトの高速ストレージ等を提供している。ABCI の取り組みは成果を上げつつあるが、データの共有と活用をさらに進めるため、産総研では新たに ABCI クラウドストレージサービスを構築し、2019 年 10 月より運用を開始した。

ABCI クラウドストレージサービスは、国内の大学や研究所を結ぶ高性能なネットワークである SINET5 [4] に直結し、ABCI の「データハーバー」の役割を担う目的で構築された ABCI のサービスの 1 つである。本サービスにより、SINET5 を通して高速かつ安全にデータを収集・蓄積してそれらを共有すると同時に、ABCI の計算能力を用いて作られた高性能な汎用学習モデルを AI の活用現場であるエッジ側に配布することを可能にする。ABCI 内だけでなく、外部とのデータの転送や共有を容易にするため、

Amazon S3 [5] 互換のインタフェースを採用し、エンタープライズクラスの通信・データ暗号化機能等を備える。

本稿では、ABCI クラウドストレージサービスの構築の背景と基本設計について述べたのち、アカウント管理やアクセス制御等の実装について述べる。また、本サービスの基本性能として、ABCI の中からと外からのアクセス経路における各性能、暗号化有効時の性能、複数クライアントによる同時アクセス時の最大性能等について報告する。

2. ABCI のストレージサービスの概要

ABCI が従来提供してきたストレージサービスは以下の通りである。これらのサービスと新しく構築した ABCI クラウドストレージサービスの構成概要を図 1 に示す。

- 計算ノードに搭載されたローカルディスク
計算ノードには NVMe 接続の SSD が搭載されている。ユーザはジョブスケジューラを介して計算ノードにジョブを投入し、各ジョブに割り当てられた計算ノードにおいてローカルディスクを利用できる。利用可能容量はジョブ投入時に指定する資源タイプとして定められている。複数の計算ノードを用いたジョブを実行する場合は、このローカルディスクを用いて、割り当てられたノード全体にまたがる共有ファイルシステムを一時的に作成可能である。これは BeeGFS On Demand (BeeOND) [6] によって実現されている。
- 各ユーザのホームディレクトリ用のファイルシステム
各ユーザに割り当てられるホームディレクトリ用のファイルシステムは、開発や小規模テストをはじめとした日常的な作業での利用を想定し、高い IOPS 性能

¹ 国立研究開発法人 産業技術総合研究所

^{a)} yusuke.tanimura@aist.go.jp

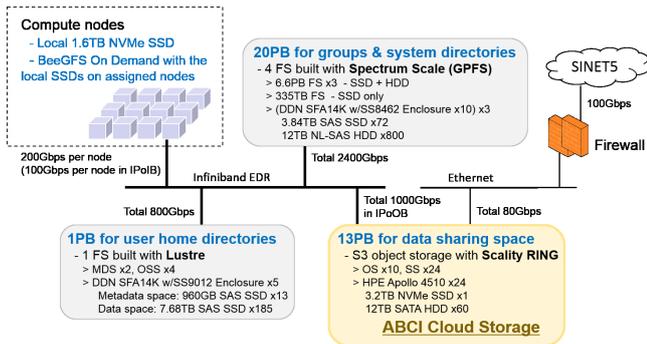


図 1 ABCI のストレージサービスの構成

を達成するため、All Flash のストレージと Lustre [7] を用いて構築されている。

- 各グループの共有ディレクトリ用のファイルシステム各グループに割り当てられる共有ディレクトリ用のファイルシステムは、グループまたは研究プロジェクト内での大規模なデータセットや共通ソフトウェアの共有のための利用を想定し、大容量のスペースを提供するため、SSD と HDD のハイブリッド構成を採用し、Spectrum Scale [8] を用いて構築されている。

上記の Lustre や Spectrum Scale は HPC 向けのシステムでよく用いられ、ABCI では、計算ノード同士の接続に用いている Infiniband を活用し、高い I/O 性能を提供するものであるが、ABCI 内での利用を想定している。外部との接続には、SSH 等のログイン経路を利用したり、ファイル転送のサービスを別途用意したりする必要がある。しかし、前者にはボトルネックが存在し、十分な性能を提供できない問題がある。また、POSIX のファイルシステムが提供するパーミッションレベルのアクセス制御だけでは、データの柔軟な共有と安全性の確保を両立させることは容易ではない。ABCI アカウントを持たない人とのデータ共有も想定する必要がある。さらに、AI に用いるデータセットの中には、ストレージサービスにエンタープライズクラスの安全性を求めるものもあり、データ保護のための暗号化やアクセス監視等にも高いレベルでの実現を要求する。これらの課題を総合的に解決するために構築されたのが、ABCI クラウドストレージサービスである。

3. クラウドストレージサービスの設計と実装

3.1 想定用途と設計方針

ABCI クラウドストレージサービスでは「データハバー」としての役割を担うために次の利用ケースを想定し、前節で述べた課題を解決できるように設計を行った。

- ABCI へのデータのインポート、および ABCI からのデータのエクスポート
- 任意の「データ利用グループ」の作成と「データ利用グループ」内でのデータ共有
- データの一般公開、パブリックリポジトリの提供

まず、データの転送や共有に関してクラウド等で利用されている既存ソフトウェア群を活用できるよう、Amazon S3 の互換インタフェースを採用した。Amazon S3 の成功により S3 のエコシステムは成熟しており、用途に合わせた S3 クライアントが利用可能である。S3 インタフェースをサポートするオブジェクトストレージは、HPC 向けの共有ファイルシステムほど高性能ではないが、スケールアウト可能なアーキテクチャにより、多数の同時アクセスに対して十分な性能を提供できる。

次に、ユーザが安全にデータをストレージに格納できるよう、クライアントとストレージ間の通信の暗号化とストレージ側での暗号化が可能な設計とした。また、ABCI 内部からのアクセスは計算ネットワークを活用して高速にストレージにアクセスできる経路とし、外部からのアクセスはファイアウォールを介して必要に応じて適切なアクセス制御と監視を行えるようにした。

アカウント管理は、ABCI におけるグループ単位のポイント購入とポイントによる利用料の支払いの仕組みを適用し、ABCI の既存サービスと同じような利用が可能となるよう、ABCI 本体のアカウント/グループ管理に組み込む設計とした。ただし、アクセス制御に関しては、グループ管理の体制を維持する一方、S3 の柔軟なアクセス制御の特徴との両立を図った。

以降では、これらの設計方針に基づく主な仕組みや機能の実装について述べる。なお、本サービスが基盤とするストレージ・ソフトウェアは Scality RING [9] である。

3.2 ネットワーク構成と暗号化

本サービスのデータ転送の経路図を図 2 に示す。本サービスへのアクセスは、ABCI の計算ノードからは Scality S3 Connector が動く 10 台のフロントエンドノード、ABCI の外からは Active-Standby 構成の HAProxy が動く 2 台のプロキシノードで受ける。前者は IPoB (IP over Infiniband) を利用し、DNS ラウンドロビンにより負荷分散がなされる。後者は、他のサービスと隔離された 10 Gigabit Ethernet を介して、プロキシノードからフロントエンドノードにアクセス要求が転送される。いずれにおいても、クライアントから本サービスまでの通信は HTTPS を利用する。

ストレージ側での暗号化は、Scality RING が提供する Server-Side Encryption (SSE) を利用し、Bucket レベルの暗号化をサポートする。これは Amazon S3 における SSE-S3 と類似しているが、Scality RING の独自 API によるものである。ただし、SSE-S3 において暗号化鍵が Amazon S3 で管理されるのと同様に、Scality RING によって暗号化鍵が管理され、ユーザからは透過的に暗号化機能が利用可能である。なお、ユーザは必要に応じて、AWS SDK 等による Client-Side Encryption (CSE) [10] を利用することもできる。

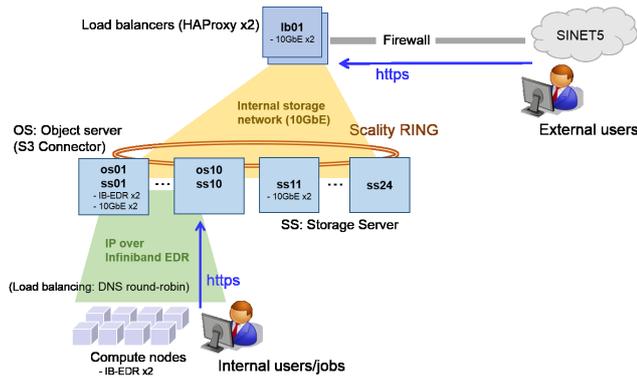


図 2 ABCI クラウドストレージのネットワーク構成

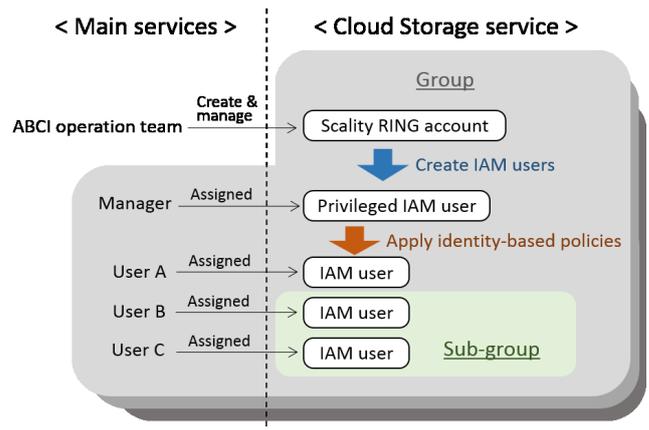


図 3 アカウント管理と ABCI 本体の管理との対応関係

3.3 アカウント管理

本サービスのアカウント管理と ABCI 本体の管理との対応関係を図 3 に示す。Scality RING では、Scality Vault により、AWS IAM (Identity and Access Management) と同様のアカウント管理機構が用意されている。本サービスでは、ABCI におけるグループを Scality RING のアカウント (AWS ルートユーザ相当) に 1 対 1 で対応させることとし、そのアカウントから作られる IAM ユーザをグループ内のユーザに割り当てる方式を採用した。ABCI のグループ管理者に割り当てる IAM ユーザには、アクセス制御について定義したアイデンティティベースのポリシーの適用やグループ内の他の IAM ユーザの削除等の特権を付与し、それ以外のユーザには S3 API による本サービスへのアクセス権限のみを付与した^{*1}。ただし、IAM ユーザの作成は ABCI の運用側でのみ行えるものとし、特権のある IAM ユーザであっても他の IAM ユーザを任意に作成できないようにした。すなわち、ABCI のユーザは自身に紐づく IAM ユーザのみを作成可能である。

これらを実現するため、グループによる本サービスの利用開始、IAM ユーザの作成、アクセスキーの生成はすべて ABCI 利用ポータルを介して行う実装とした。AWS のベストプラクティスに準じて、アプリケーション毎に IAM ユーザを用いられるよう、各 ABCI ユーザは複数の IAM ユーザを作成できるものとし、アクセスキーの更新を容易にするため、各 IAM ユーザは 2 つまでのアクセスキーを作成できるものとした。

本サービスの使用量は Scality RING の Service Utilization API (UTAPI) によって日単位で収集され、従量課金される仕組みとなっている。使用量情報は ABCI 利用ポータルや CLI によって確認が可能である。

3.4 アクセス制御機能の実装

Bucket や Object へのアクセス制御の方法としては、ACL

^{*1} ABCI ではグループを管理する者を利用管理者、ここで割り当てる IAM ユーザをクラウドストレージアカウントと呼ぶが、本稿では、ABCI のグループや AWS IAM との対応を明確にするため、これらの呼称を用いる。

による制御とアイデンティティベースのポリシーによる制御の 2 つをサポートし、データの共有や公開のための必要最小限の制御機能を実現した。前者は Bucket や Object を所有しているグループに所属するユーザが設定可能であり、後者はグループ管理者のみが設定可能である。なお、Amazon S3 では Bucket ポリシーと呼ばれるリソースベースのポリシーによる制御も可能であるが、Scality RING がサポートしていないため、本サービスでも利用できない。

ACL による制御では、誰に対して、どの操作を許可をするのかを Bucket や Object 毎に設定する。デフォルトでは、それを作成したユーザが所属するグループに対して Full-Control を許可する設定となっているが、他のグループ、本サービスの「すべての IAM ユーザ」、あるいは本サービスにアクセスできる「すべての人」に対して、Read, Write, Full-Control 等の操作権限を設定可能である。ただし、特定の IAM ユーザに対してのみ、ある操作を許可するといった細かな ACL は設定不可能である。

一方、アイデンティティベースのポリシーによるアクセス制御では、特定の IAM ユーザに対して、それぞれ操作権限を設定可能である。必要に応じて、グループ内のいくつかの IAM ユーザからなるサブグループを登録し、サブグループ毎に操作権限を設定することもできる。さらに、IP アドレス等を指定して、Bucket や Object にアクセスできるホストを限定することも可能である。

4. 基本性能評価

4.1 評価方法

本研究では、ABCI クラウドストレージサービスの基本性能として以下を明らかにするべく、評価試験を行った。

- ABCI の計算ノードからアクセスする際の性能
 - ABCI の外からアクセスする際の性能
 - 暗号化によるオーバーヘッド
 - 複数クライアントによる同時アクセス時の最大性能
- 評価試験には、オブジェクトストレージのベンチマーク

表 1 ABCI 計算ノードのスペック

CPU	Intel Xeon Gold 6148 (20 cores, 2.4GHz) × 2
GPU	NVIDIA Tesla V100 for NVLink (16GiB HBM2) × 4
Memory	384GiB
Network	Infiniband EDR (100Gbps) × 2
OS	CentOS 7.5

表 2 テスト用クライアントノードのスペック

CPU	Intel Xeon E5-2640 v4 (10 cores, 2.4GHz) × 2
Memory	256GiB
Network	10 Gigabit Ethernet × 2
OS	Red Hat Enterprise Linux 7.4

として作られた COSBench (Cloud Object Storage Benchmark) [11]を用いた。COSBenchのS3アクセスはAmazonが提供するSDKをベースとしている。そして、表 1 に示す ABCI 計算ノード、および表 2 に示すテスト用ノード上で COSBench の Driver を起動して性能を計測した。後者は ABCI データセンタ内のネットワーク上に存在するが、ABCI の外部からのアクセスと同様に、プロキシノード経由で ABCI クラウドストレージサービスにアクセスを行うサブネットワークに属している。両ノードとも Hyper-threading を有効としており、論理コア数は物理コア数の 2 倍である。

ABCI クラウドストレージを構成する主な機器のスペックは表 3 の通りである。先述のように、S3 のアクセスを受ける Scality S3 Connector (Object server) が動作するフロントエンドノードは 10 台であり、これらと別の 14 ノードとを合わせた計 24 ノードがストレージサーバ (Storage server) の役割を果たす。ストレージサーバ間、およびフロントエンドノードとプロキシノード間は他と隔離された本クラウドストレージ専用の内部ネットワークで接続されている。Scality RING のバージョンは 7.4.2 である。

これらの表に示す通り、クライアントノードやクラウドストレージのノードには Infiniband や 10GbE が 2 系統用意されている。しかし、前者は IPoIB 利用時は Active-backup の Bonding 設定となっており、後者も同様の Bonding 設定となっている。

4.2 単一ノードからのアクセス性能

単一ノードからのアクセス性能の計測結果を図 4 と図 5 に示す。この計測では、ノード内において Driver から 1 Worker のみを立ち上げて、指定したサイズの Object の Write, または Read のいずれかを 100% の負荷で実行した。グラフ中の「internal」は ABCI 計算ノードからのアクセスを表しており、「external」はプロキシノード経由のアクセスを表している。また、SSE による暗号化を有効にした

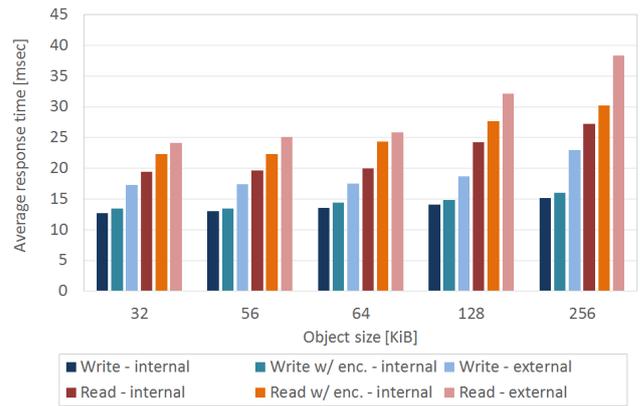


図 4 Object サイズが小さい時のアクセス性能

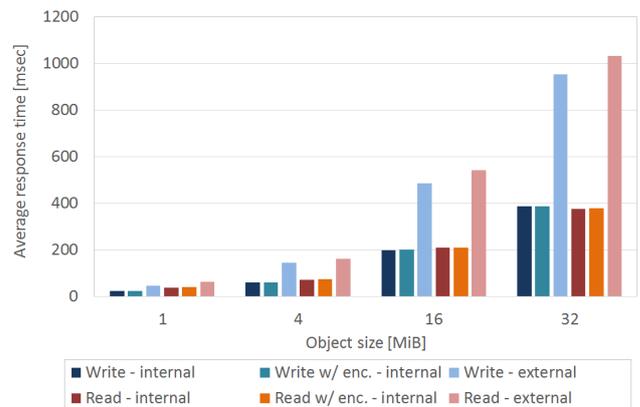


図 5 Object サイズが大きい時のアクセス性能

場合の計測も行った。

これらの結果より、まず、ABCI 計算ノードからのアクセスに比べてプロキシノード経由のアクセスの応答時間が長く、Object サイズが 4MiB 以上では、内部からのアクセスに比べて 2 倍以上の応答時間となることが分かった。次に、暗号化によるオーバーヘッドは Object サイズが小さいほど大きい傾向が見られ、最大で 21% (64KiB の Object の Read 時) であった。ただし、4MiB 以上の Object サイズではオーバーヘッドは 6.3% 以下となり、アクセス経路による応答速度の遅延よりもずっと小さいことが確認できた。また、本サービスでは Scality RING において、Object サイズが 60,000 バイト未満である場合には Replication、それ以上である場合には Erasure Coding (ARC9+3) を設定しているが、その境界前後において大きな性能差は見られなかった。

図 6 はノード内で立ち上げる Worker 数を論理コア数まで増やし、Write または Read のいずれかの負荷を一定時間かけた時の合計スループットのスケラビリティを確認した結果である。すなわち、Worker 数の上限値は ABCI 計算ノードでは 80、テスト用ノードでは 40 である。本試験においては、各 Worker がアクセスする Bucket は同じとし、1 回のアクセスでの Object サイズは 1MiB に固定

表 3 ABCI クラウドストレージサービスを構成する機器のスペック

Load balancer node	Dell PowerEdge R630 × 2
CPU	Intel Xeon E5-2640 v4 (10 cores, 2.4GHz) × 2
Memory	256GiB
Network	10 Gigabit Ethernet × 2
OS	RedHat Enterprise Linux 7.5
Object/storage server node	HPE Apollo 4510 Gen10 × 24
CPU	Intel Xeon Silver 4114 (10 cores, 2.2GHz) × 2
Memory	256GiB
Data disk	NVMe SSD (3.2TB) × 1 SATA HDD (7200RPM, 12TB) × 60, with HPE Smart Array P408i-p SR Gen10 controller
Network 1	10 Gigabit Ethernet × 2
Network 2 (Object server only)	Infiniband EDR (100Gbps) × 2
OS	Red Hat Enterprise Linux 7.5
Internal network switch	HPE FlexFabric 5940 (48XGT 6QSFP28) × 2

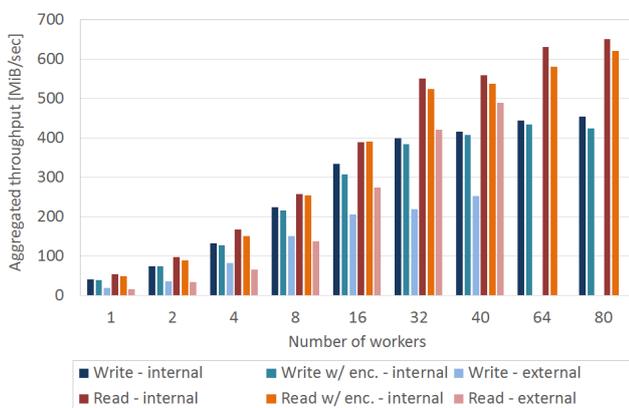


図 6 単一ノードからのアクセス性能

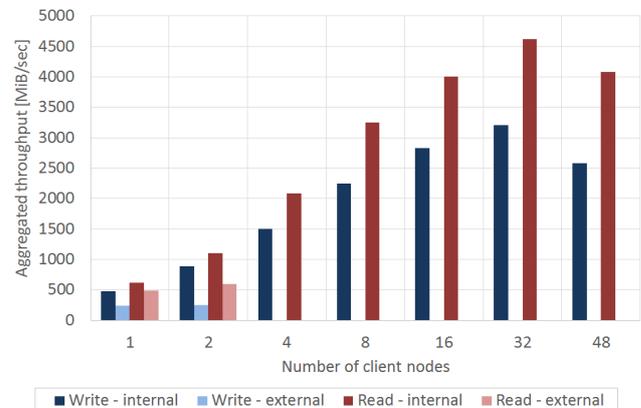


図 7 複数ノードからの同時アクセス時の性能

した。この結果より、ノード内で Worker 数を増やしてアクセスすることで、いずれのケースにおいても、より高いスループットが得られることを確認した。

4.3 複数ノードからの同時アクセス時の性能

図 7 はクライアントノード数を増やし、Write または Read のいずれかの負荷を一定時間かけた時の合計スループットのスケラビリティを確認した結果である。各ノードの Worker 数は論理コア数と同じに設定しており、ABCI 計算ノードを用いた場合には 80、テスト用ノードを用いた場合には 40 である。各クライアントノードがアクセスする Bucket はノード毎に異なるように設定し、1 回のアクセスでの Object サイズは 1MiB で固定した。

この結果より、ABCI の中からアクセスする際には、100% の Write 負荷、100% の Read 負荷のいずれにおいても 32 ノードまで合計スループットがスケールすることが確認できた。すなわち、Worker 数が 2,560 までスケールしている。一方、プロキシノードを経由したアクセスにおいては、クライアントノード数が 2 に増えた時点で合計スループットが若干上昇しているが、その伸びは小さい。これは現在、アクティブなプロキシノードの数が 1 であるためと

考える。今後の利用状況を鑑みて、ABCI の中からのアクセスと同じようにスケールさせる仕組みの導入を検討していきたい。

5. まとめと今後の課題

ABCI クラウドストレージサービスは、ABCI におけるデータの共有や公開の促進を目的としており、Amazon S3 の互換インタフェースや暗号化機能をサポートするとともに、S3 の柔軟なアクセス制御と ABCI 本体のグループ管理の体制とを両立させるアカウント管理を実現している。本稿では、本サービスのそうした設計や実装、そして評価試験による基本性能を報告した。

本サービスは既に運用を開始しているが、その目的を達成するためには、さらにいくつかの点に取り組む必要があると考えている。

まず、COSBench による評価試験だけでなく、実際にユーザが用いるクライアント・ソフトウェアを利用し、利用シナリオに沿った評価が必要である。著者らは別の研究 [12] において、S3 互換のオブジェクトストレージの可能性を調査する取り組みを進めており、同様の取り組みを ABCI において行っていきたい。特に、Worker 数を増や

すことで高いスループットが得られるため、マルチパートデータ転送の活用が重要であると考えている。また、ABCIで提供されている BeeOND 等と連携した効率的なデータステージングについても検討していきたい。

次に、データの共有や公開におけるアクセス制御についても実際的な評価が必要である。ABCI の外とのデータ共有や不特定多数へのデータ公開に関しては、セキュリティ対策や法的ルールに則った仕組みが必須であり、S3 が提供する技術的な特徴や課題を踏まえて、運用と実装の両面での評価と対策や改善の検討を行っていきたい。その上で、本サービスを用いて ABCI の内外のユーザに Public Datasets を提供することも考えていきたい。

謝辞 本研究の一部は、NEDO の委託業務「次世代ロボット中核技術開発プロジェクト」の支援を受けて実施した。また、本研究の一部は、産総研・東工大 実社会ビッグデータ活用 オープンイノベーションラボラトリ (RWBC-OIL) の活動として実施した。

参考文献

- [1] AI Bridging Cloud Infrastructure (ABCI): <https://abci.ai>.
- [2] 小川宏高, 松岡聡, 佐藤仁, 高野了成, 滝澤真一朗, 谷村勇輔, 三浦信一, 関口智嗣: 世界最大規模のオープン AI インフラストラクチャ AI 橋渡しクラウド (ABCI) の概要, 情報処理学会研究報告, Vol. 2018-HPC-165, No. 19 (2018).
- [3] 小川宏高, 松岡聡, 佐藤仁, 高野了成, 滝澤真一朗, 谷村勇輔, 三浦信一, 関口智嗣: AI 橋渡しクラウド- AI Bridging Cloud Infrastructure (ABCI) - の構想, 情報処理学会研究報告, Vol. 2017-HPC-160, No. 28 (2017).
- [4] Science Information Network 5: <https://www.sinet.ad.jp>.
- [5] Amazon S3: <http://aws.amazon.com/s3/>.
- [6] BeeGFS On Demand: <https://www.beegfs.io/wiki/BeeOND>.
- [7] Lustre: <http://lustre.org>.
- [8] IBM Spectrum Scale: <https://www.ibm.com/jp-ja/marketplace/scale-out-file-and-object-storage>.
- [9] Scality RING: <https://www.scality.com>.
- [10] Protecting Data Using Client-Side Encryption: <https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingClientSideEncryption.html>.
- [11] Zheng, Q., Chen, H., Wang, Y., Duan, J. and Huang, Z.: COSBench: A Benchmark Tool for Cloud Object Storage Services, *Proceedings of the 2012 IEEE Fifth International Conference on Cloud Computing*, pp. 998-999 (2012).
- [12] 谷村勇輔, 遊佐佳一, 高野了成, 浜西貴宏: AI・ビッグデータ処理におけるオブジェクトストレージを用いたデータステージングの評価, 情報処理学会研究報告, Vol. 2018-HPC-167, No. 14, pp. 1-7 (2018).